Universiteit
Leiden
The Netherlands

Prediction accuracy and stability of regression with optimal scaling transformations

Kooij, A.J. van der

Citation
Kooij, A. J. van der. (2007, June 27). *Prediction accuracy and stability of regression with optimal scaling transformations*. Leiden. Retrieved from https://hdl.handle.net/1887/12096

Version:	Corrected Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/12096

Note: To cite this publication please use the final published version (if applicable).

Summary in Dutch (Samenvatting)

In dit proefschrift wordt de CATREG-methode voor multipele regressie met optimale-schalingstransformaties beschreven alsmede drie belangrijke uitbreidingen van deze methode. De eerste uitbreiding betreft strategieën die suboptimale oplossingen als gevolg van lokale minima, die optreden met monotone transformaties, voorkomen of reduceren. Het lokale-minimumprobleem treedt ook op bij andere methoden die nauw verwant zijn aan CATREG en was tot dusverre in de literatuur niet bekend. Ten tweede is een procedure toegevoegd voor het beoordelen van de voorspelnauwkeurigheid van CATREG, waarbij de verwachte voorspelfout wordt geschat met de .632-bootstrapmethode, die is aangepast voor toepassing op getransformeerde variabelen. Tenslotte zijn drie regularisatiemethoden toegevoegd voor het verbeteren van de voorspelnauwkeurigheid door stabilisatie van de parameterschattingen en voor selectie van de predictoren.

In het eerste hoofdstuk wordt de CATREG-methode beschreven, waarbij de basiselementen van de methode grafisch worden geïllustreerd aan de hand van een kleine dataset met niet-lineaire regressie-effecten. De optimale-scalingsprocedure voor het transformeren van variabelen en de backfitting procedure voor het schatten van de regressieparameters worden hier uitgelegd zonder formules. Ook wordt het effect van het gekozen schalingsniveau op de vorm van de transformaties en op de kwaliteit van de voorspelling behandeld.

Met monotone transformaties eindigt CATREG soms in een lokaal minimum van de verliesfunctie, met als gevolg een suboptimale oplossing. Ook met andere methoden voor regressie met transformaties treden deze lokale minima op. Dit probleem is het onderwerp van hoofdstuk 2. De oorzaak van het eindigen in een lokaal minimum en een strategie om het globale minimum te bereiken worden beschreven. Deze strategie maakt echter gebruik van multipele systematische starts met volledige enumeratie, hetgeen niet

uitvoerbaar is met een groot aantal predictoren. Diverse strategieën die minder rekenkracht vereisen, maar die het vinden van het globale minimum niet garanderen, worden gepresenteerd alsmede resultaten van een evaluatie van de werking van deze strategieën in een simulatiestudie. Twee van deze strategieën geven goede resultaten voor het reduceren van zowel de incidentie als de ernst van lokale minima. De ene strategie gebruikt een verliespercentagecriterium en is flexibel doordat het percentage gevarieerd kan worden. De andere strategie gebruikt een hiërarchische benadering en heeft als voordeel dat een kleiner aantal starts nodig is, maar is minder flexibel. Combinatie van deze twee strategieën levert een flexibele strategie op, die de incidentie en de ernst van lokale minima aanzienlijk reduceert met een relatief klein aantal starts. Bovendien heeft de gecombineerde strategie het voordeel dat het aantal starts niet heel snel groter wordt met het toenemen van het aantal predictoren. In de simulatiestudie wordt ook het effect bestudeerd van diverse datacondities op de incidentie en ernst van lokale minima. Lokale minima blijken vaker voor te komen en ernstiger te zijn naarmate de voorspellende waarde van de variabelen lager is, naarmate de predictoren meer categorieën hebben en naarmate de predictoren onderling sterker gecorreleerd zijn.

In hoofdstuk 3 wordt de voorspelnauwkeurigheid van CATREG beoordeeld door het schatten van de verwachte voorspelfout met de .632-bootstrap. Er wordt een gedetailleerde beschrijving gegeven van de complicaties die optreden bij de implementatie van de .632-bootstrap voor optimaal geschaalde variabelen. Deze complicaties zijn gedeeltelijk te wijten aan het feit dat in de .632-bootstrapliteratuur op verschillende plaatsen verschillende definities worden gebruikt voor de leave-one-out-bootstrapschatting van de voorspelfout: drie verschillende definities werden gevonden in Efron (1983); Efron and Tibshirani (1993), and Hastie and Tibshirani (1990).

