Technische Universiteit Delft
Faculteit Elektrotechniek, Wiskunde en Informatica
Delft Institute of Applied Mathematics

Dantzig Selector: Het schatten van een sparse parametervector met weinig metingen

(Engelse titel: Dantzig Selector: Statistical estimation of a sparse vector with undersampled data)

Verslag ten behoeve van het
Delft Institute for Applied Mathematics
als onderdeel ter verkrijging

van de graad van

BACHELOR OF SCIENCE
in
TECHNISCHE WISKUNDE

door

WIKASH SEWLAL

Delft, Nederland
Augustus 2010

BSc verslag TECHNISCHE WISKUNDE

"Dantzig Selector: Het schatten van een sparse parametervector met weinig metingen"

(Engelse titel: "Dantzig Selector: Statistical estimation of a sparse vector with undersampled data)"

Wikash Sewlal

Technische Universiteit Delft

Begeleider

Prof.dr.ir. G. Jongbloed

Overige commissieleden

Prof.dr.ir. C. Vuik	Dr. G.F. Ridderbos
...	...

Augustus, 2010 Delft

Samenvatting

Dit verslag gaat over het lineaire model $y = X\beta + \varepsilon$, waarbij het doel is om de parametervector $\beta \in \mathbb{R}^p$ te schatten gegeven de metingen $y \in \mathbb{R}^n$ en de $n \times p$ matrix X.

Bij lineaire regressie geldt in het algemeen, dat het aantal metingen (n) groter is dan het aantal parameters (p). In praktische situaties kan het echter voorkomen, dat het aantal metingen veel kleiner is dan het aantal parameters. De kleinste kwadratenschatter kan in dat geval niet gebruikt worden om β te schatten, er is geen unieke oplossing voor de normaalvergelijkingen. Wanneer echter gegeven is dat een klein aantal parameters niet-nul is, is het soms toch mogelijk om de parameters te schatten.

In 2005 introduceerden Emmanuel Candès en Terence Tao de *Dantzig Selector* (DS) als een schatter voor β in deze situatie. Deze is gedefinieerd als

$$\arg\min_{\tilde{\beta}} \|\tilde{\beta}\|_{\ell_1} \qquad \text{odv} \qquad \left\|X^T(y - X\tilde{\beta})\right\|_{\ell_\infty} \leq \lambda_p \sigma. \tag{DS}$$

In dit verslag wordt uitgelegd hoe de Dantzig Selector werkt. Tevens wordt op theoretische- en praktische wijze nagegaan hoe goed de DS werkt.

VOOR u ligt het verslag van het bacheloreindproject dat door een student Technische Wiskunde van de Technische Universiteit Delft is uitgevoerd. In het bachelorproject worden de in de bachelorfase verworven kennis en vaardigheden gebruikt om een probleem te bestuderen.

Het probleem dat in dit verslag bekeken wordt, is het schatten van een parametervector onder speciale omstandigheden. Hierbij wordt er gekeken naar de Dantzig Selector als schatter van deze vector.

Graag wil ik Geurt Jongbloed bedanken voor het begeleiden van dit project. Ik heb erg veel geleerd en genoten van de vele gesprekken waar in, voor mij, nieuwe wiskundige connecties werden gelegd. Tevens wil ik u bedanken voor de vrijheid in de richting van het project. Het was een erg leerzame ervaring.

Verder wil ik de heren Ridderbos en Vuik bedanken voor hun zitting in de beoordelingscommissie.

Ten slotte wens ik u veel leesplezier bij het lezen van dit verslag.

Wikash Sewlal,

Auteur

Inhoudsopgave

1 Inleiding

Dit hoofdstuk geeft een inleiding in het bachelorproject. In dit hoofdstuk zal de oorsprong van het probleem gegeven worden. Hierbij zal vermeld worden, welke aandacht er via publicaties aan dit probleem is gegeven. Tot slot wordt de indeling van dit verslag gegeven, zodat de lezer weet wat in welk hoofdstuk te vinden is.

1.1 Achtergrond

1.1.1 Lineaire regressie Bij regressie wordt gezocht naar een samenhang tussen de responsvariabele en de verklarende variabelen. Indien aangenomen wordt dat deze samenhang volgens een lineair model te beschrijven is, praten we over lineaire regressie[1]. In dat geval is het doel om de parameters zo goed mogelijk te schatten, zodat een zo goed mogelijke weergave gegeven wordt van het verband tussen de responsvariabele en de verklarende variabelen.

In het klassieke geval is het aantal metingen (veel) groter dan het aantal parameters. In dit geval kunnen de waarden van de parameters geschat worden door gebruik te maken van de kleinste kwadratenmethode. Het complementaire geval, waarin het aantal metingen kleiner is dan het aantal parameters, krijgt traditioneel weinig aandacht. Het kleinste kwadraten minimaliseringsprobleem heeft in het complementaire geval oneindig veel oplossingen en kan dus niet gebruikt worden. De vraag is ook of het zinvol is om over een oplossing te spreken; extra eisen zijn nodig.

1.1.2 Concrete situaties met weinig metingen komen voor! Er zijn concrete situaties waarbij een lineair model aangenomen wordt en het aantal parameters (veel) groter is dan het aantal metingen. Verder bestaan er toepassingen waarbij het afnemen van een meting veel tijd en geld kost, het schatten van de parameters met zo weinig mogelijk metingen is dan gewenst.

[1] In dit verslag wordt aangenomen dat er meerdere verklarende variabelen zijn, er is dus sprake van *meervoudige* lineaire regressie.

Een voorbeeld hiervan is MRI [16], waarbij men een afbeelding (β) wil bepalen aan de hand van door ruis verstoorde indirecte metingen ($X\beta+\varepsilon$) in een ander domein. Een ander voorbeeld is te vinden in de signaalverwerking, waar een signaal in continue tijd bepaald moet worden aan de hand van eindig veel metingen met ruis. Een totaal ander voorbeeld is genen-onderzoek, waarbij er uit duizenden verschillende genen enkel die genen bepaald moeten worden die verantwoordelijk zijn voor bijvoorbeeld een erfelijke aandoening.

Merk op dat in de zojuist genoemde toepassingen er aangenomen wordt, dat een klein aantal parameters het gehele model bepaalt. Welke parameters dit zijn, is echter niet bekend; dit vormt een essentieel onderdeel van het probleem. Bij de MRI afbeeldingen van zowel het brein als angiogrammen geeft [17] aan dat deze sparse zijn ten opzichte van een ander domein (discrete cosinus transformatie (DCT) en wavelet transformatie). Over het algemeen geldt, dat veel natuurlijke signalen sparse zijn ten opzichte van een andere basis (sparse or nearly sparse spectrum), [10]. Merk op, dat ook bij het voorbeeld van genen-onderzoek wordt aangenomen, dat maar een klein aantal genen bepalend is.

Deze aanname werpt de vraag op of de parameters in deze situatie wel goed te schatten zijn. Is het mogelijk om de parameters te vinden die het model bepalen? Zo ja, kunnen alle parameters dan geschat worden en hoe kun je in dit geval alle parameters zo goed mogelijk schatten?

1.1.3 Literatuur

In de literatuur is er een lange tijd weinig aandacht gegeven aan het zojuist genoemde probleem. Er bestonden oplossingsmethoden, zoals bijvoorbeeld ridge regression, all-subset regression en stepwise regression, maar volgens [15] hadden al deze methoden tekortkomingen. Zo wordt stepwise genoemd als een greedy algoritme, kost het uitvoeren van all-subset regression te veel aan rekenkracht en zorgt ridge regression niet voor het selecteren van de juiste parameters.

Een oplossingsmethode die zowel de juiste parameters selecteerde als de parameters aardig goed wist te schatten, kwam in 1996, toen Tibshirani een paper publiceerde over de Least Absolute Shrinkage and Selection Operator (LASSO) [20]. Belangrijk hierbij was het gebruik van de ℓ_1-norm, waardoor onbelangrijke parameters op nul gezet worden.

In [14] werd het Least Angle Regression algoritme geïntroduceerd, dat met een aanpassing ook de LASSO schatter kon berekenen. Deze combinatie, genaamd LARS, zorgde ervoor dat er een snel algoritme was, waarmee de LASSO bepaald kon worden.

1.1.4 Compressed Sensing

In 2006 verscheen een reeks artikelen op het gebied van signaalverwerking, waarbij centraal stond in welke gevallen een hogerdimensionale vector — lees: een digitaal signaal of een afbeelding — met weinig metingen gereconstrueerd kan worden. Hoewel er eerder belangrijke artikelen verschenen over dit onderwerp, bijvoorbeeld [11, 13], trok het verschijnen van de artikelen [4, 5, 6, 7, 8]

veel aandacht. De introductie van het uniform uncertainty principle en de restricted isometries in [7, 6], gaven een fundament voor het mogelijk maken van een reconstructie bij weinig metingen.

Chen, Donoho en Saunders stelden in [11] voor om de meest sparse vectoren te vinden via het minimaliseren op de ℓ_1-norm, deze methode heet Basis Pursuit. Voor metingen met ruis stelden de auteurs voor om Basis Pursuit Denoising (BPDN) te gebruiken. Grappig genoeg is BPDN eigenlijk hetzelfde als de LASSO, het verschil is dat BPDN geschreven staat in de vorm met een ℓ_1-penalty.

Candès en Tao introduceerden in [8] een schatter genaamd de Dantzig Selector (DS). Deze schatter kenmerkt zich doordat deze naast het selecteren van de juiste parameters, ook de parameters zo goed mogelijk probeert te schatten. Een verrassend resultaat in het artikel was, dat de DS onder bepaalde voorwaarden een kwaliteitsgarantie kan geven.

Dit verslag beschrijft hoe de Dantzig Selector gebruikt kan worden om de parameters te schatten wanneer het aantal metingen kleiner kan zijn dan het aantal parameters; hierbij zal aangenomen worden dat het aantal invloed hebbende parameters klein is. Vanwege de vele gelijkenissen tussen de DS en de LASSO zullen overeenkomsten en verschillen tussen de schatters besproken worden.

1.2 Probleemstelling

Lineair model

We beschouwen het model

$$y = X\beta + \varepsilon, \tag{1.1}$$

waarbij

$y \in \mathbb{R}^n$ de n metingen voorstelt;
X een $n \times p$ *design*-matrix is;
$\beta \in \mathbb{R}^p$ de parametervector is;
$\varepsilon \in \mathbb{R}^n$ ruis voorstelt, bijvoorbeeld $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$.

Aanamen

In het bachelorproject is gekeken, hoe de parametervector β in dit model geschat kan worden onder enkele aannamen. Allereerst zal er worden aangenomen, dat het aantal metingen n veel kleiner is dan het aantal parameters p. Verder wordt er aangenomen, dat het aantal niet-nullen in de parametervector β kleiner dan of gelijk is aan S; dit zal de sparsity aanname worden genoemd. Homoscedasticiteit is uiteraard niet vanzelfsprekend, maar zal wel aangenomen worden. Ten slotte wordt ook aangenomen dat er geen intercept is.

Het doel is om, gegeven de metingen y en een designmatrix X, de parametervector β zo goed mogelijk te schatten.

1.2.1 Twee oplossingsstrategiën Zoals in de vorige paragraaf is vermeld, werkt de kleinste kwadratenmethode niet bij deze probleemstelling; waarom dit is, zal in hoofdstuk 2 behandeld worden. De oplossingsmethode die in dit bachelorproject bekeken is, is de Dantzig Selector. De Dantzig Selector (DS) is een schatter voor β welke gedefinieerd is als

Dantzig Selector

$$\arg\min_{\tilde{\beta}} \|\tilde{\beta}\|_{\ell_1} \qquad \text{odv} \qquad \|X^T(y - X\tilde{\beta})\|_{\ell_\infty} \leq \lambda_p \sigma. \tag{DS}$$

Hierbij is λ_p een zelf in te stellen parameter. De afkorting odv staat voor 'onder de voorwaarden'. Met arg wordt het argument genoemd; de DS is dus de $\tilde{\beta}$ die voldoet aan $\|X^T(y - X\tilde{\beta})\|_{\ell_\infty} \leq \lambda_p \sigma$ en minimaal is in de ℓ_1-norm.

Een andere schatter die veel aandacht in de literatuur heeft gekregen én een beetje lijkt op de DS, is de Least Absolute Shrinkage and Selection Operator (LASSO). In dit verslag wordt de LASSO daarom ook in bepaalde stukken meegenomen om verschillen en overeenkomsten uit te lichten. Voor een zelf gekozen waarde s is de LASSO-schatter gedefinieerd als

LASSO

$$\arg\min_{\tilde{\beta}} \|y - X\tilde{\beta}\|_{\ell_2}^2 \qquad \text{odv} \qquad \|\tilde{\beta}\|_{\ell_1} \leq s. \tag{LASSO}$$

1.2.2 Subdoelen Bij de beschrijving van de probleemstelling en de twee schatters komen veel vragen naar boven. Tijdens dit bachelorproject is er getracht een antwoord te vinden op de onderstaande vragen.

1. Hoe werkt de Dantzig Selector?
2. Werkt de DS voor elke designmatrix X?
3. Hoe wordt de waarde van λ_p gekozen?
4. Hoe sparse mag/moet β zijn om deze nog goed te kunnen schatten?
5. Hoe "meet" je hoe goed een schatter werkt?
6. Wat zijn de overeenkomsten en de verschillen tussen de LASSO en de DS?
7. Heeft de Dantzig Selector voordelen ten opzichte van de LASSO, en andersom?

In het bachelorproject — en dus ook in dit verslag — zijn niet alle vragen behandeld. Zo is er in dit verslag niet veel te vinden over de vragen 4 en 7.

1.3 Inhoud verslag

1.3.1 Voorkennis Medestudenten Technische Wiskunde van het derde jaar behoren tot de doelgroep van dit verslag. Dit betekent dat dit verslag te begrijpen is voor nominaal lopende studenten uit het derde jaar. Het is belangrijk, dat de lezer bekend is met lineaire regressie en de kleinste kwadraten methode.

