Daniel Schoberl, BSc.

Chemometrische Regressionsmodelle in der Infrarotspektroskopie

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

Masterstudium Technische Mathematik: Operations Research und Statistik

eingereicht an der

Technischen Universität Graz

Betreuer

Univ.-Prof.i.R. Dipl.-Ing. Dr.techn. Ernst Stadlober

Institut für Statistik

Graz, April 2018

Messen,
was messbar ist,
und messbar machen, was noch nicht messbar ist.

Galileo Galilei

Diese Arbeit wurde in Kooperation mit Anton Paar GmbH[1] erstellt.

[1] https://www.anton-paar.com/

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

Graz, am _____ _____

Datum Unterschrift

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Graz, _____ _____

Date Signature

Zusammenfassung

Die Infrarotspektroskopie, in Kombination mit der Fourier-Transformation (FTIR) und der Abgeschwächten Totalreflexion (ATR), ist eine äußerst zuverlässige und leistungsstarke Analysemethode zur Charakterisierung unterschiedlichster Stoffe. Basierend auf den Wellenzahlen des mittleren Infrarotbereiches werden in dieser Arbeit verschiedene chemometrische Modelle entwickelt, um unterschiedliche Inhaltsstoffe von Weinproben zu beschreiben und mit hoher Genauigkeit zu prognostizieren. Wir betrachten dafür den Alkoholgehalt, die Konzentrationen chemischer Verbindungen wie Glukose, Fruktose und verschiedener Säuren, sowie Eigenschaften der Weine wie Dichte und pH-Wert. Hierfür werden in einem ersten Schritt unterschiedliche Arten der Datenaufbereitung anhand der vorliegenden Spektren erläutert, dann wird ein modellunabhängiger, sowie heuristischer Algorithmus präsentiert, um die vorliegenden Prädiktorenanzahl (Anzahl der Wellenzahlen eines Spektrums) zu reduzieren. Im praktischen Teil liegt der Fokus auf der qualitativen Beschreibung der funktionalen Zusammenhänge mittels statistischer Verfahren, der Partial Least Square (PLS) – Regression und deren nichtlinearen Modifikation, den Kernel PLS – Modellen, sowie mithilfe von künstlichen, neuronalen Netzwerken. Anhand der zur Verfügung stehenden Stichproben wird der Einfluss von Störfaktoren wie beispielsweise der Temperatur der Weinprobe selbst, sowie der Luftfeuchtigkeit untersucht und auf mögliche Fehlerquellen hingewiesen. Weiters werden anhand dieser Resultate Aussagen über die Messqualität der in Entwicklung befindlichen Spektrometer der Firma Anton Paar GmbH abgeleitet.

Abstract

Fourier-transform infrared (FTIR) spectroscopy, combined with attenuated total reflection (ATR), provides an extremely reliable and powerful method for analyzing and characterizing a large variety of chemical substances. We establish a functional relationship between the absorption values, corresponding to the mid-infrared light, and various substances of wine as alcoholic strength, chemical concentrations like glucose and fructose and different acids. Moreover, wine characteristics like density or pH value are analyzed. The main goal is to reproduce and predict all these response variables with a high degree of accuracy. In a first step, different procedures of preprocessing the spectra are introduced and a heuristic algorithm is presented. It does not depend on a specific model type and selects relevant parts of the given wave number range. This is followed by practical chapters, in which different chemometric models are applied to the given data. We consider Partial Least Square (PLS) – regression with linear and non-linear kernels to improve the quality of predictions. Moreover artificial neural networks are used to establish benchmark values. Possible interactions between temperature or humidity are discussed and the residuals are analysed to detect disruptive factors. These results are used to evaluate the measurement quality of three different spectrometers, which are currently developed by Anton Paar GmbH.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Aufbau und chemisch-physikalische Hintergrundinformationen

Die vorliegende Arbeit untersucht funktionale Zusammenhänge zwischen Infrarotspektren von Weinen, bestimmt mit einem FTIR Spektrometer, und den zugehörenden Inhaltsstoffen und Eigenschaften der untersuchten Weine.

In Abschnitt 1.1 wird anhand des klassischen Michelson-Interferometers das Messprinzip der FTIR Spektroskopie skizziert und dessen Vor- und Nachteile erläutert, gefolgt von einem ersten Überblick über die vorliegende Datensituation in Kapitel 2. Hierzu werden zuerst alle Weinparameter (Responsevariablen) untersucht und in einen Kontext mit deren Ermittlung und deren Bedeutung gesetzt. Zusätzlich wird auf die Genauigkeit der vorliegenden Daten eingegangen, sowie die Korrelation der einzelnen Responses aufgezeigt. In weiterer Folge werden die Prädiktorvariablen in Abschnitt 2.2 diskutiert, insbesondere welche unterschiedlichen Arten an erklärenden Variablen zur Verfügung stehen (Spektren, Wiederholbarkeitsdaten, unterschiedliche Datensätze von verschiedenen Spektrometern). In Unterabschnitt 2.2.2 wird erläutert, welche dieser erklärenden Variablen bereits von vornherein von der Modellierung ausgeschlossen werden können, bevor sich der Teilbereich 2.3 mit der Datenaufbereitung beschäftigt und eventuell notwendige Adjustierungen beschreibt.

In Kapitel 3, der Modellvalidierung und -selektion, werden unterschiedliche Variationen der Kreuzvalidierung vorgestellt, welche das Ziel einer möglichst effizienten Modellfindung und -evaluierung verfolgen. Hierfür wird ein heuristischer Algorithmus für ebendieses Vorgehen präsentiert.

Die hierauf folgenden vier Kapitel gliedern sich in einen theoretischen, sowie einen praxisorientierten Teil für die vorgestellten (mathematischen) Modelle. Kapitel 4 erläutert die Grundlagen für die statistische Modellierung mit der (nichtlinearen) Partial Least Square Regression, während im anschließenden Kapitel 5 die Ergebnisse dieser Modelle für die gegebenen Responsevariablen spezifiziert und diskutiert werden. In Kapitel 6 wird der Ansatz der künstlichen neuronalen Netzwerke motiviert und die Funktionalität ebendieser komplexen Thematik vorgestellt, während das anschließende Kapitel 7 wiederum die Ergebnisse präsentiert.

Im Appendix ab Seite 207 werden ausgewählte Definitionen und Informationen, welche während der vorangegangenen Kapitel verwendet werden, detailliert beschrieben und runden diese Arbeit ab.

1.1 Chemisch-physikalische Hintergrundinformationen

Um Strukturen oder Zusammensetzungen von gewissen Stoffen zu eruieren, kann beispielsweise die Spektroskopie verwendet werden. Hierunter versteht man „das Analysieren der quantisierten Wechselwirkung von elektromagnetischer Strahlung mit Materie“[1]. Um konkrete Folgerungen zu ziehen, muss die Strahlung charakterisiert werden. Sämtliche Informationen über die allgemeine Spektroskopie sind aus [34] entnommen und für detailliertere Informationen für (speziell die organische) Spektroskopie wird an dieser Stelle auf ebendieses Werk verwiesen. Diese Charakterisierung kann grundsätzlich mit folgenden drei Beschreibungen erfolgen:

- Frequenz ν. Hierunter versteht man die Anzahl der Wellen, die durch einen fixierten Punkt binnen einer Sekunde hindurch strömt. (Einheit: beispielsweise Hertz Hz).
- Wellenlänge λ. Alternativ kann eine (periodische) Strahlung als Abstand zweier Peaks identifiziert werden. (Einheit: beispielsweise Nanometer nm).
- Wellenzahl $\bar{\nu}$. Eine weitere Charakterisierung kann durch den reziproken Wert der Wellenlänge erfolgen. (Einheit: beispielsweise cm^{-1}).

Diese Arbeit beschränkt sich auf den Infrarotbereich mit einer Wellenlänge von circa 2.5 µ ($\approx 4\,000\,\mathrm{cm}^{-1}$) bis ungefähr 15 µ ($\approx 667\,\mathrm{cm}^{-1}$), da in diesem Bereich eine hohe Anzahl von Absorptionen stattfindet.

Bei der elektromagnetischen Strahlung handelt es sich um Energie, weshalb die Moleküle durch das Zufügen dieser Strahlung diese Energie teilweise absorbieren können, indem die einzelnen Bindungen innerhalb eines Moleküls beispielsweise zu Schwingen oder sich zu Strecken beginnen. Mit Hilfe eines Spektrometers können diese Absorptionen gemessen werden. Unter Verwendung unterschiedlicher Strahlung im Sinne von unterschiedlichen Wellenzahlen entsteht ein sogenanntes Spektrum, mit welchem in dieser Arbeit versucht wird, mittels mathematischer Modelle Rückschlüsse auf die Zusammensetzung der Weine zu erhalten.

Messgerät

Beim verwendeten Messgerät handelt es sich um ein sogenanntes ATR-FTIR Spektrometer. Hierbei beschreibt FTIR (Fouriertransformations Infrarot Spektrometer) den Aufbau des Spektrometers und ATR (attenuated total reflection bzw. abgeschwächte totale Reflexion) das Probenahmeverfahren, wie in Abbildung 1.1 schemenhaft dargestellt.

Die Hauptbestandteile dieses Spektrometers bilden einerseits eine Lichtquelle, ein Strahlenteiler, zwei unterschiedliche Spiegel, die Probe mit zugehörendem Kristall sowie einem Detektor und einem Laser, welcher die Position des beweglichen Spiegels ermittelt. Diese Apparatur entspricht dem klassischen Michelson-Interferometer, wie beispielsweise in [7] beschrieben.

[1] Übersetzte Definition aus [34]

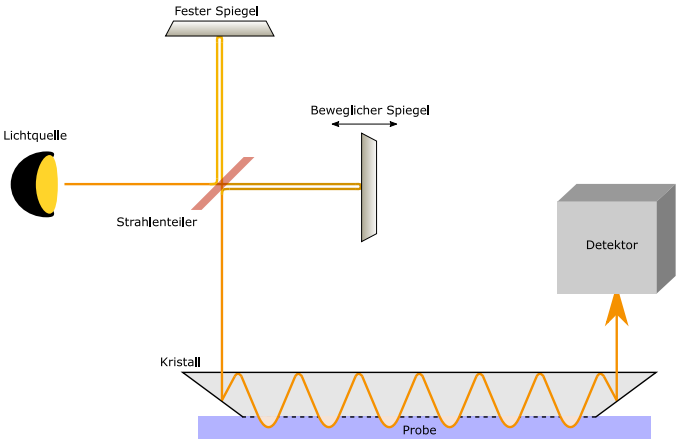

Abbildung 1.1: Schematische Darstellung eines ATR-FTIR Spektrometers

Der durch die Lichtquelle produzierte Lichtstrahl wird am Strahlenteiler in zwei unterschiedliche Lichtstrahlen unterteilt. Ein Teil des Strahls (hier durch eine dunklere Färbung des ursprünglichen Lichtstrahls) wird auf einen beweglichen Spiegel abgeleitet und zum Strahlenteiler zurückprojiziert. Der zweite Teil des Lichtstrahls (in der Grafik mit einer leicht helleren Färbung dargestellt) wird auf den festen Spiegel geleitet und ebenfalls zum Strahlenteiler zurückgeführt. Hier interferieren[2] die beiden Lichtstrahlen und werden durch das ATR-Modul in einen sogenannten Detektor weitergeleitet. Diese Interferenz hängt somit von der Position des beweglichen Spiegels ab und wirkt sich in weiterer Folge auf die im Detektor aufgezeichneten Werte aus.

Im Großen und Ganzen besteht das ATR-Modul aus einem Kristall[3], und der zu analysierenden Probe, welche sich in direktem Kontakt mit diesem befindet. Der in dieses Modul geleitete Lichtstrahl wird innerhalb des Kristalls, wie in Abbildung 1.1 skizziert, reflektiert und dringt in die Probe ein. Durch das Eindringen in die Probe können die chemischen Molekülverbindungen in Bewegung gesetzt werden und dies äußert sich in der Absorption der Infrarotstrahlung. Diese Verringerung der Intensität des Lichtstrahls bildet die Grundlage der ATR-Spektroskopie und wird im Detektor, zusammengefasst zu einem sogenannten Interferogramm, ermittelt.

Durch die Fourier-Transformation wird das Interferogramm, als Funktion in Abhängigkeit der Wellenzahlen, zu einem sogenannten Einkanalspektrum, transformiert (vgl. hierzu Abbildung 2.3 (o.)). Zusätzlich zu den Informationen über die gemessene Probe wird diese Darstellung von Eigenschaften des Spektrometers maßgeblich beeinflusst, wie beispielsweise durch Informationen über die Lichtquelle. Aus diesem Grund ist eine Referenzierung des Einkanalspektrums auf jenes einer Referenzprobe (in dieser Arbeit Wasser) notwendig und es resultieren die hier betrachteten Spektren.[4] Diese Referenzierung wird bei der Analyse der Prädiktoren in Abschnitt 2.2 erläutert.

[2] sich überlagern

[3] In dieser Arbeit wurde für die Spektrometer der Firma Anton Paar ein Zinkselenidkristall (ZnSe) verwendet.

[4] Vergleiche hierzu [12], Kapitel 2.

Die Verwendung dieser Methodik und des beschriebenen Spektrometers bringt insbesondere folgende Vorteile mit sich[5]:

- **Jaquinot-Vorteil**: Ein besseres Signal-Rausch Verhältnis[6] kann erzielt werden.
- **Fellget-/Multiplexvorteil**: Es wird nur ein geringer Zeitaufwand benötigt, da alle Frequenzen gleichzeitig gemessen werden.
- **Connes- bzw. Linearitätsvorteil**: Die Möglichkeit der präzisen Positionsbestimmung des beweglichen Spiegels mit einem Laser resultiert in einer hohen Wellenzahlengenauigkeit.

Das Signal-Rausch Verhältnis kann zudem durch die Anzahl der Reflexionen der ATR-Messtechnik verbessert werden. Zusätzlich kann hiermit grundsätzlich eine Vielzahl von unterschiedlichen Proben gemessen werden. Durch den direkten Kontakt des verwendeten Kristalls mit der Probe selbst kann es allerdings zu Kontaminationen kommen, weshalb beispielsweise die Probleme in Abschnitt 5.1 auf Seite 57 resultieren.

Mithilfe dieser Messmethode können probenspezifische Spektren ermittelt werden. Um die tatsächlichen Inhaltsstoffe der Weine zu schätzen, müssen diese Messungen in einen Zusammenhang mit den Bestandteilen der Proben gesetzt werden, was das eigentliche Ziel dieser Arbeit repräsentiert, da die Weinkomponenten nicht direkt aus den Spektren abgelesen werden können.

[5] Siehe [8], Kapitel 2.

[6] Verhältnis der Signalhöhe zu der Höhe des Rauschens. Eine zusätzliche Verbesserungsmöglichkeit stellt die Spektrenmittelung dar, vgl. Unterabschnitt 2.2.1

2 Datenvoranalyse

Für die Analyse der Funktionalität des ATR-FTIR Spektrometers stehen insgesamt 5 Datensätze zur Verfügung. Es wird folgende Unterscheidung angestellt:

- Datensatz des Jahres **2015**, V70(2015): Es stehen insgesamt $n_{2015} = 80$ analysierte Weinproben, gemessen im Jahr 2015 mit dem Spektrometer der Firma Bruker Corporation[1], dem V70, zur Verfügung.
- Datensatz des Jahres **2016**: An den vier unterschiedlichen Engines, E22, E24 und E25, sowie auf dem V70 werden $n_{2016} = 82$ unterschiedliche Weine[2] analysiert und deren Spektren erfasst.

Um die Daten in weiterer Folge identifizieren zu können, werden jene des Jahres 2015 mit V70(2015), jene des Jahres 2016, vermessen mit dem Vertex, durch V70(2016) bezeichnet.

Dieses Kapitel untersucht in einem ersten Schritt die Responsevariablen und zeigt mögliche Probleme auf, während sich der Unterabschnitt 2.2 den Spektren widmet, bevor abschließend die Datenaufbereitung erklärt wird.

2.1 Responsevariablen

Insgesamt stehen für die Weine folgende, unterschiedliche Inhaltsstoffe beziehungsweise Eigenschaften als Responsevariablen zur Verfügung:

- Ethanol in Vol.%,
- Extrakt, Glukose, Fruktose als unterschiedliche Charakterisierungen von Zucker in g/l,
- titrierbare und flüchtige Säuren als Klassifikation von Säuren, sowie die spezifizierten Säurearten Wein-, L-Äpfel-, Milch[3]- und Zitronensäure[4], je in g/l,
- der Zuckeralkohol Glyzerin in g/l,
- der pH-Wert,
- sowie die Dichte allgemein, Dichte Ethanol[5] und Dichte Extrakt[5], in g/ml.

[1] `https://www.bruker.com`

[2] Da anstelle des Weines mit der ID 21 ein anderer Wein gemessen wurde, welcher deswegen doppelt in den Daten vorkommt, reduziert sich die Kardinalität des Datensatzes auf $n_{2016} = 81$.

[3] Im Datensatz des Jahres 2016 wird zusätzlich zwischen L- und R-Milchsäure unterschieden.

[4] Die Werte sind nur für die Datensätze des Jahres 2016 verfügbar.

[5] Nur für die Datensätze des Jahres 2015.

Diese Werte wurden teilweise von Anton Paar selbst, oder durch eine zertifizierte Methode eines externen Labors bestimmt. Da sämtliche Spektren (Abschnitt 2.2) von Anton Paar selbst bestimmt werden, führt Letzteres dazu, dass die Spektren und die Inhaltsstoffe teilweise nicht mit derselben Weinprobe (aus der gleichen Weinflasche) ermittelt werden können. Für eine möglichst hohe Präzision wird allerdings auf gleiche Chargennummern geachtet. Die Variabilität der Messungen der Referenzwerte können Tabelle 2.1 entnommen werden.

Messgenauigkeit der vorliegenden Daten

Hierbei resultieren die Referenzwerte größtenteils aus einer enzymathischen Bestimmung, wie beispielsweise jene für die Äpfel- und Milchsäure, sowie für die Zuckerarten Fruktose und Glukose, während die Referenzwerte der Weinsäure auf kolorimetrischen Auswertungen beruhen.

			Variabilität	Einheit
Äpfelsäure			0.3	g/l
Ethanol			0.1	Vol.%
Flüchtige Säure			0.2	g/l
Fruktose	$<$	30	0.3	g/l
	$<$	110	3.0	
	$>$	110	5.0	
Gesamtzucker	$<$	30	0.4	g/l
	$<$	110	4.0	
	$>$	110	7.0	
Glukose	$<$	30	0.3	g/l
	$<$	110	2.5	
	$>$	110	5.0	
Glyzerin			0.2	g/l
Milchsäure			0.2	g/l
Titrierbare Säure			0.3	g/l
Weinsäure			0.3	g/l

Tabelle 2.1: Variabilität aller vorliegenden Referenzwerte. Diese entspricht der doppelten Standardabweichung der Residuen.

Bei der Variabilität in Tabelle 2.1 handelt es sich um die Genauigkeit, angegeben als doppelte Standardabweichung der Residuen. Dies bedeutet beispielsweise, dass die Residuen bei der Schätzung der Glukosekonzentration von Weinen mit einem Glukosegehalt von 30 g/l oder weniger, eine Standardabweichung von 0.15 g/l aufweisen. Auch wenn die Standardabweichung ein gefährliches Maß sein kann, wenn keine weitere Informationen über die Struktur etwaiger Residuenplots vorliegt, ist es üblich, die Genauigkeit mittels (doppelter) Standardabweichung

der Residuen auszudrücken. Für die Zitronensäure und den nicht in Tabelle 2.1 angeführten Eigenschaften (Dichte und pH-Wert) liegt keine Aussage über die Variabilität der Residuen vor. Bei der Modellierung des Extraktwertes werden die Variabilitäten des Gesamtzuckers als Vergleichswerte herangezogen.

Während Tabelle 2.1 die Genauigkeit im Sinne einer Variabilität wiedergibt, muss zusätzlich die Datengenauigkeit betrachtet werden. So liegen sämtliche Daten, mit Ausnahme von Ethanol, Extrakt, pH-Wert und Dichte mit einer Präzision von $10^{-1}\,\mathrm{g/l}$ vor. Während sich diese Genauigkeit des Ethanols, Extrakt- und pH-Wertes auf zwei Nachkommastellen erhöht, so finden sich Werte der Dichte mit einer Genauigkeit von $5 \cdot 10^{-6}\,\mathrm{g/ml}$ in den zur Verfügung stehenden Datensätzen. Dies ist aus mathematischer Sicht dahingehend von Bedeutung, weil grundsätzlich von kontinuierlichen Konzentrationswerten[6] ausgegangen wird, die geschätzt werden. Zudem wird der Wertebereich auf $\mathbb{R}_{\geq 0}$ beschränkt und daher von einem theoretischen Gesichtspunkt aus betrachtet, ist die Wahrscheinlichkeit einer Konzentration von $0\,\mathrm{g/l}$ oder $0\,\mathrm{Vol.\%}$ positiv, wohingegen die Wahrscheinlichkeit, zwei exakt gleiche Messwerte zu erhalten, vernachlässigbar ist. Aufgrund der stattfindenden Rundungen beinhalten beispielsweise 17 Weine eine Zitronensäurekonzentration von $0.2\,\mathrm{g/l}$.

In dieser Arbeit wird versucht, die abhängigen Variablen nachzubilden und man muss bedenken, dass zusätzlich zu dem Fehler der hier entwickelten Modelle noch die Messungenauigkeit gemäß Tabelle 2.1 (als davon unabhängiger Fehler) hinzugefügt werden muss.

Bedeutung der Referenzwerte

Sämtliche Zusatzinformationen dieses Unterabschnittes zu den vorliegenden Responses sind [25] entnommen.

Eine der wichtigsten Komponenten von Weinen bildet das Ethanol. Obwohl sich Ethanol entscheidend auf den Geschmack eines Weines auswirken kann, handelt es sich hierbei um eine farblose und geruchlose Verbindung. Selbst Weine ohne die beiden wichtigsten Vertreter von Zuckern in Weinen, der Glukose und Fruktose, können aufgrund eines hinreichend hohen Ethanolgehaltes, wie beispielsweise über $13.5\,\mathrm{Vol.\%}$, süßlich schmecken. Zusätzlich wirkt sich Ethanol auf die Konservierung aus. In den Weinen entsteht Ethanol durch den Fermentationsprozess von Hefepilzen des Zuckers.

Zu den wichtigsten Vertretern der Zucker zählen Glukose und Fruktose. Während am Beginn des Reifungsprozesses der Weintraube der Anteil an Glukose den Fruktoseanteil zum Teil bei Weitem übersteigt, ändert sich dieses Verhältnis mit Fortdauer des Reifungsprozesses und hat großen Einfluss auf die Süße des Weines. Bei der Fermentierung werden sowohl Fruktose als auch Glukose verbraucht und mit den richtigen Weinhefen kann sogar erreicht werden, dass nur noch geringe Spuren vom fermentierbaren Zucker im Wein feststellbar sind. Dies erklärt die

[6] Die Messmethoden können diese allerdings nicht erfassen, auch wenn die Referenzwerte (beinahe) kontinuierlich vorliegen.

große Spannweite in den vorliegenden Datensätzen (vgl. Tabelle 2.2 und Tabelle 2.3).

Diese beiden Zuckerarten zusammengefasst bilden meist den Hauptanteil von Extrakt. Dieser Weinparameter stellt einen Summenparameter dar, welcher alle nicht flüchtigen Inhaltsstoffe wie Mineralien oder verschiedene Säuren wie die Wein- oder Milchsäure, zusammenfasst. Dass die Höhe des Extraktwertes sehr stark von den beiden Zuckern Glukose und Fruktose in den vorliegenden Weinen abhängt, zeigen nachfolgende Analysen und werden beispielsweise in Abbildung 2.2 verdeutlicht.

Ein weiterer Bestandteil des Extraktwertes bildet das Glyzerin, welches sich mit einer Konzentration von meist 4 g/l bis 6 g/l vergleichsweise weniger stark auf die Höhe des Extraktwertes auswirkt. Dieser Zuckeralkohol bringt ein leicht süßliches Aroma mit sich und trägt ab einem Wert von 5.2 g/l zur Süße in Weißweinen bei. Für einen erkennbaren Einfluss auf die Viskosität benötigt man allerdings mehr als 28 g/l. Ein solcher Wein liegt ebenfalls im Datensatz vor.

Ein weiterer wichtiger Bestandteil von Weinen sind die Säuren. Hierbei ist die Responsevariable titrierbare Säure, analog zu Extrakt, wiederum als Summenparameter aller Säuren zu verstehen, welche durch das chemische Verfahren der Titrierung ermittelt werden können. Hierzu zählen insbesondere die Wein-, Äpfel-, Milch- und Zitronensäure.

Eine der wichtigsten Säuren in Weinen ist die Weinsäure. Diese trägt großteils zum Geschmack, sowie zur Stabilität der Säuren und zur Weinfarbe bei. Eine weitere wichtige Säure ist die Äpfelsäure, welche für Pflanzen und Tiere lebenswichtige Energie liefert. Während am Anfang des Reifungsprozess noch ungefähr 20 g/l in der Traube vorhanden sind, so sinkt dieser während des Reifungsprozesses auf 1 g/l bis 9 g/l, wobei sich der Gehalt an Äpfelsäure auch auf die Qualität und die Art des Weines auswirkt. Die hohe negative Korreliertheit mit Milchsäure, vgl. hierzu Abbildung 2.2, kann damit erklärt werden, dass Milchsäurebakterien einen starken Einfluss auf die Verstoffwechslung der Äpfelsäure hat (speziell bei Temperaturen im Bereich von 18 °C bis 22 °C).

Die Milchsäure bildet eine weitere Säureart, für welche Messwerte vorliegen. Wie es der Name vermuten lässt, handelt es sich hierbei um die wichtigste Säure in Milchprodukten. Im Wein entsteht die Milchsäure als Nebenprodukt der alkoholischen Fermentation und zusätzlich kann durch die Verstoffwechslung gewisser Bakterien die Äpfelsäure in Milchsäure und CO_2 umgewandelt werden.

Unter dem Terminus „flüchtige Säuren“ versteht man Säuren, welche sich mit zunehmender Zeit verflüchtigen. Der Großteil dieser Klasse von Säuren bildet die Essigsäure (in der Regel mehr als 96 %), wobei ein Essigsäuregehalt von weniger als 0.2 g/l sich nicht negativ auf den Geschmack auswirkt. Bei einer Konzentration von ca. 1.5 g/l kann der Wein ein essigähnliches Aroma erhalten, wobei diese Note nicht durch die Essigsäure selbst, sondern durch die Reaktion mit Ethanol verursacht wird.

Eine Säure mit geringeren Konzentrationen ist die Zitronensäure und dient oftmals nur zur Stabilisierung der Weine.

Eine ermittelte Eigenschaft der Weine liegt mit dem pH-Wert vor. Hierunter versteht man eine einheitenlose Kennzahl und diese berechnet sich durch den negativen Logarithmus zur Basis 10 der Aktivität von Wasserionen. Sie liefert eine Kennzahl über den Säuregehalt (der Weine), weshalb dieser eine hohe Korrelation mit dem Geschmack aufweist. Aufgrund dieser Berechnung impliziert ein um 1 höherer pH-Wert eine mit Faktor 10 erhöhte Wasserionenaktivität. Der Wertebereich für Weine kann grundsätzlich mit $[2.9, 4.2]$ angegeben werden. Im Vergleich hierzu weist Wasser einen neutralen pH-Wert in Höhe von ca. 7 auf, während die Magensäure einen Wert von ca. 2 aufweist.

Übersicht über die vorliegenden Referenzwerte

Tabelle 2.2 enthält eine Übersicht der in 2015 ermittelten Responsevariablen, während ein Überblick über den Datensatz 2016 in Tabelle 2.3 vorliegt. Diese tabellarischen Übersichten zeigen jeweils das Minimum und das Maximum, den Median und den Mittelwert, sowie das erste und dritte Quartil der Referenzwerte. Die Spalte NA gibt die Anzahl von fehlerhaften Daten wieder. Die vorletzte Spalte zählt die Anzahl der 0-Werte der entsprechenden Responsevariable und in der letzten Spalte sind die Anzahl der Proben ohne Fehlmessungen gegeben.

	Min.	$q_{0.25}$	Med.	Mw.	$q_{0.75}$	Max.	NA	0-Werte	Anz.
Äpfelsäure	0.00	0.00	0.20	0.96	1.75	5.40	0	38	80
Dichte	0.98	0.99	0.99	1.00	1.00	1.09	0	0	80
Dichte Ethanol	0.97	0.98	0.98	0.98	0.98	1.00	0	0	80
Dichte Extrakt	1.00	1.01	1.01	1.02	1.01	1.10	0	0	80
Ethanol	0.00	11.90	12.76	12.56	13.57	19.76	0	2	80
Extrakt	14.40	25.00	30.56	51.52	38.09	264.20	0	0	80
Flüchtige Säure	0.20	0.40	0.50	0.55	0.60	1.60	0	0	80
Fruktose	0.00	0.30	1.80	15.28	5.33	138.40	0	16	80
Glukose	0.00	0.00	1.30	9.09	3.62	74.60	0	22	80
Glyzerin	1.00	6.38	9.35	9.27	10.32	30.30	0	0	80
Milchsäure	0.00	0.38	1.20	1.09	1.62	2.80	0	6	80
pH-Wert	2.89	3.35	3.50	3.48	3.63	3.93	2	0	78
Titrierbare Säure	3.50	4.78	5.20	5.36	5.72	8.30	0	0	80
Weinsäure	0.00	1.60	1.90	1.94	2.20	3.70	0	1	80

Tabelle 2.2: Übersicht der Referenzwerte für den Datensatz des Jahres 2015

Beinahe die Hälfte aller Weine weist einen L-Äpfelsäuregehalt von 0 g/l auf, wobei 35 der 38 Nullwerte auf Rotweine entfallen und offensichtlich mit dieser Weinklasse korrespondieren. Dies ist auch der Grund, weshalb die Werte der L-Äpfelsäure äußerst rechtsschief verteilt sind.

Mit 16 bzw. 22 Werten weisen Fruktose bzw. Glukose den zweit und drittgrößten Wert an 0en auf. Wie in Abbildung 2.2 (li.) zu erkennen, sind diese beiden Responsevariablen auch äußerst hoch korreliert, wie auch mit Extrakt. Dies folgt aus der Tatsache, dass bei den vorliegenden Weinen geringe Glukosewerte ebenfalls Fruktosewerte von sehr kleinem Ausmaß aufweisen. Bei 13 Weinen sind sogar beide

Inhaltsstoffe nicht vorhanden. Während allerdings geringe Fruktosewerte (0 g/l bis 1 g/l) ebenfalls Glukosewerte in selbigem Bereich implizieren, so sind auch Weinmessungen vorhanden, bei welchen der Fruktose- dem dreifachen Glukosegehalt entspricht. Bei insgesamt 63 der Weine gilt die Relation Glukose $\leq$ Fruktose, wobei bei den restlichen 17 Weinen die Werte beinahe gleich sind. Ein mit diesen beiden Zuckerarten stark korrespondierender Wert ist Extrakt. Auffallend ist, dass es hier eine enorme Min-Max Spannweite gibt, welche allerdings mit jener von Glukose und Fruktose zu erklären ist, zumal der Extraktwert als Summenparameter die beiden Zuckerarten beinhaltet.

Der Wertebereich $[2.89, 3.93]$ des pH-Wertes stimmt mit dem in Abschnitt 2.1 beschriebenen Wertebereich überein, weshalb bezüglich des pH-Wertes beinahe die gesamte Spannweite abgedeckt wird, wobei der geringste Wert von 2.89 als Ausreißer der Daten aufgefasst werden kann. Diese Responsevariable ist die einzige, welche in diesem Datensatz für zwei Werte nicht verfügbar ist.

Für einen wichtigen Bestandteil von Weinen, dem Ethanol, ist eine sehr geringe Länge des interquartilen Bereichs auffällig. Dennoch finden sich sowohl alkoholfreie als auch Portweine im Datensatz (0 Vol.% bzw. 19.87 Vol.%), welche trotz der wenigen Vertreter (mit)modelliert werden. Die beiden größten Ausreißer nach unten, die gemessenen alkoholfreien Weine, weisen einen Wert von 0.0 Vol.% auf. Da das tatsächliche Erreichen dieses Wertes bei Weinen oftmals nicht möglich ist, kann an dieser Stelle ein Messfehler vorliegen.[7] Im Datensatz des Jahres 2016 sind ebenfalls diese Weine enthalten. Die entsprechenden Referenzwerte in dem aktuelleren Datensatz sind von 0 Vol.% verschieden, wenn auch nur minimal.

Darüber hinaus existieren Weine mit einem Glyzeringehalt von über 30, was, wie in Abschnitt 2.1 beschrieben, einen erkennbaren Einfluss auf die Viskosität hat.

In Abbildung 2.1 sind die Werte des 2015er Datensatzes für Milch- und Weinsäure in Abhängigkeit der Weinfarbe abgebildet. Anhand dieser Kennzahlen kann man eine Separation der Farbe erkennen und mithilfe dieser beiden Säuren kann eben diese relativ gut geschätzt werden. Man kann hier auch die Verteilung der Weinsäure erkennen. Diese scheint relativ symmetrisch um den Median zu sein, mit je einem Ausreißer nach oben und nach unten. Diese Grafik macht auch klar, dass eine Modellierung, beschränkt auf die Farbe des Weines, sinnvoll ist bzw. sein kann.

Die teilweise als Geraden erkennbaren Punktanordnungen können auf die Tatsache zurückgeführt werden, dass eine Art Diskretisierung der Werte aufgrund der Datengenauigkeit stattfindet. So liegen die Werte der Wein- und Milchsäure je mit einer Genauigkeit von 10^{-1} g/l vor.

In Tabelle 2.3 ist die analoge Aufstellung der Responsevariablen für den Datensatz des Jahres 2016 mit den 81 Weinen zu finden. Im Wesentlichen können hier dieselben Schlussfolgerungen getroffen werden wie für Tabelle 2.2. Auffallend ist, dass dieselben alkoholfreien Weine analysiert wurden, diese in diesem Datensatz

[7] Alkoholfreie Getränke dürfen einen Alkoholgehalt von bis zu 0.5 Vol.% aufweisen. Dies entspricht der erlaubten Abweichung des angegebenen Alkoholgehaltes. Getränke wie Weine müssen erst ab einem Wert von 1.2 Vol.% mit dem Alkoholgehalt gekennzeichnet sein. Vgl. hierzu [5]

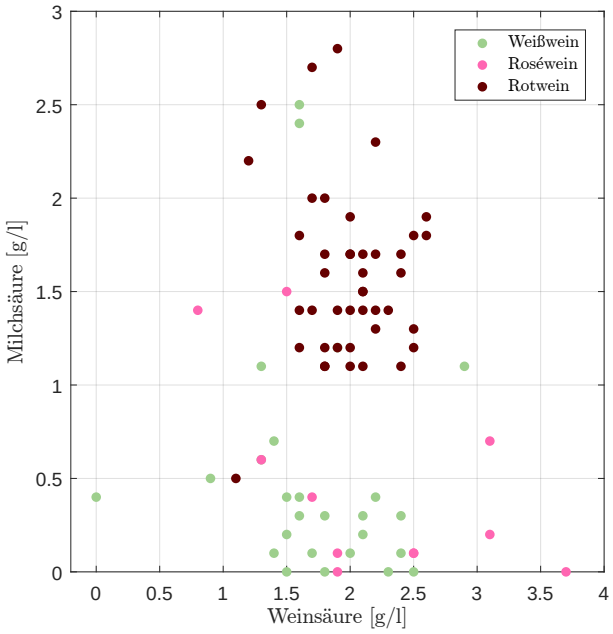

Abbildung 2.1: Die Referenzwerte der Milchsäure in g/l werden gegen jene der Weinsäure in g/l gegenübergestellt. Die Grafik zeigt den möglichen Einfluss der Weinfarbe auf die Referenzwerte.

allerdings von 0 Vol.% verschieden sind. Dies kann einerseits an dem Schwankungsbereich aus Tabelle 2.1 liegen, oder an der Tatsache, dass derselbe Wein einer anderen Charge analysiert wurde. Darüber hinaus ist die Messung des AP-Ethanols für große Extraktwerte fehlerhaft, weshalb in der weiteren Analyse dieser Arbeit für die Modellierung der Ethanolwerte stets die Kennzahl Ethanol gewählt wird.

Zusätzlich ist hier die Zitronensäure mitdokumentiert. Bei knapp mehr als der Hälfte aller Weinmessungen wurde der Wert 0 g/l festgestellt. Der höchste Wert von 1.2 g/l, stellt einen enormen Ausreißer dar, da die anderen Weine eine maximale Zitronensäure von 0.5 g/l aufweisen.

Bei den flüchtigen Säuren tritt eine zusätzliche Fehlmessung auf. Auch wenn 81 Weine zur Verfügung stehen, muss jener Wert mit einer Konzentration von 0 g/l entfernt werden, da es sich an dieser Stelle um eine Fehlmessung handelt.

Korrelation der vorliegenden Referenzwerte

Die paarweisen Korrelationen der Werte nach Pearson sind in Abbildung 2.2 dargestellt, wobei nach Möglichkeit eine abnehmende Korrelation mit der Entfernung zur Diagonale korrespondiert.

Die drei Kennzahlen Extrakt, Glukose, sowie Fruktose weisen eine äußerst hohe Korrelation auf (größer 0.9), weshalb speziell bei diesen Responses eine gemeinsame Modellierung von Vorteil sein kann.

	Min.	$q_{0.25}$	Med.	Mw.	$q_{0.75}$	Max.	NA	0-Werte	Anz.
Äpfelsäure	0.00	0.00	0.30	0.82	1.50	3.70	0	40	81
D-Milchsäure	0.00	0.10	0.90	0.77	1.30	2.80	0	19	81
Dichte	0.98	0.99	0.99	1.00	1.00	1.10	1	0	80
Ethanol	0.09	11.97	12.78	12.53	13.54	19.87	0	0	81
Extrakt	14.40	26.17	31.35	57.57	39.65	289.80	1	0	80
Flüchtige Säure	0.20	0.40	0.50	0.57	0.60	1.70	0	0	80
Fruktose	0.00	0.40	1.90	18.30	8.10	137.20	0	15	81
Glukose	0.00	0.00	1.30	10.82	5.80	95.20	0	25	81
Glyzerin	0.50	6.00	8.30	8.71	9.20	34.60	0	0	81
L-Milchsäure	0.00	0.10	0.20	0.22	0.30	1.20	0	9	81
Milchsäure	0.00	0.20	1.10	0.99	1.50	3.00	0	8	81
pH-Wert	2.91	3.34	3.51	3.47	3.59	3.95	0	0	81
Titrierbare Säure	3.30	4.70	5.30	5.34	5.90	7.90	0	0	81
Weinsäure	0.00	1.60	1.90	1.86	2.10	3.80	0	1	81
Zitronensäure	0.00	0.00	0.00	0.14	0.20	1.20	0	41	81

Tabelle 2.3: Übersicht der Referenzwerte für den Datensatz des Jahres 2016

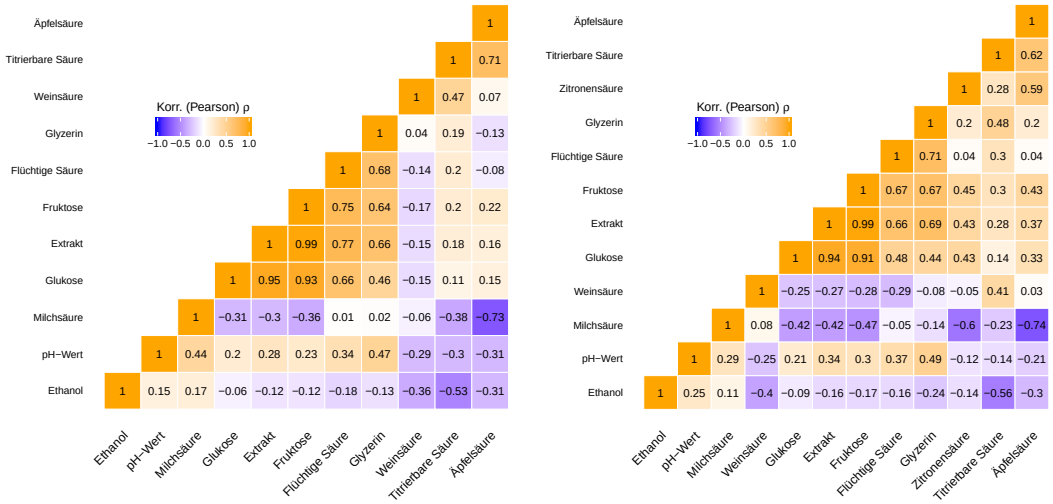

Abbildung 2.2: Die lineare Korrelation nach Pearson für die einzelnen zur Verfügung stehenden Referenzwerte, derart angeordnet, sodass die Korrelation mit der Entfernung von der Diagonale nach Möglichkeit abnimmt.

Allerdings muss beachtet werden, dass dies lediglich den linearen Zusammenhang der Responses wiedergibt und daher nicht überinterpretiert werden darf. Dies dient nur zur Übersicht um Empfehlungen für gemeinsame Modellierungen anzugeben, wie etwa Extrakt, Glukose und Fruktose. Im weiteren Verlauf der Arbeit stellt sich allerdings heraus, dass die besten Ergebnisse durch getrennte Betrachtungen aller Responsevariablen erzielt werden können.

Farbe des Weines

Die zusätzliche Klassifizierung der Weine nach deren Farbe (rot, weiß, rosé) liefert weitere wichtige Informationen über die Weinart. Da die Rotweine den Großteil der vorhandenen Stichprobe darstellen, werden stets auch speziell für diese Subklasse die dazugehörenden Werte modelliert.

Diese Faktorvariable kann einen signifikanten Einfluss auf die Referenzwerte haben, wie anhand von Milchsäure in Abbildung 2.1 gezeigt. Hierbei gibt es grundsätzlich zwei Vorgehensweisen. Für Analysezwecke können getrennte Modelle betrachtet werden und darüber hinaus kann diese Kategorisierung zur Einfärbung der Residuen in Modellen verwendet werden, um etwaige Trends innerhalb dieser Kategorien zu erkennen und um gewisse Abweichungen erklären zu können.

2.2 Prädiktoren

Mit den Spektrometern wird je ein Signal und eine Referenz gemessen und man erhält mit der Fourier-Transformation die beiden oberen Darstellungen (Einkanalspektren) in Abbildung 2.3. Das Signal entspricht der Absorption der Weinprobe, die Referenz einer Wassermessung, auf welche der Wein referenziert wird. Dies erfolgt mittels der Transformation $-\log\left\{\frac{\text{Signal}}{\text{Referenz}}\right\}$ und erhält in ebendieser Abbildung die dritte (untere) Grafik, mit dessen Werten in dieser Arbeit gearbeitet wird und im weiteren Verlauf als Spektrum bezeichnet wird.

Tatsächlich handelt es sich bei den einzelnen Daten um den gemittelten Wert von vier unterschiedlichen Messwiederholungen. Hierauf wird in Abschnitt 2.2.1 näher eingegangen.

Wie bereits anhand von Abbildung 2.3 ersichtlich, weisen die Spektren sowohl im niederen als auch im hohen Wellenzahlbereich (bis ca. $900\,\text{cm}^{-1}$ und über $3\,100\,\text{cm}^{-1}$) starke Oszillationen auf. Diese kann auf die geringen Werte (jeweils nahe bei 0 und somit kaum verwertbare Information) in den Einkanalspektren zurückgeführt werden, weshalb der Wellenzahlbereich für weitere Analysen auf $[900, 3\,100]\,\text{cm}^{-1}$ eingeschränkt wird. Dies gilt gleichermaßen für die Engines der zweiten Generation und das Spektrometer V70.

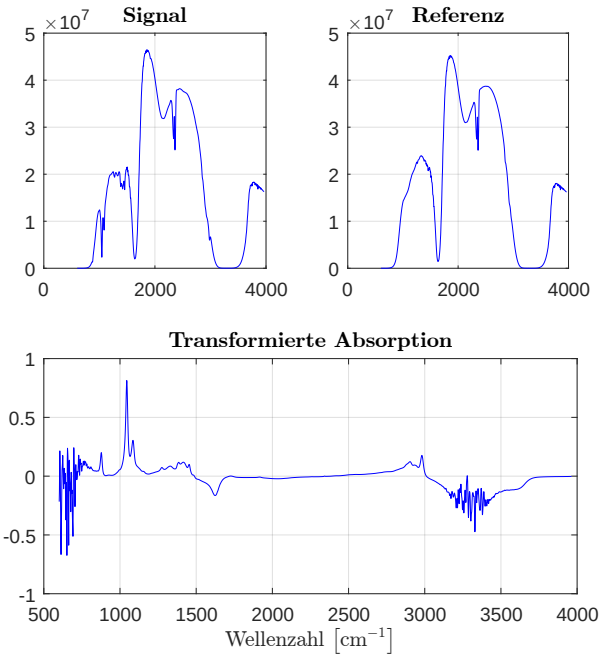

Abbildung 2.3: Zwei Einkanalspektra (o.), welche aus den ermittelten Interferogrammen mittels Fourier-Transformation resultieren: für eine Weinprobe (o., li.) und H_2O (o., re.) sowie ein durch die logarithmische Transformation resultierendes Spektrum (u.). Jeweils mit Daten des FTIR Spektrometers E22.

Datensatz, Jahr 2016

Für die Datensätze des Jahres 2016 stehen $n_{2016} = 81$ unterschiedliche Spektren von Weinproben an vier unterschiedlichen Spektrometern zur Verfügung.

Innerhalb dieser Spektrometer liegt, neben der Höhe der Absorptionswerte aufgrund der unterschiedlichen Anzahl an ATR-Reflexionen ebenfalls eine unterschiedliche Rasterung des Wellenzahlbereiches vor, wie Tabelle 2.4 zeigt. Während in etwa der gleiche Gesamtbereich abgedeckt wird, stehen bei den Engines der zweiten Generation um 44 % mehr Messwerte zur Verfügung.

	Anzahl	Schrittweite	Min.	Max.
*V*70	1 760	1.93	601.8	3 994.4
*E*2•	2 535	1.33	601.1	3 963.5

Tabelle 2.4: Übersicht der Wellenzahlen in den Datensätzen des Jahres 2016. Alle Werte in cm^{-1}.

Zusätzlich zu den Messungen an den ausgewiesenen Wellenzahlen sind die Umgebungsbedingungen wie Luftfeuchtigkeit und -druck, die Temperatur des Labors, sowie die Zellentemperatur zum Messzeitpunkt protokolliert.

Die analysierten Objekte (Weine) stammen aus 14 unterschiedlichen Ländern aus den Jahrgängen 2007 bis 2015. Am häufigsten vertreten sind Weine aus Österreich mit 24, Italien mit 14 und Frankreich mit 10 Weinen. Zusätzlich können die Weine anhand ihrer Farbe klassifiziert werden. Diese gliedern sich wie in Tabelle 2.5 in Rot-, Weiß- und Roséweine, wobei die Rotweine den größten Teil des Datensatzes bilden.

Da mehr als die Hälfte der Weine den Rotweinen zugeordnet werden kann, scheint eine Analyse der Daten, eingeschränkt auf diese Weinkategorie, zum Teil ebenfalls empfehlenswert, wie die Auswertungen in den Kapiteln 5 und 7 zeigen werden.

In Abbildung 2.4 (li.) ist der Bereich der Absorptionswerte des Messgerätes E25 dargestellt. Die eingefärbte Fläche wird durch die zwei fiktiven Spektren, dem punktweisen Maximum bzw. Minimum begrenzt. Offensichtlich handelt es sich mehrheitlich um positive Werte. Sämtliche Spektren weisen Werte im Bereich $[-0.28, 1.47]$, bei einer maximalen Spannweite von 1.62, auf. Die auffallendste negative Bande befindet sich im Bereich um $1\,600\,\mathrm{cm}^{-1}$. Da es sich hier um auf Wasser referenzierte, logarithmierte Absorptionswerte handelt, bedeutet dies lediglich, dass in den Weinproben weniger H_2O-Moleküle vorhanden sind als in reinem

	Anzahl
Roséweine	7
Rotweine	44
Weißweine	30

Tabelle 2.5: Anzahl der Weine nach Weinfarbe

Wasser, weshalb keine relevante Information für die Inhaltsstoffe in diesem Bereich zu erwarten ist.

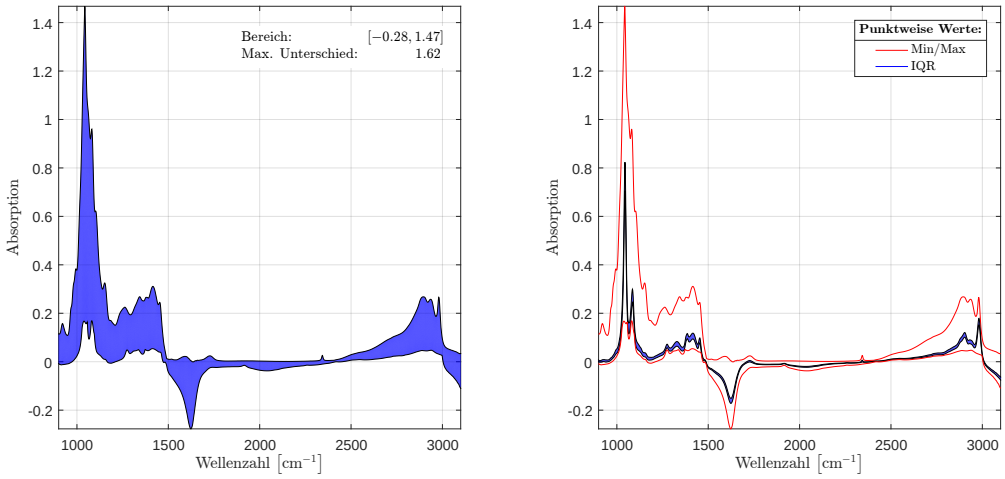

Abbildung 2.4: Punktweises Minimum und Maximum aller Weinspektren des Datensatzes E25 (li.), sowie wiederum dieselben Konturen in Rot (re.) mit zusätzlichem interquartilen Bereich der Absorptionswerte des Datensatzes E25. Jeweils in Abhängigkeit von der Wellenzahl.

Während für den V70(2016) zusätzlich TempBath zur Verfügung steht, ist dies bei den Engines E22, E24 und E25 jeweils die Temperatur des jeweiligen Lasers, Mainboards, der Lampe, des Detektors und des Interfers. Bei unterschiedlichen Temperaturen von bis zu 10 °C kann dieser Faktor unter Umständen einen Einfluss auf die Messungen (z.B. Messgenauigkeit) aufweisen und muss bei Modellanalysen ebenso berücksichtigt werden.

Datensatz, Jahr 2015

Im Vergleich zu den Datensätzen des Jahres 2016 stehen für das Jahr 2015 insgesamt $n_{2015} = 80$ Spektren des Messgerätes V70 zur Verfügung, wobei die Weine nur teilweise mit jenen der Datensätze des Jahres 2016 ident sind. Darüber hinaus sind keine zusätzlichen Informationen über das Land oder mögliche Temperaturschwankungen im Labor selbst oder des Messgerätes erfasst. Diese Weinspektren lassen sich bezüglich ihrer Weinfarbe in 26 Weiß-, 43 Rot- und 11 Roséweine klassifizieren.

Exemplarisch sind in Abbildung 2.5 (o.) 10 Spektren dargestellt. Im Bereich bis zu einer Wellenzahl von circa 600 kann man eine starke Oszillation erkennen und wiederum zeigen sich negative Banden bei den Wellenzahlen $[1\,500, 1\,700]\,\mathrm{cm}^{-1}$, sowie $[3\,000, 3\,700]\,\mathrm{cm}^{-1}$. Der höchste Peak wird im Wellenzahlbereich von circa $1\,000\,\mathrm{cm}^{-1}$ erreicht und man erkennt ein ähnliches Verhalten der Spektren, auch wenn das Rauschen in einem Wellenzahlbereich zwischen $3\,100\,\mathrm{cm}^{-1}$ und

$3\,500\,\mathrm{cm}^{-1}$ weniger stark ausgeprägt ist wie bei den Datensätzen der Engines der zweiten Generation (vgl. Abbildung 2.3 (u.)).

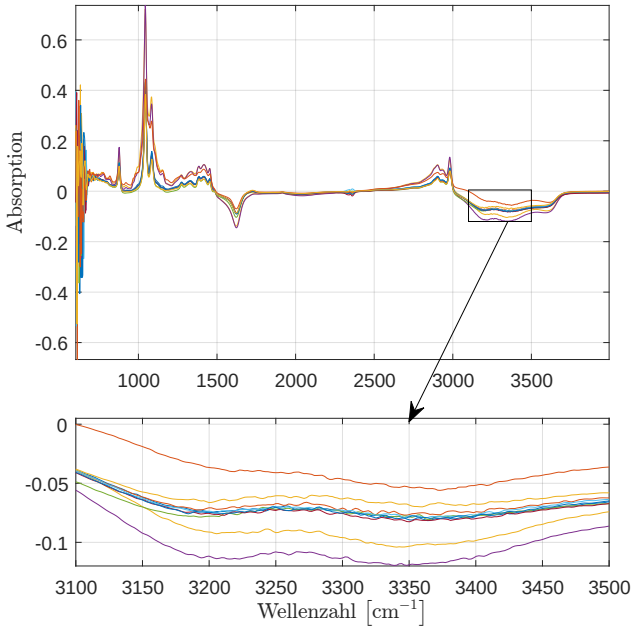

Abbildung 2.5: 10 Spektren auf dem gesamten Wellenzahlbereich mit dem Fokus auf verrauschte Bereiche im Datensatz V70(2016).

2.2.1 Reproduzierbarkeit

Ein wichtiges Kriterium neben der Modellgüte, wie diese bei Spektrometern häufig als doppelte Standardabweichung der Residuen bewertet und angegeben wird (vgl. Referenzmethode in Tabelle 2.1), stellt die Reproduzierbarkeit dar. Hierunter versteht man, dass für dieselbe Weinprobe bei wiederholten Messungen idente Spektren (resp. in weiterer Folge Schätzwerte) berechnet werden können.

Da ein Spektrum wie zuvor erwähnt, durch die Mittelung von vier unterschiedlichen, nacheinander durchgeführten Messungen, bestimmt wird, können diese vier Datensätze je Weinprobe zur Überprüfung der Reproduzierbarkeit herangezogen werden. Exemplarisch ist in Abbildung 2.6 ein einzelnes gemitteltes Weinspektrum in Rot dargestellt, wohingegen die dazugehörenden tatsächlichen Messungen in Blau abgebildet sind.

In dem unteren Teilausschnitt im Wellenzahlbereich $[1\,290, 1\,350]$ cm^{-1} erkennt man deutliche Unterschiede dieser Spektren. Speziell im Bereich des Peaks zeigen sich Messungenauigkeiten.

Konkret handelt es sich für den Datensatz des Jahres 2016 um $82 \times 4 = 328$ Weinmessungen sowie zusätzliche Messungen von zwei künstlich erzeugten Weinen,

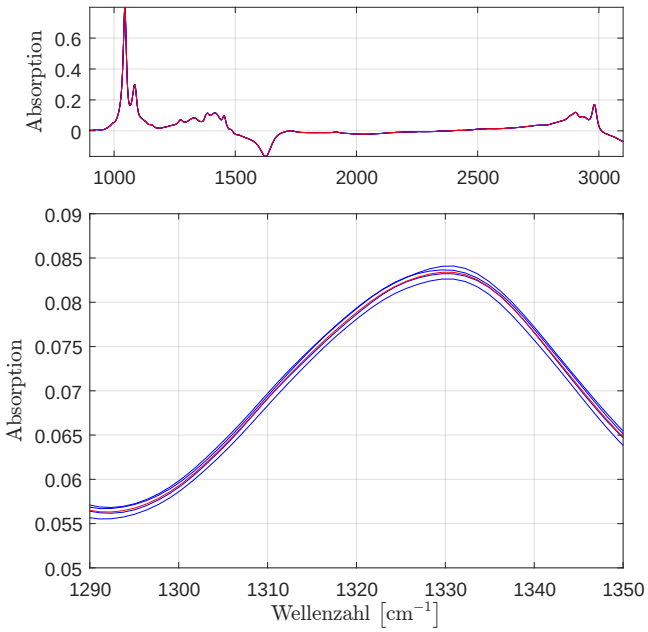

Abbildung 2.6: Darstellung von vier Einzelspektren (in Blau), die für das endgültige (rote) Spektrum gemittelt werden.

welche wiederum zu 5 unterschiedlichen Zeitpunkten mit ebenfalls je 4 Wiederholungen gemessen wurden. Diese Kunstweine stehen allerdings nur für die Responsevariablen Ethanol, Wein-, Zitronen, L-Äpfelsäure, sowie Glukose, Fruktose und Glyzerin zur Verfügung.
Im Vergleich hierzu liegen für jenen Datensatz des Jahres 2015 lediglich Mehrfachmessungen eines einzelnen Weines vor. Hierbei wurden 10 Messungen direkt aufeinanderfolgend vorgenommen, während für weitere 10 Messungen die Zelle des ATR-FTIR Spektrometers nach jeder Messung neu befüllt wurde.

Neben den Messungen für Weine liegen zusätzlich Wassermessungen vor. Während für den Datensatz aus dem Jahr 2015 insgesamt 80 unterschiedliche Wassermessungen ohne weitere Detailinformationen vorliegen, kann für den Datensatz 2016 auf 487 verschiedene Wasserspektren zurückgegriffen werden. Hierbei stehen noch zusätzliche Spezifikationen wie die verwendete Milliliteranzahl sowie die Information über Messwiederholungen analog des Reproduzierbarkeitsdatensatzes für Weine zur Verfügung.

Abbildung 2.7 vergleicht vier Weinspektren (in Rot) mit 4 Wasserspektren. Die Wassermessungen enthalten, wie zu erwarten, keine vergleichbaren Peaks und die Absorptionswerte befinden sich in einem relativ engen Intervall $[-4.8, 8.4] \cdot 10^{-3}$ um 0. Diese geringen Abweichungen können als Rauschen im Messsystem interpretiert werden, da ein ideales Wasserspektrum mit der 0-Linie übereinstimmt. Weiters kann an sämtlichen Wellenzahlbereichen ein starkes Oszillieren beobachtet werden.

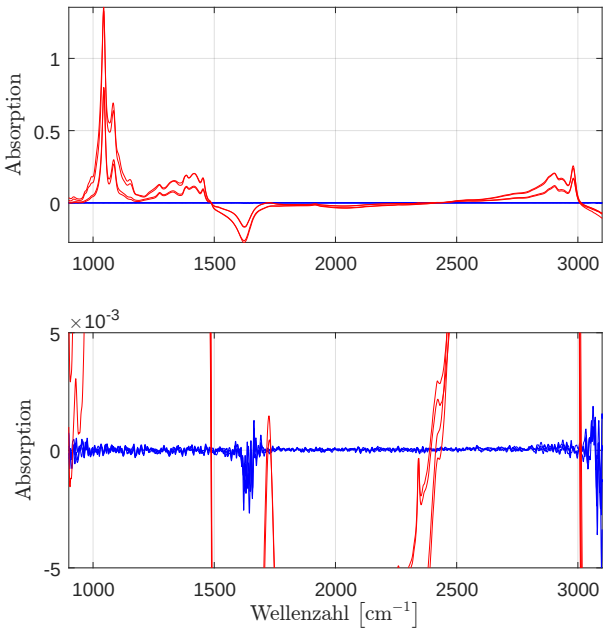

Abbildung 2.7: In roter Farbe dargestellt sind Weinspektren, während in Blau das unterschiedliche Verhalten von Wasserspektren verdeutlicht wird.

2.2.2 Variablenvorselektion

Da in weiterer Folge die Wellenzahlen reduziert werden sollen um die Referenzwerte passend modellieren zu können, muss zuerst eine Übersicht über relevante Spektrenbereiche geschaffen werden. Abbildung 2.4 (li.) zeigt, für welche Wellenzahlen die höchste Schwankungsbreite vorliegt. Hierzu sind pro Wellenzahl sowohl das Maximum, das Minimum, sowie der interquartile Bereich aufgetragen. Bis zu einer Wellenzahl von ungefähr $1\,500\,\mathrm{cm}^{-1}$, sowie ab $2\,500\,\mathrm{cm}^{-1}$ sind starke strukturelle Unterschiede zu erkennen, wohingegen im mittleren Wellenzahlbereich der Minimal- und Maximalwert nahe beieinander liegen und sich hier womöglich weniger relevante Informationen befinden.

Zusätzlich ist eine vergleichbar stark negative Bande zu erkennen. Da es sich hier um auf Wasser referenzierte, logarithmierte Absorptionswerte handelt, bedeutet dies, dass in diesem Bereich die Konzentration der OH-Gruppe geringer ist.

Weitere Erkenntnisse über den Zusammenhang der Spektren zu den Referenzwerten stellen die spektralen Korrelationsplots dar. Hierunter werden jene Grafiken bezeichnet, welche die Korrelation einer Zielgröße zu den einzelnen Prädiktoren wiedergeben, wie anhand von Abbildung 2.8 exemplarisch dargestellt. Betrachtet man beispielsweise den spektralen Korrelationsplot von drei Zuckertypisierungen Extrakt, Glukose und Fruktose, so erkennt man, dass diese an äußerst ähnlichen Wellenzahlen gleichermaßen korreliert sind.

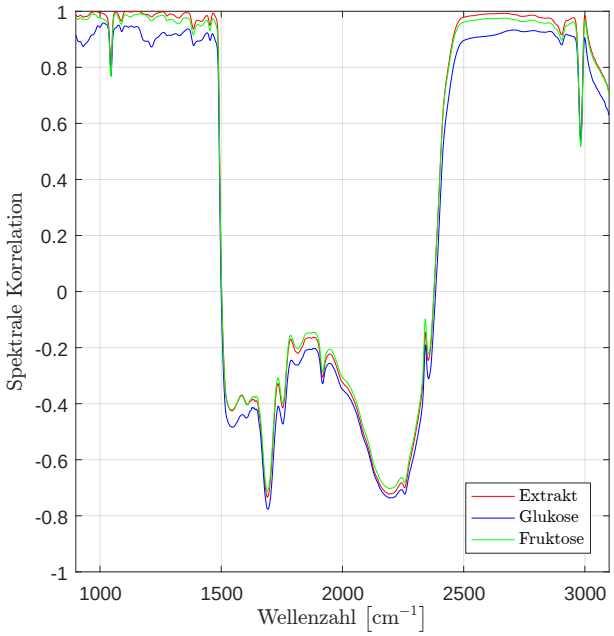

Abbildung 2.8: Die dargestellte spektrale Korrelation (Korrelation des Referenzwertes mit den Absorptionswerten an den jeweiligen Wellenzahlen) für Extrakt, Glukose und Fruktose. Vgl. Abbildung 2.2.

Oszillation und Bandenelimination

- Wie bereits erwähnt, können enorme Oszillationen am Rand auf die Werte der Einkanalspektren zurückgeführt werden. Daher kann in diesen Bereichen keine relevante physikalische bzw. chemische Information erwartet werden, weshalb eine Begrenzung des Wellenzahlbereiches auf das Intervall $[900, 3\,100]\,\mathrm{cm}^{-1}$ stattfindet.

- Weiters werden in den Spektren für die zu erklärenden Variablen die Wasser- und CO_2-Banden eliminiert. Diese befinden sich zwischen den Wellenzahlen $1\,544\,\mathrm{cm}^{-1}$ und $1\,676\,\mathrm{cm}^{-1}$ respektive $2\,300\,\mathrm{cm}^{-1}$ und $2\,400\,\mathrm{cm}^{-1}$, wie Abbildung 2.9 veranschaulicht.

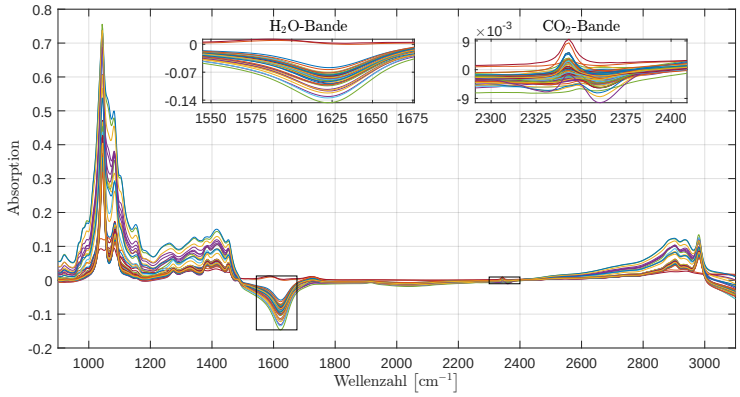

Abbildung 2.9: Spektren mit der H_2O- und CO_2-Bande, welche aus der Variablenselektion ausgeschlossen werden, da keine Information für die Referenzwerte an diesen Stellen vermutet werden kann.

2.3 Preprocessing

Bei der Messung von Absorptionen können stets Störungen auftreten. Hierzu zählt beispielsweise das Oszillieren, welches bei den hier verwendeten Datensätzen insbesondere im niederen und hohen Wellenzahlbereich beobachtet wird. Ein weiterer potentieller Störfaktor stellt eine Basislinienverschiebung dar. Während diese im Idealfall eine horizontale Linie durch den Nullpunkt ist, kann sie durch einen Offset vertikal verschoben sein oder es können auch lineare (grundsätzlich sogar beliebige) Verschiebungen auftreten. Unerwünschte Effekte können durch die Datenaufbereitung, dem sogenannten Preprocessing, reduziert werden.

Weitere Gründe für Messstörungen können vielseitig sein, wie Verunreinigungen in der Durchflusszelle[8] oder der Probe, sowie unterschiedliche Proben- oder Raumtemperaturen. Gegebenenfalls können diese als zusätzliche Prädiktoren in die Modelle aufgenommen werden.[9] Für die Bestimmung der in dieser Arbeit verwendeten Spektren wird die Temperatur der Zelle stets konstant gehalten[10].

Aufgrund dessen gibt es eine Vielzahl von Vorbehandlungsmöglichkeiten der spektralen Daten, welche zu einer Verbesserung der auf diesen Messungen basierenden Analysen führen kann, wobei grundsätzlich auch Verschlechterungen hervorgerufen oder verstärkt werden können.

In Abbildung 2.10 wird beispielsweise ein Weinspektrum von E22 und V70(2016) verglichen. In der oberen Abbildung sind die Weinspektren in ihrer originalen Form dargestellt. Um diese vergleichen zu können, wird der höchste Peak angeglichen (skaliert) und die Differenz der beiden Spektren gebildet (selbe Abbildung, unten).

[8] Dies wiederum kann zu einer Veränderung des Brechungsindex an der Grenzfläche zwischen der Probe und des Kristalls führen.

[9] Die untersuchten Modelle zeigen jedoch keine Notwendigkeit, diese in die Modellfindung miteinzubeziehen.

[10] Abweichungen von weniger als 0.002 °C, weshalb bei den verwendeten Daten die Probentemperatur keinen Einfluss aufweist

Hierbei ist der Referenzpunkt in Rot markiert (die Spektren sind derart skaliert, dass sie in diesem Punkt übereinstimmen). Trotz geringer Unterschiede der Messmethode müssten die Spektren übereinstimmen. Man erkennt aber, abgesehen von gewissem kleineren Rauschverzerrungen, einen teils starken Trend, was wiederum auf eine Basislinienverschiebung hindeuten kann.

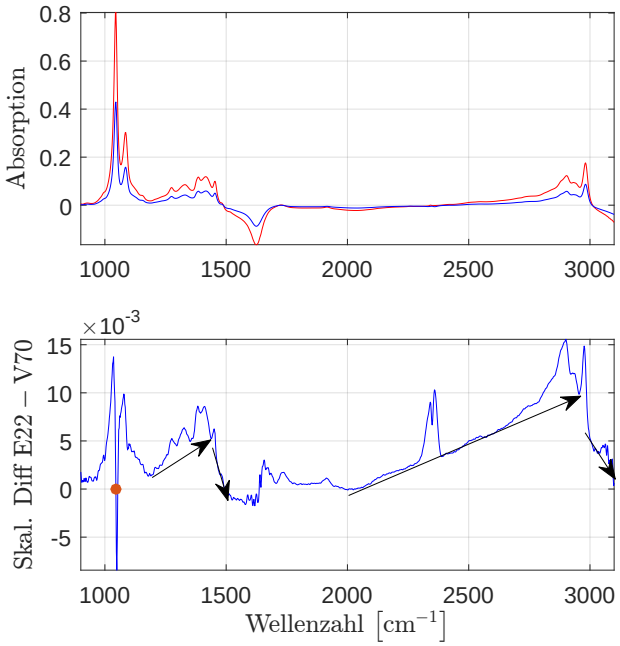

Abbildung 2.10: Vergleich zweier Spektren desselben Weines aus Datensatz E22 und V70(2016) (o.), sowie der punktweisen Differenz der beiden Spektren mit gleich skaliertem maximalen Absorptionswert (durch den Punkt markiert). Diese Differenz zeigt, dass datensatzübergreifend stückweise Trends in den Daten vorhanden sein können, da die Spektren nur geringe, allerdings überall ähnliche Abweichungen aufweisen müssten.

Peakadjustierung

Zusätzlich kann es bei den Messungen der Spektren zu (geringen) Peakverschiebungen bezüglich der Wellenzahlen kommen. Darunter versteht man, dass bei einer Absorptionsbande das Maximum der unterschiedlichen Weinmessungen nicht bei derselben Wellenzahl angenommen wird. Hierbei muss eine Anpassung bezüglich der Wellenzahlwerte durchgeführt werden, wobei im einfachsten Fall eine Verschiebung der Wellenzahlenachse ausreicht. Da in den vorliegenden Datensätzen diese Anpassung nicht von Relevanz ist, wird an dieser Stelle auf die Arbeit [31] verwiesen und nicht näher darauf eingegangen.

Datenglättung und Trendbereinigung (Basislinienkorrektur)

- Eine erste Möglichkeit bildet der **Differenzenquotient erster Ordnung**. Im Speziellen wird in dieser Arbeit der Vorwärtsdifferenzenquotient angewendet. Hierbei wird die Matrix wie folgt modifiziert: $x_{i,j} = \frac{x_{i,j} - x_{i,j+k}}{d}$, wobei i stellvertretend für die Probe, j die dazu korrespondierende Wellenzahl und d die äquidistante Differenz der Wellenzahlen zu $x_{i,j}$ und $x_{i,j+k}$ repräsentiert. $k \in \mathbb{N}_{>0}$ wird als Gap bezeichnet. Den **Differenzenquotienten zweiter Ordnung** erhält man durch wiederholte Anwendung. Mit dieser Preprocessingart können vertikale Verschiebungen beziehungsweise lineare Trends eliminiert werden.

- Der **Savitzky–Golay Filter** wird hier in dreifacher Anwendung untersucht und entspricht der Kombination von polynomialer Datenglättung und Differentiation. Einerseits als reine Glättungsfunktion und andererseits als Glättung einschließlich der ersten und zweiten Ableitung. Seien die Punkte $x_{i,j+k}, k \in \{-j, \ldots, j\}$ mit $j \in \mathbb{N}_{>0}, P_k(x) = \sum_{i=1}^{k} a_i x^i$ ein Polynom vom Grad k und Datenpaare $(w_j, y_j), j \in 1, \ldots, m$ gegeben.
  Die polynomiale Glättung minimiert in einem ersten Schritt $\sum_{j=1}^{n} \left(y_j - P_k(x_j)\right)^2$. Durch die partiellen Ableitungen nach den Koeffizienten des Polynoms P führt dies auf das simultan zu lösende Gleichungssystem

$$\sum_{j=1}^{m} x_j^l y_j \stackrel{!}{=} \sum_{j=1}^{m} x_j^l P_k(x_j)$$

  mit $l = 0, \ldots, k$, mit $x_j^l = 1$ für $l = x_j = 0$. Dies ist äquivalent mit

$$\begin{aligned} \mathcal{X}'y &= \mathcal{X}'\mathcal{X}a \\ a &= (a_0, \ldots, a_k)' \\ y &= (y_0, \ldots, y_k)' \\ \mathcal{X}_{u,v} &= x_u^{v-1} \end{aligned} \tag{2.1}$$

  mit $u = 1, \ldots, m$ und $v = 1, \ldots, k+1$. Für invertierbare Matrizen $\mathcal{X}'\mathcal{X}$ folgt aus (2.1) unmittelbar $a = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}y$ und die Koeffizienten für $P_k(y)$ können direkt berechnet werden.
  Um die polynomiale Glättung für einen Datenpunkt y_j zu vereinfachen, seien ohne Beschränkung der Allgemeinheit $j = 0$ und die x-Variablen äquidistant mit Abstand d gegeben, sowie die Datenpunkte $y_{-m}, \ldots, y_m$ und die Datenpunkte werden um x_0 zentriert und mit d normalisiert.
  Das für die Glättung relevante Polynom vom Grad k kann somit als $\widetilde{P}_k(\tilde{x}) = \sum_{j=1}^{k} \tilde{a}_j \cdot \tilde{x}^j$ mit $x \in \{-m, \ldots, m\}$ geschrieben werden. Mit dieser Umkodierung gilt $\sum_{x=-m}^{m} x^{2q+1} = 0$ für $q \in \mathbb{N}$, und es kann wegen der einfachen Berechnungsmöglichkeit der Terme $\sum_{x=-m}^{m} x^{2q+1}$ geschlossene Formeln für die Koeffizienten ermittelt werden. Mit $\widetilde{P}_k(0) = \tilde{a}_0$ folgt der polynomial geglättete Wert für y_0.

Verallgemeinert man dies mithilfe des Gleichungssystems in (2.1), so ergeben sich die Glättungen aller Punkte y_j für beispielsweise $m = 2$ und $k = 3$:

$$y_j^{\text{geglättet}} = \frac{1}{35}\left(-3y_{j-2} + 12y_{j-1} + 17y_j + 12y_{j+1} - 3y_{j+2}\right).$$

Grundsätzlich müssen die verwendeten Punkte für die Glättung nicht symmetrisch um den zu schätzenden Wert gewählt werden. Da in dieser Arbeit genügend Prädiktoren zur Verfügung stehen und ohnehin nur der Wellenzahlbereich von $[900, 3\,100]\,\text{cm}^{-1}$ verwendet wird, wird an dieser Stelle nicht näher darauf eingegangen.
Um die polynomiale Glättung mit der Differentiation zu kombinieren, kann, anstelle des Wertes $\widetilde{P}_k(0)$ die $s$te Ableitung des Glättungspolynomes an der Stelle 0

$$d^{-s}\left.\frac{\partial^s \widetilde{P}\,(\widetilde{x})}{\partial \widetilde{x}^s}\right|_{\widetilde{x}=0} = d^{-s}(s-1)! \cdot \widetilde{a}_s,$$

skaliert mit dem Differentiationsabstand d^{-s}, verwendet werden.
Für das Beispiel $m = 2$, $k = 3$ ergeben sich somit folgende Gleichungen:

$$y_j^{s=1,\text{geglättet}} = \frac{1}{12d}\left(-y_{j-2} - 8y_{j-1} + 8y_{j+1} - 1y_{j+2}\right)$$
$$y_j^{s=2,\text{geglättet}} = \frac{2}{7d^2}\left(2y_{j-2} - 1y_{j-1} - 2y_j - 1y_{j+1} + 2y_{j+2}\right),$$

wobei y_j selbst keinen Einfluss auf die Glättung mit der ersten Ableitung hat.

In Abbildung 2.11 sind die unterschiedlichen, in dieser Arbeit verwendeten, Vorbehandlungsarten abgebildet. Durch das Glätten werden kleinere Peaks abgeflacht und durch das Differenzieren kann ein möglicher Drift abgemildert werden. Es müssen allerdings zwei Punkte bei der Betrachtung dieser Plots berücksichtigt werden. Einerseits ist die Skalierung der Absorptionsachse für die erste bzw. zweite Ableitung nicht für einen direkten Vergleich mit der dem Plot rechtsstehenden Grafik vergleichbar und andererseits muss beachtet werden, dass ein eventuell vorhandenes Rauschen durch das Differenzieren verstärkt wird.

Zusätzlich kann gesagt werden, dass durch das Vorbehandeln mittels Differenzen den Banden eine höhere Bedeutung zugeschrieben werden kann, wohingegen leichte Anstiege weniger signifikante Rollen erhalten. Man vergleiche hierzu das Verhältnis in den Bereichen der Wellenzahlen $2\,400\,\text{cm}^{-1}$ bis $2\,800\,\text{cm}^{-1}$ zum Bereich $2\,900\,\text{cm}^{-1}$ bis $3\,000\,\text{cm}^{-1}$.
Mit der Behandlung ergeben sich auch Verschiebungen in der spektralen Korrelation, welche teilweise relativ stark zu oszillieren beginnt. Da diese grundsätzlich keinen Einfluss auf das Entwickeln der Modelle hat, sondern lediglich zur Interpretation der selektierten Wellenzahlen herangezogen werden kann, wird an dieser Stelle nicht weiter darauf eingegangen.

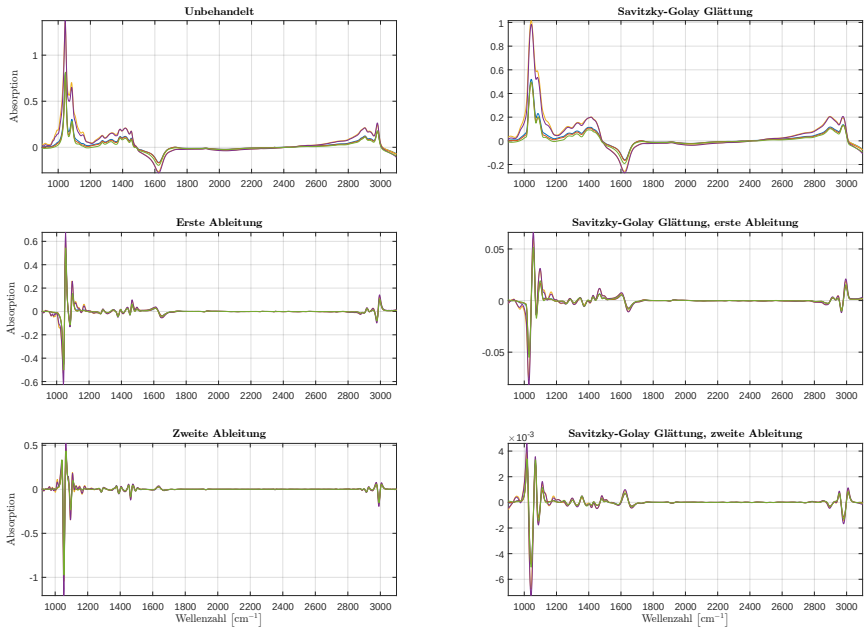

Abbildung 2.11: Vergleich von unterschiedlichen Preprocessingarten.
Links: Unbehandelte Spektren, sowie deren (unskalierte) erste und zweite Ableitung.
Rechts: Beispiel für eine Savitzky-Golay Ableitung vom Grad 0, 1 und 2.

Praktische Anwendungen des Preprocessings

Es werden jeweils Modelle entwickelt, welchen unterschiedlich aufbereitete Daten zugrunde liegen. Einerseits mit unbehandelten Spektren, Ableitungen erster und zweiter Ordnung, jeweils mit einem Gap von 5, sowie den Savitzky-Golay der 0ten (Datenglättung), 1ten und 2ten Ordnung (Glättung und Ableitung), mit den Parametern 30 für die Datenpunkte, welche zur Glättung verwendet werden, sowie einem Polynom von Grad deg $= 3$. Einerseits zeigen diese Werte eine plausible Stabilisierung im Ergebnis, und es zeigt sich, dass lediglich die Ableitung und deren Ordnung einen signifikanten Einfluss auf das Ergebnis aufweist. Insgesamt erzielt man mit den Savitzky-Golay Ableitungen erster oder zweiter Ordnung die besten Resultate, weswegen im weiteren Verlauf sich diese Arbeit auf die Savitzky-Golay Ableitung (erster oder zweiter Ordnung) fokusiert.

