Optimierung der Variablen-Selektion für die PLS-Regression

Michael Judas
Institut für Sicherheit und Qualität bei Fleisch, MRI Standort Kulmbach
E.-C.-Baumann-Str. 20
95326 Kulmbach
michael.judas@mri.bund.de

Stefaan De Smet
Laboratory for Animal Nutrition and Animal Product Quality, Ghent University
Proefhoevestraat 10
9090 Melle, Belgium
stefaan.desmet@ugent.be

Zusammenfassung

In der Lebensmittelforschung ist eine wichtige Problemstellung die Vorhersage eines relevanten Merkmals durch leicht messbare Prädiktoren. Als Beispiel dient die Schätzung der Qualität (Muskelfleisch-Anteil) eines Schweine-Schlachtkörpers durch automatisierte Videobildauswertung. Dabei werden sehr viele Längen- und Flächenmaße sowie Farbinformationen erhoben, aus denen die optimalen Prädiktoren selektiert werden müssen.
Die übliche multivariate lineare Regression (PROC REG) ist für diese Problemstellung begrenzt geeignet, da es reichlich Interkorrelationen der Prädiktoren gibt und die Zahl der Variablen die Fallzahl weit übersteigt. Eine Lösung bietet die PROC PLS, bei der in einem iterativen Verfahren Faktoren aus den Prädiktor-Variablen extrahiert werden, die in der eigentlichen Regression verwendet werden (weshalb es weder Probleme mit Interkorrelationen noch mit zuvielen Variablen gibt); gleichzeitig werden die erklärten Varianzen sowohl der abhängigen Variable als auch der Prädiktoren maximiert.
Auch bei der PLS-Regression stellt sich das Problem, aus der Masse der verfügbaren Prädiktoren einen möglichst kleinen Satz auszuwählen, der eine möglichst gute Schätzung erlaubt. Dazu bietet SAS ein Makro an, mit dem die *Variable Importance for the Projection* (VIP) berechnet werden kann. Nach einem ersten PLS-Durchlauf mit allen Variablen lassen sich dann für einen zweiten Durchlauf diejenigen Variablen ausschließen, die eine minimale VIP unterschreiten. Allerdings fällt nun nach dem zweiten Durchlauf die VIP einiger Variablen unter den zuvor gesetzten Grenzwert. Es erscheint daher sinnvoll, das Kriterium einer minimalen VIP nicht nur einmal anzuwenden, sondern in einer Schleife die VIPs so lange abzufragen, bis alle Variablen, die in dem Modell verbleiben, das gesetzte Kriterium erfüllen.
In einem Beispiel-Datensatz mit n=141 und 279 potentiellen Prädiktoren läßt sich mit dieser Methode die Schätzung optimieren, wobei 38 Prädiktoren selektiert werden. Der Schätzfehler für die Vorhersage (RMSEP, *root mean squared error of prediction*) wird dabei von 2.23 (10 PLS-Faktoren, alle 279 Variablen) auf 2.02 (4 PLS-Faktoren, 38 Variablen) gesenkt.

Schlüsselwörter: PLS-Regression, Variablen-Selektion, VIP, RMSEP.

1 Einleitung

Ein Kernproblem von Regressionsanalysen besteht in der Selektion adäquater Prädiktor-Variablen. Die etablierte Multiple Lineare Regression (MLR) bietet dazu verschiedene

Methoden, z.B. *stepwise regression*, und verschiedene Selektionskriterien, z.B. die Maximierung von R^2. Allerdings sind der Anwendung der MLR Grenzen gesetzt, wenn die Prädiktoren stark interkorreliert sind oder ihre Anzahl die der Beobachtungen weit übersteigt. Zudem ist eine vergleichende Interpretation der Regressions-Koeffizienten erschwert, wenn die Variablen auf unterschiedlichen Skalenniveaus gemessen werden.

Für Datensätze mit einer Vielzahl stark korrelierter Variablen bietet die *Partial-Least-Squares*-Regression (PLS-R) eine gute Alternative zur MLR. Das Verfahren ist in SAS mit der PROC PLS implementiert, die aber keine direkte Möglichkeit zur Selektion von Prädiktor-Variablen bietet. Hier soll aufgezeigt werden, wie der Output der PROC PLS verwendet werden kann, um eine Variablen-Selektion durchzuführen und damit ein optimiertes Regressionsmodell zu erstellen.