De theorie in het bachelorproject maakt veel gebruik van begrippen uit de lineaire algebra; er wordt bijvoorbeeld gebruik gemaakt van eigenschappen van orthogonale matrices en de definities van de DS en de LASSO maken gebruik van verschillende normen ($\ell_1, \ell_2, \ell_\infty$). Een overzicht van de meest belangrijke voorkennis is te vinden in appendix A.

1.3.2 Indeling verslag

Het verslag is opgedeeld in zes hoofdstukken.

Hoofdstuk 2 begint met het uitleggen waarom de kleinste kwadratenmethode niet werkt bij dit probleem. Vervolgens wordt de werking van de DS uitgelegd.

Hoofdstuk 3 bevat definities en uitleg over de begrippen uit de onderliggende theorie van Compressed Sensing, een techniek waarbij een hogerdimensionale vector met weinig metingen gereconstrueerd kan worden. Via de geïntroduceerde begrippen kan een voorwaarde gesteld worden, dat het schatten van de parametervector β in het ruisloze geval mogelijk maakt.

Hoofdstuk 4 begint met een theoretische beschouwing over de kwaliteit van een schatter. Het hoofdstuk behandelt enkele bewijzen over de kwaliteit van schatters en via de 'oracle inequalities' zal de DS geopperd worden als methode van variabele selectie.

Hoofdstuk 5 geeft een praktische beschouwing van de kwaliteit van de DS. Het hoofdstuk presenteert de opzet, resultaten en conclusies van een simulatiestudie, welke uitgevoerd is om meer inzicht te verkrijgen in de werking van de DS én om te verifiëren dat de DS überhaupt werkt.

Hoofdstuk 6 bespreekt in een discussie de resultaten van het bachelorproject. Het geeft aan waarom de DS gebruikt kan worden voor bepaalde toepassingen en waarom voor andere juist niet.

In de appendices is een samenvatting van de benodigde voorkennis te vinden. Naast volledige uitwerkingen van bewijzen van stellingen is de gebruikte MATLAB® code daar ook te vinden. De gebruikte literatuur is te vinden op pagina 70.

1.3.3 Leesvolgorde

Het is aan te raden om het verslag in de geschreven volgorde te lezen en om de gebruikte notatie, te vinden in hoofdstuk 2, te raadplegen. Hoofdstuk 3 behandelt begrippen uit de techniek *Compressed Sensing*. Wanneer deze begrippen bekend zijn bij de lezer, dan kan dit hoofdstuk overgeslagen worden.

2

Werking DS

In dit hoofdstuk staat de werking van de Dantzig Selector centraal. Het hoofdstuk begint met een verklaring waarom de (standaard) kleinste kwadratenmethode niet werkt bij dit probleem. Vervolgens wordt de definitie van de Dantzig Selector bekeken, waarbij wordt uitgelegd hoe de DS werkt en waarom deze gebruikt kan worden in de situatie zoals genoemd in de probleemstelling.

2.1 Notatie

Beschouw het model uit de probleemstelling (paragraaf 1.2),

$$y = X\beta + \varepsilon. \tag{1.1}$$

Laat $X = (X_1, \ldots, X_p)$ de $n \times p$ designmatrix, $\beta = \left(\beta_1, \ldots, \beta_p\right)^T$ de parametervector en J de verzameling indices $J = \{1, \ldots, p\}$ zijn. De kolommen van X worden bij gebruik met subscript als vectoren $X_1, \ldots, X_p \in \mathbb{R}^n$ beschouwd. V is de lineaire deelruimte opgespannen door deze vectoren, $V = \text{span}\left\{X_j : j \in J\right\} = \{X\tilde{\beta} : \tilde{\beta} \in \mathbb{R}^p\} \subseteq \mathbb{R}^n$.

Support

De verzameling indices $I = \{i \in J : \beta_i \neq 0\}$, de indices waar β niet gelijk is aan nul, wordt de *support* van de vector β genoemd. In dit verslag zal I altijd gebruikt worden als notatie voor de support van β.

Laat $T \subset J$, dan is $|T|$ het aantal elementen in de verzameling T. We definiëren X_T als een $n \times |T|$ submatrix van X, met kolommen $(X_j)_{j \in T}$. β_T, de notatie van matrices volgend, is een subvector van β. De ingekorte vector β_T bestaat dus uit de elementen van indices T van β. Ter verduidelijking, β_I is de ingekorte vector die bestaat uit de niet-nul elementen van β.

VOORBEELD
Dit voorbeeld dient ter illustratie van de zojuist geïntroduceerde notatie.
Laat $X = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix}$ en $\beta = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 7 \end{pmatrix}$.
De niet-nul elementen van β zijn 1 en 7, de bijbehorende indices zijn 1 en 4. Dit betekent dat I, de support van β, de verzameling $I = \{1,4\}$ is.
De submatrix X_I bestaat uit $X_1 = \begin{pmatrix} 1 \\ 5 \\ 9 \end{pmatrix}$ en $X_4 = \begin{pmatrix} 4 \\ 8 \\ 12 \end{pmatrix}$, $X_I = \begin{pmatrix} 1 & 4 \\ 5 & 8 \\ 9 & 12 \end{pmatrix}$.
Het matrixproduct $X_I\beta_I$ is dan $X_I\beta_I = \begin{pmatrix} 1 & 4 \\ 5 & 8 \\ 9 & 12 \end{pmatrix}\begin{pmatrix} 1 \\ 7 \end{pmatrix} = \begin{pmatrix} 29 \\ 61 \\ 93 \end{pmatrix}$.
Omdat I de support van β is, geldt zelfs $X_I\beta_I = X\beta$.

2.1.1 Gebruik van β, $\tilde{\beta}$ en $\hat{\beta}$

In dit verslag wordt er veelvuldig gebruik gemaakt van de vectoren β, $\tilde{\beta}$ en $\hat{\beta}$.

De 'echte' parametervector, de onderliggende parametervector die geschat moet worden, zal aangeduid worden met β.

De schatters voor β worden aangeduid met een dakje, $\hat{\beta}$. Welke schatter er met $\hat{\beta}$ bedoeld wordt, is te zien door te kijken naar de bijbehorende paragraaf. Indien meerdere schatters bekeken worden, zal een onderschrift ter verduidelijking gebruikt worden; $\hat{\beta}_{\text{DS}}$ is bijvoorbeeld de Dantzig Selector.

De kringel, $\tilde{\beta}$, wordt gebruikt als variabele. Zo wordt er bij de DS geminimaliseerd op $\sum_i |\tilde{\beta}_i|$, waarbij gekeken wordt naar alle vectoren $\tilde{\beta} \in \mathbb{R}^p$ die voldoen aan $\|X^T(Y - X\tilde{\beta})\|_{\ell_\infty} \leq \lambda_p \sigma$.

2.1.2 Probleemstelling revisited

Het introduceren van nieuwe begrippen en notaties nodigt uit om de probleemstelling opnieuw te bekijken. Zoals in paragraaf 1.2 is vermeld, is er in het bachelorproject gekeken naar het lineaire model

$$y = X\beta + \varepsilon. \tag{1.1}$$

In sommige gevallen is het handiger om (1.1) te zien als

$$y = X_1\beta_1 + \ldots X_p\beta_p + \varepsilon, \tag{2.1}$$

zodat meer nadruk gegeven wordt aan het feit, dat y gezien wordt als de som van ruis en een lineaire combinatie van de kolommen van X. In het bachelorproject is het doel om de parameters $\beta_1, \ldots, \beta_p$ te schatten.

Naast het geven van het lineaire model, is er in de probleemstelling ook een aantal aannamen gegeven. Als eerste wordt er aangenomen dat er meer parameters dan

metingen zijn. Aangezien p het aantal parameters is en n het aantal metingen, wordt deze aanname genoemd als "de situatie $n < p$". Verder is er aangenomen dat β sparse is, en wel S-sparse. Dit betekent dat β ten hoogste S elementen heeft die niet gelijk zijn aan nul. De support I heeft dus ten hoogste S elementen, $|I| \leq S$.

De overige aannamen blijven in de huidige notatie nog steeds hetzelfde. De aanname van homoscedasticiteit is zichtbaar doordat $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$ aangenomen wordt. Dat er aangenomen wordt dat er geen intercept is, betekent dat X geen extra kolom heeft, waarbij de waarde van alle elementen uit die kolom gelijk zijn.

2.2 Terugblik: Kleinste kwadratenmethode

Klassieke situatie

In deze paragraaf zal de aanname $n < p$ vervangen worden door $n > p$. Situaties waar β geschat moet worden als er meer metingen zijn dan parameters, wordt in dit verslag beschreven als de *klassieke situatie*. De reden hiervoor is, dat het schatten van de parameters in deze situatie in veel opleidingen behoort tot standaardmaterie. In deze situatie is de kleinste kwadratenmethode een veelgebruikte methode voor het schatten van de parameters.

Bij de kleinste kwadratenmethode, in het Engels 'ordinary least squares' (OLS) genoemd, wordt er naar die $\tilde{\beta}$ gezocht die de data het beste fit. Dit wordt gedaan door van alle vectoren $\tilde{\beta} \in \mathbb{R}^p$ die $\tilde{\beta}$ te kiezen, die de som van de residuen in het kwadraat minimaliseert. De kleinste kwadratenschatter $\hat{\beta}$ is die $\tilde{\beta}$ waarvoor $\sum_{i=1}^{n} \left(y - X\tilde{\beta}\right)_i^2 = \|y - X\tilde{\beta}\|_{\ell_2}^2$ minimaal is.

2.2.1 Normaalvergelijkingen Laat $V = \{X\tilde{\beta} : \tilde{\beta} \in \mathbb{R}^p\} \subseteq \mathbb{R}^n$. Omdat de kleinste kwadratenschatter ook $\|y - X\tilde{\beta}\|_{\ell_2}$ minimaliseert, is $X\hat{\beta}$ het punt in V waarvoor de afstand tussen y en V minimaal is. Dit betekent, dat $X\hat{\beta}$ de orthogonale projectie van y op V is. Merk op dat $y - X\hat{\beta}$ dus loodrecht staat op alle vectoren $X\tilde{\beta} \in V$:

$$\begin{aligned} \left\langle y - X\hat{\beta}, X\tilde{\beta} \right\rangle = 0 \qquad & \text{voor alle } \tilde{\beta} \in \mathbb{R}^p, \\ \tilde{\beta}^T X^T \left(y - X\hat{\beta}\right) = 0 \qquad & \text{voor alle } \tilde{\beta} \in \mathbb{R}^p. \end{aligned}$$

Dit brengt ons bij de normaalvergelijkingen

$$X^T X\hat{\beta} = X^T y. \tag{2.2}$$

Een kleinste kwadratenschatter $\hat{\beta}$, een $\tilde{\beta}$ die $\left\|y - X\tilde{\beta}\right\|_{\ell_2}^2$ minimaliseert, voldoet aan de normaalvergelijkingen (2.2). Als X van volle rang is (hier: rang gelijk aan p), is $X^T X$ dit ook en geldt er

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

2.3 De situatie $n < p$

2.3.1 Waarom de kleinste kwadratenmethode niet werkt bij $n < p$ Bij de kleinste kwadratenmethode wordt er naar een vector $\tilde{\beta}$ gezocht die $\|y - X\tilde{\beta}\|_{\ell_2}^2$ minimaliseert. Het probleem in de situatie $n < p$ is, dat er geen *unieke* $\tilde{\beta}$ is zodanig dat deze som minimaal is.

Dat er geen *unieke* $\tilde{\beta}$ bestaat die $\|y - X\tilde{\beta}\|_{\ell_2}^2$ minimaliseert, kan verklaard worden door te kijken naar de kolommen van X. Merk op dat er bij de kleinste kwadratenmethode gezocht wordt naar het punt $X\hat{\beta} = X_1\hat{\beta}_1 + \ldots + X_p\hat{\beta}_p$ waarvoor deze norm minimaal is. Wanneer de kolommen X_i lineair onafhankelijk zijn, dan is er een unieke combinatie $\hat{\beta} = (\hat{\beta}_1 \ldots \hat{\beta}_p)^T$ om $X\hat{\beta}$ te construeren.

De lineaire algebra leert ons dat $p > n$ vectoren $X_1, \ldots, X_p \in \mathbb{R}^n$ geen lineair onafhankelijk stelsel kunnen vormen, een lineair onafhankelijk stelsel in $\mathbb{R}^n$ heeft immers nooit meer vectoren dan een basis van $\mathbb{R}^n$. Dit betekent, dat er oneindig veel lineaire combinaties bestaan die de vector $X\hat{\beta}$ geven; er bestaat geen *unieke* $\tilde{\beta}$ die $\|y - X\tilde{\beta}\|_{\ell_2}^2$ minimaliseert.

Merk op dat alle parameters $\beta_1, \ldots, \beta_p$ bij de kleinste kwadratenmethode 'meedoen' en dus geschat worden; er wordt gezocht naar een vector, een lineaire combinatie $X\tilde{\beta}$, uit de lineaire deelruimte $V = \{X\tilde{\beta} \in \mathbb{R}^n : \tilde{\beta} \in \mathbb{R}^p\}$. De aanname dat de parametervector S-sparse is, wordt bij de kleinste kwadratenmethode niet gebruikt.

2.3.2 Brute force, een oplossing? De vorige subparagraaf eindigde met de opmerking, dat de kleinste kwadratenmethode geen gebruik maakt van de sparsity aanname. In onze probleemstelling wordt aangenomen dat er hoogstens S parameters zijn die niet gelijk zijn aan 0. Het is wenselijk dat een methode om de parameters te schatten, gebruik maakt van deze aanname.

Er zijn verschillende manieren om de aanname te gebruiken. Zo kan er een penalty gegeven worden aan het aantal parameters dat niet gelijk is aan 0, de zogeten ℓ_0-'norm'[1] van β. Hoewel het toepassen van een penalty wellicht zou kunnen werken, is het rekentechnisch gezien niet gewenst.

Een andere methode die ook gebruik maakt van de sparsity aanname, is het uitproberen van alle mogelijke selecties van parameters van grootte S. Bij deze methode wordt er bij elke iteratie een selectie van S parameters gekozen. De overige parameters worden op 0 gezet en vervolgens worden de S geselecteerde parameters geschat.