Alternative Glättungsmethoden

Alternativ können auch Glättungsmethoden wie der Moving Average oder eine lowess-Glättung verwendet werden. Für weitere Informationen hierfür wird auf [23] verwiesen, da in dieser Arbeit lediglich die sechs Aufbereitungsmöglichkeiten aus Abbildung 2.11 betrachtet werden.

Standardisierung unterschiedlicher Datensätze

Um zeitlich versetzt gemessene Datensätze vergleichen zu können, muss, abgesehen von der Eliminierung etwaiger Trends mithilfe von polynomialer Glättung bzw. durch Differenzierung der Spektren sowie der Peakadjustierung die Absorptionsstärke standardisiert werden.

Abbildung 2.12 stellt die Absorptionen der Spektren exemplarisch für einen zufällig gewählten Wein von Engine V70(2015) und V70(2016) gegenüber. Für die Wellenzahlen im Bereich $[900, 3\,100]\,\mathrm{cm}^{-1}$ scheinen die Werte der Prädiktorenmatrizen X_{E22} und X_{E25} ein lineares Verhalten aufzuweisen. Jene außerhalb dieses Wertebereiches, sowie jene, welche aus dem Bereich der CO_2-Bande stammen, scheinen einem zufälligen Muster zu folgen, wurden jedoch ohnehin als zu vernachlässigender Bereich identifiziert und spielen für weiterführende Überlegungen und praktische Anwendungen keine Rolle.

Daher werden, um Weinmessungen zu standardisieren, diese mittels Algorithmus 1 angepasst. Signifikante Peaks, wie beispielsweise jene in den Bereichen von $1\,044\,\mathrm{cm}^{-1}$ oder $2\,900\,\mathrm{cm}^{-1}$, werden als Bezugspunkte festgelegt und mit den Absorptionswerten an diesen Stellen die Parameter für eine lineare Transformation eines Spektrums angepasst. Auf diese Weise ist es möglich, beispielsweise die Messungen V70(2015) an jene von V70(2016)anzugleichen und die Güte der Datensätze zu vergleichen.

Da das primäre Ziel die Entwicklung von Modellen für die Engines der zweiten Generation ist, sowie ein Vergleich dieser Datensatzqualität, wird dieses Prozedere

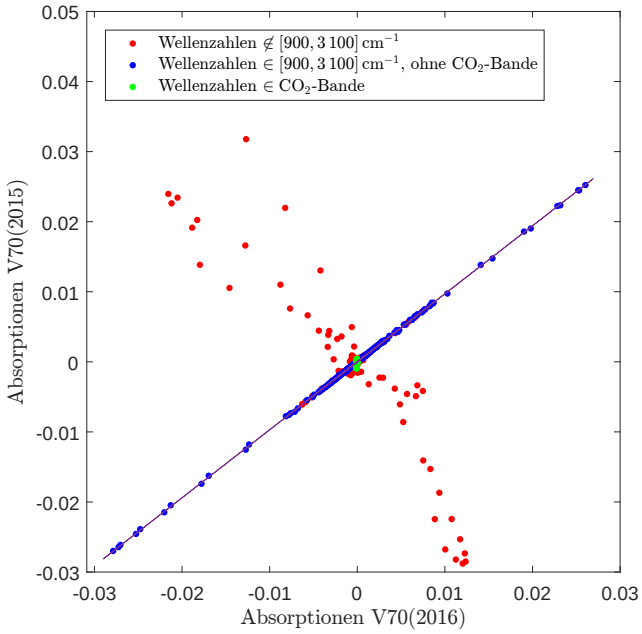

Abbildung 2.12: Die Absorptionswerte desselben Weines, gemessen mit dem Spektrometer E22 sowie E25 gegeneinander aufgetragen.

Algorithmus 1 Spektrenadjustierung

1: **procedure** MY.RD_CV
2: **Input:** $X^1_{n\times p}$, $X^2_{n\times p}$
3: Selektiere Messungen derselben Probe:
$X^1\left(idx^1_{\mathrm{M}},:\right)$ und $X^2\left(idx^2_{\mathrm{M}},:\right)$
4: Selektiere zu vordefinierten Peaks gehörende Prädiktoren
idx^1_{P} und idx^1_{P}
5: Bestimme Koeffizienten β_0 und β_1 für eine linearen Transformation von X^2 auf X^1 mit den Werten $X^1\left(idx^1_{\mathrm{M}}, idx^1_{\mathrm{P}}\right)$ und $X^2\left(idx^2_{\mathrm{M}}, idx^2_{\mathrm{P}}\right)$
6: **return** X^1 und $X^2_{\mathrm{adj}} := \beta_0 + \beta_1 \cdot X^2$
7: **end procedure**

nicht gesondert behandelt. Bei der Übertragung eines Modells von V70(2016) auf den Datensatz V70(2015) zeigen sich leicht schlechtere Kennzahlen der Ergebnisse. Zudem werden die Datensätze der zweiten Generation aufgrund derselben Weinpopulation lediglich mit V70(2016) verglichen, um eine Einschätzung der Datenqualität zu erhalten und V70(2015) wird trotz der hierfür eigens entwickelten Modelle nicht gesondert analysiert.

Das Angleichen von Spektren unterschiedlicher Spektrometer scheint mit dieser Methode nicht zu funktionieren und gleicht einer Zweckentfremdung, da nicht modellspezifische Eigenschaften angeglichen werden sollen. Trotz alledem stellt dieser Algorithmus eine durchaus effiziente und notwendige Möglichkeit dar, künftige Messungen zu standardisieren, wie unter anderem in [31] erläutert wird.

3 Modellvalidierung und -selektion

Um eine Verbindung zwischen Messdaten in Form von Prädiktorvariablen und daraus abgeleiteten Kennzahlen herzustellen, steht eine Vielzahl von Modellierungsmöglichkeiten zur Verfügung. Für diese müssen unter Umständen etliche variable Parameter bestimmt (geschätzt) werden, wobei deren tatsächliche Werte für praktische Anwendungen in der Regel unbekannt sind. Daher wird in diesem Abschnitt die Kreuzvalidierung vorgestellt, um die Güte eines Modells bewerten und diese vergleichen zu können.

Die Grundidee hierfür ist, dass man zwei oder drei, nach Möglichkeit unabhängige, Datensätze besitzt. Ein Trainingsset zur Bestimmung der Modellparameter (wie Intercept und Slope in der linearen Regressionsanalyse), ein Validierungsset um das Modell zu validieren, sowie ein davon unabhängiges Testset, um die Güte für neue Testdaten zu überprüfen. Hierbei wird durch Anwendung des durch das Trainingsset bestimmte Modell auf die Validierungsdaten ein Modell selektiert. Das Testset, sofern vorhanden, dient lediglich zur Messung der endgültigen Güte des Modells und darf, um die Unabhängigkeit dieses Fehlers zu wahren, nicht zur Entscheidungsfindung herangezogen werden. Darüber hinaus kann dieses Vorgehen zur Verminderung der Modellkomplexität, durch Reduktion der Anzahl an Parametern, führen. Eine relativ einfache Möglichkeit, dieses Vorgehen in die Praxis umzusetzen stellt die Kreuzvalidierung dar und es werden in dieser Arbeit drei unterschiedliche Varianten vorgestellt und angewendet:

- Zufällige Auswahl des Validierungssets, die Monte Carlo Kreuzvalidierung.
- Die gleichmäßige Partitionierung des Datensatzes in k gleich große Segmente, die k-fache Kreuzvalidierung mit dem bekannten Vertreter der Leave One Out Kreuzvalidierung
- Eine Kombination dieser beiden Methoden. Zusätzlich wird hierfür ein Pseudocode angegeben, welcher nach deterministischen Kriterien zusätzlich eventuelle Modellparameter selektiert. Diese Selektionsmethode kann in gleicher Weise auf die ersten beiden Kreuzvalidierungsformen übertragen werden und nennt sich Wiederholte doppelte Kreuzvalidierung.

Seien insgesamt n Datensätze gegeben.

- **MC Kreuzvalidierung**[1]: Bei der Monte-Carlo Kreuzvalidierung wird eine Stichprobe vom Umfang n_{MC} aus den gegebenen n Objekten zufällig gezogen. Mit den verbliebenen $(n - n_{\mathrm{MC}})$ Proben werden die variablen Modellparameter (Gewichte, Koeffizienten) bestimmt. Anschließend werden

[1] Vgl. hierzu [33]

Schätzungen für die Responsevariablen y_i der zufällig gezogenen n_{MC} Objekte berechnet ($\hat{y}_i, i = 1, \ldots, n_{MC}$) und zur Validierung des betrachteten Modells verwendet. Hierzu kann die mittlere quadratische Abweichung $MSE_{MC} = \frac{1}{n_{MC}} \sum_{i=1}^{n_{MC}} (y_i - \hat{y}_i)^2$ oder die Standardabweichung der n_{MC} Residuen als Gütemaß verwendet werden.

Diese Validierung wird R mal wiederholt und der Mittelwert dieser Kennzahlen gebildet. Dieser Wert wird in dieser Arbeit als Validierungsfehler (der Monte Carlo Kreuzvalidierung) bezeichnet. Unter Verwendung des MSE führt dies auf

$$MSECV_{MC} = \frac{1}{R} \sum_{i=1}^{R} MSE_{MC}^{(i)}, \tag{3.1}$$

wobei R Replikationszahl genannt wird.

- **k-fache Kreuzvalidierung**[2]: Hierzu werden die n Datensätze in k Untergruppen, nach Möglichkeit gleicher Größe, unterteilt. So werden $\lfloor \frac{n}{k} \rfloor$ Stichproben zu einem Segment zusammengefasst. Falls $k \nmid n$, müssen die restlichen $n - k\lfloor \frac{n}{k} \rfloor$ Datensätze gleichermaßen auf die k Gruppen aufgeteilt werden. Auf diese Art ist es möglich, k Segmente $S_i, i = 1, \ldots, k$ zu bilden, wobei für Anzahl $\big||S_i| - |S_j|\big| \in \{0, 1\}\ \forall i, j = 1, \ldots, k$ gilt.

  Mit diesen Segmenten wird analog zur MC Kreuzvalidierung verfahren. Die Replikationszahl entspricht hierbei k und jedes dieser Segmente dient je einmal als Validierungssegment während die restlichen $(k-1)$ Partitionen für das Bestimmen der Parameter verwendet werden.
  Analog definiert sich der MSECV für die k-fache Kreuzvalidierung

$$MSECV_{kCV} = \frac{1}{k} \sum_{i=1}^{k} MSE_{kCV}^{(i)}. \tag{3.2}$$

  Die Datensätze können einerseits zufällig den k Segmenten zugeteilt werden, oder alternativ kann darauf geachtet werden, dass alle Segmente eine ähnliche Verteilung der zu modellierenden Variablen aufweisen, oder es kann auch eine systematische und deterministische Selektion erfolgen, wie in der LOO Kreuzvalidierung.

- **LOO Kreuzvalidierung**: Eine bekannte Untergruppe der k-fachen Kreuzvalidierung stellt die „Leave One Out" Kreuzvalidierung dar. Diese entspricht dem Spezialfall $k = n$. Es werden n Modelle betrachtet mit je $n - 1$ Kalibrierungsobjekten und je einer Validierungsprobe und wiederum gemittelt.

- **Randomisierte doppelte Kreuzvalidierung**: Diese beruht auf der wissenschaftlichen Arbeit [6] und kombiniert die beiden erstgenannten Kreuzvalidierungen. Der zusätzliche Benefit ist, dass eine vom Validierungs- und Trainingsset unabhängige Kennzahl für die Güte des Modells für Testdaten,

[2] Vgl. hierzu [10], Seite 241ff

welche während der Kalibrierung nicht zur Verfügung stehen, mitberechnet wird. Hierbei wird angenommen, dass die zugrunde liegenden Daten repräsentativ sind und ein neues Objekt aus der vorhandenen Population stammt.

Sei eine Menge $\mathcal{M}$ von m Parametersets gegeben, sowie die Anzahlen der Partitionen für die Kreuzvalidierungen, sowie einem Faktor π[3] als Standardabweichungsindikator zur Wahl eines optimalen Testsets aus M. Zunächst werden die Datensätze zufällig r_{ts} Segmenten zugeordnet. Es ist darauf zu achten, dass diese nach Möglichkeit annähernd gleich groß sind (Zeile 4). Jede dieser Partitionen nimmt je einmal die Aufgabe des Testdatensatzes ein. Mit den restlichen wird mittels **for** Schleife ab Zeile 9 eine r_{cs}-fache Kreuzvalidierung durchgeführt und anschließend wird ein optimales Parametersetting bestimmt.

Definition 3.1 (π-Standardabweichung Regel zur Wahl der Parametersettings)**.** Die π-Standardabweichung Regel betrachtet das globale Minimum eines Problemes mit einer definierten Evaluierungsfunktion wie dem MSE und sucht sich jene Modelle, welche sich in einer „gewissen" Nähe dieses globalen Minimums befinden: die Evaluierungen der in Betracht kommenden Modelle dürfen maximal um den additiven Wert $\pi \cdot \frac{std_{\text{MSE,min}}}{r_{\text{cs}}}$ abweichen, wobei $std_{\text{MSE,min}}$ die Standardabweichung der Residuen für jenes Modell mit dem globalen Minimum repräsentiert. Bei den hieraus resultierenden Kandidaten für ein optimales Testset wird jenes mit der geringsten Komplexität ausgewählt.
Für $\pi = 0$ wird eben das global beste Modell gewählt, während die Wahl der Modelle mit $\pi = 2$ mit einer Art 95%-Konfidenzintervall des MSE korrespondiert.

Diese Regel ist heuristisch motiviert. Daher kann auch keine allgemein gültige Vorschrift zur Bestimmung von π angegeben werden, sondern muss an die jeweiligen Gegebenheiten angepasst werden. Für die Modellwahl in dieser Arbeit findet stets die $(\pi = 1)$-Standardabweichung Regel Anwendung, wie diese auch in [10] eingeführt wird.

Resultate der Kreuzvalidierung

Sofern sämtliche Zwischenergebnisse mitgespeichert werden, stehen am Ende der Monte Carlo Kreuzvalidierung sowie k-fache Kreuzvalidierung die geschätzten Responsevariablen mit den dazugehörenden Residuen für weitere Analysen zur Verfügung. Zusätzlich kann mit den Residuen des Validierungssets ein (Validierungs-)Fehler ermittelt werden, welcher schlussendlich die Wahl des Modells entscheidet.

Für die doppelte Kreuzvalidierung werden insgesamt r_{ts} disjunkte Testsets mit größtenteils nicht disjunkten Kalibrierungssets geschätzt. Wichtig in diesem Schritt

[3] Siehe Definition 3.1

Algorithmus 2 Doppelte Kreuzvalidierung

procedure MY.RD_CV
  Input: $X_{n\times p}, Y_{n\times q}$, Parametersets $M_{\#\text{Param.sets}\times\#\text{benotigte Param.}}$
  for $r \in \{1,\ldots,R\}$ **do**
    Splitte alle Daten zufällig in r_{ts} (nach Mögl.) gleichgroße Partitionen $P_i, i = 1,\ldots,r_{\text{ts}}$
    for $t \in \{1,\ldots,r_{\text{ts}}\}$ **do**
      $TP := P_t$ ▷ Testpartition
      $CP := P_{\{1,\ldots,r_{\text{ts}}\}\setminus t}$ ▷ Kalibrierungspartition
      Splitte CP zufällig in r_{cs} (nach Mögl.) gleichgroße Partitionen $CP_i, i = 1,\ldots,r_{\text{cs}}$
      for $s \in \{1,\ldots,r_{\text{cs}}\}$ **do**
        $VP := CP_s$ ▷ Validierungspartition
        $TrP := \text{Partition}_{\{1,\ldots,r_{\text{cs}}\}\setminus s}$ ▷ Trainingspartition
        Berechne Residuen für alle Parametersets in M
        Berechne jeweils die Güte der Residuen von VP (hier MSE) für alle Parametersets
      end for
      BerechneOptParameterSet($MSE_{r_{\text{cs}}\times\#\text{Param.sets}}$)
      Bestimme Modell mit optimalem Parameterset und mit Kalibrierungspartition als Trainingsdatensatz CP
      Wende Modell aus Zeile 16 auf Testset TP an: Güte des Modells
    end for
  end for
  return Opt. Parametersets, Residuen und Güte des Modells
end procedure

function BerechneOptParameterSet($MSE_{r_{\text{cs}}\times\#\text{Param.sets}}$)
  $m :=$ #Param.sets
  Spaltenweise Mittelwerte $m_{MSE} \in \mathbb{R}^{1\times m}$ von MSE
  $m_{MSE,\min} := \min\{m_{MSE}\}$ wird angenommen an Position $i_{\min} \in \{1,\ldots,m\}$
  Std.abw. der Residuen für $m_{MSE,\min}$ sei $std_{MSE,\min}$ (entspricht Std.abw. der $i_{\min}$ten Spalte von MSE)
  Optimales Parametersetting s sei jenes Setting mit der geringsten Komplexität und $m_{MSE}(s) \leq m_{MSE,\min} + \frac{\pi\cdot std_{MSE,\min}}{r_{\text{cs}}}$ erfüllt.
  return Setting s
end function

ist, dass das jeweilige Testset keine Anwendung in der Parametrisierung des Modells findet. Für jede Kalibrierung wird ein optimales Parametersetting bestimmt, was für die Partial Least Square Methode der Anzahl der latenten Variablen entspricht, sofern die Modelle mit denselben Prädiktoren berechnet werden. Dies führt auf insgesamt $R \times r_{\text{ts}}$ optimale Parametersettings, wobei die endgültige Festlegung der Parameterwerte aus diesen ermittelt werden muss. Dies kann beispielsweise durch Betrachtung der Auftrittswahrscheinlichkeiten der Settings bewerkstelligt werden. Der endgültige Testfehler kann infolge dessen anhand von $n \times R$ Residuen ermittelt werden.

Mit einer $n \times$ #Param.sets $\times R$ dimensionalen Matrix an Vorhersagewerten oder Residuen können statistisch wichtige Eigenschaften der Modelle ausgedrückt und veranschaulicht werden, wie

- das Verhalten der Standardabweichung der Residuen in Abhängigkeit der Parametersettings.
- die Verteilung der $n \times R$ Residuen als Qualitätsmaß für das Modell und das Ermitteln von etwaigen Messfehlern oder Ausreißern.
- die Verteilung der R Modellevaluierungen anhand der n Residuen.

Variablenreduktion

In Tabelle 2.4 wurden die verfügbaren Prädiktoren in Form von Spektren vorgestellt. Da für diese Art von Prädiktoren in der Praxis, sowie in den vorliegenden Datensätzen, gilt, dass die Anzahl der analysierten Objekte n weit kleiner ist als die Anzahl der gemessenen Wellenzahlen s, muss sowohl aus mathematischer Sicht, als auch aufgrund von chemisch-physikalischen Überlegungen deren Anzahl eingeschränkt werden, da nicht sämtliche Absorptionen in Zusammenhang mit dem zu prognostizierenden Zielwert stehen.

In dieser Arbeit wird angenommen, dass sich relevante Information in gewissen Intervallen vorfinden lässt. Beispielsweise ist bekannt, dass die CH-Bindungen von Ethanol unter anderem im Wellenzahlbereich von $[2\,900, 3\,100]\,\text{cm}^{-1}$ stärker schwingen, was sich wiederum in den Spektren bemerkbar macht und in diesem Bereich somit relevante Informationen über die Ethanolkonzentration zu finden sind. Deswegen wird versucht, Wellenzahlbereiche zu ermitteln, welche sich diese Eigenschaft zu nutze machen: es wird eine unbekannte Anzahl an Intervallen von unbekannter Länge gesucht, um einen mathematischen Zusammenhang zu den einzelnen Responsevariablen zu modellieren.

Ein Indikator für die Relevanz von Wellenzahlen stellt unter anderem die spektrale Korrelation dar. Hierunter versteht man den Korrelationskoeffizienten zwischen der Responsevariable und den dazugehörenden Spalten der Prädiktormatrix. Da dieser lediglich den linearen Zusammenhang misst und durchaus Nichtlinearitäten in den Daten erwartbar sind, wird hier auf das Heranziehen dieser Korrelation verzichtet und eine alternative Heuristik zur Wellenzahlfindung vorgestellt. Dies kann gleichermaßen auf alle mathematischen Modelle angewendet werden

- unabhängig davon, ob es sich um Regressions-, Partial Least Squares Modelle, Neuronale Netzwerke oder andere handelt.

Seien hierzu die Parametersettings für die betrachtete Modellklasse, sowie folgende drei Parameter gegeben:

- k: Die maximale Anzahl von zu selektierenden Intervallen.
- l: Die Länge der gesuchten Intervalle (Anzahl, wieviele Wellenzahlen ein Intervall umfasst).
- $m \in \mathbb{N}_{>0}$: Ein zusätzlicher Parameter, um die Laufzeit zu verringern.

Die Grundidee basiert auf dem Prinzip, sämtliche Teilintervalle der Länge l zu betrachten und die jeweils besten zu selektieren. Seien die Wellenzahlen nach Größe sortiert und von 1 bis p durchnummeriert. Zunächst werden die Wellenzahlen 1 bis l selektiert, mit diesen ein Modell erstellt und mittels Kreuzvalidierung evaluiert. Dies wird für alle Intervalle mit den Prädiktoren $(s \cdot m + 1)$ bis $\min\{s \cdot m + l, p\}$ für $s \in \left\{1, \ldots, \left\lceil \frac{p-l}{m} \right\rceil \right\}$ wiederholt. Zusätzlich muss darauf geachtet werden, dass die Wellenzahlen zusammenhängend sind. Werden wie auf Seite 20 einzelne Wellenzahlbereiche von der Selektion ausgeschlossen, so darf dieser exkludierte Bereich keine Teilmenge eines Intervalls der hier für die Kreuzvalidierung betrachteten Intervalle bilden und zusätzlich muss beachtet werden, dass die, an einen ausgeschlossenen Bereich bei der Implementierung angrenzenden Intervalle nicht als zusammenhängend betrachtet werden dürfen. Jenes Intervall, welches zu dem geringsten Validierungsfehler führt, wird in die Wellenzahlselektion für das endgültige Modell aufgenommen. Dieses Vorgehen wird k Mal wiederholt, wobei in Iterationsschritt $(w > 1)$ die $(w - 1)$ bereits ausgewählten Teilbereiche für die Kreuzvalidierung mitberücksichtigt werden. Falls sich durch die Hinzunahme eines Intervalles in der Iteration w der Länge l keine Verbesserung realisieren lässt, so bricht der heuristische Algorithmus ab und terminiert mit den bis zur Iteration $w - 1$ gefundenen Wellenzahlbereichen.

Der Grund, weshalb Intervalle von Wellenzahlen betrachtet werden liegt darin, dass die physikalische bzw. chemische Information nicht punktuell an einzelnen Messpunkten beobachtet wird, sondern dass die unterschiedlichen kovalenten Bindungen in einem Bereich von Wellenzahlen, in sogenannten Banden, eine für die Verbindung typische Absorption aufzeigt. Nach Möglichkeit, sowohl zur Interpretation als auch wegen der Sinnhaftigkeit aus physikalisch/chemischer Sicht sollten daher nicht zu viele Teilbereiche als Optimallösung betrachtet werden, um auch Einflüsse von nicht relevanten kovalenten Bindungen auf die Vorhersage der analysierten Responsevariable zu verringern.

Implementierung

Die Werte k und l, mit welchen die hier entwickelten Modelle betrachtet werden, können Tabelle 3.2 entnommen werden.

Der Parameter m dient insbesondere dem Verschieben der Intervallselektion und somit der Laufzeitverbesserung, wie Tabelle 3.1, zeigt. Hier werden unterschiedliche

Modelle, welche in dieser Arbeit Anwendung finden, bezüglich der Laufzeit, in Abhängigkeit von m verglichen und es zeigt, dass die Wahl von $m = 1$ und somit der höchstmöglichen Genauigkeit für die Wellenzahlselektion, aus Sicht der Dauer nicht gerechtfertigt ist. Insbesondere muss beachtet werden, dass zahlreiche solcher Auswertungen durchgeführt werden.

		m		
		1	5	10
PLS-Modell	linearer Kernel	6 min 30 s	1 min 18 s	40 s
PLS-Modell	Gaußkernel	9 min 14 s	1 min 51 s	55 s
Neuronales Netzwerk		10 h 57 min	1 h 56 min	1 h 5 min

Tabelle 3.1: Laufzeitenvergleich bei der Selektion von zwei Intervallen mit je 10 Wellenzahlen, zwei latenten Variablen und einem versteckten Neuron für Ethanol mit einer 100-fachen Monte Carlo Kreuzvalidierung für den Datensatz E25.

Um einen vernünftigen Trade-Off zwischen Berechnungszeit und Anzahl betrachteter Intervallbereiche zu erreichen, wurde der Parameter $m = 10$ gesetzt. Dies bedeutet eine Beschleunigung der ohnehin lange dauernden Prozedur.

Insbesondere in Hinblick auf die doppelte Kreuzvalidierung, scheint es unerlässlich, für eine Wellenzahlselektion die Anzahl der Stützstellen, d.h. der Anzahl der Spalten von X, zu reduzieren. Dies kann aufgrund der hohen Korreliertheit von Messwerten bei benachbarten Wellenzahlen ohne relevanten Wertverlust durchgeführt werden. Hierfür stehen zwei Möglichkeiten zur Verfügung. Einerseits eine deterministische Aussortierung von Informationen, wie beispielsweise der Wahl jeder dritten Spalte von X oder durch Interpolation auf Wellenzahlen mit einem gröberen Raster. Diese Methode kommt in dieser Arbeit zur Anwendung, indem die Messwerte der Datensätze des Jahres 2016 auf jene Stützstellen des Datensatzes 2015 linear interpoliert werden. Auf diese Weise geht kaum relevante Information verloren.

Die in dieser Arbeit verwendeten unterschiedlichen Kombinationen (k, l) werden wie in Tabelle 3.2 festgesetzt. Dies ermöglicht sowohl die Verwendung von sehr breiten Banden mit bis zu 100 Wellenzahlen, sowie mehrere kurze Bereiche der Länge 3.

Eine Vorselektion wurde mit der klassischen Monte Carlo Kreuzvalidierung für alle betrachteten Vorbehandlungen und unterschiedlichen Parametersettings durchgeführt. Für die in dieser Arbeit verwendeten Modelle (sowohl PLS als auch neuronale Netzwerke) bedeutet dies, dass

Anzahl Preprocessing-Arten (6)
- $\times$ Maximale Anzahl an latenten Variablen/versteckten Neuronen (8)
- $\times$ Wellenzahlkalibrierung laut Tabelle 3.2
- $\times$ Kernelarten (linear, Gauß, polynomial) mit dazugehörenden unterschiedlichen Parametersettings/unterschiedlicher Transferfunktionen etc.
- $\times$ Gesamter Datensatz vs. nur Rotweine

potentielle Wellenzahlbereiche vorselektiert wurden.

In einem nächsten Schritt wurde lediglich die Preprocessing-Art auf Savitzky-Golay eingeschränkt und anhand der bis hier hin beste dieser beiden Methoden gewählt.[4] Die Wellenzahlbereiche werden anschließend mithilfe der wiederholten doppelte Kreuzvalidierung weiter reduziert, indem der Punkt „Wellenzahlkalibrierung laut Tabelle 3.2“ damit eliminiert wird. Mit den aus der doppelten Kreuzvalidierung resultierenden Kennzahlen werden für die betrachtete Anzahl von latenten Variablen/versteckten Neuronen, die gemäß dieser Prozedur besten Wellenzahlbereiche zugeordnet. Während bis zu diesem Schritt die Reduktionen automatisiert stattfinden, wird aus den verbliebenen Optionen individuell entschieden, welches konkrete Modell vorgeschlagen werden kann. Die ausschlaggebenden Argumente sind hier der Trade-Off zwischen Komplexität des Parametersettings und der Verbesserung der Residuen sowie das Verhalten der Residuen selbst (keine Trends in den Residuen o.ä.).

Um einen Kompromiss zwischen Laufzeitoptimierung und Aussagekraft der Resultate zu erhalten, wurden in dieser Arbeit $R = 100$ für die kPLS-Modelle in Kapitel 4 und $R = 50$ für die Neuronalen Netzwerke in Kapitel 6 verwendet. Für die Anzahl der Partitionierungen wurde der Empfehlung in [6] gefolgt und auf $(r_{ts}, r_{cs}) = (4, 7)$ gesetzt.

Max. Anzahl an Intervallen	2	2	2	3	4	4
Wellenzahlen pro Intervall	25	50	10	3	5	10

Tabelle 3.2: Für diese Arbeit als 2er-Tupel definierte Kombinationspaare zur Wahl von Wellenzahlbereichen, wobei die maximale Anzahl als erlaubte Obergrenze von Wellenzahlbereichen implementiert zu verstehen ist.

Bei der manuellen Auswertung der Modelle werden jeweils die Unauffälligkeit der Residuenplots, sowie der Grafiken, welche aus der doppelten Kreuzvalidierung stammen, verwendet. Zudem muss die Reproduzierbarkeit verglichen werden. Hierfür stehen die vier Messungen pro Wein ID zur Verfügung. Hierbei wird einerseits auf die maximale Abweichung geachtet, sowie auf die Standardabweichung dieser vier Weinproben. Hierbei kann nicht von einem aussagekräftigen Standardabweichungsschätzer gesprochen werden. Diese Kennzahl dient lediglich der Information, ob Weine vorliegen, bei welchen starke Unregelmäßigkeiten beobachtbar sind.

[4] Es wurden im Schritt zuvor dennoch sämtliche Preprocessing-Arten behandelt, aber es stellte sich heraus, dass die Savitzky-Golay Methode die effektivste zu sein scheint.

4 Partial Least Square Regression

4.1 Lineare Partial Least Square Regression

Bei den Partial Least Square Regression Algorithmen wird versucht, einen Zusammenhang zwischen ein oder mehreren Responsevariablen $Y \in \mathbb{R}^{n \times q}$ und dazugehörenden Prädiktorvariablen $X \in \mathbb{R}^{n \times s1}$ mittels Least Squares herzustellen. Falls mehrere abhängige Variablen simultan betrachtet werden, so spricht man von PLS2, andernfalls von PLS1.

Sowohl X, als auch Y werden zentriert. Dies wird jeweils durch Subtraktion der Spaltenmittelwerte erreicht (Gleichungen (4.1)). Die Zentrierung der Prädiktorvariabeln dient der Herleitung des Optimierungproblemes (4.3) und zeitgleich der Interpretierbarkeit, da sich alle Abweichungen gleichzeitug auf den Referenzwert 0 beziehen. Vergleiche hierzu [27].

$$\begin{aligned} X^{(z)} &= \left(I_n - n^{-1}1_{n\times n}\right) \cdot X \\ Y^{(z)} &= \left(I_n - n^{-1}1_{n\times n}\right) \cdot Y, \end{aligned} \tag{4.1}$$

wobei I_n die n-dimensionale Einheitsmatrix und $1_{n\times n}$ die $n \times n$-dimensionale Matrix mit 1 Einträgen bezeichnen.

Im Großen und Ganzen wird bei der PLS-Regression eine Hauptkomponentenanalyse durchgeführt, wobei zusätzlich zu den Informationen der Prädiktorvariablen auch jene der Responsevariablen berücksichtigt werden.

In einem ersten Schritt werden die Prädiktoren auf lineare Vektoren, ähnlich der Hauptkomponentenanalyse projiziert. Hierzu wird zunächst das Vorgehen anhand dieser erläutert und modifiziert. Diese Herleitung orientiert sich sehr stark an [3] und für weiterführende Informationen wird auf ebendiese Arbeit verwiesen. Sei das dazugehörige Optimierungsproblem in (4.2)

$$\begin{aligned} &\min_{w\in\mathbb{R}^n} \sum_{i=1}^{n} \|x_i - x_i w w'\|^2 \\ &\text{mit } w'w = 1 \end{aligned} \tag{4.2}$$

[1] Grundsätzlich gilt $s \gg n$.

gegeben, wobei x_i als Zeileneintrag der Prädiktormatrix X ein Spektrum repräsentiert, welches auf den Vektor w' projiziert werden soll. Es gilt:

$$\begin{aligned}\min_{w\in\mathbb{R}^n}\sum_{i=1}^{n}\|x_i - x_i ww'\|^2 &= \min_{w}\sum_{i=1}^{n}\left(x_i - x_i ww'\right)\left(x_i' - ww'x_i'\right)\\ &= \min_{w\in\mathbb{R}^n}\sum_{i=1}^{n}\left(\|x_i\|^2 - 2x_i ww'x_i' + x_i w\underbrace{w'w}_{=1}w'x_i'\right)\\ &= \min_{w\in\mathbb{R}^n}\sum_{i=1}^{n}\left(\|x_i\|^2 - (x_i w)^2\right)\end{aligned}$$

Somit ist die Optimallösung des Optimierungsproblems (4.2) äquivalent zur Maximierung der empirischen Varianz von Xw in (4.3).

$$\begin{aligned}&\max_{w\in\mathbb{R}^n}\ \mathrm{Var}\,(Xw)\\ &\text{mit } w'w = 1\end{aligned} \tag{4.3}$$

Aus der Mittenzentrierung von X folgt

$$\mathrm{Var}\,(Xw) = n^{-1}\sum_{i=1}^{n}(x_i w)^2 - n^{-1}\left(\sum_{i=1}^{n}x_i w\right)^2 = n^{-1}\sum_{i=1}^{n}(x_i w)^2.$$

Dieses Optimierungsproblem kann nun mithilfe der Lagrangefunktion gelöst werden und führt aufgrund der Optimalitätskriterien zur Lösung des Eigenwertproblemes $XX'w = \lambda w$ mit $ww' = 1$, wobei hieraus λ als größter Eigenwert von XX' resultiert.

Diese Prozedur wird nun dahingehend erweitert, sodass die Verbindung zu den Responsevariablen in die Berechnung mit einfließt. Sei hierfür zuerst $q = 1$. Falls der $i$te Eintrag der Responsevariable y durch den Term $x_i w$ approximiert wird, so kann $x_i w$ durch y_i substituiert werden. Diese Modifizierung der Gleichung (4.2) führt zu

$$\begin{aligned}&\min_{w\in\mathbb{R}^n}\sum_{i=1}^{n}\|x_i - y_i w'\|^2\\ &\text{mit } w'w = 1.\end{aligned} \tag{4.4}$$

Alternativ kann das Problem (4.4) als Regressionsmodell von X auf y in der Form $X = yw'$ zuzüglich Fehlerterm mit dazugehörender Nebenbedingung interpretiert werden und bildet eine Abschätzung der linearen Regression

$$\|x_i w - y_i\|^2 = \|(x_i - y_i w')\, w\|^2 \leq \|x_i - y_i w'\|^2 \underbrace{\|w\|^2}_{=1}. \tag{4.5}$$

Analog zu vorheriger Herleitung des Optimierungsproblemes (4.3) kann (4.4) in (4.6) überführt werden.

$$\begin{aligned} &\max_{w \in \mathbb{R}^n} Cov\,(Xw, y) \\ &\text{mit } w'w = 1 \end{aligned} \tag{4.6}$$

Anstelle der Varianz von Xw wird nun die Kovarianz zwischen Xw und y maximiert, was wiederum der Grundidee der Vorgehensweise der Partial Least Square Regression entspricht und resultiert in der latenten Variable $w = X'y\|X'y\|^{-2}$ als Lösung des Optimierungsproblems. Diese sogenannten w-Loadings repräsentieren eine Verbindung zwischen X und y und werden oftmals als Gewichts-Loadings bzw. PLS-Gewichte bezeichnet. Durch die Nebenbedingung $\|w\|^2 = 1$ wird sichergestellt, dass diese orthogonal aufeinander stehen, und können als erstes PLS-Modell mit einer latenten Variable interpretiert werden.

In Hinblick auf den Kernel PLS Algorithmus muss allerdings an dieser Stelle auf die Normierung durch $\|X'y\|^{-2}$ verzichtet werden.

Im nächsten Schritt werden die X Daten auf die w-Loadings regressiert. $X = tw' + \epsilon_{n \times p}$ liefert mit der Least Square Methode die sogenannten Scores $t = Xw\|Xw\|^{-2}$ und diese werden aufgrund der Restriktion $w'w = 1$ normiert. Zur Berechnung der finalen Loadings p für X muss daher das Optimierungsproblem (4.7) gelöst werden und führt auf $p = X't$, wie Lemma 4.1 zeigt. Man beachte an dieser Stelle, dass in der traditionellen Herleitung des PLS-Algorithmus $p = X't\|t\|^{-2}$ resultiert.

$$\min_p \|X - tp'\|^2 \tag{4.7}$$

Lemma 4.1. *Das Optimierungsproblem 4.7 besitzt die Optimallösung $p = X't$.*

Beweis.

$$\begin{aligned} & \frac{\partial}{\partial p_j} \|X - tp'\|^2 \stackrel{!}{=} 0 \\ \iff & \sum_{i=1}^{p} t_i X_{ij} - t_i^2 p_j \stackrel{!}{=} 0 && \forall j = 1, \ldots, p \\ \iff & \left(X't\right)_j \stackrel{!}{=} \|t\|^2 p_j && \forall j = 1, \ldots, p \end{aligned}$$

□

Dies liefert folgende Approximation von X im PLS-Modell

$$X \approx tp' = tt'X \tag{4.8}$$

mit den orthogonalen t-Scores. Nun soll auf analoge Weise die Approximation für y verbessert werden, indem ein Minimierungsproblem für y der Form $\min_c \|y - tc'\|$ gelöst wird und zur Lösung $c = t'y$ führt. Daher kann analog die Approximation

$$y \approx tc' = tt'y. \tag{4.9}$$

verwendet werden. Um die Anzahl der latenten Variablen zu erhöhen, kann obiges Vorgehen mit erneuerten Matrizen X und y nach den Vorschriften wie (4.10) bzw. (4.11) wiederholt werden. Die Erneuerungsvorschriften resultieren aus den Gleichungen

$$X - tt'X \to X \tag{4.10}$$
$$y - tt'y \to y, \tag{4.11}$$

und die Vektoren w, t und c werden als Spalten zu den Matrizen W, T bzw. C zusammengefasst.

In einem letzten Schritt des PLS-Algorithmus wird eine lineare Regression zwischen den approximierten X bzw. y Matrizen, resultierend aus (4.8) bzw. (4.9), erstellt. Dies führt mittels der Least Square Methode zu dem Optimierungsproblem

$$\min_v \|TP'Wv - TC'\|^2 = \min_v \|T(P'Wv - C')\|^2. \tag{4.12}$$

Dies führt zur Optimallösung $v^* = (P'W)^{-1}C'$[2] und somit zu den endgültigen Koeffizienten

$$b = Wv^* = W(P'W)^{-1}C' = W(P'W)^{-1}T'y. \tag{4.13}$$

Zusammengefasst resultiert hieraus der PLS1 Algorithmus. Für die Erweiterung des Algorithmus auf PLS2 kann das Loading w nicht durch $X'Y$ berechnet werden. Der Algorithmus wird so modifiziert, dass in einem Zwischenschritt ein Vektor u zufällig initialisiert und analog verfahren wird, wie nachfolgender Code zeigt. Die Zeilennummern entsprechen den zu Algorithmus 3 korrespondierenden Bezeichnungen.

Dies wird bis zur Konvergenz von t wiederholt oder alternativ bis ein Maximum an Iterationen erreicht ist. Die Konvergenz schlägt fehl, falls die Eigenwerte zu nahe aneinander liegen.

[2] $P'W$ entspricht einer nicht singulären unteren Dreiecksmatrix, siehe Herleitung in [27].

$w = X'u$
$t = Xw\|Xw\|^{-1}$
$c = Y't$
$u = Yc\|u\|^{-1}$

Zusammengefasst resultiert somit Algorithmus 3, welcher sowohl für PLS1 als auch für PLS2 verwendet werden kann, wobei I_n die n-dimensionale Einheitsmatrix und $1_{n\times n}$ die quadratische Matrix der Dimension n mit 1-Einträgen darstellt.

Algorithmus 3 PLS

procedure MY.PLS
  Input: $X_{n\times p}$, Response $Y_{n\times q}$, #Komp. M
  initialisiere $u_{n\times 1}$ zufällig
  $X^{(z)} \leftarrow (I_n - 1_{n\times n})\, X$ ▷ Zentrierung des Kernel
  $X^{(z)}_{\text{orig}} \leftarrow X^{(z)}$
  $\overline{Y}_{n\times q} \leftarrow n^{-1}1_{n\times n}Y$ ▷ Spaltenmittelwerte$_{n\times q}$
  $Y \leftarrow Y - \overline{Y}$ ▷ Zentrierung von Y
  for $i \in \{1, \ldots, M\}$ **do**
    repeat
      $t \leftarrow X^{(z)\prime}u$
      $t \leftarrow t\|t\|^{-1}$
      $u \leftarrow YY't$
      $u \leftarrow u\|u\|^{-1}$
    until Konvergenz von t
    $X^{(z)} \leftarrow (I_n - tt')\, X^{(z)}$
    $Y \leftarrow Y - tt'Y$
  end for
  return Koeffizienten $B = X'U\,(T'XX'U)^{-1}\,T'Y$, Mittelwerte $\overline{Y}$
end procedure

4.2 Nichtlineare Partial Least Square Regression

Eine mögliche Nichtlinearisierung von Algorithmus 3 verwendet sogenannte Kernelmatrizen. Im Wesentlichen wird die Prädiktormatrix X in einen sogenannten Merkmalsraum (engl. feature space) transformiert und in diesem Raum eine PLS-Regression durchgeführt. Die theoretischen Grundlagen, respektive die Motivation, den PLS-Algorithmus um Kernel zu ergänzen, geht auf das Prinzip des reproduzierenden Kernel Hilbertraum (RKHS) zurück, sowie dem hiermit verbundenen Repräsentationssatz, welcher in einer, für die hier benötigte Verwendung, leicht vereinfachten, dennoch weitestgehend komplexen, Form wiedergegeben wird. Die Theorie ist in [29], beziehungsweise in stark verkürzter Form in [27] dargestellt.

Sei $\mathcal{X} \neq \emptyset$. Die hier präsentierte Herleitung der zugrundeliegenden Idee orientiert sich sehr stark an [29], und wird für weiterführende Informationen empfohlen.

Definition 4.2 (Kernel)**.** Sei

$$\begin{aligned} k\colon \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x, y) &\mapsto k(x, y) \end{aligned}$$

Seien $x_i \in \mathcal{X}$ für $i = 1, \ldots, n$, und sei

- $k(x_i, x_j) = k(x_j, x_i)$ symmetrisch und
- die Matrix $K = \left\{k(x_i, x_j)\right\}_{i,j}$ positiv definit.

So heißt k ein **positive definiter Kernel**.

Grundsätzlich kann der Kernel in Definition 4.2 ebenso für komplexe Zahlen definiert werden. Für die hier vorgestellten Anwendungen reicht der Wertebereich der reellen Zahlen jedoch aus.

Weiters sei $\mathcal{F}$ jener Raum von Funktionen, welcher die nichtleere Menge $\mathcal{X}$ nach $\mathbb{R}$ abbildet und definieren zusätzlich ein Mapping

$$\begin{aligned} \phi\colon \mathcal{X} &\to \mathcal{F} \\ x &\mapsto k(\cdot, x), \end{aligned}$$

welches $\mathcal{X}$ in ebendiesen Raum abbildet.

Seien nun wiederum $x_i \in \mathcal{X}$ mit Skalaren $\alpha_i \in \mathbb{R}$ für $i = 1, \ldots, n_1$, sowie $y_i \in \mathcal{X}$ mit Skalaren $\beta \in \mathbb{R}$ für $i = 1, \ldots, n_2$ mit $n_{1,2} \in \mathbb{N}$. Somit sind die beiden Funktionen

$$\begin{aligned} f(\cdot) &= \sum_{i=1}^{n_1} \alpha_i k(\cdot, x_i) \\ g(\cdot) &= \sum_{i=1}^{n_2} \beta_i k(\cdot, y_i) \end{aligned} \tag{4.14}$$

jeweils aus dem Bildraum von ϕ und es kann durch

$$\langle f, g \rangle := \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \alpha_i \beta_j k(x_i, y_j) \tag{4.15}$$

ein wohldefiniertes, inneres Produkt definiert[3] werden und es folgt die **reproduzierende Kerneleigenschaft**, da

$$\begin{aligned} f(x) &= \langle f, k(\cdot, x) \rangle \\ \text{und } k(x, y) &= \langle \phi(x), \phi(y) \rangle \end{aligned}$$

[3] Beweis findet sich in [29].

gelten. Mit dem inneren Produkt aus Gleichung (4.15) können somit die Funktionen der Form (4.14) vervollständigt werden und der hieraus resultierende Hilbertraum H_k wird in der Literatur als reproduzierender Kernel Hilbertraum (RKHS) bezeichnet.

Mit diesen Voraussetzungen kann nun das Lemma über die nichtparametrische Repräsentation, wie in [29], definiert werden.

Lemma 4.3. *Seien $\mathcal{X}$ und k wie zuvor, $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$, $g\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ streng monoton steigend und $V\colon (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$, sowie*

$$\mathcal{F} := \left\{ f \in \mathbb{R}^{\mathcal{X}} : f(\cdot) = \sum_{i=1}^{\infty} \beta_i k(\cdot, x_i), \beta_i \in \mathbb{R}, x_i \in \mathcal{X} \text{ mit } \|f\| < \infty \right\},$$

wobei $\|\cdot\|$ die durch die Kerneldarstellung induzierte Norm mit der Darstellung

$$\left\| \sum_{i=1}^{\infty} \beta_i k(\cdot, x_i) \right\|^2 = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \beta_i \beta_j k(x_i, x_j)$$

ist.

Dann kann jedes $f \in \mathcal{F}$, welches $V\left(\{x_i, y_i, f(x_i)\}_{i=1,\ldots,n}\right) + g\left(\|f\|\right)$ minimiert, dargestellt werden als

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$$

mit $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, n$.

Ein Beispiel für die Verlustfunktion V bildet die kleinste Quadrateschätzung, während g beispielsweise als Parameter für einen gewissen Trade-Off zwischen der Gestalt von f und der Modellgüte dient, so findet sich oftmals die Definition $g(\|f\|) := \lambda \|f\|$ mit $\lambda > 0$.

Wie in [27] beschrieben, kann gezeigt werden, dass die Darstellung von $k(x, y) = \langle \phi(x), \phi(y) \rangle = \sum_{i=1}^{m} \lambda_i \phi_i(x) \phi_j(x)$ mit $m \in [0, \infty]$ durch die Eigenwerte (λ_i) und dazugehörenden Eigenfunktionen (ϕ_i) des Operators T_k beschrieben werden können.

$$T_k\colon L_2(\mathcal{X}) \to L_2(\mathcal{X})$$
$$(T_k f)(x) = \int_{\mathcal{X}} k(x, y) f(y)\, \mathrm{d}y,$$

$\forall f \in L_2(\mathcal{X})$.[4]

[4] Aufgrund dessen wird k als Kernel bezeichnet, wie [29] schreibt.

Mit der Definition einer Funktion Φ, welche die Spektren x_i in eben diesen Raum abbildet,

$$\begin{aligned} \Phi \colon \mathbb{R}^s &\to \mathcal{F} \\ x_i &\mapsto \Phi(x_i) \end{aligned}$$

und der Abbildungseigenschaft

$$k(x,y) = \sum_{i=1}^{m} \sqrt{\lambda_i}\phi_i(x) \cdot \sqrt{\lambda_i}\phi_i(y) = \Phi(x)'\Phi(y)$$

wird der Kernel Partial Least Square Algorithmus motiviert. Hierbei wird aus Notationsgründen $\{\Phi(x_i)\}_{i=1,\dots,n}$ zu der Matrix Φ zusammengefasst und es resultiert der finale KPLS-Algorithmus 4, wie in [27] hergeleitet und erläutert, indem die Prädiktormatrix X durch das Mapping $\phi(X)$ substituiert wird.

Algorithmus 4 Kernel PLS

procedure MY.KPLS
  Input: Kernel $K_{n\times n}$, Response $Y_{n\times q}$, #Komp. M
  initialisiere $u_{n\times 1}$ zufällig
  $K^{(z)} \leftarrow \left(I_n - n^{-1}1_{n\times n}\right) K \left(I_n - n^{-1}1_{n\times n}\right)$ ▷ Zentrierung des Kernels
  $K^{(z)}_{\text{orig}} \leftarrow K^{(z)}$
  $\overline{Y}_{n\times q} \leftarrow n^{-1}1_{n\times n}Y$ ▷ Spaltenmittelwerte$_{n\times q}$
  $Y \leftarrow Y - \overline{Y}$ ▷ Zentrierung von Y
  for $i \in \{1,\dots,M\}$ **do**
    repeat
      $t \leftarrow K^{(z)}u$
      $t \leftarrow t/\|t\|$
      $u \leftarrow YY't$
      $u \leftarrow u/\|u\|$
    until Konvergenz von t
    $K^{(z)} \leftarrow (I_n - tt')\,K^{(z)}\,(I_n - tt')$
    $Y \leftarrow Y - tt'Y$
  end for
  return Koeffizienten $B^* = U(T'K^{(z)}_{\text{orig}}U)^{-1}T'Y$, Mittelwerte $\overline{Y}$
end procedure

Analog zum linearen PLS Algorithmus 3 ergeben sich die Koeffizienten B nach (4.13), wobei X durch deren Transformationen Φ substituiert werden. Für die gefitteten Responsevariablen kann somit wiederum auf die Kernelmatrix K zurückgegriffen werden: $\widehat{Y} = \Phi B = \Phi\Phi' B^* = KB^*$.

Für eine Schätzung der abhängigen Variablen neuer Spektren muss die zentrierte Kernelmatrix $K^{(z)}_{\text{test}}$ berechnet werden. Für $\widehat{Y} \in \mathbb{R}^u$ mit den dazugehörenden Spektren $x^u_i, i = 1,\dots,u$ ergibt sich eine Kernelmatrix mit den Einträgen

$\{K_{\text{test}}\}_{i,j} = k(x_i^u, x_j)$ mit x_j Spektren zur Bestimmung der Koeffizienten. Da die Koeffizienten mithilfe einer zentrierten Kernelmatrix ermittelt wurden, muss auch hier der Vorhersagekernel gleichwertig angepasst werden.

$$K_{\text{test}}^{(z)} = (K_{\text{test}} - n^{-1}1_s 1_n' K_{\text{orig}})(I - n^{-1}1_n 1_n') \tag{4.16}$$

Mit Gleichung (4.16) resultiert somit

$$\widehat{Y} = K_{\text{test}}^{(z)} \cdot B^* + 1_s \overline{Y}. \tag{4.17}$$

Diese Berechnung ermöglicht es somit, eine PLS-Regression in dem Merkmalsraum $\mathcal{F}$ durchzuführen, was wiederum einem nichtlinearen Algorithmus für die Daten X und Y entspricht. Hierbei ist zu beachten, dass nicht die Eigenfunktionen $\phi_i, i = 1, \ldots, m$ benötigt werden, sondern die Definition eines positiv definiten, symmetrischen Kernel[5] ausreicht.

Die Mappingfunktion ϕ muss nicht explizit gegeben sein. Alleine die Gram-Matrix $K = \phi\phi'$ wird für die kPLS Regression benötigt.

Unterschiedliche Kernel

Eine Liste einiger Vertreter von Kernelklassen finden sich in den Beispielen 4.4, 4.5 und 4.6, wie in [27] definiert.[6] Weitere Möglichkeiten, eine Nichtlinearität mit dem NIPALS-Algorithmus zu kombinieren, ist der Sigmoid-Kernel in Beispiel 4.7 und wird in [18] näher diskutiert. Eine weitere Nichtlinearisierung kann mithilfe des B-Spline Kernels in Beispiel 4.8 durchgeführt werden und wird unter anderem in [9], Seite 99, eingeführt.

Seien für die folgenden Beispiele jeweils $n \in \mathbb{N}_{>0}$ und $x, y \in \mathbb{R}^n$

Beispiel 4.4. Unter dem **lineare Kernel** versteht man das Skalarprodukt $k(x, y) = x'y$.

Beispiel 4.5. Unter dem **Gaußkernel**, auch Radial Basis Function genannt, versteht man die Funktion $k(x, y) = \exp\left\{-\sigma^{-2}\|x - y\|^2\right\}$ mit dem Parameter $\sigma \in \mathbb{R}_{>0}$.

Wird in Beispiel 4.5 die Norm nicht quadriert, so spricht man von dem **Exponentialkernel**, wie ebenfalls in [9] auf Seite 99 vorgestellt wird.

Beispiel 4.6. Seien zusätzlich die Parameter $a \in \mathbb{N}_{>0}$, $b \in \mathbb{R}_{\geq 0}$ gegeben. So kann der **polynomiale Kernel** durch die Kernelfunktion $k(x, y) = (\langle x, y\rangle + b)^a$ definiert werden.

Beispiel 4.7. Für den **Sigmoid Kernel** wie in [18] seien $a \in \mathbb{R}\setminus\{0\}$, sowie $b \in \mathbb{R}$. So wird dieser Kernel durch $k(x, y) = \tanh(ax'y + b)$ festgelegt.

[5] Wie in [27] und [29] beschrieben, kann dieses Vorgehen ebenso auf semi-positiv definite Kernel ausgeweitet werden.

[6] Es wurden primär diese Kernel verwendet.

In [18] auf Seite 7 wird aus Gründen des Verhaltens des Sigmoidkernels $a > 0$ und $r < 0$ empfohlen und die unterschiedlichen Parametrisierungen untersucht.

Ein weiterer, bekannter Vertreter der Kernelklassen ist der Spline-Kernel.

Beispiel 4.8. Sei hierfür $b \in \mathbb{N}$ und $B_{2b+1}(\cdot)$ die B-Spline Funktion vom Grad $2b+1$ wie in Definition A.10. So definiert sich der **(B-)Spline Kernel** durch

$$k(x,y) = \prod_{i=1}^{n} B_{2b+1}(x_i - y_i).$$

Hierbei beschreiben x_i bzw. y_i, $i = 1,\dots,n$ die $i$ten Komponente des Vektoren x bzw. y.

4.2.1 Verwendete Kernel

Linearer Kernel

Für eine erste Einschätzung wird der lineare Kernel $K = XX'$ verwendet. Dies entspricht dem zuvor erwähnten PLS-Algorithmus.

Gaußkernel

Den Wert des Gaußkernels zweier Vektoren x, y erhält man mit $k(x,y) = \exp\{-\sigma^{-2}\|x-y\|^2\}$. Im konkreten Fall entsprechen die beiden Vektoren x und y Zeilenvektoren der Prädiktorenmatrix X. Die Norm $\|x-y\|^2$ gibt somit den euklidschen Abstand zweier Proben an. Zusätzlich wird dieser Abstand mit dem Parameter σ^2 gewichtet.
Betrachtet man Gleichung (4.17), für ein Spektrum x_i, welches sich nicht im Trainingsset befindet, so kann die Vorhersage im PLS1-Modell geschrieben werden als

$$\hat{y}_i = \sum_{j=1}^{n} K_j^z \cdot B_j^* + \bar{y}, \tag{4.18}$$

wobei $\bar{y}$ und B^* unmittelbar aus dem Algorithmus 4 stammen und die Kernelmatrix K^z einer $1 \times n$ Matrix, zentriert und berechnet mit dem Gaußkernel, entspricht.

Dies entspricht einer Linearkombination der Kernelwerte zuzüglich des Offsets: je näher (im Sinne der Distanz gemessen mit dem Gaußkernel) das Spektrum x_i bei einem $x_j, j = 1,\dots,n$ aus dem Trainingsdatensatz liegt, desto mehr Gewichtung erhält der zu x_j gehörende Koeffizient B_j. Für die konkreten Auswertungen in dieser Arbeit wird ein großes Spektrum an σ^2-Parametern untersucht. Hierbei stellt sich heraus, dass der exakte Parameter, sofern das Spektrum abgeleitet wird, nicht von großer Relevanz zu sein scheint. Deswegen wurde nach einer anfänglich intensiven Analyse des Parameters das Raster auf „klein ($\sigma^2 = 0.1$) - mittel ($\sigma^2 = 1$) - groß ($\sigma^2 = 3$)“ eingeschränkt.

Polynomialer Kernel

Der polynomiale Kernel der Form $k(x,y) = (\langle x,y\rangle + c)^m$ setzt die Vektoren mithilfe des inneren Produktes in Relation. Dieses entspricht bei gegebenem y einer Linearkombination der Spektren. Der Parameter m gibt den Grad des Polynoms an, welcher für die Kernelbildung verwendet wird und c kann als Art Offset interpretiert werden. Für das Tupel $(c,m) = (0,1)$ entspricht dies dem linearen Kernel. Analog zum Gaußkernel wird ebenfalls eine Vielzahl an unterschiedlichen Parameterwerten für den polynomialen Kernel untersucht. Da sich allerdings für die untersuchten Daten und Referenzwerte stets bessere Ergebnisse erzielen lassen, liegt der Fokus dieser Arbeit auf einer detaillierteren Analyse des linearen und Gaußkernel.

5 Auswertungen mit der (Kernel) Partial Least Square Methode

Im diesem Kapitel werden für alle Responsevariablen, mit Ausnahme der Unterteilung der Milchsäure in L- und R-Milchsäure, ein Partial Least Square-Modell entwickelt und analysiert. Hierbei liegt der Fokus einerseits auf dem Verhalten der Residuen, sowie der Güte der Reproduzierbarkeit. Es werden jeweils Wellenzahlen selektiert, sowie die Modellparameter Kernel und Anzahl an latenten Variablen für den Datensatz E25 bestimmt und auf die vergleichbaren Datensätze E22, E24, sowie V70(2016) angewendet, um die Güte zu vergleichen. Sofern vorteilhafte Erkenntnisse aus Submodellen (Modell für Rotweine oder einen eingeschränkten Wertebereich) erzielt werden können, wird ebenfalls auf diese verwiesen.

Zu Beginn wird anhand von Ethanol ein erstes Modell vorgestellt, um die Methodik zu verstehen, sowie erste Erkenntnisse über die Funktionalität der PLS-Modelle zu erhalten. Zudem wird ein Modell vorgestellt, welches sich Datensätze eines spezifischen Wertebereiches als Grundgesamtheit selektiert, wie beispielsweise durch Ausschluss von alkoholarmen Weinen, sowie deren Pendant, zum Beispiel den Portweinen, mit einem hohen Alkoholgehalt. Die Modelle werden stets mit dem Datensatz E25 entwickelt und für die Qualität des Modells, sowie der Güte der unterschiedlichen Datensätze, auf diese angewendet. Hierbei werden, um eine Vergleichbarkeit zu erhalten, lediglich die Datensätze aus dem Jahr 2016, d.h. E22, E24 und V70(2016) verwendet. In weiterer Folge wird zusätzlich die Reproduzierbarkeit analysiert.

In einem nächsten Schritt werden Modelle für Extrakt, welcher unter anderem die Zucker Glukose und Fruktose beeinhaltet, entwickelt, und in weiterer Folge wird auf diese beiden Zuckerarten eingegangen. Das Unterkapitel über Glukose kann zusätzlich in einen Unterabschnitt für Rotweine, einen reduzierten Wertebereich, sowie deren Kombination gegliedert werden, sowie der gesonderten Datensatzanalyse. Ebenso wird ein Vergleich mit dem Datensatz V70(2015) angestellt.

Der hierauf folgende thematische Block befasst sich mit der Modellierung unterschiedlicher Säuren. An erster Stelle wird der Summenparameter titrierbare Säure untersucht, gefolgt von den vorliegenden Daten zu den einzelnen Bestandteilen, absteigend geordnet nach der Größe des Wertebereiches, welche die unterschiedlichen Säuren zu der Gesamtsäurekonzentration beitragen. Hierzu zählen neben der Weinsäure und der L-Äpfelsäure, die Milchsäure, gefolgt von den flüchtigen Säuren und der Zitronensäure.

Glyzerin wird als abschließender Inhaltsstoff der Weinproben modelliert, während zur Abrundung des Analysekapitels versucht wird, zwei Eigenschaften der Weine

zu modellieren. Einerseits die Dichte, sowie zusätzlich den pH-Wert als Maßstab für die saure bzw. basische Charakteristik der Weine.

Die Analysen bezüglich unterschiedlicher Temperaturen wie Raumtemperatur oder Probentemperatur entfallen in den folgenden Kapiteln, da in keinem Modell ein Zusammenhang zwischen den unterschiedlichen Temperaturmessungen und den Schätzwerten beziehungsweise den Residuen beobachtet werden konnte. Selbiges gilt ebenfalls für die Kennzahlen des Luftdrucks und der Luftfeuchtigkeit.

Grundsätzlich gilt, dass kein Modell für Weine entwickelt werden konnte, welches zeitgleich auch für Wasserspektren angewendet werden kann. Beispielsweise werden bei dem vorgeschlagenen Modell für Extrakt bei Anwendung auf Wasser alle Extraktkonzentrationen in ein Intervall von $[6, 9]\,\mathrm{g/l}$ geschätzt. Das Modell für titrierbare Säuren unterstellt Wasser stets eine Konzentration von $[1.90, 2.17]\,\mathrm{g/l}$. Bei dem einzig passenden Modell, welches ebenfalls für Wasser akzeptable Werte liefert, handelt es sich um das Ethanolmodell. Hierbei werden 149 Werte unterschätzt (theoretisch ein negativer Ethanolgehalt), welche auf 0 Vol.% gekappt werden müssen. Für die restlichen 338 Wasserspektren liefert das Modell Ethanolwerte von maximal 0.11 Vol.%. Man muss hier allerdings auch beachten, dass die beiden alkoholarmen Weine im Datensatz eine Ethanolkonzentration von weniger als 0.1 Vol.% aufweisen und das Wasserspektrum teilweise auch diese Weine überschätzt. Bezüglich der Vorhersagequalität für Wasser kann somit kein Modell als passend betrachtet werden. Um dennoch Wasserproben zuzulassen, muss somit ein Workaround geschaffen werden, wie die Interpretation als Wasser in Abhängigkeit vom höchsten Peak eines Spektrums.

5.1 Ethanol

In einer ersten Analyse werden die Partial Least Square-Modelle mit unterschiedlichen Kernels auf die Spektren in Zusammenhang mit dem Alkoholgehalt der Weinproben angewendet. Die Nichtlinearisierung der PLS-Methode liefert keine signifikanten Verbesserungen für die hier betrachteten Responsevariable Ethanol, dennoch wird an dieser Stelle der mittels Gaußkernel ermittelte Wellenzahlbereich vorgestellt und in weiterer Folge mit dem linearen Modell verglichen. Der Hauptgrund für diese Wahl liegt im Verlauf des SEP-Plots in Abbildung 5.1 (li.). Obwohl der lineare Kernel ebenfalls die Verwendung von 2 (oder 3) latenten Variablen nahelegt, scheint bei einer Erhöhung der Variablenanzahl ebendiese Grafik zu divergieren. Darüber hinaus kann das derartig einheitliche Abfallen durch die Erhöhung der latenten Variablen von 1 auf 2 nicht in gleichem Maße bei linearen Kernels in der zugehörenden Wellenzahlwahl beobachtet werden.

Betrachtet man zusätzlich den aus der Kreuzvalidierung resultierenden Residuenplot mit je 100 Schätzungen für jeden der Weine als Testset in Abbildung 5.1 (re.), so kann die größte Variabilität in den alkoholfreien Weinen festgestellt werden. Dies kann einerseits damit begründet werden, dass 79 von 81 Weinen einen Alkohlgehalt von mindestens 7.9 Vol.% aufweisen und dies somit den Großteil des Trainingsdatensatzes repräsentiert. Zudem muss berücksichtigt werden, da

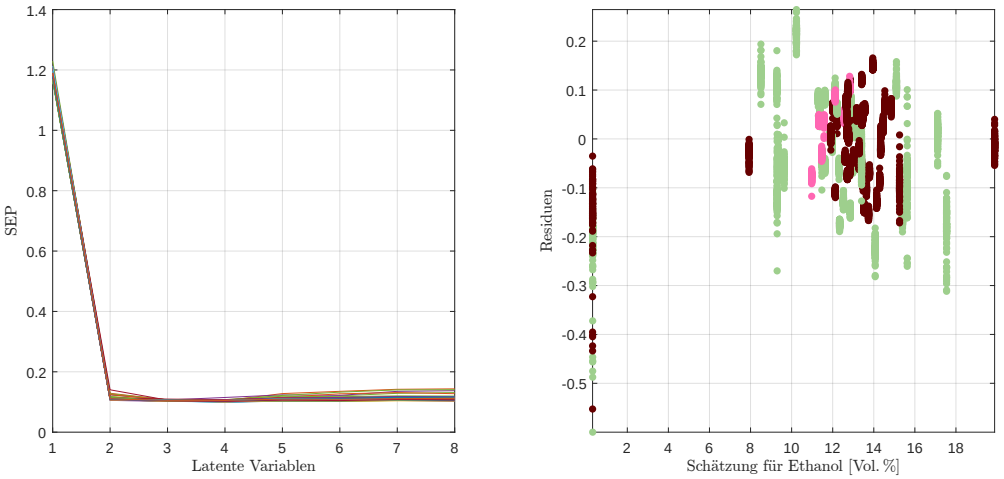

Abbildung 5.1: Verlauf des SEP in Abhängigkeit von latenten Variablen (li.) und die dazu korrespondierenden Weinresiduen der doppelten Kreuzvalidierung [Vol.%] (re.) für Ethanol im PLS-Modell.

für jedes Residuum eines alkoholfreien Weines in dieser Grafik maximal ein alkoholfreier Wein im Trainingsdatensatz berücksichtigt wurde, dass diese somit als eine Art Extrapolation aufgefasst werden können. Andererseits selektiert das Modell Wellenzahlen, (vgl. Abbildung 5.2 (li.)), welche nicht ausschließlich typische Ethanolbanden (wie die CO- oder die CH-Bande) repräsentieren.

Für diese Auswertungen werden folgende Modellparameter gewählt:

- 4×5 Wellenzahlen, beim höchsten Peak im Fingerprintbereich, beim Anstieg nach der H_2O-Bande, jenem Peak um $2900\,cm^{-1}$ sowie ein am Rande des zugelassenen Wellenzahlbereiches.
- 2 latente Variablen, wie durch Abbildung 5.1 (li.) motiviert.
- Ein Gaußkernel mit Parameter $\sigma^2 = 3$.
- Savitzky-Golay Ableitung zweiter Ordnung.

Weitere nachträglich erfolgte Auswertungen zeigen, dass die Wahl des konkreten Parameters des Gaußkernels unter der Verwendung von Ableitungen beim Preprocessing keinen nennenswerten Effekt aufweisen (nicht jedoch, wenn kein Preprocessing angewendet wird), ebenso wie die Anzahl der verwendeten Punkte zur Glättung bei der Savitzky-Golay Ableitung.

Um die tatsächliche Qualität des erstellten Modells beurteilen zu können, muss notwendigerweise der klassische Residuenplot gegen die Schätzungen in Abbildung 5.2 (re.) analysiert werden. Wie bereits aus den Grafiken der doppelten Kreuzvalidierung zeigen sich keine Ausreißer. Selbst die Überschätzung der beiden alkoholarmen Weine können nicht als auffällig bezeichnet werden (hierbei befinden sich beide Weinproben im Trainingsdatensatz).

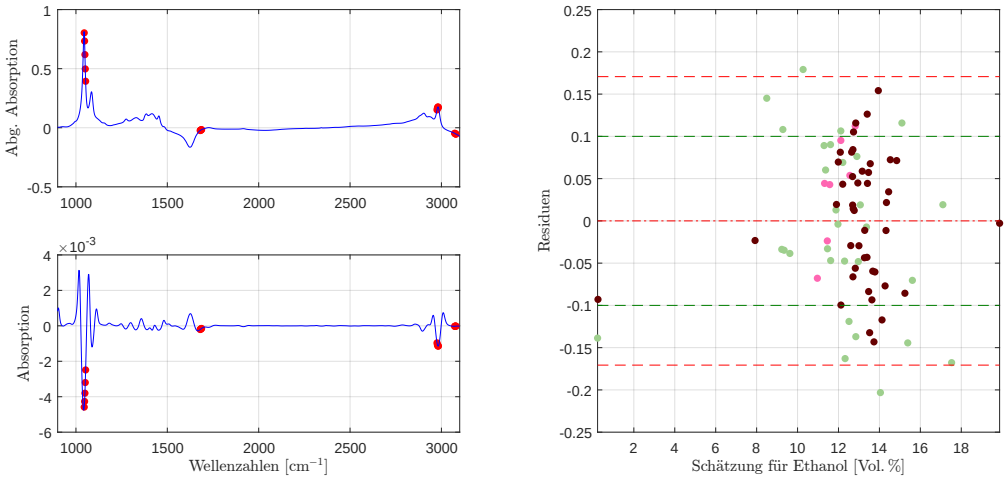

Abbildung 5.2: Die selektierten Wellenzahlen im PLS-Modell für Ethanol mit der zweiten Savitzky-Golay Ableitung (li.) und der dazugehörende Residuenplot mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (re.).

Durch die Konstruktion dieses statistischen Modells tritt kein Bias auf. Die Standardabweichung der Residuen beläuft sich auf lediglich 0.0853 Vol.% und liegt somit geringfügig über jener der Vergleichsmethode.

Teilmodelle für Farbe und eingeschränkten Wertebereich

Klassifiziert man die Daten nach der Farbe und versucht, für die Rotweine ein eigenes Modell zu entwickeln, oder beschränkt man den Ethanolbereich auf $[8, 16]$ Vol.%, beziehungsweise eine Kombination dieser Spezifikationen, so können teilweise Verbesserungen von bis zu 30 %, im Sinne einer Reduzierung der Variabilität der Residuen, erzielt werden, wie Tabelle 5.1, mit den wichtigsten Kennzahlen der besten untersuchten PLS-Modelle für Ethanol, zeigt. In der ersten Zeile wird auf die Performance des Gesamtmodells mit allen 81 Datensätzen hingewiesen. Die hieraus resultierenden Residuen variieren mit einer Standardabweichung in Höhe von 0.0853 Vol.%, welche mit 0.0841 Vol.% minimal über jener der Residuen dieses Modells, eingeschränkt auf Weine mit einem tatsächlichen Ethanolgehalt im Bereich $[8, 16]$ Vol.%, liegt.[1] In einem vergleichbaren linearen Modell zeigt sich mit einer Variabilität in Höhe von 0.0992 Vol.% für sämtliche Weinproben eine unwesentlich höhere Kennzahl, sowie ein ebenso unauffälliges Verhalten im Residuenplot.

Darüber hinaus sind drei weitere Modelle mit einigen Vergleichswerten angeführt. Einerseits ein Modell, welches speziell für den Bereich $[8, 16]$ Vol.% entwickelt wurde, welches ebenfalls einen Gaußkernel[2] als optimal interpretiert. Durch die-

[1] Diese Differenzierung gewinnt insbesondere in Hinblick auf die Modellierung des Alkoholgehaltes mit den neuronalen Netzwerken an Bedeutung, wie Abschnitt 7.1 zeigt.

[2] Modelle mit einem linearen Kernel liefern dieselben Residuen und somit auch dieselbe Güte.

ses Spezialmodell kann, verglichen mit den vorangegangenen Modellen eine geringfügige Verbesserung erzielt werden, da sich die Variabilität der Residuen von 0.0841 Vol.% auf 0.0724 Vol.% reduziert. Hierbei gilt, dass jeweils modellspezifische Wellenzahlen selektiert wurden, sowie unter Umständen eine andere Anzahl (eine oder zwei) von latenten Variablen.[3]

Weitere Submodelle beziehen sich auf die Weinfarbe. Da die Rotweine mit 44 Weinen die Mehrheit der Population beschreiben, wurden insbesondere Modelle für Rotweine untersucht. Um einen Vergleich mit dem Ausgangsmodell anstellen zu können, wird an dieser Stelle dieses Modell mit denselben Parametern, allerdings ausschließlich mit den Rotweinen als Datensätze, kalibriert und es resultiert eine Standardabweichung der Rotweine in Höhe von 0.0794 Vol.%, respektive 0.0761 Vol.% für den zusätzlich eingeschränkten Wertebereich $[8, 16]$ Vol.%, während im Vergleich hierzu die Standardabweichung der Rotweine im Gesamtmodell 0.0751 Vol.% beträgt. Eine weitere, deutliche Verbesserung kann hierbei durch die Neukalibrierung des für den Wertebereich $[8, 16]$ Vol.% entwickelten Modells mit Rotweinen erzielt werden. So reduziert sich die Standardabweichung auf 0.0560 Vol.%. Auch hier können mit einem linearen Kernel dieselben Kennzahlen erzielt werden. Es bleibt zu erwähnen, dass keines der hier angeführten Modelle strukturelle Fehler oder Abhängigkeiten in den Residuen erkennen lässt, welche auf die Modellierung bzw. die Methodik zurückzuführen ist. Detailliertere Informationen können der Kennzahlenübersicht in Tabelle 5.1 entnommen werden.

	Anz.	Mw.	Std.	Min.	Med.	Max.	IQR	MSE
Modell, Gauß	81	0.00	0.0853	−0.20	0.01	0.18	0.13	0.0072
$[8, 16]$ Vol.%	75	0.01	0.0841	−0.20	0.01	0.18	0.12	0.0070
Modell, linear	81	0.00	0.0992	−0.24	0.01	0.22	0.15	0.0097
$[8, 16]$ Vol.%	75	0.01	0.0981	−0.24	0.01	0.22	0.15	0.0096
Modell, Gauß, $[8, 16]$ Vol.%	75	0.00	0.0724	−0.24	−0.01	0.24	0.09	0.0052
Modell, Gauß, Kal: Rot	44	0.00	0.0794	−0.17	0.01	0.13	0.10	0.0062
$[8, 16]$ Vol.%	41	0.01	0.0761	−0.17	0.01	0.13	0.10	0.0057
Modell, Gauß, Kal: Rot, $[8, 16]$ Vol.%	41	0.00	0.0560	−0.10	−0.01	0.12	0.07	0.0031

Tabelle 5.1: Modellvergleich der PLS-Modelle für Ethanol mit unterschiedlichem Kernel, Weinfarbe und dem auf $[8, 16]$ Vol.% eingeschränkten Wertebereich. Alle Werte in Vol.%.

Abbildung 5.3 zeigt einen beispielhaften SEP-Verlauf eines Modells mit ähnlicher Güte für Rotweine bei einem Wertebereich von $[8, 16]$ Vol.% Ethanol. Hierbei werden lediglich 9 Prädiktoren mit einer latenten Variable verwendet (gesamtes Modell: 2 lat. Variablen und 20 unterschiedliche Wellenzahlen), ebenfalls mit einem Gaußkernel. Auffallend ist, dass durch die Hinzunahme von latenten Variablen kein (bzw. nur ein minimaler) Effekt bezüglich des SEP beobachtbar ist, wohingegen ab 5 latenten Variablen die SEP zu divergieren scheinen. Dies bedeutet, dass in diesem Modell die Hinzunahme von 4 latenten Variablen sich nicht auf die Güte auswirkt und erst ab 6 Variablen gewisse Eigenheiten von Trainingssets kompensiert wer-

[3] Vergleiche hierzu die vollständige Übersicht in Tabelle 5.36

den (können), was sich wiederum auf die Vorhersage bzw. die Generalisierung des Modells negativ auswirkt, beziehungsweise der NIPALS-Algorithmus keine Konvergenz in diesen Variablen zeigt.

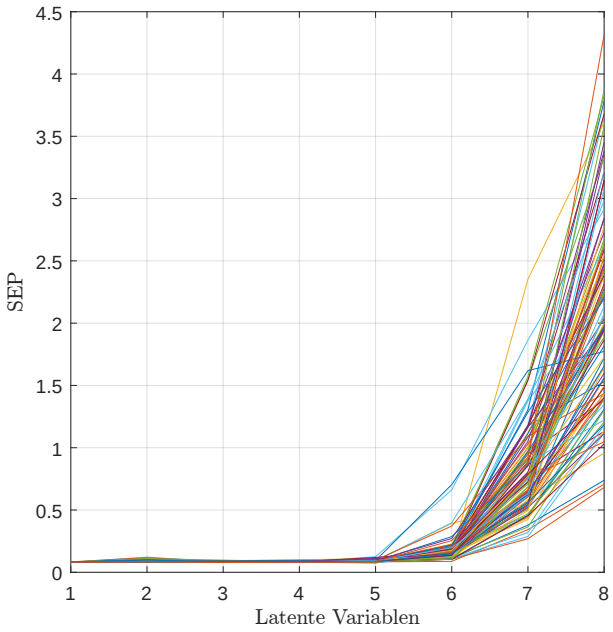

Abbildung 5.3: Verlauf des SEP in Abhängigkeit von latenten Variablen für Ethanol im PLS-Modell für Rotweine.

Reproduzierbarkeit

Zusätzlich muss für die Güte des Modells die Reproduzierbarkeit überprüft werden. Hierbei stehen pro Wein jeweils vier Messungen zur Verfügung und die Residuen dieser Einzelmessungen sind in analoger Weise zum klassischen Residuenplot in Abhängigkeit der Wein ID in Abbildung 5.4 dargestellt. Bei den künstlich erzeugten Weinen, hier in Grau, weichen jeweils wenige einzelne Messungen stärker ab. Darüber hinaus konnte eine entartete Weißweinmessung beobachtet werden. Da diese einzelne Messung auch einen ähnlichen, nicht plausiblen Schätzwert für andere Responsevariablen wie zum Beispiel Glukose oder Fruktose liefert, deutet dies sehr stark auf einen Messfehler (von 328 Messungen echter Weine) hin und wird in Abbildung 5.4 bzw. Tabelle 5.2, sowie den nachfolgenden Auswertungen, nicht weiter berücksichtigt. Der Trend in Abhängigkeit der Wein ID, wie in dieser Grafik beobachtbar, wird im Unterabschnitt Beobachtbare Auffälligkeiten gesondert behandelt.

Insgesamt zeigt sich für die Rot- und Weißweine ein ähnliches Bild, jeweils mit einem Mittelwert von knapp über 0.01 Vol.%. Die davon abweichende Gestalt des Boxplots der Roséweine kann unter anderem damit begründet werden, dass hier lediglich 7 Weine im Datensatz vorliegen. Zusätzlich muss auch auf die Genauigkeit der Daten hingewiesen werden, welche auf zwei Nachkommastellen genau

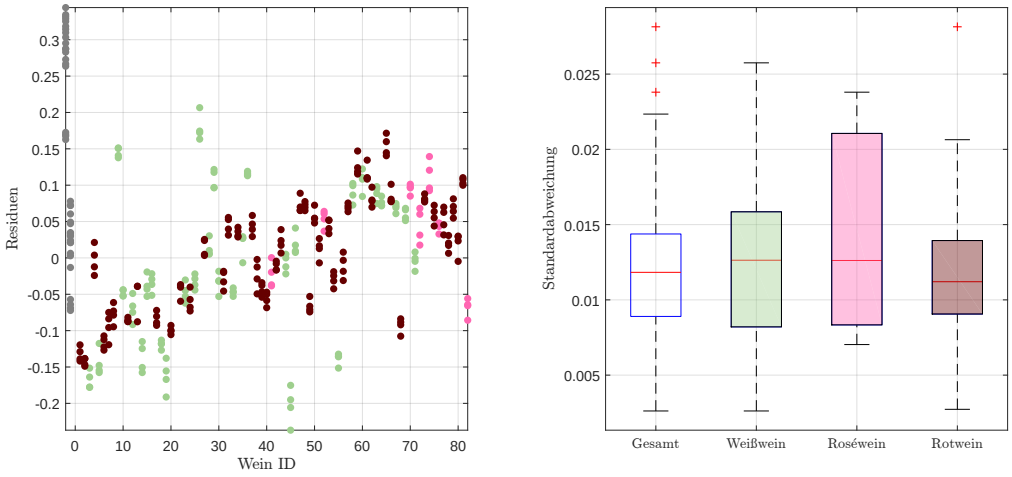

Abbildung 5.4: Residuen in Vol.% der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung des Ethanolgehaltes im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

vorliegen und somit eine einigermaßen plausible Reproduzierbarkeit ausweisen, da durch die Mittelung der vier Spektren kleinere Ungereimtheiten ausgeglichen werden, was bei den hier betrachteten Wiederholungsmessungen im Einzelfall nicht möglich ist.

