High-dimensional time series analysis

Nederlandse Samenvatting

We bevinden ons momenteel in een nieuw tijdperk van data-analyse, dat gekarakteriseerd wordt door de beschikbaarheid van grote, ongestructureerde datasets. U kunt hierbij denken aan data die wordt verzameld door grote tech-bedrijven zoals Google en Facebook, maar ook gegevens die verzameld worden via de klantenkaart van de lokale supermarket en de betaalpas waarmee afgerekend wordt. Omdat traditionele statistische modellen vaak het beste werken wanneer er rekening gehouden dient te worden met de effecten van *slechts enkele* variabelen, zijn er de laatste jaren veel nieuwe statistische methoden ontwikkelt die beter toepasbaar zijn op grote datasets. Deze nieuwe methoden worden ook wel hoog-dimensionale statistieken genoemd. Echter, binnen economische en financiële sectoren, werkt men met name met tijdreeksen, zoals bijvoorbeeld de Nederlandse werkloosheidcijfers of het bruto binnenlands product. Tijdreeksen vertonen vaak unieke eigenschappen, zoals trendmatig gedrag waarbij toekomstige waardes sterk afhangen van het verleden, waarvan we weten dat ze de uitkomsten van traditionele statistieken sterk beïnvloeden. Het is daarom niet verstandig om hoog-dimensionale statistieken toe te passen op grote verzamelingen van tijdreeksen zonder theoretische verificatie of praktische aanpassingen. Dit onderwerp staat centraal in mijn proefschrift.

In dit proefschrift, richten we ons enkel op statistische methoden welke onder te verdelen zijn in drie algemene categorieën: (1) factor modellen, (2) geregulariseerde regressie en (3) hybride modellen. Het idee achter factormodellen is dat alle waargenomen variabelen worden aangedreven door enkele latente (niet geobserveerde) variabelen. Zo kunnen we bijvoorbeeld werkloosheid observeren binnen verschillende industrieën, of rentetarieven voor verschillende looptijden, maar worden al deze variabelen mogelijk (deels) verklaard door de onderliggende bedrijfsconjunctuur. Factor modellen proberen deze latente variabelen, de factoren, te schatten en daarmee de

data samen te vatten met een minimum verlies aan informatie. Op deze manier hoeft er geen complex model met honderden geobserveerde variabelen geschat te worden. Een alternatieve methode is om de data niet samen te vatten, maar om ervan uit te gaan dat veel variabelen simpelweg irrelevant zijn voor het verklaren van de afhankelijke variabele waar men in geïnteresseerd is. Zo is het aannemelijk dat de grondstofprijzen voor thee van invloed zijn op de verkoop van koffie, maar dat de grondstofprijzen voor ketchup hier weining in verklaren. Voor dit soort applicaties is geregulariseerde regressie uitermate geschikt. Deze vorm van regressie schat een lineair model en zorgt er automatisch voor dat de geschatte bijdrages van irrelevante variabelen omlaag geschaald worden. Sommige vormen van geregulariseerde regressie, zoals de Least Absolute Shrinkage and Selection Operator (LASSO) welke een belangrijke rol in dit proefschrift heeft, hebben de wenselijke eigenschap dat ze irrelevante variabelen geheel automatisch uit het geschatte model kunnen verwijderen. Als laatste optie komen in dit proefschrift hybride methoden aan bod, welke irrelevante variabelen verwijderen en de relevante variabelen middels het schatten van factoren samenvatten.

In Hoofdstuk 2 vergelijken we de voorspellingsprestaties van statistische methoden welke onder te verdelen zijn middels de bovenstaande categorisatie. Door het uitvoeren van gecontroleerde simulaties waarin we bepaalde data eigenschappen doelbewust vastleggen, vinden we dat factor modellen en geregulariseerde regressie goed presteren in het kader waar ze voor ontwikkeld zijn, maar ontdekken we ook dat geregulariseerde regressie beter kan voorspellen indien er factoren in de data aanwezig zijn met "veel ruis".[1] In een empirische toepassing vinden we dan ook dat voor sommige Amerikaanse economische indicatoren geregulariseerde regressie nauwkeuriger voorspelt dan factor modellen, ondanks dat de aanwezigheid van factoren in een macro-economische toepassing zeer aannemelijk is.

Gemotiveerd door de gunstige prestaties van geregulariseerde regressie, ontwikkelen we in Hoofdstuk 3 de Single-equation Penalized Error-Correction Selector (SPECS). SPECS is een gespecializeerde methode waarmee geregulariseerde lineaire modellen geschat kunnen worden die rekening houden met het trendmatige gedrag van de beschouwde variabelen. Zo komt het in economische toepassingen geregeld voor dat individuele variabelen een stochastische (willekeurige) trend bevatten, maar dat deze trend verdwijnt na het nemen van een bepaalde lineaire combinatie. Dit welbekende fenomeen heet cointegratie en heeft grote invloed op het gedrag van statistieken. Wij

[1] Dit is een simplificatie ter bevording van de leesbaarheid. De preciezere omschrijving is dat cross-sectionele correlatie in het idiosyncratische component de nauwkeurige schatting van factoren belemmert.

leiden theoretische (asymptotische) resultaten af die laten zien dat onze methode zich wenselijk gedraagt wanneer de steekproefgrootte groeit. Ter demonstratie van de toepasbaarheid van SPECS, gebruiken we onze nieuwe methode om de werkloosheid in Nederland te voorspellen aan de hand van de populariteit van 100 verschillende Google zoektermen, waaronder bijvoorbeeld "werkloosheidsuitkering" en "solliciteren". In lijn der verwachtingen, overtreft SPECS de voorspellingsprestaties van hoog-dimensionale statistieken welke cointegratie negeren.

In Hoofdstuk 4 leiden we vergelijkbare theoretische resultaten af onder minder restrictieve aannames. Zo laten we toe dat het aantal variabelen in het model mag toenemen wanneer de steekproefgrootte toeneemt. Dit is van belang om een duidelijk inzicht te geven in het gedrag van SPECS bij toepassingen op datasets met een groot aantal variabelen.

Ten slotte, in Hoofdstuk 5 vergelijken we (1) statistische testen om het trendmatig gedrag van tijdreeksen te classiferen en (2) een selectie aan hoog-dimensionale voorspellingsmethoden welke cointegratie al dan niet in acht nemen. Middels simulaties vinden we dat het uitermate belangrijk is om de trend in de afhankelijke variabele juist te classificeren, gezien de nauwkeurigheid waarmee deze variabele voorspeld kan worden sterk van deze classificatie afhangt. In een macro-economische toepassing op een Amerikaanse dataset vinden we dat geen enkel model consistent het nauwkeurigst voorspelt en is er ook geen definitief antwoord op de vraag of cointegratie belangrijk is voor het maken van voorspellingen. Gezien er gevallen zijn waarin SPECS beter presteert dan de andere methodes in de vergelijking, bevestigen we dat onze methode zowel theoretische als toegepaste waarde heeft. Echter, zal de keuze voor de optimale methode altijd van de specifieke toepassing afhankelijk zijn.