Door alle mogelijke selecties uit te proberen, kan de beste selectie parameters gekozen worden; de schatter voor β is dan de geschatte waarde bij de S geselecteerde parameters en 0 elders. Ter verduidelijking zal de eerste iteratie, de eerste selectie

[1] Merk op dat de ℓ_0-'norm' geen norm is, $\|\lambda v\|_0 \neq |\lambda| \|v\|_0$ voor alle $v \in \mathbb{R}^p$ en $\lambda \in \mathbb{R}$.

parameters, in het volgende voorbeeld uitgewerkt worden.

VOORBEELD

Stel dat $n = 20$, $S = 3$ en $p = 40$. Er wordt dus aangenomen dat er maar 3 niet-nul elementen zitten in β. In de eerste selectie/combinatie wordt aangenomen dat β_1, β_2 en β_3 niet gelijk zijn aan nul, dus $\beta_4 = \ldots = \beta_p = 0$. Er wordt nu gezocht naar de 'beste' β_1, β_2 en β_3 zodanig dat

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & x_{34} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}.$$

Doordat $\beta_4 = \ldots = \beta_p = 0$, betekent dit eigenlijk dat de kolommen 4 tot en met p van X buiten spel staan. De kolommen $X_4, \ldots X_p$ dragen in dit geval niets bij voor het bepalen van y, er wordt immers aangenomen dat *alleen* β_1, β_2 en β_3 niet nul zijn. Merk op dat we in deze situatie de parametervector β en de designmatrix X kunnen inkorten naar

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_n \end{pmatrix}.$$

Dezelfde inzicht kan ook verkregen worden door y op te schrijven als som van de ruis en een lineaire combinatie van de kolommen van X:

$$\begin{aligned} y &= X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + \ldots + X_p\beta_p + \varepsilon \\ &= X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + 0 + \ldots + 0 + \varepsilon \\ &= X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon. \end{aligned}$$

Merk op dat het schatten van β_1, β_2 en β_3 nu gedaan zou kunnen worden met behulp van de kleinste kwadratenmethode.

Bij de tweede iteratie van deze methode zal aangenomen worden dat alleen β_2, β_3 en β_4 niet gelijk zijn aan nul. Vervolgens worden deze geschat met behulp van de kleinste kwadratenmethode. De betere van de twee combinaties zou dan, bijvoorbeeld, de selectie zijn waarvoor $\|y - X\tilde{\beta}\|_{\ell_2}^2$ het kleinste is.

Verrassend genoeg laat het voorbeeld zien, dat het schatten gedaan kan worden met behulp van de kleinste kwadratenmethode. Het vergelijken van verschillende selecties van parameters, het vergelijken van verschillende iteraties, zou gedaan kunnen worden aan de hand van het kleinste kwadratencriterium.

Het testen van alle mogelijke deelverzamelingen $I^* \subseteq J = \{1,2,\ldots,p\}$ met $|I^*| = S$ duurt echter heel lang. Het aantal verzamelingen dat bij deze methode uitgetest zou moeten worden, is gelijk aan $\binom{p}{S}$; het aantal combinaties bij $p = 100$ en $S = 5$ is bijvoorbeeld $\binom{100}{5} = 75.287.520$. Merk op dat $|I^*| \leq S$ in de probleemstelling aangenomen wordt, niet $|I^*| = S$. Het testen van alle mogelijke deelverzamelingen met $|I^*| \leq S$ is nog groter! Ideaal zou zijn als er een methode bestaat die met een degelijke snelheid parameters zou kunnen selecteren en tegelijkertijd β goed zou kunnen bepalen.

Merk op dat er bij het schatten van de coëfficienten gebruik gemaakt wordt van de kleinste kwadratenmethode, wat als gevolg heeft dat alle mogelijke deelverzamelingen met $|I^*| \leq S$ uitgeprobeerd moeten worden. Indien andere methoden worden gebruikt, zoals bijvoorbeeld gebruik te maken van Akaike's Information Criterion of Bayesian Information Criterion, kan het aantal uit te proberen combinaties gelijk zijn aan het aantal mogelijke deelverzamelingen met $|I^*| = S$. Merk op dat dit getal nog steeds zeer groot is.

2.4 Dantzig Selector

De Dantzig Selector (DS) is door E.J. Candès en T. Tao geïntroduceerd als schatter voor β in de situatie zoals beschreven in de probleemstelling. De naam vertelt veel over de schatter: omdat de schatter de oplossing is van een lineair programma en G.B. Dantzig, de vader van het Lineair Programmeren, overleed in de tijd dat deze schatter bedacht werd, hebben de auteurs van [8] de schatter naar hem vernoemd; het woord *Selector* is hieraan toegevoegd omdat de schatter parameters selecteert.

Zoals in paragraaf 1.2.1 is vermeld, is de DS gedefinieerd als

$$\arg\min_{\tilde{\beta}} \|\tilde{\beta}\|_{\ell_1} \qquad \text{odv} \qquad \left\|X^T(y - X\tilde{\beta})\right\|_{\ell_\infty} \leq \lambda_p \sigma. \tag{DS}$$

Het is voorstelbaar dat het niet meteen duidelijk is, hoe de DS werkt en waarom deze zo is gedefinieerd. In deze paragraaf zal hier aandacht aan gegeven worden. Door de definitie (DS) anders op te schrijven staat er

$$\arg\min_{\tilde{\beta}} \sum_{i=1}^{p} |\tilde{\beta}| \qquad \text{odv} \qquad \sup_i \left\{\left|\left(X^T(y - X\tilde{\beta})\right)_i\right|\right\} \leq \lambda_p \sigma. \tag{DS-2}$$

De definitie van de DS zoals gegeven in (DS-2) biedt meer inzicht in de werking hiervan. Bij een gekozen waarde λ_p zijn alle vectoren $\tilde{\beta} \in \mathbb{R}^p$ die voldoen aan de eis $\sup_i \left\{\left|\left(X^T(y - X\tilde{\beta})\right)_i\right|\right\} \leq \lambda_p \sigma$, kandidaat om de Dantzig Selector te worden. Van al deze $\tilde{\beta}$ die hier aan voldoen, wordt die $\tilde{\beta}$ gekozen waarvoor $\sum_{i=1}^{p} |\tilde{\beta}|$ minimaal is. De gekozen $\tilde{\beta}$ is dan de Dantzig Selector.

2.4.1 Een lineair programma Een van de voordelen van de Dantzig Selector is dat deze de oplossing is van het lineair programma

$$f = \min \sum_i u_i \qquad \text{odv}$$
$$-u \le \tilde{\beta} \le u$$
$$-\lambda_p \sigma \mathbf{1} \le X^T (y - X\tilde{\beta}) \le \lambda_p \sigma \mathbf{1}$$

In appendix B.1 wordt een bewijs gegeven dat dit lineair programma inderdaad de DS geeft als oplossing. Dat de Dantzig Selector middels een lineair programma verkregen kan worden, betekent dat deze vrij eenvoudig en met een degelijke snelheid bepaald kan worden.

Kijkende naar de definitie van de DS zoals gegeven in (DS-2), kan gezegd worden dat $\sum |\tilde{\beta}_i|$ de doelfunctie is welke geminimaliseerd moet worden, en de voorwaarde $\sup_i \{|(X^T(y - X\tilde{\beta}))_i|\} \le \lambda_p \sigma$ het toegelaten gebied van alle mogelijke $\tilde{\beta}$ definieert.

2.4.2 Toegelaten gebied: een relaxatie van de normaalvergelijkingen Het toegelaten gebied bestaat uit alle vectoren $\tilde{\beta} \in \mathbb{R}^p$ die voldoen aan

$$\sup_i \{|(X^T(y - X\tilde{\beta}))_i|\} \le \lambda_p \sigma \tag{2.3}$$

voor een gekozen waarde van λ_p. Merk op dat indien (2.3) waar is, er geldt dat

$$|(X^T(y - X\tilde{\beta}))_i| \le \lambda_p \sigma \qquad \text{voor alle } i \in \mathbb{N},\ 1 \le i \le p.$$

Het toegelaten gebied van de methode van de DS kan dus ook gezien worden als alle vectoren $\tilde{\beta}$ waarvoor geldt dat

$$-\lambda_p \sigma \le (X^T(y - X\tilde{\beta}))_i \le \lambda_p \sigma \qquad \text{voor alle } i \in \mathbb{N},\ 1 \le i \le p. \tag{2.4}$$

Deze voorwaarde laat zien wat de overeenkomst is tussen het schatten van β met behulp van de kleinste kwadratenmethode en de DS. Merk op dat de kleinste kwadratenschatter voldoet aan de normaalvergelijkingen (2.2) of in een herschreven vorm

$$X^T(y - X\hat{\beta}) = \mathbf{0}. \tag{2.5}$$

De randvoorwaarden van het toegelaten gebied van de DS zoals in (2.4) gegeven, lijken heel erg op (2.5). Waar bij (2.5) voor alle i geldt dat $(X^T(y - X\tilde{\beta}))_i = 0$, geldt bij (2.4) voor alle i dat $(X^T(y - X\tilde{\beta}))_i$ niet gelijk aan 0 hoeft te zijn, maar wel dicht bij 0 moet zitten: $-\lambda_p \sigma \le (X^T(y - X\tilde{\beta}))_i \le \lambda_p \sigma$.

De randvoorwaarden van het toegelaten gebied van de methode van de DS zijn dus een relaxatie van de normaalvergelijkingen; tot op een vastgesteld niveau wordt er aan de normaalvergelijkingen voldaan.

De DS voldoet aan de normaalvergelijkingen tot op een vastgesteld niveau, maar wat betekent dit? Herschrijf (2.4) als

$$-\lambda_p\sigma \le \left\langle X_i, y - X\tilde{\beta}\right\rangle \le \lambda_p\sigma \qquad \text{voor alle } i \in \mathbb{N},\ 1 \le i \le p. \tag{2.6}$$

Bij de normaalvergelijkingen is $\lambda_p\sigma = 0$ en staan de residuen $y - X\tilde{\beta}$ loodrecht op alle kolommen van X. Bij de DS is $\lambda_p\sigma$ vaak niet gelijk aan 0, maar wel klein. Meetkundig gezien is $\lambda_p\sigma$ de maximale lengte van de projectie van de residuen $y - X\tilde{\beta}$ op de kolommen van X. Als $\lambda_p\sigma$ klein is, dan staan de residuen weliswaar niet loodrecht op de kolommen van X, maar wel 'bijna'; de kolommen van X correleren dan weinig met de residuen.

Samenvattend bestaat het toegelaten gebied van de DS dus uit de vectoren $\tilde{\beta} \in \mathbb{R}^p$ waarvoor geldt dat ze tot op een vastgesteld niveau aan de normaalvergelijkingen voldoen. Dit betekent dat de residuen $y - X\tilde{\beta}$ 'bijna' loodrecht staan op de kolomruimte van X en dus dat de kolommen van X weinig correleren met de residuen $y - X\tilde{\beta}$.

2.4.3 Toegelaten gebied: invariantie bij orthogonale transformatie In [8, 9] en [1] wordt ook een ander argument gegeven waarom $X^T\left(y - X\tilde{\beta}\right)$ begrensd wordt in plaats van bijvoorbeeld $y - X\tilde{\beta}$.

Stel dat er een orthogonale transformatie U op de metingen wordt losgelaten. Laat $y^\circ = Uy$, $X^\circ = UX$ en $\varepsilon^\circ = U\varepsilon$. Dit geeft

$$Uy = UX\beta + U\varepsilon; \tag{2.7}$$
$$y^\circ = X^\circ\beta + \varepsilon^\circ. \tag{2.8}$$

Omdat U een orthogonale matrix is, geldt $U^T\sigma^2 I_n U = \sigma^2 I_n$. Dit geeft $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$, waardoor (2.8) voldoet aan het modelvoorschrift van de probleemstelling. β kan dus ook geschat worden bij (2.8), dit kan ook gedaan worden door de Dantzig Selector uit te rekenen. Dit roept de vraag op, of de DS die berekend is met de originele y en X, verschilt met de DS die verkregen is via getransformeerde de metingen y° en designmatrix X°.

Merk op dat als het toegelaten gebied van beide problemen hetzelfde is, de beide schatters gelijk aan elkaar zijn; de doelfunctie blijft immers ongewijzigd. Beschouw $(X^\circ)^T\left(y^\circ - X^\circ\tilde{\beta}\right)$. Het toepassen van de definitie van y°, X° en ε° en het feit dat U een orthogonale matrix is, geeft

$$\begin{aligned}\left(X^\circ\right)^T\left(y^\circ - X^\circ\tilde{\beta}\right) &= (UX)^T\left(Uy - UX\tilde{\beta}\right)\\ &= X^T U^T U\left(y - X\tilde{\beta}\right)\\ &= X^T I_n\left(y - X\tilde{\beta}\right)\\ &= X^T\left(y - X\tilde{\beta}\right).\end{aligned} \tag{2.9}$$

Vergelijking (2.9) geeft

$$\left\|\left(X^\circ\right)^T\left(y^\circ - X^\circ\tilde{\beta}\right)\right\|_{\ell_\infty} = \left\|X^T\left(y - X\tilde{\beta}\right)\right\|_{\ell_\infty},$$

waaruit volgt dat de toegelaten gebieden hetzelfde zijn. Dit betekent, dat beide schatters aan elkaar gelijk zijn. De Dantzig Selector wijzigt dus niet onder een orthogonale transformatie, het optimalisatieprobleem wijzigt niet. De metingen kunnen dus bijvoorbeeld gespiegeld of geroteerd worden.

2.4.4 Doelfunctie Van alle mogelijke vectoren $\tilde{\beta}$ uit het toegelaten gebied (2.3) is de Dantzig Selector $\hat{\beta}$ die $\tilde{\beta}$ waarvoor geldt dat $\sum|\tilde{\beta}_i|$ minimaal is. Van alle mogelijke vectoren die tot op een vastgesteld niveau aan de normaalvergelijkingen voldoen, is de Dantzig Selector die vector met de kleinste ℓ_1 norm.