Die Kennzahlen der maximalen Abweichung in Tabelle 5.2 zeigen ein ähnliches Verhalten. Während die höchste Auslenkung mit 0.06 Vol.% bemessen wird, beträgt diese Auslenkung im Mittel nur 0.03 Vol.% bei einer hierfür ermittelten Standardabweichung von 0.0119 Vol.% bei Betrachtung des Gesamtmodells.

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0125	0.0126	0.0616	0.0281	0.0268	0.0133
Roséwein	0.0147	0.0126	0.0506	0.0321	0.0298	0.0140
Rotwein	0.0114	0.0112	0.0489	0.0253	0.0241	0.0104
Gesamt	0.0121	0.0118	0.0616	0.0269	0.0261	0.0119

Tabelle 5.2: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Ethanol im PLS-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in Vol.%.

Für die Submodelle können ähnliche Reproduzierbarkeitszahlen beobachtet werden, weshalb diese in diesem Abschnitt nicht gesondert behandelt werden. Nähere Informationen finden sich hierzu in Tabelle 5.36.

Verwendet man dieses Setting mit den Kerneleinstellungen, den Wellenzahlen, dem Preprocessing sowie der Anzahl der latenten Variablen, und verwendet jeden

einzelnen Datensatz als Trainingsdatensatz, so ergeben sich die Kennzahlen in Tabelle 5.3. Das für E25 gefundene Modell liefert für den Datensatz E22 bei Betrachtung der Standardabweichung, sowie des interquartilen Bereichs eine geringfügig bessere Struktur der Residuen, während eine umso stärkere Verschlechterung bei der Engine E24 zu beobachten ist.
Verwendet man den Datensatz V70(2016) um eine Vergleichbarkeit mit den anderen Datensätzen zu erhalten (enthält dieselben Weine wie die Engines der zweiten Generation), so weist dieses Modell eine höhere Standardabweichung auf.

Ebenso deutliche Unterschiede fallen bei der Betrachtung des interquartilen Bereiches auf. Während sich dieser im Datensatz E22 mit dem für die Engine E25 entwickelten, optimalen Modell, auf 0.10 Vol.% beläuft, so nimmt dieser im Datensatz V70(2016) einen um die Hälfte größeren Wert an. Ebenso liefern die Datensätze E22 und E25 die beste Reproduzierarkeit bei einer (auf die Genauigkeit der Daten gerundeten) maximalen Abweichung von 0.06 Vol.% bzw. 0.07 Vol.%, wobei deren Mittelwert bei lediglich 0.01 Vol.% liegt.

Bei allen vier Datensätzen kann allerdings ein leichter Trend in den Residuen, abhängig von der Messreihenfolge wie in Abbildung 5.5 festgestellt werden, wobei bei V70(2016) ein vergleichsweise starker Trend beobachtbar ist.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.0791	0.10	−0.20	0.19	0.0127	0.0112	0.0704
E24	0.1052	0.14	−0.26	0.15	0.0136	0.0119	0.1999
E25	0.0853	0.13	−0.20	0.18	0.0121	0.0118	0.0616
V70(2016)	0.1128	0.15	−0.27	0.18	0.0154	0.0115	0.1178

Tabelle 5.3: Performance des entwickelten PLS-Modells für Ethanol, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in Vol.%.

Betrachtet man das auf den Wertebereich [8, 16] Vol.% und Rotweine eingeschränkte Modell wie in Tabelle 5.1, so zeigt sich ein ähnliches Verhalten in den Datensätzen mit geringerer Schwankung der Residuen, allerdings erhöhter Reproduzierbarkeit. Dies kann unter anderem durch das Fehlen von Information begründet werden, da nur die Hälfte der Daten und somit weniger Informationen über einen funktionalen Zusammenhang für die Kalibrierung zur Verfügung stehen. Auffallend ist, dass in diesem Datensatzvergleich V70(2016) die beste mittlere Reproduzierbarkeit aufweist, bei derselben (wiederum auf die Genauigkeit der vorliegenden Daten gerundeten) maximalen Auslenkung von 0.09 Vol.%. Während in diesem Submodell die maximale Abweichung der Einzelspektren sich für E24 und V70(2016) verringert, erhöht sich diese für die Datensätze E25 und E22 auf einen Wert, welcher für letzteres Spektrometer sogar mehr als doppelt so hoch ist. Dies zeigt den schmalen Grad der Modellfindung und der Definition, für welche Weine dieses Modell anzuwenden ist.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.0523	0.06	−0.08	0.13	0.0277	0.0271	0.1554
E24	0.0770	0.11	−0.17	0.14	0.0250	0.0247	0.1228
E25	0.0560	0.07	−0.10	0.12	0.0255	0.0257	0.0869
V70(2016)	0.0673	0.11	−0.15	0.10	0.0190	0.0176	0.0905

Tabelle 5.4: Performance des entwickelten PLS-Modells für Ethanol, entwickelt für den Wertebereich $[8, 16]$ Vol.% und kalibriert mit Rotweinen, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in Vol.%.

Aufgrund all dieser Erkenntnisse kann, von etwaigen Messfehlern und der Abhängigkeit der Residuen von der Wein ID, ein akzeptables Modell für Ethanol mithilfe der PLS-Methodik, mit einem minimal besseren Verhalten durch eine Nichtlinearisierung, ermittelt werden.

Beobachtbare Auffälligkeiten

Betrachtet man Abhängigkeiten der Residuen von anderen Inhaltsstoffen oder der gegebenen Temperaturen während der Messung, so können keine Auffälligkeiten beobachtet werden. Allerdings kann eine ungewünschte Korrelation in Form eines steigenden Trends der Residuen festgestellt werden, wie Abbildung 5.4 (li.) der Wiederholbarkeitsmessungen zeigt und ist somit gleichbedeutend mit einer Abhängigkeit der Messreihenfolge. Dies wird in den Grafiken in Abbildung 5.5 verdeutlicht.

Diese Abhängigkeit bezüglich der Wein ID (und somit der Messreihenfolge) kann in dieser Form nur auf Messstörungen zurückgeführt werden.[4] Die anfängliche Vermutung, dass ein spezifisches Wellenzahlenintervall wie beispielsweise $[2\,500, 3\,100]\,\mathrm{cm}^{-1}$ existiert[5], welches einen derartigen Trend verursacht, kann verworfen werden. Es finden sich beispielsweise Modelle in dieser Arbeit, welche ebenfalls Wellenzahlen aus diesem Intervall selektieren und dennoch keine beziehungsweise kaum Auffälligkeiten in der Grafik der Residuen gegen die ID zeigen und vice versa. Da es sich hier um einen unmodellierbaren Effekt handelt, wird im weiteren Verlauf der Auswertungen nicht weiter darauf eingegangen.

[4] Als Ursache kann eine zunehmende Verschmutzung der Kristalloberfläche, welche in Kontakt zur Probenfläche steht und zu einer Veränderung des Brechungsindex führt, identifiziert werden.

[5] Diese Hypothese entstammt aus den Beobachtungen erster Modellierungen für Ethanol, auch wenn die Störung in diesem Bereich verstärkt auftritt.

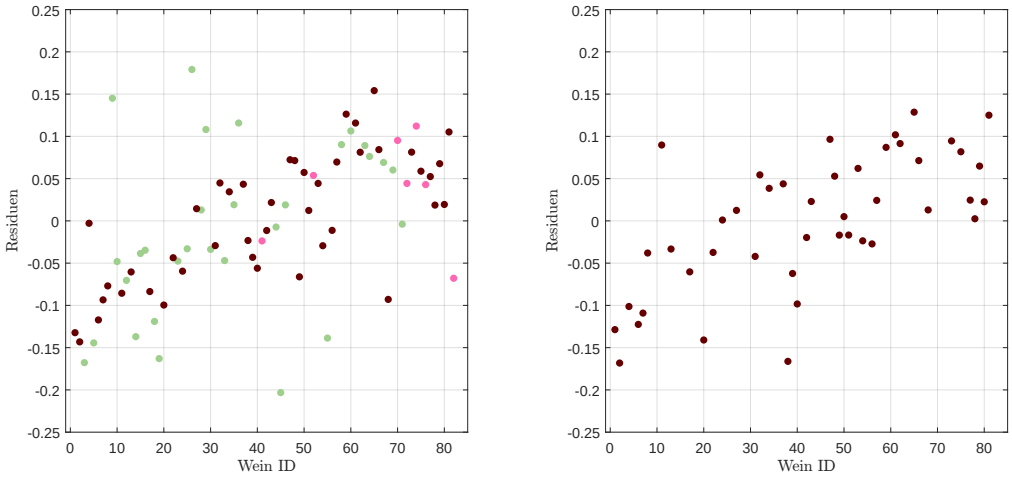

Abbildung 5.5: Trend in den Ethanolresiduen, abhängig von der Wein ID, d.h. der Messreihenfolge, im Gesamtmodell (li.), bzw. dem Modell mit einer Neukalibrierung für Rotweinen (re.) mittels PLS-Modellen.

5.2 Extrakt

Im Vergleich zu Ethanol aus dem vorangegangenen Abschnitt handelt es sich bei Extrakt nicht um eine explizite chemische Verbindung, sondern um einen Summenparameter, welcher unter anderem Glukose, Fruktose, Glyzerin und flüchtige Säuren gemeinsam betrachtet. Um die Extraktwerte aus den vorliegenden Spektren bestimmen zu können, resultieren die verwendeten Heuristiken in einem Modell mit Gaußkernel und Parameter $\sigma^2 = 3$, wobei wiederum der exakte Parameterwert des Kernels lediglich einen nachrangigen Einfluss hat, mit insgesamt 4 latenten Variablen. Die Wellenzahlselektion zur heuristischen Maximierung der Vorhersagegenauigkeit ergibt eine Auswahl von insgesamt 40 unterschiedlichen Wellenzahlen, wie in Abbildung 5.6 visualisiert. Insgesamt verteilen sich diese Wellenzahlen auf 4 Blöcke, jeweils mit einer Kardinalität von 10 Messpunkten. Einer dieser verwendeten Prädiktorenblöcke liegt knapp unter $1\,500\,\mathrm{cm}^{-1}$ und somit direkt vor der H_2O-Bande. Während jene im Fingerprintbereich in der ersten Savitzky-Golay Ableitung jeweils einen kleineren Peak beschreiben, beinhalten die restlichen 20 Wellenzahlen lediglich geringe lokale Anhebungen und werden durch die Heuristik als drittes beziehungsweise viertes Intervall für Wertkorrekturen selektiert. Diese Hinzunahme führt zu einer Verbesserung der Variabilität der Residuen um beinahe 30 %, von $1.5715\,\mathrm{g/l}$ auf $1.1455\,\mathrm{g/l}$.

Die Wahl der Anzahl an latenten Variablen begründet sich durch den Verlauf der SEP Kurve in Abbildung 5.7 (li.) und legt diese eindeutig fest. Auch wenn der steilste Abfall von 1 auf 2 zu beobachten ist, so können durch die Hinzunahme von zwei weiteren latenten Variablen signifikante Genauigkeitsverbesserungen erzielt werden. Um dieses Modell zu erhalten, wurden diese zuerst mit der ersten Savitzky-Golay Ableitung vorbehandelt. Für die aus der Kreuzvalidierung resultierenden Residuen kann mit steigendem Extraktwert, wie Abbildung 5.7 (re.) visualisiert,

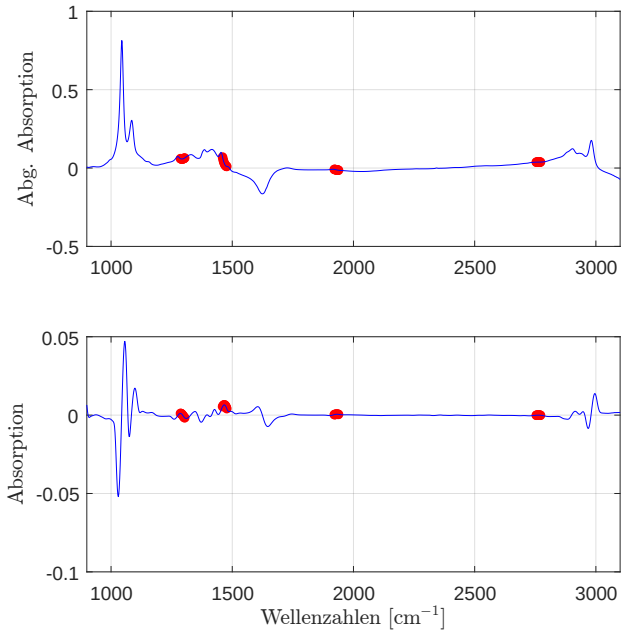

Abbildung 5.6: Die selektierten Wellenzahlen im PLS-Modell für Extrakt mit der ersten Savitzky-Golay Ableitung.

eine teilweise höhere Variabilität beobachtet werden, und begründet sich durch eine Art Extrapolation der Dateninformation, da sich im Bereich unter 100 g/l der größte Teil an Informationen befindet und für das Schätzen der Residuen in dieser Abbildung teilweise Weine mit hoher Extraktkonzentration im Trainingsset nicht berücksichtigt werden.

Der klassische Residuenplot in Abbildung 5.8 zeigt für geringe Extraktwerte eine etwas höhere Variabilität, wobei man die in diesem Bereich sehr hohe Anzahl von gegebenen Stichproben mitberücksichtigen muss. Starke Ausreißer können zudem nicht beobachtet werden. Betrachtet man die Punkte abhängig von der Farbe des Weines, so zeigt sich ein unauffälliges Bild der Rotweinresiduen, während die Weiß- und Roséweine womöglich einen leichten Trend aufweisen. Dies liegt unter anderem an dem großen/größeren Wertebereich der Weißweinresiduen und es existiert ein farbspezifischer Bias. Während die Rotweine einen Offset in Höhe von -0.09 g/l bei einer Standardabweichung von 1.1239 g/l aufweisen, so belaufen sich diese Kennzahlen der übrigen Residuen auf 0.10 g/l beziehungsweise 1.1772 g/l. Man beachte, dass aufgrund der Skalierung die Abweichung der Labormethode für hohe Extraktwerte nicht dargestellt werden (kann) und somit die Fortsetzungen der grünen Linien ab 110 g/l fehlen.

Aufgrund der, auch wenn nur leichten, Über- bzw. Unterschätzung der Rot- bzw. Weiß-/Roséweine, sind in Abbildung 5.9 die Residuen mit der entsprechenden Neukalibrierung dargestellt und es zeigen sich hier keine Auffälligkeiten, selbst für große Extraktwerte nicht. Der Bias kann durch diese Modellmodifikation jeweils eliminiert werden. Die Standardabweichung der Residuen erhöht sich für Rotweine

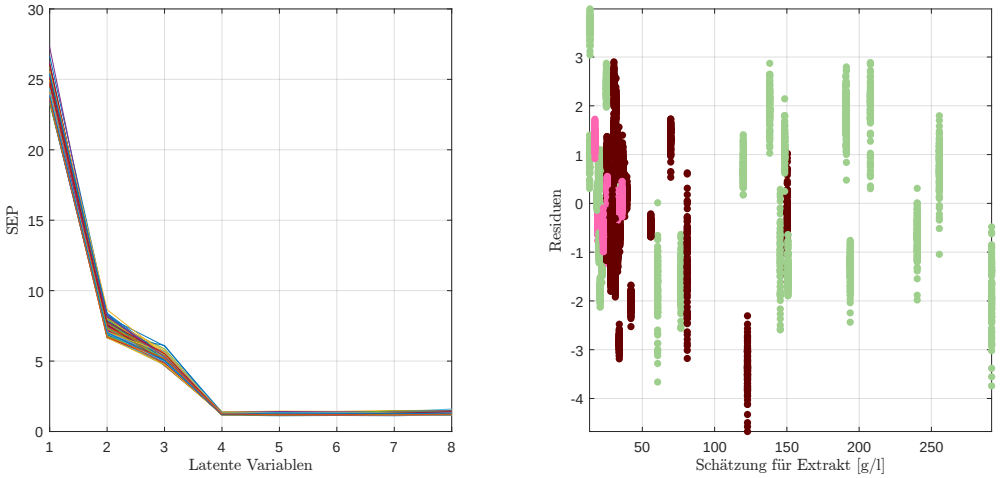

Abbildung 5.7: Verlauf des SEP in Abhängigkeit von latenten Variablen (li.) und die dazu korrespondierenden Weinresiduen der doppelten Kreuzvalidierung [g/l] (re.) für Extrakt im PLS-Modell.

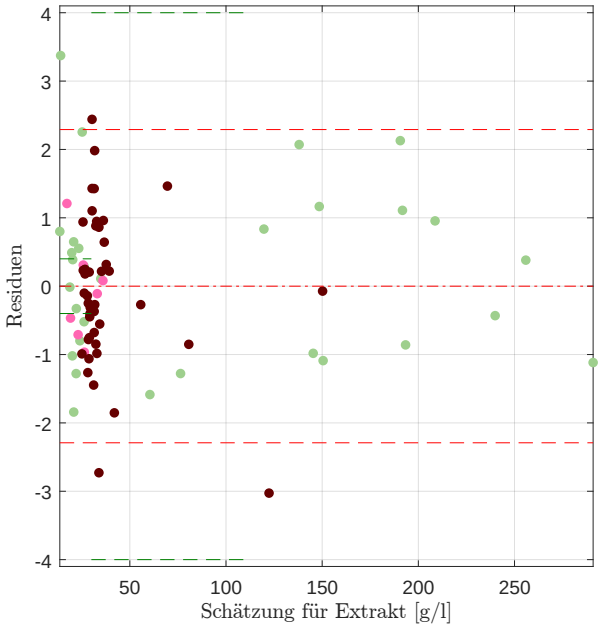

Abbildung 5.8: Residuenplot des PLS-Modells für Extrakt mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode.

von 1.1239 g/l auf 1.211 g/l, wohingegen sie für die Weiß- und Roséweine bei gemeinsamer Betrachtung von 1.1772 g/l auf 1.0813 g/l fällt und kein Trend mehr auffindbar ist.

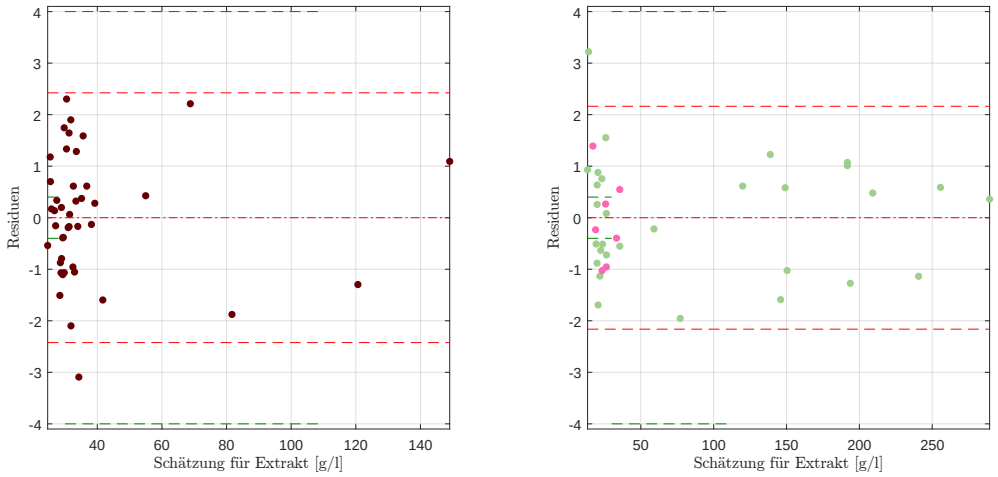

Abbildung 5.9: Residuenplot des PLS-Modells für Extrakt mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode. Kalibrierung des Gesamtmodells mit Rotweinen (li.) und Rosé-/Weißweinen (re.).

Betrachtet man die Einschränkung des Wertebereiches auf $[0.0, 63.5]$ g/l, so kann die Variabilität in diesem Bereich auf Kosten der Erhöhung der Wellenzahlen bei gleichbleibender Anzahl von latenten Variablen verbessert werden. Da in diesem Modell allerdings ein leichter Trend ersichtlich ist, wird die weitere Diskussion dieses Modells unterlassen.

Reproduzierbarkeit

Die Reproduzierbarkeit weist genau einen einzelnen Ausreißer bezüglich der Variabilität der Schätzung mit Einzelspektren auf. Zwischen den einzelnen Weinfarben zeigen sich nur geringe Unterschiede, wie die Abbildung 5.10, mit zugehörender tabellarischer Übersicht in Tabelle 5.5 zeigt. Dies gilt insbesonders für das Gros der Population, den Rot- und Weißweinen. Die Mittelwerte und Mediane der Standardabweichungen stehen stets in einem ähnlichen Verhältnis.

Vergleicht man das Modell durch die Anwendung auf die Daten der übrigen drei Datensätze, so zeigt sich das stabilste Verhalten wiederum in den Datensätzen E22 und E25. Auffallend ist die hohe Reproduzierbarkeit für das Spektrometer E24, insbesondere bei Betrachtung der maximalen Reproduzierbarkeit. Hierbei können keine Rückschlüsse auf eine etwaige Fehlmessung gezogen werden, da speziell die maximale Abweichung bei diesem Datensatz in mehreren Analysen, auch anderer Weinkomponenten, Auffälligkeiten zeigt und diese meist von unterschiedlichen Weinen verursacht werden.

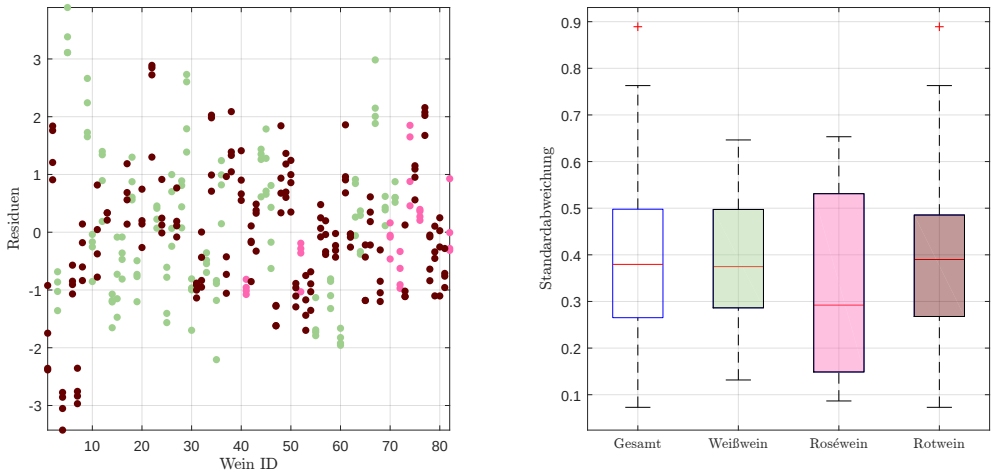

Abbildung 5.10: Residuen in g/l der Einzelspektren pro Wein, ohne Kunstweine, zur Schätzung von Extrakt im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.3889	0.3747	1.4370	0.8539	0.8153	0.3295
Roséwein	0.3382	0.2922	1.3928	0.7413	0.6375	0.4551
Rotwein	0.3966	0.3900	2.0216	0.8843	0.8926	0.4029
Gesamt	0.3886	0.3795	2.0216	0.8604	0.8372	0.3788

Tabelle 5.5: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Extrakt im PLS-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	1.2078	1.57	−3.01	3.16	0.3836	0.3582	1.9258
E24	1.2153	1.73	−3.41	3.03	0.4480	0.4003	4.3639
E25	1.1455	1.64	−3.03	3.37	0.3886	0.3795	2.0216
V70(2016)	1.3740	1.64	−3.58	4.71	0.5151	0.4426	4.1454

Tabelle 5.6: Performance des entwickelten PLS-Modells für Extrakt, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

5.3 Glukose

Für die Bestimmung des Glukosegehaltes ist dieser Abschnitt wie folgt gegliedert:

(a) Ein globales Modell für alle Weine wird vorgestellt.
(b) Es wird versucht, Modellverbesserung für Rotweine vorzunehmen.
(c) Modellierung des Wertebereiches $[0, 10]\,\mathrm{g/l}$.
(d) Kombination von (b) und (c).
(e) Übertragung der entwickelten Modelle auf die restlichen Datensätze.

(a) Modell für E25

Das hier vorgestellte Modell für Glukose verwendet ein PLS-Modell mit linearem Kernel bei einer Selektion von $4 \times 5 = 20$ Wellenzahlen, wie in Abbildung 5.11. Ein Block der Länge 5 startet bei dem Peak bei circa $1\,044\,\mathrm{cm}^{-1}$, ein weiterer Block in der darauf folgenden Senke. Die weiteren beiden Abschnitte verteilen sich in die Region links und rechts des Peaks bei $2\,900\,\mathrm{cm}^{-1}$. Bei Verwendung der zweiten Savitzky-Golay Ableitung korrespondieren diese Teilbereiche jeweils mit Peaks/Senken respektive mit einem Knick im Spektrum.

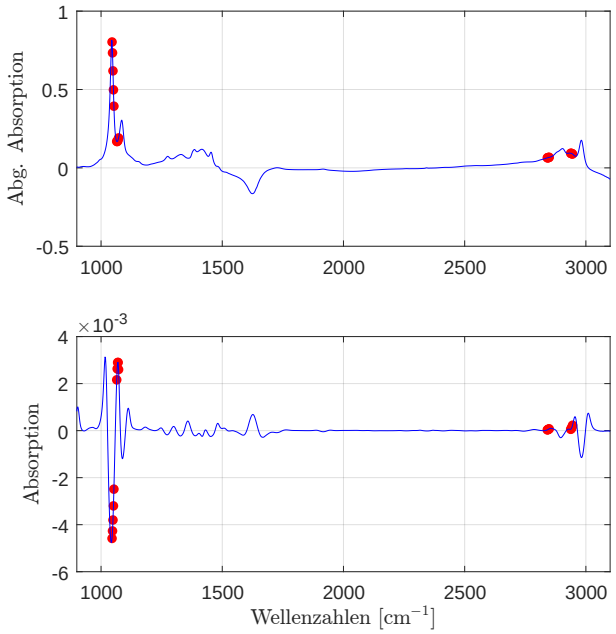

Abbildung 5.11: Die selektierten Wellenzahlen im PLS-Modell für Glukose mit der ersten Savitzky-Golay Ableitung.

Die doppelte Kreuzvalidierung, grafisch aufbereitet in Abbildung 5.12 (li.), legt hierbei eine Wahl von 6 latenten Variablen nahe. Durch die Schätzungen der hieraus resultierenden Residuen in derselben Abbildung, rechts, scheint es, dass Weine mit hohem Glukosewert eine teils höhere Variabilität aufweisen. Die drei Weiß- bzw. Rotweinresiduen in einer Region um $20\,\mathrm{g/l}$, welche als einzelne, verhältnismäßig

große Ausreißer zu identifizieren sind, können auf eine ungünstige Aufteilung der verwendeten Trainins-, Validierungs- und Testsets zurückgeführt werden. Zusätzlich muss die Tatsache beachtet werden, dass mehrere Weißweine beinahe idente Schätzwerte aufweisen, was wiederum eine höhere Variabilität suggeriert.

Die geringste Streubreite tritt erwartungsgemäß bei geringen Glukosewerten auf. Dies kann wiederum damit begründet werden, dass hier die höchste Datendichte vorliegt.

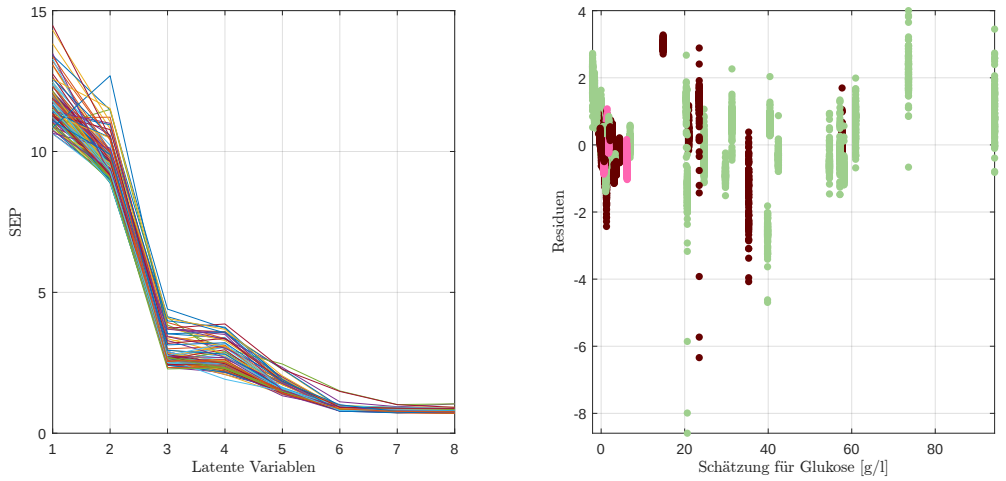

Abbildung 5.12: Verlauf des SEP in Abhängigkeit von latenten Variablen (li.) und die dazu korrespondierenden Weinresiduen der doppelten Kreuzvalidierung [g/l] (re.) für Glukose im PLS-Modell.

Betrachtet man den klassischen Residuenplot in Abbildung 5.13 (li.), so zeigt sich ein homogenes Bild mit großer Punktemenge um 0 g/l. Zudem kann ein Rotweinausreißer beobachtet werden, welcher sich in anderen Analysen weitestgehend unauffällig verhält und nicht auf eine Fehlmessung schließen lässt. Zusätzlich sieht man für hohe Glukosewerte einigermaßen gute Schätzungen. Wie zu erwarten, kann eine abfallende Gerade mit knapp einem Drittel der Datenpunkte beobachtet werden. Dies geht insbesondere aus dem vergrößerten Abschnitt des Wertebereiches $[0, 10]$ g/l hervor, wie der Unterpunkt (c) Einschränkung des Wertebereiches mit den dazugehörenden Residuenplots zeigt. Diese Gerade resultiert aus dem Häufungspunkt von Daten mit einem Glukosegehalt von 0 g/l. Abgesehen von dieser begründbaren Teilstruktur zeigen die Residuen keine weiteren atypischen Verhaltensmuster und die Residuen bewegen sich in einem Bereich zwischen -1.88 g/l bis zu einem Wert von 2.82 g/l bei einer Länge des interquartilen Bereichs von 0.72 g/l. Das eingezeichnete Intervall in Rot berechnet sich mittels der residualen Standardabweichung in Höhe von 0.7099 g/l, welche somit einen dem IQR ähnlichen Wert annimmt. Der hierzu gehörende MSE beläuft sich lediglich auf $0.4977(\text{ g/l})^2$. Insgesamt wird 12 Weinen ein negativer Schätzwert zugeordnet. Diese setzen sich zusammen aus 8 Rot- und 4 Roséweinen.

In Abbildung 5.13 (re.) sind die Schätzwerte gegen die tatsächlichen Werte aufge-

tragen und es zeigt sich eine, aus dem Residuenplot erwartungsgemäß, äußerst gute Approximation - die Diagonale in Schwarz und die rot gestrichelte Linie als linearer Trend der Datenpunkte stimmen beinahe überein.

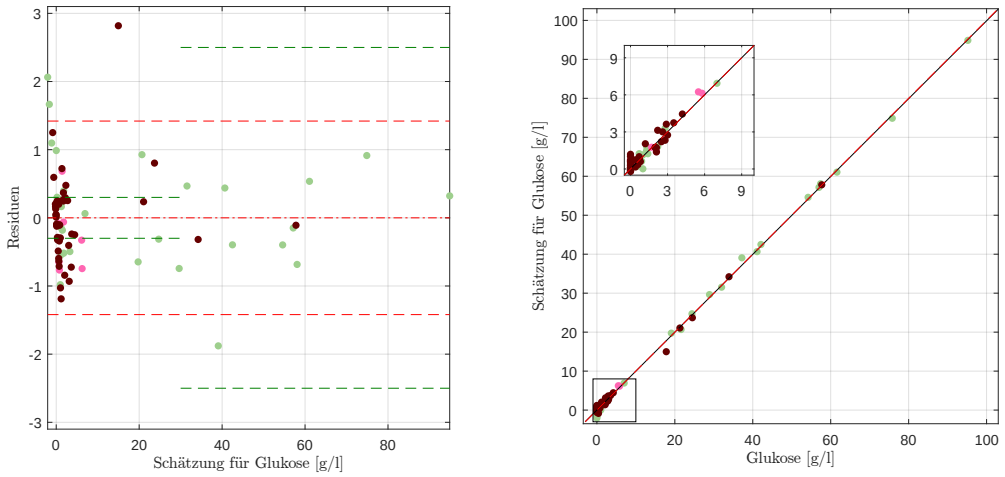

Abbildung 5.13: Residuenplot des PLS-Modells für Glukose mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.), sowie die Schätzungen gegen die tatsächlichen Werte mit optimaler Schätzline, d.h. der rot gestrichelten Diagonale, sowie der durch die Daten motivierte lineare Zusammenhang in Schwarz (re.).

Reproduzierbarkeit

Betrachtet man die Reproduzierbarkeit des Glukosemodells in Abbildung 5.14 beziehungsweise Tabelle 5.7, so zeigen sich, mit Ausnahme eines Ausreißers bei den Weißweinen, durchwegs vergleichbare Reproduzierbarkeitswerte für alle drei Weinfarben. Hierbei handelt es sich um einen Honigwein, welcher allerdings keine erkennbaren Auffälligkeiten in den Daten zeigt und die vergleichsweise schlechte Reproduzierbarkeit kann nicht auf eine explizite Fehlmessung zurückgeführt werden, insbesondere deswegen, weil die Residuen aller Messungen ähnlich weit auseinander liegen.

Insgesamt kann eine gemittelte Standardabweichung der Reproduzierbarkeit von 0.42 g/l bestimmt werden, bedingt wiederum durch Ausreißer. So resultiert der Median als robuste Mittelwertschätzung in 0.35 g/l und es kann in allen Weinfarben, teils auch nur eine minimale, Unterschreitung des Medians gegenüber dem Mittelwert beobachtet werden. Beim Wert mit der maximalen Auslenkung von 5.22 g/l handelt es sich um jenen Wein, welcher auch bezüglich der Variabilität das schlechteste Verhalten zeigt. Eine mögliche Erklärung ist der äußerst hohe Extrakt- und Fruktoseanteil in diesem Wein.

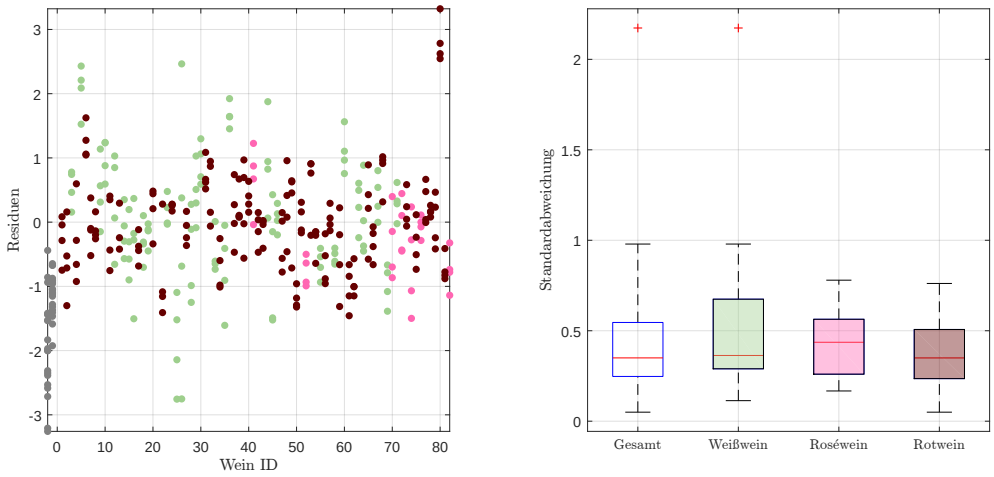

Abbildung 5.14: Residuen in g/l der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung von Glukose im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.4948	0.3634	5.2162	1.1316	0.8270	0.9093
Roséwein	0.4369	0.4363	1.7358	0.9802	0.8952	0.4737
Rotwein	0.3688	0.3497	1.6887	0.8178	0.7351	0.3966
Gesamt	0.4213	0.3499	5.2162	0.9480	0.8045	0.6505

Tabelle 5.7: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Glukose im PLS-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

(b) Restringierung auf Rotweine

Die Rotweine stellen den Großteil des Datensatzes dar. Daher wird versucht, das vorangegangene Modell, eingeschränkt auf die Rotweine, zu verbessern.

Ein hierfür gefundenes Modell verwendet wiederum 6 latente Variablen, ebenfalls mit der zweiten Savitzky-Golay Ableitung der Spektren. Im Gegensatz zum Gesamtmodell wird die Information aus 40 Wellenzahlen benötigt, wobei diese teilweise in derselben Region wie im Gesamtmodell liegen, und werden wie in Abbildung 5.15 (li.) gewählt. Der Gaußkernel mit Parameter $\sigma^2 = 1$ dient zur Nichtlinearisierung des Modells. Für Weine mit einer Glukosekonzentration von über 20 g/l steigt die Variabilität der Residuen der doppelten Kreuzvalidierung eindeutig, kann jedoch auf den Mangel an vorhandenen Daten in diesem Bereich zurückgeführt werden, da die Methodik für das Erzeugen des Residuenplots in Abbildung 5.15 (re.) die Stichproben in drei unterschiedliche Sets unterteilen muss und insbesondere im Trainingsset Informationen zur Schätzung hoher Glukosewerte verloren gehen.

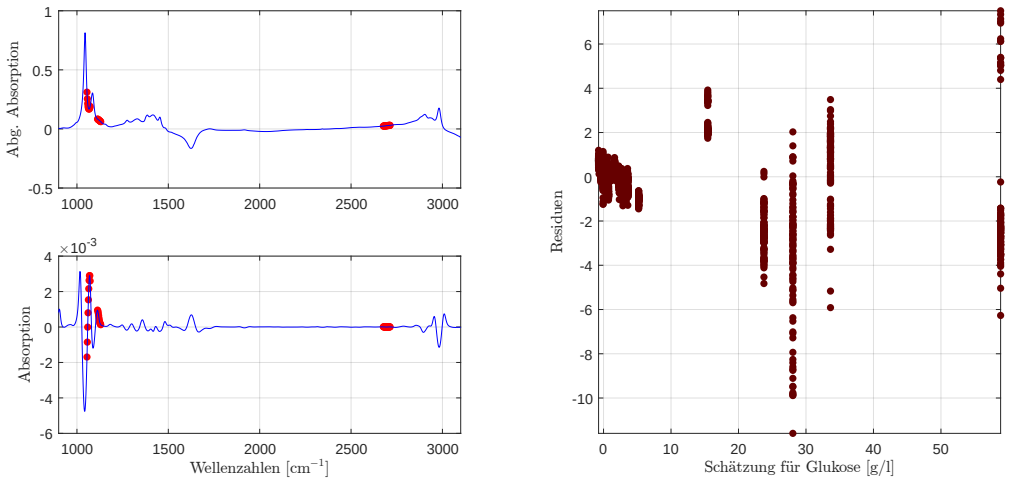

Abbildung 5.15: Die selektierten Wellenzahlen im PLS-Modell für Glukose mit der zweiten Savitzky-Golay Ableitung, eingeschränkt auf Rotweine, (li.) und der aus der doppelten Kreuzvalidierung resultierende Residuenplot (re.).

Im Vergleich der Rotweinresiduen des Gesamtmodells mit dem Residuenplot des hier entwickelten Modells in Abbildung 5.16[6] zeigt sich ein ähnliches Verhalten der Residuen selbst. Weine mit einem Glukosegehalt von über 20 g/l werden trotz mangelnder Daten sehr genau geschätzt. Ein Vorteil des reduzierten Modells ist, dass der Rotweinausreißer erheblich reduziert werden kann, was auf eine gewisse Robustheit gegenüber potentiellen Ausreißern schließen lässt. Mit 11 Rotweinen werden unmerklich weniger negativ geschätzt.

Durch die Verwendung dieses Modells kann die Variabilität der Rotweinresiduen von 0.6626 g/l auf 0.4306 g/l, sowie in ähnlichem Maße der interquartile Bereich,

[6]Man beachte die unterschiedliche Skalierung der Residuenachse.

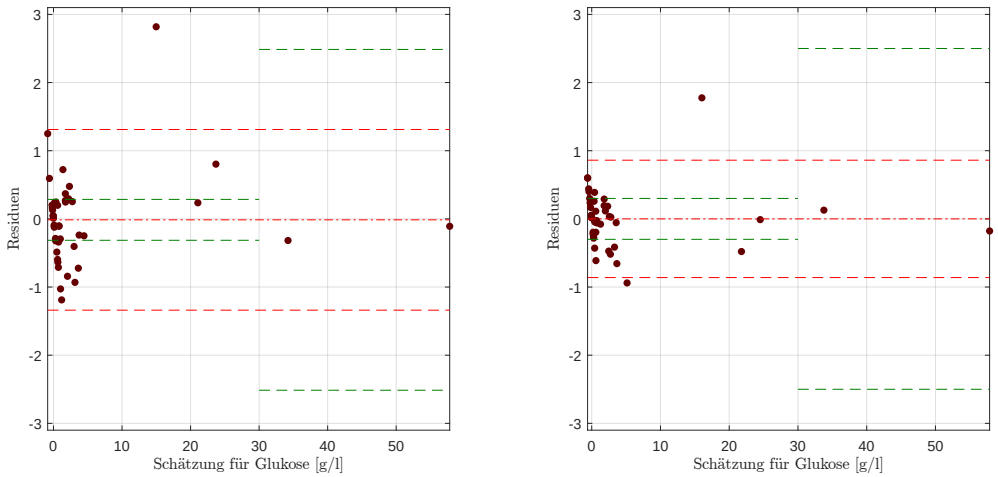

Abbildung 5.16: Extrahierte Rotweinresiduen des Residuenplots des PLS-Modells für Glukose mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.), sowie der Residuenplot des rotweinspezifischen PLS-Modells mit den entsprechenden Kennzahlen (re.).

reduziert werden. Insbesondere verringert sich das Residuum des Ausreißers von 2.82 g/l auf 1.78 g/l.

Ein entscheidender Nachteil dieses Modells ist allerdings das Vorhandensein eines leichten Trends im Bereich unterhalb von 10 g/l, weshalb dieses Modell, trotz Verbesserung der Reproduzierbarkeit für Rotweine hier nicht weiter diskutiert wird und hierfür auf das Modell im folgenden Abschnitt (c) Einschränkung des Wertebereiches verwiesen wird, zumal lediglich vier Rotweine außerhalb des Wertebereiches $[0, 10]$ g/l liegen. Zusätzlich finden sich die wesentlichen Kennzahlen zur Wiederholbarkeit in tabellarischer Form auf den Seiten 120ff.

(c) Einschränkung des Wertebereiches

Da die Datenpunkte eine starke Konzentration im Bereich bis zu 10 g/l aufweisen, wird in diesem Unterabschnitt versucht, für geringere Konzentrationen von Glukose ein eigenes Modell zu entwickeln. Hierbei werden die Daten auf den Wertebereich $[0, 10]$ g/l eingeschränkt. Somit stehen noch 62 Daten zur Verfügung, wobei diese wiederum jene 26 Weine mit einem Glukosehalt von 0 g/l beinhalten.

Für dieses spezifische Modell resultiert aus der heuristischen Methode zur Modellkalibrierung eine Wahl von ebenfalls 20 Wellenzahlen und 6 latenten Variablen bei gleicher Nichtlinearisierung wie für das Rotweinmodell. Das Preprocessing mittels Savitzky-Golay Ableitung zweiter Ordnung mit den hier verwendeten Modellen ist in Abbildung 5.17 grafisch dargestellt. Der Großteil der Information wird wiederum aus dem Bereich des maximalen Peaks gewonnen, wohingegen andere

(weniger) Wellenzahlen außerhalb des Fingerprintbereiches vom PLS-Modell zur Kompensation kleinerer Fehler hinzugenommen werden.

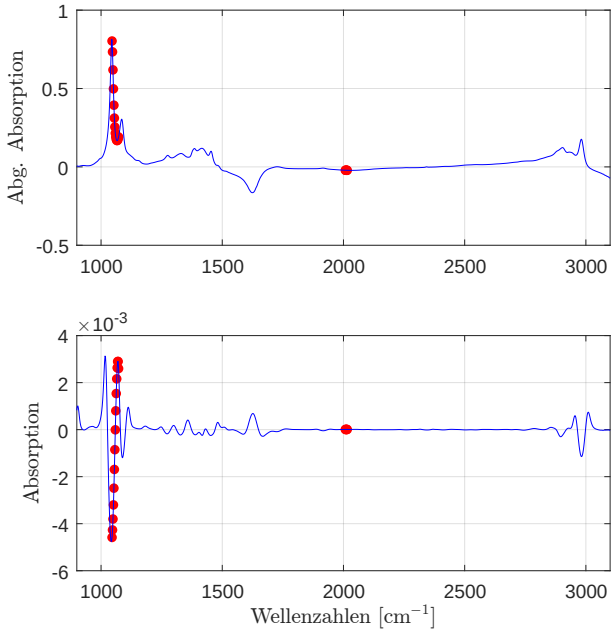

Abbildung 5.17: Die selektierten Wellenzahlen im PLS-Modell für Glukose, eingeschränkt auf den Wertebereich $[0, 10]$ g/l mit der zweiten Savitzky-Golay Ableitung.

Wiederum sind keine unerklärlichen Auffälligkeiten aus den Residuenplots abzulesen. Die größte Schwankungsbreite und der im Residuenplot auffallende Ausreißer resultieren aus den glukosefreien Weinen. Insgesamt kann die Standardabweichung im Vergleich zu jener des Gesamtmodells für denselben Wertebereich von 0.6221 g/l auf 0.2384 g/l mehr als halbiert werden. Das verbesserte Verhalten des Residuenplots zeigt sich in Abbildung 5.18. Die Anzahl der negativ geschätzten Weine reduziert sich auf 2 Weiß- und 7 Rotweine.

Reproduzierbarkeit

Eine ähnlich gute Verbesserung kann bei der Reproduzierbarkeit beobachtet werden. Die maximale Standardabweichung der Einzelmessungen pro Wein ID reduziert sich auf unter 0.55 g/l, wohingegen ein Ausreißer bei den Rotweinen beobachtet werden kann (vgl. Abbildung 5.14). Insgesamt zeichnet sich wie zuvor ein ähnliches Bild aller drei Weintypen ab, wobei bei dieser Datenreduktion bei den Weißweinen der Median über dem Mittelwert liegt, da einige Weißweine eine überproportional gute Reproduzierbarkeit mit beinahe identen Schätzungen der Einzelspektren aufweisen.

Betrachtet man zusätzlich die maximale Abweichung, wie in Tabelle 5.8 aufgelistet, so nimmt diese mit bis zu 1.18 g/l im Vergleich zum betrachteten Wellenzahlbereich

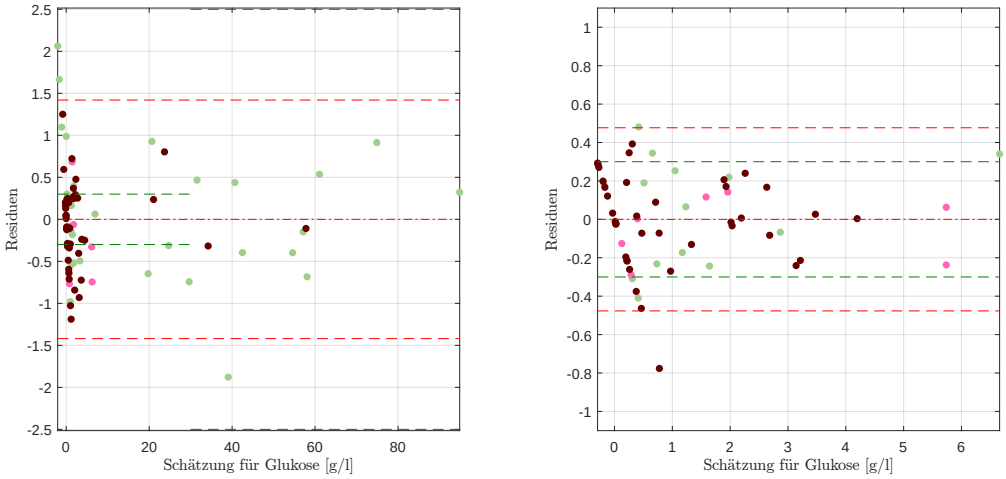

Abbildung 5.18: Extrahierte Residuen für den Wertebereich $[0, 10]$ g/l des Residuenplots des PLS-Modells für Glukose mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.), sowie der Residuenplot des wertebereichspezifischen PLS-Modells mit den entsprechenden Kennzahlen (re.).

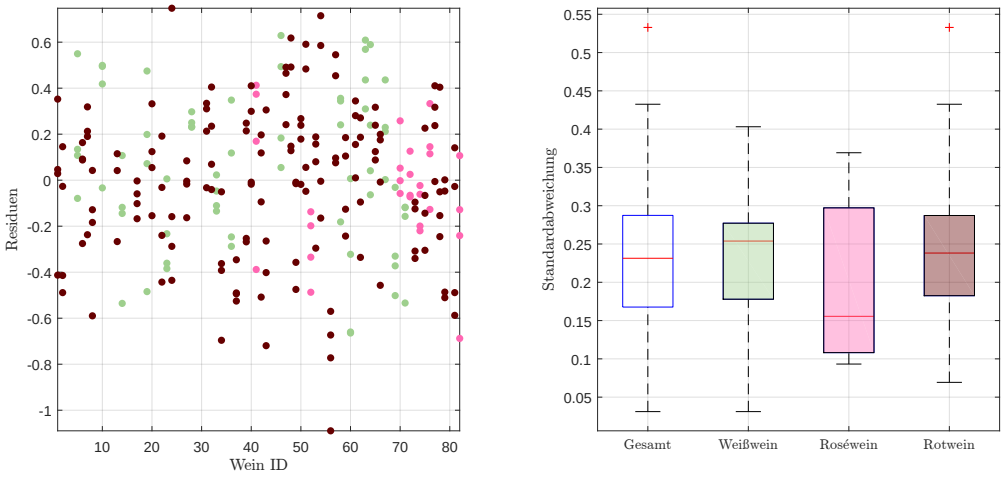

Abbildung 5.19: Residuen in g/l der Einzelspektren pro Wein, ohne Kunstweine, zur Schätzung der Glukose im PLS-Modell, eingeschränkt auf den Wertebereich $[0, 10]$ g/l (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

einen enorm hohen Wert an.

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.2260	0.2538	0.9590	0.5042	0.5349	0.2154
Roséwein	0.1966	0.1556	0.8009	0.4452	0.3498	0.2573
Rotwein	0.2385	0.2382	1.1829	0.5193	0.4897	0.2179
Gesamt	0.2305	0.2314	1.1829	0.5070	0.4881	0.2192

Tabelle 5.8: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Glukose im PLS-Modell, eingeschränkt auf den Wertebereich $[0, 10]\,g/l$, anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

(d) Modell für Rotweine mit geringer Glukosekonzentration

Da für die Modellierung von Rotweinen mit einem Glukosegehalt von weniger als $10\,g/l$ lediglich 39 Proben zur Kalibrierung, von welchen für die Ermittlung des SEP einige Testdaten weggelassen werden müssen, verwendet werden können, wird für diese Modellspezifikation nur der Wellenzahlplot vorgestellt. Bei einer latenten Variablenanzahl von 5 und einem Gaußkernel mit Parameter $\sigma^2 = 3$ werden wiederum 20 Wellenzahlen wie in Abbildung 5.20 (li.) selektiert. Auch wenn das Modell einigermaßen stabil erscheint und mit einer latenten Variablen weniger auskommt, kann einerseits eine enorme Verbesserung gegenüber dem Gesamtmodell erzielt werden, nicht allerdings im Vergleich zum Modell mit demselben Wertebereich mit allen Weinproben aus Unterpunkt (c) Einschränkung des Wertebereiches, wie der Residuenplot in Abbildung 5.20 (re.) zeigt. Aufgrund dessen entfällt, auch wenn der Residuenplot unauffällig aussieht[7], eine weiterführende Analyse dieser Modellspezifikation.

Die wichtigsten Kennzahlen können wiederum in tabellarischer Form der Gesamtübersicht der Kennzahlen der PLS-Modelle entnommen werden.

(e) Datensatzvergleich

Betrachtet man Tabelle 5.9 mit der Übersicht über das Verhalten des Gesamtmodells aus (a) Modell für E25 für die Datensätze des Jahres 2016, so weist E25 die geringste Standardabweichung der Residuen auf und zeugt somit von einer hohen Datenqualität. Vergleicht man die Reproduzierbarkeit, so zeigt sich ein auffälliges Verhalten für den äußerst fruktoselastigen Wein, mit einer maximalen Differenz der Schätzungen der Einzelresiduen in Höhe von $5.22\,g/l$, während bei den restlichen Datensätzen dieser Unterschied maximal $2.66\,g/l$ misst. Es kann allerdings keine

[7] Mit der Ausnahme, dass die meisten Weine ohne Glukose überschätzt werden und somit, um den Bias zu eliminieren, die restlichen Weinresiduen tendentiell positiv sind.

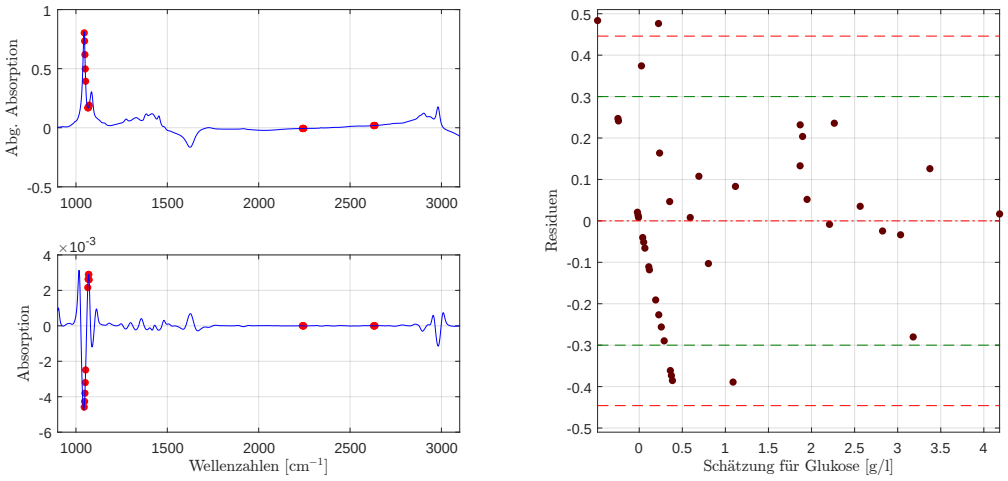

Abbildung 5.20: Die selektierten Wellenzahlen im PLS-Modell für Glukose, eingeschränkt auf den Wertebereich $[0, 10]$ g/l und Rotweine, mit der zweiten Savitzky-Golay Ableitung (li.) und der dazugehörende Residuenplot mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (re.).

Fehlmessung als solche identifiziert werden, wie im entsprechenden Abschnitt erläutert.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.7952	0.97	−2.97	3.26	0.4071	0.3639	2.4587
E24	0.9114	0.87	−3.18	2.58	0.3876	0.3423	2.6574
E25	0.7099	0.72	−1.88	2.82	0.4213	0.3499	5.2162
V70(2016)	0.7170	0.70	−1.74	2.89	0.3277	0.3166	1.8471

Tabelle 5.9: Performance des entwickelten PLS-Modells für Glukose, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

Für das Modell mit einem auf $[0, 10]$ g/l reduzierten Wertebereich zeigt Tabelle 5.10 wiederum das (leicht) bessere Residuenverhalten für den Datensatz E25. Insbesondere reduziert sich das Verhältnis der maximalen Abweichung, verglichen mit den restlichen Weinen, enorm. Auffallend ist dennoch, dass der Datensatz E22 insbesondere bezüglich der Reproduzierbarkeit erhebliche Vorteile, mit einer maximalen Differenz in den Einzelspektren je Wein von 0.78 g/l, zeigt, während das zur Modellentwicklung herangezogene Spektrometer E25 einen Wert in Höhe von 1.18 g/l aufweist. Weiters gilt, dass insbesondere der interquartile Bereich deutlich über der Standardabweichung liegt.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.2536	0.36	−0.57	0.54	0.1623	0.1607	0.7885
E24	0.2726	0.34	−0.62	0.69	0.2286	0.2324	1.3015
E25	0.2384	0.39	−0.78	0.48	0.2305	0.2314	1.1829
V70(2016)	0.3080	0.42	−0.73	0.66	0.2057	0.1952	0.9210

Tabelle 5.10: Performance des entwickelten PLS-Modells für Glukose, eingeschränkt auf den Wertebereich $[0, 10]$ g/l, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

Da das letzte Modell vergleichsweise wenige Datensätze beinhaltet, wird an dieser Stelle der Datensatzvergleich für das auf $[0, 10]$ g/l Rotweinmodell nicht angestellt, da sich zudem ein ähnliches Bild wie beim reduzierten Modell in Tabelle 5.10 mit etwas geringeren Werten zeigt. Weiters verweist das Modell für Rotweine auf jenes mit dem eingeschränkten Wertebereich für alle Weine, weshalb auch dieser Datensatzvergleich an dieser Stelle obsolet ist.

5.4 Fruktose

Die zweite wichtige Zuckerart bildet der Fruchtzucker, die sogenannte Fruktose. Bei dem Versuch, ein funktionales Modell zwischen den Spektren und den Referenzwerten herzustellen, verwendet das PLS-Modell lediglich 2 latente Variablen in Kombination mit einem Gaußkernel und Parameter $\sigma^2 = 0.1$. Die hierfür optimalen Wellenzahlen sind in Abbildung 5.21 dargestellt. Insgesamt unterteilen sich diese in 4 Blöcke der Länge 10, wobei sich hiervon zwei im Fingerprintbereich befinden, jeweils zu beiden Seiten der Bande, welche den maximalen Peak einschließt. Die restliche Information filtert das PLS-Modell aus dem Anstieg zum Peak vor $2\,890\,\mathrm{cm}^{-1}$ heraus, wobei wiederum die zweite Savitzky-Golay Ableitung zur Datenaufbereitung dient.

Bei dieser Wellenzahlselektion folgt unmittelbar die Wahl von zwei latenten Variablen, wie Abbildung 5.22 (li.) zeigt. Während die Verdoppelung dieses Parameters von 1 auf 2 eine signifikante Verbesserung für etwaige Schätzungen neuer Proben herbeiführt, kann keine weitere Information durch Hinzunahme von latenten Variablen aus den Spektren gewonnen werden. Für die aus der Kreuzvalidierung resultierenden Residuen in Abbildung 5.22 (re.) zeigt sich eine erhöhte Variabilität für hohe Fruktosekonzentrationen. Dies liegt daran, dass für die Schätzung dieser Residuen Weine mit hohem Fruchtzuckergehalt aufgrund der Vorgehensweise teils weggelassen werden (müssen) und zudem mit über drei Viertel der Daten die höchste Konzentration im Bereich $[0, 10]$ g/l liegt. Neben der Variabilität innerhalb der Residuen, bedingt auf einen Wein mit hohem Fruchtzucker, kann zusätzlich eine erhöhte Streuung der gemittelten Residuen beobachtet werden.

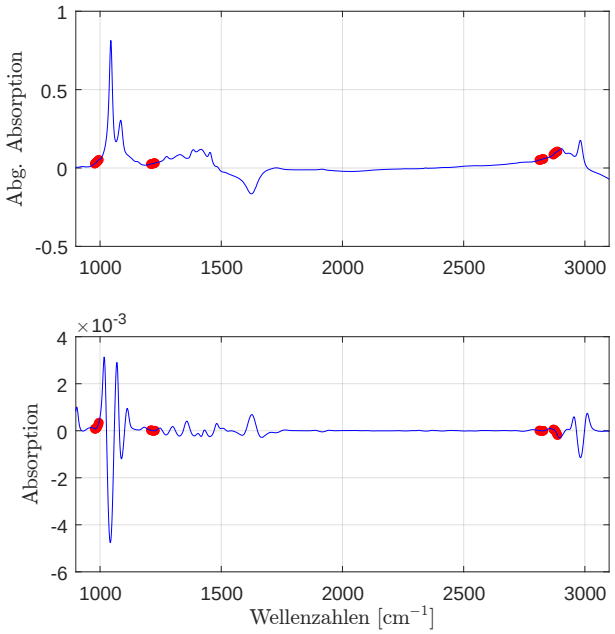

Abbildung 5.21: Die selektierten Wellenzahlen im PLS-Modell für Fruktose mit der zweiten Savitzky-Golay Ableitung.

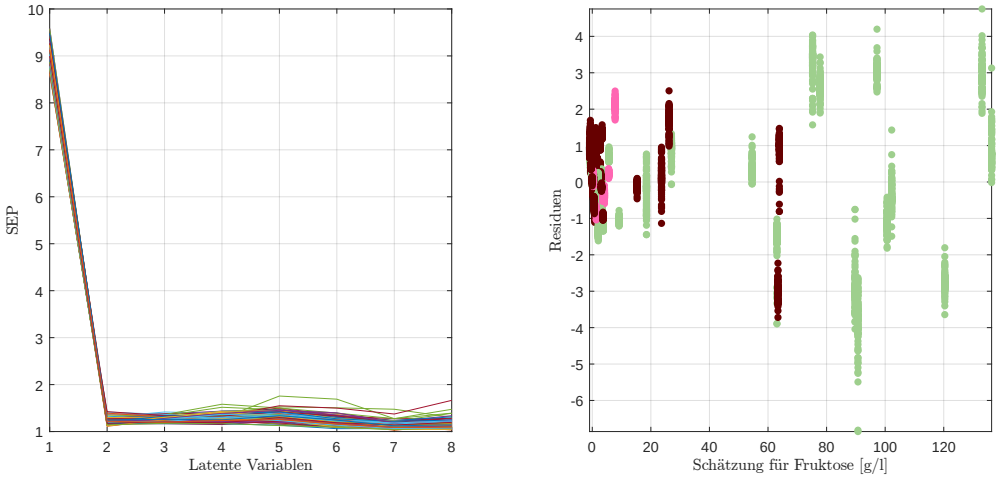

Abbildung 5.22: Verlauf des SEP in Abhängigkeit von latenten Variablen (li.) und die dazu korrespondierenden Weinresiduen der doppelten Kreuzvalidierung [g/l] (re.) für Fruktose im PLS-Modell.

Diese Erkenntnis kann auch auf den klassischen Residuenplot in Abbildung 5.23 (li.) übertragen werden. Für die Modellbildung von Süßweinen mit hohem Fruchtzuckeranteil scheint das Modell an Variabilität zuzunehmen, auch wenn die Standardabweichung der Referenzmethode mit einem Wert von über 30 g/l unterboten werden kann, obwohl die Aussagekraft an dieser Stelle aus Mangel an Stichproben in diesem Wertebereich stark begrenzt ist. Abgesehen von der erhöhten Variabilität können keine Auffälligkeiten wie Ausreißer oder Trends beobachtet werden. Betrachtet man alle Residuen gemeinsam, so beläuft sich die Standardabweichung auf 1.051 g/l, wobei wie bei der PLS-Methode üblich, kein Bias beobachtet werden kann. Die Residuen finden sich in einem Bereich von -3.00 g/l bis 3.21 g/l wieder, wobei der interquartile Bereich eine Länge von etwas mehr als einer Standardabweichung, 1.23 g/l, misst. Die Abweichung, bemessen mittels des MSEs, beläuft sich für dieses Modell auf $1.1016(\,\mathrm{g/l})^2$. Ein Indiz für die Güte des Modells ist eine erklärte Varianz von über 99 %. Zusätzlich muss erwähnt werden, dass insgesamt 9 Weinproben derart unterschätzt werden, dass diesen ein negativer Fruktosegehalt unterstellt wird. Für praktische Zwecke muss der Referenzbereich angepasst und die Werte auf 0 g/l korrigiert werden.

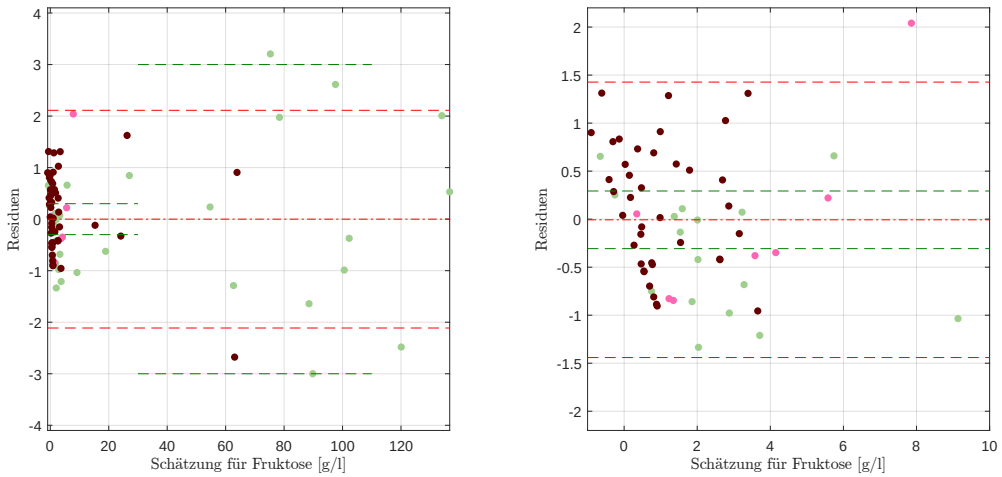

Abbildung 5.23: Residuenplot des PLS-Modells für Fruktose mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.), sowie der vergrößerte Ausschnitt $[0, 10]$ g/l mit Bias und Abweichungen für diesem Bereich.

In Abbildung 5.23 (re.) sind die Residuen dieses Modells für den Bereich $[0, 10]$ g/l vergrößert dargestellt, mittels der hierfür eigens berechneten Standardabweichung. Der für dieses Segement auftretende Bias ist verschwendend gering und liegt unter der betrachteten Genauigkeit von 10^{-1} g/l. Verglichen mit der Variabilität des Gesamtmodells reduziert sich die Standardabweichung dieser Weinresiduen auf 0.7170 g/l. Ebenfalls deutlich erkennbar sind jene 9 Weine, welche einen negativen Schätzwert aufweisen, wobei es sich hierbei nicht ausschließlich um Weine mit einer Fruktosekonzentration von 0 g/l handelt. Die bemerkenswerteste Reduktion der Kennzahlen, wie in Tabelle 5.11 dargestellt, bildet die Abweichung des MSE,

welcher sich um mehr als die Hälfte reduziert und es zeigt sich ein, restringiert auf den Wertebereich $[0, 10]$ g/l, unauffälliges Verhalten, auch wenn selbst in diesem Unterausschnitt des Gesamtmodells die Standardabweichung der Referenzmethode mehr als doppelt so hoch ausfällt.

	Anz.	Mw.	Std.	Min.	Med.	Max.	IQR	MSE
Gesamtmodell	81	0.00	1.0561	−3.00	−0.01	3.21	1.23	1.1016
$[0, 10]$ g/l	62	−0.01	0.7170	−1.33	0.00	2.04	1.05	0.5059

Tabelle 5.11: Kennzahlenübersicht für Fruktose mit den Kennzahlen des PLS-Modells mit zusätzlicher Einschränkung auf den Wertebereich $[0, 10]$ g/l. Alle Werte in g/l.

Betrachtet man ausschließlich die Rotweinresiduen des Gesamtmodells wie in Abbildung 5.24 (li.), sowie jene, welche sich durch eine Neukalibrierung des Gesamtmodells mit den Rotweinen ergeben, so reduziert sich auch in diesem Submodell die Standardabweichung, indem insbesondere die Überschätzung des Ausreißers bei einem Schätzwert von 63.9 g/l enorm reduziert wird und somit als robustes Teilmodell aufgefasst werden kann.

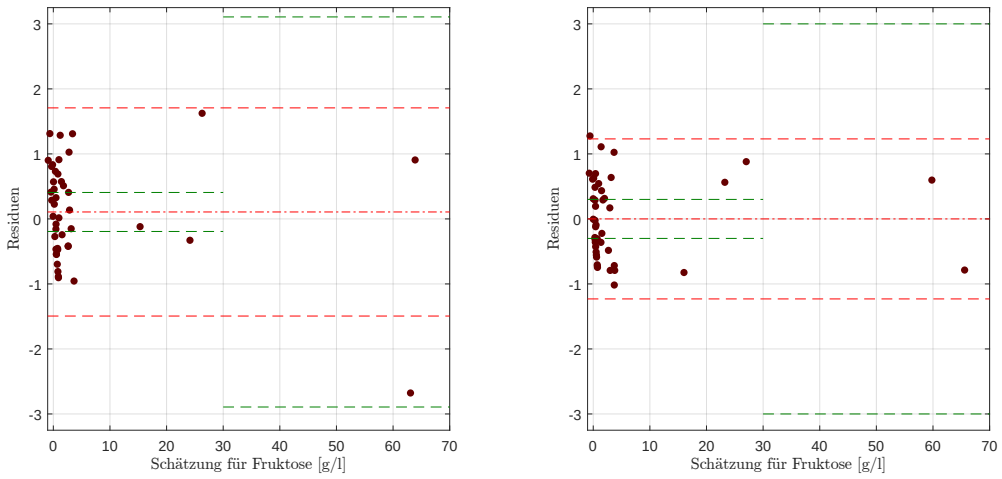

Abbildung 5.24: Residuenplot des PLS-Modells für Fruktose mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode, mit Fokus auf die Rotweine des Gesamtmodells (li), sowie den Rotweinresiduen durch die Neukalibrierung der Rotweinen.

Zusätzlich wurden Modelle ermittelt, welche bei Betrachtung der Kennzahlen niedrigere Werte aufweisen, als das hier vorgestellte. Diese Modelle wurden bei Betrachtung der klassischen Residuenplots jedoch ausgeschlossen, da die Grafiken für eine Selektion nicht die notwendigen Merkmale wie Unabhängigkeit der Strukturen oder einer klaren Unter-/Überschätzung der Rot-/Weißweine mit sich bringen.

Reproduzierbarkeit

Die Reproduzierbarkeit in Abbildung 5.25 zeigt ein einigermaßen homogenes Verhalten, auch wenn der Boxplot der Weißweinresiduen etwas längere Tails, einschließlich zweier Ausreißer bei der Betrachtung des Gesamtmodells, aufweisen. Die mittlere Standardabweichung beträgt hier jeweils weniger als 0.20 g/l, bei ähnlichem Median. Größere Unterschiede können allerdings bei den Kennzahlen der maximalen Abweichung in Tabelle 5.12 beobachtet werden. Während sich die vier Residuen der Roséweine lediglich um maximal 0.58 g/l unterscheiden, sind im Gesamtmodell Weine mit einer Abweichung von bis zu 0.96 g/l beobachtbar. Dies wiederum bedeutet, dass die Mittlung mehrerer gemessener Spektren von entscheidender Wichtigkeit sein kann, zumal keine Auffälligkeiten in diesen Weinen ersichtlich sind.

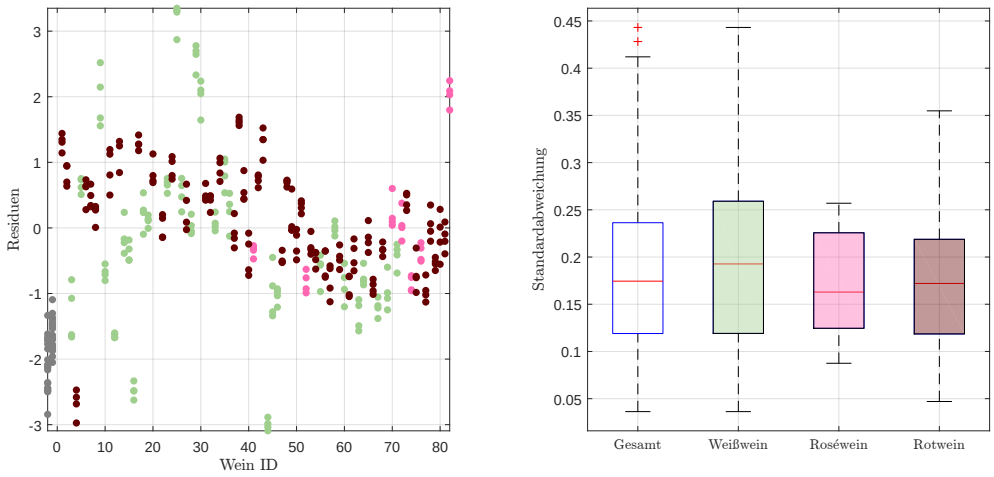

Abbildung 5.25: Residuen in g/l der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung der Fruktosekonzentration im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.1992	0.1927	0.9618	0.4499	0.4415	0.2247
Roséwein	0.1698	0.1629	0.5767	0.3787	0.3588	0.1531
Rotwein	0.1798	0.1720	0.8372	0.3954	0.3640	0.1682
Gesamt	0.1861	0.1745	0.9618	0.4141	0.3743	0.1899

Tabelle 5.12: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Fruktose im PLS-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

Bei dem Datensatzvergleich zeigt sich in allen Datensätzen ein ähnliches Verhalten

der Residuen, wohingegen die Spektren von E22 und E25 leicht bessere Wiederholbarkeit zeigen.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	1.1203	1.23	−3.47	3.34	0.1812	0.1714	1.0265
E24	1.1375	1.26	−3.43	3.24	0.2072	0.1904	1.4276
E25	1.0561	1.23	−3.00	3.21	0.1861	0.1745	0.9618
V70(2016)	1.0891	1.30	−2.81	3.67	0.2187	0.2088	1.1603

Tabelle 5.13: Performance des entwickelten PLS-Modells für Fruktose, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

5.5 Titrierbare Säure

Ähnlich zum Abschnitt über ein Modell für Extraktwerte handelt es sich bei den titrierbaren Säuren wiederum um einen Summenparameter, wessen hier analysierte Bestandteile Wein-, L-Äpfel-, Milch- und Zitronensäure sind. In einem ersten Schritt wird versucht, ein PLS-Modell an die titrierbaren Säuren anzupassen, während in den nachfolgenden Abschnitten 5.6 bis 5.10 spezifische Säuren untersucht und modelliert werden.

Ein für die titrierbaren Säuren heuristisch optimales Modell schätzt die Konzentrationen mit einem linearen Kernel, wobei ein Gaußkernel teilweise zu demselben Kalibrierungssetting[8] führt. Insgesamt liegen die hierfür selektierten 20 Wellenzahlen in den Regionen $2\,260\,\mathrm{cm}^{-1}$ und $2\,225\,\mathrm{cm}^{-1}$, sowie $1\,875\,\mathrm{cm}^{-1}$ und $1\,740\,\mathrm{cm}^{-1}$, jeweils in Blöcken der Länge 5, wie in Abbildung 5.26, vor.

Die weiteren Modellparameter setzen sich aus der ersten Savitzky-Golay Ableitung, sowie zwei latenten Variablen, deren Wahl durch den SEP Verlauf in Abbildung 5.27 (li.) gerechtfertigt wird. In dieser Grafik kann ein starker Abfall durch die Erhöhung auf 2 latente Variablen beobachtet werden, wohingegen der weitere Verlauf der SEP Kurve lediglich (leichte) Verschlechterungen zeigt. Dieser Verlauf ist gleichbedeutend mit den beiden Tatsachen, dass keine Zusatzinformation mit mehreren Variablen modelliert werden kann, da einerseits nur für den Trainingsdatensatz spezifische Eigenheiten mitmodelliert würden, sowie der PLS-Algorithmus für die latenten Variablen an den entsprechenden Stellen nicht konvergiert.

Betrachtet man das Histogramm der aus dieser doppelten Kreuzvalidierung stammenden 100 Residuen pro Wein, so zeigt sich eine gewisse Bimodalität. Die beiden Häufungspunkte deuten auf einen zusätzlichen (nicht berücksichtigten) Einflussfaktor auf dieses Modell hin und der Residuenplot in Abbildung 5.28 liefert die

[8] Sowohl die Anzahl an latenten Variablen, als auch die Wellenzahlselektion stimmen überein.

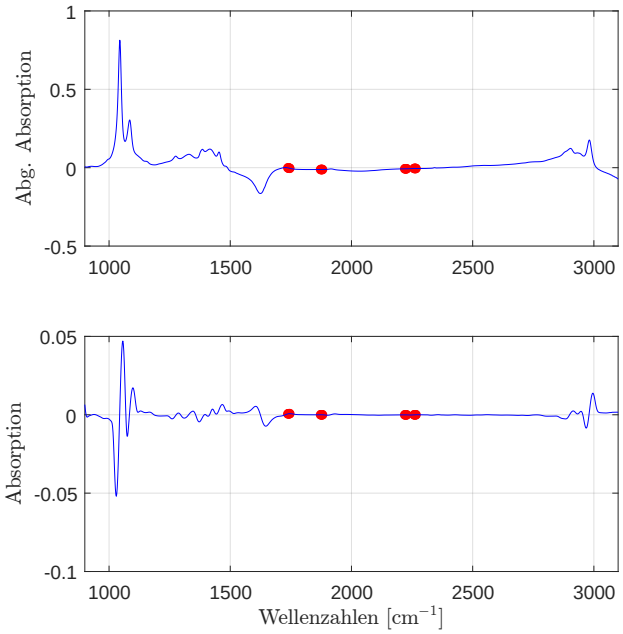

Abbildung 5.26: Die selektierten Wellenzahlen im PLS-Modell für titrierbare Säuren mit der ersten Savitzky-Golay Ableitung.

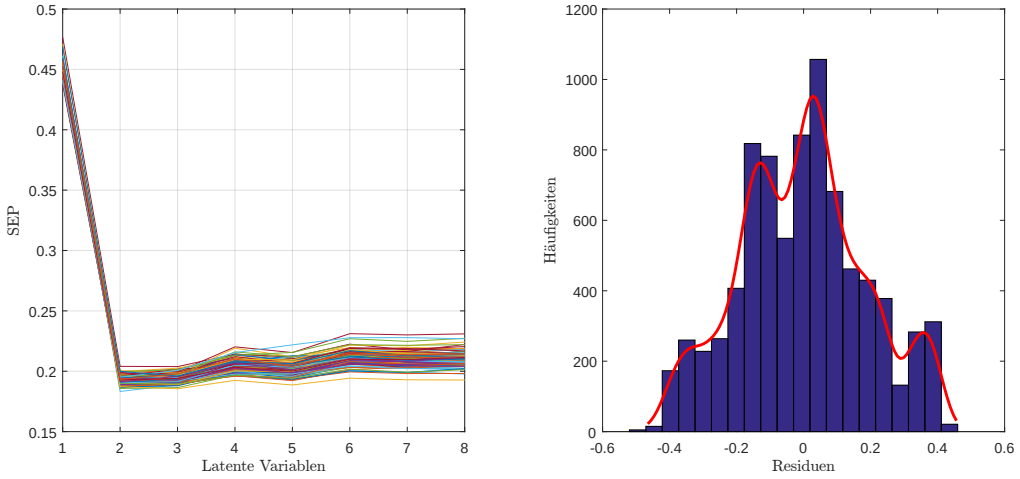

Abbildung 5.27: Verlauf des SEP in Abhängigkeit von latenten Variablen (li.) und die dazu korrespondierenden Weinresiduen der doppelten Kreuzvalidierung [g/l] für titrierbare Säure im PLS-Modell, dargestellt als Histogramm mit einer empirischen Dichteschätzung.

zugehörige Erklärung. Auch wenn auf den ersten Blick im gesamten Erscheinungsbild diese Grafik unauffällig wirkt und keine Strukturen wie Trends aufweist, so legt diese Grafik nahe, als zusätzliche Komponente die Weinfarbe in das Modell aufzunehmen, da die Rotweine tendenziell unterschätzt, sowie die Weiß- und Roséweine in gleichem Ausmaß überschätzt werden. Für das Gesamtmodell, welches keinen Bias aufweist, bedeutet dies, dass für Rotweine ein positiver, wohingegen für Weiß- und Roséweinen ein negativer Bias auftritt. Dennoch scheint das Modell als Gesamtmodell gut zu funktionieren und die Genauigkeit der Referenzmethode kann ebenfalls erreicht werden. Während die im Residuenplot abgebildeten Residuen eine Standardabweichung von 0.1788 g/l aufweisen, beläuft sich der MSE auf $0.0316(\,\mathrm{g/l})^2$.