Hoofdstuk 3 begint met een vergelijking van de voorspelnauwkeurigheid van CATREG met die van vijf andere regressiemethoden die transformaties gebruiken om niet-lineaire relaties te fitten. Resultaten van het schatten van de verwachte voorspelfout van deze vijf methoden met de .632-bootstrap voor de Ozone data (een in de literatuur regelmatig gebruikte dataset) worden gegeven in Hastie and Tibshirani (1990, sec. 10.3). De voorspelnauwkeurigheid van CATREG wordt ook vergeleken met die van ACE (Breiman and Friedman 1985), een methode die nauw verwant is aan CATREG. De resultaten tonen aan dat voor de Ozone data de voorspelnauwkeurigheid van CATREG even goed of iets beter is dan die van de methoden waarmee CATREG vergeleken wordt.

Vervolgens worden de schalingsniveaus van CATREG met elkaar vergeleken in termen van voorspelnauwkeurigheid. Deze vergelijking bevestigt de alge-

mene veronderstelling dat, hoewel de geobserveerde voorspelfout (per definitie) groter wordt naarmate transformaties restrictiever zijn, restricties de verwachte voorspelfout verkleinen. Als echter de relatie tussen de te voorspellen variabele en een predictor in aanzienlijke mate niet-lineair is, kan een restrictievere transformatie tot minder nauwkeurige voorspellingen leiden. Of in zulke gevallen een restrictievere transformatie de voorspelnauwkeurigheid vermindert en in welke mate, is afhankelijk van het belang van de predictor en de relatie van die variabele met de andere predictoren. Voor monotone transformaties is de .632-bootstrap toegepast op CATREG met en zonder multipele systematische starts. Vergelijking van de voorspelnauwkeurigheid in beide gevallen werpt een nieuw licht op lokale minima: voor de Ozone data is de verwachte voorspelfout lager met lokale-minimaoplossingen dan met globale-minimumoplossingen. Dus hoewel de globale-minimumoplossing de geobserveerde voorspelfout minimaliseert (per definitie), gaat dit niet altijd op voor de verwachte voorspelfout.

Hoofdstuk 3 besluit met een evaluatie van het effect van het aantal vrijheidsgraden op de voorspelnauwkeurigheid met niet-monotone transformaties (die minder restrictief zijn) door het aantal observaties en het aantal categorieën van de predictoren te variëren. Zoals verwacht verbetert de voorspelnauwkeurigheid naarmate het aantal effectieve vrijheidsgraden afneemt, d.w.z. naarmate het aantal observaties toeneemt en/of het aantal categorieën afneemt.

Hoofdstuk 4 behandelt methoden voor geregulariseerde regressie en de implementatie daarvan in CATREG. In de context van lineaire regressie wordt regularisatie vaak toegepast om de voorspelnauwkeurigheid te verbeteren door het stabiliseren van de regressiecoefficiënten. Met sommige regularisatiemethoden kan tegelijkertijd selectie van predictoren plaatsvinden. Er zijn vele regularisatiemethoden en -algoritmes waarvan er drie (Ridge Regression, Lasso en Elastic Net) zijn opgenomen in de CATREG-methode, waardoor regularisatie mogelijk wordt voor niet-lineaire regressieproblemen. In de literatuur zijn diverse nogal complexe en/of tijdrovende algoritmes voorgesteld voor het schatten van lineaire Lasso-modellen, hoewel een recente ontwikkeling een elegante en efficiënte procedure (LARS, Efron et al. 2004) levert. Al deze algoritmes hebben de beperking dat, in het geval het aantal predictoren (P) het aantal observaties (N) overtreft, maximaal N predictoren geselecteerd kunnen worden. Door gebruik van de backfitting-procedure vormt de CATREG-Lasso, hoewel iets minder efficiënt dan LARS, eveneens een simpele en niet-tijdrovende methode die bovendien toepasbaar is op niet-lineaire regressieproblemen en niet beperkt is tot selectie van maximaal N predictoren.

Het effect van regularisatie is dat de regressiecoefficiënten krimpen, waardoor ze stabieler zijn. Krimping van de coefficiënten wordt bereikt door een