Merk op dat de aanname $|I| \leq S$ niet direct door de DS wordt gebruikt. Wel wordt er geminimaliseerd op de ℓ_1 norm. Het minimaliseren op $\sum|\tilde{\beta}_i|$, zorgt er schijnbaar voor, dat de DS sparse is. Een verklaring hiervoor is te vinden in een deelresultaat van stelling 9, ongelijkheid (C.4); een klein gedeelte van deze uitleg volgt nu, maar is ook te vinden in hoofdstuk 4. Stel dat er een zeer sparse vector $\tilde{\beta}$ in het toegelaten gebied van de DS ligt met $\tilde{I}$ als zijn support. Dan geldt

$$\sum_{i\in\tilde{I}^c}|\hat{\beta}_i| \leq \sum_{i\in\tilde{I}}|\hat{\beta}_i - \tilde{\beta}_i|.$$

Als $\tilde{\beta}$ zeer sparse is, dan geldt dat $\tilde{I}$ weinig elementen heeft en $\tilde{I}^c$ veel; gemiddeld genomen moet $\hat{\beta}_i$ dan klein zijn voor $i \in \tilde{I}^c$. Wanneer $\hat{\beta}$ dicht bij $\tilde{\beta}$ ligt, wordt $\sum_{i\in\tilde{I}}|\hat{\beta}_i - \tilde{\beta}_i|$ ook klein, wat ervoor zorgt dat $\hat{\beta}_i$ zeer klein moet zijn voor veel $i \in \tilde{I}^c$.

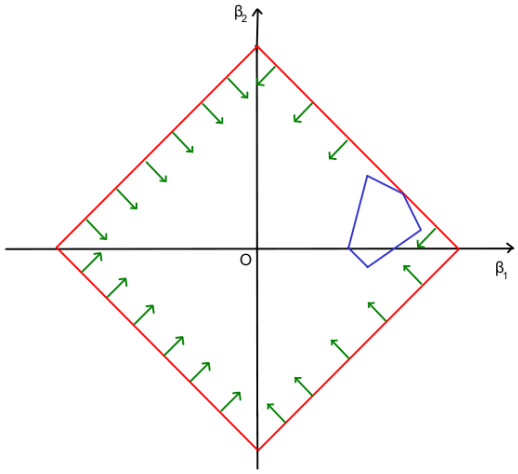
De doelfunctie van de DS is een convexe functie: Zij $\tilde{\beta}_1 \in \mathbb{R}^p$, $\tilde{\beta}_2 \in \mathbb{R}^p$ willekeurig gegeven, dan geldt voor elke $t \in [0,1]$ dat $\|t\tilde{\beta}_1 + (1-t)\tilde{\beta}_2\|_{\ell_1} \leq \|t\tilde{\beta}_1\|_{\ell_1} + \|(1-t)\tilde{\beta}_2\|_{\ell_1} = |t|\|\tilde{\beta}_1\|_{\ell_1} + |(1-t)|\|\tilde{\beta}_2\|_{\ell_1} = t\|\tilde{\beta}_1\|_{\ell_1} + (1-t)\|\tilde{\beta}_2\|_{\ell_1}$. Vanwege de driehoeksongelijkheid geldt per definitie dat een norm een convexe functie is.

Het toegelaten gebied van de Dantzig Selector is een convex gebied. Zij $\tilde{\beta}_1 \in \mathbb{R}^p$, $\tilde{\beta}_2 \in \mathbb{R}^p$ en $t \in [0,1]$ willekeurig gegeven. Dan geldt

$$\begin{aligned}
\left\|X^T(y - X(t\tilde{\beta}_1 + (1-t)\tilde{\beta}_2))\right\|_{\ell_\infty} &= \left\|X^T(y - Xt\tilde{\beta}_1) + X^T(y - X(1-t)\tilde{\beta}_2)\right\|_{\ell_\infty} \\
&= \left\|tX^T(y - X\tilde{\beta}_1) + (1-t)X^T(y - X\tilde{\beta}_2)\right\|_{\ell_\infty} \\
&\leq |t|\left\|X^T(y - X\tilde{\beta}_1)\right\|_{\ell_\infty} + |1-t|\left\|X^T(y - X\tilde{\beta}_2)\right\|_{\ell_\infty} \\
&\leq t\left\|X^T(y - X\tilde{\beta}_1)\right\|_{\ell_\infty} + (1-t)\left\|X^T(y - X\tilde{\beta}_2)\right\|_{\ell_\infty}.
\end{aligned}$$

Het toegelaten gebied van de DS is dus een convex gebied. De DS is dus een oplossing is van een convex lineair programma. Van een convex lineair programma weten we dat de oplossing een randpunt is van het toegelaten gebied. Dit betekent dat wanneer (DS) een unieke oplossing heeft, er geldt dat de oplossing een randpunt is van het toegelaten gebied.

Een visualisatie van de doelfunctie en een convex gebied in $\mathbb{R}^2$ is te zien in figuur 2.1. Het is duidelijk te zien dat de DS een hoekpunt is van het toegelaten gebied. Wanneer twee of meer hoekpunten de ℓ_1-norm minimaliseren, dan zijn er oneindig veel oplossingen.



Figuur 2.1: Een voorstelling van de doelfunctie min $\|\tilde{\beta}\|_{\ell_1}$ (rood) en een convex gebied (blauw) in $\mathbb{R}^2$

2.4.5 Keuze λ_p Nu is er nog niets gezegd over de keuze van λ_p. De waarde van λ_p geeft aan hoe groot het toegelaten gebied van de Dantzig Selector is. In een subparagraaf is al aangegeven, dat de grootte van $\lambda_p\sigma$ controleert wat de maximale lengte is van de projectie van de residuen op de kolommen van X.

In [18] wordt opgemerkt[2] dat $\mathbf{0}$ in het toegelaten gebied ligt als $\|X^T y\|_{\ell_\infty} \leq \lambda_p\sigma$, dit is dan meteen Dantzig Selector. Het probleem is om λ_p niet al te groot te nemen, maar wel groot genoeg zodat β in het toegelaten gebied van de DS zit.

In [8] wordt $\lambda_p = \sqrt{2\log(p)}$ gebruikt in het bewijs van de efficiëntie van de schatter; in hetzelfde artikel wordt echter ook genoemd, dat deze waarde iets te groot is. Het verkrijgen van een betere λ_p kan gedaan worden via Monte Carlo-simulatie; hierbij wordt de ruisvector telkens gesimuleerd zodat de waarde van de optimale λ_p geschat kan worden. Om meer inzicht te verkrijgen in de keuze van λ_p, is er in het bachelorproject een simulatie uitgevoerd (hoofdstuk 5), waarbij verschillende waarden van λ_p worden uitgeprobeerd.

Een andere methode om λ_p te bepalen, is via V-fold cross validation [1]. Deze methode is in dit bachelorproject niet bekeken en wordt daarom niet in dit verslag behandeld.

[2] Hoewel de opmerking ging over de duale vorm van de LASSO, is deze ook hier van toepassing.

2.5 LASSO

De Least Absolute Shrinkage and Selection Operator (LASSO) is in [20] door Tibshirani geïntroduceerd als een schattingsmethode die coëfficienten kleiner maakt ('shrinkage') door gebruik te maken van de ℓ_1-norm ('absolute'); sommige coëfficienten worden zelfs op 0 gezet ('selection operator').

Zoals in paragraaf 1.2.1 is aangegeven, is de LASSO gedefinieerd als

$$\arg\min_{\tilde{\beta}} \left\| y - X\tilde{\beta} \right\|_{\ell_2}^2 \qquad \text{odv} \qquad \left\| \tilde{\beta} \right\|_{\ell_1} \leq s, \tag{LASSO}$$

waarbij de waarde van s zelf in te stellen is. Het verband tussen de kleinste kwadratenmethode en de LASSO is vrij duidelijk. Zowel de LASSO als de kleinste kwadratenschatter minimaliseren $\|y - X\beta\|_{\ell_2}^2$, de LASSO beperkt ook de ℓ_1-norm van $\tilde{\beta}$. Dit laatste zorgt ervoor, dat de LASSO sparse oplossingen geeft bij geschikte waarden van s.

In [12] en [21] is er gekeken naar het ruisloze probleem. In beide artikelen wordt gesteld, dat indien β sparse genoeg is, de oplossing van het programma $\min \left\| \tilde{\beta} \right\|_{\ell_1}$ odv $y - X\tilde{\beta} = \mathbf{0}$ ook de oplossing van het probleem $\min \left\| \tilde{\beta} \right\|_{\ell_0}$ odv $y - X\tilde{\beta} = \mathbf{0}$ is.

2.5.1 Basis Pursuit denoising

Een equivalent probleem is

$$\arg\min_{\tilde{\beta}} \left\| y - X\tilde{\beta} \right\|_{\ell_2}^2 + \lambda \left\| \tilde{\beta} \right\|_{\ell_1}, \tag{BPDN}$$

dat beter bekend staat als het Basis Pursuit DeNoising (BPDN) [11]. [18] geeft aan de oplossingen van beide problemen gelijk aan elkaar zijn: voor een gegeven $\lambda \geq 0$ bestaat er een $s \geq 0$ zodanig dat de oplossingen van beide programma's hetzelfde zijn (andersom ook). Één richting van de equivalentie is bewezen in appendix B.2.

2.5.2 Connecties met de Dantzig Selector

Lemma 3.1 uit [3] bevat een interessant resultaat, het bewijst dat de LASSO schatter voldoet aan $\|X^T(y - X\tilde{\beta})\|_{\ell_\infty} \leq \lambda_p \sigma$. In dit artikel wordt de LASSO schatter gegeven in de BPDN vorm met $\lambda = \lambda_p \sigma$. Het lemma laat zien dat de LASSO schatter, bij gelijke λ_p, in het toegelaten gebied van de DS ligt.

In [18] wordt aangetoond, dat de LASSO ook een oplossing is van het alternatieve probleem

$$\arg\min_{\tilde{\beta}} \frac{1}{2} \left\| X\tilde{\beta} \right\|_{\ell_2}^2 \text{odv} \left\| X^T \left(y - X\tilde{\beta} \right) \right\|_{\ell_\infty} \leq \lambda_p \sigma. \tag{2.10}$$

Deze vorm lijkt heel erg op de DS, de doelfunctie is alleen anders. Bij de LASSO wordt er dus geminimaliseerd op de kwadratische vorm $\tilde{\beta}^T X^T X \tilde{\beta}$ terwijl er bij de DS geminimaliseerd op de ℓ_1-norm van β. De equivalentie van de LASSO en zijn alternatieve vorm is niet in dit verslag bewezen.

3

Onderliggende theorie

Het artikel [8] waarin de Dantzig Selector wordt geïntroduceerd, is een vervolg op een reeks artikelen [2, 4, 6]. In deze reeks artikelen staat centraal wanneer — of wanneer juist niet — een hogerdimensionale vector met weinig metingen gereconstrueerd kan worden. Op toepassingsniveau kan zo'n hogerdimensionale vector worden gezien als een signaal of een afbeelding.

De aanname dat de parametervector zeer sparse is, is niet voldoende om te verzekeren dat deze geschat kan worden. Door enkele begrippen te introduceren, kan een voorwaarde aan de designmatrix gesteld worden, waardoor het bepalen van β in het ruisloze geval zeker mogelijk is.

3.1 Unieke sparse representatie

3.1.1 Waarom de designmatrix aan een voorwaarde moet voldoen. Doordat het aantal parameters p groter is dan het aantal metingen n, vormen de kolommen van X *geen* lineair onafhankelijk stelsel. Dit gegeven geeft problemen voor het schatten van β. Door te kijken naar de ruisloze situatie zal er in deze paragraaf aangegeven worden waarom het nodig is, dat er een voorwaarde aan de designmatrix gesteld wordt.

In paragraaf 2.1.2 wordt via (2.1) opgemerkt dat y een lineaire combinatie is van de ruis ε en de kolommen van X. Aangezien er in deze paragraaf gekeken wordt naar het ruisloze geval, is de ruis gelijk aan $\mathbf{0}$. Dit geeft

$$y = X_1\beta_1 + \ldots X_p\beta_p. \tag{3.1}$$

Omdat $\{X_1, \ldots, X_p\}$ geen lineair onafhankelijk stelsel vormen, zijn er oneindig veel lineaire combinaties die y geven. Bij het schatten van β zijn alleen de metingen y gegeven. Gegeven de metingen y moet één van de oneindig veel lineaire combinaties gekozen worden als β.

Gelukkig is er de sparsity aanname: omdat $|I| \leq S$, vallen alle lineaire combinaties af met meer dan S niet-nullen. Als er dan maar één lineaire combinatie overblijft, dan is de gezochte lineaire combinatie β gevonden.

Als er, na het gebruik maken van de sparsity aanname, meerdere lineaire combinaties overblijven, dan is het niet mogelijk om β te schatten; er zijn dan immers meerdere sparse lineaire combinaties die mogelijk de 'juiste' kunnen zijn. Hoewel de echte β hoort bij een van de overgebleven lineaire combinaties, zijn de andere lineaire combinaties even goede kandidaten. In dit geval zeggen we, dat er meerdere sparse representaties zijn voor y.

Dit hoofdstuk introduceert enkele begrippen waarmee een voorwaarde aan de designmatrix X gesteld kan worden, zodat er bij elke y maar één lineaire combinatie 'overblijft'. De voorwaarde dient als een verzekering, dat er een unieke sparse representatie is voor y.

3.2 Restricted isometries

Het zoeken naar goede designmatrices X heeft gezorgd voor de introductie van een restricted isometrie aanname in [6]. Het idee hiervan is dat een willekeurig "sparse" stelsel kolommen van X zich zoveel mogelijk gedraagt als een orthonormaal stelsel. Dit moet voorkomen, dat er meerdere sparse representaties zijn voor een gegeven vector y.

3.2.1 Restricted isometrie constante

Càndes en Tao introduceren in [6] een 'maat' voor matrices X, de restricted isometrie constante (RIC). De naam 'restricted isometrie constante' verraadt al veel over dit getal. Een isometrie is een normbehoudende afbeelding. Het woord 'restricted' geeft aan, dat er gekeken wordt naar sparse lineaire combinaties; er wordt gekeken naar een willekeurige verzameling van maximaal S kolommen van X.