Für die Methodik der PLS-Modelle bedeutet dies, Modellparameter für die Weinfarben getrennt zu kalibrieren, was anhand folgender Überlegung ersichtlich ist. Erweitert man die Matrix mit den relevanten Absorptionswerten $X \in \mathbb{R}^{81\times 20}$ um die Weinfarbe auf $X \in \mathbb{R}^{81\times(20+1)}$, so würde auch mit der Faktorvariable Farbe eine Art Distanzmessung stattfinden, wie diese auf Seite 46 beschrieben wird, weshalb das Modell auf diese Art nicht kalibriert werden kann und getrennte Modellkoeffizienten ermittelt werden müssen.

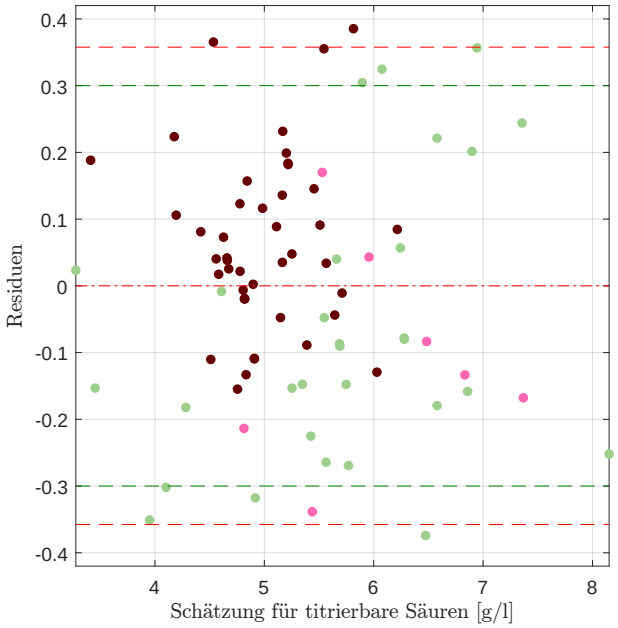

Abbildung 5.28: Residuenplot des PLS-Modells für titrierbare Säuren mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode.

Extrahiert man aus Abbildung 5.28 die Residuen aller Rotweine, so führt dies zu einem Offset in Höhe von 0.06 g/l und einer Standardabweichung der Residuen im Ausmaß von 0.1304 g/l und die Vermutung der Bimodalität der Verteilung der Residuen kann tatsächlich auf die Weinfarbe reduziert werden. Abbildung 5.29 stellt die

Rotweinresiduen, einerseits durch die Neukalibrierung des Modells mit den Spezifikationen des Gesamtmodells und andererseits jene, welche durch Schätzungen in einem eigens für Rotweine entwickelten Modell, resultieren.

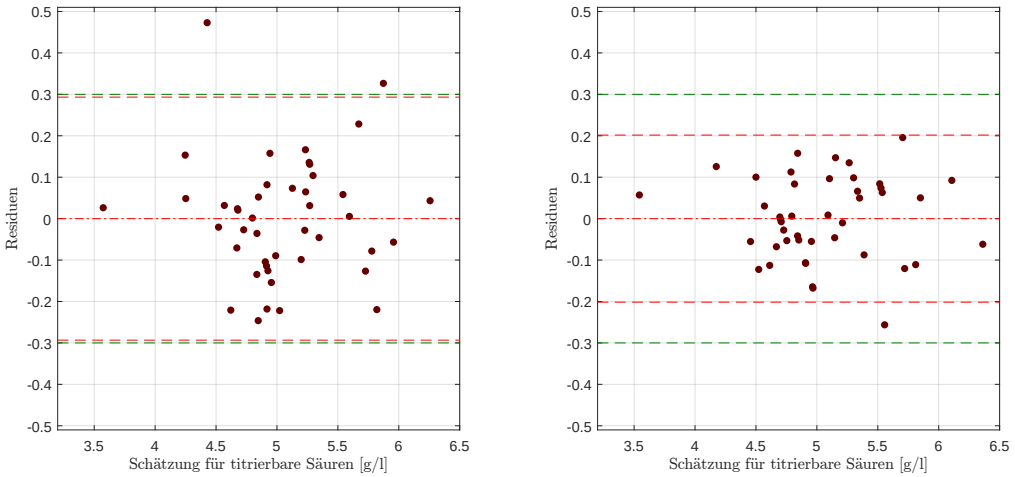

Abbildung 5.29: Residuenplot des PLS-Modells für titrierbare Säuren mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode, mit Neukalibrierung des Gesamtmodells für Rotweine (li), sowie der Residuenplot des rotweinspezifischen PLS-Modells mit den entsprechenden Kennzahlen (re.).

Bei der aus der Neukalibrierung stammenden Grafik kann im Gegensatz zum Residuenplot des Gesamtmodells ein Ausreißer beobachtet werden, welcher ein Residuum von über 0.4 g/l besitzt, während im Gesamtmodell nur geringere Residuen beobachtet werden konnten. Zusätzlich zeigt sich eine geringfügige Erhöhung der Variabilität auf 0.1467 g/l. Im Umkehrschluss bedeutet dies, dass Weiß-/Roséweinspektren relevante Informationen für die Rotweine beinhalten.

Vergleicht man diese Residuen mit einem eigens für Rotweine kalibrierten Modell, so können Verbesserungen in der Volatilität, allerdings auf Kosten der Modellparameter, erzielt werden. So verwendet dieses Modell eine zusätzliche latente Variable bei insgesamt 40, und somit doppelt so vielen, Wellenzahlen.

Betrachtet man das für Weiß- und Roséweine neukalibrierte Gesamtmodell, so zeigt sich ein unauffälliger Residuenplot mit einer erhöhten Variabilität von 0.2068 g/l und einem, verglichen mit dem Gesamtmodell, minimal längeren Intervall, welches die betragsmäßig größten Residuen umschließt. Die Unterschiede in der farbspezifischen Kalibrierung resultieren aus der Datensituation für die titrierbaren Säuren, da bei annähernd gleicher Anzahl von Datensätzen die Verteilung des Wertebereichs signifikant unterschiedlich ist, wie folgender Abschnitt zeigt.

Datensituation für titrierbare Säuren

Das unterschiedliche Verhalten und insbesondere die Bimodalität kann aufgrund der vorliegenden Datenverteilung der titrierbaren Säuren anhand der Weinfarbe, grafisch dargestellt in Abbildung 5.30, erklärt werden. Die Rotweine weisen ein kleineres Wertespektrum mit einer einhergehend hohen Konzentration bei 5 g/l auf, während sich die Weiß- und Roséweine einerseits primär im Bereich um 5.5 g/l, bei gleichzeitig hoher Streuung, befinden und somit weniger konzentriert vorliegen. Aufgrund dieser ungleichen Datenverteilung resultiert das unterschiedliche Schätzverhalten innerhalb der Weinfarben, sowie der Neukalibrierung und die daraus resultierende Bimodalität, wie eingangs in diesem Kapitel beschrieben.

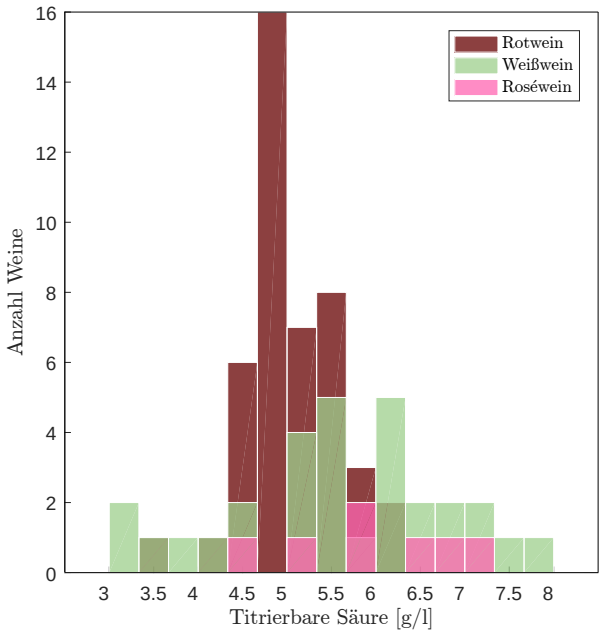

Abbildung 5.30: Datensituation der titrierbaren Säuren, mit drei übereinander gelegten Histogrammen zur Visualisierung der unterschiedlichen Datenverteilung nach Weinfarbe.

Reproduzierbarkeit

In einem ersten Überblick der Reproduzierbarkeit des Gesamtmodells in Abbildung 5.31 zeigt sich eine leicht bessere Wiederholbarkeit der Weißweinschätzungen mit einer gemittelten Standardabweichung in Höhe von 0.0246 g/l. Zudem kann lediglich ein Ausreißer bei Betrachtung der vereinten Daten beobachtet werden, welche bei geringen 0.055 g/l liegt.

Zieht man zusätzlich die maximale Abweichung als Gütemaß hinzu, so ergibt sich eine Auslenkung von höchstens 0.1 g/l bei Rundung auf die vorliegende Datengenauigkeit. Dies gilt über alle Weinfarben hinweg, sowie für die Betrachtung aller Weinproben gemeinsam. Mit einem Median der maximalen Differenz zweier

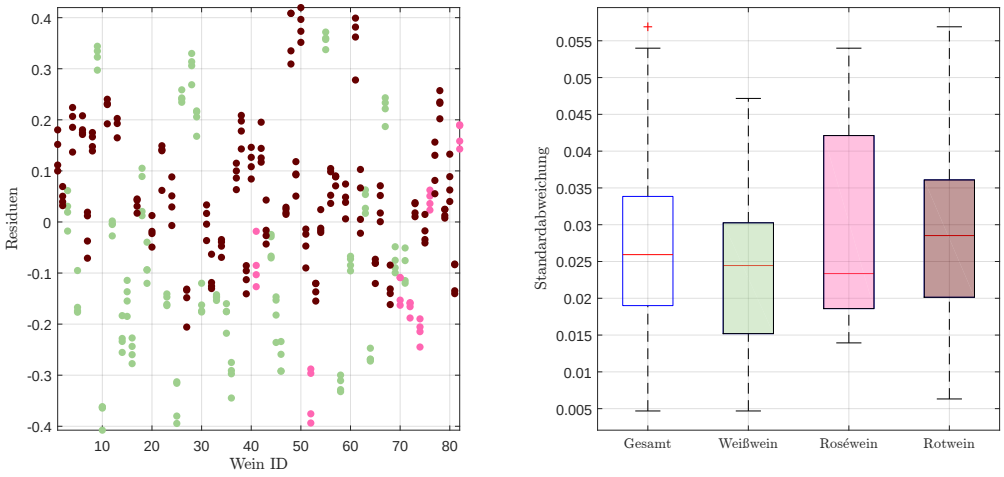

Abbildung 5.31: Residuen in g/l der Einzelspektren pro Wein, ohne Kunstweine, zur Schätzung der titrierbaren Säuren im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

Messungen, gegeben einer Wein ID, von mehr als 0.05 g/l folgt, dass bei mehr als der Hälfte der Weine ein Messunterschied in Höhe der maximalen Abweichung, d.h. 0.1 g/l, festgestellt werden kann.

Für die Submodelle der Rotweine gilt gleichermaßen, dass die Reproduzierbarkeit verbessert werden kann, auch wenn die maximale Abweichung (gerundet) bei 0.1 g/l bleibt, so reduziert sich diese durch die Neukalibrierung auf 0.09 g/l, während sie im rotweinspezifischen Modell auf circa die Hälfte (0.06 g/l) der ursprünglichen maximalen Differenz sinkt. Der entscheidende Vorteil zeigt sich im Vergleich der Mediane der maximalen Abweichung. So reduzieren sich beide auf unter 0.5 g/l, wobei somit bei mehr als der Hälfte der Rotweinproben keine Unterschiede innerhalb der Messungen festgestellt werden können (Median der maximalen Abweichung bei Neukalibrierung: 0.04 g/l, beim spezifischen Modell: 0.03 g/l).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0246	0.0244	0.0929	0.0540	0.0536	0.0217
Roséwein	0.0295	0.0234	0.1084	0.0629	0.0543	0.0314
Rotwein	0.0285	0.0285	0.1249	0.0632	0.0665	0.0264
Gesamt	0.0271	0.0259	0.1249	0.0597	0.0579	0.0253

Tabelle 5.14: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für titrierbare Säuren im PLS-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

Vergleicht man die Datensätze mit den hier ermittelten Werten, so zeigt sich in allen drei Modellen ein ähnliches Verhalten. E22 und E25 weisen jeweils die besten Kennzahlen für die Residuen auf, während das Spektrometer E25 die vergleichsweise besten Reproduzierbarkeiten liefert, auch wenn alle diese Kennzahlen, mit Ausnahme der maximalen Abweichung in den Datensätzen E24 und V70(2016), vergleichbar sind.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.1767	0.28	−0.35	0.37	0.0326	0.0303	0.1396
E24	0.1998	0.29	−0.53	0.45	0.0363	0.0332	0.3017
E25	0.1788	0.25	−0.37	0.39	0.0271	0.0259	0.1249
V70(2015)	0.2190	0.29	−0.51	0.53	0.0464	0.0378	0.5706

Tabelle 5.15: Performance des entwickelten PLS-Modells für titrierbare Säuren, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

5.6 Weinsäure

Bei der Modellierung jener Säure mit dem höchsten Wertebereich, der Weinsäure, müssen unterschiedliche Aspekte betrachtet werden. So kann beispielsweise mit zwei latenten Variablen und Messstellen an 20 Wellenzahlen eine Standardabweichung der Residuen in Höhe von 0.3416 g/l bei einer maximalen Abweichung der Einzelmessungen in Höhe von $0.06 \approx 0.1$ g/l erreicht werden. Da die Intention besteht, ein Modell mit möglichst geringer Variabilität in den Residuen zu entwickeln (bei einer plausiblen Reproduzierbarkeit), so kann dies auf 0.2338 g/l, bei einer Wiederholbarkeit, wiederum im Sinne der höchsten Abweichung der Einzelschätzungen, von immerhin $0.18 \approx 0.2$ g/l, auf Kosten der Modellparameter, erreicht werden: hierfür werden 7 latente Variablen und 100 Messpunkte benötigt. Bei beiden kann über die Heuristik der Selektion mit der doppelten Kreuzvalidierung argumentiert werden. Aufgrund des Minimalitätsprinzip im Sinne von Ockhams Rasiermesser ist allerdings ersteres vorzuziehen, weshalb das Modell mit zwei latenten Variablen präsentiert wird und sich lediglich der grafische Vergleich der Residuen in Abbildung 5.34 wiederfindet.

Das beschriebene beste Modell zur Bestimmung der Konzentration von Weinsäure verwendet zwei latente Variablen und 20 Wellenzahlen, aufgeteilt in vier gleich große Blöcke, beidseitig in unmittelbarer Nähe der ausgeschlossenen H_2O-Bande angesiedelt. Alle selektierten Wellenzahlen liegen zwischen $1\,411\,\mathrm{cm}^{-1}$ und $1\,804\,\mathrm{cm}^{-1}$, wie in Abbildung 5.32 mit der zweiten Savitzky-Golay Ableitung dargestellt.

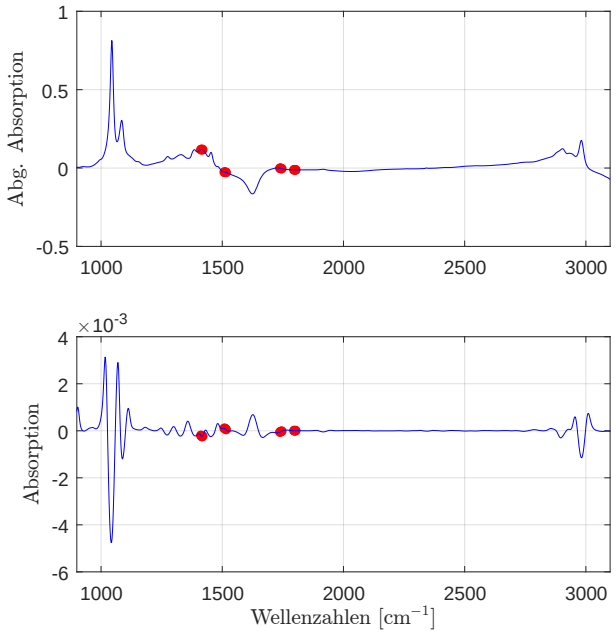

Abbildung 5.32: Die selektierten Wellenzahlen im PLS-Modell für Weinsäure mit der zweiten Savitzky-Golay Ableitung.

Die Wahl der latenten Variablen rechtfertigt sich durch Abbildung 5.33 (li.). Grundsätzlich kann für eine latente Variablenzahl in Höhe von 4 argumentiert werden, da es zu einem Abfall der SEP Kurve zwischen 2 und 4 kommt, wobei diese Verbesserungen stark abhängig von der Wahl der jeweiligen Partitionen zu sein scheint. Aus diesem Grund, sowie der Unauffälligkeit des dazu korrespondierenden Residuenplots scheint die Wahl von 2 latenten Variablen plausibel. In ebendieser Grafik können für fast alle Weine ähnliche Variabilitäten beobachtet werden. Zusätzlich gibt es jeweils zwei Ausreißer bezüglich der vertikalen Position. Dies kann auf die Datenbeschaffenheit zurückgeführt werden, da es sich hier wiederum um eine Art Extrapolation der Daten des Wertebereiches $[1.0, 2.5]$ g/l handelt und der Methodik der doppelten Kreuzvalidierung geschuldet ist, wie der klassische Residuenplot in Abbildung 5.34 (li.) bestätigt.

Betrachtet man die Residuen des hieraus resultierenden Modells (Abbildung 5.34 (li.)), so können keine Auffälligkeiten identifiziert werden. Zunächst scheint eine mögliche konische Form der Residuen vorzuliegen, wobei dies lediglich durch die Datenpunkte zwischen 1.00 g/l und 1.25 g/l, sowie des Rotweinausreißers suggeriert wird. Selbst jener Ausreißer der doppelten Kreuzvalidierung mit einer Weinsäurekonzentration von 0 g/l kann nicht durch einen außergewöhnlich hohen Schätzfehler identifiziert werden. In der rechten Grafik sind die Residuen des eingangs beschriebenen Modells mit 7 Variablen dargestellt. Man beachte hierbei die unterschiedliche Skalierung der vertikalen Achse. Auch hier kann kein auffälliges Verhalten bei geringerer Variabilität beobachtet werden. Aufgrund des Minimalitätsprinzipes wird allerdings ersteres Modell präferiert.

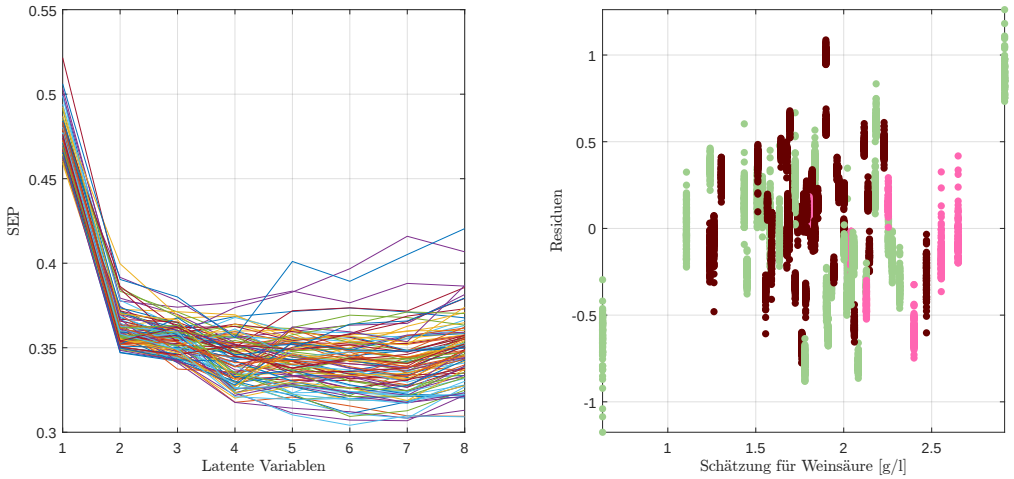

Abbildung 5.33: Verlauf des SEP in Abhängigkeit von latenten Variablen (li.) und die dazu korrespondierenden Weinresiduen der doppelten Kreuzvalidierung [g/l] (re.) für Weinsäure im PLS-Modell.

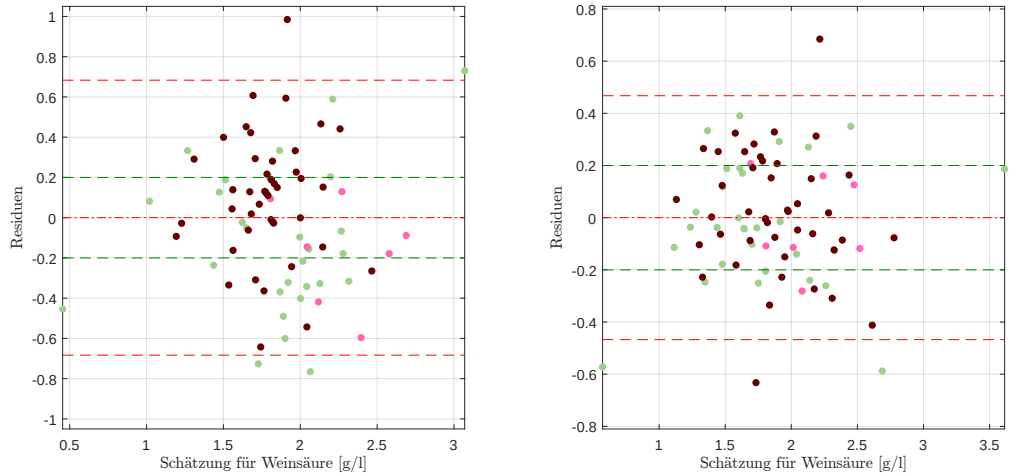

Abbildung 5.34: Residuenplot des PLS-Modells für Weinsäure mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode mit linearem Kernel (li.) und Gaußkernel mit Parameter $\sigma^2 = 1$ (re.).

Reproduzierbarkeit

Die Reproduzierbarkeitsplots zeigen, insbesondere bei der nicht nach Farbe differenzierten Betrachtung des Boxplots in Abbildung 5.35 (re.) außergewöhnlich viele Ausreißer, wobei diese gemittelte Standardabweichung nicht auf eine schlechte Modellgüte schließen lässt, insbesondere wenn man die Höhe dieser Kennzahlen in ein Verhältnis zur vorliegenden Datengenauigkeit von $10^{-1}\,\mathrm{g/l}$ setzt.

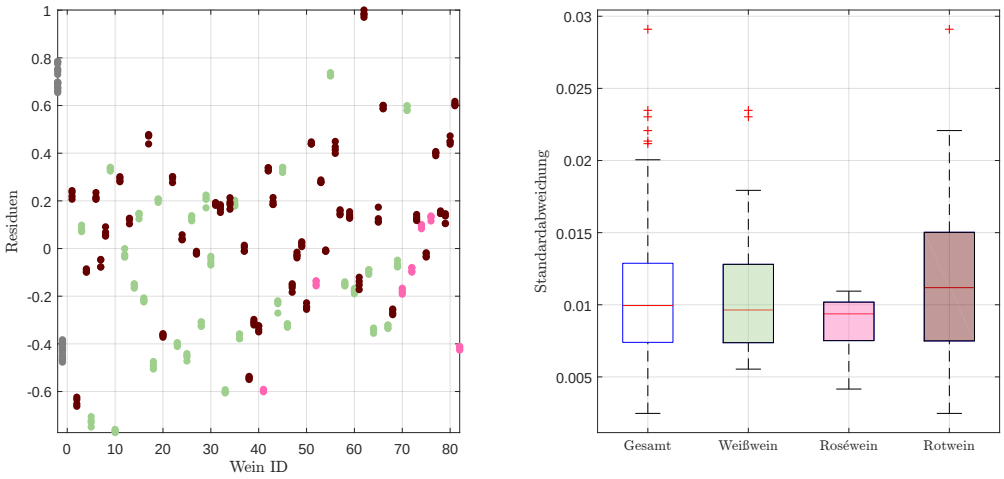

Abbildung 5.35: Residuen in g/l der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung der Weinsäurekonzentration im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

Ebenso gute Werte zeigen sich selbst bei der maximalen Abweichung, welche für alle Weine gemeinsam unter $0.063 \approx 0.1\,\mathrm{g/l}$ liegt. Betrachtet man ebendiese Genauigkeit, so resultiert eine maximale Abweichung von $0.1\,\mathrm{g/l}$ für lediglich 5 Weine, wohingegen für die restlichen 76 Einzelspektren für jeden Wein dieselbe Weinsäurekonzentration, gerundet auf die Genauigkeit der gegebenen Daten, berechnet wird.

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0108	0.0096	0.0533	0.0240	0.0209	0.0107
Roséwein	0.0085	0.0094	0.0248	0.0183	0.0187	0.0048
Rotwein	0.0118	0.0112	0.0629	0.0261	0.0242	0.0126
Gesamt	0.0111	0.0099	0.0629	0.0246	0.0219	0.0116

Tabelle 5.16: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Weinsäure im PLS-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

Ebenfalls bei der Betrachtung der unterschiedlichen Datensätze ergeben sich für sämtliche Kennzahlen, mit Ausnahme einer womöglichen Fehlmessung im Datensatz E24, beinahe idente Kennzahlen und das gefundene Modell kann als äußerst stabil und robust angesehen werden.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.3440	0.45	−0.79	0.96	0.0112	0.0107	0.0578
E24	0.3423	0.47	−0.74	0.96	0.0153	0.0128	0.2620
E25	0.3416	0.44	−0.77	0.98	0.0111	0.0099	0.0629
V70(2016)	0.3411	0.44	−0.74	0.95	0.0114	0.0099	0.0597

Tabelle 5.17: Performance des entwickelten PLS-Modells für Weinsäure, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

5.7 L-Äpfelsäure

Neben der Weinsäure weist die Äpfelsäure den zweit größten Wertebereich aller untersuchten Säuren auf, auch wenn mit 40 von 81 Weinen beinahe die Hälfte des Datensatzes eine L-Äpfelsäurekonzentration in Höhe von 0 g/l aufweist. Dies bedeutet, dass im Residuenplot die Hälfte der Residuen aufgrund deren Berechnung mit $r = y - \hat{y}$ auf einer fallenden Gerade liegen. Hierauf wird beispielsweise in nachfolgendem Abschnitt 5.9 etwas näher eingegangen.

Für eine Schätzung der L-Äpfelsäurekonzentration verwendet das PLS-Modell 40 Wellenzahlen, aufgeteilt in 4×10 Blöcke, wie Abbildung 5.36 (li.) zeigt. Diese finden sich zur Hälfte in der Fingerprintregion bei 1 120 cm^{-1}, sowie 1 200 cm^{-1}, wobei es sich hier um zwei aufeinanderfolgende, allerdings nicht direkt angrenzende Bereiche handelt, sowie ein Block, welcher unmittelbar auf die H_2O-Bande bei 1 780 cm^{-1} folgt, beziehungsweise ein Intervall, welches vor der CO_2-Bande bei 2 180 cm^{-1} liegt.

Hierbei wurde die Savitzky-Golay Ableitung zweiter Ordnung und 4 latente Variablen verwendet. Der Gaußkernel mit Parameter $\sigma^2 = 10^{-1}$ erzielt die besten Ergebnisse und dient somit als Nichtlinearisierung des Kernels. Abgesehen von der Tatsache, dass sich der Großteil der Rotweindaten auf der eingangs angesprochenen Gerade wiederfindet, können keine Auffälligkeiten im Residuenplot 5.36 (re.) beobachtet werden. Weder ein Trend, noch ein Offset oder Ausreißer sind ersichtlich.

Für die mehrheitlich roten Weine ohne L-Äpfelsäure gilt, dass diese trotz des geringen Wertebereichs mit einer Standardabweichung von 0.4459 g/l beinahe doppelt so stark streuen wie die Weiß-/Roséweine mit 0.2369 g/l, für welche vergleichsweise viele unterschiedliche Referenzwerte vorliegen. Zusammengefasst

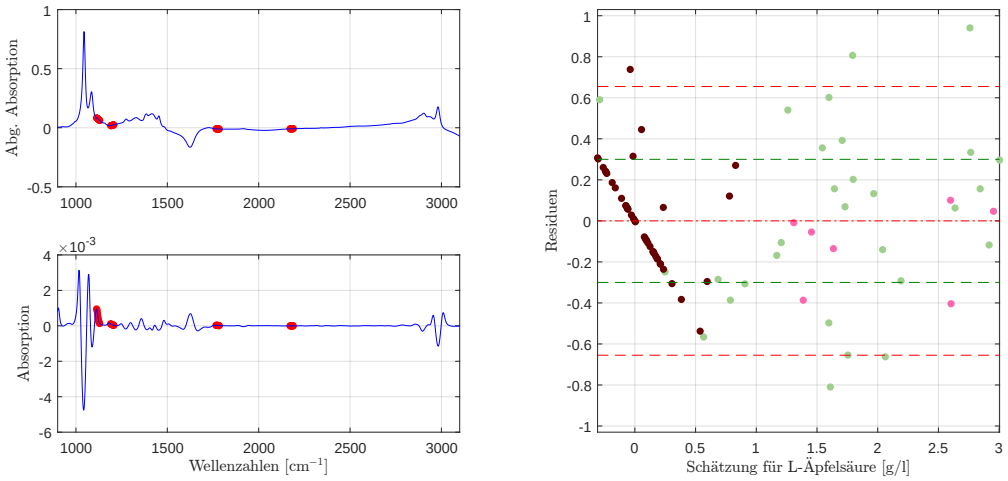

Abbildung 5.36: Die selektierten Wellenzahlen im PLS-Modell für L-Äpfelsäure mit der zweiten Savitzky-Golay Ableitung (li.) und der dazugehörende Residuenplot mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (re.).

ergibt sich eine Gesamtvariabilität in Höhe von 0.3275 g/l. Die Residuen befinden sich in einem Intervall von $[-0.81, 0.94]$ g/l, wobei der interquartile Bereich lediglich eine Länge von 0.37 g/l misst. Für die Abweichung der Schätzungen zu den tatsächlichen Werten ergibt sich ein MSE von $0.1059(\text{ g/l})^2$. Hierbei fallen allerdings 20 Schätzungen negativ aus und müssen für realitätsnahe Anwendungen auf 0 g/l aufgerundet werden. Hierunter fallen allerdings nicht ausschließlich Rotweine mit einer L-Äpfelsäurekonzentration von 0 g/l, sondern auch drei Weine, bei welchen eine Konzentration von 0.3 g/l (je ein Rot- und Weißwein) bzw. 0.7 g/l (Rotwein) vorliegt.

Der Verlauf des SEP in Abbildung 5.37 begründet die Variablenanzahl. Aus dem, aus der doppelten Kreuzvalidierung resultierenden, Residuenplot kann, neben des linearen Verlaufes der Residuen mit L-Äpfelsäurekonzentration von 0 g/l eine teilweise breite Streuung für Weißweine beobachtet werden, während die Roséweine mit dem Modell eine vergleichsweise geringe Streubreite aufweisen. Die teilweise erhöhte Variabilität für größere Konzentrationen kann wiederum durch die große Punktmasse um 0 g/l begründet werden.

Aufgrund der vorliegenden Daten kann es an dieser Stelle als unpassend erachtet werden, ein Modell eigens für die Rotweine zu entwickeln, während ein weiß-/roséweinspezifisches Modell womöglich zu einer Verringerung der Residuen führen kann. In Abbildung 5.38 sind die Residuen aus dem Gesamtmodell der Neukalibrierung mit den Weißweinen gegenübergestellt und es zeigt sich, abgesehen von der Reduzierung des minimalen Bias kaum ein verändertes Verhalten, weshalb ein Weißweinmodell an dieser Stelle keine signifikante Verbesserung bringt.

Die Nichtlinearisierung bewirkt für die L-Äpfelsäure insbesondere eine Verbesse-

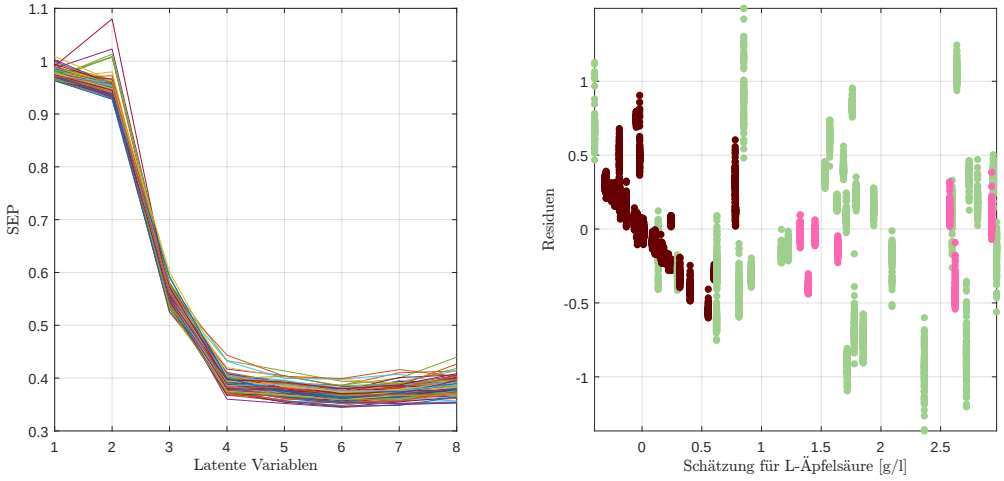

Abbildung 5.37: Verlauf des SEP in Abhängigkeit von latenten Variablen (li.) und die dazu korrespondierenden Weinresiduen der doppelten Kreuzvalidierung [g/l] (re.) für L-Äpfelsäure im PLS-Modell.

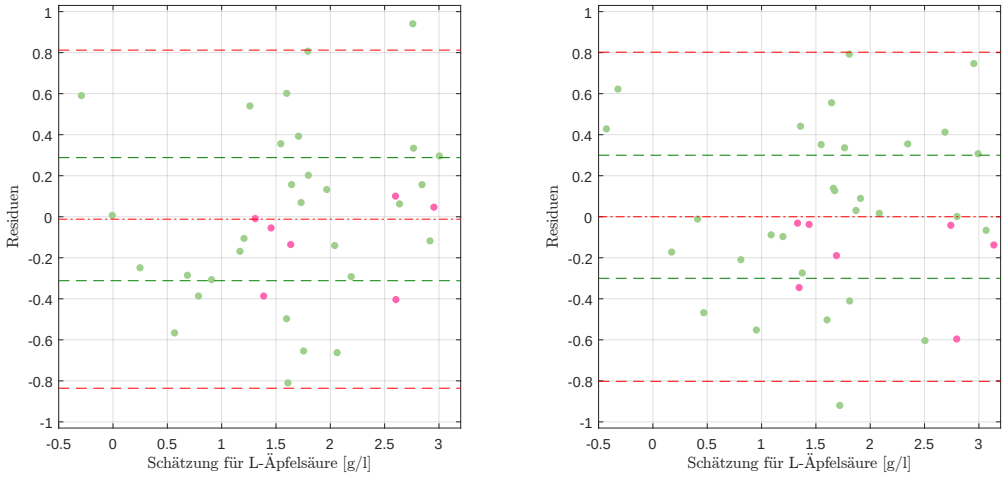

Abbildung 5.38: Extrahierte Weiß-/Roséweineresiduen des Residuenplots des PLS-Modells für L-Äpfelsäure mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.), sowie der Residuenplot des weiß-/roséweinspezifischen PLS-Modells durch Neukalibrierung mit den entsprechenden Kennzahlen (re.).

rung des Verhaltens der Weiß- und Roséweinresiduen. Während für beispielsweise den linearen Kernel und unterschiedliche Preprocessingmethoden die entwickelten Modelle stets einen teils starken Trend der Residuen erkennen lassen, muss trotz vergleichbarer Kennzahlen ein Modell mit Gaußkernel empfohlen werden.

Reproduzierbarkeit

Anhand der Residuen aller Einzelspektren in Abbildung 5.39 (li.) können nur geringe Messunterschiede innerhalb eines Weines beobachtet werden. Bei den Ausreißern in Grau handelt es sich um Kunstweine. Daraus kann gefolgert werden, dass das Modell Wechselwirkungen unterschiedlicher Inhaltsstoffe mitberücksichtigt, welche in ähnlicher Form nicht durch die Weinbestandteile der Kunstweine auftreten.

Dieser erste Eindruck wird durch die Boxplots der Standardabweichungen zur Übersicht über die Reproduzierbarkeit in Abbildung 5.39 (re.) bekräftigt. Hierbei kann lediglich ein Ausreißer identifiziert werden, während Rot- und Weißweine ein ähnliches Verhalten aufweisen, und es sind keine weiteren Auffälligkeiten, mit Ausnahme der überproportional guten Wiederholbarkeit der Roséweine, sichtbar.

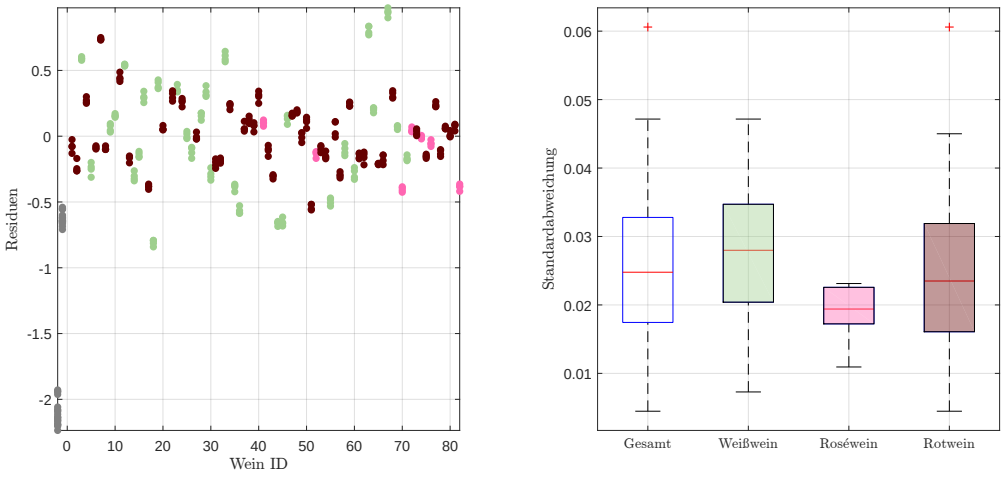

Abbildung 5.39: Residuen in g/l der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung der L-Äpfelsäurekonzentration im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

Betrachtet man wiederum die Genauigkeit der vorliegenden Daten von 10^{-1} g/l, so kann über alle Weinfarben hinweg eine beinahe idente Reproduzierbarkeit beobachtet werden und die maximale Abweichung beträgt im Gesamtmodell lediglich 0.1 g/l.

In Tabelle 5.19 werden wiederum die Modellkennzahlen des Gesamtmodells, angewendet auf die unterschiedlichen Datensätze, gegenübergestellt. Es zeigt sich

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0275	0.0280	0.1124	0.0619	0.0628	0.0229
Roséwein	0.0190	0.0194	0.0551	0.0439	0.0471	0.0102
Rotwein	0.0246	0.0235	0.1182	0.0547	0.0516	0.0250
Gesamt	0.0252	0.0248	0.1182	0.0564	0.0532	0.0236

Tabelle 5.18: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für L-Äpfelsäure im PLS-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

in den Kennzahlen der Residuen ein ähnliches Verhalten, ebenso wie bei der Reproduzierbarkeit, wohingegen der Datensatz E25 mit der geringsten maximalen Abweichung leichte Vorteile aufweist.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.3311	0.35	−0.87	0.96	0.0231	0.0210	0.1668
E24	0.3306	0.38	−0.75	0.94	0.0293	0.0255	0.3655
E25	0.3275	0.37	−0.81	0.94	0.0252	0.0248	0.1182
V70(2016)	0.3498	0.42	−0.85	1.03	0.0285	0.0279	0.1487

Tabelle 5.19: Performance des entwickelten PLS-Modells für L-Äpfelsäure, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

5.8 Milchsäure

Ein dritter wichtiger Bestandteil der Gesamtsäure bildet die Milchsäure. Während in den vorliegenden Datensätzen zusätzlich zwischen der L- und R-Milchsäure differenziert wird, fokusiert sich dieser Abschnitt, sowie die gesamte Arbeit, auf die Milchsäure in ihrer Gesamtheit.

Wendet man die vorgestellten Heuristiken und den PLS-Algorithmus auf die Milchsäurewerte an, so bringt die Nichtlinearisierung keine Verbesserung mit sich und die Wahl fällt wiederum auf einen linearen Kernel. Zusätzlich zur Savitzky-Golay Ableitung zweiter Ordnung werden 40 Wellenzahlen in vier gleichgroßen Blöcken wie in Abbildung 5.40 selektiert. Wiederum findet sich der Großteil der Wellenzahlen im sogenannten Fingerprintbereich, wobei die zusätzlich benötigte Information an die ausgeschlossene H_2O-Bande anschließt und sich auf den Wellenzahlbereich $[1\,745, 1\,728]\,\mathrm{cm}^{-1}$ verteilt. Nach dem Preprocessing repräsentieren sämtliche Intervallblöcke die Information von Peaks.

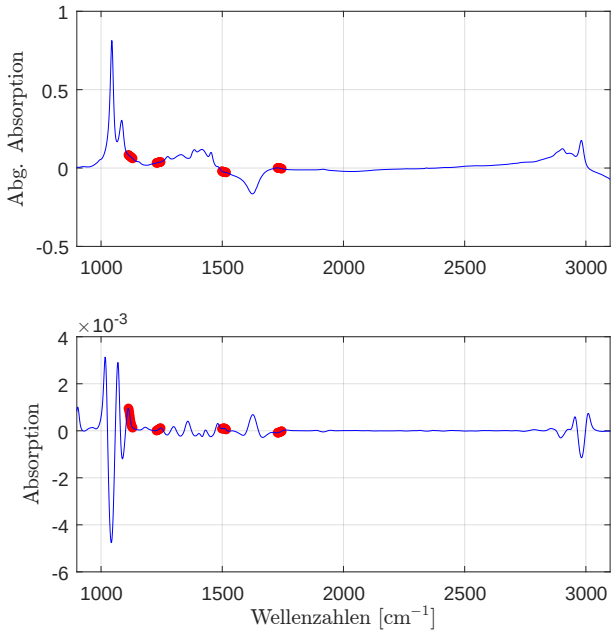

Abbildung 5.40: Die selektierten Wellenzahlen im PLS-Modell für Milchsäure mit der zweiten Savitzky-Golay Ableitung.

Die hierfür optimale Anzahl an latenten Variablen ergibt sich aus Abbildung 5.41 (li.). Während die Hinzunahme einer weiteren Variable von 3 auf 4 eine einigermaßen große Reduktion der Schwankung bringt, zeigen sich keine signifikanten Verbesserungen bei einer Hinzunahme einer weiteren latenten Variable. In den Residuen der doppelten Kreuzvalidierung in Abbildung 5.41 (re.) zeigen sich bezüglich der Schwankungen innerhalb eines Weines kaum Ausreißer - man beachte hierbei, dass sich die Residuen jenes Weißweines mit der vermeintlich höchsten Streuung aus mehreren Weinen zusammensetzen. In dieser Grafik darf den Weißweinen nicht fälschlicherweise ein fallender Trend unterstellt werden, da es sich hier um eine zusätzliche Häufung von Daten mit einer Konzentration von $0\,\mathrm{g/l}$ handelt und somit die hierzu korrespondierenden Residuen um eine gedachte Gerade mit Steigung $-1\,\mathrm{g/l}$ streuen, wie der klassische Residuenplot in Abbildung 5.42 bestätigt.

Betrachtet man nun, um die Güte des Modells festzustellen, den Residuenplot, so ist dieser bei einer Gesamtbetrachtung unauffällig und kein relevanter Trend bei einer Standardabweichung der Residuen von $0.2275\,\mathrm{g/l}$ sichtbar. Weiters ist kein Bias erkennbar, wohingegen der Median mit einem Wert von $-0.04\,\mathrm{g/l}$ leicht negativ ausfällt und die Residuen bewegen sich in einem Intervall $[-0.40, 0.73]\,\mathrm{g/l}$. Obwohl die Spannweite im Vergleich zum Wertebereich relativ groß zu sein scheint, ist der interquartile Bereich lediglich $0.29\,\mathrm{g/l}$. Von den 81 Weinen werden 44 über-, und $81 - 44 = 37$ unterschätzt, wohingegen insgesamt 8 Weinen ein negativer Säuregehalt zugeordnet wird. Dennoch weist dieses Modell mit einer erklärten Varianz von über 90 % einen beachtlichen Wert auf.

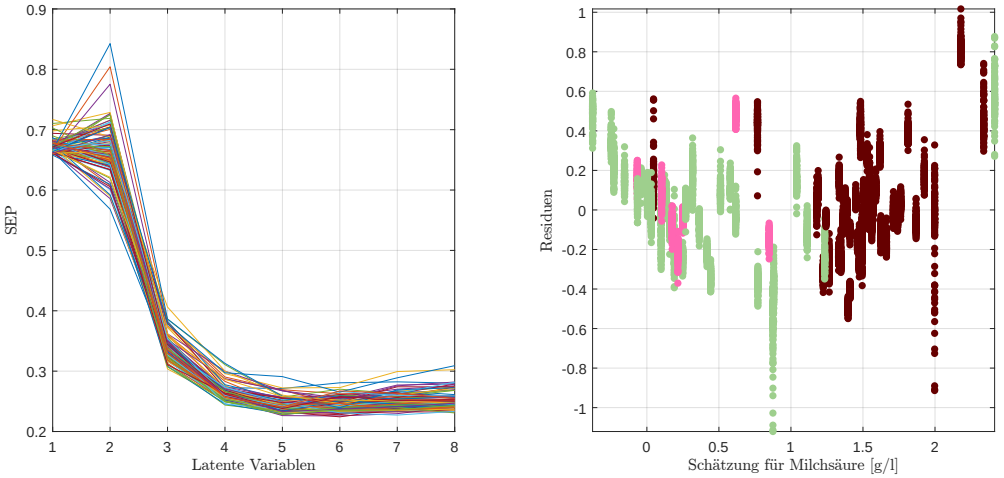

Abbildung 5.41: Verlauf des SEP in Abhängigkeit von latenten Variablen (li.) und die dazu korrespondierenden Weinresiduen der doppelten Kreuzvalidierung [g/l] (re.) für Milchsäure im PLS-Modell.

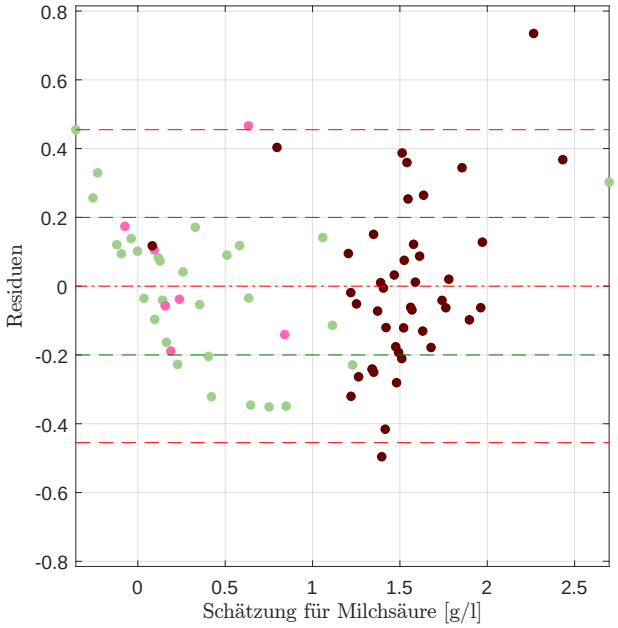

Abbildung 5.42: Residuenplot des PLS-Modells für Milchsäure mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode.

Modell für Rotweine

Auch wenn lediglich zwei Punkte die Wahrnehmung eines Trends verstärken beziehungsweise suggerieren, kann es durchaus Sinn machen, für die Weinfarben spezifische Modelle zu entwickeln, da einerseits die Wertebereiche für Rotweine und Weiß-/Roséweine sich nur mit wenigen Ausnahmen überlappen und andererseits die Rotweinresiduen aufgrund ihrer Darstellung womöglich nicht unabhängig sind.

Das beste für Rotweine gefundene Modell verwendet lediglich 20 Wellenzahlen mit einer Savitzky-Golay Ableitung ersten Grades und insgesamt 6 latenten Variablen. Im Gegensatz zu Abbildung 5.41 (li.) verläuft die SEP Kurve nicht derart harmonisch und steigt mit Hinzunahme von weiteren Variablen (ab 6), es kann jedoch ein eindeutiges Minimum für $M = 6$ latente Variablen beobachtet werden. Da sich die Modellgüte, insbesondere auf Kosten der Erhöhung von M nicht wesentlich verbessert, wird an dieser Stelle auf das Modell verzichtet und stattdessen eine Neukalibrierung des Gesamtmodells mit den Rotweinen durchgeführt. Die hieraus resultierenden Residuen, im Vergleich zu den dazu korrespondierenden Werten des Gesamtmodells, finden sich in Abbildung 5.43.

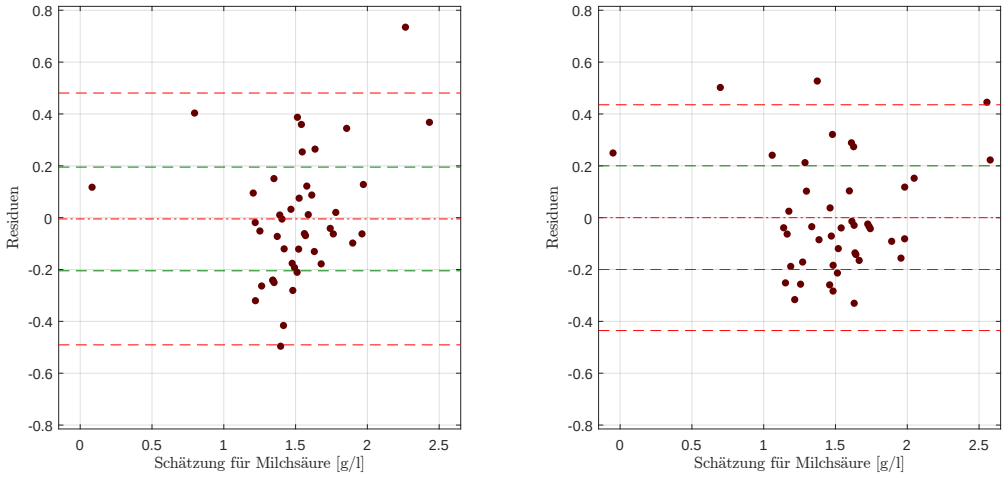

Abbildung 5.43: Residuenplot des PLS-Modells für Milchsäure mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode, mit Fokus auf die Rotweine des Gesamtmodells (li), sowie den Rotweinresiduen durch die Neukalibrierung der Rotweinen.

Die Variabilität reduziert sich nur leicht, allerdings kann der Ausreißer des Gesamtmodells in dieser Form nicht mehr beobachtet werden. Insgesamt gilt, dass die Residuen ein einigermaßen unauffälliges Verhalten zeigen, auch wenn insbesondere im neukalibrierten Modell die unterschätzten Weine etwas mehr streuen als jene mit negativen Residuen.

Reproduzierbarkeit

In Abbildung 5.44 (re.), der Boxplotübersicht über die Standardabweichungen der Einzelmessungen zeigt sich insbesondere für Rot- und Roséweine eine äußerst gute Reproduzierbarkeit, wohingegen die Weißweine stärker zu streuen scheinen. So können bei der Gesamtbetrachtung aller Weinfarben, sämtliche Ausreißer den Weißweinen zugeschrieben werden. Betrachtet man jedoch die Datengenauigkeit der vorliegenden Referenzwerte von 10^{-1} g/l, so können kaum unterschiedliche Schätzwerte für die Milchsäurekonzentration eines Weines, bei einer maximalen Diskrepanz innerhalb dieser Messungen von maximal 0.12 g/l, wie Tabelle 5.20 zeigt, beobachtet werden.

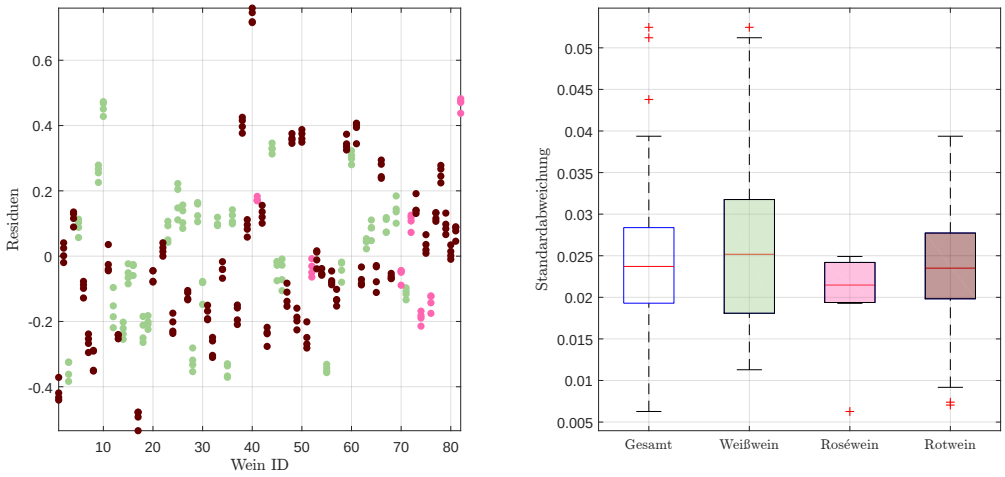

Abbildung 5.44: Residuen in g/l der Einzelspektren pro Wein, ohne Kunstweine, zur Schätzung der Milchsäurekonzentration im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

Bei einer Rundung auf die erste Nachkommastelle und somit der Genauigkeit der gegebenen Referenzwerte, ergibt sich eine maximale Abweichung innerhalb der Messungen eines Weines von 0.0 g/l für 33 Weine, während für die restlichen 48 Weine bei der Milchsäurekonzentration der maximale Unterschied bei 0.1 g/l liegt. Aus diesem Grund kann das hier entwickelte Modell als äußerst robust betrachtet werden.

Im Vergleich zu dem für E25 vorgestellten Modell können ähnliche Eigenschaften auch für die Messungen der anderen Geräte beobachtet werden. Die geringste Standardabweichung besitzen die Residuen, welche aus ebendiesem Datensatz resultieren. Für die Reproduzierbarkeit gibt es im Bereich der mittleren maximalen Abweichung kaum Unterschiede, während in den Datensätzen E24 und V70(2016) eine vergleichbar hohe maximale Abweichung beobachtet werden kann.

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0262	0.0252	0.1231	0.0581	0.0575	0.0236
Roséwein	0.0199	0.0215	0.0564	0.0446	0.0460	0.0144
Rotwein	0.0234	0.0235	0.0823	0.0515	0.0541	0.0159
Gesamt	0.0241	0.0237	0.1231	0.0533	0.0535	0.0193

Tabelle 5.20: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Milchsäure im PLS-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.2315	0.28	−0.51	0.73	0.0250	0.0221	0.1307
E24	0.2340	0.28	−0.47	0.75	0.0297	0.0284	0.2087
E25	0.2275	0.29	−0.50	0.73	0.0241	0.0237	0.1231
V70(2016)	0.2454	0.30	−0.53	0.81	0.0254	0.0242	0.1961

Tabelle 5.21: Performance des entwickelten PLS-Modells für Milchsäure, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

5.9 Flüchtige Säuren

Nachdem bereits drei Komponenten der Gesamtsäure modelliert sind, sollen in diesem Abschnitt die flüchtigen Säuren näher untersucht werden. Hierfür ergeben sich die in Abbildung 5.45 dargestellten Wellenzahlen für eine optimale Kalibration eines PLS-Modells. Hierbei handelt es sich um vier unterschiedliche Wellenzahlenblöcke mit je fünf Messpunkten. Diese Wahl der Prädiktoren erfolgt für einen linearen Kernel und insgesamt 3 latente Variablen unter Verwendung der ersten Savitzky-Golay Ableitung für das Preprocessing der einzelnen Spektren.

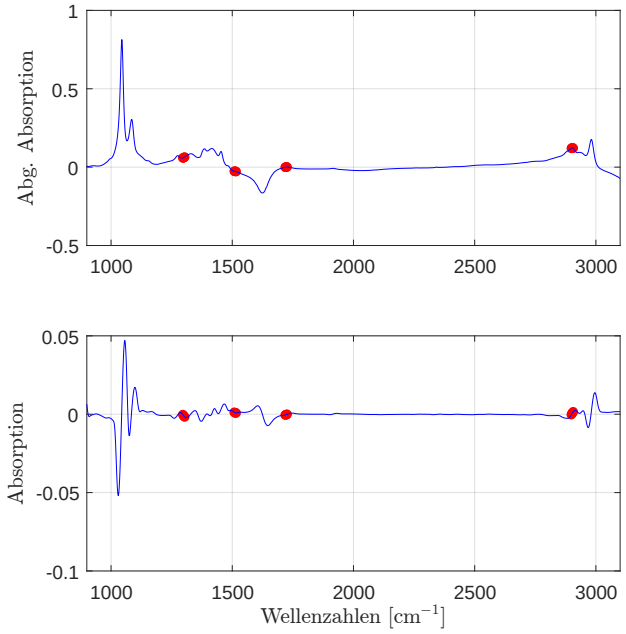

Abbildung 5.45: Die selektierten Wellenzahlen im PLS-Modell für die flüchtigen Säuren mit der ersten Savitzky-Golay Ableitung.

Während sich die Hälfte der Absorptionswerte bei Wellenzahlen im Fingerprintbereich befinden, werden zusätzlich zwei weitere, einmal im Bereich von 1 720 cm^{-1} und somit unmittelbar in der Nähe der ausgeschlossenen H_2O-Bande und kurz vor dem Peak der Bande um 2 900 cm^{-1} verwendet.

Mittels der doppelten Kreuzvalidierung wird wiederum die Anzahl der latenten Variablen bestimmt. Auch wenn in Abbildung 5.46 (li.) das Verwenden einer zusätzlichen latenten Variable argumentierbar ist, scheint es hier dennoch plausibel, 3 zu verwenden, sofern man den Trade-Off zwischen der Hinzunahme einer latenten Variable und der Abnahme der SEP Kurve betrachtet. Der Residuenplot, resultierend aus der doppelten Kreuzvalidierung, zeigt ebenfalls keine Auffälligkeiten. Ein eventuell leicht fallender Trend kann aufgrund der gegebenen Datensituation in Tabelle 5.22 erklärt werden. Auf diese Tatsache wird bei der Analyse des klassischen Residuenplots näher eingegangen. Anhand dieser Grafiken scheint das Modell adäquat zu sein.

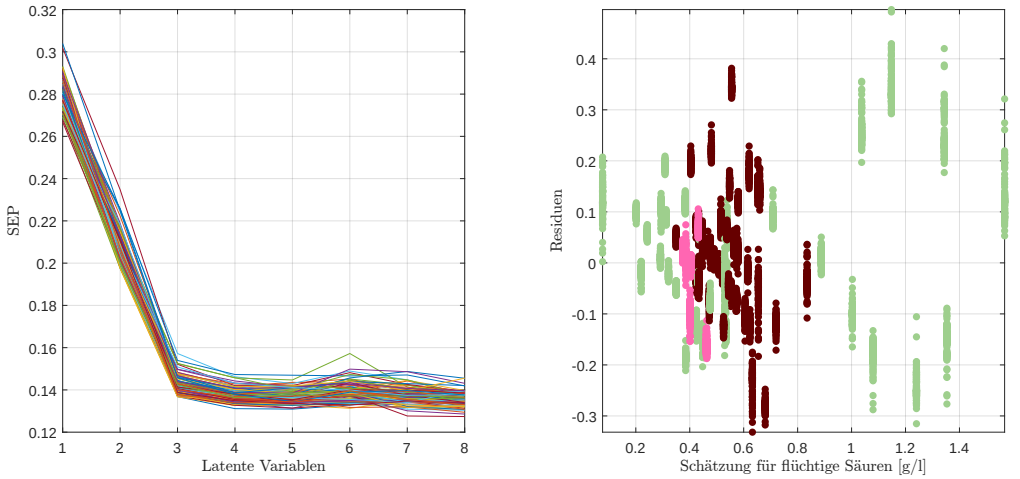

Abbildung 5.46: Verlauf des SEP in Abhängigkeit von latenten Variablen (li.) und die dazu korrespondierenden Weinresiduen der doppelten Kreuzvalidierung [g/l] (re.) für flüchtige Säuren im PLS-Modell.

Betrachtet man als nächstes den klassischen Residuenplot in Abbildung 5.47 (li.), so kann man einige parallel verlaufende Linien erkennen.

Die Gestalt und der scheinbar teilweise lineare Trend folgt direkt aus der Berechnung der Residuen. Für eine detailliertere Betrachtung seien hier r die Residuen, y die vorliegenden Referenzwerte und $\hat{y}$ die dazugehörenden Schätzungen des betrachteten Modells. Aus

$$r = y - \hat{y}$$

folgt unmittelbar, dass sämtliche Residuen von Objekten mit identer Konzentration y_0 sich als Geraden im Residuenplot wiederfinden. Betrachtet man die empirischen Auftrittswahrscheinlichkeiten der Werte y_0, so ergibt sich die Übersicht in Tabelle 5.22, Rubrik Datenverteilung. Hieraus kann man schließen, dass (mind.) 4 Geraden im Residuenplot ersichtlich sein müssen, jeweils mit Intercept $\beta_0 \in \{0.3, 0.4, 0.5, 0.6\}$ und Slope $\beta_1 = -1$. Beispielsweise liegt etwas mehr als ein Viertel aller hier berücksichtigten 80 Weine auf jener Gerade mit Intercept $\beta_0 = 0.5$. Betrachtet man einen modifizierten Residuenplot, indem man die Residuen gegen die tatsächlichen Konzentrationen aufträgt, so transformieren sich die Geraden mit Steigung -1 zu vertikalen Punkteansammlungen, wie Abbildung 5.47 (re.) zeigt. Aus diesem Grund können die Geraden des klassischen Residuenplot als Variabilität bei gegebenem Messwert interpretiert werden und zusätzlich die Schwankungen, bedingt auf einen spezifischen Säuregehalt, untersucht werden und ist wiederum in Tabelle 5.22 zu finden. Es treten kaum signifikante Biaswerte auf und die Standardabweichungen der Residuen, bedingt auf einen der Häufungspunkte bleibt in etwa konstant, insbesondere wenn man den korrespondierenden Residuenplot, sowie die Datengenauigkeit der vorliegenden Daten mit 10^{-1} g/l berücksichtigt.

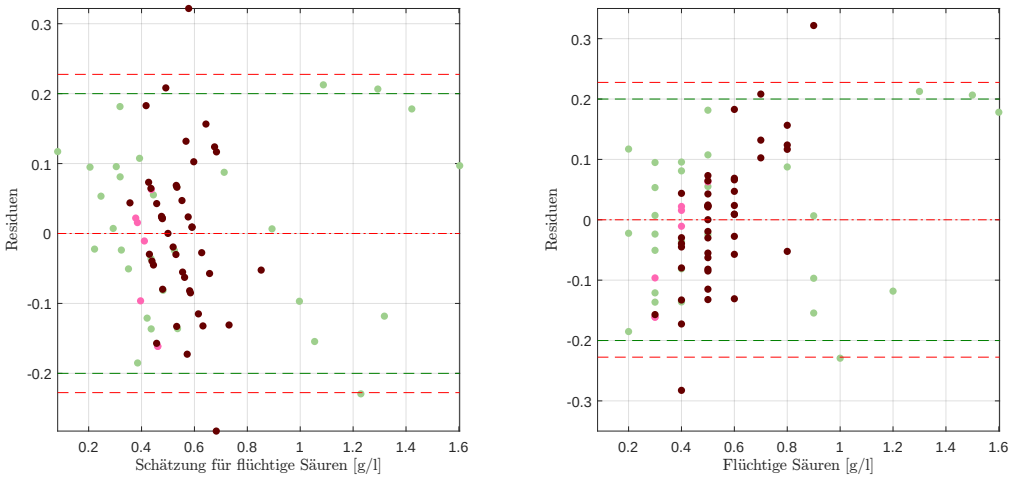

Abbildung 5.47: Residuenplot des PLS-Modells für flüchtige Säuren mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.), sowie dieselben Residuen gegen die tatsächlichen Werte zur Betrachtung der Variabilität, gegeben einer bestimmten Konzentration von flüchtigen Säuren.

	Datenverteilung		Residuen	
g/l	Anzahl	Anteil	Mittelwert	Std.
0.3	11	13.8 %	−0.07	0.0910
0.4	16	20.0 %	−0.05	0.0980
0.5	21	26.3 %	0.00	0.0769
0.6	11	13.8 %	0.02	0.0806
Andere	je ≤ 5	26.3 %		

Tabelle 5.22: Die größten Häufungspunkte im vorliegenden Datensatz des Jahres 2016, sowie das Verhalten der Residuen im PLS-Modell. Alle Werte in g/l.

Aufgrund der geringen unterschiedlichen (diskreten) Werte macht das Modellieren einzelner Weinfarben keinen Sinn und es wird an dieser Stelle nicht weiter darauf eingegangen, zumal auch keine signifikant besseren Modelle entwickelt werden können.

Reproduzierbarkeit

Betrachtet man in einer letzten Analyse dieser Kennzahl Abbildung 5.48 für die Reproduzierbarkeit, so ergibt sich eine relativ geringe Streuung, gegeben den wiederholten Messungen eines einzelnen Weines. Es können bei der Gesamtbetrachtung lediglich zwei Ausreißer mit einer vergleichsweise großen Schwankung beobachtet werden. Hierbei muss allerdings wiederum auf die Datengenauigkeit von 10^{-1} g/l hingewiesen werden.

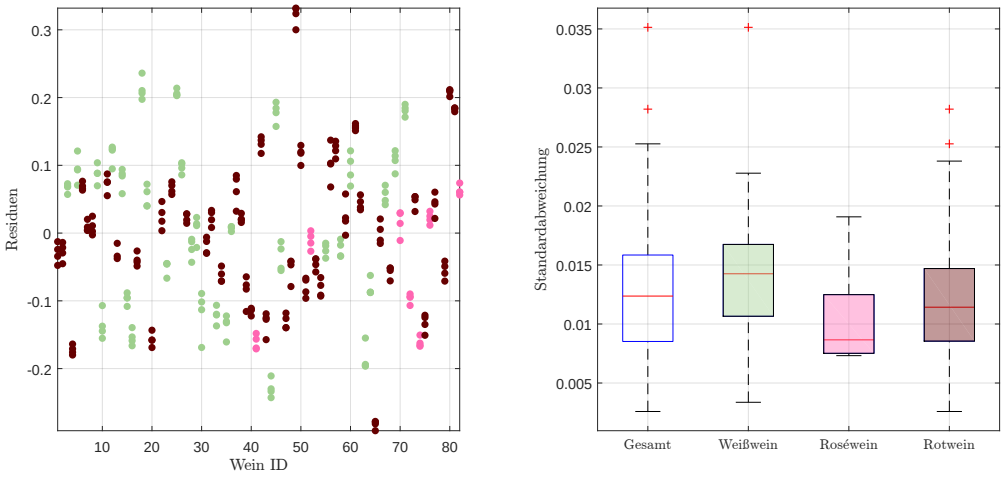

Abbildung 5.48: Residuen in g/l der Einzelspektren pro Wein, ohne Kunstweine, zur Schätzung der flüchtigen Säuren im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

Wie Tabelle 5.23 zeigt, weisen die Schätzungen eines jeden Weines eine maximale Abweichung von 0.1 g/l auf und das Modell kann als einigermaßen robust betrachtet werden.

Wendet man das Modell auf die jeweiligen Datensätze des Jahres 2016 an, so zeigen sich hier kaum Unterschiede. Das Modell und die Datenqualität scheint in allen verwendeten Datensätzen ähnlich zu sein, mit Ausnahme eines Ausreißers für das Spektrometer E24.

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0144	0.0143	0.0796	0.0325	0.0321	0.0143
Roséwein	0.0106	0.0087	0.0409	0.0236	0.0205	0.0090
Rotwein	0.0121	0.0114	0.0691	0.0272	0.0252	0.0135
Gesamt	0.0128	0.0124	0.0796	0.0289	0.0271	0.0136

Tabelle 5.23: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für flüchtige Säuren im PLS-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.1166	0.15	−0.31	0.35	0.0135	0.0124	0.0633
E24	0.1236	0.17	−0.29	0.36	0.0146	0.0131	0.2103
E25	0.1138	0.15	−0.28	0.32	0.0128	0.0124	0.0796
V70(2016)	0.1269	0.19	−0.29	0.34	0.0111	0.0106	0.0877

Tabelle 5.24: Performance des entwickelten PLS-Modells für flüchtige Säuren, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

5.10 Zitronensäure

Für die Zitronensäure, welche den geringsten Teil zur Gesamtsäure beiträgt, zeigt sich ein ähnliches Bild wie beispielsweise für die flüchtigen Säuren, da mit über 50 % der vorliegenden Weinproben mehr als die Hälfte aller Weine keine Zitronensäure beinhalten, bei weiteren 17 ($\approx 21\,\%$) Weinen lediglich eine Konzentration von 0.2 g/l feststellbar ist. Ein einzelner Wein enthält 1.2 g/l, während sich die restlichen 22 Weinproben auf die Werte $\{0.1, 0.3, 0.4, 0.5\}$ verteilen. Für den Roséwein mit 1.2 g/l Zitronensäure können die Werte wiederum als eine Art Extrapolation aufgrund der Verteilung der Responses betrachtet werden. Zusätzlich gilt aufgrund der Datenbeschaffenheit und insbesondere wegen dieses einen Wertes, dass keine Konvergenz im SEP Verlauf auftritt. Wählt man beispielsweise eine latente Variable bei 20 Wellenzahlen, wie in Abbildung 5.49 (li.), gepaart mit einem linearen Kernel, so resultieren die Residuen in einem Residuenplot mit einer Standardabweichung in Höhe von 0.1530 g/l. Abgesehen von dem zu erwartenden Ausreißer und den Geraden können keine weiteren Auffälligkeiten anhand von Abbildung 5.49 beobachtet werden.

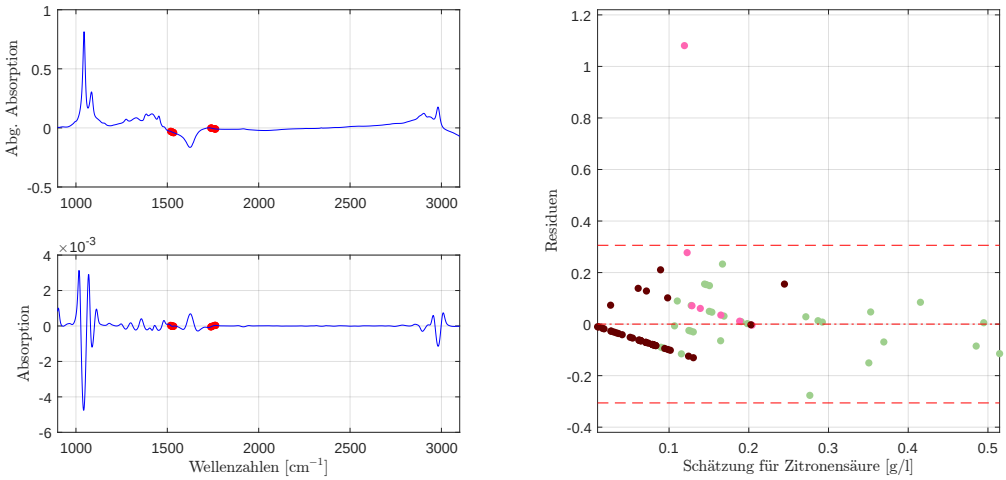

Abbildung 5.49: Die selektierten Wellenzahlen im PLS-Modell für Zitronensäure mit der zweiten Savitzky-Golay Ableitung (li.) und der dazugehörende Residuenplot mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (re.).

Um die tatsächliche Güte des Modells, insbesondere unter Breücksichtigung der Tatsache, dass sich sämtliche Zitronensäurekonzentrationen auf lediglich 7 Werte verteilen, zu untersuchen, liefert Tabelle 5.25, aufgeschlüsselt nach dem tatsächlichen Vorkommen von Zitronensäure, die wichtigsten Kennzahlen, wobei die Anzahl der Weinprobe pro Cluster mitberücksichtigt werden muss. Tatsächlich aussagekräftige Kennzahlen können lediglich für 0.0 g/l und 0.2 g/l abgelesen werden, wobei je ein Datenpunkt eine große Hebelwirkung auf die Standardabweichung besitzt. Aufgrund dessen, sowie der geringen Datenmengen in den anderen Konzentrati-

onskategorien kann von einer einigermaßen homogenen Streuung ausgegangen werden.

g/l	Anz.	Mw.	Med.	Std.	IQR	Min.	Max.
0.00	41	−0.07	−0.07	0.0440	0.03	−0.28	−0.01
0.10	6	−0.01	−0.03	0.0465	0.02	−0.06	0.07
0.20	17	0.04	0.05	0.0661	0.07	−0.15	0.14
0.30	8	0.08	0.09	0.0983	0.14	−0.07	0.21
0.40	6	0.09	0.10	0.1638	0.32	−0.11	0.28
0.50	2	0.04				0.01	0.08
1.20	1	1.08					

Tabelle 5.25: Verhalten der Residuen für gegebene Zitronensäurekonzentration im PLS-Modell. Alle Werte in g/l.

Betrachtet man die Genauigkeit der Daten, so resultieren 72 der 81 Weine in einem Residuum von 0.0 g/l oder $|0.1|$ g/l und das Modell gilt als einigermaßen genau. Der als Ausreißer zu erkennende Roséwein mit einem Zitronensäuregehalt von 1.20 g/l kann nicht als Fehlmessung aufgefasst werden, wie das auch von Anton Paar bestätigt wird. Aufgrund dessen muss dieser in der Modellbildung inkludiert und darf nicht vernachlässigt werden.

Reproduzierbarkeit

Sowohl für Rotweine als auch für Weißweine kann eine vergleichbare Reproduzierbarkeit festgestellt werden, auch wenn bei den Rotweinen zwei Ausreißer (bei den Weißweinen einer) auftreten. Bei den Roséweinen tritt eine vergleichsweise überproportional kleine Standardabweichung auf, insbesondere bei Betrachtung des interquartilen Bereichs, d.h. der Länge des Körpers des Boxplots, und der Tatsache, dass die Roséweine trotz der geringen Probenanzahl, einen aufgrund des signifikant hohen Wertes von 1.2 g/l den größten Wertebereich abdecken.

Betrachtet man zudem die maximale Abweichung in Höhe von 0.03 g/l, so bedeutet dies bei Betrachtung der vorliegenden Datengenauigkeit von einer Nachkommastelle eine idente Schätzung aller drei Weinfarben. Dennoch können die relevanten Kennzahlen der Reproduzierbarkeit, aufgelistet nach den Farben, Tabelle 5.26 entnommen werden.

Selbst bei der Übertragung des Modells auf die anderen Datensätze können exakt dieselben Schlüsse in allen Punkten gezogen werden, da alle Datensätze in den in Tabelle 5.27 dargestellten Kennzahlen, sowie in den graphisch visualisierten Residuen eine gewisse Homogenität zeigen.

Bemerkung 5.1. Jener Roséwein mit einer Zitronensäurekonzentration von 1.2 g/l hat aufgrund der allgemein guten Reproduzierbarkeit kaum einen Einfluss auf diese. Bei einer Neukalibrierung ohne diesen Wein ändert sich die Wiederholbarkeit nicht signifikant (maximale Abweichung von unter 0.03 g/l bleibt erhalten) und die

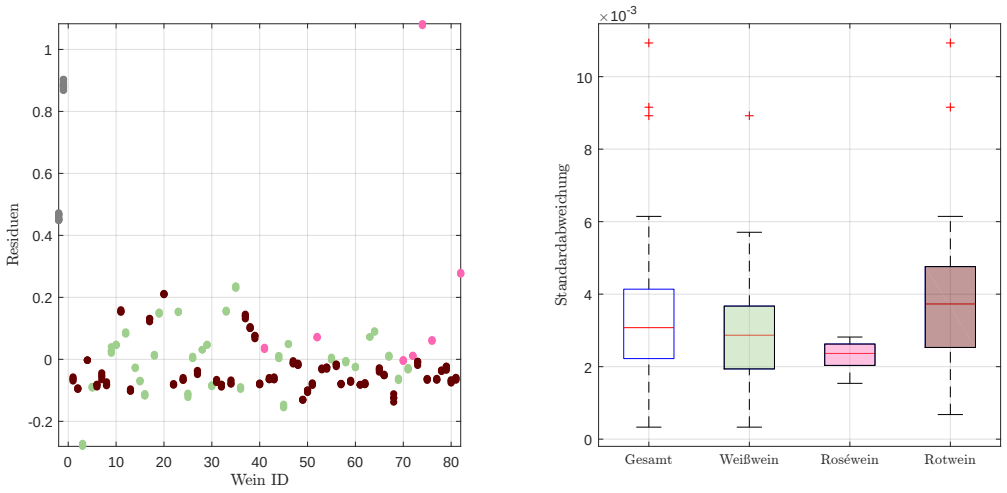

Abbildung 5.50: Residuen in g/l der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung der Zitronensäurekonzentration im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0030	0.0029	0.0213	0.0068	0.0062	0.0039
Roséwein	0.0023	0.0024	0.0069	0.0051	0.0052	0.0011
Rotwein	0.0039	0.0037	0.0268	0.0088	0.0080	0.0048
Gesamt	0.0034	0.0031	0.0268	0.0078	0.0069	0.0044

Tabelle 5.26: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Zitronensäure im PLS-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.1534	0.12	−0.27	1.09	0.0032	0.0028	0.0201
E24	0.1523	0.11	−0.28	1.08	0.0039	0.0035	0.0195
E25	0.1530	0.12	−0.28	1.08	0.0034	0.0031	0.0268
V70(2016)	0.1512	0.11	−0.29	1.07	0.0039	0.0030	0.0242

Tabelle 5.27: Performance des entwickelten PLS-Modells für Zitronensäure, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

Standardabweichung in den Residuen reduziert sich erwartungsgemäß auf knapp unter 0.10 g/l. Hierbei stimmen die Standardabweichungen der Neukalibrierung mit jenem des vorgestellten Modells, bei Nichtberücksichtigung des Roséweines zur Berechnung der Variabilität, beinahe überein. Dieses Verhalten kann in allen Datensätzen beobachtet werden und dieser Wein hat somit kaum bis keinen Einfluss auf die restlichen Schätzungen.