zogenaamde "penaltyparameter" toe te voegen aan het model, die een sanctie oplegt aan de (absolute) waarden van de regressiecoefficiënten. Met Ridge Regression wordt een sanctie opgelegd aan de kwadratensom van de coefficiënten (L_2-penalty), met de Lasso aan de som van de absolute waarden (L_1-penalty), en met Elastic Net aan zowel de kwadratensom als de sum van de absolute waarden. In lineaire regressie worden de coefficiënten allemaal tegelijk geschat uit de correlatiematrix van de predictoren, waardoor het schatten van coefficiënten die worden gekrompen door een L_1-penalty op te leggen een gecompliceerde zaak wordt. Met backfitting daarentegen kunnen coefficiënten die worden gekrompen door een L_1-penalty wel rechttoe rechtaan geschat worden omdat in deze procedure de correlatiematrix geen rol speelt: met backfitting worden de coefficiënten één voor één geschat, waarbij voor het schatten van de coefficiënt voor een bepaalde predictor de te voorspellen variable wordt gecorrigeerd voor het effect van de andere predictoren. Met backfitting kan de penaltyterm rechtstreeks worden opgenomen in de vergelijking voor de schatting van een coefficiënt. Voor gewone lineare regressie is backfitting natuurlijk een onnodig inefficiënte procedure en daarom niet gebruikelijk. Maar voor geregulariseerde regressie met een L_1-penalty levert backfitting een rechttoe rechtaan en eenvoudige schattingsmethode voor gekrompen coefficiënten, toepasbaar op zowel lineaire als niet-lineaire regressieproblemen.

Beginnend met een penalty van nul (resulterend in het volledige model: het model dat alle predictoren bevat met ongekrompen coefficiënten) en stapsgewijze verhoging van de waarde van de penalty, volgen de Ridge-coefficiënten paden die met verschillende snelheden naar nul gaan, maar nul nooit helemaal bereiken. Met andere woorden: hogere waarden van de penalty krimpen de coefficiënten sterker, en sommige coefficiënten sterker dan andere, maar geen van de coefficiënten wordt naar nul gekrompen. Dus voor elke penaltywaarde bevat een Ridge-model alle variabelen. Om de Ridge paden te construeren moet een groot aantal geregulariseerde modellen geschat worden met penalties tussen nul en een waarde die hoog genoeg is om alle variabelen naar bijna nul te krimpen.

In tegenstelling tot de Ridge-paden gaan de Lasso- en Elastic-Net-paden met verschillende snelheden naar exact nul, dus verschillende waarden van de penalty resulteren in modellen met een verschillend aantal predictoren. Op het punt waar de coefficiënt voor een predictor naar nul gekrompen is, verandert het aantal predictoren in het model en daarom worden zulke punten transitiepunten genoemd. Met lineaire geregulariseerde regressie zijn de Lasso- en Elastic-Net-paden lineair tussen twee aangrenzende transitiepunten en kunnen daarom geconstrueerd worden door de transitiepunten te verbinden met een rechte lijn, vooropgesteld dat de transitiepunten gevonden kunnen worden.

Het vinden van de transitiepunten, met een rekenbelasting in dezelfde orde van grootte als voor gewone regressie, is gerealiseerd in de LARS methode (Efron et al. 2004). De CATREG-Lasso voor lineaire regressie vindt de transitiepunten door gebruik te maken van de verandering in de helling van de paden op een transitiepunt. Hoewel minder efficiënt dan de LARS-Lasso, is deze aanpak aanzienlijk minder rekenintensief dan het schatten van modellen voor een groot aantal waarden van de penalty. Met de niet-lineaire CATREG-Lasso bestaan de paden echter niet uit rechte lijnen tussen twee aangrenzende transitiepunten en moeten er dus wel modellen geschat worden voor een groot aantal waarden van de penalty. Hierbij hangt het aantal te schatten modellen af van de stapgrootte die gekozen wordt om de penalty te verhogen.

Voor het selecteren van het optimale geregulariseerde model (hetgeen equivalent is aan het selecteren van de optimale waarde van de penalty) worden gewoonlijk analytische selectiecriteria als BIC, AIC en GCV gebruikt. Voordat het LARS-algoritme was ontwikkeld vereisten resampling methoden zoals kruisvalidatie en de bootstrap te veel tijd. Met zowel de LARS-Lasso als de CATREG-Lasso zijn resampling methoden wel uitvoerbaar binnen een redelijk tijdsbestek. Het voordeel van resampling boven een analytische benadering is dat het aantal effectieve vrijheidsgraden niet geschat hoeft te worden. De toepassing van analytische selectiecriteria wordt vergeleken met modelselectie via de .632-bootstrap voor zowel lineaire als niet-lineaire CATREG-Lasso.