Merk op dat een orthogonale afbeelding normbehoud heeft. De RIC kwantificeert hoe dicht een willekeurige verzameling van maximaal S kolommen van X zich gedraagt als een orthonormaal stelsel.

Definitie 1 (Restricted isometrie constante (RIC))**.** Laat $S \in \mathbb{N}$, $1 \leq S \leq |J| = p$. De restricted isometrie constante is het kleinste getal $\delta_S(X)$ zodanig dat voor alle deelverzamelingen $A \subset J$ met $|A| \leq S$ geldt, dat

$$(1-\delta_S(X))\,\|c\|_{\ell_2}^2 \leq \|X_A c\|_{\ell_2}^2 \leq (1+\delta_S(X))\,\|c\|_{\ell_2}^2$$

voor alle vectoren $c \in \mathbb{R}^{|A|}$.

Op het eerste oog lijkt het vrij lastig om, gegeven een S, de restricted isometrie constante $\delta_S(X)$ te bepalen: alle deelverzamelingen $|A| \leq S$ en alle vectoren $c \in \mathbb{R}^{|A|}$ moeten getest worden. Door gebruik te maken van een eigenschap van het Rayleigh

quotiënt, hoeven alle vectoren $c \in \mathbb{R}^{|A|}$ echter niet uitgetest te worden om de RIC te bepalen.

Stelling 2 (Rayleigh quotiënt)**.** *Zij X een $n \times p$ matrix en $A \subset J$ met $|A| \leq S$. Laat $\lambda_{\min} \leq \ldots \leq \lambda_{\max}$ de eigenwaarden van $X_A^T X_A$ zijn. Dan geldt*

$$\lambda_{\min} \|c\|^2 \leq \|X_A c\|^2 \leq \lambda_{\max} \|c\|^2.$$

Bewijs. Zie appendix B.3. □

Stelling 2 kan gebruikt worden om het bepalen van de RIC eenvoudiger te maken. Kijkende naar de definitie van de RIC geldt voor elke verzameling A, $|A| \leq S$ en elke vector $c \in \mathbb{R}^{|A|}$ dat

$$(1-\delta_S(X))\|c\|^2 \leq \lambda_{\min}\|c\|^2 \leq \|X_A c\|^2 \leq \lambda_{\max}\|c\|^2 \leq (1+\delta_S(X))\|c\|^2.$$

Voor alle deelverzamelingen $A \subset J$ met $|A| \leq S$ moet gelden, dat $1-\lambda_{\min}(X_A) \leq \delta_S(X)$ en $\lambda_{\max}(X_A) - 1 \leq \delta_S(X)$. Dit geeft

$$\max\{1-\lambda_{\min}(X_A), \lambda_{\max}(X_A)-1\} \leq \delta_S(X). \tag{3.2}$$

Ongelijkheid (3.2) geeft een ondergrens voor $\delta_S(X)$ voor elke deelverzameling $A \subset J$ met $|A| \leq S$. Door het maximum van al deze ondergrenzen te nemen, verkrijgen we de waarde van $\delta_S(X)$. Merk op, dat het bepalen van de RIC nog steeds veel rekentijd kost, van alle matrices $X_A^T X_A$ moeten de eigenwaarden uitgerekend worden.

VOORBEELD

In dit voorbeeld wordt de waarde van $\delta_S(X)$ voor $S = 2$ uitgerekend bij een matrix met genormaliseerde kolommen. Laat

$$Z = \begin{pmatrix} 60 & 68 & 8 \\ 26 & 74 & 22 \end{pmatrix}$$

en normaliseer de kolommen (op de ℓ_2-norm met norm 1), laat X deze matrix zijn. X is dus de matrix

$$X = \begin{pmatrix} 0.9176 & 0.6766 & 0.3417 \\ 0.3976 & 0.7363 & 0.9398 \end{pmatrix}.$$

Voor $S = 1$ geldt $c \in \mathbb{R}$ en voor alle kolommen X_i geldt $\|X_i\|_{\ell_2}^2 = 1$. $\delta_1(X)$ is het kleinste getal zodanig dat $(1-\delta_1(X))c^2 \leq \|X_i c\|_{\ell_2}^2 \leq (1+\delta_1(X))c^2$ voor alle $i \in \mathbb{N}$, $1 \leq i \leq p$ en voor alle $c \in \mathbb{R}$. Omdat $\|X_i c\|_{\ell_2}^2 = \|X_i\|_{\ell_2}^2 c^2 = c^2$ geeft dit $\delta_1(X) = 0$.

Het bepalen van $\delta_2(X)$ wordt gedaan door de eigenwaarden uit te rekenen van alle submatrices X_A, bestaande uit deelverzamelingen $A \subseteq J = \{1,2,3\}$ met $|A| \leq 2$. Laat $A_1 = \{1,2\}$, $A_2 = \{2,3\}$ en $A_3 = \{1,3\}$.

De matrix $X_{A_1}^T X_{A_1} = \begin{pmatrix} 1.0000 & 0.9136 \\ 0.9136 & 1.0000 \end{pmatrix}$ heeft eigenwaarden 0.0864 en 1.9136, $X_{A_2}^T X_{A_2}$ heeft eigenwaarden 0.0768 en 1.9232 en $X_{A_3}^T X_{A_3}$ heeft eigenwaarden 0.3128 en 1.6872. De grootste gevonden eigenwaarde is 1.9232 en de kleinste is 0.0768. Omdat $1 - 0.0768 = 0.9232$, $1.9232 - 1 = 0.9232$ en $\delta_1(X) = 0$, verkrijgen we $\delta_2(X) = 0.9232$.

De kolommen van X zijn genormaliseerd op de ℓ_2-norm, dan en slechts dan als $\delta_1(X) = 0$. Het vinden van matrices met kleine $\delta_S(X)$ waarden gaat dus makkelijker, wanneer de kolommen van X zijn genormaliseerd.

Merk op dat voor een orthogonale matrix geldt dat $\|X_T c\|_{\ell_2} = \|c\|_{\ell_2}$. De RIC laat dus zien hoe "orthogonaal" een sparse combinatie van kolommen van X zich gedraagt. In de probleemstelling van het bachelorproject bestaan matrices met $\delta_S(X) = 0$ niet. Voor matrices met $\delta_S(X) = 0$ zou in ieder geval moeten gelden, dat de kolommen van X lineair onafhankelijk zijn; in de situatie $n < p$ geldt dit niet. We zoeken dus wel naar matrices met een zo klein mogelijke $\delta_S(X)$.

3.2.2 Restricted orthogonaliteitsconstante Naast de restricted isometrie constante is ook de restricted orthogonaliteitsconstante in [6] geïntroduceerd. Het getal is een bovengrens voor de projectie van elk paar vectoren uit alle kolomruimten van X met een grootte van maximaal S en S'.

Definitie 3 (Restricted orthogonaliteitsconstante)**.** Laat $S, S' \in \mathbb{N}$ met $S + S' \leq |J|$. De restricted orthogonaliteit constante is het kleinste getal $\theta_{S,S'}(X)$ zodanig dat

$$|\langle X_A c, X_{A'} c' \rangle| \leq \theta_{S,S'}(X) \|c\|_{\ell_2} \|c'\|_{\ell_2}$$

voor alle disjuncte verzamelingen $A, A' \subseteq J$ met $|A| \leq S$ en $|A'| \leq S'$ en alle vectoren $c \in \mathbb{R}^{|A|}$ en $c' \in \mathbb{R}^{|A'|}$.

Merk op, dat

$$\left| \left\langle X_A \frac{c}{\|c\|_{\ell_2}}, X_{A'} \frac{c'}{\|c'\|_{\ell_2}} \right\rangle \right| = \left| \frac{\langle X_A c, X_{A'} c' \rangle}{\|c\|_{\ell_2} \|c'\|_{\ell_2}} \right| = \frac{|\langle X_A c, X_{A'} c' \rangle|}{\|c\|_{\ell_2} \|c'\|_{\ell_2}}. \tag{3.3}$$

Voor het bepalen van $\theta_{S,S'}(X)$ hoeft dus alleen gekeken te worden naar alle eenheidsvectoren u en u'. De restricted orthogonaliteitsconstante is dus het kleinste getal $\theta_{S,S'}(X)$ zodanig dat

$$|\langle X_A u, X_{A'} u' \rangle| \leq \theta_{S,S'}(X) \tag{3.4}$$

geldt voor alle disjuncte verzamelingen $A, A' \subseteq J$ met $|A| \leq S$ en $|A'| \leq S'$ en alle vectoren $u \in \mathbb{R}^{|A|}$ en $u' \in \mathbb{R}^{|A'|}$ met $\|u\|_{\ell_2} = \|u'\|_{\ell_2} = 1$.

(3.3) en (3.4) laten zien dat er gekeken wordt naar een hoek, $\theta_{S,S'}(X)$ geeft het grootst mogelijke inproduct dat twee eenheidsvectoren van twee verschillende kolomruimten A en A' kunnen maken. Twee willekeurige eenheidsvectoren van twee verschillende kolomruimten A en A' staan bijna loodrecht op elkaar wanneer $\theta_{S,S'}(X)$ dicht bij 0 is, $\theta_{S,S'}(X)$ controleert de minimale hoek die de vectoren maken.

Voor een designmatrix zijn kleine waarden van $\theta_{S,S'}(X)$ gewenst. Kleine waarden zorgen er voor dat de minimale hoek tussen twee vectoren van verschillende kolomruimten groot is. Hierdoor wordt het minder goed mogelijk gemaakt, dat er verschillende sparse representaties voor y zijn.

Het is geen toeval dat de restricted isometrie constante en de restricted orthogonaliteitsconstante in naam op elkaar lijken en achter elkaar gedefinieerd worden; er bestaan connecties tussen de RIC en de ROC. In [6] wordt de onderstaande stelling bewezen. De uitwerking van het bewijs van de onderstaande stelling is in appendix B.4 te vinden.

Stelling 4 (Restricted isometrie constante – restricted orthogonaliteitsconstante)**.** *Voor alle $S, S' \in \mathbb{N}$ met $S + S' \leq p$ geldt $\theta_{S,S'}(X) \leq \delta_{S+S'}(X)$.*

Bewijs. Zie appendix B.4. □

3.3 Unieke sparse representatie (2)

Door gebruik te maken van de RIC kan een voorwaarde aan de designmatrix X gesteld worden, zodat een unieke sparse representatie voor y gegarandeerd kan worden.

Stelling 5 (Unieke sparse representatie)**.** *Zij X een $n \times p$ designmatrix en $S \geq 1$ zodanig dat $\delta_{2S}(X) < 1$. Zij $I \subset J$ met $|I| \leq S$ en $\beta \in \mathbb{R}^p$. Laat $y = X_I \beta_I$. Als y en X gegeven zijn, dan geldt dat I en β te bepalen zijn en dat deze uniek zijn.*

Bewijs. Laat I de support van β zijn, dan $|I| \leq S$. Laat $c = \beta \in \mathbb{R}^p$. Merk op dat $c_I = \beta_I$, $c_i \neq 0$ voor alle $i \in I$ en $c_i = 0$ voor alle $i \notin I$. Merk op dat er dus zeker één sparse representatie bestaat, de support I met bijbehorende vector β_I. Er wordt aangenomen dat er twee verschillende sparse representaties I en I' voor y zijn. Vervolgens komen we op een tegenspraak uit.

Neem aan dat er een verzameling $I' \subsetneq I$, $|I'| \leq S$, bestaat zodanig dat er een vector $c' \in \mathbb{R}^p$ bestaat met $y = Xc = X_I c_I = X_{I'} c'_{I'} = Xc'$ met $c'_{i'} \neq 0$ voor alle $i' \in I'$ en $c'_{i'} = 0$ voor alle $i' \notin I'$. Dan bestaat er een element $a \in I$ met $a \notin I'$. Omdat $a \in I$, geldt $a \neq 0$. Maar $a \notin I'$, dus moet $a = 0$. Tegenspraak! Zo een verzameling I' bestaat niet.

Neem aan dat er een verzameling $I' \nsubseteq I$, $|I'| \leq S$, bestaat zodanig dat er een vector $c' \in \mathbb{R}^p$ bestaat met $y = Xc = X_I c_I = X_{I'} c'_{I'} = Xc'$ met $c'_{i'} \neq 0$ voor alle $i' \in I'$ en $c'_{i'} = 0$ voor alle $i' \notin I'$. Er wordt dus aangenomen dat er twee verschillende sparse representaties zijn voor y. Merk op dat

$$Xc - Xc' = X_I c_I - X_{I'} c'_{I'} = \mathbf{0}.$$

Bekijk de vector $d = c - c' \in \mathbb{R}^p$. De vector d is dus zo gedefinieerd, dat $X(c - c') = Xd = X_{I \cup I'} d_{I \cup I'} = \mathbf{0}$ en $d_i = 0$ voor alle $i \notin I \cup I'$. Omdat $\|X_{I \cup I'} d_{I \cup I'}\|_{\ell_2} = \|\mathbf{0}\|_{\ell_2} = 0$ en aangezien $d_{I \cup I'} \in \mathbb{R}^{|I \cup I'|}$ en $|I \cup I'| \leq 2S$, geldt

$$\begin{aligned} (1 - \delta_{2S}(X)) \|d_{I \cup I'}\|_{\ell_2}^2 \leq \|X_{I \cup I'} d_{I \cup I'}\|_{\ell_2}^2 \leq (1 + \delta_{2S}(X)) \|d_{I \cup I'}\|_{\ell_2}^2, \\ (1 - \delta_{2S}(X)) \|d_{I \cup I'}\|_{\ell_2}^2 \leq 0 \leq (1 + \delta_{2S}(X)) \|d\|_{\ell_2}^2. \end{aligned} \tag{3.5}$$

Het gegeven $\delta_{2S}(X) < 1$ geeft $0 < (1 - \delta_{2S}(X))$. Dit gegeven tezamen met (3.5) kan enkel waar zijn, indien $d_{I \cup I'} = \mathbf{0}$. Er is echter aangenomen, dat $I' \nsubseteq I$; er bestaat een

$a \in I'$ met $a \notin I$. Aangezien $d_{I \cup I'} = \mathbf{0}$, betekent dit dat $d_a = c_a - c'_a = -c'_a = 0$. Dit is in tegenspraak met de aanname, dat $c'_{i'} \neq 0$ voor alle $i' \in I'$. Deze tegenspraak geeft ons, dat er geen verzameling I' bestaat zodanig dat $|I'| \leq S$, $I' \not\subseteq I$, $y = X_I c = X_{I'} c'_{I'} = X c'$ en $c'_{i'} \neq 0$ voor alle $i' \in I'$ en $c'_{i'} = 0$ voor alle $i' \notin I'$.