2 Datenbasis

Die Datengrundlage für die hier vorgestellte Analyse stammt aus der Fleischproduktion. In Schlachthöfen werden in der laufenden Schlachtlinie Geräte eingesetzt, die Parameter der Fleischqualität erfassen sollen, z.B. den Anteil des Muskelfleisches am Schlachtkörper (*lean meat percentage*, LMP). 2006 wurde in belgischen Schlachthöfen das VCS2000[1] installiert, das auf Basis von Videobildanalysen die Qualität von Schweine-Schlachtkörpern bestimmt [1]. Dabei wird eine Vielzahl von Längen, Flächen und Winkeln sowie daraus abgeleiteten Größen gemessen, dazu wird eine Reihe von Farbinformationen erfasst. Insgesamt stehen 279 Variablen zur Verfügung, um den LMP zu schätzen. Die Größenordnung der einzelnen Variablen schwankt zwischen 10^{-3} und 10^{5}. An einem Datensatz von 141 Schlachtkörpern wurde das vorgestellte PLS-Modell entwickelt. Für diesen Datensatz war der LMP mit einer Referenzmethode bestimmt worden.

3 PLS-Regression

Die PLS-Regression analysiert den Zusammenhang zwischen Prädiktoren (X-Matrix) und einer oder mehreren Zielvariablen (Y-Matrix) durch iterative Hauptkomponenten-Analysen der beiden Matrizen [2] [3]. Diese erfolgen aber nicht unabhängig sondern beeinflussen sich gegenseitig: die Faktorladungen der jeweils anderen Matrix gehen in die PCA der X- bzw. Y-Matrix ein. Dadurch wird gleichzeitig die erklärte Varianz von X- *und* Y-Matrix maximiert.

Die PLS-Regression bietet eine Reihe von Eigenschaften, die sie geradezu als ideale Regressionsmethode erscheinen lassen:

- Die Regression erfolgt mit den extrahierten Faktoren, hat also kein Problem mit Kollinearitäten der Variablen.

[1] e+v Technologies, Oranienburg; <http://www.eplusv.de/vcsd.htm>

- Die Anzahl der Prädiktor-Variablen ist nicht begrenzt, da — unabhängig von der Anzahl der Variablen im Modell — eine relativ kleine Zahl von Faktoren extrahiert wird, maximal 10–15, meist deutlich weniger.
- Die Variablen werden vor der Analyse zentriert und skaliert, daher ist ihre relative Bedeutung direkt ist aus den resultierenden Regressionskoeffizienten ablesbar (gewichtete Regressionskoeffizienten, b_w).
- Nach der Analyse werden ungewichtete (rohe) Koeffizienten rekalkuliert, mit denen das Regressionsmodell auf den originalen Datenraum angewendet werden kann.

3.1 PROC PLS

Die SAS-Hilfe bietet neben Informationen zu den Optionen der PROC PLS auch einige Programmbeispiele und Makros, die für eine Interpretation der Ergebnisse sehr nützlich sind (u.a. den Code zur Berechnung der VIP, vgl. unten). Zusätzlich ist ein SUGI-Beitrag zugänglich [4], der PLS-Anwendungen beschreibt und nützliche Makros bietet (u.a. einen IML-Code zur Berechnung der VIP).

PROC PLS bietet neben der PLS-Regression (default) als alternative Regressions-Methoden auch eine PCR (*principal components regression*), welche die Varianzerklärung nur für die X-Matrix maximiert, sowie die RRR (*reduced ranks regression*), welche die Varianzerklärung nur für die Y-Matrix maximiert.

3.2 Verwendete Parameter der PROC PLS

Die allgemeine Syntax für die PLS-Prozedur lautet:

```
PROC PLS < options > ;
     BY variables ;
     CLASS variables < / option > ;
     MODEL dependent-variables = effects < / options > ;
     OUTPUT OUT= SAS-data-set < options > ;
```

In der hier vorgestellten Anwendung werden weder BY- noch CLASS-Variablen verwendet. In den anderen Statements steht eine Reihe von Optionen zur Verfügung, um den Modus der Berechnung zu beeinflussen bzw. einen gewünschten Output zu erzeugen. Die hier verwendeten Optionen sind in Tabelle 1 aufgelistet.