Recent hebben Yuan and Lin (2006) de Grouped Lasso ontwikkeld, een Lasso methode voor variantieanalyse en additive modellen. Kim et al. (2006) hebben deze methode uitgebreid naar toepassing op algemenere verliesfuncties en noemen de ruimer toepasbare methode Blockwise Sparse Regression (BSR). In deze methoden wordt een categorische predictor vertegenwoordigd door een groepen (blok) dummievariabelen. Deze dummievariabelen worden in de analyse behandeld als een blok door een normrestrictie op te leggen aan hun coefficiënten. In Hoofdstuk 4 wordt aangetoond dat deze normrestrictie equivalent is aan de gewogen standaardisatie van nominale categoriekwantificaties in CATREG. Dus voor nominale predictoren zijn de Grouped Lasso en BSR equivalent aan de CATREG-Lasso. Echter, in de CATREG-Lasso worden nominale variabelen rechtstreeks getransformeerd en hun coefficiënten gekrompen, zonder dat dummievariabelen hoeven te worden gebruikt. Met Grouped Lasso en BSR kunnen ook niet-monotone transformaties van continue variabelen worden toegepast door die variabelen te vertegenwoordigen door een groep basisfuncties, zoals polynomen. Met de CATREG-Lasso kunnen continue variabele rechtstreeks (niet-monotoon of monotoon) getransformeerd worden met krimping van de coefficiënten door een spline transformatie toe te passen. Als dummievariabelen worden gebruikt voor nominale

variabelen zonder normrestrictie op de coefficiënten, worden de categorieën gekrompen in plaats van de variabele als geheel. Hoofdstuk 4 illustreert ook het effect van deze vorm van krimping.

In hoofdstuk 5 worden de CATREG-Lasso en de .632-bootstrap toegepast op een nieuwe dataset uit de psychotherapeutische praktijk, gevormd door een steekproef van patiënten die in verschillende mate lijden aan bulimia nervosa. De data werden verzameld door het Departement Psychosomatische Geneeskunde en Psychotherapie van de Universiteit van Freiburg en bestaan uit een beoordeling door een expert van de ernst van de aandoening (op een vijfpuntsschaal van niet tot zeer ernstig) en een standaardbatterij vragenlijsten voor zowel expert- als zelfrapportagemetingen als predictoren. De onderzoeksdoelen waren: het selecteren van een optimale kleine subset van predictoren om de ernst van de aandoening te voorspellen; het bepalen van het scoregebied van de predictoren dat bijdraagt aan de voorspellingen; en het empirisch bepalen van een drempelwaarde voor het classificeren van de patiënten in twee groepen: klinische gevallen (vrij tot zeer ernstig) en gezonde/subklinische gevallen (niet en weinig ernstig). In het bereiken van deze doelen wilden de onderzoekers er rekening mee houden dat de variabelen niet op intervalniveau waren gemeten en dat mogelijk niet-lineaire verbanden tussen de variabelen bestonden. Hun zoektocht naar deskundigheid op deze gebieden leidde tot samenwerking met de auteur van dit proefschrift, die de statische analyses heeft uitgevoerd en grotendeels de “Statische Analyse” en “Resultaten” secties van het resulterende artikel heeft geschreven.

De CATREG-Lasso wordt in deze toepassing niet gebruikt om een gekrompen model te selecteren, maar om een subset van predictoren te selecteren. Aan de hand van de resultaten van een CATREG-Lasso-analyse worden modellen met alle mogelijke aantallen predictoren (van één tot en met alle behalve één) opgesteld met ordinale en monotone spline transformaties. Vervolgens wordt de .632-bootstrap toegepast op deze modellen om de optimale subset van predictoren te selecteren. Uit de transformaties blijkt dat de geselecteerde variabelen informatief zijn voor het maken van onderscheid tussen klinische en gezonde/subklinische gevallen, alsmede voor het maken van onderscheid tussen gezonde en subklinische gevallen, maar minder informatief voor het maken van onderscheid binnen de klasse van klinische gevallen. Het beperkte vermogen van het geselecteerde model in het maken van onderscheid tussen de zwaardere en minder zware klinische gevallen wordt ook gevonden met modellen waarin meer of andere predictoren zijn opgenomen. Voor evaluatie van de diagnostische validiteit van het geselecteerde model worden de patiënten verdeeld in twee groepen - klinisch en gezond/subklinisch - en worden drempelwaarden bepaald met ROC-curve-analyse op de voorspelde ernst van de

aandoening. Met het geselecteerde model blijken drempelwaarden bepaald te kunnen worden voor drie soorten goede classificaties: ten eerste een classificatie met zeer hoge *sensitiviteit* (proportie correct-positief) en acceptabele *specificiteit* (proportie correct-negatief), ten tweede een classificatie met zeer hoge specificiteit en acceptabele sensitiviteit en ten derde een classificatie met zowel hoge sensitiviteit als hoge specificiteit.

De CATREG methode is opgenomen in de Categories module van het statistische softwarepakket SPSS (Meulman et al. 1999, 2004). De auteur van dit proefschrift heeft in belangrijke mate bijgedragen aan de implementatie en de documentatie van het CATREG progamma. In Appendix A wordt een gedetailleerde beschrijving van het CATREG algoritme gegeven. Appendix B bevat de CATREG hoofdstukken uit de SPSS Categories handleiding.