Omdat $\delta_S(X) \leq \delta_{2S}(X) < 1$ weten we, dat de kolommen van X_I lineair onafhankelijk zijn. Dit sluit de mogelijkheid van twee verschillende sparse representaties $I = I'$ uit met verschillende vectoren c en c'. De kern van X_I is dan gelijk aan $\{\mathbf{0}\}$, en de coëfficiënten c_i kunnen dan bepaald worden door $y = X_I c_I$ op te lossen.

□

Door van een designmatrix X te eisen dat $\delta_{2S}(X) < 1$, kan een garantie voor een unieke sparse representatie gegeven worden. In het ruisloze geval kunnen de coefficiënten dan meteen bepaald worden. Wanneer er ruis zit in de metingen, kan een iets sterkere aanname zoals $\delta_{2S}(X) + \theta_{S,2S}(X) < 1$ gewenst zijn. De aanname $\delta_{2S}(X) + \theta_{S,2S}(X) < 1$ wordt bijvoorbeeld in het volgende hoofdstuk gebruikt in een stelling waarbij een kwaliteitsgarantie voor de DS gegeven wordt.

3.4 Coherence property

In [3] staan enkele interessante bewijzen die niet gebruik maken van de restricted isometrie eigenschap. Bij deze bewijzen wordt er een *coherence property* aangenomen. Deze eigenschap zal verder niet gebruikt worden in dit verslag, het wordt hier alleen vermeld.

Laat X een $n \times p$ matrix zijn waarvan de kolommen zijn genormaliseerd op de ℓ_2-norm met norm 1. Definieer $\mu(X)$ als

$$\mu(X) = \sup_{1 \leq i < j \leq p} |\langle X_i, X_j \rangle|.$$

Definitie 6 (Coherence property)**.** Een matrix X voldoet aan de 'coherence property' als voor een vooraf opgegeven constante $A_0 \geq 0$ geldt dat

$$\mu(X) \leq \frac{A_0}{\log(p)}.$$

De coherence property kan veel eenvoudiger gecontroleerd worden dan dat de restricted isometrie constante berekend kan worden. Merk op dat de matrix $X^T X$ bestaat uit de inproducten van de kolommen van X. Het element uit $X^T X$ met de grootste absolute waarde is $\mu(X)$. Wanneer $\mu(X)$ berekend is, kan gecontroleerd worden of de matrix voldoet aan de coherence property.

4

Theoretische eigenschappen

In dit hoofdstuk wordt er ingegaan op de kwaliteit van de DS. Na een theoretische beschouwing van de kwaliteit van een schatter zullen enkele bewijzen over de kwaliteit van bepaalde schatters behandeld worden. Via de oracle inequalities zal de DS geopperd worden als methode van variabele selectie, de zogeten Gauss-DS methoed; het schatten van de parameters via deze methode kan mogelijk efficiënter zijn dan het gebruiken van de DS als schatter.

4.1 De kwaliteit van een schatter

Wanneer de kwaliteit van verschillende schatters voor β wordt bekeken, dan moet er een maat zijn om deze kwaliteit aan te geven. Bij het kiezen van een juiste maat is het belangrijk, dat het einddoel duidelijk zichtbaar is.

Het artikel waar de Dantzig Selector in is geïntroduceerd, gebruikt $\|\hat{\beta}-\beta\|_{\ell_2}^2$ om de kwaliteit van de schatter te beoordelen. Ritov geeft in [19] een reactie op het gebruik van $\|\hat{\beta}-\beta\|_{\ell_2}^2$ om de kwaliteit van een schatter te meten: de prediction error zou zich hier beter voor lenen. In dit artikel wordt opgemerkt, dat het gebruik van een ℓ_2 verlies functie in lagerdimensionale gevallen nog logisch is, maar in de hogerdimensionale gevallen niet.

In [9] wordt een reactie gegeven op [19]. Er wordt gezegd dat de prediction error van belang is wanneer men geïnteresseerd is in het zo goed mogelijk voorspellen (bepaal $X\beta$ zo goed mogelijk). Als voorbeeld worden enkele toepassingen genoemd waarvoor de DS ontworpen was en waar $\|\hat{\beta}-\beta\|_{\ell_2}^2$ een juiste maat is om de kwaliteit aan te geven. Zo kan men bij afbeeldingen (MRI) geïnteresseerd zijn in de gemiddelde fout bij elke pixel.

Hoewel voor beide reacties iets te zeggen is, zal $\|\hat{\beta}-\beta\|_{\ell_2}^2$ gebruikt worden als maat voor de kwaliteit van de schatter. Toch moet benadrukt worden dat Ritov enkele belangrijke punten heeft genoemd. Wanneer p groot is en de data bestaat uit verschillende grootheden (lengte, gewicht, inkomen, ...), lijkt $\|\hat{\beta}-\beta\|_{\ell_2}^2$ weinig voor te stellen.

De vraag is echter of $\|X(\hat{\beta}-\beta)\|_{\ell_2}^2$ wel iets voorstelt; dit is uiteraard afhankelijk van de concrete situatie. Verder is de invloed van sparsity niet te zien.

Tot slot moet opgemerkt worden, dat de Dantzig Selector minimaliseert op de ℓ_1-norm. Bij de DS wordt er gezocht naar een vector met een ℓ_1-norm die dicht bij de ℓ_1-norm van β zit. Toch kan er een kwaliteitsgarantie voor de DS gegeven worden, wanneer de β in het toegelaten gebied ligt en X voldoet aan een voorwaarde, dan kan er een bovengrens aan $\|\hat{\beta}-\beta\|_{\ell_2}^2$ gegeven worden.

4.2 Kwaliteit Dantzig Selector

Het bijzondere aan de DS is dat deze, indien de designmatrix X aan bepaalde voorwaarden voldoet, met een hoge kans een kwaliteitsgarantie kan geven. Dit werd aangetoond in de onderstaande stelling uit [8].

Stelling 7 (Kwaliteitsgarantie Dantzig Selector)**.** *Laat $\beta \in \mathbb{R}^p$ een S-sparse vector zijn en X een matrix die voldoet aan $\delta_{2S}(X)+\theta_{S,2S}(X)<1$. Kies $\lambda_p = \sqrt{2\log p}$. Dan is de kans op de gebeurtenis*

$$\left\|\beta-\hat{\beta}_{\text{DS}}\right\|_{\ell_2}^2 \leq \left(\frac{4}{1-\delta_{2S}(X)-\theta_{S,2S}(X)}\right)^2 (2\log p)S\sigma^2$$

groter of gelijk aan $1-\left(\sqrt{\pi\log p}\right)^{-1}$.

In deze paragraaf wordt stelling 7 op een andere wijze gepresenteerd. Door de nadruk te leggen op een orthogonaliteitsconditie, kan het verband tussen de kwaliteitsgarantie en de keuze van λ_p zichtbaar gemaakt worden. Het volledige bewijs is in appendix C.1 uitgewerkt, de subparagraaf 4.2.3 zal in grote lijnen de bewijsvoering geven.

4.2.1 Aannamen In deze paragraaf wordt aangenomen dat $\sigma = 1$ en dat de kolommen van X genormaliseerd zijn met ℓ_2-norm 1. Doordat $\sigma = 1$ geldt $\varepsilon \sim \mathcal{N}(0, I_n)$. Het bewijs is ook te voeren met een willekeurige σ en een matrix X waarvan de kolommen niet genormaliseerd zijn; de aanname maakt het bewijs iets overzichtelijker.

4.2.2 Orthogonaliteitsconditie Het is niet gegeven dat β in het toegelaten gebied van de DS ligt. De orthogonaliteitsconditie is een aanname die gesteld kan worden, om te verzekeren dat de 'echte' β in het toegelaten gebied van de DS ligt.

Definitie 8 (Orthogonaliteitsconditie)**.** Voor alle $j \in \mathbb{N}, 1 \leq j \leq p$ geldt

$$\left|\langle \varepsilon, X_j \rangle\right| \leq \lambda_p \sigma.$$

Gevolg. *Onder de orthogonaliteitsconditie geldt dat β in het toegelaten gebied van de methode van de Dantzig Selector ligt.*

Bewijs. Neem de orthogonaliteitsconditie aan. Het toegelaten gebied van de DS bestaat uit alle vectoren $\tilde{\beta} \in \mathbb{R}^p$ die voldoen aan

$$\left\| X^T \left(y - X\tilde{\beta} \right) \right\|_{\ell_\infty} \leq \lambda_p \sigma. \tag{2.3}$$

Beschouw β. Door op te merken dat $y = X\beta + \varepsilon$ verkrijgen we

$$\begin{aligned} \left\| X^T \left(y - X\beta \right) \right\|_{\ell_\infty} &= \left\| X^T \left(X\beta + \varepsilon - X\beta \right) \right\|_{\ell_\infty} \\ &= \left\| X^T \varepsilon \right\|_{\ell_\infty} \\ &= \sup_i |\langle \varepsilon, X_i \rangle| \end{aligned} \tag{4.1}$$

Het toepassen van de orthogonaliteitsconditie op (4.1) geeft het gevraagde, (2.3). □

Onder de orthogonaliteitsconditie geldt dat β in het toegelaten gebied van de DS ligt. Van alle vectoren $\tilde{\beta}$ uit het toegelaten gebied is de DS die $\tilde{\beta}$, waarvoor de ℓ_1-norm minimaal is. Een belangrijk gevolg hiervan is dat $\|\hat{\beta}\|_{\ell_1} \leq \|\beta\|_{\ell_1}$ geldt onder de orthogonaliteitsconditie.

4.2.3 Kwaliteitsgarantie DS Deze subparagraaf presenteert stelling 7 op een andere wijze. Door meer nadruk te leggen op de orthogonaliteitsconditie verdwijnt de kans uit stelling 7 en verhuist deze naar subparagraaf 4.2.4. Het volledige bewijs is te vinden in appendix C.1, de nummering van de vergelijkingen en ongelijkheden corresponderen met die van de appendix. In deze subparagraaf worden tussenstappen overgeslagen.

Stelling 9 (Kwaliteitsgarantie Dantzig Selector — versie 2)**.** *Zij $\beta \in \mathbb{R}^p$ een S-sparse vector en X een matrix die voldoet aan $\delta_{2S}(X) + \theta_{S,2S}(X) < 1$. Laat $\sigma = 1$ en $\hat{\beta} = \hat{\beta}_{\text{DS}}$ de Dantzig Selector zijn. Onder de orthogonaliteitsconditie geldt*

$$\left\| \beta - \hat{\beta}_{\text{DS}} \right\|_{\ell_2}^2 \leq \left(\frac{4}{1 - \delta_{2S}(X) - \theta_{S,2S}(X)} \right)^2 \lambda_p^2 \cdot S\sigma^2.$$

Bewijs. Vanwege de orthogonaliteitsconditie geldt dat β in het toegelaten gebied van de DS zit. Dit geeft $\|\hat{\beta}\|_{\ell_1} \leq \|\beta\|_{\ell_1}$. Laat $\hat{\beta} = \beta + h$, dan $h = \hat{\beta} - \beta$.

In subparagraaf C.1.1 van appendix C.1 is aangetoond, dat

$$\|h_{I^c}\|_{\ell_1} \leq \|h_I\|_{\ell_1}. \tag{C.4}$$

Wat dit impliceert, kan het beste uitgelegd worden door (C.4) anders op te schrijven en op te merken dat $\beta_i = 0$ voor alle $i \in I^c$:

$$\sum_{i \in I^c} |\hat{\beta}_i| = \left\| \hat{\beta}_{I^c} \right\|_{\ell_1} = \|h_{I^c}\|_{\ell_1} \leq \|h_I\|_{\ell_1} = \left\| \left(\hat{\beta} - \beta \right)_I \right\|_{\ell_1}.$$

Dit verklaart enigszins waarom de DS variabelen selecteert. Als β zeer sparse is, dan heeft I^c veel elementen en I niet zo veel. Omdat $\sum_{i \in I^c} |\hat{\beta}_i|$ begrensd is door

de som van het absolute verschil van β en $\hat{\beta}$ over de support, moet gelden dat veel coëfficiënten $\hat{\beta}_i$ klein moeten zijn of zelfs gelijk aan 0. Merk op dat dit nadrukkelijk te maken heeft met het feit, dat de DS minimaliseert op de ℓ_1-norm.

In subparagraaf C.1.1 van de appendix wordt, via de driehoeksongelijkheid en de orthogonaliteitsconditie, de onderstaande ongelijkheid aangetoond.

$$\left\| X^T X h \right\|_{\ell_\infty} \leq 2\lambda_p. \tag{C.5}$$

Door het onderstaande lemma te introduceren en te gebruiken, kan vervolgens de gezochte bovengrens voor $\|h\|_{\ell_2}^2$ gegeven worden. Het lemma zelf wordt in appendix C.2 bewezen. De tussenstappen van de verdere uitwerking van het bewijs zijn te vinden in appendix C.1.3.