Die Anzahl der zu extrahierenden Faktoren wird mit `nfac` vorgegeben (default=15). Mit `cvtest` wird ein Test angefordert, der die PRESS-Statistik (*predicted residual sum of squares*) für alle PLS-Modelle mit 1 bis `nfac` Faktoren berechnet, wobei ein Modell eine minimale PRESS aufweist. Dasjenige Modell wird auswählt, das die wenigsten Faktoren hat, ohne eine signifikant höhere PRESS aufzuweisen als das Minimum-PRESS-Modell. Der Modell-Vergleich beruht auf einer vollen Kreuzvalidierung (`cv =`

one), wobei auch andere CV-Optionen möglich sind. Die Ergebnisse der PLS-Schätzung (predicted) werden in eine OUTPUT-Datei geschrieben.

Tabelle 1: Verwendete Optionen der PROC-PLS-Statements

PROC PLS	nfac = …	max. Anzahl der PLS-Faktoren
PROC PLS	details	ODS-Tabellen für VIP-Berechnung
PROC PLS	cvtest	Test auf signifikante PLS-Faktoren
PROC PLS	cv = one	Methode des CV-Tests
MODEL	solution	Gewichtete Regressions-Koeffizienten
OUTPUT	out = …	Output-Datei
OUTPUT	predicted = …	PLS-Schätz-Ergebnisse

Die Selektion von Variablen beruht auf zwei Statistiken:

- Den gewichteten Regressionskoeffizienten, b_w, die durch solution in eine ODS-Tabelle CenScaleParms geschrieben werden.
- Der *Variable Importance for the Projection* (VIP), für deren Berechnung die ODS-Tabellen Xweights und PercentVariation benötigt werden (details).

3.3 Variablen-Selektion

3.3.1 Gewichtete Regressionskoeffizienten, b_w

Die gewichteten Regressionskoeffizienten variieren zwischen -1 und 1. Ein Grenzwert für einen relevanten Beitrag zum Modell ist am Besten iterativ zu finden. In dem Beispiel erwies sich $| b_w |>.01$ als optimal.

3.3.2 Variable Importance for the Projection (VIP)

In der SAS-Online-Hilfe zu PROC PLS ist ein Beispielprogramm[2] beschrieben, das die Berechnung der VIP in folgenden Schritten beschreibt:

- Summierung der normalisierten quadrierten Gewichte (pro Faktor für jede Variable, aus Xweights), multipliziert mit den normalisierten R^2s (pro Faktor, aus PercentVariation).
- Die VIP ergibt sich als Quadratwurzel aus dem Produkt dieses gewichteten Mittels mit der Anzahl der Prädiktor-Variablen.

Ein VIP-Wert von 0.8 wird als sinnvoller Grenzwert für einen relevanten Beitrag der Variablen zum PLS-Modell angesehen.

3.3.3 Selektion der PLS-Faktoren

Durch die Option cvtest kann ein Test auf signifikante Faktoren durchgeführt werden. Dieser Test schließt keine Variablen aus, erhöht aber die nicht erklärte Varianz.

[2] Example 56.1: Examining Model Details

Grundsätzlich lassen sich Variablen- und Faktoren-Selektion kombinieren. In der vorgestellten Anwendung hat es sich aber als sinnvoll erwiesen, nur die Variablen-Selektion wiederholt durchzuführen, während die Faktoren-Selektion nur einmal in einem abschließenden PLS-Schritt erfolgt (Der komplette Programmcode befindet sich auf der Begleit – CD zu diesem Tagungsband).

4 Das PLS-Modell

Nach einem PLS-Durchlauf werden zunächst die b_ws und VIPs jeder Variablen berechnet. Dies erlaubt eine Selektion derjenigen Variablen, die die jeweiligen Grenzwerte überschreiten. Mit einem zweiten PLS-Durchlauf könnte damit der Prozess der Modellierung abgeschlossen werden. Es zeigte sich jedoch, dass auch in dem zweiten Durchlauf einige Variablen unter die zuvor festgelegten Grenzwerte fallen. Daher lag es nahe, die PLS-Modellierung nicht als einen zweistufigen Vorgang mit einmaliger Variablen-Selektion durchzuführen, sondern solange zu wiederholen, bis alle verbleibenden Variablen die Grenzwerte für b_w und VIP überschreiten.

Zusätzlich musste das SAS-Beispiel-Programm zur Berechnung der VIP modifiziert werden, da es auf 2 Faktoren und 15 Variablen festgelegt ist. Für eine allgemeine Anwendbarkeit mussten die Anzahl der erlaubten Variablen und Faktoren flexibel gestaltet werden.