5.11 Glyzerin

Für die Schätzung des Zuckeralkohols Glyzerin mittels eines PLS-Modells genügt ein linearer Kernel, um die heuristisch besten Resultate zu erzielen. Insgesamt setzt sich das hier ermittelte Modell aus folgenden Parametern zusammen:

- 50 selektierte Wellenzahlen, aufgeteilt in zwei gleichgroße Blöcke. Einer dieser beiden zusammenhängenden Prädiktorenbereiche umschließt den maximalen Peak bei 1 045 cm^{-1}, wohingegen sich der zweite am anderen Ende des zugelassenen Wellenzahlbereiches befindet und verwendet die Information durch das Verhalten der Spektren zwischen den beiden Peaks bei 2 900 cm^{-1} und 2 980 cm^{-1} und sind in Abbildung 5.51 (li.) graphisch dargestellt.
- Ein linearer Kernel mit 5 latenten Variablen wird für die Kalibrierung verwendet.
- Die Spektren werden mit Savitzky-Golay zweifach abgeleitet.

Die, im Vergleich zu den restlichen analysierten Inhaltsstoffen, hohe Anzahl an latenten Variablen, sowie die Verwendung von vergleichsweise vielen Wellenzahlen deutet auf eine komplexe Modellierung der Glyzerinkonzentration hin, wie dies insbesondere im zweiten thematischen Schwerpunkt, den neuronalen Netzwerken, aufgezeigt wird.

Betrachtet man den aus der doppelten Kreuzvalidierung stammenden Residuenplot in Abbildung 5.52 (re.), so kann man ein allgemein gutes Verhalten der Residuen feststellen. Dass jener Weißwein mit hohem Glyzeringehalt eine erhöhte Streuung aufweist, lässt sich wiederum auf das Fehlen von Daten mit ähnlichen Glyzerinkonzentrationen zurückgeführen. Es sind lediglich zwei Proben auffällig, ein Weiß- und ein Rotwein, mit relativ großen Residuen, welche in dem klassischen Residuenplot in Abbildung 5.51 (re.) nicht als derart entartete Ausreißer identifizierbar und daher der Methodik zur Modellfindung geschuldet sind.

Bei dem hier dargestellten Verlauf des SEP ist ab einer Variablenanzahl von 5 ein gewisser Gap zwischen unterschiedlichen Verläufen festgestellbar und kann an dieser Stelle nur mit der Methodik und den drei auffallenden Weinproben erklärt werden. Dies wiederum hat großen Einfluss auf die Variabilität des SEP, wie der Boxplot in Abbildung 5.53 mit Ausreißern nach oben zeigt.

Betrachtet man den klassischen Residuenplot in Abbildung 5.51 (re.), so treten keinerlei Auffälligkeiten auf. Die zwei Weine mit einem betragsmäßig größeren Residuum als 1 g/l können beobachtet werden, wobei diese nicht in gleicher Weise entartet sind, wie jene Residuen, welche die Grafik der doppelten Kreuzvalidierung

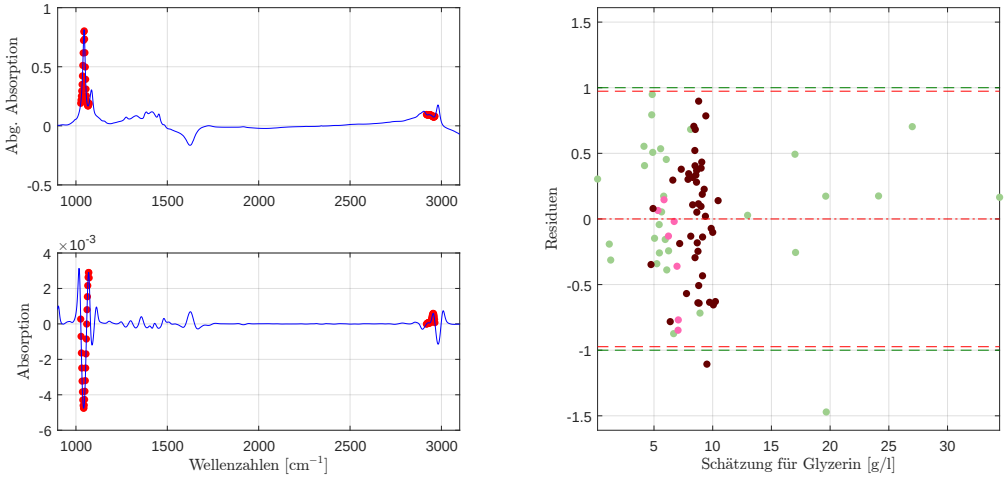

Abbildung 5.51: Die selektierten Wellenzahlen im PLS-Modell für Glyzerin mit der zweiten Savitzky-Golay Ableitung (li.) und der dazugehörende Residuenplot mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (re.).

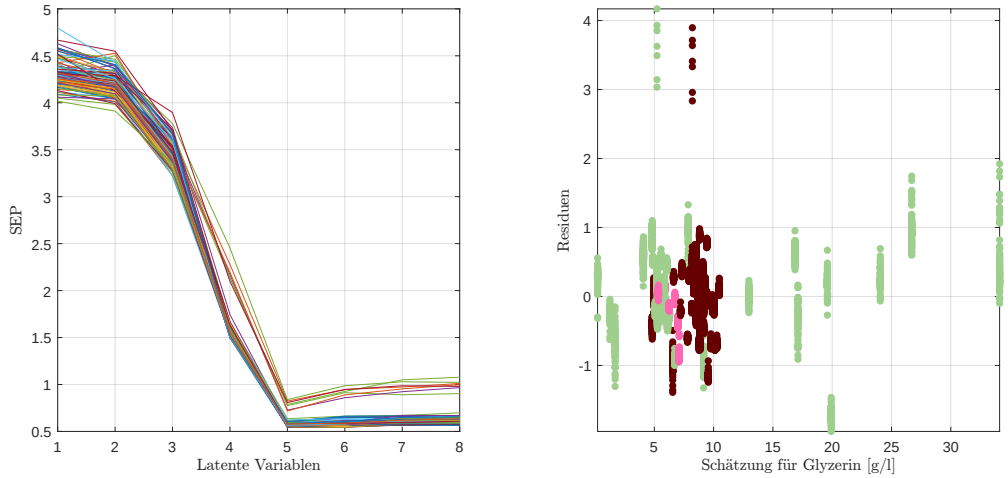

Abbildung 5.52: Verlauf des SEP in Abhängigkeit von latenten Variablen (li.) und die dazu korrespondierenden Weinresiduen der doppelten Kreuzvalidierung [g/l] (re.) für Glyzerin im PLS-Modell.

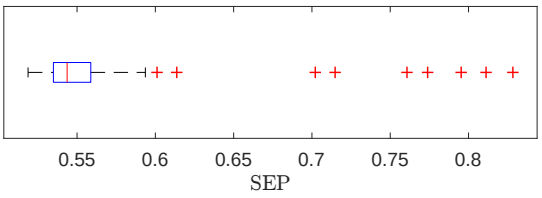

Abbildung 5.53: Übersicht über die SEP-Werte in Abbildung 5.52 (li.) für $M = 5$ latente Variablen.

und den SEP verzerren. Darüber hinaus scheint das gesamte Kalibrierungsmodell bei Verwendung aller Daten auch für große Glyzerinwerte plausibel. Hierbei muss man allerdings ebenfalls festhalten, dass das Anwenden des Modells für einen vergleichsweise hohen Wertebereich (z.B. $\geq 20\,\mathrm{g/l}$) wiederum einer Art Extrapolation gleicht. Dass die Residuen für Schätzungen ab $11\,\mathrm{g/l}$ für die Weißweine (mit Ausnahme des einen Ausreißers) tendenziell leicht unterschätzt werden, kann nicht auf einen leichten Trend oder Bias für hohe Werte zurückgeführt werden, sondern begründet sich lediglich durch die Art der Extrapolation. Betrachtet man die Residuenplots getrennt nach Farbe, so zeigen auch diese keine Auffälligkeiten bei einer Standardabweichung von $0.4867\,\mathrm{g/l}$, was in etwa der Genauigkeit der Referenzmethode entspricht. Auch wenn das Intervall, welches sämtliche Residuen einschließt, eine Länge von $2.4\,\mathrm{g/l}$ misst, so beläuft sich der interquartile Bereich lediglich auf $0.65\,\mathrm{g/l}$.

Da, wie aus dem Residuenplot ersichtlich, die Spannweiten von den tatsächlichen Glyzerinwerten zwischen Weiß- und Rotweinen stark differieren, wie aus der Übersicht 5.28 hervorgeht, kann es durchaus von Vorteil sein, diese getrennt zu modellieren.

	Anzahl	Minimum	Maximum
Roséwein	7	5.4	6.7
Rotwein	44	4.4	10.6
Weißwein	30	0.5	34.6

Tabelle 5.28: Übersicht über die Konzentrationen nach Weinfarbe, einschließlich dem jeweiligen Minimum und Maximum. Alle Werte in g/l.

Da bereits das Gesamtmodell die Glyzerinwerte gut beschreibt, wird an dieser Stelle auf ein weinfarbenspezifisches Modell verzichtet und lediglich die Neukalibrierung für Rot- bzw. Weiß-/Roséweine vorgestellt. Bei keinem dieser Modelle können Unregelmäßigkeiten beobachtet werden und die Standardabweichung der Residuen reduziert sich auf $0.4053\,\mathrm{g/l}$ für Rotweine, bzw. $0.5176\,\mathrm{g/l}$ für die restlichen Weine[9]. Trotz des ungleichen Wertebereiches zeigt sich hierbei kaum ein Mehrwert.

[9]Im Vergleich hierzu variieren die Rotweine im Gesamtmodell mit einer Standardabweichung von $0.4630\,\mathrm{g/l}$, während die Weiß- und Roséweine gemeinsam eine Standardabweichung von $0.5200\,\mathrm{g/l}$ besitzen.

Reproduzierbarkeit

Ein weiteres Maß für die Güte des gewählten Modells stellt die Reproduzierbarkeit dar. Hierbei kann eine relativ geringe Variabilität in den 4 Teilmessungen beobachtet werden, wie Abbildung 5.54 in Kombination mit Tabelle 5.29 bestätigt. Betrachtet man diese Grafiken, so können die beiden Ausreißer der Rotweine problemlos identifiziert werden und die Datenpunkte scheinen vergleichsweise außergewöhnlich weit voneinander entfernt zu liegen, deuten allerdings nicht auf womögliche Fehlmessungen hin. Weiters erkennt man an den maximalen Abweichungen, dass es bei allen Weinfarben jeweils Messungen gibt, welche von den restlichen verhältnismäßig stark abweichen. Dies zeigt wiederum die Notwendigkeit der Mittlung mehrerer Spektren zur tatsächlichen Schätzung einer Konzentration wie beispielsweise Glyzerin.

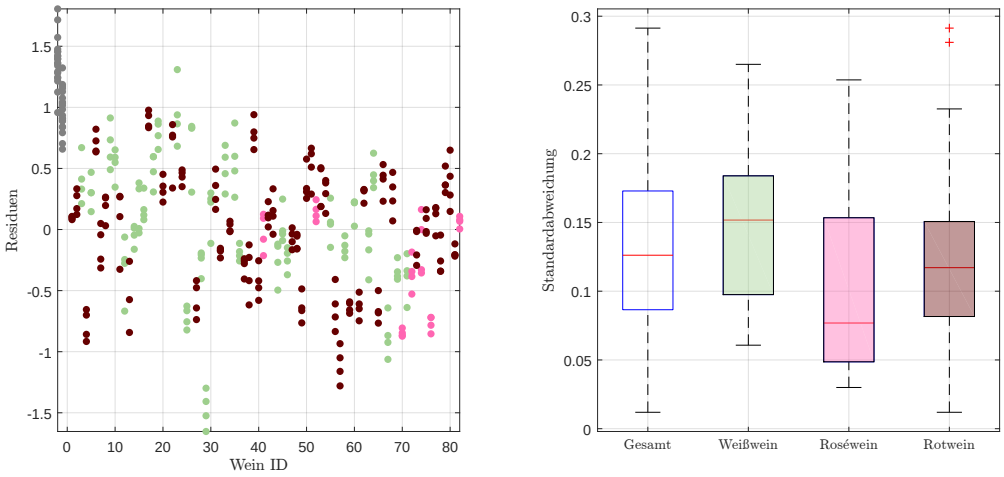

Abbildung 5.54: Residuen in g/l der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung von Glyzerin im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.1531	0.1517	0.6269	0.3472	0.3468	0.1358
Roséwein	0.1095	0.0769	0.5189	0.2407	0.1794	0.1641
Rotwein	0.1231	0.1171	0.5972	0.2728	0.2680	0.1348
Gesamt	0.1330	0.1262	0.6269	0.2976	0.2864	0.1415

Tabelle 5.29: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Glyzerin im PLS-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

Betrachtet man den Datensatzvergleich in Tabelle 5.30, so können die Glyzerin-

werte mit den Spektren des Spektrometers E25, mit welchen die Modellbildung durchgeführt wird, am besten modelliert werden. Dies zeigt sich sowohl bei den Kennzahlen der Residuen, als auch bei der Reproduzierbarkeit mit Ausnahme des Datensatzes V70(2016), welcher eine deutlich bessere Reproduzierbarkeit aufweist.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.4593	0.47	−1.42	1.13	0.1502	0.1268	1.0580
E24	0.5674	0.77	−1.42	1.23	0.1559	0.1443	1.0118
E25	0.4867	0.65	−1.47	0.95	0.1330	0.1262	0.6269
V70(2016)	0.5597	0.76	−1.51	1.03	0.0809	0.0807	0.3714

Tabelle 5.30: Performance des entwickelten PLS-Modells für Glyzerin, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

5.12 Dichte

Da die Dichte in engem Zusammenhang mit Extrakt (Korrelation von über 99 %) steht, wird an dieser Stelle nur kurz ein mögliches Modell zur Prädiktion von Dichtewerten vorgestellt. Im Großen und Ganzen haben sich hier zwei unterschiedliche Zusammensetzungen für die Kalibrierungsparameter herauskristallisiert, jeweils mit zwei latenten Variablen. Einerseits ein linearer Kernel mit lediglich 3×3 unterschiedlichen Wellenzahlen, wohingegen dem Gaußkernel mit Parameter $\sigma^2 = 0.1$ zwei Wellenzahlblöcke der Länge 50 zugeordnet werden. Hierbei handelt es sich um die Wellenzahlen symmetrisch um $1\,411\,\mathrm{cm}^{-1}$ bzw. $3\,014\,\mathrm{cm}^{-1}$.

Obwohl beide Modelle einen ähnlichen Verlauf des SEP aufweisen, kann letztgenanntem Modell eine höhere Stabilität und somit eine bessere Qualität unterstellt werden. Daher wird an dieser Stelle, trotz der hohen Anzahl an selektierten Wellenzahlen, für dieses Kalibrierungssetup plädiert und der Vergleich auf Seite 113 angestellt, der die Wahl rechtfertigt.

Betrachtet man den Residuenplot in Abbildung 5.55 (re.), so scheint dieser weitestgehend unauffällig. Es existieren lediglich zwei größere Ausreißer, je ein unter- und ein überschätzter Weißwein, und es werden alle Roséweine überschätzt. Mit Ausnahme des erwähnten Ausreißers liegen alle Abweichungen in einem sehr knappen Intervall von $[-2.5, 2.5] \cdot 10^{-3}\,\mathrm{g/ml}$ bei einem gegebenem Wertebereich von $[0.98, 1.10]\,\mathrm{g/ml}$.

Betrachtet man die Farbe als zusätzlichen Inputparameter, so kann durch eine Neukalibrierung keine Verbesserung in den Residuen beobachtet werden. Es findet lediglich keine Überschätzung aller Roséweine mehr statt, sondern eine teilweise leichte Unterschätzung tritt auf. Versucht man beispielsweise für die Rotweine

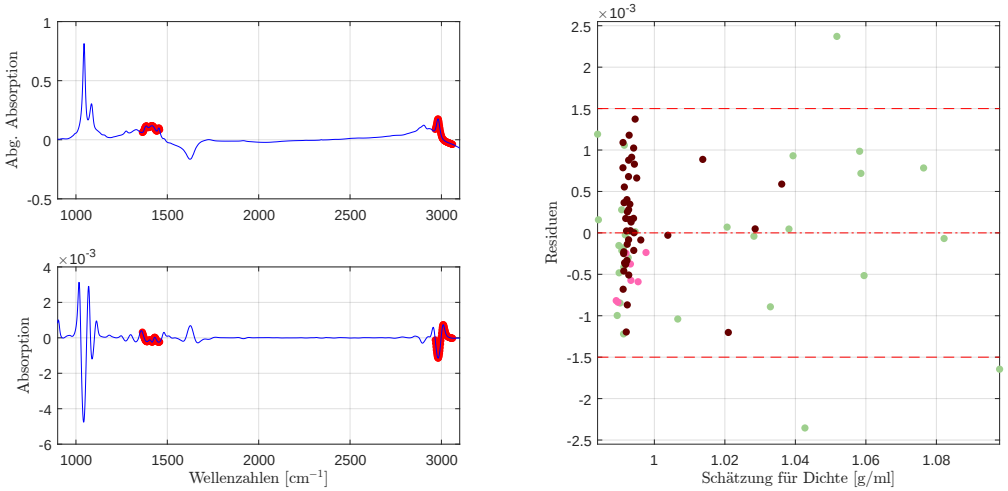

Abbildung 5.55: Die selektierten Wellenzahlen im PLS-Modell für die Dichte mit der zweiten Savitzky-Golay Ableitung (li.) und der dazugehörende Residuenplot mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) (re.).

ein eigenes Modell zu entwickeln, so kann, wiederum mit dem Hinweis auf das Minimalitätsprinzip kein alternatives, farbenspezifisches Modell entwickelt werden.

Reproduzierbarkeit

Für die Dichte weist das Modell eine einigermaßen gute Reproduzierbarkeit auf, auch wenn diese eine leicht schlechtere Performance für Rotweine als für die beiden anderen Weinklassen zeigt, wie der Boxplot in Abbildung 5.56 (re.) nahelegt. Sämtliche Ausreißer der totalen Reproduzierbarkeit resultieren aus den Abweichungen innerhalb der Rotweine, auch wenn sich diese Werte unter $2.2 \cdot 10^{-4}$ befinden.

Während für die Roséweine der Median über dem Mittelwert der Standardabweichungen liegt, gilt für die Rot- bzw. Weißweine das genaue Gegenteil, auch wenn die tatsächlichen Unterschiede nur minimal sind. Als auffallend kann die maximale Abweichung betrachtet werden, welche für Rotweine deutlich erhöhte Werte aufweist. Mit einem Wert von $5 \cdot 10^{-4}\,\mathrm{g/ml}$ kann dieser Rotwein nicht als alleiniger Ausreißer betrachtet werden, da weitere 8 Weine eine höhere maximale Abweichung als die höchste Differenz der Einzelmessungen pro Wein ID der anderen beiden Weinklassen, aufweisen.

Wiederum nur geringe Unterschiede zeigen sich beim Vergleich der unterschiedlichen Datensätze für die Dichte. Weder das Verhalten der klassischen Residuen noch jenes der Reproduzierbarkeit zeigt signifikante Unterschiede bezüglich der Datensätze, mit Ausnahme einer leicht erhöhten maximalen Auslenkung im Datensatz E24, wie Tabelle 5.32 zeigt.

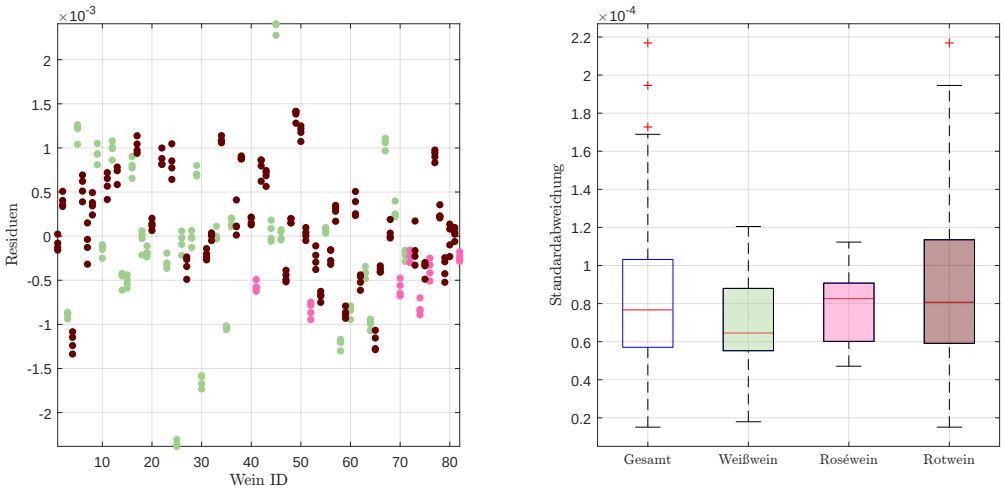

Abbildung 5.56: Residuen in g/ml der Einzelspektren pro Wein, ohne Kunstweine, zur Schätzung der Dichte im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0693	0.0645	0.2748	0.1574	0.1467	0.0605
Roséwein	0.0782	0.0826	0.2601	0.1767	0.1928	0.0517
Rotwein	0.0909	0.0807	0.5008	0.2044	0.1796	0.1056
Gesamt	0.0817	0.0767	0.5008	0.1844	0.1723	0.0893

Tabelle 5.31: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Dichte im PLS-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in 10^{-3} g/ml.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.7486	0.91	−2.30	2.48	0.0865	0.0840	0.4456
E24	0.7412	0.87	−2.38	2.49	0.0903	0.0779	0.8667
E25	0.7500	0.87	−2.35	2.37	0.0817	0.0767	0.5008
V70(2016)	0.7592	0.82	−2.30	2.63	0.0824	0.0716	0.4805

Tabelle 5.32: Performance des entwickelten PLS-Modells für Dichte, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in 10^{-3} g/ml.

Modellvergleich

Wie zu Beginn dieser Analyse diskutiert, kann alternativ zu diesem Modell mit einem Gaußkernel und 100 Wellenzahlen ein für den linearen Kernel optimale Kalibrierung bei gleichbleibender Variabilität in den Residuen gefunden werden, wie Tabelle 5.33 verdeutlicht. Es können zudem die betragsmäßig größten Residuen reduziert werden. Zieht man zeitgleich die Reproduzierbarkeit in Betracht, so zeigen sich enorme Vorteile bei dem vorgestellten Modell, welches den Gaußkernel verwendet. Betrachtet man exemplarisch die maximalen Auslenkungen, nehmen diese einen teils vierfach so hohen Wert an. Dies legt die Interpretation nahe, dass durch den Gaußkernel für die Dichte einzelne, kleinere Messfehler besser abgefangen werden können, was schlussendlich den Entscheidungsgrund für die vorgestellte Kalibrierung liefert.

		Residuen				Reproduzierbarkeit		
						Std.	Max. Abweichung	
	Anz.	Std.	IQR	Minimum	Maximum	Mittelwert	Maximum	Mittelwert
Dichte, linear K.	80	0.7980	1.0045	−1.8601	1.9289	0.2631	1.1869	0.5846
Weiß	30	1.0039	0.9702	−1.8601	1.9289	0.2510	1.1197	0.5559
Rosé	7	0.2206	0.4086	−0.4102	0.0718	0.2445	0.8975	0.5576
Rot	43	0.6999	1.1080	−1.5100	1.5596	0.2745	1.1869	0.6091
Dichte, Gauß K.	80	0.7500	0.8743	−2.3545	2.3712	0.0817	0.5008	0.1844
Weiß	30	0.9317	0.7941	−2.3545	2.3712	0.0693	0.2748	0.1574
Rosé	7	0.2492	0.4823	−0.8367	−0.2368	0.0782	0.2601	0.1767
Rot	43	0.6122	0.8883	−1.2017	1.3731	0.0909	0.5008	0.2044

Tabelle 5.33: Vergleich der wichtigsten Kennzahlen des linearen Modells und einem PLS-Modell mit Gaußkernel. Bei ähnlicher Variabilität der Residuen können Verbesserungen in der Wiederholbarkeit durch zusätzliche Wellenzahlen erzielt werden.

Da sich das Minimalitätsprinzip an Modellen gleicher Güte orientiert, steht diese Wahl nicht im Widerspruch zu dieser Orientierungshilfe.

5.13 pH-Wert

Da es sich bei dem pH-Wert um eine durch die Aktivität der Wasserstoffionen berechnete Kennzahl handelt und die basische bzw. saure Eigenschaft der Weinprobe beschreibt, kommen viele Wellenzahlbereiche in Frage, da grundsätzlich alle Inhaltsstoffe zum pH-Wert beitragen können.

Ein Modell für den pH-Wert selektiert 20 Wellenzahlen, aufgeteilt in 4 Blöcke mit je 5 Messpunkten, wie in Abbildung 5.57 visualisiert, wobei sich der Großteil im Fingerprintbereich unter $1\,500\,\mathrm{cm}^{-1}$ befindet, sowie ein kleiner Teilbereich unmittelbar nach der Erhöhung, welcher auf die H_2O-Bande folgt. Zusätzlich wurde ein Gaußkernel mit Parameter $\sigma^2 = 0.1$ für die Nichtlinearisierung, sowie die erste Savitzky-Golay Ableitung und 4 latente Variablen verwendet.

In Abbildung 5.58 (li.) kann bis zu einer Anzahl von 4 latenten Variablen eine beinahe lineare Abnahme des SEP, mit einer leichten Zunahme der Kurve ab dieser

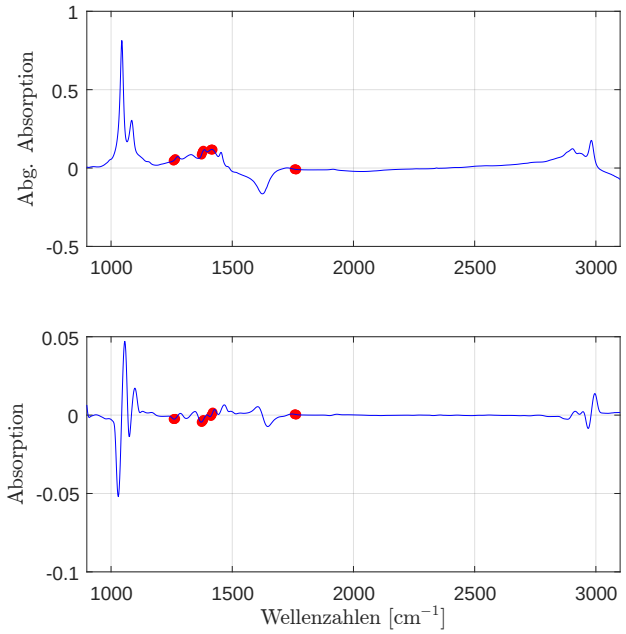

Abbildung 5.57: Die selektierten Wellenzahlen im PLS-Modell für den pH-Wert mit der ersten Savitzky-Golay Ableitung.

Grenze, beobachtet werden, womit eine eindeutige Empfehlung von 4 latenten Variablen bei dieser Wellenzahlselektion begründet und bestätigt wird. Bei den daraus resultierenden Residuen in derselben Abbildung (re.) können ebenfalls keine Auffälligkeiten beobachtet werden. Mit Ausnahme eines Weißweines sind keine Residuen durch eine hohe Streuung bzw. Ausreißer geprägt. Hierbei muss zusätzlich berücksichtigt werden, dass unmittelbar bei einem pH-Wert von 3.3 sich vier Weißweinschätzungen befinden, was eine höhere Variabilität vortäuscht.

Da für den pH-Wert keine Referenzgenauigkeit vorliegt, entfallen im Residuenplot in Abbildung 5.59 die in Grün dargestellten Referenzlinien. Es kann lediglich ein Ausreißer nach unten, mit einem Residuum in Höhe von -0.27, beobachtet werden. In dieser Grafik scheint ein minimal steigender Trend aufzutreten. Die Begründung hierfür liegt darin, dass anhand der Weinfarbe jeweils ein geringer Offset berechnet werden kann. So besitzen die Rotweinresiduen einen Bias von 0.01, während die Weiß- und Roséweine im Mittel um 0.01 überschätzt[10] werden, innerhalb der Weinfarben allerdings keinen erkennbaren Trend aufweisen. Aufgrund dieser Tatsache vermittelt der Residuenplot den Anschein eines vorhandenen Trends. Zudem scheinen die Rotweinresiduen mit steigendem Schätzwert eine geringfügig erhöhte Varianz aufzuzeigen, während die restlichen Residuen, abgesehen von dem minimalen Offset, sich unauffällig verhalten.

Selektiert man Wellenzahlen für ein rotweinspezifisches PLS-Modell, so erhält man mit 50 Wellenzahlen, einem linearen Kernel und ebenfalls 4 latenten Variablen ein leicht besseres Ergebnis (Standardabweichung der Rotweinresiduen in

[10] Bias in Höhe von -0.01.

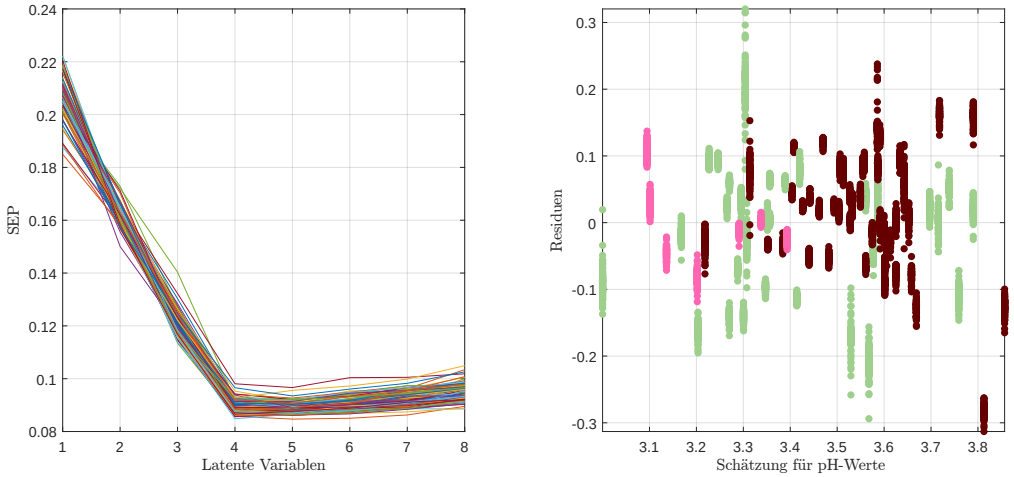

Abbildung 5.58: Verlauf des SEP in Abhängigkeit von latenten Variablen (li.) und die dazu korrespondierenden Weinresiduen der doppelten Kreuzvalidierung (re.) für den pH-Wert im PLS-Modell.

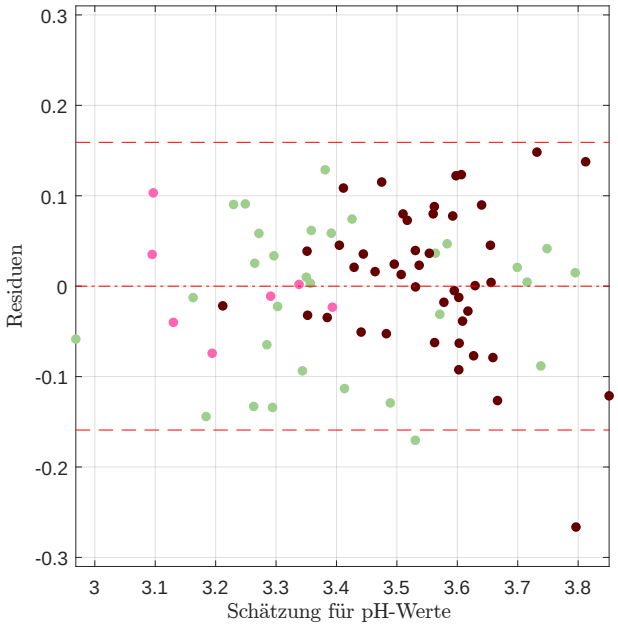

Abbildung 5.59: Residuenplot des PLS-Modells für den pH-Wert mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode.

Höhe von 0.0590 im Vergleich zu den Rotweinresiduen im Gesamtmodell mit 0.0815). Aufgrund des Minimalitätsprinzips wie in Abschnitt[11] erklärt, empfiehlt es sich allerdings, für ebendiese Rotweine eine Neukalibrierung des Gesamtmodells durchzuführen, was in einer Standardabweichung von 0.0706 resultiert und zeigt somit eine ähnliche Genauigkeit[12]. Darüber hinaus zeigt sich ein unauffälliges Verhalten der Rotweinresiduen, wie Abbildung 5.60 (re.), auch im Vergleich zu den Rotweinresiduen des Gesamtmodells (li.), demonstriert.

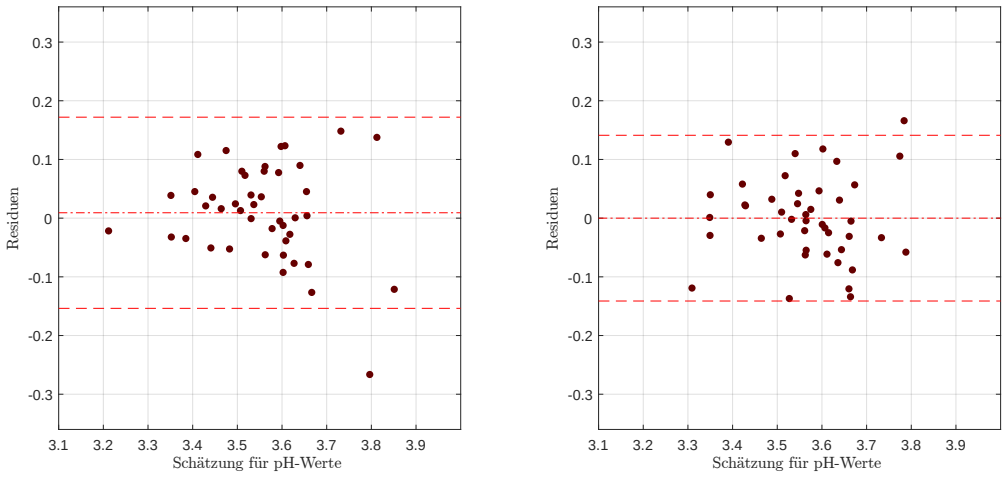

Abbildung 5.60: Extrahierte Rotweinresiduen des Residuenplots des PLS-Modells für den pH-Wert mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) (li.), sowie der Residuenplot des rotweinspezifischen PLS-Modells mit den entsprechenden Kennzahlen (re.).

Durch eine Neukalibrierung der Weiß- und Roséweine zeigt sich einerseits eine kaum merkbare Verbesserung in der Variabilität, wobei ein Trend in den Residuen auftritt, weswegen diese Neukalibrierung an dieser Stelle nicht durchgeführt wird.

Reproduzierbarkeit

Betrachtet man die Reproduzierbarkeit des Gesamtmodells, so können für dieses Parametersetting keine Auffälligkeiten bezüglich der Wiederholbarkeit beobachtet werden, wie Abbildung 5.61 in Kombination mit Tabelle 5.34 zeigt. Lediglich die Tails des Boxplots der Rotweine messen geringfügig mehr, allerdings kann eine gemittelte Standardabweichung von weniger als $1.4 \cdot 10^{-2}$ als gute Wiederholbarkeit interpretiert werden.

Die maximale Abweichung ist ebenfalls über alle Weinfarben hinweg vergleichbar und misst maximal 0.03. Vergleicht man die Wiederholbarkeit des Gesamtmodells

[11] Verweis auf Appendix, kurze Erklärung
[12] Der pH-Wert wird auf die zweite Nachkommastelle genau angegeben.

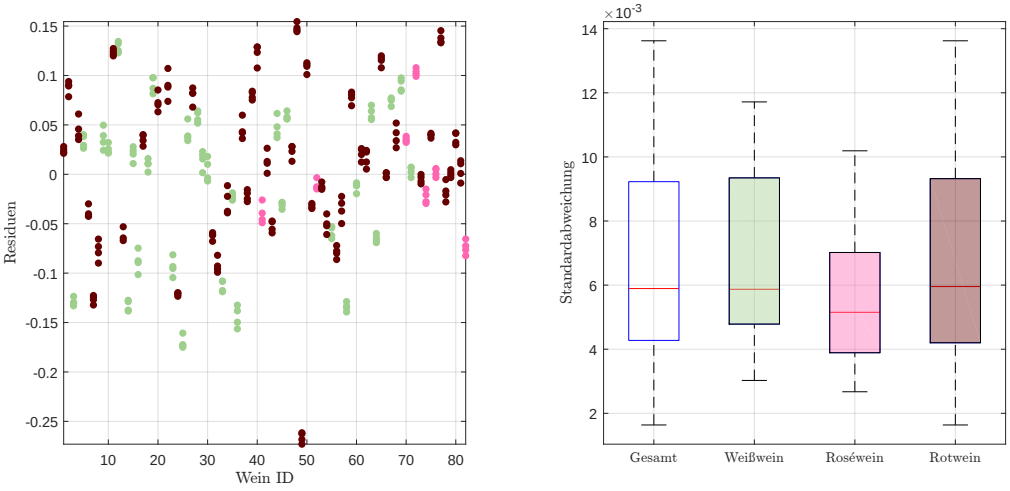

Abbildung 5.61: Residuen der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung des pH-Wertes im PLS-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, ohne künstliche Weine (re.).

mit jenen des Modells aus Abbildung 5.60, so reduzieren sich sämtliche Kennzahlen der Reproduzierbarkeit um circa ein Drittel.

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0067	0.0059	0.0268	0.0148	0.0129	0.0061
Roséwein	0.0057	0.0052	0.0230	0.0128	0.0114	0.0058
Rotwein	0.0067	0.0060	0.0333	0.0151	0.0131	0.0073
Gesamt	0.0066	0.0059	0.0333	0.0148	0.0125	0.0067

Tabelle 5.34: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für den pH-Wert im PLS-Modell anhand der Einzelspektren pro Wein ID.

Bei der Adaptierung des entwickelten Gesamtmodells für die Engines E22 und E24, sowie das Spektrometer V70(2016) zeigt sich für alle Datensätze ein ähnliches Verhalten, mit Ausnahme des Datensatzes E24, welcher einen Ausreißer bezüglich der Reproduzierbarkeit aufweist.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.0800	0.10	−0.25	0.15	0.0064	0.0064	0.0364
E24	0.0810	0.11	−0.25	0.16	0.0081	0.0072	0.1056
E25	0.0796	0.10	−0.27	0.15	0.0066	0.0059	0.0333
V70(2016)	0.0821	0.11	−0.26	0.15	0.0059	0.0058	0.0281

Tabelle 5.35: Performance des entwickelten PLS-Modells für den pH-Wert, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung..

5.14 Übersicht der Kennzahlen der PLS-Modelle

Für Ethanol kann in dem gefundenen PLS-Modell ein Ausreißer im Residuenplot mit einer der Referenzmethode vergleichbaren Standardabweichung gefunden und ein stabiles und plausibles Modell kalibriert werden. Eine Spezifizierung des Gesamtmodells kann hierfür zusätzlichen Informationsgehalt liefern und birgt teilweise Verbesserungspotential, auch wenn die PLS-Methodik grundsätzlich nicht sehr sensibel auf Ausreißer reagiert. Das Modell für Rotweine und einen auf $[8, 16]$ Vol.% eingeschränkten Wertebereich ohne die isolierten[13] Weine liefert vergleichsweise sehr niedrige Kennzahlen, verglichen mit den hierzu korrespondierenden Residuen des Gesamtmodells, wohingegen eine Neukalibrierung oder die Modellentwicklung für ausschließlich Rotweine nicht zielführend ist.

Der Extraktwert der Stichproben kann mit einer doppelten Standardabweichung von knapp über 2 g/l einigermaßen gut beschrieben werden. Auch wenn sich Weine mit hohen Extraktwerten einigermaßen gut replizieren lassen, zeigt sich eine große Schwankungsbreite für den Bereich $[0, 10]$ g/l, was jenem Wertebereich entspricht, welcher die meisten der vorliegenden Daten umfasst. Zudem scheinen die Rotweine tendenziell überschätzt, wohingegen die restlichen Weinproben eher unterschätzt werden. Deutliche Verbesserungen zeigen sich hierfür durch die Neukalibrierung für Rotweine bzw. Weiß-/Roséweine. Die Hauptkomponenten des Extraktwertes, die Zuckerarten Glukose und Fruktose lassen sich nur mit einer deutlich größeren Variabilität im Vergleich zur Labormethode modellieren. Erstgenannter Zucker liefert für den auf $[0, 10]$ g/l eingeschränkten Wertebereich einigermaßen plausible Ergebnisse, nicht jedoch bei der Entwicklung von ausschließlich Rotweinmodellen. Insgesamt zeigt sich bei den für E25 entwickelten Modellen eine bessere Datenqualität in den Datensätzen von Engine E22. Die für den Fruchtzucker entwickelten Modelle zeigen sich einigermaßen trotz der, mit der Variabilität der Referenzmethode verglichenen hohen Standardabweichung der Residuen, unauffällig. Zudem kann durch die Neukalibrierung des Fruktosemodells eine Art Robustifizierung durchgeführt werden.

[13] Hierunter werden sowohl die alkoholarmen, sowie die äußerst alkoholhaltigen Weine wie Portweine verstanden.

Bei den titrierbaren Säuren zeigt sich eine grundsätzliche Unterschätzung der Rotweine und Überschätzung der gegebenen Weißweine, weshalb eine starke Verbesserung bei Modellen für die Rotweine in einem hierfür spezifischen Modell erzielt werden kann. Das unterschiedliche Verhalten der Residuen kann mittels der vorliegenden Datensituation erklärt werden, da die Rotweine konzentrierter vorliegen, während die Weiß- und Roséweine mehr streuen. Dies spiegelt sich ebenfalls in den Residuen wider. Zudem weisen insbesondere die Datensätze E22 und E25 eine einigermaßen gute Reproduzierbarkeit auf.

Der Residuenplot der Weinsäure weist eine mehr als doppelt so hohe (beziehungsweise, abhängig von der Modellwahl, dreifache) Standardabweichung der Daten als die vorliegenden Laborwerte auf und kann somit nur mit einer enormen Streuung modelliert werden. Selbst eine Einschränkung auf Rotweine liefert kaum eine signifikante Verbesserung, ebenso wenig die Neukalibrierung mit den Weinfarben. Ein positiver Effekt dieser Modellierung stellt die hohe Reproduzierbarkeit dar. Dieselben Erkenntnisse können auch auf die L-Äpfelsäure übertragen werden. Betrachtet man das Modell für die Milchsäure, so kann ähnlich wie bei der Weinsäure im Vergleich mit der Referenzmethode diese Säure ebenfalls nur unzureichend, trotz der hohen Güte in der Reproduzierbarkeit, modelliert werden. Weder die Neukalibrierung noch Submodelle können die Variabilität nicht weiter reduzieren.

Sowohl bei den flüchtigen Säuren, als auch bei der Zitronensäure nehmen die tatsächlichen Konzentrationen lediglich wenige Werte an, was wiederum eine Modellierung, insbesondere, da von einer theoretischen Betrachtung einer kontinuierlichen Datenverteilung auszugehen ist, mit einer möglichen Auftrittswahrscheinlichkeit der Säuren am Rand der Wertebereiche, erschwert. Dennoch erweist sich die PLS-Methode als äußerst effektiv und es können plausible Modelle entwickelt werden. Selbst bei der Zitronensäure nimmt der im entsprechenden Kapitel analysierte enorme Ausreißer nur einen wesentlichen Einfluss auf die Standardabweichung der Residuen, nicht jedoch auf die Qualität des Modells.

Beim Zuckeralkohol Glyzerin liefert das PLS-Modell eine der Referenzmethode ähnliche Güte, bei einer teils hohen Reproduzierbarkeit von bis zu einer maximalen Differenz der Einzelschätzungen pro Wein ID von 0.63 g/l, obwohl keine Fehlmessungen als solche erkenntlich sind. Die beiden untersuchten Eigenschaften Dichte und pH-Wert zeigen sich als gut modellierbar. Dies gilt sowohl für das Verhalten der Residuen, als auch für die Güte der Reproduzierbarkeit.

		Residuen				Reproduzierbarkeit		
						Std.	Max. Abw.	
	Anz.	Std.	IQR	Min.	Max.	Mw.	Max.	Mw.
Ethanol, lin, M: 1, WZ: 50	81	0.0992	0.1548	−0.2423	0.2225	0.0121	0.0569	0.0267
Weiß	30	0.1172	0.1743	−0.2423	0.2225	0.0116	0.0569	0.0261
Rosé	7	0.0674	0.0959	−0.0605	0.1365	0.0150	0.0461	0.0328
Rot	44	0.0870	0.1532	−0.1569	0.1808	0.0120	0.0542	0.0261
[8, 16] Vol.%	75	0.0981	0.1549	−0.2423	0.2225	0.0118	0.0569	0.0259
Ethanol, gauss, M: 2, WZ: 20, [8, 16] Vol.%	75	0.0724	0.0931	−0.2361	0.2359	0.0252	0.1168	0.0562
Weiß	27	0.0884	0.1020	−0.2361	0.2359	0.0255	0.0995	0.0565
Rosé	7	0.0837	0.1212	−0.1052	0.1351	0.0225	0.1168	0.0491
Rot	41	0.0575	0.0788	−0.1115	0.1151	0.0255	0.0865	0.0573
Ethanol, gauss, M: 2, WZ: 20	81	0.0853	0.1290	−0.2032	0.1791	0.0121	0.0616	0.0269
Weiß	30	0.1021	0.1466	−0.2032	0.1791	0.0125	0.0616	0.0281
Rosé	7	0.0634	0.0918	−0.0679	0.1121	0.0147	0.0506	0.0321
Rot	44	0.0751	0.1232	−0.1432	0.1541	0.0114	0.0489	0.0253
[8, 16] Vol.%	75	0.0841	0.1250	−0.2032	0.1791	0.0119	0.0616	0.0264
Ethanol, gauss, M: 2, WZ: 20, Kal: Rot	44	0.0794	0.1035	−0.1682	0.1286	0.0108	0.0585	0.0241
[8, 16] Vol.%	41	0.0761	0.1039	−0.1682	0.1286	0.0103	0.0441	0.0229
Ethanol, gauss, M: 2, WZ: 20, [8, 16] Vol.%, Kal: Rot	41	0.0560	0.0707	−0.0959	0.1229	0.0255	0.0869	0.0572
Ethanol, gauss, M: 1, WZ: 9, Rot, [8, 16 Vol.%]	41	0.0766	0.1077	−0.1444	0.1764	0.0142	0.0593	0.0308
Ethanol, lin, M: 2, WZ: 20, [8, 16] Vol.%, Kal: Rot	41	0.0560	0.0707	−0.0959	0.1229	0.0255	0.0870	0.0572
Extrakt, lin, M: 2, WZ: 100	80	1.7530	1.7306	−4.2700	5.0592	0.1781	1.3179	0.3999
Weiß	30	2.2395	1.9267	−4.2700	5.0592	0.1462	0.5307	0.3307
Rosé	7	0.7794	0.4106	−0.7567	1.7605	0.1392	0.4112	0.3075
Rot	43	1.3172	1.6378	−3.3372	2.7918	0.2066	1.3179	0.4632
[0, 63.5] g/l	63	1.2479	1.4083	−3.1872	4.0112	0.1799	1.3179	0.4050

		Residuen				Reproduzierbarkeit		
						Std.	Max. Abw.	
	Anz.	Std.	IQR	Min.	Max.	Mw.	Max.	Mw.
Extrakt, gauss, M: 4, WZ: 40	80	1.1455	1.6412	−3.0271	3.3736	0.3886	2.0216	0.8604
Weiß	30	1.2647	1.8176	−1.8414	3.3736	0.3889	1.4370	0.8539
Rosé	7	0.7268	0.9011	−0.9657	1.2092	0.3382	1.3928	0.7413
Rot	43	1.1239	1.5812	−3.0271	2.4404	0.3966	2.0216	0.8843
$[0, 63.5]$ g/l	63	1.0813	1.3635	−2.7290	3.3736	0.3667	2.0216	0.8118
Extrakt, gauss, M: 4, WZ: 40, Kal: Rot	43	1.2117	1.6138	−3.0920	2.3029	0.4066	1.9887	0.9051
$[0, 63.5]$ g/l	39	1.1513	1.4654	−3.0920	2.3029	0.3994	1.9887	0.8830
Extrakt, gauss, M: 4, WZ: 40, Kal: Weiß, Rosé	37	1.0813	1.5650	−1.9547	3.2208	0.3962	1.5207	0.8731
Weiß	30	1.1387	1.6393	−1.9547	3.2208	0.4047	1.5088	0.8884
Rosé	7	0.8620	1.2918	−1.0267	1.3910	0.3598	1.5207	0.8075
$[0, 63.5]$ g/l	24	1.0762	1.3736	−1.6937	3.2208	0.3555	1.5207	0.7892
Extrakt, lin, M: 4, WZ: 100, $[0, 63.5]$ g/l	63	0.7646	1.0448	−1.7048	1.7696	0.1124	0.5045	0.2490
Weiß	17	0.7095	1.1262	−1.4878	0.8197	0.0984	0.4268	0.2150
Rosé	7	0.6515	0.9578	−1.1434	0.7357	0.1280	0.3804	0.2869
Rot	39	0.7630	0.9730	−1.7048	1.7696	0.1158	0.5045	0.2571
Glukose, lin, M: 6, WZ: 20	81	0.7099	0.7212	−1.8783	2.8183	0.4213	5.2162	0.9480
Weiß	30	0.8106	0.9622	−1.8783	2.0632	0.4948	5.2162	1.1316
Rosé	7	0.5175	0.6541	−0.7661	0.6836	0.4369	1.7358	0.9802
Rot	44	0.6626	0.5784	−1.1887	2.8183	0.3688	1.6887	0.8178
$[0, 10]$ g/l	62	0.6221	0.7400	−1.1887	2.0632	0.3617	1.7358	0.8085
Glukose, gauss, M: 6, WZ: 40, rot	44	0.4306	0.4140	−0.9399	1.7759	0.2002	0.9897	0.4440
$[0, 10]$ g/l	39	0.3441	0.4167	−0.9399	0.6027	0.1939	0.9897	0.4276

		Residuen				Reproduzierbarkeit		
						Std.	Max. Abw.	
	Anz.	Std.	IQR	Min.	Max.	Mw.	Max.	Mw.
Glukose, gauss, M: 6, WZ: 20, $[0, 10]$ g/l	62	0.2384	0.3859	−0.7763	0.4808	0.2305	1.1829	0.5070
Weiß	16	0.2657	0.4386	−0.4100	0.4808	0.2260	0.9590	0.5042
Rosé	7	0.1724	0.3127	−0.2893	0.1419	0.1966	0.8009	0.4452
Rot	39	0.2387	0.3497	−0.7763	0.3924	0.2385	1.1829	0.5193
Glukose, gauss, M: 5, WZ: 20, Rot, $[0, 10\,\mathrm{g/l}]$	39	0.2230	0.2475	−0.3891	0.4835	0.2463	1.0784	0.5529
Glukose, lin, M: 7, WZ: 50, $[0, 10]$ g/l	62	0.2334	0.3108	−0.5207	0.5095	0.1815	0.7793	0.3945
Weiß	16	0.2564	0.3320	−0.4908	0.5095	0.1691	0.6257	0.3714
Rosé	7	0.1537	0.2527	−0.3157	0.0883	0.1898	0.7793	0.4301
Rot	39	0.2323	0.3205	−0.5207	0.3342	0.1850	0.6378	0.3977
Fruktose, gauss, M: 2, WZ: 40	81	1.0561	1.2328	−2.9999	3.2063	0.1861	0.9618	0.4141
Weiß	30	1.3719	1.5175	−2.9999	3.2063	0.1992	0.9618	0.4499
Rosé	7	0.9901	0.8949	−0.8461	2.0409	0.1698	0.5767	0.3787
Rot	44	0.8004	1.1491	−2.6774	1.6245	0.1798	0.8372	0.3954
$[0, 10]$ g/l	62	0.7170	1.0503	−1.3348	2.0409	0.1732	0.6933	0.3844
Fruktose, gauss, M: 2, WZ: 40, Kal: Rot	44	0.6153	1.0856	−1.0168	1.2766	0.1464	0.6211	0.3221
$[0, 10]$ g/l	39	0.5967	0.9755	−1.0168	1.2766	0.1480	0.5685	0.3243
titrierbare Säuren, lin, M: 2, WZ: 20	81	0.1788	0.2511	−0.3741	0.3853	0.0271	0.1249	0.0597
Weiß	30	0.2082	0.2653	−0.3741	0.3566	0.0246	0.0929	0.0540
Rosé	7	0.1679	0.2137	−0.3384	0.1700	0.0295	0.1084	0.0629
Rot	44	0.1304	0.1600	−0.1548	0.3853	0.0285	0.1249	0.0632
titrierbare Säuren, lin, M: 2, WZ: 20, Kal: Rot	44	0.1467	0.1704	−0.2462	0.4730	0.0224	0.0908	0.0500
titrierbare Säuren, lin, M: 3, WZ: 40, rot	44	0.1008	0.1486	−0.2563	0.1956	0.0148	0.0614	0.0323

			Residuen				Reproduzierbarkeit		
							Std.	Max. Abw.	
		Anz.	Std.	IQR	Min.	Max.	Mw.	Max.	Mw.
Weinsäure, lin, M: 2, WZ: 20		81	0.3416	0.4351	−0.7654	0.9844	0.0111	0.0629	0.0246
	Weiß	30	0.3587	0.4822	−0.7654	0.7303	0.0108	0.0533	0.0240
	Rosé	7	0.2613	0.4067	−0.5958	0.1293	0.0085	0.0248	0.0183
	Rot	44	0.3103	0.3308	−0.6427	0.9844	0.0118	0.0629	0.0261
	$[1, 2.9]$ g/l	79	0.3320	0.4249	−0.7654	0.9844	0.0112	0.0629	0.0247
Weinsäure, lin, M: 4, WZ: 9		81	0.3021	0.4170	−0.6855	0.7588	0.0219	0.1224	0.0490
	Weiß	30	0.3492	0.4075	−0.6855	0.7448	0.0216	0.0929	0.0477
	Rosé	7	0.2253	0.3980	−0.3786	0.1621	0.0147	0.0586	0.0344
	Rot	44	0.2807	0.3950	−0.6009	0.7588	0.0233	0.1224	0.0522
	$[1, 2.9]$ g/l	79	0.2919	0.3821	−0.6855	0.7588	0.0220	0.1224	0.0490
Weinsäure, gauss, M: 7, WZ: 100		81	0.2338	0.3113	−0.6325	0.6846	0.0338	0.1847	0.0762
	Weiß	30	0.2449	0.3651	−0.5875	0.3904	0.0364	0.1458	0.0816
	Rosé	7	0.1826	0.2688	−0.2812	0.2076	0.0231	0.0890	0.0524
	Rot	44	0.2360	0.2952	−0.6325	0.6846	0.0337	0.1847	0.0763
	$[1, 2.9]$ g/l	79	0.2267	0.3060	−0.6325	0.6846	0.0334	0.1847	0.0752
L-Äpfelsäure, lin, M: 4, WZ: 15		81	0.3670	0.5561	−0.8715	0.8738	0.1651	0.9130	0.3652
	Weiß	30	0.4112	0.6304	−0.8715	0.7776	0.1738	0.9130	0.3858
	Rosé	7	0.3073	0.3237	−0.4364	0.5288	0.1723	0.4957	0.3962
	Rot	44	0.3504	0.5133	−0.6058	0.8738	0.1581	0.6195	0.3462
L-Äpfelsäure, gauss, M: 4, WZ: 40		81	0.3275	0.3743	−0.8099	0.9408	0.0252	0.1182	0.0564
	Weiß	30	0.4459	0.6262	−0.8099	0.9408	0.0275	0.1124	0.0619
	Rosé	7	0.2021	0.3566	−0.4038	0.1005	0.0190	0.0551	0.0439
	Rot	44	0.2391	0.3298	−0.5377	0.7382	0.0246	0.1182	0.0547

			Residuen				Reproduzierbarkeit		
							Std.	Max. Abw.	
		Anz.	Std.	IQR	Min.	Max.	Mw.	Max.	Mw.
L-Äpfelsäure, gauss, M: 4, WZ: 40, Kal: Weiß, Rosé		37	0.4011	0.5657	−0.9198	0.7927	0.0291	0.1182	0.0654
	Weiß	30	0.4232	0.5644	−0.9198	0.7927	0.0307	0.1182	0.0688
	Rosé	7	0.2089	0.2678	−0.5959	−0.0312	0.0222	0.0623	0.0509
Milchsäure, lin, M: 4, WZ: 40		81	0.2275	0.2875	−0.4959	0.7348	0.0241	0.1231	0.0533
	Weiß	30	0.2104	0.2840	−0.3509	0.4550	0.0262	0.1231	0.0581
	Rosé	7	0.2252	0.2768	−0.1886	0.4663	0.0199	0.0564	0.0446
	Rot	44	0.2428	0.2968	−0.4959	0.7348	0.0234	0.0823	0.0515
flüchtige Säuren, lin, M: 3, WZ: 20		80	0.1138	0.1526	−0.2825	0.3218	0.0128	0.0796	0.0289
	Weiß	30	0.1212	0.1925	−0.2293	0.2126	0.0144	0.0796	0.0325
	Rosé	7	0.0919	0.1655	−0.1614	0.0628	0.0106	0.0409	0.0236
	Rot	43	0.1122	0.1271	−0.2825	0.3218	0.0121	0.0691	0.0272
Zitronensäure, lin, M: 1, WZ: 20		81	0.1530	0.1168	−0.2766	1.0807	0.0034	0.0268	0.0078
	Weiß	30	0.1041	0.1189	−0.2766	0.2330	0.0030	0.0213	0.0068
	Rosé	7	0.3915	0.2085	−0.0041	1.0807	0.0023	0.0069	0.0051
	Rot	44	0.0757	0.0567	−0.1305	0.2107	0.0039	0.0268	0.0088
Glyzerin, lin, M: 5, WZ: 50		81	0.4867	0.6502	−1.4709	0.9485	0.1330	0.6269	0.2976
	Weiß	30	0.5301	0.7484	−1.4709	0.9485	0.1531	0.6269	0.3472
	Rosé	7	0.3998	0.7107	−0.8478	0.1464	0.1095	0.5189	0.2407
	Rot	44	0.4630	0.6606	−1.1064	0.8975	0.1231	0.5972	0.2728
Dichte, lin, M: 2, WZ: 9		80	0.7980	1.0045	−1.8601	1.9289	0.2631	1.1869	0.5846
	Weiß	30	1.0039	0.9702	−1.8601	1.9289	0.2510	1.1197	0.5559
	Rosé	7	0.2206	0.4086	−0.4102	0.0718	0.2445	0.8975	0.5576
	Rot	43	0.6999	1.1080	−1.5100	1.5596	0.2745	1.1869	0.6091

			Residuen				Reproduzierbarkeit		
							Std.	Max. Abw.	
		Anz.	Std.	IQR	Min.	Max.	Mw.	Max.	Mw.
Dichte, gauss, M: 2, WZ: 100		80	0.7500	0.8743	−2.3545	2.3712	0.0817	0.5008	0.1844
	Weiß	30	0.9317	0.7941	−2.3545	2.3712	0.0693	0.2748	0.1574
	Rosé	7	0.2492	0.4823	−0.8367	−0.2368	0.0782	0.2601	0.1767
	Rot	43	0.6122	0.8883	−1.2017	1.3731	0.0909	0.5008	0.2044
pH-Werte, lin, M: 3, WZ: 9		81	0.0930	0.1069	−0.2662	0.2362	0.0098	0.0459	0.0220
	Weiß	30	0.1150	0.1537	−0.2662	0.1988	0.0108	0.0414	0.0241
	Rosé	7	0.0824	0.0562	−0.1555	0.1207	0.0080	0.0246	0.0178
	Rot	44	0.0774	0.1052	−0.1435	0.2362	0.0094	0.0459	0.0212
pH-Werte, gauss, M: 4, WZ: 20		81	0.0796	0.1010	−0.2664	0.1482	0.0066	0.0333	0.0148
	Weiß	30	0.0815	0.1351	−0.1704	0.1287	0.0067	0.0268	0.0148
	Rosé	7	0.0572	0.0626	−0.0742	0.1032	0.0057	0.0230	0.0128
	Rot	44	0.0814	0.1119	−0.2664	0.1482	0.0067	0.0333	0.0151

Tabelle 5.36: Übersicht über die wichtigsten Kennzahlen der behandelten PLS-Modelle.

6 Neuronale Netzwerke

In diesem Kapitel werden die künstlichen neuronalen Netzwerke als alternative Modellierungsmöglichkeit für den Zusammenhang der Absorptionswerte und den Weinkomponenten vorgestellt. Zuerst wird erklärt, wie die biologischen Vorbilder in die Sprache der Mathematik übertragen werden können. Als Folge dessen werden unterschiedliche Gestaltungsmöglichkeiten wie die Modellparameter vorgestellt und insbesondere der Algorithmus, wie ein funktionales Modell zustande kommt (hierunter versteht man speziell den Lernalgorithmus, welcher das neuronale Netzwerk kalibriert), erklärt. Weiters wird auf die mathematische Motivation und die Mächtigkeit dieser Modelle hingewiesen, bevor die spezifischen Modellparameter, welche in dieser Arbeit Anwendung finden, erklärt werden.

6.1 Motivation und Namensgebung

Wie bereits die Bezeichnung KÜNSTLICHE NEURONALE NETZWERKE vermuten lässt, versuchen diese ein natürliches Netz von Neuronen, wie man dieses beispielsweise im menschlichen Gehirn oder Nervensystem vorfindet, in die Sprache der Mathematik zu transformieren. Ein derartiges Netz setzt sich im Wesentlichen aus einzelnen Neuronen (Nervenzellen) zusammen und eine schematische Darstellung findet sich in Abbildung 6.1[1].

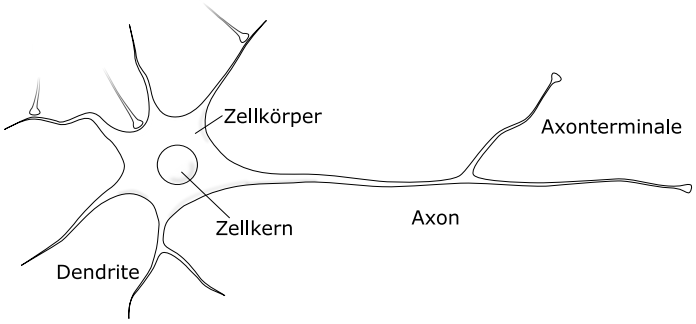

Abbildung 6.1: Schematische Darstellung eines biologischen Neurons

Die abgebildete Nervenzelle erhält Informationen in Form von elektrischen Impulsen, und diese werden durch die Empfänger, den sogenannten Dendriten, von anderen Zellen erhalten. Diese Informationen werden im Zellkern modifiziert und über das Axon an die sogenannten Axonterminale weitergegeben. Von diesen Enden eines Neurons werden Impulse an die Dendriten anderer Neuronen weitergeleitet. Bei dieser Transformation und Übermittlung von Impulsen handelt

[1] Die Erlaubnis für das Verwenden der Grafik vom Ersteller Mashio erteilt und stammt von [22].

es sich um komplexe, chemische Prozesse, in welchen unter anderem die Stärke der elektrischen Impulse erhöht oder vermindert wird. Sofern die Intensität eines solchen Impulses einen gewissen Schwellenwert übersteigt, findet ein Informationsaustausch statt. Die Aneinanderreihung solcher Nervenzellen bezeichnet man als neuronales Netzwerk.

Um diesen Prozess in mathematischer Sprache darzustellen, müssen einige Annahmen getroffen werden, welche auf die Arbeit von Warren S. McCulloch aus dem Jahre 1943 zurückgehen. Hierbei wird dem Netzwerk unter anderem unterstellt, dass sich die Struktur des Netzwerkes mit Fortdauer der Zeit nicht ändert, sowie es sich um einen sogenannten „Alles-Oder-Nichts“[2] Prozess handelt. Dies bedeutet, sofern ein gewisser Schwellenwert überschritten wird, feuert das Neuron einen Impuls: der Output eines Neurons kann somit als Indikatorfunktion interpretiert werden. Sei pre_i jene Menge von Neuronen, welche eine direkte Verbindung mit den Dendriten der Nervenzelle i besitzen. Auf diese Art kann ein erstes simplifiziertes Modell veranschaulicht werden, welches den Output out_i des Neurons i wie in Gleichung (6.1) interpretiert.

$$\text{out}_i := \Theta\Bigg(\underbrace{\sum_{j\in\text{pre}_i} w_{j\to i}\cdot \text{out}_j, \mu_i}_{=:\text{inp}_i}\Bigg) \in \{0,1\}, \tag{6.1}$$

mit Gewichten $w_{j\to i}\in\mathbb{R}$ und der Funktion $\Theta(x,y) := 1_{\{x\geq y\}}$. $w_{j\to i}$ repräsentiert die Intensität des Impulses von Neuron j an i und μ_i jenen Grenzwert, welcher überschritten werden muss, sodass die Zelle i über dessen Axonterminale den Impuls weiterleitet.

Nicht berücksichtigt werden hierbei die Tatsachen, dass biologische Neuronen nicht ausschließlich lineare Transformationen durchführen, sondern zugleich komplexe Prozesse abbilden können, wie beispielsweise logische Operatoren oder nichtlineare Kombinationen der empfangenen Impulse. Ein weiterer Informationsverlust findet durch das Nichtberücksichtigen von Impulsphasen statt, sowie durch den Wertebereich des Outputs, da dieser durchaus stetige Werte annehmen kann. Letzterer Kritikpunkt kann durch alternative Definitionen von Θ eliminiert werden und die Funktionenklasse Θ wird als Klasse der Aktivierungsfunktionen bezeichnet.

Fügt man die künstlichen Neuronen zu einem Netzwerk zusammen, so bilden die wichtigsten Bestandteile die Knotenstruktur sowie das Vernetzungsmuster. Hierunter versteht man die Wahl der Aktivierungsfunktion, die Anzahl der Neuronen mit den entsprechenden Verbindungen[3], sowie die dazugehörenden Schwellenwerte. Ein weiterer wichtiger Bestandteil ist die Bestimmung der Gewichte $w_{\cdot\to\cdot}$, welche durch sogenannte Lernalgorithmen bestimmt werden, wie dies beispielsweise durch die Backpropagation in Abschnitt 6.2.2 erfolgt. Hierbei kann μ als zusätzliches Neuron mit $\text{out}_\mu = 1$ und zugehörenden Gewichten $w_{\mu\to\cdot} = \mu_i$ aufgefasst werden.

[2] „all-or-none“ in [30]

[3] Es müssen nicht alle Neuronen miteinander verbunden sein und es bedarf deswegen einer Definition dieser Verbindungen.

Mit der Interpretation des Schwellwertes als zusätzliches Neuron können sämtliche Anpassungen dieser Schwellenwerte in Lernalgorithmen zur Aktualisierung der Gewichte $w_{\cdot\to\cdot}$ eingepflegt werden.

6.2 Framework

Gegeben seien analog zu Abschnitt 4.1 n Objekte mit Messwerten an s unterschiedlichen Prädiktoren, zusammengefasst zu $X \in \mathbb{R}^{n\times s}$, bestehend aus Absorptionswerten für n unterschiedliche Weinproben zu den korrespondierenden s Wellenzahlen, sowie die dazugehörenden Referenzwerte $Y \in \mathbb{R}^{n\times q}$.

Definition 6.1 (Künstliches, neuronales vorwärts gerichtetes k-Layer Netzwerk)**.** Unter einem künstlichen, neuronalen Netzwerk versteht man einen gerichteten Graphen $G = (V, E)$ mit Knotenmenge V und Kantenmenge E. Zusätzlich wird allen Knoten $i \in V$ je eine Aktivierungsfunktion $f_i(\cdot)$, sowie allen Kanten $e \in E$ eine Gewichtsfunktion $\omega\colon E \to \mathbb{R}$ mit folgenden Eigenschaften zugeordnet:

Sei $\text{pre}_i := \{j \in V : (j,i) \in E\}$.

- Die Knotenmenge V kann in drei disjunkte Knotenmengen unterteilt werden: $V = V_{\text{Input}} \dot\cup V_{\text{Inneres}} \dot\cup V_{\text{Output}}$ mit $|V_{\text{Input}}| = p$ und $|V_{\text{Output}}| = q$. Die Kardinalität von V_{Inneres} sei m. Hierbei korrespondieren die Knoten $i \in V_{\text{Input}}$ mit den Prädiktoren und $j \in V_{\text{Output}}$ mit den Responsevariablen.
- Der Wert (Output) eines Knotens $V \ni i = 1,\dots,s$ sei bei Betrachtung des Objektes $j \in \{1,\dots,n\}$ definiert als

$$\text{out}_i\left(x_j, W\right) :- 1_{\{i\in V\setminus V_{\text{Input}}\}} f_i\left(\text{inp}_i\left(x_j, W\right)\right) + 1_{\{i\in V_{\text{Input}}\}} x_{j,i}. \tag{6.2}$$

  Hierbei entspricht x_j den Messwerten des Objektes $j = 1,\dots,n$, d.h. der $j$ten Datenzeile der Prädiktormatrix X und $x_{j,i}$ dem Eintrag $X_{j,i}$ als dem Knoten i zugeordneter Messwert des Objektes j. W beschreibt den Zusammenhang der einzelnen Knoten des Netzwerkgraphen G und wird in weiterer Folge als Matrix aufgefasst, welche die Inputs gewichtet. Der Eintrag $W_{j,i} =: w_{j\to i}$ entspricht dem Skalierungsfaktor des von Knoten j an Knoten i gefeuerten Outputs. Der Input inp_i des Knotens i definiert sich durch

$$\text{inp}_i\left(x_j, W\right) := \sum_{k\in\text{pre}_i} w_{k\to i}\text{out}_k.$$

Zusätzlich müssen noch Annahmen für die Startwerte der rekursiven Darstellung in Gleichung (6.2) getroffen werden. Diese hängen von der tatsächlich verwendeten Art des Netzwerkes ab. Für die hier verwendeten Netzwerke wird stets $\text{pre}_i = \emptyset, \forall\, i \in V_{\text{Input}}$ und $\nexists (i,j) \in E$ mit $i \in V_{\text{Output}}$.

Bei azyklischen Graphen G spricht man von einem künstlichen neuronalen Feed-Forward Netzwerk. Wenn zusätzlich die Knotenmenge V in $k+1$ disjunkte Partitionen $V = V_0 \dot\cup V_1 \dot\cup \ldots \dot\cup V_{k-1} \dot\cup V_k$ mit $V_0 := V_{\text{Input}}$, $V_k := V_{\text{Output}}$ und der Eigenschaft

$$\forall e = (i,j) \in E, i \in V_l,\ l \in \{1,\dots,k-1\}\colon j \in V_{l+1}$$

zerlegt werden kann, so spricht man von einem künstlichen, vorwärts gerichteten neuronalen k-Layer Netzwerk. In diesem speziellen Fall werden alle Aktivierungsfunktionen einer Partition als identische Funktion angenommen.

Bemerkung 6.2. Alternativ kann ein vorwärts gerichtetes Netzwerk durch die Nummerierung, welche in vorangegangener Notation impliziert wird, festgelegt werden. So definiert [24] ein neruonales vorwärts gerichtetes Netzwerk derart, dass die Knoten nummeriert werden können und für sämtliche Verbindungen $e = (i, j)$ die Bedingung $i < j$ gilt. In der Praxis werden diese Knoten in sogenannten Layern angeordnet, sodass die Verbindungen stets zu einem höher nummerierten Layer führen.

Bemerkung 6.3. Die Knoten $v \in V$ entsprechen den künstlichen Neuronen.

Handelt es sich bei dem betrachteten Konstrukt um ein künstliches, neuronales vorwärts gerichtetes k-Layer Netzwerk, so kann die Schätzung der Responsevariable als deterministische und rekursive Funktion angegeben werden. Konkret bedeutet dies für einen Datensatz $i \in \{1, \dots, n\}$, wobei x_i der $i$ten Zeile der Matrix X entspricht. Seien hierzu die Aktivierungsfunktionen, sowie W gegeben und vollständig spezifiziert.

$$\begin{aligned} \text{out}_j(x_i, W) &:= x_{i,j} && j \in V_0 = V_{\text{Input}} \\ \text{out}_h(x_i, W) &= f_h\left(\sum_{l \in \text{pre}_h} w_{l \to h} \cdot \text{out}_l(x_i, W)\right) && h \in V_v,\ v = 1, \dots, k \\ \hat{y}(x_i, W) &:= \{\text{out}_h(x_i, W)\}_{h \in V_{\text{Output}}} \end{aligned} \tag{6.3}$$

Hierbei werden den Neuronen der Inputpartition in einer festgelegten Reihenfolge die Prädiktorenwerte $(x_i)_j$ zugeordnet. Für ein Netzwerk mit gegebener Gewichtsmatrix W können somit mithilfe der Gleichungen (6.3) die abhängigen q Responses des Objektes i geschätzt werden.

Bemerkung 6.4. In diesem Kapitel werden, stets künstliche, vorwärts gerichtete neuronale k-Layer Netzwerke behandelt. Daher werden diese in weiterer Folge stets als (neuronale) Netzwerke (mit Tiefe k) bezeichnet. Zudem werden die Aktivierungsfunktionen des $i$ten Layers mit f_i bezeichnet. Weiters sind alle Knoten zweier aufeinanderfolgender Layer miteinander verbunden:

$$\forall\, l = 1, \dots, k \colon \forall\, i \in V_l \colon \text{pre}_i = V_{l-1}.$$

Ein neuronales Netzwerk mit Tiefe 2 kann wie in Abbildung 6.2 visualisiert werden. Hierbei entpricht der Inputlayer den gegebenen Prädiktoren (grün) und werden via der versteckten Schicht (blau) hin zum Output (rot) transferiert. Hierbei kann die Anzahl der zu einem Layer gehörenden mathematischen Neuronen im Wesentlichen beliebig gewählt werden und jeder dargestellte Knotenpunkt kann wie in Abbildung 6.3 interpretiert werden.

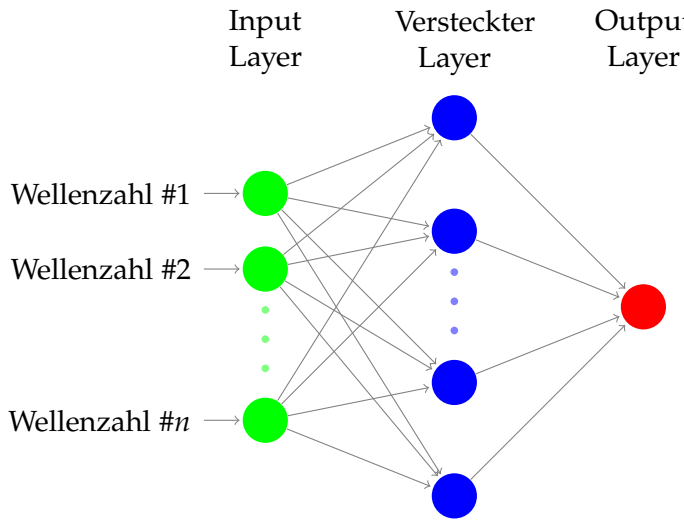

Abbildung 6.2: Schematische Darstellung eines vorwärts gerichteten neuronalen Netzwerkes mit einem versteckten Layer

Bei der Betrachtung eines mathematisch interpretierten Neurons in Abbildung 6.3 kann der Input x_i sowohl die Werte der Prädiktormatrix X annehmen, sowie als Output der Neuronen des vorangegangenen Layers aufgefasst werden. Dieser Input wird mit Gewichten w skaliert und gemeinsam mit einem Bias b aufsummiert. Mithilfe der Aktivierungsfunktion wird diese Summe transformiert und bildet den Output, welcher als Schätzung oder wiederum als Input für einen weiteren Layer dient.

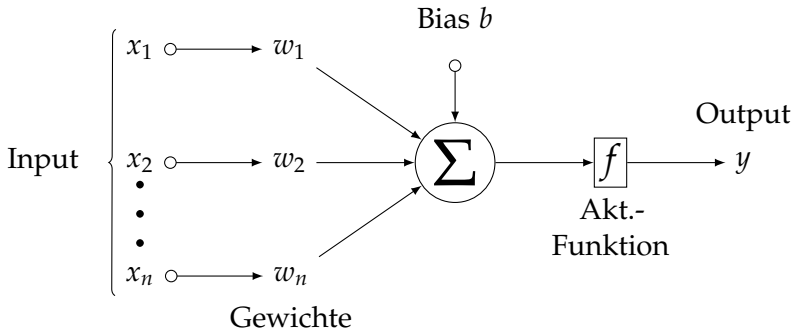

Abbildung 6.3: Schematische Darstellung eines mathematischen Neurons

Aktivierungsfunktionen

Eine essentielle Eigenschaft der neuronalen Netzwerke bilden die Aktivierungsfunktionen. Für die hier betrachteten Anwendungen seien diese definiert als Abbildung $f\colon \mathbb{R} \to \mathbb{R}$. Durch die biologische Motivation, wie durch Abschnitt 6.1 beziehungsweise [30] inspiriert, bietet sich die Indikaterfunktion $f(x) = 1_{\{x \geq 0\}}$ an. Die bei $x = 0$ befindliche Sprungstelle kann mit einem Biasknoten verschoben werden. Aus praktischen Gründen, um besondere und günstige Eigenschaften

wie Stetigkeit oder Differenzierbarkeit zu erhalten, wird die Aktivierungsfunktion modifiziert. Um eine stetige Transferfunktion verwenden zu können, kann diese auf eine stückweise lineare Funktion der Form

$$f(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases}$$

für $a < b$ erweitert werden. Da diese Funktion an den Stellen $x_1 = a$ beziehungsweise $x_2 = b$ nicht differenzierbar ist, findet sich in der Literatur die Sigmoidfunktion oftmals als stetig differenzierbare Erweiterung wieder. Hierbei handelt es sich um eine der am meisten verwendeten Aktivierungsfunktionen (vgl. hierzu [11], Seite 34ff). Sei $\alpha > 0$. Die Sigmoidfunktion $\mathcal{S}(x)$ besitzt die Form

$$\mathcal{S}(x) := (1 + \exp\{-\alpha \cdot x\})^{-1}. \tag{6.4}$$

Eine Besonderheit dieser Variante ist, dass die Ableitung als Funktion in $\mathcal{S}(x)$ dargestellt werden kann:

$$\begin{aligned} \frac{\partial \mathcal{S}(x)(x)}{\partial x} &= \frac{\alpha \cdot \exp\{\alpha \cdot x\}}{(1 + \exp\{-\alpha \cdot x\})^2} \\ &= \frac{1}{1 + \exp\{-\alpha \cdot x\}} \cdot \left(\frac{\alpha \cdot \exp\{\alpha \cdot x\}}{1 + \exp\{-\alpha \cdot x\}} \right) \\ &= \mathcal{S}(x) \cdot (1 - \mathcal{S}(x)). \end{aligned}$$

Diese Eigenschaft wird beispielsweise im Lernalgorithmus in Abschnitt 6.2.2 zur Vereinfachung der Berechnung verwendet.

Zusätzlich gilt, dass für $\alpha \to \infty$ die Sigmoidfunktion gegen die Indikatorfunktion konvergiert. Modifikationen der Sigmoidaktivierungsfunktion sind in [11] auf Seite 190 und 191 beschrieben.