Lemma. *Zij I een verzameling met $|I| = S$ en $\delta_{2S}(X) + \theta_{S,2S}(X) < 1$. Laat I_1 de indexverzameling zijn die correspondeert met de S grootste posities van h buiten I. Laat $I_2 = I \cup I_1$. Dan geldt*

$$\left\| h_{I_2} \right\|_{\ell_2} \leq \frac{1}{1-\delta_{2S}(X)} \left\| X_{I_2}^T X h \right\|_{\ell_2} + \frac{\theta_{S,2S}(X)}{(1-\delta_{2S}(X))\sqrt{S}} \|h_{I^c}\|_{\ell_1} \tag{C.6}$$

en

$$\|h\|_{\ell_2}^2 \leq \left\| h_{I_2} \right\|_{\ell_2}^2 + \frac{1}{S} \|h_{I^c}\|_{\ell_1}^2. \tag{C.7}$$

Het toepassen van de ongelijkheid van Cauchy, (C.4) en het feit dat $I \subseteq I_2$ geeft de onderstaande ongelijkheid

$$\|h_{I^c}\|_{\ell_1} \leq \|h_I\|_{\ell_1} \leq \sqrt{S}\|h_I\|_{\ell_2} \leq \sqrt{S} \left\| h_{I_2} \right\|_{\ell_2}. \tag{C.8}$$

Door (C.5) te gebruiken, op te merken dat $X_{I_2}^T$ uit $2S$ kolommen bestaat en de driehoeksongelijkheid te gebruiken komen we op de onderstaande ongelijkheid (zie de appendix voor de volledige uitwerking)

$$\left\| X_{I_2}^T X h \right\|_{\ell_2} \leq \sqrt{2S} \cdot 2\lambda_p. \tag{C.9}$$

Gebruikmakende van (C.8) en (C.9) kan (C.6) van het lemma gebruikt worden om een bovengrens voor $\|h_{I_2}\|_{\ell_2}$ te geven:

$$\left\| h_{I_2} \right\|_{\ell_2} \leq \frac{1}{1-\delta_{2S}(X)-\theta_{S,2S}(X)} \sqrt{2S} \cdot 2\lambda_p. \tag{C.13}$$

Via (C.8) en (C.7), de tweede ongelijkheid uit het lemma, wordt aangetoond dat

$$\|h\|_{\ell_2}^2 \leq 2 \left\| h_{I_2} \right\|_{\ell_2}^2.$$

Het invullen van (C.13) hierin geeft dat wat te bewijzen was,

$$\left\| \beta - \hat{\beta}_{\text{DS}} \right\|_{\ell_2}^2 \leq \left(\frac{4}{1-\delta_{2S}(X)-\theta_{S,2S}(X)} \right)^2 \lambda_p^2 \cdot S\sigma^2.$$

□

In bepaalde toepassingen is het mogelijk om de designmatrix X zelf in te stellen. De stelling laat opnieuw zien, waarom het gewenst is wanneer voor de designmatrix X geldt, dat $\delta_{2S}(X)+\theta_{S,2S}(X)$ zo klein mogelijk is. Wanneer $\delta_{2S}(X)+\theta_{S,2S}(X)$ klein is, gaat $1-\delta_{2S}(X)-\theta_{S,2S}(X)$ naar 1 toe en wordt $\frac{4}{1-\delta_{2S}(X)-\theta_{S,2S}(X)}$ kleiner. De bovengrens voor de kwaliteitsgarantie van de DS wordt hiermee kleiner.

Stelling 9 is waar wanneer de designmatrix X voldoet aan $1-\delta_{2S}(X)-\theta_{S,2S}(X)<1$ en wanneer de orthogonaliteitsconditie waar is. Neem aan dat we een designmatrix X tot onze beschikking hebben die voldoet aan $1-\delta_{2S}(X)-\theta_{S,2S}(X)<1$. Dan moet gekeken worden wanneer de orthogonaliteitsconditie waar is. Dit wordt gedaan in de volgende subparagraaf.

4.2.4 Kans op de orthogonaliteitsconditie Of er aan de orthogonaliteitsconditie voldaan wordt, hangt onder andere af van de keuze van λ_p. In het bewijs van stelling 9 gaat het erom, dat β in het toegelaten gebied ligt. Het toegelaten gebied mag echter niet te groot worden, wanneer $\|X^T y\|_{\ell_\infty} \leq \lambda_p \sigma$, dan geldt zelfs dat de nulvector in het toegelaten gebied van de DS ligt, wat dan ook meteen de DS is.

Het probleem is dus om λ_p zo te kiezen, dat β met een hoge kans in het toegelaten gebied ligt, maar dat het toegelaten gebied niet te groot wordt. Aangezien de DS de oplossing is van een convex lineair programma, zou de ideale λ_p zo gekozen moeten zijn, zodat β een hoekpunt is in het toegelaten gebied. De ℓ_1-norm van de DS zou in dat geval gelijk zijn aan de ℓ_1-norm van β.

De orthogonaliteitsconditie stelt dat $\|X^T\varepsilon\|_{\ell_\infty} \leq \lambda_p\sigma$. In de probleemstelling wordt aangenomen, dat $\varepsilon \sim \mathcal{N}\left(0,\sigma^2 I_n\right)$. Dit betekent, dat

$$X^T\varepsilon \sim \mathcal{N}\left(0,\sigma^2 X^T X\right). \tag{4.2}$$

Merk op dat er geen λ_p bestaat zodanig dat $\|X^T\varepsilon\|_{\ell_\infty} \leq \lambda_p\sigma$ geldt voor alle ε. Wel kan er gezocht worden naar een waarde voor λ_p zodat de orthogonaliteitsconditie met een hoge kans waar is. Dit betekent dat er gezocht moet worden naar een λ_p zodanig dat de kans dat het maximum van $X^T\varepsilon \sim \mathcal{N}\left(0,\sigma^2 X^T X\right)$ kleiner is dan λ_p.

In dit hoofdstuk is er aangenomen, dat $\sigma = 1$. Dit betekent dat $X^T\varepsilon \sim \mathcal{N}\left(0, X^T X\right)$. Laat $\varepsilon \sim \mathcal{N}\left(\mathbf{0}, I_n\right)$ en laat $Z_j = \langle \varepsilon, X_j \rangle$. Dan geldt

$$X_j^T\varepsilon = Z_j \sim \mathcal{N}\left(X_j^T\mathbf{0}, X_j^T I_n X_j\right).$$

Omdat er is aangenomen dat de kolommen van X genormaliseerd zijn op de ℓ_2-norm met norm 1, geldt voor elke i geldt dat $X_j^T I_n X_j = X_j^T X_j = \|X_i\|_{\ell_2} = 1$. Dit betekent dat de diagonaal van $X^T X$ bestaat uit enen. Vanwege het feit dat $X_j^T\mathbf{0} = 0$, weten we dat elke Z_j standaardnormaal verdeeld is; er geldt dus $Z_j \sim \mathcal{N}\left(0,1\right)$. Merk op dat er gezocht wordt naar $\mathbb{P}\left(\sup_j |Z_j| \leq \lambda_p\right)$. Laat $1-\Phi(a) = \mathbb{P}\left(Z \geq a\right)$ voor een stochast Z

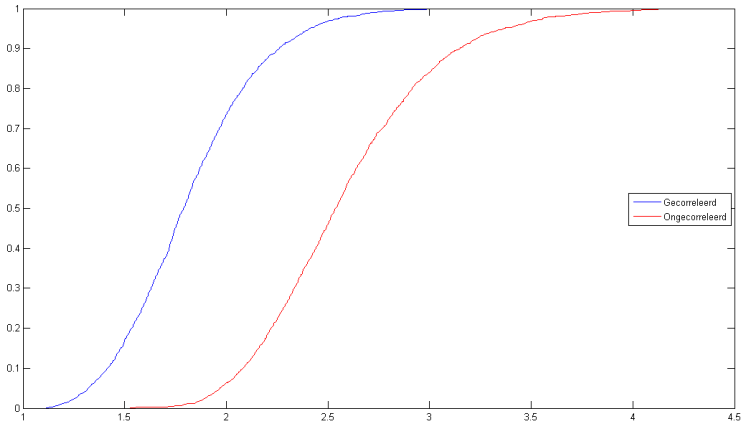
die standaardnormaal verdeeld is. Dit geeft

$$\begin{aligned}\mathbb{P}\left(|Z_j| \leq \lambda_p\right) &= \mathbb{P}\left(-\lambda_p \leq Z_j \leq \lambda_p\right) \\ &= 1 - 2 \cdot \mathbb{P}\left(Z_j \geq \lambda_p\right) \\ &= 1 - 2 \cdot \left(1 - \Phi(\lambda_p)\right) \qquad (4.3) \\ &= 2\Phi(\lambda_p) - 1. \qquad (4.4)\end{aligned}$$

Het probleem bij het bepalen van het maximum van $X^T\varepsilon \sim \mathcal{N}\left(0, X^T X\right)$ is dat $(X^T\varepsilon)_i$ correleert met $(X^T\varepsilon)_j$ voor $i \neq j$.

In [8] wordt $\lambda_p = \sqrt{2\log(p)}$ genomen en wordt gezegd dat de kans dat $\|X^T\varepsilon\|_{\ell_\infty} \leq \lambda_p\sigma$ groter is dan $1 - (\sqrt{\pi\log(p)})^{-1}$. Deze kans wordt uitgerekend door de gezochte kans uit te rekenen als aangenomen wordt, dat alle Z_i's onafhankelijk zijn van elkaar.

Om deze stap beter te begrijpen is er een kleine simulatie uitgevoerd waarbij het maximum van $M = 2000$ vectoren $y, z \in \mathbb{R}^n$ met $n = 60$ bekeken werd. Hierbij is er gekeken naar vectoren y die een realisatie zijn van ε, hierbij wordt dus aangenomen dat $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{60}$ kunnen correleren. Verder is er gekeken naar vectoren y waarbij $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{60}$ ongecorreleerd zijn. Om de verdelingsfunctie van het maximum te berekenen, wordt er gebruik gemaakt van de MATLAB®-functie `ecdf` (empirische verdelingsfunctie). Figuur 4.1 laat de gevonden resultaten zien.



Figuur 4.1: De gevonden empirische verdelingsfunctie bij $M = 1000$ simulaties.

Figuur 4.1 geeft de indruk, dat het maximum van 'gecorreleerde' vectoren lager is dan het maximum van 'ongecorreleerde' vectoren. Uiteraard moet opgewerkt worden dat niet alle correlatie matrices getest zijn en alle vectoren gesimuleerd zijn; het geeft echter wel een vermoeden.

4.3 Oracle inequalities

In deze paragraaf zal worden nagegaan hoe 'goed' de gevonden kwaliteitsgarantie van de DS is. Dit wordt gedaan door de gevonden bovengrens te vergelijken met het schatten van β in een veel gunstigere situatie, namelijk dat je weet waar de niet-nul coëfficienten zijn.

Neem aan dat er een orakel tot onze beschikking staat die exact kan vertellen waar de niet-nul elementen uit de parametervector β zijn; de support I is dus gegeven en met dit gegeven zal β zo goed mogelijk geschat worden. De gebruikte schatter zal de orakelschatter genoemd worden.

Aangezien gegeven is welke elementen van β gelijk zijn aan nul, geldt $\beta^\star_{I^c} = \beta_{I^c} = \mathbf{0}$. Resteert het schatten van $\beta^\star_I$ uit de metingen y,

$$y = X\beta + \varepsilon.$$

Het feit dat $\beta_{I^c} = 0$ heeft als gevolg dat veel kolommen van X "niet gebruikt worden" bij het berekenen van $X\beta$, zie het voorbeeld op bladzijde 13. We kunnen y dus herschrijven als

$$y = X_I\beta_I + \varepsilon. \tag{4.5}$$

Aangezien X_I hier een $n \times |I|$ submatrix is en we aannemen dat $|I| < n$, verkrijgen we het klassieke geval van lineaire regressie; er moeten $|I|$ parameters geschat worden uit n metingen. We gebruiken de kleinste kwadratenmethode. Merk op dat de $\beta^\star_I$ gezocht is die voldoet aan de normaalvergelijkingen

$$X_I^T X_I \beta^\star_I = X_I^T y.$$

Aangenomen dat $\left(X_I^T X_I\right)^{-1}$ bestaat, dit kan verzekerd worden door $\delta_{2S}(X) < 1$ te eisen, verkrijgen we een unieke oplossing.

$$\beta^\star_I = \left(X_I^T X_I\right)^{-1} X_I^T y. \tag{4.6}$$

Stelling 10 (Kwaliteit orakelschatter)**.** *Neem aan dat I, de support van β, gegeven is. Neem verder aan dat $\delta_{2S}(X) < 1$ en $|I| = S$. Laat $\beta^\star$ de orakelschatter zijn, $\beta^\star$ is de schatter voor β zijn wanneer I gegeven is. Dan geldt*

$$\mathbb{E}\left[\left\|\beta^\star - \beta\right\|_{\ell_2}^2\right] \geq \frac{S\sigma^2}{1+\delta_{2S}(X)}. \tag{4.7}$$

Bewijs. Zie appendix D. □

4.3.1 Kwaliteit DS versus de orakelschatter Door het resultaat van stelling 9 met (4.7) te vergelijken, komen overeenkomsten en verschillen meteen naar boven. Op een constante na, levert de Dantzig Selector een factor λ_p^2 in op $\|\hat{\beta} - \beta\|_{\ell_2}^2$.

Neem $\lambda_p = \sqrt{2\log(p)}$ zoals in [8] is gebruikt. Dan geldt dus dat er een factor $(2\log p)$ ingeleverd moet worden ten opzichte van de orakelschatter. De orakelschatter heeft echter meer informatie tot zijn beschikking, de orakelschatter weet immers waar de niet-nullen zijn. Schijnbaar is $(2\log p)$ de prijs in kwaliteit die de DS moet betalen, omdat niet gegeven is waar de nullen zijn. Dit is een zeer lage prijs, de grootte van p heeft hier weinig invloed op. Zo is $\sqrt{2\log(100000)} = 3.16227766$.

4.4 'Verbetering' van de Dantzig Selector

Het grote probleem met het orakel is dat deze in de praktijk niet bestaat. De orakelschatter kan dus niet gebruikt worden. Wanneer het doel is om $\|\beta - \hat{\beta}\|_{\ell_2}^2$ te minimaliseren, zou je echter de verzameling I kunnen schatten.

Stel dat de Dantzig Selector gebruikt wordt om I te schatten. We reduceren het probleem dan weer tot de klassieke situatie, β kan dan geschat worden met behulp van de kleinste kwadratenmethode.