4.1 Programm-Aufbau

Das Programm besteht aus folgenden Modulen:

- Einem allgemeinen Block zur Definition von Makrovariablen mit Grenzwerten und Parametern, sodass leicht mit verschiedenen Grenzwerten experimentiert werden kann.
- Einem Makro (PLS) zur Durchführung der PLS mit oder ohne Test auf signifikante Faktoren (*&cvoptions*). Darin werden die b_ws und VIPs für eine beliebige Anzahl von Variablen (*&npred*) und für bis zu 20 Faktoren (*&nfac*) berechnet.
- Einem Makro (SCHLEIFE) zur iterativen Durchführung der Variablen-Selektion. Dieses Makro ruft das PLS-Makro solange auf, wie Variablen die Selektions-Kriterien nicht erfüllen. Das Makro selektiert anschließend an eine PLS die Variablen und plottet b_ws und VIPs pro Variable. Abschließend erfolgt eine PLS mit Test auf signifikante Faktoren (was die selektierten Variablen nicht mehr beeinflusst).
- In einem abschließenden Schritt (MACRO RMSEP) wird die Vorhersage-Qualität des endgültigen Modells berechnet und geplottet.

4.2 Modell-Qualität

Der vollständige Datensatz mit n=141 wurde zur Kalibrierung des PLS-Modells verwendet. Damit ergibt sich aus den Differenzen der Referenzwerte und der PLS-Schätzungen ein *root mean squared error of calibration* (RMSEC). Wichtiger für die Beur-

teilung eines Regressions-Modells ist aber der Fehler der Vorhersage (*root mean squared error of prediction*, RMSEP). Dieser wird durch die PRESS-Statistik, die mit PROC PLS abrufbar ist, aber zu optimistisch berechnet. Daher wurde eine vollständige (*leave-one-out*) Kreuzvalidierung des endgültigen PLS-Modells als Makro in das Programm eingebaut: jeweils ein Referenzwert wird bei einer PLS mit den selektierten Variablen und Faktoren auf fehlend gesetzt. Dadurch wird dieser Wert geschätzt, ohne in die Modellbildung einzugehen. Der resultierende RMSEP ist größer als der RMSEC, stellt aber eine akzeptable Annäherung an die Vorhersage-Genauigkeit dar, wenn keine unabhängigen Testdaten zu Verfügung stehen.

4.3 Ergebnis der Modellierung

Die Ergebnisse der iterativen Modellbildung sind in Tabelle 2 dargestellt. Die erste PLS basiert auf allen 279 Prädiktor-Variablen und hat einen minimalen RMSEC von 1.57 (bei der festgelegten maximalen Anzahl von 10 PLS-Faktoren). Nach 7 PLS-Durchläufen bleiben nur noch Variablen im Modell, die die Kriterien von $|b_w|>.01$ und VIP>0.8 erfüllen. Durch diese Variablen-Selektion steigt der RMSEC zwar auf 1.71, der RMSEP fällt jedoch von 2.22 auf 2.07 (jeweils mit 10 PLS-Faktoren). Der abschließende Durchlauf ändert nichts mehr an den selektierten 38 Variablen, reduziert jedoch die Zahl der Faktoren auf signifikante vier. Dadurch erhöht sich der RMSEC weiter auf 1.80, der RMSEP hat jedoch mit 2.02 ein Minimum.

Damit führt die iterative Variablen-Selektion zu einem optimalen Regressionsmodell: sowohl die Anzahl der Prädiktor-Variablen und PLS-Faktoren als auch der Vorhersage-Fehler sind minimiert.

Tabelle 2: Ergebnisse der Modellierungs-Schritte

PLS #	NPred	NFac	RMSEC	RMSEP
1	279	10	1.57	2.22
2	83	10	1.57	2.09
3	59	10	1.60	2.03
4	50	10	1.61	2.13
5	45	10	1.71	2.10
6	41	10	1.71	2.07
7	38	10	1.71	2.07
8	38	4	1.80	2.02

Literatur

[1] W. Branscheid, A. Dobrowolski, R. Höreth: Video-Image-Analyse. Methode zur automatischen Handelswertbestimmung von Schweinehälften. Fleischwirtschaft 12: 93–95, 1999.

[2] W. Kessler: Multivariate Datenanalyse für die Pharma-, Bio- und Prozessanalytik. Wiley-VCH Verlag, 2007.

[3] S. Wold, M. Sjöström, L. Eriksson: PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58: 109–130, 2001.

[4] R. D. Tobias: An Introduction to Partial Least Squares Regression, SUGI Proceedings, 1995. URL <http://support.sas.com/rnd/app/papers/plsex.pdf>