Die hier erwähnten Funktionen haben alle einen Wertebereich von $[0, 1]$ und ein monoton steigendes Verhalten. Dieser Bereich kann durch das Verwenden einer linearen Transferfunktion $f(x) = ax + b$ erweitert werden. Selbst die Monotonie muss nicht gegeben sein. Beispielsweise können auch die Radiale Basisfunktion $f(x) = \exp\{-x^2\}$, sowie eine große Anzahl von Funktionen theoretisch als Aktivierungsfunktion verwendet werden. Die hier vorgestellten Vertreter sind in Abbildung 6.4 zusammengefasst.

In dieser Abbildung wird auf die explizite Darstellung der linearen Aktivierungsfunktion verzichtet. Im oberen dieser drei Plots finden sich die Indikatorfunktion mit Sprungstelle 0, die stückweise lineare Abbildung mit $(a, b) = (-4, 4)$, sowie die Sigmoidfunktion mit Parameter $\alpha = 1$. Für $(a, b) = (-t, t)$ mit kleinem t bzw. mit α sehr groß, können diese Variationen der Aktivierungsfunktion als stetige bzw. stetig differenzierbare Version der Indikatorfunktion interpretiert werden.

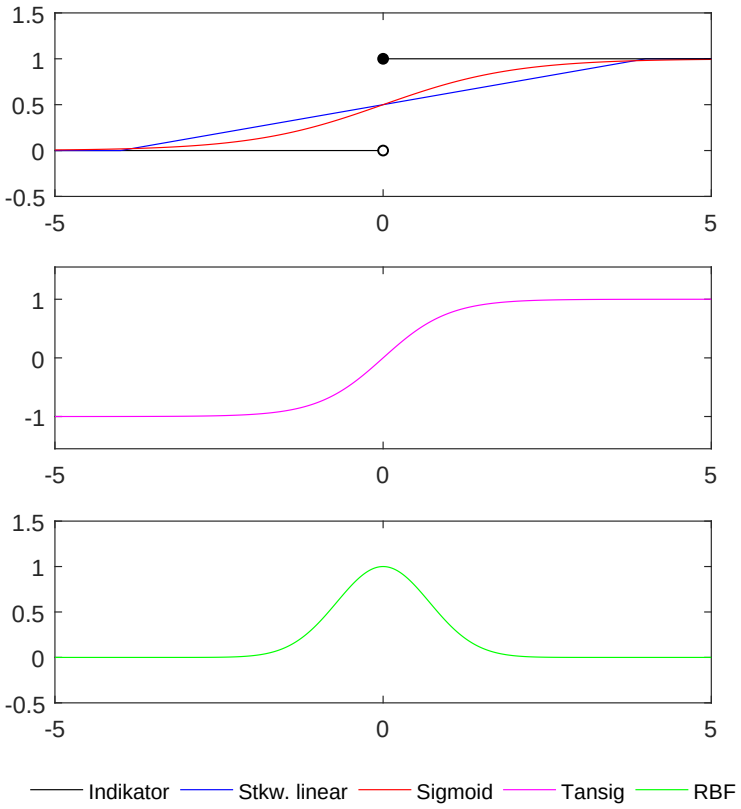

Abbildung 6.4: Visualisierung unterschiedlicher Aktivierungsfunktionen: Die klassische Indikatorfunktion, deren stetige Adaptierung, sowie eine stetig differenzierbare Version, jeweils mit einem Wertebereich $[0,1]$ (o.), einer Aktivierungsfunktion mit modifiziertem Wertebereich (m.) und zusätzlich eine nicht monotone Aktivierungsfunktion (u.).

Während die Funktionen in der oberen dieser drei Grafiken einen Wertebereich $[0,1]$ aufweisen, so wird dieser Bereich in der mittleren Abbildung mithilfe der hyperbolischen Tangens-Sigmoidfunktion auf $[-1,1]$ erweitert. Diese definiert sich als

$$\mathrm{TS}(x) := \frac{2}{1+\exp\{-\alpha x\})} - 1 \tag{6.5}$$

und wird in dieser Arbeit oftmals wegen der schnellen Berechnungsgeschwindigkeit bevorzugt (vgl. hierzu [19]). Offensichtlich gilt $\mathrm{TS}\left(\frac{2x}{\alpha}\right) = \tanh x$. In der mittleren Abbildung wird diese Aktivierungsfunktion mit Parameter $\alpha = 2$ dargestellt.

Analog kann eine Relation zwischen der Sigmoidfunktion und des Tangenshyperbolikus hergestellt werden, da $2 \cdot \mathcal{S}(\frac{2x}{\alpha}) - 1 = \tanh x$ gilt.

Während die ersten vier Aktivierungsfunktionen streng monoton steigend sind, kann alternativ auch eine Funktion mit hoher Dichte um den Biasknoten als solche definiert werden. In der mittleren Grafik findet sich die Radial Basisfunktion $\mathrm{RBF}(x) := \exp\{-x^2\}$ wieder.

Biasknoten

Um die Aktivierungsfunktion horizontal verschieben zu können, wird ein Biasknoten mit $\mathrm{out}_{\mathrm{Bias}} := 1$, wie zuvor motiviert, eingeführt. Im Gegensatz zu den anderen Neuronen besitzt dieser Knoten keinen Input aus den Netzwerklayern, sondern feuert konstant den Wert 1 an die verbundenen Knotenpunkte. Um nun die Sprungstelle zu verschieben, beziehungsweise allgemein die Aktivierungsfunktionen entlang einer horizontalen Achse zu shiften, wird das Gewicht dieses Impulses analog zu sämtlichen anderen Outputs der Neuronen variabel definiert. Dieser Wert entspricht genau dem Offset und diese zusätzliche Kante unterliegt in sämtlichen Überlegungen und Herleitungen den gleichen Annahmen und Bedingungen wie die restlichen Gewichte. Daher wird dieser Biasknoten nicht explizit angeführt. Für interpretatorische Überlegungen kann dieses Neuron, insbesondere für den Outputlayer, analog zu Regressionsmodellen als eine Art Intercept aufgefasst werden.

Typische Anwendungen und andere Netzstrukturen

Bei den hier vorgestellten künstlichen neuronalen Netzwerken handelt es sich um eine Methode, einem Datenpaar (Input/Output) eine gewisse, deterministische Beziehung zu unterstellen, welche in weiterer Folge auf einen neuen Input-Datensatz mit vordefiniertem Format (ohne bekannten Output) für dessen Outputschätzung angewendet werden kann, um das Paar (Input/Output) zu vervollständigen beziehungsweise Letzteres zu schätzen.

Dieses zugrundeliegende Modell ermittelt sich bei gegebenen Modellparametern wie versteckten Neuronen oder Aktivierungsfunktionen ausschließlich mit Regeln,

welche aus den Optimierungsvorschriften, wie beispielsweise dem Lernalgorithmus Backpropagation (siehe nachfolgenden Abschnitt 6.2.2), entstammen und somit keine besonderen Eigenschaften an die Stichprobenverteilungen oder an die Ausgangsdaten gestellt werden müssen. Hierbei handelt es sich um die Modellierung eines funktionalen Zusammenhangs (vgl. 6.3) und kann gerade deswegen auf eine Vielzahl von Aufgabenfeldern angewendet werden.

Ein klassisches Beispiel hierfür ist die Mustererkennung, wie sie beim Auffinden von Regelmäßigkeiten in Textdokumenten oder der Erkennung von Sprachnachrichten sowie in Bildern auftreten (können). Weiterführende Informationen können beispielsweise in [26] oder [24] gefunden werden. So können bei geeigneter Wahl der Modellparameter die einzelnen Bestandteile der vorliegenden Daten mit hoher Zuverlässigkeit klassifiziert werden.

Eine Möglichkeit der Erweiterung der vorwärts gerichteten neuronalen Netzwerke, wie sie in dieser Arbeit angewendet werden, sind beispielsweise rekursive Netzwerke, indem die Netzwerkstruktur um Schleifen erweitert wird. Diese Konstruktion ermöglicht unter anderem die Addition von (zum Beispiel binären) Zahlen, da hier Überträge möglich sind. Während bei klassischen neuronalen Netzwerken stets ein Input in festgelegter Form angenommen wird, so ermöglichen die rekursiven (rekurrenten) Netzwerke auch Inputs unbestimmter Länge (siehe [26] Seite 44ff).

Während in der Definition 6.1 von vorhandenen Datenpaaren ausgegangen wird, spricht man von überwachtem Lernen, da für jedes Datenpaar ein Fehler mittels der Evaluierungsfunktion, wie in Gleichung (6.3), ermittelt wird. Aufgrund dessen findet eine gewisse Art von „Überwachung“ statt. Im Vergleich hierzu können auch neuronale Netzwerke ohne das Vorhandensein von Referenzwerten kalibriert werden. Hierbei spricht man von unüberwachtem Lernen, wie dieses unter anderem in [2], Seite 91ff beschrieben wird und dient beispielsweise der Modellierung von Datendichten oder zur Reduzierung der Datendimension und kann, auch wenn in dieser Arbeit nicht untersucht, eine Art Preprocessing für die Spektren darstellen.

Für weitere Informationen wird auf [2], [26], [14] sowie [24] als interessante Literatur verwiesen.

6.2.1 Mathematische Motivation

Für diesen Abschnitt und die mathematische Rechtfertigung, künstliche neuronale Netzwerke als Modell für die Abbildung des Zusammenhanges einer Prädiktormatrix X und den dazu korrespondierenden abhängigen Variablen Y zu verwenden, reicht es aus, neuronale Netzwerke der Tiefe 2 zu betrachten: bestehend aus einer Input-, einer versteckten und einer Output-Schicht mit Kardinalität (Anzahl Neuronen der Layer) $V_{\text{Input}} = p$, $V_{\text{Hidden}} = m$ und $V_{\text{Output}} = 1$. Das Resultat kann für Netzwerke mit beliebig vielen Outputneuronen und versteckten Layern verallgemeinert werden. Sämtliche Resultate dieses Unterabschnittes sind nachzulesen in [13].

Folgende Notation verwendet die Bezeichnungen i_l, $l = 1, \ldots, p$ für die Inputneuronen, m_l, $l = 1, \ldots, m$ für die Neuronen des versteckten Layers und o sei

das Outputneuron. Die Klasse der hier betrachteten Netzwerke mit m versteckten Neuronen kann mit der Darstellung aus den Gleichungen (6.3) zur Menge $\mathcal{R}^m(f)$, abhängig von der Transferfunktion im versteckten Layer und der Verwendung der linearen Transformation im Outputlayer, zusammengefasst werden.

$$\mathcal{R}^m(f) := \left\{ \nu : \mathbb{R}^n \to \mathbb{R} \;\middle/\; \nu(x) = \sum_{j=1}^{m} w_{m_j \to o} f\left(\sigma_j\right), \right.$$
$$\left. \text{wobei } \sigma_j := \sum_{l=1}^{p} w_{i_l \to h_j} x_{i_l} - \mu_j \right\}$$
$$\mathcal{R}(f) := \bigcup_{l \geq 1} \mathcal{R}^l(f)$$

mit x_{i_l} dem Knoten i_l zugeordnetem Inputwert, zusammengefasst werden. μ_j repräsentiert hier den Biasknoten.

Lemma 6.5 (Unbeschränkte Aktivierungsfunktion)**.** *Sei μ ein endliches Maß auf $\mathbb{R}^n$ und die Aktivierungsfunktion f unbeschränkt und nicht konstant. Dann liegt die Menge $\mathcal{R}(f)$ dicht in $L^p(\mu)$.*

Lemma 6.6 (Beschränkte Aktivierungsfunktion)**.** *Sei die Aktivierungsfunktion f beschränkt, stetig und nicht konstant. Dann liegt die Menge $\mathcal{R}(f)$ dicht in $C(X, \mathbb{R})$ für alle kompakten Mengen $X \subset \mathbb{R}^n$.*

Diese beiden Lemmata bedeuten, dass mit einer bereits einfachen Struktur von vorwärtsgerichteten Netzwerken unter milden Voraussetzungen alle L^p bzw. stetigen Funktionen, für m ausreichend groß, beliebig genau approximiert werden können. Aufgrund dieser Tatsache muss beachtet werden, dass, sofern statistische Messungen mit möglichen Messfehlern vorliegen, eine zu hohe Anzahl an Neuronen im versteckten Layer zu einem Overfitting führt.

6.2.2 Backpropagation

Bei der Rückwärtspropagierung (Backpropagation) handelt es sich um die Optimierung entlang des steilsten Abstieges. Der Name leitet sich davon ab, dass die Gewichte, beginnend beim letzten Layer, dem Outputlayer, rückwärts bis zum Inputlayer angepasst werden.

Ziel ist es, den Fehler der Schätzungen aller betrachteten Objekte nach Gleichung (6.3), welcher durch eine Evaluierungsfunktion $\mathcal{E}$ bestimmt wird, zu minimieren. Hierbei sind die zu berücksichtigenden Variablen die Gewichtsparameter W, deren Änderung den Fehler verkleinern soll. Dies führt auf ein Optimierungsproblem ohne Restriktionen. Ohne Beschränkung der Allgemeinheit soll die Evaluierungsfunktion derart definiert sein, sodass diese minimiert werden soll um den Fehler zu verringern.

Algorithmus zur Minimierung der Zielfunktion

Sei $\mathcal{E}\colon \mathbb{R}^k \to \mathbb{R}$ eine reellwertige Zielfunktion, welche einem unrestringierten Optimierungsproblem zugrunde liegt. Eine Möglichkeit, um diese Problemstellung zu lösen, ist jene des iterativen Gradientenabstieges. Hierbei wird eine Zahlenfolge $x_i \in \mathbb{R}^m, i \geq 0$ mit der Eigenschaft $\mathcal{E}(x_i) > \mathcal{E}(x_{i+1})\, \forall i \geq 0$ wie folgt ermittelt.

Sei $\mathcal{E}$ differenzierbar und $x \in \mathbb{R}^k$ mit $\nabla\mathcal{E}(x) \neq 0$, so folgt insbesondere für $\forall\alpha > 0$ mit $x_\alpha = x - \alpha\nabla\mathcal{E}(x)$ durch Betrachtung der Taylorentwicklung:

$$\begin{aligned}\mathcal{E}(x_\alpha) &= \mathcal{E}(x) + \nabla\mathcal{E}(x)' \cdot (x_\alpha - x) + o(\|x_\alpha - x\|) \\ &= \mathcal{E}(x) - \alpha\|\nabla\mathcal{E}(x)\|^2 + o(\alpha\|\nabla\mathcal{E}(x)\|) \\ &= \mathcal{E}(x) - \underbrace{\alpha\|\nabla\mathcal{E}(x)\|^2}_{①} + \underbrace{o(\alpha)}_{②}.\end{aligned} \tag{6.6}$$

Die formale Erklärung des Operators ∇ kann der Definition Ableitung, Gradient entnommen werden. Die verwendete Darstellung der Taylorreihe wird in Abschnitt A.1 erläutert.

Es gilt, dass für ausreichend kleine Werte von α der Term ② durch ① dominiert wird. Überträgt man dieses Konzept auf eine etwas allgemeinere Darstellung, so gilt analog für $x_\alpha = x + \alpha \cdot d$, wiederum für alle $\alpha \in \mathbb{R}_{>0}$ mit der zusätzlichen Eigenschaft $\nabla\mathcal{E}(x)'d < 0$, dass sich der Funktionswert $\mathcal{E}(x_\alpha)$ für ausreichend kleine Parameterwerte α im Vergleich zu $\mathcal{E}(x)$ verringert, wie analog zu Gleichung (6.6) aus (6.7) gefolgert werden kann.[4]

$$\mathcal{E}(x_\alpha) = f(x) + \underbrace{\alpha\nabla\mathcal{E}(x)'d}_{<0} + o(\alpha) \tag{6.7}$$

Auf diese Weise kann die anfangs beschriebene Folge von Punkten $x_i, i \geq 0$ zur iterativen Minimierung der Evaluierungsfunktion $\mathcal{E}$ erzeugt werden:

$$\begin{aligned}&x_{i+1} \coloneqq x_i + \alpha_i \cdot d_i, \\ &\text{mit } \alpha_i \in \mathbb{R}_{>0} \\ &\quad \nabla\mathcal{E}(x_i)' \cdot d_i < 0 \\ &\quad x_0 \text{ zufällig,}\end{aligned} \tag{6.8}$$

mit dem Abbruchkriterien $\nabla\mathcal{E}(x) = 0$ und es folgt $\mathcal{E}(x_{k+1}) < \mathcal{E}(x_k)$ für $k \geq 0$. Möglichkeiten für die Wahl der Schrittweiten α_i werden in Abschnitt Schrittweite in Gradientenverfahren vorgestellt.

[4] Für die Definition des Taylorpolynoms, sowie der klein-o Notation vgl. Abschnitt Taylorentwicklung und o-Notation.

Adaptierung des Algorithmus für neuronale Netzwerke

Diese Methode wird nun für den weit verbreiteten Lernalgorithmus für neuronale Netzwerke, die Backpropagation, adaptiert. Hierzu werden folgende Notationen eingeführt:

- Für das betrachtete Netzwerk seien die Knoten durchnummeriert, von 1 bis $|V|$.
- $y(x_i, W) \in \mathbb{R}^{1\times q}, i = 1, \ldots, n$ sei jene Funktion, welche die zum Objekt i gehörenden abhängigen Variablen schätzt. Die Realisierungen hiervon seien $\hat{y}_i \in \mathbb{R}^{1\times q}$. Die zu i gehörenden, tatsächlichen, Responses werden mit $y_i \in \mathbb{R}^{1\times q}$ bezeichnet und entsprechen der $i$ten Zeile der gegebenen Matrix Y.
- Sei $\mathcal{E}(W)$[5] die Evaluierungsfunktion. Für die hier betrachteten Anwendungsgebiete und im speziellen den Optimierungsalgorithmus für den Lernprozess wird die quadratische Abweichung als Bewertungsfunktion betrachtet. Konkret ist der MSE für die Datensätze $i = 1, \ldots, n$ definiert als

$$\mathcal{E}(W) := \sum_{i=1}^{n} \|y_i - y(x_i, W)\|_2^2. \tag{6.9}$$

- Analog zu pre_i sei $\text{nex}_i := \{j \in V : (i,j) \in E\}$ die Menge der Neuronen, an welche das Neuron i feuern kann. Für die hier betrachteten neuronalen Netzwerke mit $i \in V_s$, $s = 0, \ldots, k-1$: $\text{nex}_i = V_{s+1}$.

Da $\mathcal{E}$ in (6.9) als Summe des quadratischen Fehlers über allen Eingabedaten gebildet wird, reicht es an dieser Stelle aus, die partiellen Ableitung nach einem einzelnen Gewicht $w_{i\to j}$ mit $i \in \text{pre}_j$ für einen einzelnen Datensatz herzuleiten. Da sich die partielle Ableitung mit der endlichen Summe vertauschen lässt, kann die tatsächliche Änderung als Summe der partiellen Ableitungen ermittelt werden.

Um den zuvor skizzierten Gradientenalgorithmus anzuwenden, wird die partielle Ableitung von $\mathcal{E}$ unter Berücksichtigung der Kettenregel nach den Gewichten $w_{u\to j}$ mit $u \in \text{pre}_j$ ermittelt:

$$\frac{\partial \mathcal{E}(W)}{\partial w_{u\to j}} = \underbrace{\frac{\partial \mathcal{E}(W)}{\partial \text{out}_j(x_i, W)}}_{①} \cdot \underbrace{\frac{\partial \text{out}_j(x_i, W)}{\partial w_{u\to j}}}_{②}. \tag{6.10}$$

Einerseits wird die Evaluierungsfunktion nach dem Output des $j$ten Neurons in ① abgeleitet, und andererseits der Output nach dem tatsächlichen Gewicht in ②.

Diese beiden Punkte werden nun etwas differenzierter betrachtet. Um die Gleichung (6.11) übersichtlicher darstellen zu können und somit die Lesbarkeit zu

[5] Die Fehler hängen grundsätzlich von mehreren Faktoren und nicht nur von der Gewichtsmatrix W ab. Da für den Lernalgorithmus lediglich die Gewichte des Netzwerkes als Veränderliche dienen, welche zur Fehlerreduzierung erneuert werden können, wird aus notationstechnischen Gründen $\mathcal{E}(W)$ gewählt.

erhöhen, wird hier die verkürzte Notation $o_{j|i} := \text{out}_j(x_i, W)$ und dem Input $\sigma_{j|i} = \sum_{l \in pre(j)} w_{l \to j} o_{l|i}$ verwendet. Für den Multiplikanden (1) ergibt sich mit dieser Notation für $j \notin V_k (= V_{\text{Output}})$:

$$\begin{aligned}\frac{\partial \mathcal{E}(W)}{\partial o_{j|i}} &= \sum_{l \in \text{nex}_j} \frac{\partial \mathcal{E}(W)}{\partial o_{l|i}} \frac{\partial o_{l|i}}{\partial o_{j|i}} \\ &= \sum_{l \in \text{nex}_j} \frac{\partial \mathcal{E}(W)}{\partial o_{l|i}} \frac{\partial f_l(\sigma_{l|i})}{\partial \sigma_{l|i}} \frac{\partial \sigma_{l|i}}{\partial o_{j|i}} \\ &= \sum_{l \in \text{nex}_j} \frac{\partial \mathcal{E}(W)}{\partial o_{l|i}} \frac{\partial f_l(\sigma_{l|i})}{\partial \sigma_{l|i}} w_{j \to l}.\end{aligned} \tag{6.11}$$

Man beachte, dass die Kettenregel auf den Output der nachfolgenden Layerschicht angewendet wird, da sich eine Änderung der Gewichte auf den Output des nachfolgenden Layers auswirkt. Für $j \in V_k$ gilt

$$\frac{\partial \mathcal{E}(W)}{\partial o_{j|i}} = -2\left((y_i)_j - o_{j|i}\right) \tag{6.12}$$

mit $(y_i)_j$ dem Neuron j zugeordneten Eintrag des Zeilenvektors y_i.

In einem zweiten Schritt kann (2) vereinfacht werden zu

$$\begin{aligned}\frac{\partial o_{j|i}}{\partial w_{u \to j}} &= \frac{\partial f_j\left(\sigma_{j|i}\right)}{\partial \sigma_{j|i}} \frac{\partial \sigma_{j|i}}{\partial w_{u \to j}} \\ &= o_{u|i} \frac{\partial f_j(\sigma_{j|i})}{\partial \sigma_{j|i}}\end{aligned} \tag{6.13}$$

Die Gleichungen (6.11), (6.12) sowie (6.13) bilden die Grundlagen, um die Lernmethode anzuwenden. Es ergibt sich somit folgender Algorithmus:

Man beachte, dass diese Prozedur die Matrizen X und Y für Gewichtsupdates wiederholt verwendet, bis eines der Abbruchkriterien erfüllt ist.

Bemerkung 6.7. Sofern eine Sigmoid Transferfunktion verwendet wird, vereinfacht sich der Term (für die Lesbarkeit sei $\alpha := 1$):

$$\begin{aligned}\frac{\partial f\left(\sigma_{j|i}\right)}{\partial \sigma_{j|i}} &= \frac{\partial}{\partial \sigma_{j|i}} \left(1 + \exp\left\{-\sigma_{j|i}\right\}\right)^{-1} \\ &= f\left(\sigma_{j|i}\right)\left(1 - f\left(\sigma_{j|i}\right)\right) = o_{j|i}(1 - o_{j|i}).\end{aligned}$$

Algorithmus 5 Backpropagation

procedure BACKPROPAGATION
  Input: Netzwerk mit zufällig initialisierter Gewichtsmatrix W
  Wähle Schrittweite α
  while Abbruchkriterium nicht erfüllt (vgl. Abschnitt 6.2.3) **do**
    for $i = 1, \ldots, n$ **do**
      Berechne $\hat{y}_i$
      Berechne rekursiv $\beta_{j|i} = \frac{\partial \mathcal{E}(W)}{\partial o_{j|i}}, j \in V$ mit (6.11) und (6.12)
      Berechne für alle Gewichte $\Delta w_{(u \to j)|i} := \alpha \gamma_{j|i} \beta_{j|i}$
      mit $\gamma_{j|i} = \frac{\partial o_{j|i}}{\partial w_{u \to j}}$ laut (6.13).
    end for
    Summiere über alle Gewichts-Δs und erneuere die Gewichte nach Vorschrift (6.8).
  end while
  return Netzwerk mit erneuerter Gewichtsmatrix W
end procedure

Modifikation des Backpropagation Algorithmus

Um praktischen Ansprüchen gerecht zu werden, wird an dieser Stelle eine Modifikation des Optimierungsalgorithmus vorgestellt. Hierbei handelt es sich um die Anwendung der sogenannten Gauß-Newton Methode. Aufgrund der langsamen Konvergenz (vgl. [20]) und der in dieser Arbeit hohen Anzahl an simulierten neuronalen Netzwerken scheint dies unabdinglich.

Dieser Algorithmus ist speziell für die Minimierung des Kleinstequadrateschätzers konzipiert. Sei $\mathcal{E}\colon \mathbb{R}^q \to \mathbb{R}$ mit $x \mapsto \frac{1}{2}\|g(x)\|_2^2$ die zu minimierende Funktion.[6] Hierbei wird $g(x)$ durch

$$\tilde{g}(x, x_k) = g(x_k) + \nabla g(x_k)'(x - x_k)$$

als Taylorpolynom ersten Grades linearisiert. Um eine Sequenz für das Gradientenverfahren herzuleiten, wird die Folge $\{x_k\}$ mit der rekursiven Darstellung in (6.14)

$$\begin{aligned} x_{k+1} &= \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|\tilde{g}(x, x_k)\|_2^2 \\ &= \arg\min_{x \in \mathbb{R}^n} \|g(x_k)\|_2^2 + 2(x - x_k)'\nabla g(x_k) g(x_k) \\ &\qquad + \left\|\nabla g(x_k)'(x - x_k)\right\|_2^2 \end{aligned} \tag{6.14}$$

[6] Der Vorfaktor von 2^{-1} dient lediglich zur Vereinfachung der Notation in der Herleitung. $g(x)$ sei wie in Gleichung (6.9).

und einem beliebigen Startwert x_0 mit $\nabla g(x_0) \neq 0$ erzeugt. Unter der Annahme der Invertierbarkeit der Matrix $G_k := \nabla g(x_k)\nabla g(x_k)'$ führt (6.14) auf die Rechnungsvorschrift (6.15).

$$x_{k+1} = x_k - G_k^{-1}\nabla g(x_k)g(x_k). \tag{6.15}$$

Da die Invertierbarkeit von G_k in der Praxis nicht immer garantiert ist, wird das Gauss-Newton Verfahren dahingehend modifiziert, dass G_k durch die Matrix $\widetilde{G}_k := G_k + \mu_k \cdot I_q$ mit einer geschickten Wahl von μ_k substituiert wird, wobei I_q der q-dimensionalen Einheitsmatrix entspricht. Diese Abänderung ist wie folgt motiviert:

Seien hierzu $\lambda_i, i = 1, \ldots, q$ die Eigenwerte von G_k, sowie $z_i, i = 1, \ldots, q$ die dazugehörenden Eigenvektoren. So gilt $\forall i = 1, \ldots, q$:

$$(G_k + \mu_k \cdot I_q)z_i = G_k z_i + \mu_k z_i = \lambda_i z_i + \mu_k z_i = (\lambda_i + \mu_k)z_i$$

Die Wahl von μ_k mit der Eigenschaft $\lambda_i + \mu_k > 0 \; \forall i$ stellt sicher, dass die modifizierte Matrix $\widetilde{G}_k$ nicht singulär ist und daraus resultiert das sogenannte Levenberg-Marquardt Verfahren mit Schrittweite $\alpha_k = 1$. Für alternative Schrittweiten wird $\widetilde{G}_k^{-1}$ mit dem Vorfaktor α_k versehen.

Die Levenberg-Marquardt Methode stellt somit eine Kombination folgender Punkte dar:

- Gauß-Newton. Für μ_k sehr klein wird der Levenberg-Marquardt Algorithmus zum Gauß-Newton Verfahren.
- Steilster Abstieg. Mit großen Werten von μ_k ist dies eine Approximation des Algorithmus des steilsten Abstieges, wie anfangs in Abschnitt 6.2.2 beschrieben:

$$x_{k+1} \approx x_k - \frac{1}{\mu_k}\nabla g(x_k)g(x_k)$$

Um die Levenberg-Marquardt Prozedur auf die Problemstellung der Bestimmung der Gewichte für ein neuronales Netzwerk anzuwenden, wird mit μ_k sehr klein begonnen. Falls durch den Algorithmus eine verbesserte Lösung erreicht werden kann, wird die adaptierte Lernrate $\mu_k \leftarrow \mu_k \cdot \eta$ mit einem multiplikativen Faktor $\eta > 1$ angepasst. Sofern sich an dieser Stelle der Zielfunktionswert verschlechtert, wird für den nächsten Schritt μ_k durch den Faktor η dividiert, um unter anderem die Konvergenzgeschwindigkeit zu erhöhen (vgl. hierzu [20]).

Da die Thematik der Optimierung der Backpropagation nicht den Hauptbestandteil dieser Arbeit darstellt, wird an dieser Stelle auf weiterführende Literatur zu dieser Modifikation verwiesen ([21], [20]).

6.2.3 Modellierung

Als mögliche Modelle werden künstliche neuronale Netzwerke der Tiefe 2 betrachtet, bestehend aus je einem Input-, einem versteckten und einem Outputlayer und begründet sich einerseits durch die Populationsgröße von lediglich $n_{2016} = 81$ Daten, bzw. durch die beiden Lemmata in Abschnitt 6.2.1.

Transformation der Daten

Also zusätzliches Preprocessing werden die Daten für die Modellentwicklung auf den Wertebereich $[-1, 1]$ transformiert. Dies bedeutet, dass die einzelnen Datensätze der Prädiktormatrix $X^{n\times s}$, sowie die dazugehörenden Referenzwerte $Y \in \mathbb{R}^{n\times q}$ jeweils auf diesen Bereich transformiert werden, und die Netzwerke somit versuchen, einen funktionalen Zusammenhang $X \rightsquigarrow [-1, 1]^{n\times s} \to [-1, 1]^{n\times q} \rightsquigarrow Y$ zu erzeugen. Die Schätzungen werden dann wiederum in den tatsächlichen Wertebereich rücktransformiert. Dies ist auch der Grund, weshalb bei den vorliegenden Daten mit wenigen Ausnahmen keine negativen Konzentrationen geschätzt werden.

Abbruchkriterium

Ein wichtiger Bestandteil des Lernalgorithmus bildet das Abbruchkriterium, wie in Algorithmus 5, angeführt. Ohne eine sinnvolle Vorschrift resultiert der Lernalgorithmus in einer Endlosschleife. Einerseits kann ein neuronales Netzwerk so lange trainiert werden, bis eine vorgegebene Genauigkeit erfüllt wird. Dies führt allerdings in der Regel zu dem unerwünschten Effekt des Overfittings. Daher wird in dieser Arbeit wie folgt vorgegangen:

- Teile die zur Verfügung stehenden Daten in zwei (zufällige) Partitionen, den Trainings- und den Validierungsdatensatz.
- Modelliere das Netzwerk mit den Trainingsdaten so lange, bis sich keine Verbesserung im Validierungsdatensatz mehr zeigt. Um nicht in lokalen Minima festzustecken, wird hier ein Parameter u eingeführt, welcher angibt, wie viele Lernschritte mit einer Verschlechterung im Validierungsdatensatz in Kauf genommen werden dürfen. Führt dies nach diesen u Schritten zu keiner neuerlichen Verbesserung, so terminiert der Algorithmus mit entsprechenden Schätzungen und Fehlerkennzahlen.
- Um Kennzahlen aufgrund von äußerst (un)günstigen Partitionierungen zu reduzieren, wird diese Prozedur R[7] mal wiederholt und die R Schätzungen gemittelt. Hieraus resultiert der endgültige Fehler.

Dieses Vorgehen ist mit der Monte Carlo Kreuzvalidierung vergleichbar. Offen bleibt die Frage, was eine „Verbesserung“ bedeutet. Hierbei wurde der MSE als entscheidende Kennzahl betrachtet.

[7] Aus laufzeittechnischen Gründen wurde von anfangs $R = 100$ auf $R = 30$ reduziert.

Weitere Parameter der neuronalen Netzwerke

Hierbei werden sämtliche Modelle für $|V_1| \in \{1, \ldots, 8\}$[8] versteckte Neuronen, die die Hyperbolic Tangent Sigmoid Aktivierungsfunktion, wie aus Gleichung (6.5) mit Parameter $\alpha = 2$, sowie einer linearen Aktivierungsfunktion des Outputlayers verwenden, betrachtet.

Da die tatsächlichen Aktivierungsfunktionen aus theoretischen Überlegungen nur eine untergeordnete Rolle spielen, wird dieses Standardmodell auf alle Responsevariablen angewendet. Hierfür werden unterschiedliche Wellenzahlbereiche selektiert und in weiterer Folge bewertet.

Insbesondere für Glukose stehen weitere Modelle zur Verfügung, welche von obigen, grundlegenden Einstellungen für neuronale Netzwerke abweichen.

- Ein Glukosemodell wird für eine Aktivierungsfunktion des versteckten Layers der Form $\mathcal{RBF}(x) = \exp\{\{-x^2\}\}$, der $\mathcal{TS}$ entwickelt.
- Weiters wurde insbesondere für Glukose die Aktivierungsfunktion im Outputlayer auf eine $\mathcal{TS}$ geändert.
- Zuletzt wird ein Glukosemodell mit einer S Aktivierungsfunktion (Gleichung (6.4)), sowohl im versteckten, als auch im Outputlayer kalibriert.
- Es wird versucht, die Transformation der Werte in einem Glukosemodell auf das Intervall $[-1, 1]$ (sowohl von X als auch von Y) zu vernachlässigen.
- Für Ethanolwerte wird ebenfalls versucht, ein Modell mit einer Aktivierungsfunktion des versteckten Layers mit einer $\mathcal{LS}$ zu entwickeln.
- Für Ethanol wird zusätzlich der Bereich $[8, 16]$ Vol.% untersucht, während für die Glukosekonzentrationen Modelle mit dem Wertebereich $[0, 10]$ g/l entwickelt werden.
- Eine Modellierung für Datensätze unter Ausschluss der 0-Werte führte nach anfänglichen Untersuchungen auf keine akzeptablen Ergebnisse und wird daher nicht weiter betrachtet.

Grundsätzlich führten lediglich die Standardkalibrationen zum Erfolg.

[8] Aufgrund von langen Laufzeiten, wie aus Tabelle 5 ersichtlich, werden jedoch nur die Wellenzahlbereiche, welche mit $m \leq 4$ versteckten Neuronen ermittelt werden, mit der doppelten Kreuzvalidierung vollständig analysiert.

7 Auswertungen mit den künstlichen neuronalen Netzwerken

In diesem Abschnitt werden die neuronalen Netzwerke, wie in Kapitel 6 definiert, auf insgesamt 11 mögliche Bestandteile von Weinen, sowie die Dichte und den pH-Wert der untersuchten Weine angewendet und versucht, Modelle von hoher Güte zu kalibrieren. Im Wesentlichen sind die Auswertungen analog zu Abschnitt 5 gegliedert.

Wiederum beginnend mit der Präsentation eines ersten Modells für den Alkoholgehalt, sowie für einen reduzierten Teilbereich des gesamten Wertebereiches, um ein Grundverständnis für die Anwendung der neuronalen Netzwerke zu erhalten. Wie in jedem dieser 13 Unterabschnitte wird zusätzlich zum Modell die Reproduzierbarkeit untersucht, sowie die unterschiedlichen Daten der Spektrometer E22, E24 und V70(2016) mit dem zur Kalibrierung verwendeten Datensatz der Engine E25 verglichen, um etwaige Messungenauigkeiten oder Indizien für eine bessere/schlechtere Datenqualität zu identifizieren.

Nachdem ein erster Eindruck über die Praktikabilität der neuronalen Netzwerke gegeben wurde, wird versucht, ein Modell für Extrakt, dem Summenparameter für (unter anderem) die beiden Zuckerarten Glukose und Fruktose zu entwickeln, bevor für ebendiese Bestandteile eigene Modelle betrachtet werden.

Als nächsten inhaltlichen Schwerpunkt werden die unterschiedlichen Säuren, beginnend mit dem Summenparameter titrierbaren Säure, untersucht. In weiterer Folge werden fünf unterschiedliche Säuren, absteigend in deren Wichtigkeit, modelliert. Als Hierarchie versteht diese Arbeit den Anteil im Sinne der Größe des zugrundeliegenden Wertebereiches im Datensatz des Jahres 2016, welchen diese zur Gesamtsäure beitragen. Beginnend mit der Weinsäure und der L-Äpfelsäure, gefolgt von der Milchsäure, wobei an dieser Stelle nicht zwischen L- und R-Milchsäure differenziert wird, über die flüchtigen Säuren bis hin zur Zitronensäure, welche als einigermaßen preisgünstige Säure den geringsten Teil an den titrierbaren Säuren ausmacht.

Zur Abrundung des Kapitels wird zudem der Bestandteil Glyzerin und die beiden Eigenschaften Dichte und pH-Wert modelliert und analysiert.

In einem letzten, abschließenden Abschnitt werden die wichtigsten Erkenntnisse zusammengefasst und eine Übersicht über diese Modellkennzahlen gegeben.

7.1 Ethanol

Bereits mit einer Anzahl von 9 Prädiktoren und einem versteckten Neuron kann ein erstes, einigermaßen plausibles Modell für die Schätzung des Alkoholgehaltes kalibriert werden. Hierbei werden 3×3 Wellenzahlen im Bereich der Bande bei $2\,950\,\mathrm{cm}^{-1}$ selektiert. Bei einer Verwendung der zweiten Savitzky-Golay Ableitung entsprechen die selektierten Bereiche somit zwei Senken und dem Anstieg zum letzten Peak des zugelassenen Wertebereiches der Wellenzahlen, wie in Abbildung 7.1 (li.) dargestellt. Mit einem Richtwert für den Testfehler von 0.3139 Vol.% erscheint dieser Wert auf einen ersten Blick einigermaßen hoch. Dies kann allerdings, wie im Abschnitt der PLS-Modellierung damit begründet werden, dass jeweils die alkoholfreien Weine diese Kennzahl maßgeblich beeinflussen, wie die Residuen der doppelten Kreuzvalidierung in Abbildung 7.1 (re.) bestätigen, da, um einen Wein zu schätzen, dieser in dieser Grafik nicht im Kalibrierungsset enthalten ist. Aufgrund dessen befindet sich somit maximal ein alkoholfreier Wein im Kalibrierungsset, welcher entweder im Trainings- oder Validierungsdatensatz oder in keinem der beiden verwendet wird. Dadurch verliert das Modell eine Vielzahl an Information, insbesondere für die Daten am Rande des verfügbaren Wertebereiches. Darüber hinaus zeigt sich, dass kein Modell adäquate Schätzwerte für Wassermessungen liefert, und somit ein weiteres Indiz gegeben ist, weshalb die Modelle für alkoholfreie Weine derartige Überschätzungen liefern.

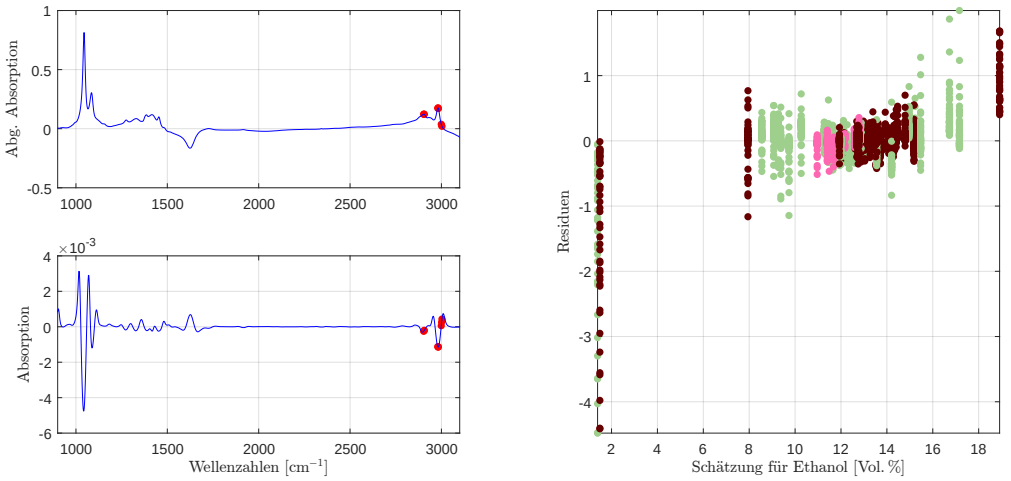

Abbildung 7.1: Die selektierten Wellenzahlen im NN-Modell für Ethanol mit der zweiten Savitzky-Golay Ableitung (li.) und der aus der doppelten Kreuzvalidierung resultierende Residuenplot (re.).

Der klassische Residuenplot findet sich in Abbildung 7.2 (li.). Die große Variabilität der alkoholfreien Weine, auch wenn diese in dieser Grafik im Trainingsdatensatz verwendet wurden, spiegelt sich in der enormen Überschätzung der alkoholfreien Weine wider. Des Weiteren kann ein Ausreißer nach oben, und somit eine starke Unterschätzung des tatsächlichen Ethanolgehaltes für diesen Rotwein beobachtet werden. Hierbei handelt es sich um einen Portwein mit einem Alkoholgehalt von

beinahe 20 Vol.%. Analog zu den alkoholarmen Weinen kann dieses große Residuum auf den Mangel an Informationen außerhalb eines Bereiches von $[8, 16]$ Vol.% zurückgeführt werden. Im Vergleich zu dem Modell aus Abschnitt 5.1 zeigt sich bei dieser Kalibrierung des neuronalen Netzwerkes eine enorme Sensibilität für Datenpunkte am Rande des Wertebereiches. Dies führt daher unweigerlich zur Frage nach dem Verhalten des Modells ohne diese Ausreißer. In einem ersten Schritt werden aus Abbildung 7.2 (li.) lediglich die Weine des Wertebereiches $[8, 16]$ Vol.% extrahiert (keine Neukalibrierung des Modells) und es resultiert der Residuenplot in Abbildung 7.2 (re.)[1]. Mit der Einschränkung auf ebendiesen Wertebereich reduziert sich die Standardabweichung der Residuen auf unter 0.1 Vol.%, liegt allerdings trotzdem über jener der Referenzmethode. Es treten keine Residuenmuster oder Ausreißer, abgesehen von einer größeren Überschätzung des Alkoholgehaltes eines Weißweines, auf. Ebenso können auch bei Betrachtung dieses Ausschnittes keine Unregelmäßigkeiten für das selektierte Modell beobachtet werden.

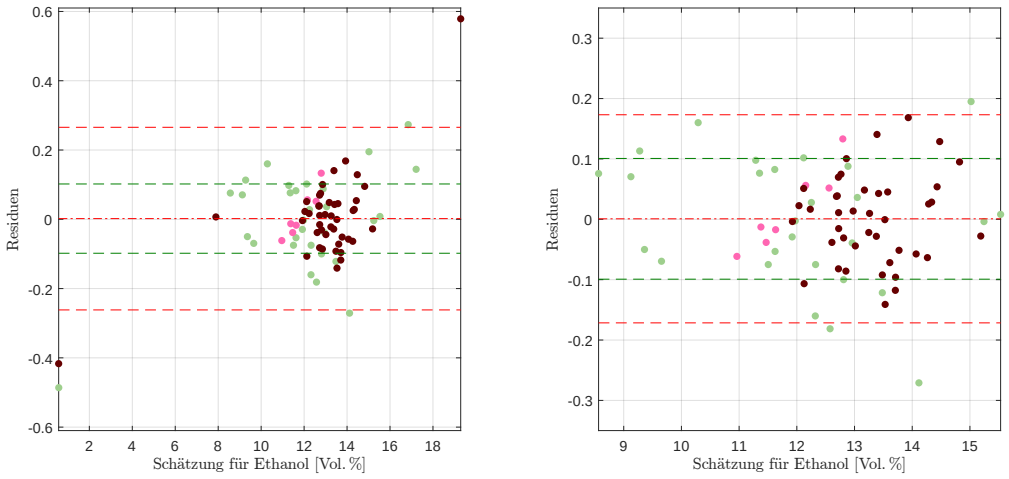

Abbildung 7.2: Residuenplot des NN-Modells für Ethanol mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.), sowie der vergrößerte Ausschnitt $[8, 16]$ Vol.% mit Bias und Abweichungen für diesem Bereich.

Insgesamt beläuft sich die Standardabweichung der Residuen im Gesamtmodell auf 0.1318 Vol.%. Betrachtet man wie in Abbildung 7.2 jenen Ausschnitt mit einer konzentrierten Punktmasse, so reduziert sich die Standardabweichung auf 0.0862 Vol.% mit einem jeweils marginalen und vernachlässigbaren Bias. Die Residuen befinden sich allesamt in einem Intervall von $[-0.50, 0.60]$ Vol.%, bedingt durch die Ausreißer und der IQR beläuft sich auf 0.13 Vol.%. Wie aus den Residuenplots ersichtlich, kann mit diesem Modell die Datensituation einigermaßen plausibel abgebildet werden. Die erklärte Varianz beläuft sich hierbei auf über 99% bei einem MSE von $0.017(\text{ Vol.\%})^2$.

[1] Man betrachte hierbei die unterschiedliche Skalierung. Während die Residuen im Gesamtmodell in einem Bereich von $[-0.6, 0.6]$ Vol.% dargestellt sind, so zeigt die rechte Abbildung den Ausschnitt $[-0.35, 0.35]$ Vol.%.

Betrachtet man den Ethanolbereich von $[8, 16]$ Vol.%, indem man alkoholarme und äußerst alkoholstarke Weine wie Portweine exkludiert, so weisen diese Residuen im Gesamtmodell, wie zuvor erwähnt, eine Standardabweichung von 0.0862 Vol.% auf. In einem nächsten Schritt soll hierfür ein eigenes Modell entwickelt werden, um die Fehler weiter zu minimieren. Selektiert man die Wellenzahlen für ebendiesen Bereich, so resultiert ein Modell mit 6 Wellenzahlen mit denselben Parametern wie Savitzky-Golay Ableitung, Aktivierungsfunktionen und versteckten Neuronen. Die Residuen des Submodells weisen keine signifikant verringerte Standardabweichung auf (0.0827 Vol.%), noch reduziert sich das Intervall, in welchem sich diese befinden, wie der Überblick in Tabelle 7.1 bestätigt. Abbildung 7.3 visualisiert die Verbesserung bei einer spezifischeren Wahl für den reduzierten Bereich. Während in Abbildung 7.3 (li.) ein vergleichsweise starker Ausreißer nach unten beobachtbar ist, so wird in Abbildung 7.3 (re.) ein Weißwein in gleichem Maße unterschätzt. In diesem reduzierten Modell kann eine Überschätzung aller 4 (Weiß-)Weine, welche eine Ethanolkonzentration von weniger als 10 Vol.% aufweisen, beobachtet werden. Da es sich lediglich um 5 von 75 in dem Intervall $[8, 10]$ Vol.% Weinproben handelt, kann an dieser Stelle nicht von einer Struktur im Modell gesprochen werden, da für eine derartige Aussage in diesem Bereich zu wenige Daten zur Verfügung stehen. Zusätzlich kann ein Weißweinausreißer nach oben beobachtet werden, welcher im Gesamtmodell nicht auffällig ist. Hierbei handelt es sich um einen Süßwein mit einem Extraktwert von $256\,\mathrm{g/l}$[2]. Vergleicht man den durch die Referenzmethode ermittelten Ethanolwert für Weine mit hohen Extraktwerten wie beispielsweise $\geq 150\,\mathrm{g/l}$, so weisen alle diese Weine einen um mindestens 3 Vol.% oder einen um 1 Vol.% bis 1.5 Vol.% geringeren Ethanolgehalt auf. Dies stellt somit eine mögliche Begründung für den Ausreißer im reduzierten Modell dar.

In Tabelle 7.1 sind diese Ergebnisse zusammengefasst. Keines der drei Modelle weist einen relevanten Offset auf und starke Ausreißer können lediglich bei dem Gesamtmodell in dem Portwein und den alkoholarmen Weinen festgestellt werden. Insgesamt erscheint eine spezifische Modellentwicklung mit einem reduzierten Wertebereich wie $[8, 16]$ Vol.% nicht notwendig, da die Ergebnisse nicht signifikant verbessert werden können und der Bereich der minimalen und maximalen Residuen lediglich nach oben geshiftet wird.

	Anz.	Mw.	Std.	Min.	Med.	Max.	IQR	MSE
Gesamtmodell Ethanol	81	0.00	0.1318	−0.49	0.01	0.58	0.13	0.0172
für $[8, 16]$ Vol.%	75	0.00	0.0862	−0.27	−0.00	0.20	0.11	0.0073
Modell für $[8, 16]$ Vol.%	75	0.00	0.0827	−0.17	−0.01	0.31	0.12	0.0067

Tabelle 7.1: Modellvergleich des NN-Modells für Ethanol mit jenem für den Wertebereich $[8, 16]$ Vol.% eigens entwickelten Modell. Alle Werte in Vol.%.

[2] Dies entspricht dem zweithöchsten Extraktwert im vorliegenden Datensatz.

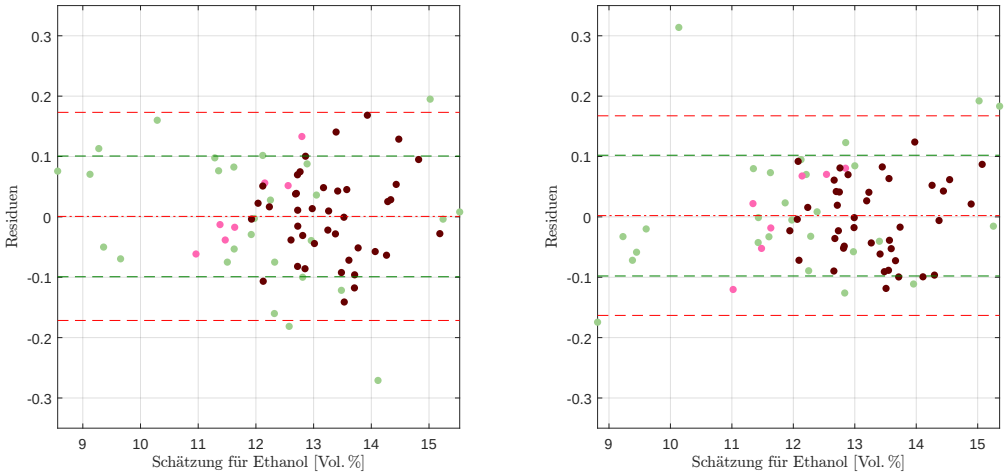

Abbildung 7.3: Extrahierte Residuen für den Wertebereich [8, 16] Vol.% des Residuenplots des NN-Modells für Ethanol mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.), sowie der Residuenplot des wertebereichspezifischen NN-Modells mit den entsprechenden Kennzahlen (re.).

Reproduzierbarkeit und Datensatzvergleich

Betrachtet man in einem ersten Schritt die Grafiken in Abbildung 7.4, so können im Gegensatz zum Verlauf der Residuen in Abhängigkeit der Messreihenfolge keine Auffälligkeiten bezüglich der Reproduzierbarkeit festgestellt werden und man erkennt eine einigermaßen homogene Wiederholbarkeit. Lediglich zwei Ausreißer (nach oben) existieren bei den Weißweinen. Hierbei handelt es sich um jene zwei Weine mit der ID 14 bzw. 19, welche einerseits in sämtlichen bestimmten Inhaltsstoffen vergleichbar sind, aber keine Unregelmäßigkeiten gegenüber den anderen Proben beobachtet werden können. Ansonsten zeigen die Boxplots der Standardabweichungen der Messungen von den einzelnen Weinproben ein ähnliches Verhalten, insbesondere jene mit den meisten vorliegenden Proben, den Rot- und Weißweinen.

Betrachtet man neben den beiden Kennzahlen der Standardabweichung zusätzlich die maximale Abweichung, so zeigt sich erwartungsgemäß ein ähnliches Verhalten. Mit einer maximalen Abweichung von 0.0639 Vol.% weisen jedoch die Rotweine den hierfür geringsten Wert auf, was sich wiederum in der mittleren maximalen Abweichung widerspiegelt. Das Modell der neuronalen Netzwerke für sämtliche Daten resultiert somit in einer Standardabweichung der größten Differenzen der wiederholten Analysen von weniger als 0.018 Vol.%.

Wendet man das Gesamtmodell auf alle Datensätze des Jahres 2016 an, so zeigt sich mit Ausnahme der deutlichen Verschlechterung der Reproduzierbarkeit und der leicht höheren Standardabweichung bei Engine E24 innerhalb eines jeden Datensatzes eine ähnliche Güte der Residuen. Da das für E25 optimale Modell

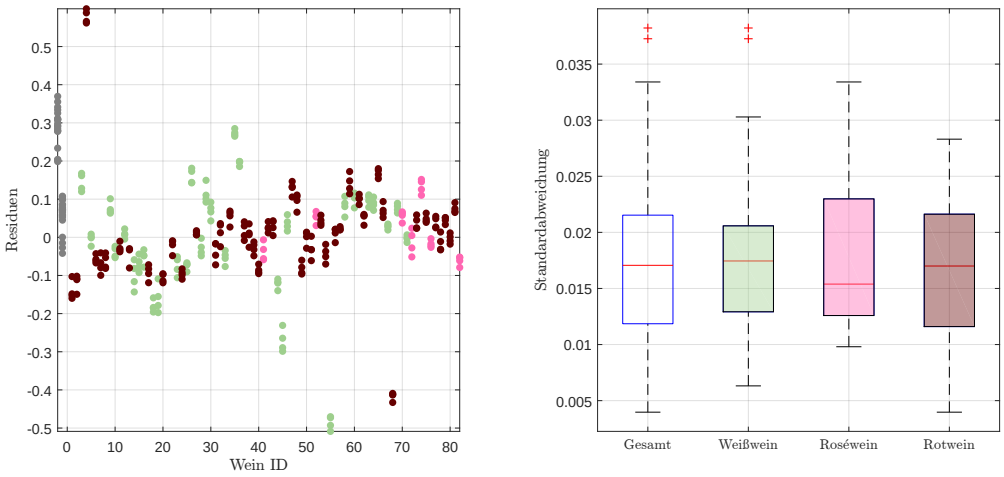

Abbildung 7.4: Residuen in Vol.% der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung des Ethanolgehaltes im NN-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0180	0.0175	0.0885	0.0407	0.0384	0.0179
Roséwein	0.0182	0.0154	0.0751	0.0406	0.0369	0.0181
Rotwein	0.0169	0.0170	0.0639	0.0376	0.0367	0.0137
Gesamt	0.0174	0.0171	0.0885	0.0390	0.0377	0.0156

Tabelle 7.2: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Ethanol im NN-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in Vol.%.

mit dem Datensatz von E22 minimal bessere Kennzahlen der Residuen produziert, zeigt sich ein konträres Bild bei der Reproduzierbarkeit. So ergeben sich sowohl geringere Kennzahlen bei der Betrachtung des Mittelwertes, des Medians, als auch der maximalen Abweichung der einzelnen Messungen der Weine und insgesamt die beste Reproduzierbarkeit. Auffallend ist ebenfalls die mehr als doppelt so hohe maximale Abweichung im Datensatz E24.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.1276	0.13	−0.53	0.55	0.0205	0.0203	0.0938
E24	0.1306	0.13	−0.49	0.50	0.0232	0.0221	0.1802
E25	0.1318	0.13	−0.49	0.58	0.0174	0.0171	0.0885
V70(2016)	0.1410	0.11	−0.56	0.57	0.0209	0.0188	0.1155

Tabelle 7.3: Performance des entwickelten NN-Modells für Ethanol, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in Vol.%.

7.2 Extrakt

Das für Extrakt ermittelte, beste Modell, welches den gesamten Wertebereich abzudecken versucht, resultiert in einer Prädiktorenwahl der drei aufeinanderfolgenden Wellenzahlen 2810 cm^{-1}, 2808 cm^{-1} sowie 2806 cm^{-1} mit einem versteckten Neuron. Dies bedeutet, dass das Modell in diesen drei Messstellen den maximalen Informationsgehalt vermutet (bei einer Glättung mit der ersten Savitzky-Golay Ableitung) und durch Hinzunahme von weiteren Wellenzahlen keine Verbesserung erzielt werden kann. Konkret bedeutet dies, dass durch die Hinzunahme von drei weiteren Prädiktoren die Variabilität in den Residuen nicht reduziert wird.

Bei diesem Modell realisieren die Residuen in einem Intervall von $[-6.9, 7.9]$ g/l bei einem interquartilen Bereich von 3.6 g/l und einer Standardabweichung von 2.6897 g/l und liefert, verglichen mit dem Pendant aus Abschnitt 5.2 keine plausiblen Ergebnisse, insbesondere das Verhalten der Reproduzierbarkeitskennzahlen, ebenso wie beim Vergleich mit der zertifizierten Messmethode. Diese breite Streuung kann nicht damit erklärt werden, dass ein steigender Extraktwert eine womöglich größere Variabiltät impliziert, sondern bereits die Werte von bis zu 30 g/l mit einer beobachteten Standardabweichung von über 2 g/l beeinflussen die Gesamtvariabilität enorm. Auch wenn die Residuenplots, abgesehen von der enormen Streubreite keine Auffälligkeiten zeigen, wird an dieser Stelle auf eine detailliertere Modellanalyse mit Verweis auf das PLS-Modell verzichtet.

Selbst bei der Einschränkung des Wertebereichs, indem der Bereich mit hoher Extraktkonzentration aufgrund der weiten Bandbreite an Referenzwerten nach oben beschränkt wird, können mit der hier vorgestellten Methode keine zufriedenstellenden Modelle entwickelt werden. Dies zeigt sich in einer, mit Ausnahme

der heuristischen Charakteristik der Suchmethode, nicht für die gegebenen Daten erklärbaren, nicht plausiblen Variabilität und zusätzlich einem augenscheinlichen Trend der Residuen.

Dieser beschriebene Trend zeigt sich beispielsweise bei der Modellbildung für Extraktkonzentrationen bis zu einer Konzentration von circa 63.5 g/l, wie in Abbildung 7.5 (re.) visualisiert. Hierbei weisen die Residuen eine Standardabweichung von vergleichsweise niedrigen 0.6805 g/l auf, und resultieren bei 30 selektierten Wellenzahlwerten in ebendieser Abbildung.

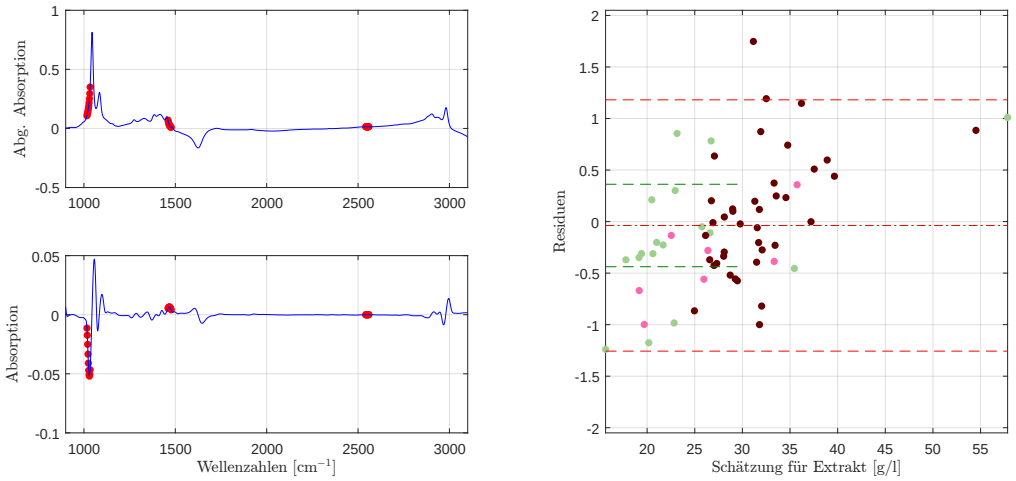

Abbildung 7.5: Die selektierten Wellenzahlen im für $[0.0, 63.5]$ g/l wertebereichspezifischen NN-Modell für Extrakt mit der ersten Savitzky-Golay Ableitung (li.) und der dazugehörende Residuenplot mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (re.).

Aufgrund der offensichtlichen Struktur in den Residuen scheint auch dieses Submodell für die Extraktkonzentration nicht brauchbar, weshalb eine weitere, im Speziellen auch grafische Analyse, insbesondere jene der Reproduzierbarkeit, an dieser Stelle entfällt. Auch zeigt sich mit der hier gewählten Parametrisierung in den anderen zur Verfügung stehenden Datensätzen keine Verbesserung.

Bemerkung 7.1. Der Trend zeigt sich insbesondere bei Betrachtung der einzelnen Weinfarben in Abbildung 7.5 (re.). Selbst bei ausschließlicher Betrachtung der Rotweine (oder Weißweine), kann kein trendfreies Modell gefunden werden.

Bemerkung 7.2. Analysiert man dieses Modell mit dem Datensatz E22, so reduziert sich der Trend etwas, auch wenn dies keine Modellbildung mit dieser Parametrisierung der neuronalen Netzwerke für Extraktkonzentrationen rechtfertigt.

7.3 Glukose

Ein wesentlicher Bestandteil des Extraktwertes, welcher in vorherigem Abschnitt 7.2 diskutiert und unzureichend modelliert werden konnte, bildet Glukose. Das hier selektierte Modell für Glukose verwendet insgesamt 20 Wellenzahlen, aufgeteilt in 4×5 jeweils zusammenhängende Teilbereiche. Die Hälfte dieser Prädiktoren befindet sich in der Abflachung der Bande mit dem höchsten Peak in der Region von $1\,120\,\mathrm{cm}^{-1}$, welcher nach dem Preprocessing in der zweiten Savitzky-Golay Ableitung direkt auf die doppelte Oszillation folgt und einem (kleinen) Peak in ebendiesem Wellenzahlbereich entspricht. Die restlichen beiden Blöcke mit je 5 Wellenzahlen befinden sich im Bereich von $2\,780\,\mathrm{cm}^{-1}$ und $2\,980\,\mathrm{cm}^{-1}$, wie in Abbildung 7.6 (li.) dargestellt. Hierfür wurden zwei versteckte Neuronen in Zusammenhang mit einer Transferfunktion der Sigmoid-Klasse wie in Gleichung (6.4) verwendet. Der primäre Grund für die Änderung der Aktivierungsfunktion liegt in der Tatsache begründet, dass mit der Tangens-Sigmoidfunktion aus Gleichung (6.5) ein Modell mit 9 Wellenzahlen selektiert wurde, welches einen auftretenden Trend in den Residuen verstärkt.

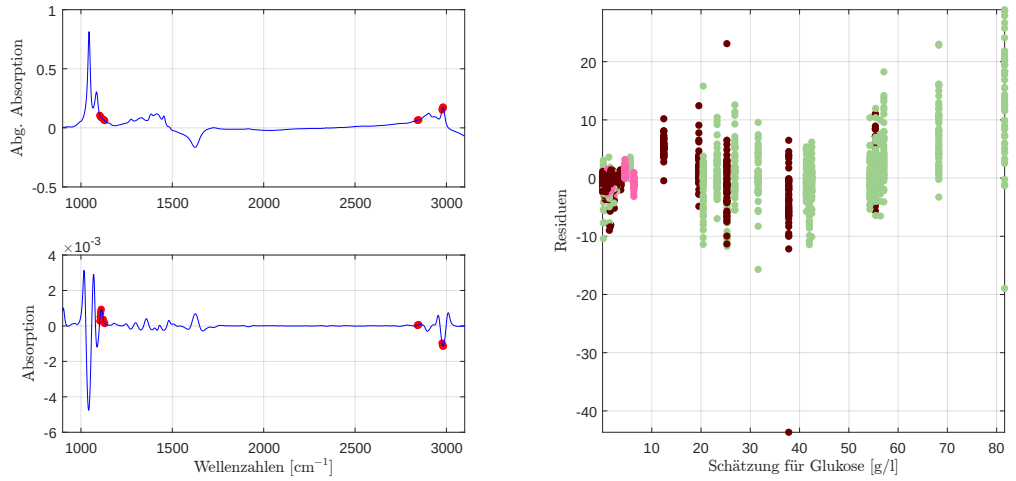

Abbildung 7.6: Die selektierten Wellenzahlen im NN-Modell für Glukose mit der zweiten Savitzky-Golay Ableitung (li.) und der aus der doppelten Kreuzvalidierung resultierende Residuenplot (re.).

Betrachtet man den Residuenplot aus der doppelten Kreuzvalidierung in Abbildung 7.6 (re.), so können geringe Streuungen im niedrigen Wellenzahlbereich festgestellt werden, wohingegen mit Zunahme der Glukoseschätzung die Variabilität mehr zu streuen scheint. Dies kann wiederum auf die Datenkonzentration im Bereich von weniger als $10\,\mathrm{g/l}$ zurückgeführt werden. Darüber hinaus scheinen die Schätzungen in einem Bereich von $[10, 40]\,\mathrm{g/l}$ leicht zu fallen, wohingegen in dem daran anschließenden Wertebereich $(40, 90]\,\mathrm{g/l}$ die Residuen zu steigen beginnen.

Im klassischen Residuenplot in Abbildung 7.7 (li.) können zwei Ausreißer beobachtet werden, je ein Weiß- und ein Rotwein. Auch in dieser Grafik zeigt sich ein

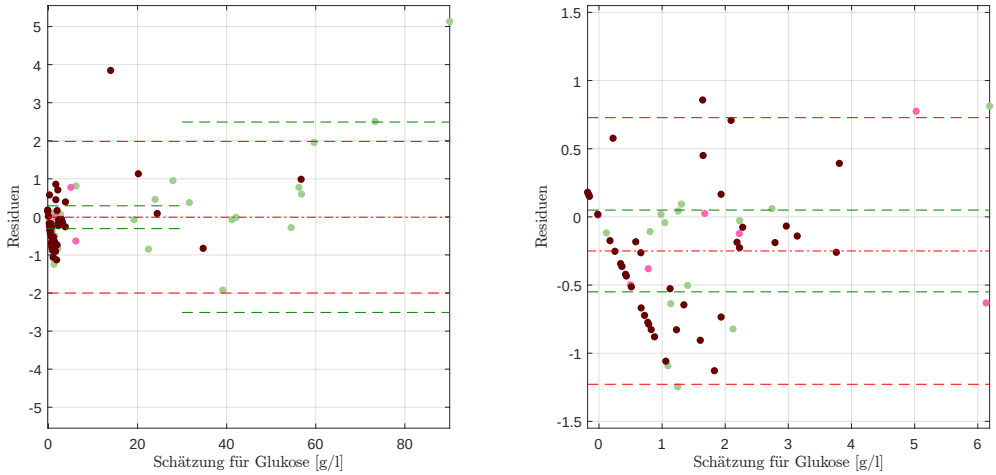

Abbildung 7.7: Residuenplot des NN-Modells für Glukose mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.), sowie der vergrößerte Ausschnitt $[0, 10]\,\mathrm{g/l}$ mit Bias und Abweichungen für diesem Bereich.

ähnliches Verhalten wie in Abbildung 7.6 (re.). An dieser Stelle wird ersichtlich, wie sensibel die neuronalen Netzwerke auf das Vorenthalten von Informationen reagieren. So reduzieren sich die Residuen im Vergleich zu jenen der doppelten Kreuzvalidierung auf ein Intervall von $[-1.92, 5.13]\,\mathrm{g/l}$. Aufgrund der vergleichsweise wenigen Information, welche für Weine mit einem Glukosegehalt von mehr als $40\,\mathrm{g/l}$ vorliegt, wird zusätzlich jener Ausschnitt des Wertebereiches, welcher die meisten Stichproben beinhaltet, näher untersucht. In Abbildung 7.7 (re.) finden sich die aus dem Gesamtmodell extrahierten Weine mit einem Schätzwert von weniger als $10\,\mathrm{g/l}$ mit deren eingezeichneter Standardabweichung. Man beachte, dass hierfür keine eigene Kalibrierung durchgeführt wurde. Jene auffallende Gerade, auf welcher der Großteil der Residuen in dieser Grafik liegt, entspricht wiederum den glukosefreien Weinen. Da durch die Transformierung der Schätzungen innerhalb der Vorgehensweise von neuronalen Netzwerken diese tendenziell in einen positiven Wertebereich abgebildet werden, führt dies meist zu einer Überschätzung des Glukosewertes der Weine mit einer Konzentration bei $0\,\mathrm{g/l}$. Dies begründet die Tatsache, dass die restlichen Weine mit einem Glukosegehalt von $(0, 10]\,\mathrm{g/l}$ einen leichten, positiven Trend aufzuweisen scheinen. Während die Residuen des Gesamtmodells eine Standardabweichung von $0.9959\,\mathrm{g/l}$ aufweisen, reduziert sich dies im eingeschränkten Abschnitt auf $0.4890\,\mathrm{g/l}$ mit einem MSE in Höhe von $0.2978(\,\mathrm{g/l})^2$ und einem Bias von $-0.25\,\mathrm{g/l}$, welcher sich durch das nicht eigens für diesen Wertebereich durchgeführte Kalibrieren begründen lässt.

Aufgrund dieser Erkenntnisse scheint eine Neukalibrierung, insbesondere mit einem eigens für den Wertebereich $[0, 10]\,\mathrm{g/l}$ entwickelten Modell, unumgänglich. Die verwendeten Wellenzahlen ändern sich somit von den Bereichen in Abbildung 7.6 (li.) zu jenen aus der Visualisierung in Abbildung 7.8, wobei für dieses

konkrete Modell die Standardkalibrierung[3] verwendet wurde. Ebenfalls mit der zweiten Savitzky-Golay Ableitung mit einem zusammenhängenden Wellenzahlenblock der Länge 25 im Fingerprintbereich, welcher den Peak im Originalspektrum bei $1\,045\,\mathrm{cm}^{-1}$ einschließt, bei der Wahl von einem versteckten Neuron. Die im Gesamtmodell beobachtete Struktur kann mit diesem Parametersetting weitestgehend eliminiert werden und es tritt lediglich ein Ausreißer auf, bei welchem es sich um jenen Wein mit der höchsten Glukosekonzentration im reduzierten Wertebereich handelt. Offensichtlich wird mit diesem die Verringerung eines Bias erzielt, welcher sich auf einen Wert von circa $-0.01\,\mathrm{g/l}$ beläuft. Insgesamt kann durch dieses spezifische Modell die Standardabweichung reduziert werden und insbesondere ergibt sich eine auffallend starke Reduzierung des MSEs auf $0.0478(\,\mathrm{g/l})^2$.

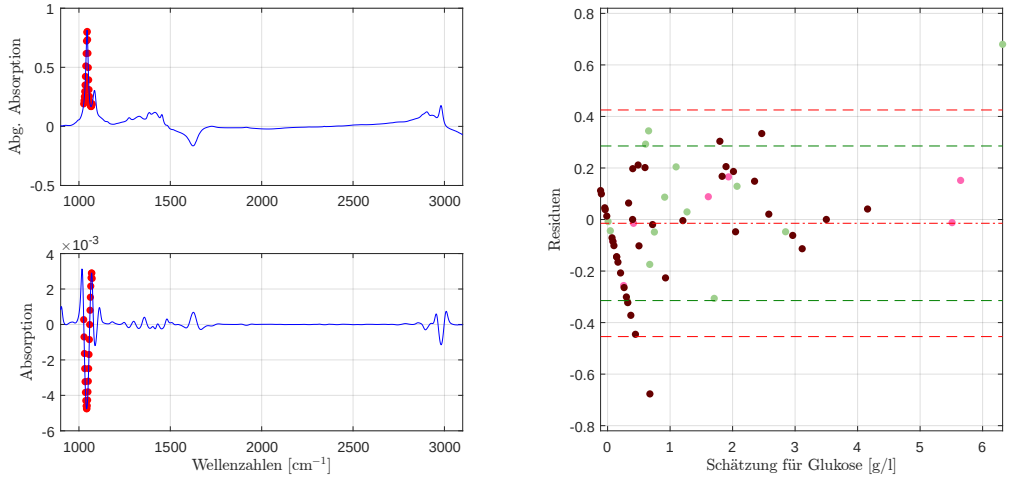

Abbildung 7.8: Die selektierten Wellenzahlen im für $[0,10]\,\mathrm{g/l}$ wertebereichspezifischen NN-Modell für Glukose mit der zweiten Savitzky-Golay Ableitung (li.) und der dazugehörende Residuenplot mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (re.).

	Anz.	Mw.	Std.	Min.	Med.	Max.	IQR	MSE
Gesamtmodell Glukose	81	−0.01	0.9959	−1.92	−0.12	5.13	0.72	0.9797
für $[0,10]\,\mathrm{g/l}$	62	−0.25	0.4890	−1.25	−0.21	0.86	0.66	0.2978
Modell für $[0,10]\,\mathrm{g/l}$	62	−0.01	0.2199	−0.68	−0.01	0.68	0.27	0.0478

Tabelle 7.4: Modellvergleich des NN-Modells für Glukose mit jenem für den Wertebereich $[0,10]\,\mathrm{g/l}$ eigens entwickelten Modell. Alle Werte in $\mathrm{g/l}$.

Zusammenfassend kann für die neuronalen Netzwerke eine gute Modellgüte der Glukosekonzentrationen für den auf $[0,10]\,\mathrm{g/l}$ reduzierten Wertebereich unterstellt werden. Für eine Modellierung des restlichen Stichprobenraumes ohne die

[3] Die Aktivierungsfunktion des versteckten Layers ist die Tangens-Sigmoidfunktion wie in Gleichung (6.5).

Berücksichtigung dieser Daten liegen jedoch insbesondere für eine Kalibrierung mit den neuronalen Netzwerken zu wenige Datensätze (19 Objekte) vor.

Reproduzierbarkeit

In einem ersten Schritt wird die Reproduzierbarkeit des Gesamtmodells trotz der leichten, strukturellen Schwächen analysiert und es zeigt sich ein, aufgrund der Residuenplots zu erwartendes Bild, auch wenn es sich bei jenem Weißwein mit der größten Variabilität von 1.5556 g/l und zugleich der höchsten maximalen Abweichung mit 3.22 g/l nicht um jenen Wein mit dem maximalen Residuum handelt. Die Unauffälligkeit dieser Weinprobe (mit Identifikantionsnummer 16) kann durch die Residuen der vier Wiederholungen erklärt werden (-1.74 g/l, -1.44 g/l, 0.54 g/l und 1.49 g/l). Zudem zeigen sich speziell bei den Weißweinen auffällig viele Ausreißer in Bezug auf die Gesamtheit der vorliegenden Daten wie in den Wiederholbarkeitsplots in Abbildung 7.9 grafisch dargestellt, selbst wenn jener Rotwein mit der Identifikationsnummer 81 eine schlechte Reproduzierbarkeit für ebendiesen Wein aufzeigt (vgl. Abbildung 7.9 die großen Unterschiede der Schätzungen für ebendiesen Wein mit ID 81).

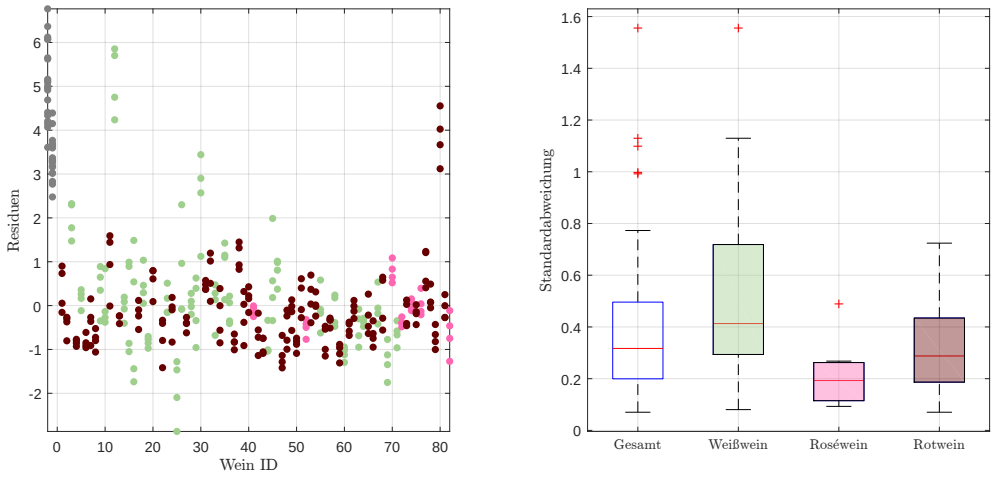

Abbildung 7.9: Residuen in g/l der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung der Glukosekonzentration im NN-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

Zudem kann aus Tabelle 7.5 abgelesen werden, dass die Weißweine eine signifikant schlechtere Reproduzierbarkeit aufweisen.

Im Vergleich zu Tabelle 7.5, welche die Übersicht über die Reproduzierbarkeit des Gesamtmodells zeigt, können auch hier für das auf den Wertebereich $[0, 10]$ g/l reduzierte Modell deutlich bessere Ergebnisse erzielt werden.[4]

[4]Man beachte an dieser Stelle, dass die Kennzahlen der beiden Modelle nicht direkt miteinander verglichen werden können, da in Tabelle 7.5 sämtliche Weinproben enthalten sind. Es zeigt sich

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.5198	0.4130	3.2230	1.1382	0.9686	0.7255
Roséwein	0.2181	0.1928	1.1569	0.5013	0.4503	0.3280
Rotwein	0.3153	0.2878	1.6029	0.7006	0.6108	0.3549
Gesamt	0.3826	0.3169	3.2230	0.8454	0.6785	0.5662

Tabelle 7.5: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Glukose im NN-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.1769	0.1681	0.7411	0.3947	0.3632	0.1662
Roséwein	0.1645	0.1993	0.6408	0.3723	0.4536	0.2010
Rotwein	0.1904	0.1930	0.8881	0.4183	0.4252	0.1841
Gesamt	0.1840	0.1866	0.8881	0.4070	0.4142	0.1793

Tabelle 7.6: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Glukose im NN-Modell, eingeschränkt auf den Wertebereich $[0, 10]$ g/l, anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

Für den Vergleich der unterschiedlichen Datensätze, wie aus den Tabellen 7.7 und 7.8 ersichtlich, zeigt sich in beiden Modellen ein ähnliches Bild. Während die Residuen für den Datensatz E25 die geringsten Kennzahlen aufweisen, so weist ebendieser Datensatz mit den hier verwendeten Modellen eine vergleichsweise hohe Reproduzierbarkeit auf.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	1.1955	0.93	−2.61	6.61	0.3013	0.2550	2.8035
E24	1.0640	0.96	−2.21	5.24	0.3314	0.2667	3.8919
E25	0.9959	0.72	−1.92	5.13	0.3826	0.3169	3.2230
V70(2016)	1.0977	0.86	−1.87	5.28	0.2749	0.2592	1.6937

Tabelle 7.7: Performance des entwickelten NN-Modells für Glukose, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.2756	0.41	−0.49	0.74	0.1253	0.1238	0.5598
E24	0.2748	0.37	−0.53	0.65	0.1593	0.1531	0.7648
E25	0.2199	0.27	−0.68	0.68	0.1840	0.1866	0.8881
V70(2016)	0.3291	0.41	−0.72	1.19	0.1563	0.1578	0.9376

Tabelle 7.8: Performance des entwickelten NN-Modells für Glukose, eingeschränkt auf den Wertebereich $[0, 10]$ g/l, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

7.4 Fruktose

In vorliegendem Datensatz befinden sich Weine mit einem Fruktosegehalt von 0 g/l (insgesamt 15 Weine) bis hin zu einer Konzentration von 137.2 g/l. Diese Daten liegen allerdings mit einer Anzahl von 64 Weinen stark konzentriert in einem Bereich bis zu 10 g/l vor. Aufgrund dessen sind für Süßweine mit einem hohen Fruchtzuckeranteil, welche für Vorhersagen als eine Art Extrapolation interpretiert werden können, tendenziell größere Residuen zu erwarten, da versucht werden muss, nicht vorhandene Informationen in diesem Bereich zu modellieren, wie dies beispielsweise auf den Residuenplot der doppelten Kreuzvalidierung zutrifft, wie bereits für Glukose analysiert.

dennoch eine starke Verbesserung, wie der Kennzahlenübersicht in Tabelle 7.35 entnommen werden kann.

Die neuronalen Netzwerke versuchen den funktionalen Zusammenhang zwischen den Spektren und Fruktosekonzentrationen durch einen zusammenhängenden Wellenzahlbereich mit 50 Prädiktoren und insgesamt 3 versteckten Neuronen abzubilden. Diese selektierten Vorhersagevariablen befinden sich zur Gänze im Fingerprintbereich und umschließen zwei kleinere Peaks im originalen Spektrum. Bei der verwendeten zweiten Savitzky-Golayableitung entspricht dies der Senke bei $1\,267\,\mathrm{cm}^{-1}$ bis unmittelbar nach dem Peak bei $1\,362\,\mathrm{cm}^{-1}$ und ist in Abbildung 7.10 (li.) grafisch dargestellt.

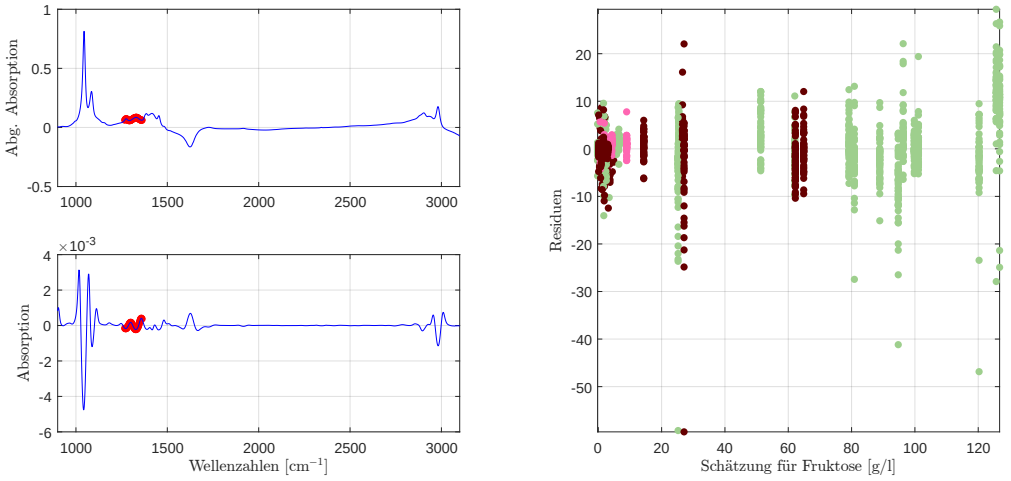

Abbildung 7.10: Die selektierten Wellenzahlen im NN-Modell für Fruktose mit der zweiten Savitzky-Golay Ableitung (li.) und der aus der doppelten Kreuzvalidierung resultierende Residuenplot (re.).

Erwartungsgemäß nimmt die Variabilität mit der Zunahme der Fruktosekonzentration (über $20\,\mathrm{g/l}$) zu, wie die Grafik der Residuen der doppelten Kreuzvalidierung in Abbildung 7.10 (re.), mit Ausnahme jener Weine in dem Bereich $[40, 80]\,\mathrm{g/l}$, bestätigt. Diese Residuen scheinen, ausgenommen der Weine mit einem Fruktosegehalt über $120\,\mathrm{g/l}$, gleichermaßen um den Nullpunkt zu streuen. Betrachtet man den dazugehörenden klassischen Residuenplot in Abbildung 7.11 (li.), so können mit Ausnahme der starken Unterschätzung der Weine mit einem Fruktosegehalt von $136\,\mathrm{g/l}$ bzw. $137.2\,\mathrm{g/l}$ keine Auffälligkeiten in dieser Grafik beobachtet werden. Auch wenn die Streuung der Daten mit einer Standardabweichung in Höhe von $0.9795\,\mathrm{g/l}$ im Verhältnis zur weitreichenden Spanne des Fruktosegehaltes gering ausfällt, kann die Genauigkeit der zertifizierten Laborwertmessmethode (in Grün dargestellt) insbesondere im Bereich geringer Schätzungen von unter $30\,\mathrm{g/l}$ nicht erreicht werden. Selbst bei einer Nichtberücksichtigung der beiden Ausreißer reduziert sich die Standardabweichung ohne eine erneute Kalibrierung lediglich auf knapp unter $0.6000\,\mathrm{g/l}$ und entspricht in etwa der Standardabweichung der Residuen mit einem Fruchtzuckergehalt von weniger als $30\,\mathrm{g/l}$.

Auch wenn diese beiden Ausreißer ein hohes Residuum aufweisen, so besitzen alle gemeinsam dennoch einen negativen Mittelwert in Höhe von $-0.10\,\mathrm{g/l}$.

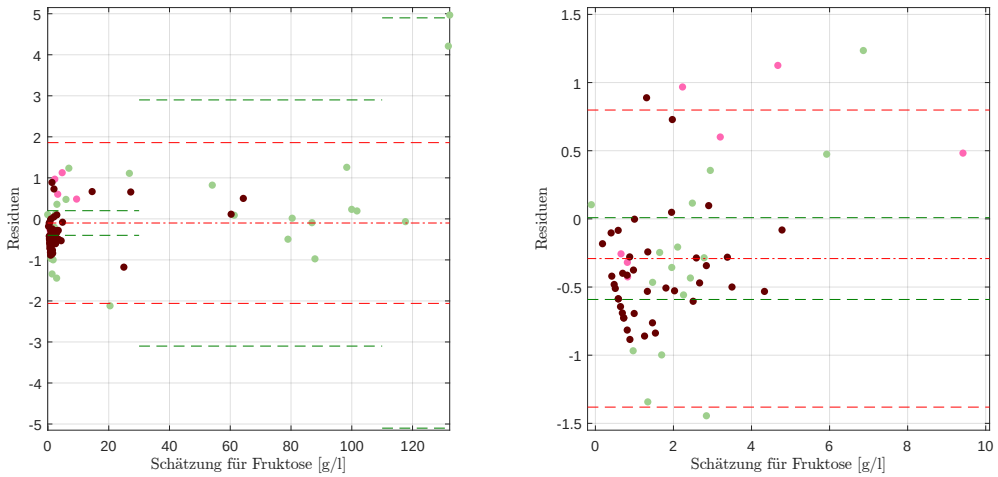

Abbildung 7.11: Residuenplot des NN-Modells für Fruktose mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.), sowie der vergrößerte Ausschnitt $[0, 10]\,\mathrm{g/l}$ mit Bias und Abweichungen für diesem Bereich.

Betrachtet man die Residuen in dem Bereich mit der höchsten Punktedichte, $[0, 10]\,\mathrm{g/l}$, in Abbildung 7.11 (re.), so scheint das Modell in diesem Bereich einem steigenden Trend zu folgen. Dieser Eindruck wird dadurch erweckt, da das hier verwendete Modell wie bereits bei der Modellierung der Glukose den Großteil der Daten in einen durch die Trainingsdaten vordefinierten Bereich transformiert. Dies hat zur Folge, dass aufgrund der hohen Punktedichte im Bereich $[0, 1]\,\mathrm{g/l}$ diese grundsätzlich überschätzt werden[5] und kann als Ursache für das Verhalten des Residuenplots identifiziert werden. Zudem verstärken die Residuen mit einem Wert von über $0.5\,\mathrm{g/l}$ die Wahrnehmung eines Trends enorm. Auch für die Fruktose kann somit kein adäquates Modell mit unauffälligen Residuen gefunden werden.

Schränkt man die Datenpunkte auf jene ein, welche unter einer gewissen Grenze liegen, wie beispielsweise auf die Punktwolke zwischen $0\,\mathrm{g/l}$ und $10\,\mathrm{g/l}$ oder $30\,\mathrm{g/l}$, so entfernt sich die Biaslinie von der idealen Nulllinie, aufgrund der zwei sehr hohen Residuen der beiden Süßweine. Grundsätzlich bedeutet die Beschränkung des Wertebereiches nach oben eine Reduzierung der Kennzahlen der Residuen wie Standardabweichung oder Bias, wobei das Gesamtmodell durch Ausschluss der Süßweine zu einem insgesamt einigermaßen stabilen Ergebnis führt, wie in Tabelle 7.9 veranschaulicht.

Für ein Modell, insbesondere durch den beobachtbaren Trend und im Hinblick auf die Extrapolation für Süßweine scheint daher ein PLS-Modell geeigneter zu sein.

[5] In diesem konkreten Beispiel wird lediglich ein Wein mit einer Konzentration von $0\,\mathrm{g/l}$ mit einem negativen Fruktosewert unterschätzt.

	g/l	Anz.	Mw.	Std.	Min.	Med.	Max.	IQR	MSE	Neg.
		81	−0.1006	0.9795	−2.12	−0.2867	4.97	0.68	0.9577	1
≤	130	79	−0.2193	0.6341	−2.12	−0.3189	1.26	0.68	0.4452	1
≤	30	67	−0.2824	0.6276	−2.12	−0.3991	1.24	0.58	0.4678	1
≤	10	62	−0.2912	0.5448	−1.44	−0.4061	1.24	0.50	0.3768	1

Tabelle 7.9: Verhalten der Residuen für konkrete Bereiche der Fruktosekonzentration im NN-Modell. Alle Werte in g/l bzw. die Anzahl an Weinen in der letzten Spalte.

Reproduzierbarkeit

Betrachtet man wiederum die Reproduzierbarkeitsplots in Abbildung 7.12, so liegt die größte Variabilität der Streuung in den einzelnen Messungen bei den Weißweinen vor. Dies ist unter anderem darauf zurückzuführen, dass hierbei relativ viele Süßweine vorhanden sind und durch eine Art Extrapolation kleinere Abweichungen in den Spektren verstärkt werden können. Auf den ersten Blick scheinen die Roséweine die beste Reproduzierbarkeit aufzuweisen, allerdings muss hier wiederholt auf die Tatsache hingewiesen werden, dass lediglich 7 Weine im Datensatz vorhanden sind. Auffallend ist bei den Rotweinen, dass oftmals zwei Messungen auftreten, welche ein sehr ähnliches Residuum aufweisen, wohingegen eine oder die beiden anderen Wiederholungsmessungen hiervon relativ stark abweichen. Insgesamt kann jedoch ein durchwegs homogenes Verhalten der Rot- und Weißweine, gemeinsam mit dem Gesamtüberblick beobachtet werden (Abbildung 7.12, re.).

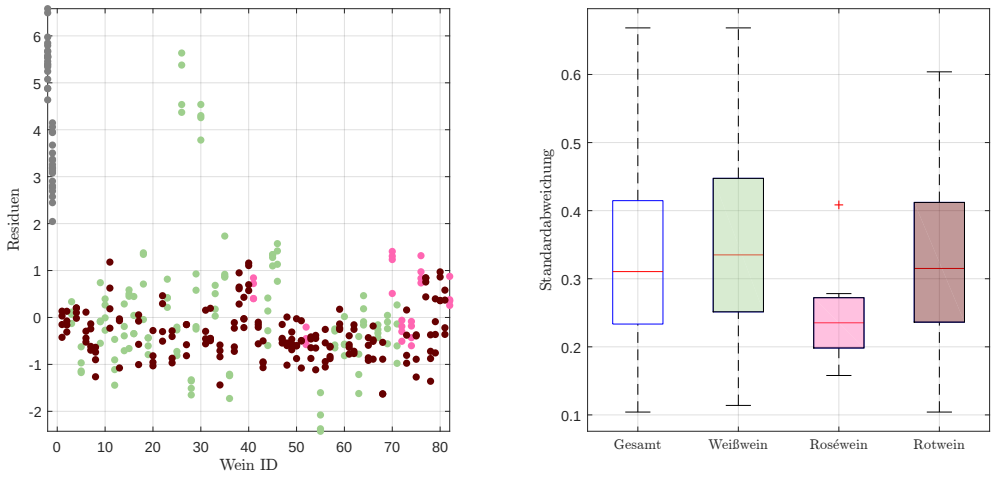

Abbildung 7.12: Residuen in g/l der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung der Fruktosekonzentration im NN-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

Betrachtet man zusätzlich den Überblick über die maximalen Abweichungen, so ergibt sich ein ähnliches Bild wie jenes durch die Grafiken der Wiederholbarkeit

induzierte. Die Rotweine scheinen eine leicht bessere Reproduzierbarkeit als die Weißweine aufzuweisen. Hierbei gilt, dass von 7 Weinen mit einer maximalen Differenz der Einzelmessungen von über 1.20 g/l sechs auf Weißweine entfallen und lediglich ein Rotwein ein derartiges Maximum aufweist. Die teilweise hohen Reproduzierbarkeitswerte selbst können jedoch nicht ausschließlich auf einen hohen Fruktoseanteil zurückgeführt werden: selbst bei weniger als 2 g/l tritt eine maximale Abweichung von über 1.2 g/l auf und das Modell selbst kann für die Reproduzierbarkeit als unzureichend erachtet werden.

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.3565	0.3350	1.4210	0.7871	0.7755	0.3366
Roséwein	0.2498	0.2353	0.8957	0.5535	0.5174	0.1730
Rotwein	0.3240	0.3151	1.4084	0.7089	0.6643	0.2653
Gesamt	0.3296	0.3105	1.4210	0.7244	0.6635	0.2921

Tabelle 7.10: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Fruktose im NN-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

Stellt man zusätzlich den Vergleich der Datensätze in Tabelle 7.11 an, so weist das für E25 ermittelte, optimale Modell für den Datensatz E22 niedrigere Kennzahlen beim Residuenvergleich aus. Selbiges kann für die Reproduzierbarkeit mit Ausnahme der maximalen Abweichung beobachtet werden.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.9016	0.73	−1.31	4.72	0.3218	0.3008	2.1189
E24	1.0820	0.95	−2.27	5.53	0.3073	0.2885	1.8155
E25	0.9795	0.68	−2.12	4.97	0.3296	0.3105	1.4210
V70(2016)	1.0220	0.89	−1.97	4.97	0.2739	0.2571	1.4190

Tabelle 7.11: Performance des entwickelten NN-Modells für Fruktose, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

7.5 Titrierbare Säure

Bei den titrierbaren Säuren handelt es sich um einen Summenparameter, weshalb, analog zu Dichte, Extrakt oder pH-Wert, den titrierbaren Säuren keine spezifische chemische Strukturformel zugeordnet werden kann. Dennoch resultiert das Modell für die titrierbaren Säuren in der Wahl von lediglich 10 Wellenzahlen und einem versteckten Neuron. Im Vergleich zum Modell mittels der PLS-Methode werden für

vergleichbare Ergebnisse somit nur die Hälfte der Prädiktoren verwendet, obwohl die Information aus ähnlichen Bereichen herausgelesen wird, wie Abbildung 7.13 zeigt. Die Wahl des ersten kleinen Peaks bei 1 742 cm^{-1} findet sich auch im Modell in Abschnitt 5.5 wieder. Ebenso wird der Anstieg im Bereich 2 262 cm^{-1} zur Schätzung der Konzentration von titrierbaren Säuren herangezogen.

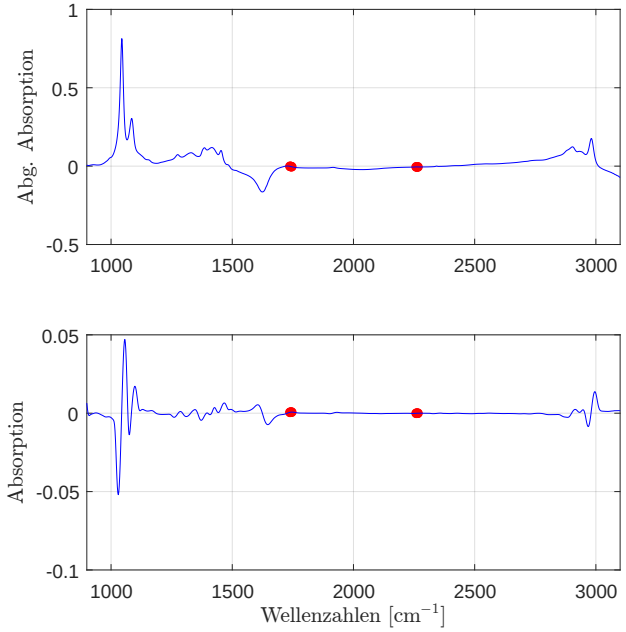

Abbildung 7.13: Die selektierten Wellenzahlen im NN-Modell für titrierbare Säuren mit der ersten Savitzky-Golay Ableitung.

Betrachtet man die aus der doppelten Kreuzvalidierung stammenden Residuen in Abbildung 7.14 (li.), so kann eine für die Weißweine vergleichsweise starke Streuung innerhalb der einzelnen Weinproben beobachtet werden. Dies gilt insbesondere für Weißweine, welche grundsätzlich einen höheren Gehalt an titrierbaren Säuren aufweisen. Es kann allerdings anhand dieser Grafik kein struktureller Mangel der Methode beobachtet werden, noch ist ein Trend oder eine andere systematische Abweichung erkennbar. Hierbei dürfen die drei/vier überschätzten Weißweine im Bereich bis 4.5 g/l nicht überbewertet werden. Betrachtet man zusätzlich das zugehörende Histogramm, so scheinen diese Residuen einigermaßen symmetrisch um 0 g/l verteilt zu sein, insbesondere wenn man der Beeinflussung der drei überschätzten Weißweine mit einem Säuregehalt von weniger als 4.5 g/l nicht zu starkes Gewicht beimisst.

Betrachtet man den Residuenplot für die titrierbaren Säuren in Abbildung 7.15, so befinden sich die Schätzfehler zwischen -0.47 g/l und 0.38 g/l in einem relativ kompakten Bereich. Der interquartile Bereich misst hier eine Länge von 0.22 g/l. Betrachtet man die Variabilität der Residuen, so kann beinahe jene der Referenzmethode erreicht werden. Die einzige hier beobachtbare Auffälligkeit ist, dass der Bias des gesamten Modells (nicht signifikant) unter 0 g/l liegt, während jener der Rotweine geringfügig positiv ausfällt. Aus dieser Tatsache kann gefolgert werden,

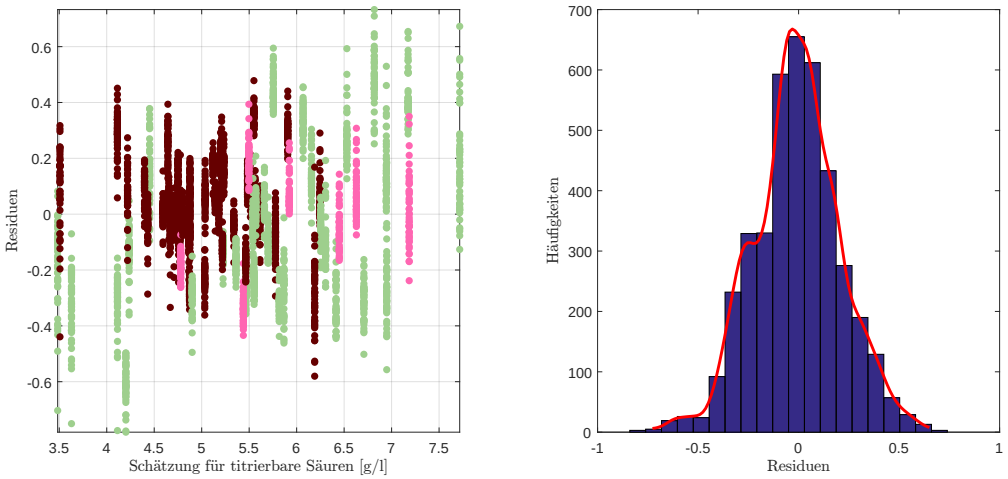

Abbildung 7.14: Die Weinresiduen der doppelten Kreuzvalidierung als Residuenplot (li.) und zusammengefasst zu einem Histogramm (re.) mit einer empirischen Dichteschätzung im NN-Modell für titrierbare Säuren.

dass der Säuregehalt der Rotweine tendenziell überschätzt, während jener der Weiß- und Roséweine gleichermaßen unterschätzt wird. Dieser Unterschied kann somit neben dem PLS-Modell auch bei den neuronalen Netzwerken beobachtet werden. Es kann daher durchaus Sinn machen, für die Weintypen ein getrenntes Modell zu kalibrieren.

Die konkrete Standardabweichung der hier dargestellten Residuen beläuft sich auf 0.1728 g/l bei einem MSE in Höhe von $0.0295(\,\mathrm{g/l})^2$ und in Anbetracht der Genauigkeit der gegebenen Referenzwerte von lediglich einer Nachkommastelle liefert der Mittlere Fehler eine durchaus akzeptablen Wert.

Die beobachtbare Über-/Unterschätzung der titrierbaren Säuren, geclustert nach den Weinfarben „rot“ und „nicht rot“, kann durch die zugrundeliegende Population erklärt werden, wie bereits zuvor in der Analyse mittels PLS-Methode in Abschnitt 5.5 erläutert. Aus diesem Grund scheint ein gesondertes Modell für Rotweine als sinnvolle Alternative. Eine hierfür spezifizierte Kalibrierung verwendet insgesamt 6 Wellenzahlen. Auch bei diesem Modell können keine Auffälligkeiten beobachtet werden - weder Trends noch Ausreißer sind erkennbar. Der Vorteil der Modellspezialisierung liegt in der beinahen Eliminierung des Bias in den Residuen und der Reduzierung der absoluten Streubreite, liefert allerdings keine signifikanten Verbesserungen, während die Neukalibrierung des Gesamtmodells mit ausschließlich den Rot- bzw. den Weiß- und Roséweinen ein, mit Ausnahme eines minimalen Bias, vergleichbares Bild zeigen, wie in Abbildung 7.16 dargestellt.

Bei dieser Kalibrierung kann eine geringere empirische Standardabweichung der Rotweinresiduen als jene der Referenzmethode bezüglich des Gesamtmodells erzielt werden, während diese für die Weiß- und Roséweine in gleichem Maße überschritten wird.

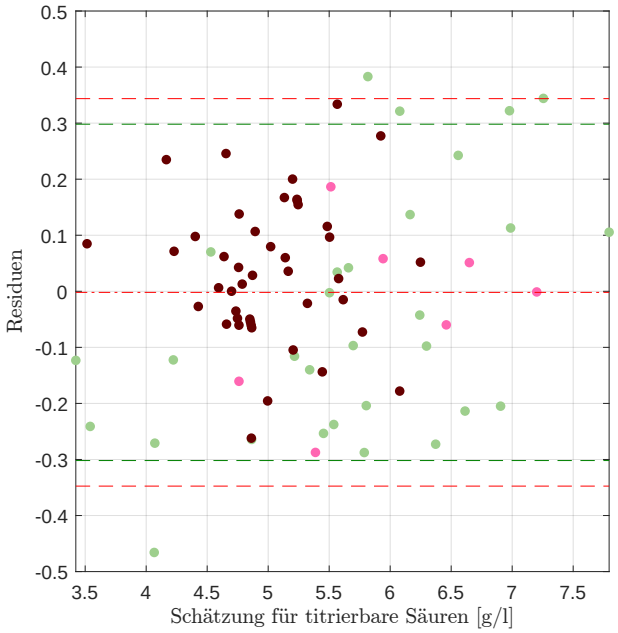

Abbildung 7.15: Residuenplot des NN-Modells für titrierbare Säuren mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode.

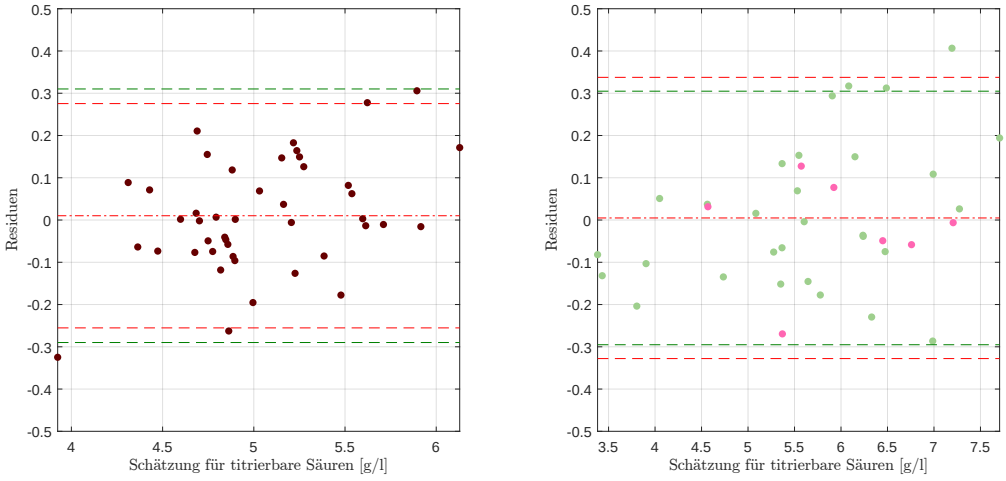

Abbildung 7.16: Residuenplot des NN-Modells für titrierbare Säuren mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode. Kalibrierung des Gesamtmodells mit Rotweinen (li.) und Rosé-/Weißweinen (re.).

Reproduzierbarkeit

Den Abbildungen 7.17 kann wiederum die grafische Darstellung der Reproduzierbarkeit entnommen werden. Die Boxplots der Weißweine, sowie jener aller Weinproben sind beinahe deckungsgleich, wohingegen die Darstellung der Rotweine mit einem kürzeren interquartilen Bereich und etwas kürzeren Tails eine minimal bessere Wiederholbarkeit suggerieren. Die Gestalt des Boxplots für Roséweine kann auf deren geringe Anzahl zurückgeführt werden und wird deshalb wenig Bedeutung zugeschrieben.

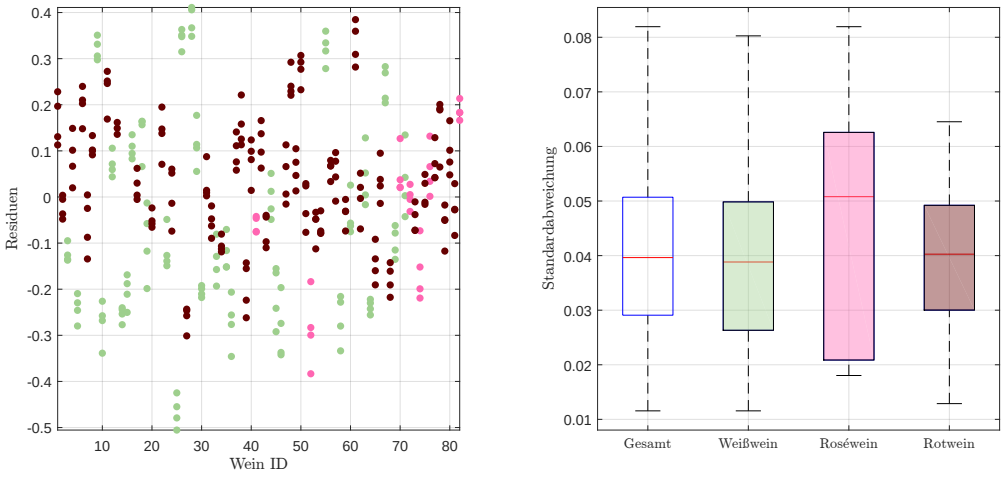

Abbildung 7.17: Residuen in g/l der Einzelspektren pro Wein, ohne Kunstweine, zur Schätzung der titrierbaren Säuren im NN-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, ohne künstliche Weine (re.).

Insgesamt kann allen Weinklassen eine vergleichsweise gute Reproduzierbarkeit unterstellt werden, insbesondere bei Berücksichtigung der Tatsache, dass die Referenzwerte lediglich auf eine Nachkommastelle genau vorliegen.

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0408	0.0388	0.1855	0.0919	0.0815	0.0416
Roséwein	0.0451	0.0508	0.1997	0.1032	0.1064	0.0603
Rotwein	0.0396	0.0403	0.1389	0.0890	0.0916	0.0298
Gesamt	0.0405	0.0396	0.1997	0.0913	0.0915	0.0373

Tabelle 7.12: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für titrierbare Säuren im NN-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

Stellt man analog zu den vorangegangenen Abschnitten den Datensatzvergleich an, so kann ein ähnliches Verhalten der Residuen aller Datensätze beobachtet werden,

wiederum mit geringfügig besserer Standardabweichung der Residuen, welche auf der Schätzung mittels den Spektren der Engines E22 und E25 basieren. Mit dem Verweis auf die Genauigkeit der vorliegenden Daten können allerdings, insbesondere bei Betrachtung der Reproduzierbarkeit dieser beiden Spektrometer, keine signifikanten Unterschiede identifiziert werden. Mit einer maximalen Abweichung in den Reproduzierbarkeitsdaten von 0.4932 g/l weist der Datensatz E24 einen mehr als doppelt so hohen Wert wie die vergleichbaren Datensätze E22 und E25 aus.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.1786	0.20	−0.46	0.48	0.0384	0.0369	0.2036
E24	0.1829	0.24	−0.41	0.49	0.0460	0.0433	0.4932
E25	0.1728	0.22	−0.47	0.38	0.0405	0.0396	0.1997
V70(2016)	0.1939	0.22	−0.56	0.41	0.0483	0.0462	0.2830

Tabelle 7.13: Performance des entwickelten NN-Modells für titrierbare Säuren, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

7.6 Weinsäure

Ein wichtiger Bestandteil der (titrierbaren) Säuren bildet die Weinsäure mit dem größten Wertebereich aller Säurearten der vorliegenden Datensätze von $[0.00, 3.80]$ g/l.

Die in dieser Arbeit verwendeten Heuristiken bestimmen für eine optimale Schätzung der Weinsäure in den Proben des Datensatzes E25, insgesamt 40 Wellenzahlen, aufgeteilt in vier unterschiedliche und gleich große Prädiktorenintervalle. Mit drei dieser vier Intervalle befindet sich der Großteil der von den neuronalen Netzwerken verwendeten Messstellen im Fingerprintbereich und umschließt einerseits den Peak bei circa 1 400 cm^{-1}, sowie mittels zwei dicht beieinander liegenden Intervallen die Senke um 1 180 cm^{-1}. Die Restinformation vermutet das neuronale Netzwerk in einem Bereich um 1 750 cm^{-1}, wie in Abbildung 7.18 (li.) visualisiert. Der aus der doppelten Kreuzvalidierung resultierende Residuenplot in ebendieser Abbildung (re.) zeigt, von zwei Weißweinen abgesehen, ein einigermaßen unauffälliges Bild. Bei jenem Weißwein am unteren Ende des Wertebereiches handelt es sich um einen Honigwein, welcher im Datensatz mit 0 g/l aufgeführt wird. Bei jenem Weißwein, welcher am oberen Ende des Wertebereiches liegt und vergleichsweise stark unterschätzt wird, handelt es sich um den alkoholfreien Weißwein mit einer Weinsäure in Höhe von 3.80 g/l, während die Werte der restlichen 79 Weine in $[1.00, 2.90]$ g/l liegen. Daher kann an dieser Stelle ein möglicher Messfehler (keine Weinsäure) vermutet bzw. als extrapolierten Wert angesehen werden.

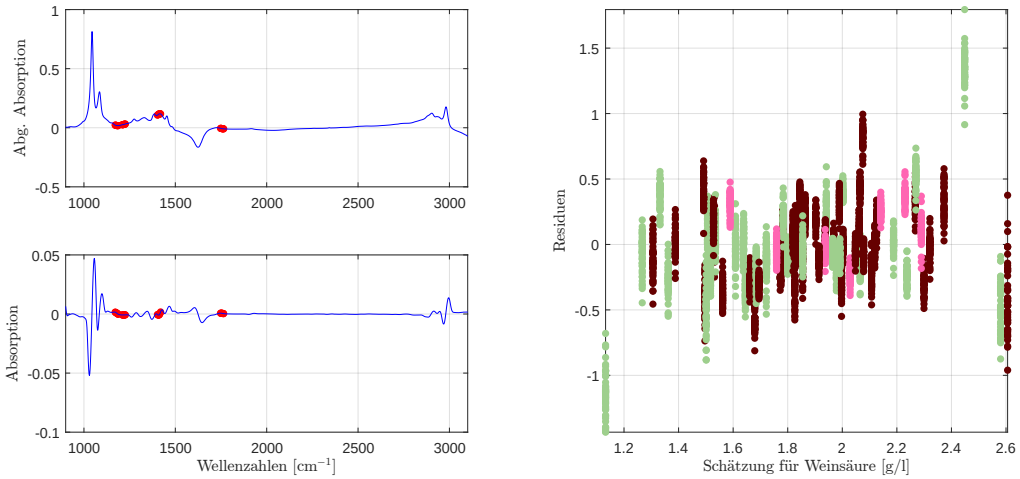

Abbildung 7.18: Die selektierten Wellenzahlen im NN-Modell für Weinsäure mit der ersten Savitzky-Golay Ableitung (li.) und der aus der doppelten Kreuzvalidierung resultierende Residuenplot (re.).

Analysiert man den klassischen Residuenplot in Abbildung 7.19 (li.), so kann wiederum ein geringfügiger Bias beobachtet werden, sowie, verglichen mit der Referenzmethode, eine mehr als doppelte Standardabweichung. Zudem scheint ein leichter Trend in den Residuen vorhanden zu sein, sowie eine minimale Verringerung der Variabilität im mittleren Bereich um 1.80 g/l. Die systematische Über- bzw. Unterschätzung im Sinne eines Trendes wird in Abbildung 7.19 (re.) etwas entkräftet. Ein perfektes Modell würde die Schätzungen auf die rot gestrichelte Linie projizieren, wobei sich die tatsächliche, lineare Schätzgerade in Schwarz davon deutlich unterscheidet, und stark von den beiden Weinen außerhalb des Wertebereichs $[1.00, 2.90]$ g/l beeinflusst wird. Bei Vernachlässigung dieser beiden Punkte für die Berechnung der Gerade ergibt sich die blaue Linie als möglicher Trend, wobei dieser deutlich weniger von der Diagonale abweicht.

Betrachtet man die Unterklasse der Rotweine, so kann eine geringfügig bessere Kalibrierung ermittelt werden. Hierfür werden wiederum 40, teilweise übereinstimmende, Wellenzahlen selektiert. Auch wenn sich die Standardabweichung der Residuen um 30 % reduziert, gibt es zugleich einen ansteigenden Trend, weshalb an dieser Stelle nicht näher auf das rotweinspezifische Modell eingegangen wird.

Reproduzierbarkeit

Auch wenn die Residuen eine im Vergleich zur Referenzmethode mehr als doppelt so hohe Variabilität aufweisen, zeigt sich im Datensatz E25 eine einigermaßen plausible Reproduzierbarkeit, insbesondere, da die Vergleichswerte auf die erste Nachkommastelle genau zur Verfügung stehen. So liegt die mittlere, maximale Abweichung mit $0.09 \approx 0.1$ g/l exakt bei der Genauigkeit von 10^{-1} g/l (vgl. Tabelle 7.14).

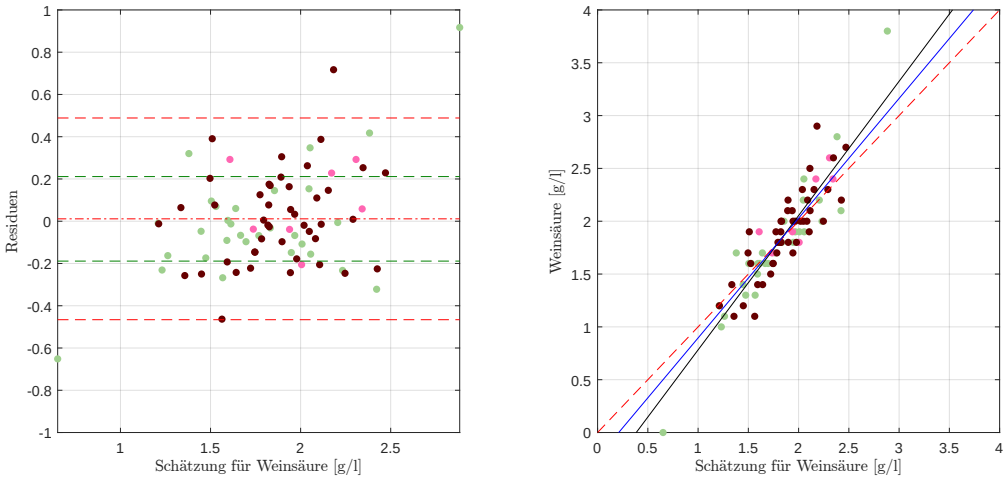

Abbildung 7.19: Residuenplot des NN-Modells für Weinsäure mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.), sowie die Schätzungen gegen die tatsächlichen Werte mit optimaler Schätzline, d.h. der rot gestrichelten Diagonale, sowie der durch die Daten motivierte lineare Zusammenhang in Schwarz, sowie der lineare Zusammenhang unter Vernachlässigung der beiden Hebelpunkte, dargestellt als blaue Gerade (re.).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0424	0.0420	0.1901	0.0940	0.0885	0.0365
Roséwein	0.0531	0.0566	0.1904	0.1229	0.1269	0.0536
Rotwein	0.0404	0.0373	0.1700	0.0896	0.0818	0.0393
Gesamt	0.0423	0.0416	0.1904	0.0941	0.0883	0.0401

Tabelle 7.14: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Weinsäure im NN-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.2687	0.33	−0.76	1.09	0.0336	0.0305	0.1609
E24	0.2772	0.32	−0.77	1.17	0.0357	0.0338	0.2088
E25	0.2387	0.30	−0.65	0.92	0.0423	0.0416	0.1904
V70(2016)	0.2729	0.29	−0.74	0.98	0.0311	0.0287	0.1520

Tabelle 7.15: Performance des entwickelten NN-Modells für Weinsäure, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

Betrachtet man zusätzlich die Standardabweichungen der wiederholten Messungen, so treten lediglich zwei Ausreißer bei der nach Weinfarben getrennten Betrachtung auf, wobei diese in der gesamten Population nicht mehr als solche identifizierbar sind.

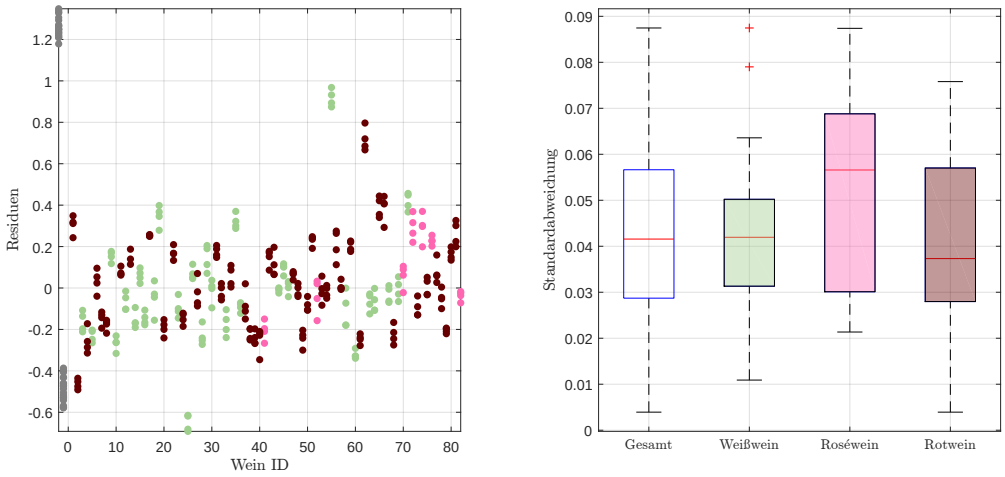

Abbildung 7.20: Residuen in g/l der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung der Weinsäurekonzentration im NN-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

Wie bei den bereits zuvor analysierten Bestandteilen zeigen sich beim Datensatzvergleich leichte Vorteile bei der Kalibrierung für die Engine E25, wobei die Reproduzierbarkeit in ebendiesem Datensatz vergleichsweise hohe Werte ausweist, auch wenn diese, wiederum mit dem Verweis auf die Genauigkeit der Referenzwerte, kaum relevante Unterschiede zeigen (vgl. Tabelle 7.15).

7.7 L-Äpfelsäure

Den zweitgrößten Wertebereich der Einzelsäure des Datensatzes 2016, mit einem der Weinsäure vergleichbaren Wertebereich von $[0.00, 3.70]\,\mathrm{g/l}$, bildet die L-Äpfelsäure. Wie nachfolgende Analysen zeigen, liefern die neuronalen Netzwerke ein plausibles Modell, welches eine Kalibrierung anhand von vier Wellenzahlblöcken mit je 10 Messstellen und zwei versteckten Neuronen, sowie der zweiten Savitzky-Golay Ableitung durchführt. Dies ist in Abbildung 7.21 (li.) grafisch dargestellt. Drei Viertel dieser Wellenzahlen liegen im Fingerprintbereich, angesiedelt bei circa $1\,220\,\mathrm{cm}^{-1}$, sowie nach dem Peak bei $1\,420\,\mathrm{cm}^{-1}$. Außerhalb des Wertebereiches selektiert die Heuristik einen Bereich unmittelbar nach der H_2O-Bande. In letztgenanntem Segment scheint aufgrund des Verhaltens der Spektren kaum chemische Information vorhanden zu sein, weshalb dieses als eine Art Kompensation für unbekannte Effekte dient. Bei den Residuen der doppelten Kreuzvalidierung kann für die Rotweine wiederum eine relativ geringe Schwankungsbreite, geschuldet der geringen Konzentrationen in dieser Stichprobe im Vergleich zur vorliegenden Population, beobachtet werden. Mit zunehmender L-Äpfelsäurekonzentration (bei den Weiß- bzw. Roséweinen) zeigt sich eine meist größere Streuung. Hierbei muss berücksichtigt werden, dass sich in Abbildung 7.21 (re.) teilweise Weißweine mit einem ähnlichen, gemittelten Residuum überlagern, was zu einer grafischen Überschätzung der Variabilität führt, allerdings nichts an der Tatsache ändert, dass hier erwartungsgemäß die Variabilität in einem höheren Bereich liegt.

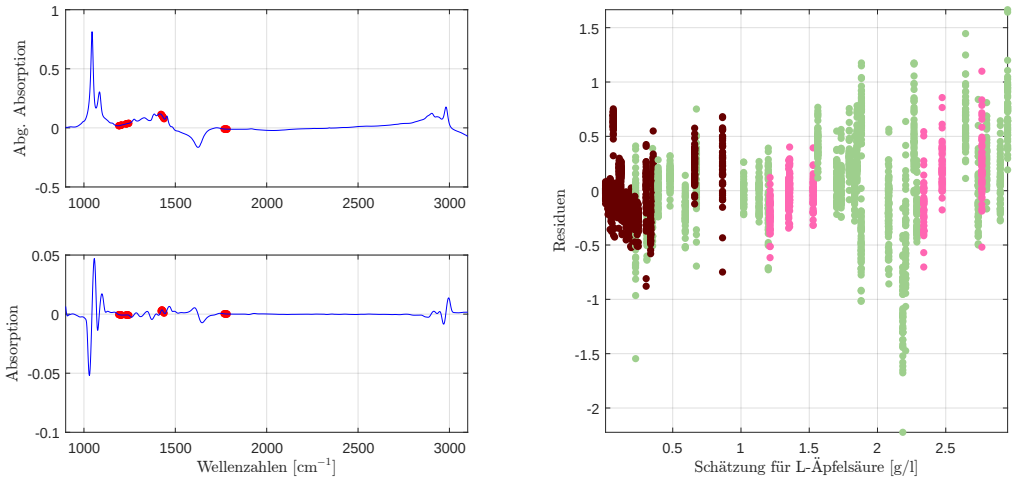

Abbildung 7.21: Die selektierten Wellenzahlen im NN-Modell für L-Äpfelsäure mit der ersten Savitzky-Golay Ableitung (li.) und der aus der doppelten Kreuzvalidierung resultierende Residuenplot (re.).

Da die Hälfte aller Weine (mehrheitlich Rotweine) eine L-Äpfelsäurekonzentration von $0\,\mathrm{g/l}$ aufweisen, finden sich diese im klassischen Residuenplot in Abbildung 7.22 (li.) auf einer Geraden wieder. Da in der Regel die neuronalen Netzwerke aufgrund der Einstellungen in dieser Arbeit keine negativen Konzentrationen

schätzt, führt dies zu einer systematischen Überschätzung der Rotweine, während die von 0 g/l verschiedenen Rotweine unterschätzt werden. Hierbei kann sogar ein starker Ausreißer in dieser Weinfarbe beobachtet werden. Die Weißweine scheinen einer leichten Struktur zu folgen, wobei Weine, welche sich in einem nahen Wertebereich der höchsten Datenkonzentration (bis circa 1.5 g/l) befinden, hauptsächlich über-, respektive ab dieser Grenze mehrheitlich unterschätzt werden. Dies kann allerdings nur teilweise auf die Tatsache der Transformation der Schätzungen bzw. auf die Datenverteilung zurückgeführt werden, da sich bei einer gesonderten Kalibrierung mit den Weiß- und Roséweinen (in allen Datensätzen) ein durchwegs klar zu erkennender steigender (linearer) Trend widerspiegelt. Betrachtet man hingegen in diesem Gesamtmodell die tatsächlichen L-Äpfelsäurekonzentrationen gegen die Schätzwerte in Abbildung 7.22 (re.), so liegen die Werte allesamt relativ nahe bei der in rot gestrichelten (Ideal-)Linie, welche eine perfekte Schätzung repräsentiert.

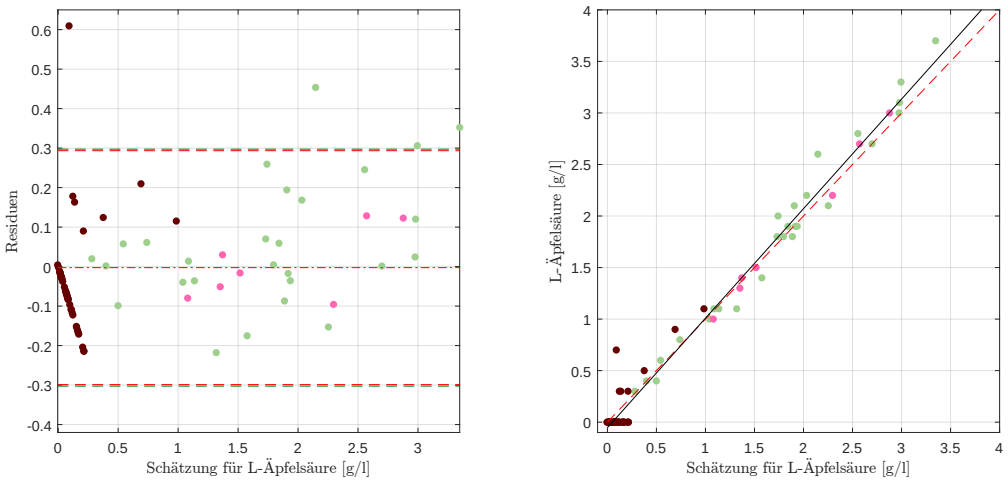

Abbildung 7.22: Residuenplot des NN-Modells für L-Äpfelsäure mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.), sowie die Schätzungen gegen die tatsächlichen Werte mit optimaler Schätzline, d.h. der rot gestrichelten Diagonale, sowie der durch die Daten motivierte lineare Zusammenhang in Schwarz (re.).

Mit einer Standardabweichung der Residuen in Höhe von 0.1481 g/l zeigt das Gesamtmodell eine der Referenzmethode ähnliche Variabilität, wobei der interquartile Bereich einen beinahe identischen Wert aufweist. Im Vergleich hierzu beläuft sich der MSE auf $0.0217(\,\mathrm{g/l})^2$, was eine einigermaßen genaue Schätzung bedeutet.

Reproduzierbarkeit

Betrachtet man die Grafiken mit einem Überblick über die Reproduzierbarkeit in Abbildung 7.23, so zeigt sich für den gesamten Datensatz lediglich ein Ausreißer, welcher den Weißweinen zuzuordnen ist. Wie aus den Residuenplots zu erwarten, zeigt sich bei den Rotweinen die mit Abstand beste Reproduzierbarkeit mit

einer mittleren, maximalen Abweichung von 0.0421 g/l, während die maximale Abweichung sich hierbei, gerundet auf die Genauigkeit der vorliegenden Daten, auf lediglich 0.1 g/l beläuft.

Die auf eine Nachkommastellen gerundeten Genauigkeiten der maximalen Abweichung mit der jeweiligen Anzahl der Weine sind in Tabelle 7.16 gelistet. So unterscheiden sich die (höchstens) vier Messungen der Rotweine um maximal 0.1 g/l, wohingegen die Messunterschiede bei Rosé- und Weißweinen sich mehrheitlich auf $\{0.1\,\mathrm{g/l}, 0.2\,\mathrm{g/l}\}$ belaufen.

	Anzahl nach Weinfarben			
g/l	Weißwein	Roséwein	Rotwein	$\sum$
0.0	4	0	31	35
0.1	10	3	13	26
0.2	13	3	0	16
0.3	3	1	0	4

Tabelle 7.16: Auf die Datengenauigkeit (10^{-1} g/l) gerundete, maximale Abweichung der Einzelspektren pro Wein ID für L-Äpfelsäure, aufgeschlüsselt nach Weinfarbe.

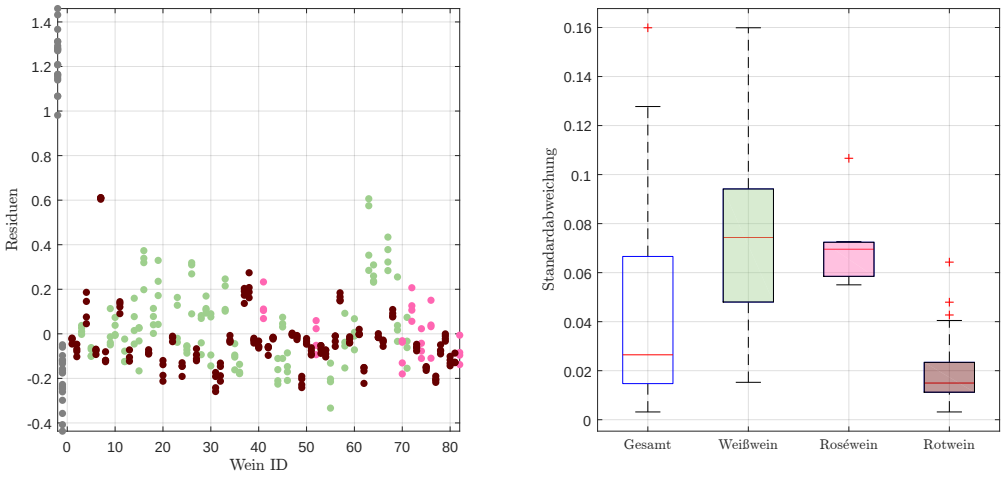

Abbildung 7.23: Residuen in g/l der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung der L-Äpfelsäurekonzentration im NN-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

Im Datensatzvergleich zeigt sich auch für die L-Äpfelsäure das beste Verhalten der Residuen innerhalb des Datensatzes E25, während die Reproduzierbarkeit minimale Verschlechterungen gegenüber den Spektrometern E22 und V70(2016), nicht aber E24 aufweist. Berücksichtigt man zusätzlich die Genauigkeit der vorliegenden Daten (10^{-1} g/l), so sind die Unterschiede nicht/kaum wahrnehmbar.

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0712	0.0743	0.3216	0.1576	0.1566	0.0787
Roséwein	0.0707	0.0696	0.2602	0.1632	0.1507	0.0443
Rotwein	0.0190	0.0150	0.1409	0.0421	0.0329	0.0274
Gesamt	0.0428	0.0265	0.3216	0.0953	0.0613	0.0788

Tabelle 7.17: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für L-Äpfelsäure im NN-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.1850	0.17	−0.32	0.60	0.0385	0.0244	0.2809
E24	0.1573	0.13	−0.35	0.55	0.0428	0.0266	0.4690
E25	0.1481	0.14	−0.22	0.61	0.0428	0.0265	0.3216
V70(2016)	0.1872	0.17	−0.30	0.66	0.0408	0.0276	0.4302

Tabelle 7.18: Performance des entwickelten NN-Modells für L-Äpfelsäure, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

7.8 Milchsäure

Im Gegensatz zur L-Äpfelsäure beinhalten Rotweine tendenziell mehr Milchsäure als Weißweine. Anhand dieser beiden Säurekonzentrationen kann somit die Klassifizierung nach der Weinfarbe (Rotwein oder kein Rotwein) mit einer hohen Zuverlässigkeit bestimmt werden. Es zeigt sich im Residuenplot der doppelten Kreuzvalidierung in Abbildung 7.24 (re.) ein ähnliches Bild wie in jenen für L-Äpfelsäure für die Farbtrennung. Bezüglich der Variabilität zeigt sich wiederum für einen geringen Bereich, hier bis circa 0.4 g/l, sowie für die restlichen Weine eine jeweils homogene Variabilität. Während für ebendiese Säure aus dem vorangegangenen Abschnitt 7.7 zwei versteckte Neuronen mit insgesamt 40 Wellenzahlen zur Modellbildung herangezogen wurden, so reduziert sich die Anzahl von zwei auf lediglich ein verstecktes Neuron, wohingegen sich die Wellenzahlen auf zwei Blöcke der Länge 25 erhöhen, welche im Vergleich zur L-Äpfelsäure ähnlich positioniert sind und somit eine vergleichbare Information liefern, wie Abbildung 7.24 (li.) zeigt. Darüber hinaus geschieht die Spektrenaufbereitung mit der Savitzky-Golay Ableitung zweiter Ordnung.

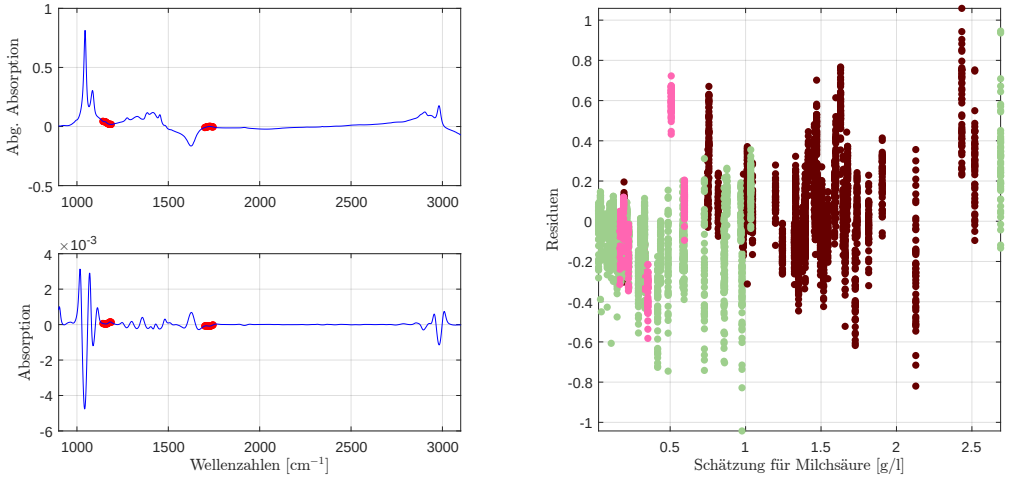

Abbildung 7.24: Die selektierten Wellenzahlen im NN-Modell für Milchsäure mit der zweiten Savitzky-Golay Ableitung (li.) und der aus der doppelten Kreuzvalidierung resultierende Residuenplot (re.).

Im Residuenplot in Abbildung 7.25 tritt ein geringfügiger Bias in Höhe von 0.02 g/l auf, wohingegen ein klar positiver (negativer) Offset der Rotweine (Weißweine) beobachtet werden kann und wie in vorigem Abschnitt durch die Beschaffenheit der neuronalen Netzwerke zum Teil erklärt wird. Zudem tritt eine leicht steigende Tendenz in den Rotweinen, sowie ein fallendes Verhalten in den Weißweinen auf. Aus diesem Grund scheint eine Kalibrierung, getrennt nach den Weinfarben für die Milchsäure unumgänglich zu sein, trotz einer residualen Standardabweichung von 0.1380 g/l des Gesamtmodells, welche nur geringfügig über jener der Referenzmethode liegt.

Betrachtet man zunächst das bestehende Gesamtmodell mit einer Kalibrierung

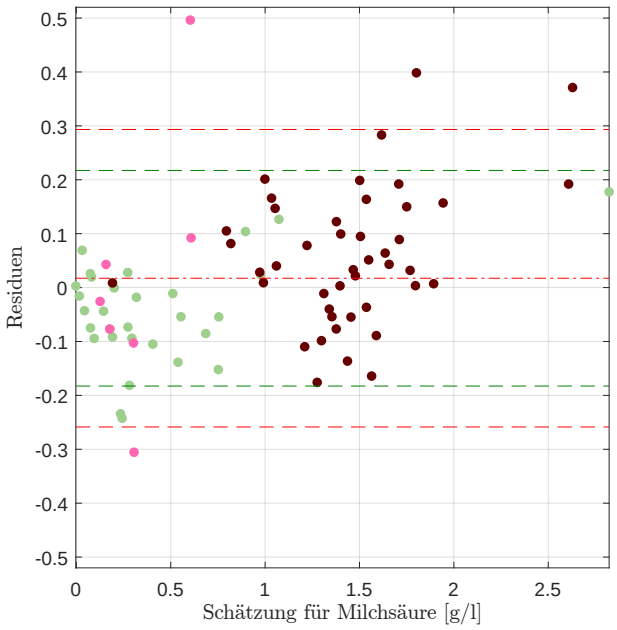

Abbildung 7.25: Residuenplot des NN-Modells für Milchsäure mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode.

durch die Rotweindaten, so kann in diesem reduzierten Modell der Bias (beinahe) eliminiert werden, wobei sich ein klarer, mit der Schätzung der Milchsäurekonzentration steigender Trend in den Residuen manifestiert, welcher nicht nur durch die beiden Ausreißer mit den betragsmäßig größten Residuen suggeriert wird, wie Abbildung 7.26 (li.) zeigt. Insgesamt reduziert sich durch diese Neukalibrierung der Bias der Rotweine von 0.06 g/l auf unter 10^{-2} g/l, mit einer Erhöhung der dazugehörenden Standardabweichung von 0.1277 g/l[6] auf 0.1593 g/l. Ein möglicher Mitgrund für das Versagen dieses Teilmodells bildet das Verhältnis von einer großen Anzahl von Prädiktoren zu einer äußerst geringen Datenmenge[7], welche zur Kalibrierung herangezogen wird. Vergleicht man dieses Submodell mit jenem für die Rotweine spezifizierten Modell, indem eine erneute Wellenzahlbestimmung durchgeführt wird, so zeigt Abbildung 7.26 (re.), dass die Standardabweichung in den Residuen bei einem Bias von 0.01 g/l auf 0.1877 g/l steigt. Dieses Modell kommt jedoch mit deutlich weniger Prädiktoren (4×5 Blöcken) aus, welche ähnlich verteilt sind, wie die im Modell der L-Äpfelsäure mit den neuronalen Netzwerken. Daher ist dieses Teilmodell für Rotweine vorzuziehen, zumal sich einerseits die absolut größten Ausreißer etwas reduzieren und insbesondere kein Trend in den Residuen mehr beobachtbar ist.

Für das Gesamtmodell, kalibriert ausschließlich mit Weiß- und Roséweinen zeigt

[6] Standardabweichung der Rotweine des Gesamtmodells ohne Neukalibrierung

[7] Tatsächlich verringert sich die Anzahl der zur Verfügung stehenden Weine durch das Absplitten eines Validierungdatensatzes ein weiteres Mal.

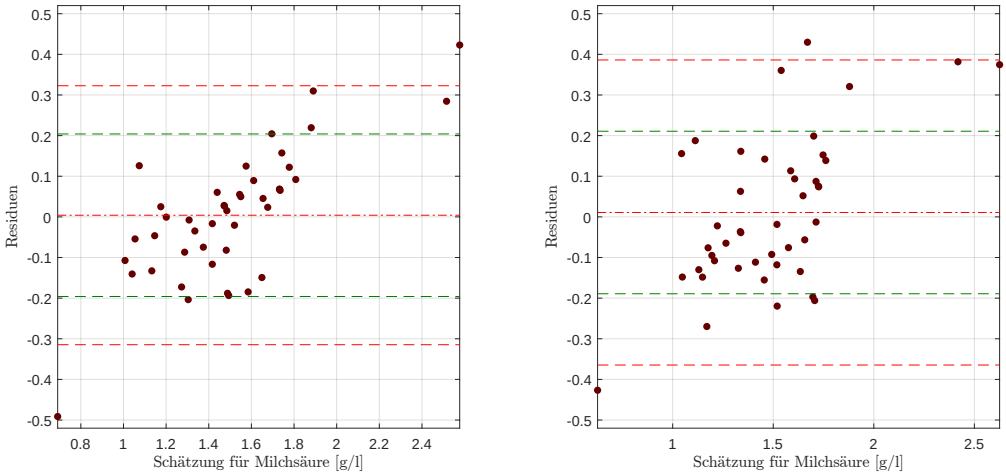

Abbildung 7.26: Residuenplot des NN-Modells für Milchsäure mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode, mit Neukalibrierung des Gesamtmodells für Rotweine (li), sowie der Residuenplot des rotweinspezifischen NN-Modells mit den entsprechenden Kennzahlen (re.).

sich im Vergleich zu Abbildung 7.25 ein noch deutlicher Trend. Es kann daher kein akzeptables Modell für diese Weinklasse gefunden werden.

Reproduzierbarkeit

Für die Reproduzierbarkeit ergibt sich eine gemittelte Standardabweichung von 0.0673 g/l im Gesamtmodell, was wiederum in Anbetracht der Genauigkeit der Daten als akzeptabel aufgefasst werden kann, auch wenn die maximale Abweichung eines Rotweines bei 0.4321 g/l liegt, wobei es insgesamt 3 Rotweine gibt, welche eine vergleichbar hohe maximale Abweichung von gerundet 0.4 g/l aufweisen, wie zusätzlich zu Tabelle 7.19 aus der Übersicht 7.20 entnommen werden kann. Im Vergleich zu nachfolgender, empirischer Analyse über die Modellierung der Zitronensäure, Unterabschnitt Datenmanipulation, mittels neuronaler Netzwerke kann diese hohe Abweichung nicht durch einzelne Weine begründet werden und es zeigt sich bei Betrachtung der maximalen Differenz eines Weines eine im Vergleich zur Streuung schlechte Wiederholbarkeit, insbesondere für Rotweine.

Für das empfohlene Submodell für Rotweine reduziert sich die maximale Abweichung in den Reproduzierbarkeitsdaten auf 0.27 g/l und liefert, trotz der Verschlechterung der Standardabweichung der Residuen eine stark verbesserte Reproduzierbarkeit.

In Tabelle 7.21 findet sich der Datensatzvergleich des Gesamtmodells. Hierbei weisen die Residuen des Datensatzes E25 wiederum die geringste Standardabweichung auf, wohingegen die Reproduzierbarkeit nicht zufriedenstellende Werte aufweist,

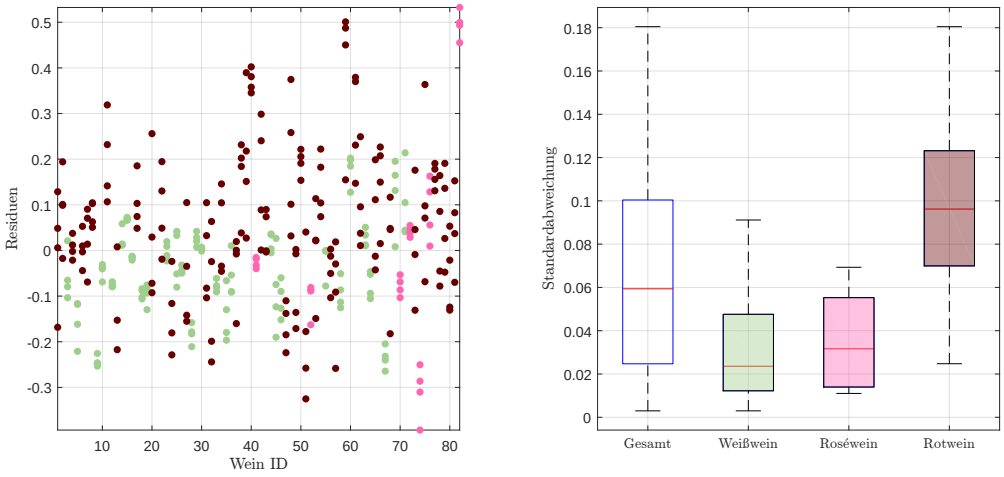

Abbildung 7.27: Residuen in g/l der Einzelspektren pro Wein, ohne Kunstweine, zur Schätzung der Milchsäurekonzentration im NN-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0314	0.0236	0.2165	0.0693	0.0534	0.0549
Roséwein	0.0349	0.0316	0.1532	0.0793	0.0772	0.0521
Rotwein	0.0969	0.0962	0.4321	0.2155	0.2130	0.0943
Gesamt	0.0673	0.0594	0.4321	0.1496	0.1304	0.1064

Tabelle 7.19: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Milchsäure im NN-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

	Anzahl nach Weinfarben			
g/l	Weißwein	Roséwein	Rotwein	$\sum$
0.0	14	2	0	16
0.1	13	4	11	28
0.2	3	1	19	23
0.3	0	0	11	11
0.4	0	0	3	3

Tabelle 7.20: Auf die Datengenauigkeit (10^{-1} g/l) gerundete, maximale Abweichung der Einzelspektren pro Wein ID für Milchsäure, aufgeschlüsselt nach Weinfarbe.

da die maximale Auslenkung einen Wert von knapp über 0.4 g/l annimmt. Zugleich weist dieser Datensatz mit einer mittleren Standardabweichung in der Höhe von 0.0673 g/l die vergleichsweise höchste Variabilität auf.

In dem für Rotweine akzeptableren und aus mathematischer Sicht gerechtfertigten Modell zeigt sich bei allen Spektrometern eine Verschlechterung im Verhalten der Residuen (im Sinne einer größeren Streuung), wohingegen die Reproduzierbarkeit im Datensatz E25, sowie E24 verbessert wird, allerdings zu einer Verschlechterung für E22 und V70(2016) führt.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.1611	0.17	−0.29	0.60	0.0498	0.0408	0.3529
E24	0.1642	0.14	−0.30	0.59	0.0519	0.0501	0.3454
E25	0.1380	0.17	−0.31	0.50	0.0673	0.0594	0.4321
V70(2016)	0.1875	0.18	−0.34	0.72	0.0464	0.0387	0.2838

Tabelle 7.21: Performance des entwickelten NN-Modells für Milchsäure, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.1949	0.22	−0.42	0.47	0.0703	0.0654	0.3748
E24	0.2453	0.32	−0.48	0.59	0.0454	0.0456	0.2028
E25	0.1877	0.26	−0.43	0.43	0.0571	0.0557	0.2651
V70(2016)	0.2281	0.33	−0.49	0.58	0.0544	0.0551	0.2705

Tabelle 7.22: Performance des rotweinspezifischen NN-Modells für Milchsäure, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

7.9 Flüchtige Säuren

Mit einem Wertebereich von $[0.20, 1.70]$ g/l weisen die flüchtigen (volatilen) Säuren den viertgrößten Wertebereich innerhalb der hier untersuchten Bestandteile der titrierbaren Säuren auf. Hierfür selektieren die Heuristiken in Kombination mit den neuronalen Netzwerken 20 Wellenzahlen, aufgeteilt in 4×5 Intervalle. Zwei Intervalle befinden sich unmittelbar vor bzw. nach der ausgeschlossenen H_2O-Bande, wohingegen die restlichen 10 Messstellen einerseits im Fingerprintbereich bei 1 300 cm^{-1} bzw. außerhalb dieses Bereiches bei einer Wellenzahl von 2 610 cm^{-1}

liegen und sind in Abbildung 7.28 anhand eines Spektrums mit dessen ersten Savitzky-Golay Ableitung dargestellt.

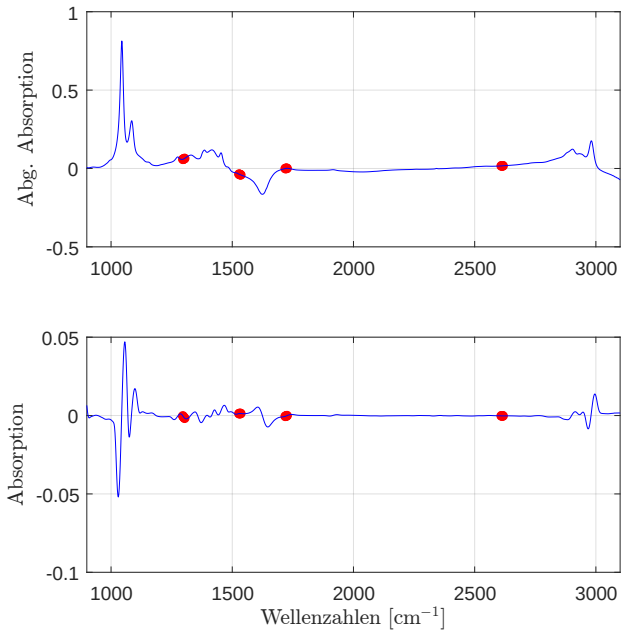

Abbildung 7.28: Die selektierten Wellenzahlen im NN-Modell für die flüchtigen Säuren mit der ersten Savitzky-Golay Ableitung.

Die neuronalen Netzwerke verwenden hierfür lediglich ein verstecktes Neuron und es resultiert der Residuenplot aus der doppelten Kreuzvalidierung in Abbildung 7.29 (li.), mit den Residuen als Histogramm in der rechten Grafik. Es zeigt sich bei einer geringen Konzentration von flüchtigen Säuren ein einigermaßen unauffälliges Verhalten, wohingegen jene Weine mit größeren Schätzwerten stärker streuen. Dies ist wiederum zurückzuführen auf die hohe Datenkonzentration im Punkt $0.5 \pm 0.25\,\mathrm{g/l}$ in Kombination mit dem Vorgehen der doppelten Kreuzvalidierung. Insgesamt sind die Residuen einigermaßen symmetrisch um $0\,\mathrm{g/l}$ verteilt und zeigen für den Wertebereich bis $1\,\mathrm{g/l}$ kein auffälliges Verhalten.

Analysiert man den klassischen Residuenplot in Abbildung 7.30, so findet sich erwartungsgemäß der Großteil der Residuen auf 4 parallel verlaufenden Geraden (75 % der Daten), wie in Abschnitt 5.9 erläutert. Ausreißer können für dieses Modell nicht beobachtet werden, ebenso wie keine ungewöhnliche Struktur der Residuenplots aufzufinden ist. Es wird mit einer Standardabweichung von $0.0841\,\mathrm{g/l}$ die Genauigkeit der Referenzmethode leicht unterschritten, bei einem Bias in Höhe von weniger als $10^{-2}\,\mathrm{g/l}$. Mit einem MSE in einer geringen Höhe von $0.0070(\,\mathrm{g/l})^2$ liegt dieser somit knapp unter der Genauigkeit der vorliegenden Daten von $10^{-1}\,\mathrm{g/l}$, da $\sqrt{0.0070} \approx 0.0837\,\mathrm{g/l}$ gilt.

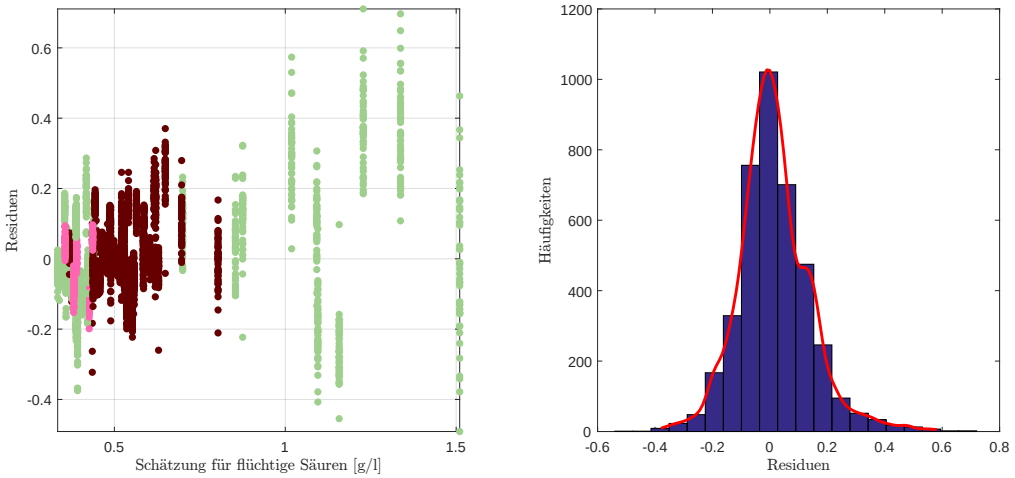

Abbildung 7.29: Die Weinresiduen der doppelten Kreuzvalidierung als Residuenplot (li.) und zusammengefasst zu einem Histogramm (re.) mit einer empirischen Dichteschätzung im NN-Modell für flüchtige Säuren.

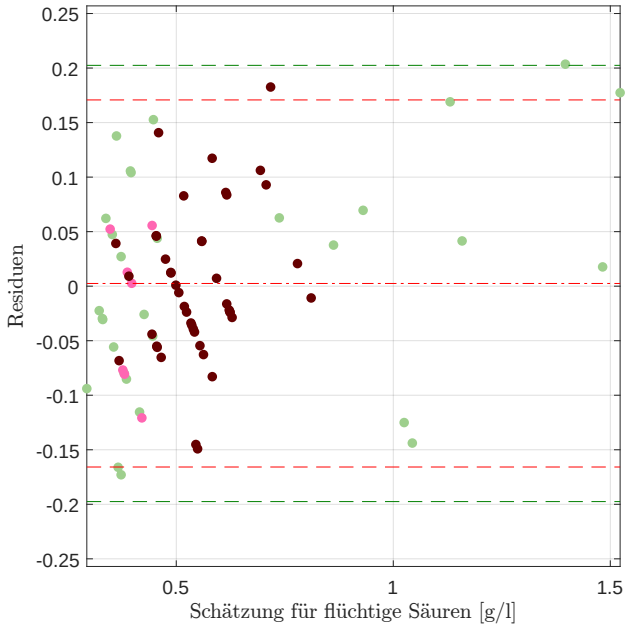

Abbildung 7.30: Residuenplot des NN-Modells für flüchtige Säuren mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode.

Reproduzierbarkeit

Ein ähnlich unauffälliges Bild wie die Residuen zeigen die zugehörenden Reproduzierbarkeitsplots in Abbildung 7.31. Es können lediglich zwei Weißweine als Ausreißer bezüglich der Standardabweichung beobachtet werden, wobei diese jeweils unter 0.06 g/l liegen. Die höchste maximale Auslenkung der Wiederholungsmessungen misst 0.1327 g/l, wie in Tabelle 7.23 nachgelesen werden kann. Auf die Genauigkeit von 10^{-1} g/l zurückprojiziert, bedeutet dies eine maximale Abweichung von 0.1 g/l. Insgesamt gliedert sich die gerundete Reproduzierbarkeit in 34 mal 0 g/l und 46 mal 0.1 g/l.

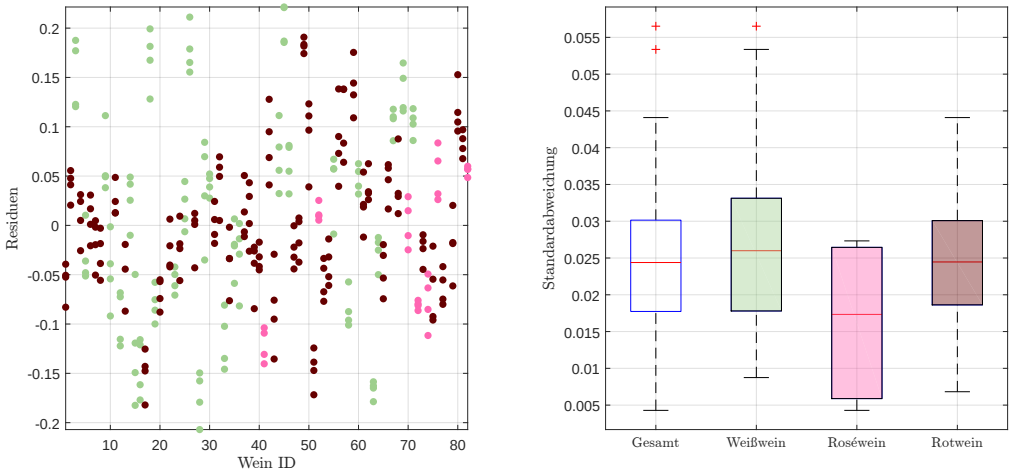

Abbildung 7.31: Residuen in g/l der Einzelspektren pro Wein, ohne Kunstweine, zur Schätzung der flüchtigen Säure im NN-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0262	0.0260	0.1327	0.0582	0.0566	0.0259
Roséwein	0.0163	0.0173	0.0623	0.0360	0.0365	0.0223
Rotwein	0.0247	0.0245	0.1062	0.0561	0.0552	0.0193
Gesamt	0.0246	0.0244	0.1327	0.0552	0.0550	0.0228

Tabelle 7.23: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für flüchtige Säuren im NN-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

Beim Vergleich aller Datensätze zeigt sich ein ähnliches Verhalten innerhalb aller Engines E22, E24 und E25, sowohl bei den Residuen als auch bezüglich der Reproduzierbarkeit. Aus diesem Grund kann von einem äußerst adäquaten und guten

Modell an dieser Stelle gesprochen werden, sofern die Fehlmessungen ausgeschlossen werden.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.0845	0.12	−0.19	0.21	0.0251	0.0227	0.1522
E24	0.0836	0.10	−0.19	0.19	0.0258	0.0228	0.1971
E25	0.0841	0.10	−0.17	0.20	0.0246	0.0244	0.1327
V70(2016)	0.0880	0.13	−0.16	0.22	0.0371	0.0316	0.2089

Tabelle 7.24: Performance des entwickelten NN-Modells für flüchtige Säuren, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

7.10 Zitronensäure

Bei der letzten in dieser Arbeit analysierten Säure, mit dem geringsten Wertebereich, handelt es sich um die Zitronensäure. Das hier vorgestellte neuronale Netzwerk versucht mithilfe einer latenten Variable und 4×5 unterschiedlichen Wellenzahlblöcken die Konzentration dieser Fruchtsäure aus den vorhandenen Spektren abzuleiten. Die Glättung für dieses Modell erfolgt durch die zweite Savitzky-Golay Ableitung. Bei der Verteilung der Wellenzahlenblöcke in die vier unterschiedlichen Bereiche fällt auf, dass grundsätzlich keine Peaks selektiert werden, wie Abbildung 7.32 zeigt. Einer dieser Blöcke liegt unmittelbar am Rand des zugelassenen Wertebereiches bei $3\,030\,\mathrm{cm}^{-1}$, welcher durch analoge Überlegungen wie in Abschnitt 2.2 zur Begrenzung des Wellenzahlbereiches auf $[900, 3\,100]\,\mathrm{cm}^{-1}$ ebenfalls ausgeschlossen werden hätte können.

Ein ähnliches Bild wie bei der Verwendung eines PLS-Modells ergibt sich für die Modellbildung mit den neuronalen Netzwerken. Es kann wiederum ein einzelner Ausreißer beobachtet werden (Roséwein mit 1.2 g/l) und die restlichen Residuen zeigen, abgesehen von den, durch die Verteilung der vorliegenden Werte auf 7 unterschiedliche Konzentrationen bedingten, Geraden im Residuenplot, keine Auffälligkeiten. Zudem kann auch kein Trend oder eine Struktur beobachtet werden. Hierbei darf die Tatsache nicht außer Acht gelassen werden, dass die Weine mit einer Konzentration von 0 g/l von den neuronalen Netzwerken zumeist überschätzt werden und lediglich durch diese Tatsache die Residuen fälschlicherweise einen Trend suggerieren.

Um einen besseren Überblick über die Variabilität zu erhalten, können die wichtigsten Informationen der Residuen, geclustert nach deren tatsächlicher Zitronensäurekonzentration, der Tabelle 7.25 entnommen werden, wobei die Standardabweichung aufgrund der geringen Probenanzahl, insbesondere für Weine mit

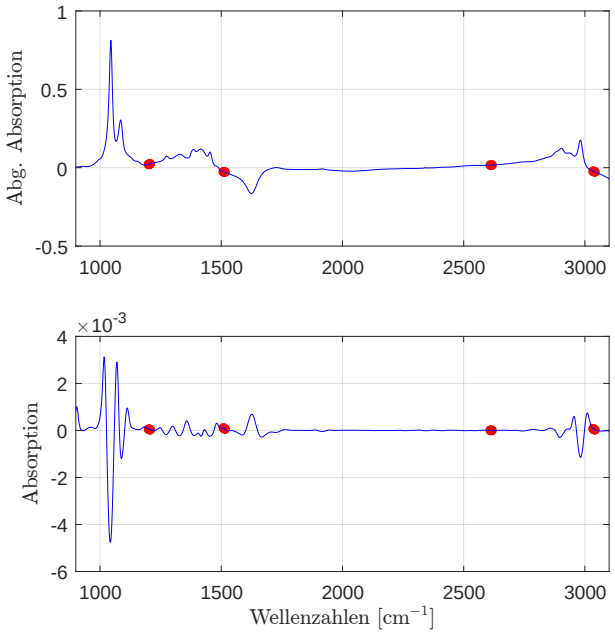

Abbildung 7.32: Die selektierten Wellenzahlen im NN-Modell für Zitronensäure mit der zweiten Savitzky-Golay Ableitung.

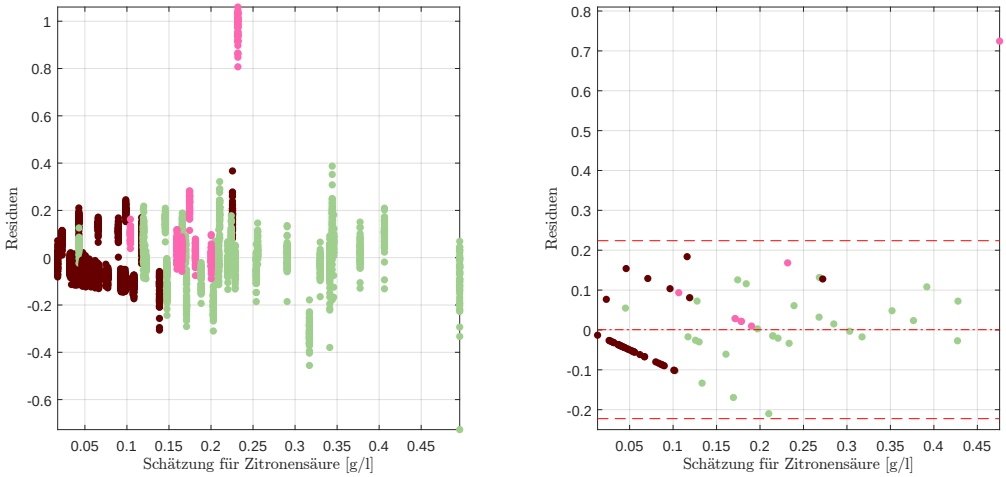

Abbildung 7.33: Die Weinresiduen der doppelten Kreuzvalidierung als Residuenplot (li.) und der Residuenplot des NN-Modells für Zitronensäure mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) (re.).

einer Konzentration von 0.10 g/l, 0.20 g/l, sowie 0.30 g/l, an Aussagekraft verliert. So ergeben sich einigermaßen plausible Schätzungen, insbesondere mit der Berücksichtigung der Genauigkeit der vorliegenden Daten (eine Nachkommastelle).

g/l	Anz.	Mw.	Med.	Std.	IQR	Min.	Max.
0.00	41	−0.06	−0.05	0.0393	0.05	−0.21	−0.01
0.10	6	−0.00	−0.02	0.0538	0.08	−0.06	0.08
0.20	17	0.04	0.03	0.0561	0.09	−0.03	0.15
0.30	8	0.06	0.05	0.0711	0.11	−0.02	0.18
0.40	6	0.08	0.09	0.0754	0.11	−0.03	0.17
0.50	2	0.09				0.07	0.11
1.20	1	0.72					

Tabelle 7.25: Verhalten der Residuen für gegebene Zitronensäurekonzentration im NN-Modell. Alle Werte in g/l.

Bemerkung 7.3. Passt man die Schätzungen an die Messgenauigkeit der tatsächlichen Konzentrationen an, indem diese auf die erste Nachkommastelle gerundet werden, so realisieren beispielsweise die Schätzungen für $y = 0\,\mathrm{g/l}$ insgesamt 20 Mal in 0 g/l und 19 Mal in 0.10 g/l. Für $y = 0.20\,\mathrm{g/l}$ resultieren diese mehrheitlich mit einer Anzahl von 10 Weinen in einer Konzentration von 0.20 g/l, während 6 Weinen ein Zitronensäuregehalt in Höhe von 0.10 g/l zugeschrieben wird.

Da der Großteil der Rotweindaten eine Konzentration von 0 g/l aufweisen, wird auch hier kein spezifisches Modell für Rotweine vorgestellt.

Reproduzierbarkeit

Betrachtet man die Reproduzierbarkeitsplots in Abbildung 7.34, so kann eine vergleichsweise hohe Standardabweichung in der Reproduzierbarkeit beobachtet werden, mit Ausnahme der Rotweine, deren gemittelte Standardabweichung bei 0.02 g/l liegt. Insgesamt ergibt sich die Übersicht wie in Tabelle 7.26.

Mit Abstand die schlechteste Reproduzierbarkeit weisen die Weißweine, mit einer Standardabweichung der beiden Ausreißer von über 0.15 g/l auf, auch wenn die mit Ausnahme eines weiteren Rotweines ebendiese Kennzahl unter der Messgenauigkeit von 10^{-1} g/l der Daten liegt, weshalb an dieser Stelle eine einigermaßen plausible Wiederholbarkeit vorliegt, auch wenn die maximale Abweichung eines Weißweines bei beinahe 0.5 g/l liegt. Hierbei handelt es sich allerdings um den Einfluss jenes Roséweines mit einer Konzentration von 1.20 g/l, wie der folgende Unterabschnitt Datenmanipulation, Seite 187 nahelegt.

Während im Datensatzvergleich die Residuen der Engine E25 die besten Kennzahlen aufweisen, so zeigt dieses Spektrometer die mit Abstand schlechteste Reproduzierbarkeit. Mit einer maximalen Abweichung von 0.45 g/l im Vergleich zu den restlichen Datensätzen von maximal 0.16 g/l.

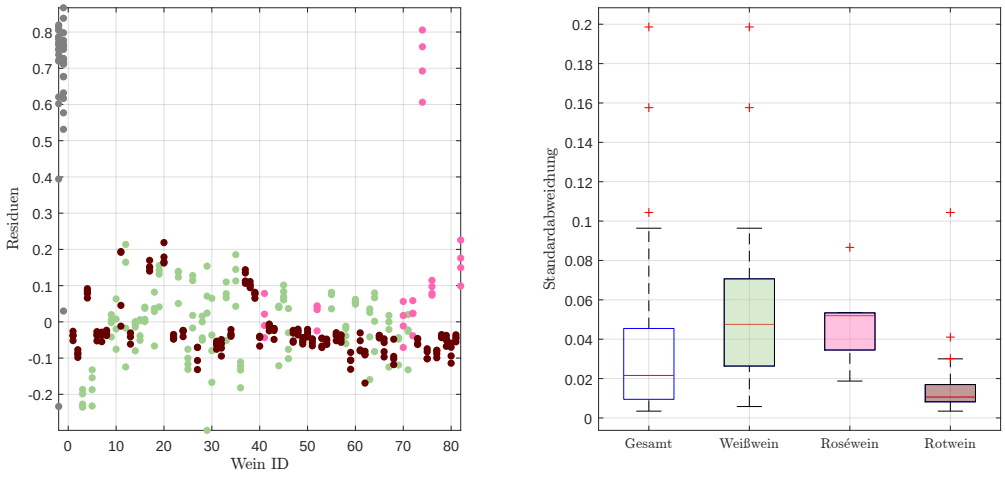

Abbildung 7.34: Residuen in g/l der Einzelspektren pro Wein, zuzüglich der Kunstweine in Grau, zur Schätzung der Zitronensäurekonzentration im NN-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0558	0.0475	0.4533	0.1240	0.1068	0.0924
Roséwein	0.0481	0.0519	0.1994	0.1117	0.1221	0.0505
Rotwein	0.0153	0.0106	0.2059	0.0333	0.0256	0.0316
Gesamt	0.0332	0.0216	0.4533	0.0737	0.0466	0.0761

Tabelle 7.26: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Zitronensäure im NN-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.1391	0.09	−0.31	0.97	0.0184	0.0180	0.0914
E24	0.1196	0.09	−0.17	0.80	0.0260	0.0231	0.1585
E25	0.1116	0.09	−0.21	0.72	0.0332	0.0216	0.4533
V70(2016)	0.1367	0.09	−0.25	0.97	0.0181	0.0160	0.1152

Tabelle 7.27: Performance des entwickelten NN-Modells für Zitronensäure, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

Datenmanipulation

Die vergleichbar hohe, maximale Abweichung der Reproduzierbarkeit im Datensatz E25, wie diese durch Tabelle 7.27 ausgedrückt wird, kann unter anderem auf den Einfluss des Roséweines mit einer Zitronensäurekonzentration von 1.2 g/l zurückgeführt werden, wie durch die Tabellen 7.28 bzw. 7.29 verifiziert wird.

Grundsätzlich dürfen einzelne Stichproben (insbesondere Ausreißer) nicht ausgeschlossen werden, um ein besseres Modell zu erhalten, sofern diese nicht auf (Mess-)Fehlern beruhen, wie in dieser Situation. Aus diesem Grund wird obiges Modell, kalibriert mit allen 82 Weinen, abzüglich des inkorrekten Spektrums (ID 21), sowie des Roséweines mit der ID 74 (1.2 g/l), in einem getrennten Unterabschnitt über „Datenmanipulation“ mit Hinblick auf die Reproduzierbarkeit untersucht.

Betrachtet man die Tabelle 7.28 mit deren Reproduzierbarkeitsgütemaßen, so kann im Vergleich zum eigentlichen Modell in Tabelle 7.26 insbesondere in den Kategorien Weiß- und Roséweinen eine signifikante Verbesserung festgestellt werden. Beispielsweise kann in allen drei Weinklassen die maximale Abweichung innerhalb der vier Messwiederholungen mehr als halbiert werden. Dies gilt auch insbesondere für Weiß- und Rotweine, welche im Gesamtmodell allesamt mitmodelliert werden. So reduziert sich die maximale Abweichung bei den Weißweinen von 0.45 g/l auf 0.11 g/l.

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0260	0.0259	0.1108	0.0582	0.0578	0.0234
Roséwein	0.0221	0.0220	0.0626	0.0519	0.0518	0.0078
Rotwein	0.0141	0.0124	0.0856	0.0310	0.0276	0.0149
Gesamt	0.0191	0.0162	0.1108	0.0428	0.0355	0.0223

Tabelle 7.28: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Zitronensäure im NN-Modell, wobei der Roséwein mit 1.2 g/l für die Kalibrierung und Auswertung ignoriert wird, um den Einfluss dieses Extremwertes zu demonstrieren, anhand der Einzelspektren pro Wein ID. Alle Werte in g/l.

Während sich die Gütemaße der Residuen, aufgrund der entarteten Form des ausgeschlossenen Roséweines, verbessern, können ebenso Verbesserungen im Datensatz E24 bezüglich der Reproduzierbarkeit beobachtet werden, wohingegen sich die maximale Wiederholbarkeit im Datensatz E22 nicht merklich verschlechtert und somit die Güte von E22 im Allgemeinen unterstreicht.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.0840	0.08	−0.29	0.24	0.0193	0.0201	0.0977
E24	0.0747	0.08	−0.19	0.23	0.0221	0.0198	0.1163
E25	0.0802	0.09	−0.22	0.20	0.0191	0.0162	0.1108
V70(2016)	0.0819	0.08	−0.25	0.23	0.0166	0.0149	0.1160

Tabelle 7.29: Performance des entwickelten NN-Modells, wobei der Roséwein mit 1.2 g/l für die Kalibrierung und Auswertung ignoriert wird, um den Einfluss dieses Extremwertes zu demonstrieren, für Zitronensäure, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

7.11 Glycerin

Für Glycerin kann mit den hier beschriebenen Heuristiken und Parametereinstellungen in Kombination mit den neuronalen Netzwerken kein passendes Gesamtmodell ermittelt werden. Hierbei versucht das neuronale Netzwerk, sämtliche Informationen aus einem Bereich von 9 Wellenzahlen mit lediglich einer latenten Variable herauszulesen. Da sich zwei Drittel dieser Prädiktoren am Rand des erlaubten Spektrums bei 900 cm^{-1} befinden, kann ein Restrauschen, welches Einfluss auf die Güte des Modells nimmt, nicht ausgeschlossen werden. Darüber hinaus zeigen die Weißweinresiduen einen Trend bei einer Standardabweichung von circa 0.80 g/l und einem interquartilen Bereich von knapp unter 1 g/l.

Aus diesem Grund wird für Glycerin an dieser Stelle lediglich ein Modell für die Rotweine vorgestellt. Das Parametersetting setzt sich aus zwei Wellenzahlbereichen mit je 10 Prädiktoren zusammen. Dies ist in Abbildung 7.35 (li.) dargestellt, wobei die erste Savitzky-Golay Ableitung sowie eine latente Variable verwendet werden. Wie bereits im unpassenden Gesamtmodell, selektiert der Algorithmus die Hälfte der Prädiktoren in jenem Bereich von 900 cm^{-1}. Obwohl dieses Modell eine geringere Standardabweichung als die Referenzmethode aufweist, lässt sich eine (leicht) konische Form der Residuen erkennen. Es scheint die Variabilität der Residuen ab einer Konzentration von 7.5 g/l zu steigen, weshalb auch an dieser Stelle das für Rotweine heuristisch beste Modell nicht als für die Daten adäquat aufgefasst werden kann, weshalb an dieser Stelle auf eine weitere detailliertere Beschreibung und Analyse der Residuen verzichtet wird.

Reproduzierbarkeit

Eine mögliche Begründung, weshalb das Modell als für die Daten nicht passend zu sein scheint, kann anhand der Reproduzierbarkeit eine mögliche Begründung für das Versagen dieser Parametrisierung beobachtet werden. In Abbildung 7.36 weichen die Residuen der künstlichen Weine auffallend von den Rotweinen ab.

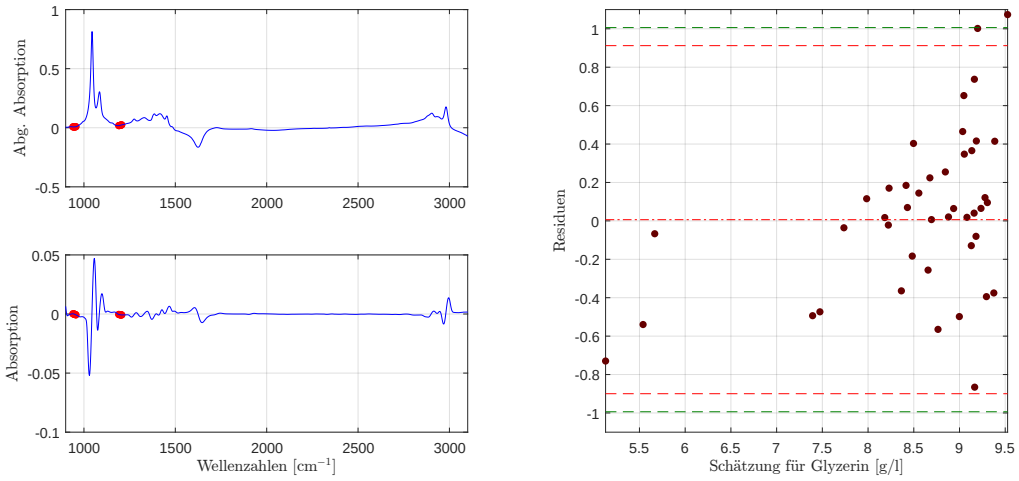

Abbildung 7.35: Die selektierten Wellenzahlen im rotweinspezifischen NN-Modell für Glyzerin mit der ersten Savitzky-Golay Ableitung (li.) und der dazugehörende Residuenplot mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (re.).

Dies deutet darauf hin, dass das neuronale Netzwerk einen wichtigen Teil der Informationen zur Kalibrierung aus den Wechselwirkungen einzelner Bestandteile filtert, was wiederum das konische Verhalten der Residuen erklären kann.[8] Die größere Variabilität der Kunstweine kann hierbei auf die unterschiedlichen Messzeitpunkte der Mehrfachmessungen zurückgeführt werden.

Eine weitere Auffälligkeit bildet die Tatsache, dass das Modell, kalibriert und ausgewertet mit dem Datensatz V70(2016) die geringsten Residuenkennzahlen erzielen (mit Ausnahme des geringfügig größeren Residuums nach oben, verglichen mit E25), allerdings die schlechteste Reproduzierbarkeit in sämtlichen Kennzahlen liefert. Da der hierfür ermittelte Median von 0.1736 g/l einen zum Mittelwert mit 0.1918 g/l vergleichbar hohen Wert aufweist, bedeutet, dass, im Gegensatz zu den Datensätzen der Engines der zweiten Generation nicht wenige Ausreißer die Reproduzierbarkeit beeinflussen und daher eine vergleichsweise schlechte Wiederholbarkeit vorliegt.

[8] Die Kunstweine beinhalten nicht alle in dieser Arbeit analysierten Bestandteile.

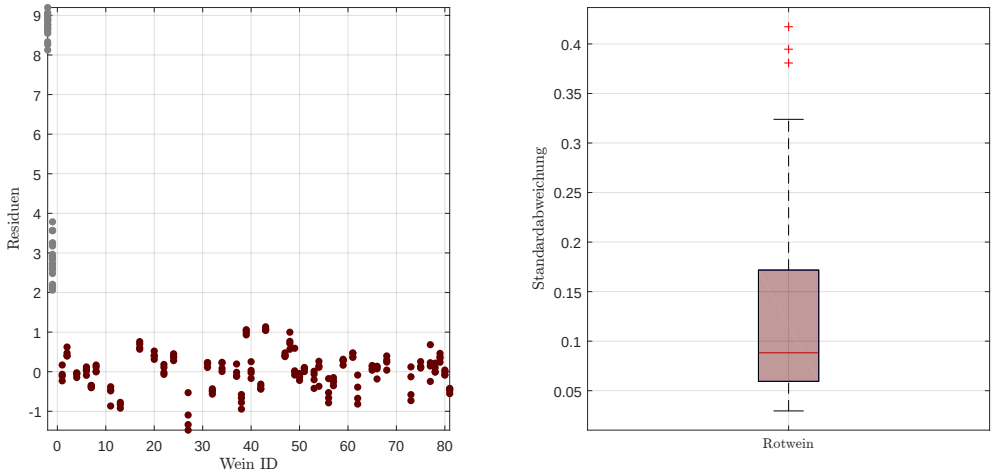

Abbildung 7.36: Residuen in g/l der Einzelspektren pro Rotein, zuzüglich der Kunstweine in Grau, zur Schätzung der Glyzerin im NN-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot ohne künstliche Weine (re.).

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.5031	0.55	−1.08	1.25	0.1115	0.0935	0.8685
E24	0.4845	0.73	−0.94	1.22	0.1226	0.0980	0.9165
E25	0.4530	0.55	−1.14	1.07	0.1339	0.0883	0.9462
V70(2016)	0.4111	0.49	−0.93	1.11	0.1918	0.1736	1.4663

Tabelle 7.30: Performance des entwickelten NN-Modells für Glyzerin, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in g/l.

7.12 Dichte

Während bei vorangegangenen Analysen stets Komponenten von Weinen betrachtet wurden, stellt die Dichte eine Eigenschaft eines Weines dar. Dieser lässt sich somit keine bestimmte chemische Struktur zuordnen, weshalb auch kein bestimmter Wellenzahlbereich aus chemisch-physikalischer Sicht erwartet werden kann.

Das Modell der neuronalen Netzwerke zur Schätzung der Dichte von Weinen verwendet insgesamt zwei Blöcke, jeweils mit einer Länge von 25 an Wellenzahlen, wie in Abbildung 7.37 (li). Einerseits wird der Abfall nach dem zweithöchsten Peak des Originalspektrums, welcher einer ganzen Bande in der Savitzky-Golay Ableitung der ersten Ordnung entspricht, und andererseits einen Anstieg bei $2\,730\,\mathrm{cm}^{-1}$ zum Peak im Bereich von $2\,900\,\mathrm{cm}^{-1}$ verwendet. Wiederum begnügt sich das statistische Modell mit nur einem versteckten Neuron. Die hierbei erwarteten Residuen bewegen sich in einem äußerst geringen Bereich, was wiederum durch den kleinen Wertebereich von $[0.9843, 1.0963]\,\mathrm{g/ml}$ begründet ist.

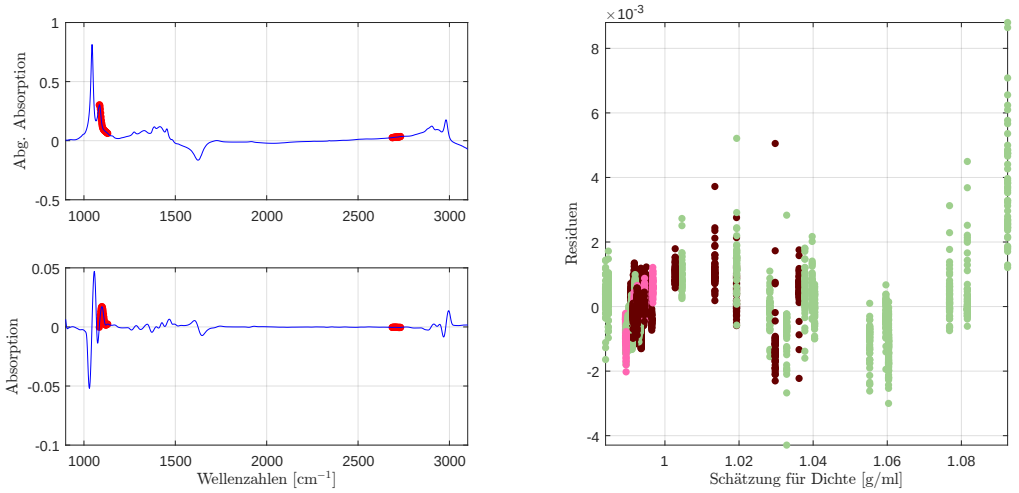

Abbildung 7.37: Die selektierten Wellenzahlen im NN-Modell für die Dichte mit der ersten Savitzky-Golay Ableitung (li.) und der aus der doppelten Kreuzvalidierung resultierende Residuenplot (re.).

Betrachtet man die aus der doppelten Kreuzvalidierung resultierenden Residuen in Abbildung 7.37 (re.), so kann ein teilweise leichter Anstieg der Schwankungsbreite für dichtere Residuen mit einer maximalen Auslenkung des Weißweines mit einer Dichte in Höhe von $1.0963\,\mathrm{g/ml}$ beobachtet werden. Dies begründet sich dadurch, dass bei der Erzeugung dieser Grafik durch die Aufteilung in unterschiedliche Datenpartitionen viele Informationen für dichtere Weine verloren gehen. Eben dieser Weißwein ist es auch, welcher die höchste Unterschätzung im klassischen Residuenplot in Abbildung 7.38 aufweist, auch wenn ein geringfügig größerer Ausreißer bei einem Roséwein mit einer vergleichsweise starken Überschätzung beobachtet wird. Im Residuenplot der doppelten Kreuzvalidierung ist dieser hingegen nicht auffällig, was wiederum auf das Vorhandensein von ähnlichen Daten, und

somit ähnlicher Information bei der Anwendung der doppelten Kreuzvalidierung, zurückgeführt werden kann. Trend oder andere Strukturen sind im klassischen Residuenplot nicht zu erkennen, wobei eine einigermaßen homogene Streuung zu beobachten ist. Gravierende Unterschiede bezüglich der Weinfarbe können hierbei nicht festgestellt werden.

Der Residuenplot in Abbildung 7.38 zeigt einen erkennbaren Bias in Höhe von $3 \cdot 10^{-5}\,\mathrm{g/ml}$ mit einer residualen Standardabweichung von $3.8 \cdot 10^{-4}\,\mathrm{g/ml}$ und einem betragsmäßig größten Residuum mit einem Fehler, welcher sich lediglich auf $8.6 \cdot 10^{-4}\,\mathrm{g/ml}$ beläuft. Es sind auch hier wiederum zwei Ausreißer zu beobachten.

Aufgrund der weitestgehenden Unauffälligkeit, sowie den für den geringen Wertebereich adäquaten Kennzahlen der Residuen, wird an dieser Stelle kein für Rotweine spezifisches Modell vorgestellt, zumal keine signifikanten Verbesserungen bezüglich der Darstellung der Residuen mit einem auf Rotweine reduzierten Modell erreicht werden können.

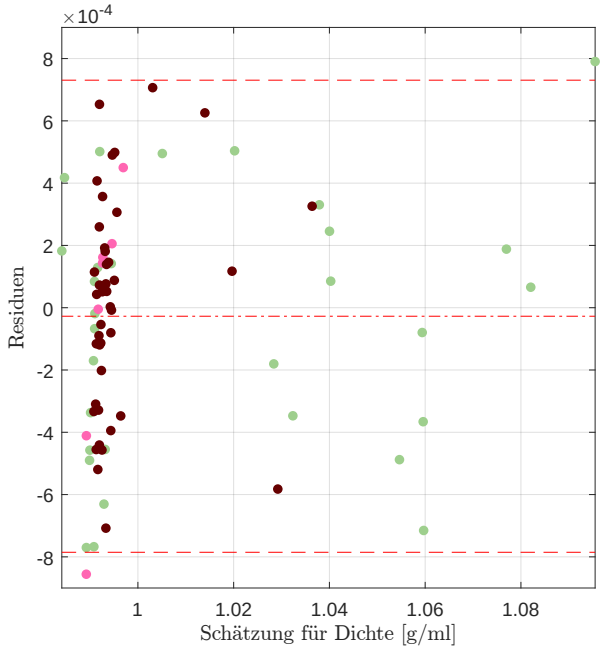

Abbildung 7.38: Residuenplot des NN-Modells für Dichte mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode.

Zudem zeigt sich für die Dichte, im Vergleich zum Extraktwert, trotz der hohen Korrelation von über 0.99 eine einfachere Modellierbarkeit.

Reproduzierbarkeit

Betrachtet man die Reproduzierbarkeit, so weisen die nicht repräsentativen Roséweine das beste Verhalten auf, wobei wiederum auf die geringe Anzahl der hier zur Verfügung stehenden Weine aufmerksam gemacht werden muss. Sowohl Tabelle 7.31 als auch Abbildung 7.39 zeigen trotz des annähernd vergleichbaren Verhaltens der Boxplots für Rot- und Weißweine, für erstgenannte eine geringfügig bessere Wiederholbarkeit, auch wenn letztere einen Ausreißer aufzeigen, welcher zugleich im Gesamtmodell als solcher identifiziert werden kann, beeinflusst dieser die Reproduzierbarkeit nicht wesentlich.

An dieser Stelle darf die unterschiedliche Interpretation der Kennzahlen der maximalen Abweichung in Tabelle 7.31 nicht mit der Darstellung der Boxplots vermischt werden. Während in Abbildung 7.39 (re.) ein Ausreißer bei den Weißweinen mit $5.8 \cdot 10^{-4}$ auffällt, so findet sich die maximale Abweichung von $1.15 \cdot 10^{-3}$ wieder. Hierbei handelt es sich nicht um einen Widerspruch, da die Residuen des Rotweines trotz der höheren maximalen Differenz weniger streuen.

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.2418	0.2395	1.1358	0.5357	0.5197	0.2431
Roséwein	0.1918	0.1821	0.7415	0.4396	0.4350	0.2573
Rotwein	0.2216	0.2014	1.1520	0.4947	0.4407	0.2273
Gesamt	0.2266	0.2124	1.1520	0.5053	0.4638	0.2345

Tabelle 7.31: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für Dichte im NN-Modell anhand der Einzelspektren pro Wein ID. Alle Werte in $10^{-3}\,\mathrm{g/ml}$.

Wendet man wiederum das neuronale Netzwerk mit den ausgewählten Prädiktoren, gekoppelt mit den restlichen Parametern, auf die vier vergleichbaren Datensätze mit der jeweils vom Datensatz abhängigen Kalibrierung, an, so ergibt sich die Übersicht in Tabelle 7.32. Wiederum zeigen die beiden Engines E22 und E25 der zweiten Generation eine vergleichbare Güte, auch wenn sich im Datensatz E25 ein leicht besseres Verhalten der Residuen beobachten lässt, wohingegen sich ein gegensätzliches Bild in der Reproduzierbarkeit zeigt. Zudem liefert der Datensatz E24 die beste Wiederholbarkeit, wohingegen die klassischen Residuen innerhalb der Enginegruppe der zweiten Generation um circa 15 % mehr streuen.

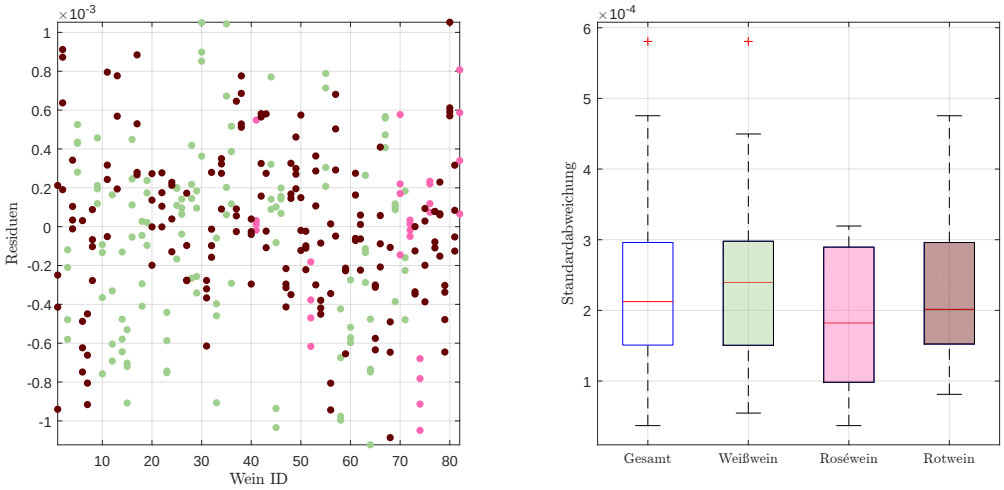

Abbildung 7.39: Residuen in g/ml der Einzelspektren pro Wein, ohne Kunstweine, zur Schätzung der Dichte im NN-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.3906	0.48	−0.95	0.80	0.2108	0.1999	1.1874
E24	0.4406	0.63	−0.88	1.20	0.1627	0.1598	1.0439
E25	0.3790	0.53	−0.86	0.79	0.2266	0.2124	1.1520
V70(2016)	0.4623	0.60	−1.04	1.16	0.1956	0.1858	0.9500

Tabelle 7.32: Performance des entwickelten NN-Modells für Dichte, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung. Alle Werte in 10^{-3} g/ml.

7.13 pH-Wert

Eine zweite Eigenschaft, welche in dieser Arbeit analysiert wird, ist der pH-Wert. Wie auch der Dichte kann dem einheitenlosen pH-Wert keine chemische Struktur zugrunde gelegt werden. Bei der hier durchgeführten Modellbildung mit Hilfe der neuronalen Netzwerke werden für das Modell bei der Betrachtung aller Weinproben insgesamt 40 Wellenzahlen selektiert und gliedern sich in vier unterschiedliche Teilbereiche gleicher Länge. Diese Prädiktorenselektion, dargestellt in Abbildung 7.40, beschreibt in der Savitzky-Golay Ableitung der ersten Ordnung einerseits zwei kleinere Peaks im Fingerprintbereich und geringfügige An- bzw. Abstiege in einem Bereich um die ausgeschlossene H_2O-Bande. Vergleicht man die Variablenselektion mit dem hierfür äquivalenten PLS-Modell für pH-Werte, so kann eine Teilübereinstimmung der Wellenzahlen beobachtet werden.

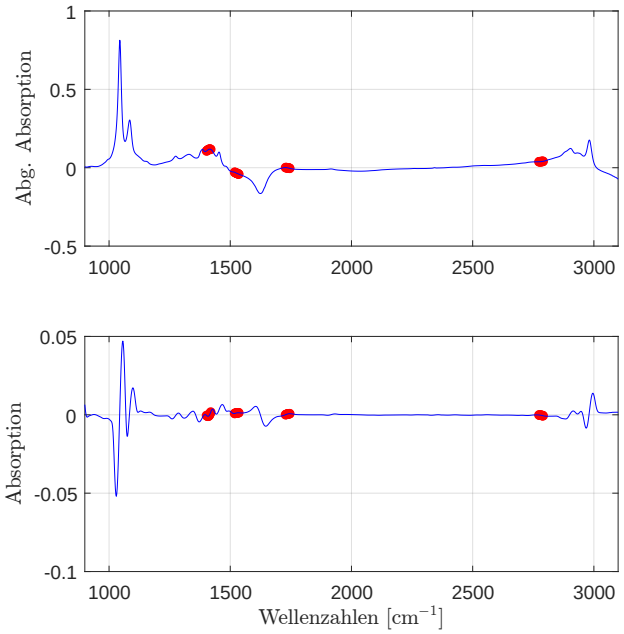

Abbildung 7.40: Die selektierten Wellenzahlen im NN-Modell für den pH-Wert mit der ersten Savitzky-Golay Ableitung.

Im neuronalen Netzwerk wird für dieses Modell ein verstecktes Neuron mit einer Aktivierungsfunktion des Typs Tangenssigmoid verwendet, wobei das Outputneuron wiederum eine lineare Transferfunktion besitzt. Somit genügt auch dieses Modell den Voraussetzungen für die Approximation beliebiger Funktionen im Abschnitt 6.2.1 über die mathematische Motivation von neuronalen Netzwerken.

Betrachtet man die Residuen aus der doppelten Kreuzvalidierung in Abbildung 7.41 (li.), so erkennt man keine bemerkenswerten Auffälligkeiten bezüglich der Variabilität, obwohl vereinzelte Weine etwas mehr streuen als andere. Hier kann jedoch kein struktureller Fehler aufgrund von heterogener Variabilität, bedingt durch Kennzahlen wie beispielsweise Schätzwert oder Farbe beobachtet werden.

Man kann lediglich einen leichten Trend in dieser Grafik feststellen, auch wenn dieser nicht signifikant erscheint. Darüber hinaus scheinen diese Residuen einigermaßen symmetrisch um 0 verteilt zu sein, wie das Histogramm dieser Residuen in Abbildung 7.41 (re.) nahelegt.

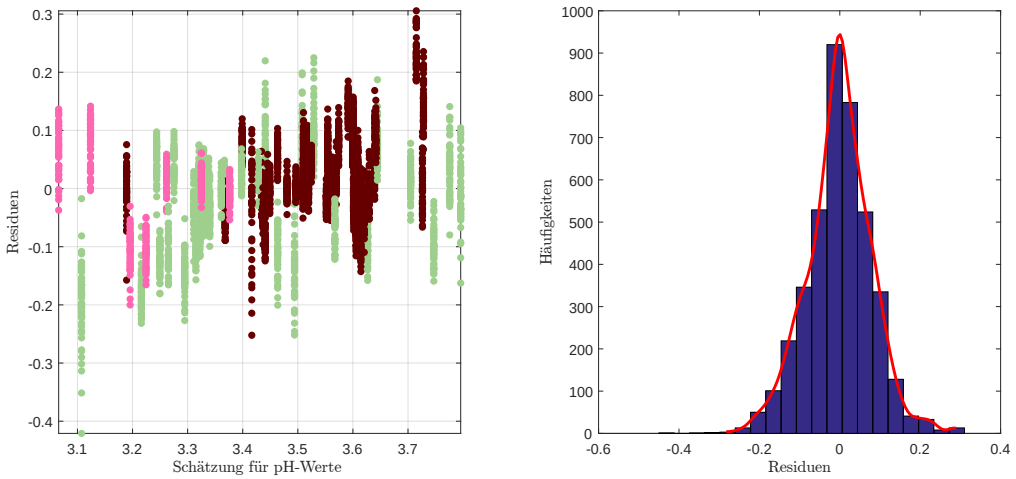

Abbildung 7.41: Die Weinresiduen der doppelten Kreuzvalidierung als Residuenplot (li.) und zusammengefasst zu einem Histogramm (re.) mit einer empirischen Dichteschätzung im NN-Modell für den pH-Wert.

Auch wenn die doppelte Kreuzvalidierung vergleichsweise akzeptable Werte und einigermaßen homogene Residuen hervorbringt, so zeigen die Grafiken in Abbildung 7.42 strukturelle Schwächen des Modells. Bei, bezüglich der Weinfarben, kumulierter Betrachtung erkennt man eine leicht steigende Tendenz der Residuen, abhängig von den pH-Schätzungen. Zudem kann den in der Mitte des Wertebereiches liegenden Weinproben eine etwas geringere Variabilität der Residuen unterstellt werden. Dies alles deutet auf strukturelle Mängel des Modells hin, auch wenn keine gesonderten Voraussetzungen im Sinne von Verteilungsannahmen an die Werte, abgesehen von deren Unabhängigkeit, durch die neuronalen Netzwerke verlangt wird. Während den Rotweinresiduen ein minimaler Trend unterstellt werden kann, zeigt sich in der Darstellung der übrigen Weine eine relativ deutliche Struktur. Abbildung 7.42 (re.) visualisiert ebendiese Residuen der Weiß- und Roséweine und diese können, abgesehen von wenigen Ausnahmen, in zwei Gruppen klassifiziert werden. Eine dieser Klassifizierungen folgt einem steigendem, während bei den restlichen Weinen, trotz der geringen Anzahl an Proben, ein linear fallender Trend beobachtet werden kann.

Trotz eingehender Analyse dieser zwei Weinklassifizierungen kann keine relevante Gemeinsamkeit festgestellt werden. Es findet sich jeweils ein breites Spektrum der unterschiedlichen Inhaltsstoffe wie beispielsweise Ethanol, Glukose oder Extrakt in den beiden Untergruppierungen. Zusätzlich sind keine offensichtlichen Messfehler der Spektren an den selektierten Wellenzahlen erkennbar.

Als mögliche Fehlerquelle kann eine ungeschickte Wahl der Prädiktoren außerhalb

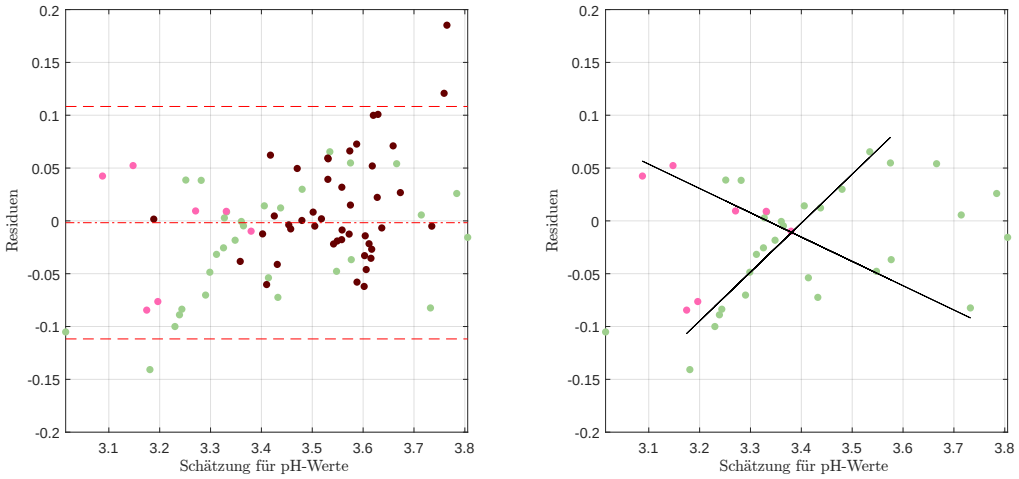

Abbildung 7.42: Residuenplot des NN-Modells für den pH-Wert mit Bias (mittlere rote Linie), der symmetrisch um die Biaslinie eingezeichneten (doppelten) empirischen Standardabweichung (in Rot) bzw. (in Grün) jene der Referenzmethode (li.) und den extrahierte Weiß-/Roséweinresiduen mit zwei erkennbaren Strukturen.

des Fingerprintbereiches vermutet werden, da an diesen Stellen kaum signifikante chemisch-physikalische Informationen vermutet wird. Da der pH-Wert die Aktivität der Wasserstoffionen beschreibt, findet sich womöglich relevante Information in der ausgeschlossenen H_2O Bande, auch wenn selbst bei Betrachtung dieses Wellenzahlbereichs kein adäquates Modell mit den zur Verfügung gestellten Datensätzen und der hier vorgestellten Methodik gefunden werden kann. Eine weitere Schwierigkeit stellt die Modellselektion dar, da prinzipiell sämtliche Weinbestandteile einen Einfluss auf den pH-Wert ausüben (können), und somit relativ viele Prädiktoren betrachtet werden sollten, was allerdings im Widerspruch zu der beschränkten Anzahl von 80 Weinproben steht.

Aufgrund der hier angeführten Mängel des Modells wird nicht weiter auf die Kennzahlen der Residuen eingegangen, auch wenn diese mit einer Standardabweichung in Höhe von 0.0550 bei gegebenen Werten mit einer Genauigkeit von zwei Nachkommastellen eine einigermaßen geringe Schwankungsbreite ausweisen. Diese finden sich dennoch in der Kennzahlenübersicht in Tabelle 7.35 wieder.

Reproduzierbarkeit

Betrachtet man dennoch die Reproduzierbarkeit in diesem Modell, so können abgesehen von dem Rotweinausreißer keine Auffälligkeiten beobachtet werden, wie Abbildung 7.43 zeigt. Die Rotweine zeigen hierbei eine, den Weißweinen gegenüber, leicht bessere Wiederholbarkeit, auch wenn die maximale Abweichung mit 0.1359, was wiederum beinahe der $2\,1/2$ Standardabweichung des Residuenplots in Abbildung 7.42 entspricht, einigermaßen hoch ausfällt. Da das Modell allerdings ohnehin nicht den mathematischen Ansprüchen genügt (Darstellung der Residuen), wird

auch an dieser Stelle auf eine detailliertere Analyse verzichtet und zur Übersicht der Reproduzierbarkeit auf Tabelle 7.33 verwiesen.

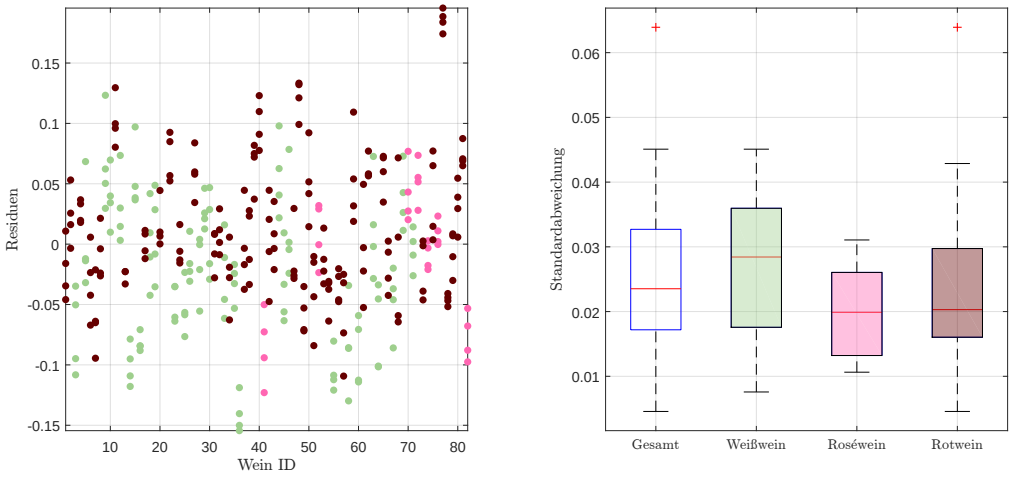

Abbildung 7.43: Residuen in g/l der Einzelspektren pro Wein, ohne Kunstweine, zur Schätzung des pH-Wertes im NN-Modell (li.), sowie die Standardabweichungen der Residuen pro Wein als Boxplot, gegliedert nach Farbe, ohne künstliche Weine (re.).

	Standardabweichung		Max. Abweichung			
	Mittelwert	Median	Maximum	Mittelwert	Median	Std.
Weißwein	0.0282	0.0284	0.1026	0.0635	0.0638	0.0235
Roséwein	0.0205	0.0199	0.0729	0.0460	0.0455	0.0179
Rotwein	0.0231	0.0203	0.1359	0.0515	0.0469	0.0259
Gesamt	0.0248	0.0235	0.1359	0.0555	0.0531	0.0250

Tabelle 7.33: Kennzahlen der Reproduzierbarkeit, aufgeschlüsselt nach Standardabweichung und maximaler Abweichung aller zur Verfügung stehenden Schätzungen für den pH-Wert im NN-Modell anhand der Einzelspektren pro Wein ID.

Vergleicht man die Datensätze, so können aus den vorangegangenen Unterabschnitten keine neuen Erkenntnisse gewonnen werden. Lediglich die Reproduzierbarkeit scheint im Datensatz E24 wiederum besser zu sein, während die Residuen lediglich eine vergleichsweise minimal höhere Standardabweichung aufweisen. Bei der Betrachtung der Residuenplots der restlichen Datensätze, insbesondere der analogen Darstellung der Weiß- und Roséweinresiduen in Abbildung 7.42 (re.), zeigen sich diese Residuen aller drei restlichen, einschließlich des V70(2016) Datensatzes, einigermaßen unauffällig, auch wenn bei der Gesamtbetrachtung aller Weinfarben wiederum ein steigender Trend mit Zunahme der Schätzwerte auftritt. Dies kann als Indiz von minimalen Verfälschungen in den Spektren des Datensatzes E25 aufgefasst werden.

	Residuen				Reproduzierbarkeit		
					Standardabweichung		
	Std.	IQR	Minimum	Maximum	Mittelwert	Median	Max. Abw.
E22	0.0563	0.09	−0.12	0.16	0.0269	0.0254	0.1372
E24	0.0655	0.09	−0.18	0.20	0.0157	0.0154	0.0753
E25	0.0550	0.07	−0.14	0.19	0.0248	0.0235	0.1359
V70(2016)	0.0599	0.08	−0.15	0.14	0.0164	0.0160	0.0811

Tabelle 7.34: Performance des entwickelten NN-Modells für den pH-Wert, angewendet auf die zur Verfügung stehenden Datensätze des Jahres 2016. Jeweils unter Verwendung des jeweiligen Datensatzes zur Kalibrierung und Auswertung..

7.14 Überblick über die entwickelten Modelle und Vergleich zur PLS-Methode

Anhand der ersten modellierten Komponente, dem Ethanol, zeigte sich bereits die hohe Sensibilität der neuronalen Netzwerke in Bezug auf Datenpunkte, welche in einer gewissen Form isoliert auftreten. Im Datensatz liegen zwei Weine mit einem Alkoholgehalt von weniger als 0.2 Vol.% vor, während der restliche Datensatz aus Weinen mit einem Alkoholgehalt von mindestens 7.9 Vol.% besteht. Da die neuronalen Netzwerke in den hier entwickelten Modellen äußerst sensibel auf derartige Stichproben reagieren, wird das Gesamtbild des Modells verzerrt, und daher ist es unumgänglich, den Wertebereich für eine plausible Vorhersage beziehungsweise Datenbeschreibung einzuschränken oder Teilmodelle zu entwickeln. Verglichen mit dem PLS-Modell für Ethanol stellt dies einen Nachteil dar. Zudem kann, auch wenn nur geringfügig, die Kennzahl Ethanol mit der PLS-Methode adäquater abgebildet werden. Dies gilt sowohl für das Verhalten der Residuen, als auch für die zweite beleuchtete Eigenschaft, der Reproduzierbarkeit.

Ähnlich verhält sich dies für die Kombination Extraktwerte und neuronale Netzwerke. Während zumindest mit der Neukalibrierung des Gesamtmodells für die Weinfarben im PLS-Modell für Extrakt plausible Ergebnisse und ein unauffälliges Verhalten der Residuen erzeugt werden kann, so kann der Trend innerhalb der Residuen, welcher aus den neuronalen Netzwerken entstammt, nicht eliminiert werden. Auch die Submodelle versagen in diesem Punkt. Bei den Extraktkomponenten, den beiden Zuckerarten Glukose und Fruktose muss ebenfalls für passendere Modelle auf die PLS-Modelle verwiesen werden, da stets leichte Strukturen erkennbar sind. Eine Ausnahme bildet das Untermodell des auf $[0, 10]$ g/l eingeschränkten neuronalen Netzwerkmodells für Glukose, auch wenn nur geringfügige Verbesserungen in den Kennzahlen erzielt werden können (sowohl im Verhalten der Residuen, als auch bezüglich der Reproduzierbarkeit).

Für den Repräsentanten der Gesamtsäure, den sogenannten titrierbaren Säuren, zeigt sich ein homogeneres Bild als im vergleichbaren Gesamtmodell mit den PLS-Modellen. Allerdings macht es auch an dieser Stelle Sinn, eine Neukalibrierung für die Weinfarben durchzuführen, was nicht durch Schwächen des Modells motiviert

wird, sondern durch die vorliegende Datensituation, wie auf Seite 82 beschrieben. Während die Variabilität bei beiden betrachteten mathematischen Modellen beinahe ident ist, wird diese bei den neuronalen Netzwerken durch einen betragsmäßig größeren Ausreißer in den Residuen etwas verstärkt und die restlichen Residuen zeigen ein vergleichsweise kompakteres Erscheinungsbild, während beim PLS-Modell das entwickelte Rotweinmodell zu bevorzugen ist. Zudem fällt auf, dass die neuronalen Netzwerke mit der Hälfte der Wellenzahlen auskommt und sich diese beiden Wellenzahlbereiche ebenfalls im PLS-Modell wiederfinden.

Zu den wichtigsten Bestandteilen der Gesamtsäure zählt die Weinsäure. Auch wenn diese, als Teil der titrierbaren Säure, sich im Verhältnis zur Referenzmethode nicht mit gleich guter Qualität modellieren lässt, kann dennoch eine geringere Streubreite als im PLS-Modell erzielt werden. Insbesondere wenn man berücksichtigt, welchen Einfluss isolierte Punkte außerhalb des Bereiches mit der größten Datendichte auf die Güte des Gesamtmodells (und insbesondere als Ausreißer auf die Standardabweichung), haben. Für die L-Äpfelsäure kann hingegen ein adequates Modell gefunden werden und die Residuen variieren mit einer ähnlichen Standardabweichung wie die Referenzmethode, trotz des Ausreißers, welcher circa die vierfache Standardabweichung misst und zeigt somit Vorteile gegenüber dem PLS-Modell, auch wenn die Reproduzierbarkeit in den neuronalen Netzwerken nicht auf gleiche Weise erreicht werden kann. Für die Milchsäure zeigt sich ein unterschiedliches Schätzverhalten der Weinfarben (negativer resp. positiver Bias), was eine getrennte Modellierung anhand der Farben erzwingt. So kann einerseits der Trend in den Rotweinresiduen eliminiert werden, bei zeitgleicher Erhöhung der Streubreite und es zeigt sich wiederum die Sensitivität der neuronalen Netzwerke. Zusätzlich verschlechtert sich die Reproduzierbarkeit um ein Vielfaches gegenüber den PLS-Modellen.

Wie bereits in den einzelnen Kapiteln zu den flüchtigen Säuren und der Zitronensäure respektive den Pendants der PLS-Modelle diskutiert, verteilen sich die Referenzwerte auf wenige diskrete Konzentrationen. Dennoch können für beide Klassen gute Modelle für alle Daten entwickelt werden. Insbesondere für die flüchtigen Säuren zeigt sich eine Standardabweichung der Residuen ähnlich jener der Referenzmethode. Zudem funktioniert das vorgestellte Modell auf allen Datensätzen gleichermaßen, mit Ausnahme der Reproduzierbarkeit für das Spektrometer E24. Im Vergleich zum PLS-Modell zeigen sich für das hierfür entwickelte Modell Schwächen in der Reproduzierbarkeit, auch wenn diese im Allgemeinen dennoch als gut erachtet werden kann. Für die Zitronensäure zeigt sich ein ähnliches Bild wie im PLS-Modell mit der Ausnahme, dass jener Roséwein mit einer Konzentration von 1.2 g/l einen starken Einfluss auf das Modell, insbesondere auf die Reproduzierbarkeit, aufweist.

Bei Glyzerin handelt es sich um den am schwersten modellierbaren Inhaltsstoff der Weine, da dieser mit den neuronalen Netzwerken und den verwendeten Heuristiken nur unzureichend beschrieben werden kann. Deswegen wird auch lediglich ein Rotweinmodell vorgestellt, welches diese Unterkategorie des Datensatzes bezüglich der Standardabweichung der Residuen ähnlich der Referenzmethode beschreibt, allerdings eine Struktur in den Residuen vermuten lässt. Daher wird an dieser Stelle wiederum auf das PLS-Modell verwiesen, welches plausiblere Resultate zeigt.

Die beiden Eigenschaften Dichte und pH-Wert lassen sich beide mit den neuronalen Netzwerke einigermaßen gut modellieren, wobei die Unterklasse der Weißweine nicht unabhängiger scheinen. Während im Vergleich zu den zugehörenden PLS-Modellen die Standardabweichung der Residuen teilweise halbiert werden kann (Dichte), so zeigt sich in beiden eine erhöhte maximale[9] Reproduzierbarkeit und eine damit einhergehende Ungenauigkeit.

[9]Sowohl bei der maximalen Reproduzierbarkeit als Maximalwert, als auch die grestlichen Kennzahlen zeigen erhöhte Werte.

		Residuen				Reproduzierbarkeit		
						Std.	Max. Abweichung	
	Anz.	Std.	IQR	Minimum	Maximum	Mittelwert	Maximum	Mittelwert
Ethanol	81	0.1318	0.1284	−0.4861	0.5788	0.0174	0.0885	0.0390
Weiß	30	0.1487	0.1628	−0.4861	0.2733	0.0180	0.0885	0.0407
Rosé	7	0.0677	0.0881	−0.0615	0.1331	0.0182	0.0751	0.0406
Rot	44	0.1292	0.1012	−0.4168	0.5788	0.0169	0.0639	0.0376
[8, 16] Vol.%	75	0.0862	0.1120	−0.2709	0.1950	0.0174	0.0885	0.0391
Ethanol, [8,16] Vol.%	75	0.0827	0.1156	−0.1744	0.3138	0.0387	0.1966	0.0853
Weiß	27	0.1073	0.1322	−0.1744	0.3138	0.0392	0.1966	0.0877
Rosé	7	0.0749	0.1135	−0.1203	0.0806	0.0389	0.1530	0.0819
Rot	41	0.0646	0.1000	−0.1185	0.1240	0.0383	0.1759	0.0844
Extrakt	80	2.6897	3.6189	−6.9059	7.8929	1.6461	11.0703	3.7441
Weiß	30	2.8902	3.6178	−6.9059	7.8929	1.8034	11.0703	4.1497
Rosé	7	2.9446	4.1933	−2.2509	6.0327	1.8667	6.9989	4.2770
Rot	43	2.5318	3.8338	−4.8638	4.6991	1.5004	8.6668	3.3743
[0, 63.5] g/l	63	2.4704	3.3946	−6.9059	6.0327	1.4819	6.9989	3.3645
Extrakt, [0,63.5] g/l	63	0.6095	0.6805	−1.2398	1.7477	0.6498	2.9805	1.4391
Weiß	17	0.6495	0.6254	−1.2398	1.0105	0.6094	2.3476	1.3230
Rosé	7	0.4306	0.4710	−0.9984	0.3575	0.6117	1.8680	1.3653
Rot	39	0.5988	0.7850	−0.9997	1.7477	0.6741	2.9805	1.5029
Glukose	81	0.9959	0.7182	−1.9243	5.1317	0.3826	3.2230	0.8454
Weiß	30	1.2729	0.8252	−1.9243	5.1317	0.5198	3.2230	1.1382
Rosé	7	0.4678	0.4908	−0.6316	0.7753	0.2181	1.1569	0.5013
Rot	44	0.8229	0.8530	−1.1286	3.8465	0.3153	1.6029	0.7006
[0, 10] g/l	62	0.4890	0.6617	−1.2464	0.8575	0.2891	1.2374	0.6468

		Residuen				Reproduzierbarkeit		
						Std.	Max. Abweichung	
	Anz.	Std.	IQR	Minimum	Maximum	Mittelwert	Maximum	Mittelwert
Glukose, [0,10] g/l	62	0.2199	0.2735	−0.6768	0.6799	0.1840	0.8881	0.4070
Weiß	16	0.2646	0.2780	−0.3221	0.6799	0.1769	0.7411	0.3947
Rosé	7	0.1479	0.2033	−0.2563	0.1655	0.1645	0.6408	0.3723
Rot	39	0.2119	0.2542	−0.6768	0.3337	0.1904	0.8881	0.4183
Fruktose	81	0.9795	0.6792	−2.1187	4.9682	0.3296	1.4210	0.7244
Weiß	30	1.4369	0.8512	−2.1187	4.9682	0.3565	1.4210	0.7871
Rosé	7	0.6416	1.1792	−0.4255	1.1258	0.2498	0.8957	0.5535
Rot	44	0.4635	0.5305	−1.1767	0.8886	0.3240	1.4084	0.7089
[0, 10] g/l	62	0.5448	0.5026	−1.4443	1.2350	0.3044	1.2445	0.6729
tirierbare Säuren	81	0.1728	0.2232	−0.4659	0.3830	0.0405	0.1997	0.0913
Weiß	30	0.2206	0.3430	−0.4659	0.3830	0.0408	0.1855	0.0919
Rosé	7	0.1561	0.1918	−0.2874	0.1865	0.0451	0.1997	0.1032
Rot	44	0.1264	0.1633	−0.2620	0.3338	0.0396	0.1389	0.0890
titrierbare Säuren, Kal: Weiß, Rosé	37	0.1664	0.2236	−0.2864	0.4065	0.0408	0.2113	0.0903
Weiß	30	0.1754	0.2652	−0.2864	0.4065	0.0393	0.2113	0.0869
Rosé	7	0.1282	0.1216	−0.2695	0.1275	0.0473	0.1543	0.1047
titrierbare Säuren, Kal: Rot	44	0.1327	0.1776	−0.3248	0.3058	0.0355	0.1543	0.0783
Weinsäure	81	0.2387	0.2982	−0.6515	0.9169	0.0423	0.1904	0.0941
Weiß	30	0.2724	0.2257	−0.6515	0.9169	0.0424	0.1901	0.0940
Rosé	7	0.1923	0.3146	−0.2055	0.2924	0.0531	0.1904	0.1229
Rot	44	0.2224	0.3287	−0.4634	0.7171	0.0404	0.1700	0.0896
[1, 2.9] g/l	79	0.2056	0.2938	−0.4634	0.7171	0.0423	0.1904	0.0943
Weinsäure, rot	44	0.1659	0.2225	−0.5589	0.3422	0.0519	0.2159	0.1165
[1, 2.9] g/l	44	0.1659	0.2225	−0.5589	0.3422	0.0519	0.2159	0.1165

			Residuen				Reproduzierbarkeit		
							Std.	Max. Abweichung	
		Anz.	Std.	IQR	Minimum	Maximum	Mittelwert	Maximum	Mittelwert
L-Äpfelsäure		81	0.1481	0.1424	−0.2179	0.6095	0.0428	0.3216	0.0953
	Weiß	30	0.1617	0.1867	−0.2179	0.4536	0.0712	0.3216	0.1576
	Rosé	7	0.0919	0.1722	−0.0957	0.1285	0.0707	0.2602	0.1632
	Rot	44	0.1400	0.0935	−0.2142	0.6095	0.0190	0.1409	0.0421
Milchsäure		81	0.1380	0.1697	−0.3055	0.4963	0.0673	0.4321	0.1496
	Weiß	30	0.0963	0.1137	−0.2424	0.1775	0.0314	0.2165	0.0693
	Rosé	7	0.2467	0.1762	−0.3055	0.4963	0.0349	0.1532	0.0793
	Rot	44	0.1277	0.1723	−0.1759	0.3984	0.0969	0.4321	0.2155
Milchsäure, Kal: Weiß, Rosé		37	0.2354	0.1287	−0.2328	1.1016	0.0218	0.1377	0.0485
	Weiß	30	0.2343	0.1273	−0.2133	1.1016	0.0206	0.1377	0.0460
	Rosé	7	0.2583	0.1766	−0.2328	0.5700	0.0266	0.1267	0.0594
Milchsäure, Kal: Rot		44	0.1593	0.1759	−0.4913	0.4230	0.0751	0.4420	0.1682
Milchsäure, rot		44	0.1877	0.2627	−0.4267	0.4298	0.0571	0.2651	0.1277
flüchtige Säuren		80	0.0841	0.1046	−0.1729	0.2035	0.0246	0.1327	0.0552
	Weiß	30	0.1054	0.1490	−0.1729	0.2035	0.0262	0.1327	0.0582
	Rosé	7	0.0702	0.1222	−0.1208	0.0556	0.0163	0.0623	0.0360
	Rot	43	0.0694	0.0827	−0.1491	0.1826	0.0247	0.1062	0.0561
Zitronensäure		81	0.1116	0.0892	−0.2100	0.7244	0.0332	0.4533	0.0737
	Weiß	30	0.0823	0.0881	−0.2100	0.1317	0.0558	0.4533	0.1240
	Rosé	7	0.2585	0.1280	0.0098	0.7244	0.0481	0.1994	0.1117
	Rot	44	0.0696	0.0297	−0.1016	0.1839	0.0153	0.2059	0.0333
Glyzerin, rot		44	0.4530	0.5496	−1.1420	1.0739	0.1339	0.9462	0.3023

			Residuen				Reproduzierbarkeit		
							Std.	Max. Abweichung	
		Anz.	Std.	IQR	Minimum	Maximum	Mittelwert	Maximum	Mittelwert
Dichte		80	0.3790	0.5320	−0.8558	0.7905	0.2266	1.1520	0.5053
	Weiß	30	0.4198	0.6431	−0.7702	0.7905	0.2418	1.1358	0.5357
	Rosé	7	0.4433	0.5041	−0.8558	0.4499	0.1918	0.7415	0.4396
	Rot	43	0.3430	0.4719	−0.7079	0.7066	0.2216	1.1520	0.4947
pH-Werte		81	0.0550	0.0692	−0.1408	0.1852	0.0248	0.1359	0.0555
	Weiß	30	0.0538	0.0845	−0.1408	0.0654	0.0282	0.1026	0.0635
	Rosé	7	0.0537	0.0939	−0.0845	0.0523	0.0205	0.0729	0.0460
	Rot	44	0.0523	0.0711	−0.0621	0.1852	0.0231	0.1359	0.0515

Tabelle 7.35: Übersicht über die wichtigsten Kennzahlen der behandelten NN-Modelle.

A Anhang

A.1 Grundlagen für Nichtlineare Optimierung

In diesem Abschnitt werden die wichtigsten Definitionen, welche in dieser Arbeit benutzt werden und einer Erklärung bedürfen, genau definiert beziehungsweise erläutert.

Ableitung und Gradient

Folgende Definitionen stammen aus [4]. Weiterführende und detailliertere Informationen können diesem Werk entnommen werden.

Definition A.1 (Ableitung, Gradient)**.** Sei $f\colon \mathbb{R}^n \to \mathbb{R}$, sowie $x \in \mathbb{R}^n$ und $e^i \in \{0,1\}^n$, mit $e^i_j = 1_{\{i=j\}}$ jener Vektor mit dem Wert 1 an der $i$te Stelle und 0 sonst. Falls der Grenzwert

$$\lim_{\alpha \to 0} \frac{f(x + \alpha \cdot e_i) - f(x)}{\alpha} \tag{A.1}$$

existiert, so bezeichnet Gleichung (A.1) die partielle Ableitung der Funktion f nach der $i$ten Komponente und wird mit $\frac{\partial f(x)}{\partial x_i}$ bezeichnet.

Falls die partielle Ableitung $\frac{\partial f(x)}{\partial x_i} \forall i \in \{1, \ldots, n\}$ existiert, so wird der Operator

$$\nabla := \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n} \right)'$$

als Gradient bezeichnet. Folglich ergibt sich der Gradient der Funktion f als

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)'.$$

Um zusammengesetzte Funktionen ableiten zu können, kann folgendes Lemma nützlich sein. Vergleiche hierzu [1], Seite 107, beziehungsweise die Seiten 114 und 353 für die Vektor- bzw. Matrixversion der Kettenregel.

Lemma A.2 (Kettenregel)**.** *Sei $f\colon I \to \mathbb{R}$ mit I offen und $g\colon f(I) \to \mathbb{R}$. Dann kann die Zusammensetzung dieser beiden Funktionen $(f \circ g) = g(f(x))$ unter den Bedingung*

- *c ein innerer Punkt von I,*
- *$f(c)$ ein innerer Punkt von $f(I)$ und*

- $\left.\frac{\partial f(x)}{\partial x}\right|_{x=c}$ *und* $\left.\frac{\partial g(x)}{\partial x}\right|_{x=f(c)}$ *existieren*

gefolgert werden, dass

$$\left.\frac{\partial \left(f \circ g\right)(x)}{\partial x}\right|_{x=c} = \left.\frac{\partial g(x)}{\partial x}\right|_{x=f(c)} \cdot \left.\frac{\partial f(x)}{\partial x}\right|_{x=c}$$

gilt.

Taylorentwicklung und o-Notation

Für diesen Abschnitt und weiterführende Informationen vergleiche beispielsweise [28]. Sei eine Funktion $f\colon I \subset \mathbb{R}$ eine $n+1$ fach differenzierbare Funktion. Das Taylorpolynom $P_n(x)$ vom Grad n an der Stelle $x_0 \in I$ ist definiert als

$$P_n(x) := \sum_{i=0}^{n} \left.\frac{\partial^n f(x)}{\partial x^n}\right|_{x=x_0} \cdot \frac{(x-x_0)^i}{i!}.$$

Mit Zuhilfenahme von $P_n(x)$ kann die Funktion $f(x)$ dargestellt werden als

$$\begin{aligned} f(x) &= P_n(x) + R_n(x) \\ \text{mit } R_n(x) &= \left.\frac{\partial^{n+1} f(x)}{\partial x^{n+1}}\right|_{x=\xi} \cdot \frac{(x-x_0)^i}{i!} \end{aligned}$$

an einer Zwischenwertstelle ξ zwischen x_0 und x. Hierbei gibt das Restglied $R_n(x)$ einen Wert für die Güte der polynomialen Approximation von f an. Offensichtlich gilt $R_n(x) = o(x^n)$ mit der folgender Definition von o.

Definition A.3 (Klein o-Notation)**.** Gegeben sind zwei Funktionen $f(x)$ und $g(x)$, mit $g(x) \neq 0$ für eine Umgebung von x_0, wobei $g(x_0) = 0$ erlaubt sei. Dann ist $f(x) = o(g(x))$ genau dann, wenn

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = 0$$

gilt.

Schrittweite in Gradientenverfahren

Für das Gradientenverfahren beziehungsweise deren Modifikationen, wie in den Gleichungen (6.8) beschrieben, muss die Schrittweite α_i bestimmt werden. Seien die Definitionen sowie Notationen wiederum wie in (6.8). Einige Auswahlmöglichkeiten für α_i sind in [4], Seite 29ff, gegeben. Hierzu zählen

- **Minimierungsregel**. Wähle α_i derart, dass die Schrittlänge die zu minimierende Funktion bezüglich der gewählten Richtung des Abstiegs d_i minimiert.

$$\alpha_i = \arg\min_{\alpha \geq 0} f(x_i + \alpha d_i) \tag{A.2}$$

- **Beschränkte Minimierungsregel**. In (A.2) wird α auf den Wertebereich $[0, s]$ beschränkt, wobei $s \in \mathbb{R}_{>0}$ ein fester Skalar ist.
- **Konstante Schrittweite**. Hierbei wird $\alpha_i := \alpha \in \mathbb{R}_{>0}$ gewählt.
- Weitere Wahlmöglichkeiten können [4], Seite 29ff, entnommen werden, wie beispielsweise die **Goldstein** oder der **Armijo Regel**.

A.2 Maßtheoretische Definitionen und Räume von Funktionen

Definition A.4 (σ-Algebra und Messraum)**.** Sei $\Omega \neq \emptyset$. Dann ist $\mathcal{A}$ als Teilmenge der Potenzmenge von Ω eine σ-Algebra, wenn folgende Punkte erfüllt sind:

- $\Omega \in \mathcal{A}$
- $A \in \mathcal{A}$ impliziert, dass das Komplement bezüglich der Grundmenge Ω ebenfalls in $\mathcal{A}$ enthalten ist: $\Omega \backslash A \in \mathcal{A}$.
- Für abzählbar viele $A_i, i \geq 1 \in \mathcal{A}$ gilt, dass deren abzählbare Vereinigung wiederum in $\mathcal{A}$ enthalten ist: $\bigcup_{i \geq 1} A_i \in \mathcal{A}$.

$A \in \mathcal{A}$ wird als messbare Mengen bezeichnet und das Paar $(\Omega, \mathcal{A})$ wird als Messraum bezeichnet.

Definition A.5 (Maß)**.** Sei $(\Omega, \mathcal{A})$ ein Messraum. Eine σ-additive Funktion $\mu \colon \mathcal{A} \to [0, \infty]$ mit der Eigenschaft $\mu(\emptyset) = 0$ wird als Maß bezeichnet. Als σ-Additivität wird folgende Eigenschaft verstanden:

Für $A_i, i \geq 1$, paarweise disjunkt gilt:

$$\mu\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} \mu(A_i).$$

Das Tripel $(\Omega, \mathcal{A}, \mu)$ wird als Wahrscheinlichkeitsraum bezeichnet.

Definition A.6 (endliches Maß)**.** Sei $(\Omega, \mathcal{A}, \mu)$ wie in Definition A.5. Wenn zusätzlich $\mu(\Omega) < \infty$ gilt, so ist μ ein endliches Maß.

Definition A.7 ($L^p(\mu)$)**.** Sei $(\Omega, \mathcal{A}, \mu)$ wie in Definition A.5. Der L^p-Raum mit $p \in \mathbb{N}_{>0}$ wird definiert als

$$\mathcal{L}^p := \{ f \colon \Omega \to \mathbb{R} \,/\, f \text{messbar} \wedge \|f\|_{L^p} < \infty \}$$

mit der Halbnorm für $f \in \mathcal{L}^p$

$$\|f\|_{\mathcal{L}^p} := \begin{cases} \left(\int\limits_{\Omega} |f(\omega)|^p \, \mathrm{d}\mu(\omega) \right)^{\frac{1}{p}} & p < \infty \\ \operatorname*{ess\,sup}_{\omega \in \Omega} |f(\omega)| & p = \infty \end{cases},$$

wobei das essentielle Supremum definiert ist als

$$\operatorname*{ess\,sup}_{\omega \in \Omega} |f(\omega)| := \inf_{\substack{N \in \mathcal{A} \\ \mu(N)=0}} \sup_{\omega \in \Omega \setminus N} |f(\omega)| .$$

Der normierte Vektorraum L^p ergibt sich durch die Faktorisierung $L^p := \mathcal{L}^p / \{f \in \mathcal{L}^p / f \equiv 0, \ \mu \text{ fast überall}\}$, wobei μ-fast überall bedeutet, dass $f \neq 0$ nur auf Mengen mit μ-Maß 0 ist.

Alternativ können diese Räume auch mit $\mathbb{C}$ anstelle von $\mathbb{R}$ definiert werden.

Definition A.8 ($C(X, Y)$)**.** Der Raum der stetigen Funktionen mit Definitionsbereich X und Y wird mit $C(X, Y) := \{ f \colon X \to Y / f \text{ stetig} \}$.

Definition A.9 (Dichte Mengen)**.** Sei (Ω, d) ein normierter Raum und $\Omega_1 \subset \Omega$. Dann liegt Ω_1 dicht in Ω, wenn $\forall \omega_1 \in \Omega_1 \forall \epsilon > 0 \exists \omega \in \Omega \colon d(\omega, \omega_1) < \epsilon$.

A.3 Weitere Definitionen

Definition A.10 (B-Spline)**.** Sei $x \in \mathbb{R}$ und $n \in \mathbb{N}$. So wird nach [9] mit

$$\mathrm{B}_n(x) := \sum_{i=0}^{n+1} \frac{(-1)^i}{n!} \binom{n+1}{i} \left[\max \left\{ 0; x + \frac{n+1}{2} - i \right\} \right]^n$$

der eindimensionale **B-Spline** definiert. Für den höherdimensionalen Fall können die B-Splines durch das Produkt der Funktionen $\mathrm{B}_n(\cdot)$, angewendet auf die einzelnen Komponenten eines Vektors, erweitert werden.

Literatur

[1] Tom M. Apostol. *Mathematical Analysis*. Reading, Massachusetts: Addison Wesley Publishing Company, 1981.

[2] S. Behnke. *Hierarchical Neural Networks for Image Interpretation*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2003. ISBN: 9783540451693.

[3] K. P. Bennett und M. J. Embrechts. »An Optimization Perspective on Kernel Partial Least Squares Regression«. In: *Advances in Learning Theory: Methods, Models and Applications*. Bd. 190. IOS Press Amsterdam, 2003, S. 227–250.

[4] Dimitri P. Bertsekas. *Nonlinear Programming*. Belmont, Massachusetts: Athena Scientific, 1995.

[5] Council of European Union. *Verordnung (EU) Nr. 1169/2011*. `http://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32011R1169`. 2011.

[6] Peter Filzmoser, Bettina Liebmann und Kurt Varmuza. »Repeated double cross validation«. In: *Journal of Chemometrics* 23.4 (2009), S. 160–171. ISSN: 1099-128X. DOI: 10.1002/cem.1225.

[7] Jörn Güldenhaupt. »ATR-FTIR-spektroskopische Untersuchungen von membrangebundenem Ras«. Diss. Ruhr-Universität Bochum, 2010.

[8] Jörn Güldenhaupt. »ATR-FTIR-spektroskopische Untersuchungen von membrangebundenem Ras«. Diss. Ruhr-Universität Bochum, 2010.

[9] Bart Hamers. »Kernel Models for Large Scale Applications«. Diss. Katholieke Universiteit Leuven Faculteit Toegepaste Wetenschappen Departement Elektrotechniek, Juni 2004. ISBN: 90-5682-509-7. URL: `ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/hamers/PhD_bhamers.pdf`.

[10] Trevor Hastie, Robert Tibshirani und Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[11] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, New Jersey: Prentice Hall PTR, 1999.

[12] Jens Hoffmann. »Methanol-Oxidation an getragenen Pd-Modellkatalysatoren«. Diss. Freie Universität Berlin, 2003.

[13] Kurt Hornik. »Approximation Capabilities of Multilayer Feedforward Networks«. In: *Neural Netw.* 4 (2 1991), S. 251–257. DOI: 10.1016/0893-6080(91)90009-T.

[14] Richard G. Palmer John Hertz Anders Krogh. *Introduction To The Theory Of Neural Computation*. Redwood City, California: Addison Wesley Publishing Company, 1991.

[15] Yehuda Lindell Jonathan Katz. *Introduction to Modern Cryptography: Principles and Protocols*. Boca Raton, Florida: Chapman & Hall/CRC, 2007.

[16] W. Kessler. *Multivariate Datenanalyse: für die Pharma, Bio- und Prozessanalytik*. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA., 2006. ISBN: 9783527312627.

[17] Jeannette Lawrence. *Neuronale Netze*. München: Systhema Verlag GmbH, 1992.

[18] Hsuan tien Lin und Chih-Jen Lin. *A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods*. Techn. Ber. Taipei 106, Taiwan: National Taiwan University, 2003.

[19] Howard Demuth Mark Beale Martin Hagan. *Neural Network Toolbox™ Reference*. Hrsg. von Natick MA MATLAB R2017a The MathWorks. 1992.

[20] Mark Hudson Beale Orlando De Jesús Martin T. Hagan Howard B. Demuth. *Neural network design*. Oklahoma: Martin Hagan, 1996.

[21] Mohammad B. Menhaj Martin T. Hagan. »Training Feedforward Networks with the Marquardt Algorithm«. In: *IEEE Transactions on Neural Networks* 5 (1994), S. 989 –993. DOI: 10.1109/72.329697.

[22] Kairi Mashio. Hrsg. von The University of Tokyo. Graduate School of Frontier Sciences. Department of Complexity Science und Engineering. [Online; besucht 26.03.2017]. 2009-2010. URL: http://mns.k.u-tokyo.ac.jp/~mashio/product.html.

[23] G. Neuhaus und J.P. Kreiß. *Einführung in die Zeitreihenanalyse*. Skripten zur mathematischen Statistik. Ges. zur Förderung d. Math. Statistik, 1985.

[24] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press, 1996.

[25] J. Robinson und J. Harding. *The Oxford Companion to Wine*. Oxford Companions. OUP Oxford, 2015. ISBN: 9780191016073.

[26] Raúl Rojas. *Theorie der neuronalen Netze: eine systematische Einführung*. Berlin: Springer-Verlag, 1993.

[27] Roman Rosipal und Leonard J. Trejo. »Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space«. In: *J. Mach. Learn. Res.* 2 (März 2002), S. 97–123. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944790.944806.

[28] Vladimir I. Rotar. *Actuarial Models*. Boca Raton, Florida: Chapman und Hall/CRC, 2015.

[29] Bernhard Schölkopf, Ralf Herbrich und Alex J. Smola. »A Generalized Representer Theorem«. In: *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*. COLT '01/EuroCOLT '01. London, UK, UK: Springer-Verlag, 2001, S. 416–426. ISBN: 3-540-42343-5. URL: http://dl.acm.org/citation.cfm?id=648300.755324.

[30] Walter Pitts Warren S. McCulloch. »A Logical Calculus Of The Ideas Immanent In Nervous Activity«. In: *Bulletin Of Mathematical Biophysics* 5 (1943), S. 115–133.

[31] Hanne Winning. *Standardization of FT-IR instruments*. Techn. Ber. A White Paper from FOSS P/N 1026672, März 2014. URL: http://www.foss.dk/~/media/114CD4E50AD14D649DB8672606C50ED5.ashx.

[32] Patrick Henry Winston. *Artificial Intelligence*. Reading, Massachusetts: Addison Wesley Publishing Company, 1993.

[33] Qing-Song Xu und Yi-Zeng Liang. »Monte Carlo cross validation«. In: *Chemometrics and Intelligent Laboratory Systems* 56.1 (2001), S. 1 –11. ISSN: 0169-7439. DOI: http://dx.doi.org/10.1016/S0169-7439(00)00122-2. URL: http://www.sciencedirect.com/science/article/pii/S0169743900001222.

[34] L.D.S. Yadav. *Organic Spectroscopy*. Springer Netherlands, 2004. ISBN: 9781402025747.