De zojuist genoemde methode is door de auteurs van het originele artikel Gauss-DS genoemd. Hoewel deze 'tweestaps' schatter in het artikel de afkorting GDS gegeven is, zal ik de schatter benoemen als Gauss-DS. De reden hiervoor is dat er een recent artikel is uitgekomen waarin de 'Group Dantzig Selector' geïntroduceerd wordt, welke ook als afkorting GDS is gegeven. Om verwarring te voorkomen zal daarom Gauss-DS worden gebruikt.

Tot slot moet er opgemerkt worden, dat de Gauss-DS niet per definitie een betere schatter is dan de DS. Allereerst is de kwaliteit van de Gauss-DS schatter afhankelijk van de 'orakel'-kwaliteit van de DS. In de simulatie van hoofdstuk 5 is aandacht gegeven aan de vraag, hoe goed de DS kan voorspellen waar de niet-nul elementen zitten. Nog belangrijker is wellicht de gebruikte maat voor kwaliteit. De Dantzig Selector zoekt een vector met een ℓ_1-norm die dicht bij de ℓ_1-norm van β ligt. De keuze van λ_p geeft immers ook aan of β in het toegelaten gebied ligt en hoe dicht deze aan de rand van het toegelaten gebied ligt[1]. Als het minimaliseren van $\|\beta - \hat{\beta}\|_{\ell_2}^2$ het doel is, dan kan overwogen worden om de Gauss-DS te gebruiken.

[1] Wanneer β dicht bij de rand van het toegelaten gebied ligt, wordt het verschil in ℓ_1-norm tussen de DS en β ook kleiner.

5

Simulatie

Een bachelorproject wijden aan een schatter zonder te verifiëren dat deze werkt is niet wijs. In dit hoofdstuk wordt getest of de Dantzig Selector daadwerkelijk werkt. Ideaal zou zijn om de schatter ook te testen bij een toepassing, maar dit was vanwege de tijd niet haalbaar. Er is met een numeriek experiment getest of de DS in staat is om een sparse vector β te schatten.

5.1 Inleiding

In [8] en in dit bachelorproject wordt verondersteld, dat in bepaalde gevallen een sparse parametervector β met behulp van de methode van de DS te schatten is, wanneer het aantal metingen n kleiner is dan het aantal parameters p. Het is echter aannemelijk, dat wanneer er *zeer* weinig metingen zijn, de parametervector niet meer te schatten is.

Neem als extreem voorbeeld $n = 1$, $p = 100$ en S, het aantal niet-nul elementen in β, gelijk aan 10. Het schatten van de parameters met enkel één meting is niet mogelijk. Dit roept de vraag op voor welke verhoudingen tussen n, p en S β niet meer te schatten is. Een van de doelen van het uitvoeren van de simulaties is om meer inzicht te krijgen in de verhoudingen $S : n : p$ waarvoor β te schatten is.

De methode van de DS gebruikt een gekozen waarde λ_p, die het toegelaten gebied definieert. Waar [8] in eerste instantie voor λ_p de waarde $\sqrt{2\log p}$ voorstelt, zijn er andere geluiden, waaronder [9], die een andere waarde en of methode voor het verkrijgen van de best mogelijke λ_p voorstellen. In de simulaties worden verschillende waarden van λ_p, verkregen via verschillende methoden, uitgeprobeerd om meer inzicht te verkrijgen in het bepalen van de ideale waarde voor λ_p.

In paragraaf 4.4 wordt een mogelijke verbetering van de Dantzig Selector genoemd door de DS te gebruiken als een orakel. Deze methode zou kunnen werken als de DS de invloed hebbende parameters zou kunnen selecteren; het zou kunnen werken als de support van β goed wordt bepaald. Aangezien de echte support I bij een simulatie

bekend is, kan er gekeken worden hoe goed de DS de juiste parameters kan selecteren. Verder kan de orakelschatter bepaald worden.

Kortom zal er voor verschillende verschoudingen tussen S, n en p en voor verschillende waarden van λ_p gekeken worden hoe goed β geschat kan worden via de DS en zijn verbetering de Gauss-DS.

5.2 Analyse

Het simuleren bestaat uit het zo goed mogelijk schatten van β via de methode van de Dantzig Selector. Per verhouding $S : n : p$ en per λ_p zullen $M = 1000$ simulatieruns gehouden worden. Per simulatierun zal een gegenereerde vector β geschat worden aan de hand van de metingen $y = X\beta + \varepsilon$.

Om vergelijkbare omstandigheden te creëren zullen enkele parameters niet veranderd worden in de simulatie. Zo zal in elke simulatie $p = 100$ en $\sigma = 1$. De designmatrix X zal bestaan uit trekkingen van de standaardnormale verdeling, waarna de kolommen van X genormaliseerd worden op de ℓ_2-norm. Zoals het voorbeeld in subparagraaf 3.2.1 laat zien, zorgt dit er voor, dat $\delta_1(X) = 0$.

In de simulatie zal er gekeken worden naar enkele verhoudingen $S : n : p$. Hierbij zal $S = 10$, $p = 100$ en n varieren van 20 tot 150. Omdat sparsity niet exact gedefinieerd is, is er voor $S = 10$ gekozen; 10% van de parameters is niet gelijk aan nul. Merk op dat er situaties $n < p$, $n = p$ en $n > p$ worden uitgetest.

Bij de simulaties zullen verschillende waarden van λ_p uitgeprobeerd worden. Per simulatierun zal er een nieuwe matrix X gegenereerd worden, verschillende waarden van λ_p zullen dus niet op dezelfde matrix X getest worden. Achteraf gezien zou het wellicht verstandiger zijn geweest om verschillende waarden van λ_p uit te testen op dezelfde matrix X. zodat de verschillende waarden onder iets meer gelijke omstandigheden met elkaar worden vergeleken.

De verschillende waarden voor λ_p die geprobeerd worden, zijn $\sqrt{2\log p}$, $\sqrt{2\log n}$, 0 en een λ_p die afhangt van X en bepaald wordt via Monte-Carlo simulatie. De laatstgenoemde λ_p zal dus per gegenereerde X bepaald worden. Er zullen 1000 vectoren $z \sim \mathcal{N}(0, I_n)$ getrokken worden die de ruis ε moeten simuleren. Per trekking zal $\sup_i \langle X_i, z\rangle$ uitgerekend worden, het gemiddelde van de suprema zal de λ_p van die simulatierun zijn.

Het uitvoeren van de simulatie zal met behulp van het programma MATLAB® gedaan worden. Bij het bepalen van de DS zal er gebruik worden gemaakt van de MATLAB®-functie `linprog`. Deze functie lost een gegeven lineair programma op met behulp van een interior point algoritme.

Om de simulaties niet af te laten hangen van de keuze van één specifieke parametervector β, wordt er elke simulatierun een parametervector β gegenereerd. Deze wordt vervolgens geschat met de Dantzig Selector $\hat{\beta}_{\text{DS}}$. Paragraaf 4.4 stelt echter voor om

de DS te gebruiken als een 'orakel': laat $\hat{I}$ een schatter zijn voor I. Omdat er gebruik gemaakt wordt van een interior point algoritme en omdat MATLAB® wellicht afrond, wordt $\hat{I}$ gedefinieerd als $\{\hat{\beta}_i : i \in \mathbb{N}, 1 \leq i \leq p \text{ en } \hat{\beta}_i \leq 0.1\sigma\}$. Door $\hat{I}$ als schatter voor I te gebruiken, kan de Gauss-DS schatter $\hat{\beta}_{Gauss-\text{DS}}$ nu berekend worden.

Merk op dat er elke simulatierun een parametervector β en een support I gegeneerd worden. De support is dus gegeven en hierdoor is er een echte orakel tot onze beschikking. Dit betekent dat de orakelschatter $\beta^\star$ ook berekend kan worden.

5.2.1 Presentatie van de resultaten Paragraaf 4.1 volgend zal de kwaliteit van de schatter beoordeeld worden op $\|\beta - \hat{\beta}\|_{\ell_2}^2$. Per simulatierun zal deze waarde voor alle schatters uitgerekend en opgeslagen worden. Aan het einde van elke simulatie (elke combinatie $S : n : p$ en λ_p) is er 1000 keer uitgerekend hoe groot $\|\beta - \hat{\beta}\|_{\ell_2}^2$ is. Aan de hand van deze gegevens zal het gemiddelde en de 5de en 95ste kwantiel van $\|\beta - \hat{\beta}\|_{\ell_2}^2$ berekend worden.

Per simulatierun zal er voor elke parameter $i \in \mathbb{N}$, $1 \leq i \leq p = 100$ gekeken worden of i in de echte support I zit, en of i ook in de geschatte support $\hat{I}$ zit.

	$i \in \hat{I}$	$i \in \hat{I}^c$
$i \in I$	True positive	False negative
$i \in I^c$	False positive	True negative

Tabel 5.1: Definitie van False positive en False negative.

Tabel 5.1 geeft aan of i een true positive, false negative, false positive of true negative oplevert. Om weer te geven hoe goed de DS de support kan schatten, zal het percentage false positives en false negatives uitgerekend en opgeslagen worden. Ook hier zal er aan het einde van elke simulatie het gemiddelde percentage en de 5de en 95ste kwantiel berekend worden.

5.3 Simulatieopzet

Deze simulatieopzet is samengevat in algoritme 1. Een uitgeschreven versie van deze opzet volgt hieronder.

Per simulatierun zal er een designmatrix X, parametervector β en ruis ε gegenereerd worden. Vervolgens wordt y berekend en wordt er een λ_p bepaald of gekozen. Tot slot zal, gegeven X, y en λ_p, de Dantzig Selector $\hat{\beta}_{\text{DS}}$ berekend worden.

De DS selecteert parameters, het kan gebruikt worden om de support I te bepalen. Nadat $\hat{\beta}_{\text{DS}}$ berekend is, kan $\hat{I}_{\text{DS}}$, de schatter voor I via de DS, bepaald worden. Deze

$\hat{I}_{\text{DS}}$ zal gebruikt worden als een voorspeller zodat $\hat{\beta}_{Gauss-\text{DS}}$ bepaald kan worden. Tot slot wordt de orakelschatter $\beta^\star$ nog berekend.

Algorithm 1: Simulatieopzet voor het bepalen van $\|\beta - \hat{\beta}_{\text{DS}}\|_{\ell_2}^2$, $\|\beta - \beta^\star\|_{\ell_2}^2$ en $\|\beta - \hat{\beta}_{Gauss-\text{DS}}\|_{\ell_2}^2$ bij verschillende waarden van λ_p

```
initialization;
Kies een p, S en een σ, vast;
for n = [20,30,40,50,60,80,100,150]
    for i = 1:M
        Genereer een X;
        Bepaal λ_p;
        Kies een support indices = I met |I| = S;
        Genereer een β (beta_origineel);
        Bereken y = Xβ + ε;
        Bepaal β̂_DS en Î aan de hand van X, y en λ_p;
        Bepaal β̂_Gauss−DS en β*;
```

5.3.1 Lineair programma De Dantzig Selector zoals gegeven in (DS) op bladzijde 7 is nog niet klaar om door MATLAB® bepaald te worden. In appendix B.1 wordt aangetoond dat de oplossing van (DS) gevonden kan worden door het onderstaande lineaire programma op te lossen:

$$\begin{aligned} \min \sum_i u_i \qquad & \text{odv} \\ -u \le \tilde{\beta} &\le u \\ -\lambda_p \sigma \mathbf{1} \le X^T(y - X\tilde{\beta}) &\le \lambda_p \sigma \mathbf{1}. \end{aligned}$$

De functie `linprog` van MATLAB® kan de bovenstaande versie echter niet interpreteren. Door de randvoorwaarden van het bovenstaande programma om te schrijven naar de randvoorwaarden in het onderstaande programma, kan MATLAB® het programma wel oplossen.

$$\min \sum_i u_i \qquad \text{odv} \tag{5.1}$$

$$-u + \tilde{\beta} \le 0 \tag{5.2}$$

$$-u - \tilde{\beta} \le 0 \tag{5.3}$$

$$X^T X\tilde{\beta} \le \lambda_p \sigma \mathbf{1} + X^T y \tag{5.4}$$

$$-X^T X\tilde{\beta} \le \lambda_p \sigma \mathbf{1} - X^T y \tag{5.5}$$

Merk op dat (5.1) eigenlijk gezien moet worden als een functie in u en β. De doelfunctie die geminimaliseerd moet worden is $Z = \sum_i u_i + 0 \cdot \tilde{\beta}_i$.

5.4 Resultaten

De resultaten van de simulaties zijn in de figuren en tabellen van de komende pagina's te zien.

5.4.1 Tabellen De tabellen zijn gegroepeerd op de waarde van λ_p.

		n							
		20	30	40	50	60	80	100	150
$\|\beta-\beta^\star\|^2_{\ell_2}$	mean	0.01109	0.005411	0.003607	0.002769	0.00213	0.001548	0.001214	0.000799
	$q_n(0.05)$	0.003424	0.001861	0.00131	0.001068	0.0008094	0.0005885	0.0004533	0.0003044
	$q_n(0.95)$	0.0253	0.01095	0.007217	0.0053	0.004148	0.002875	0.002303	0.001498
$\|\beta-\hat{\beta}_{\text{DS}}\|^2_{\ell_2}$	mean	487.2	92.87	4.075	0.1357	0.05703	0.02108	0.0125	0.006812
	$q_n(0.05)$	32.93	0.1214	0.03637	0.01846	0.01293	0.007947	0.006056	0.0038
	$q_n(0.95)$	1290	473.3	8.758	0.4375	0.1549	0.04842	0.02335	0.01071
$\|\beta-\hat{\beta}_{Gauss-\text{DS}}\|^2_{\ell_2}$	mean	543.8	105.9	3.338	0.01049	0.005195	0.002596	0.001728	0.001007
	$q_n(0.05)$	29.56	0.01248	0.003884	0.002191	0.001388	0.000829	0.0005635	0.000319
	$q_n(0.95)$	1306	504.9	2.721	0.02365	0.01123	0.005552	0.003561	0.002113
% false-negative	mean	4.576	1.415	0.193	0.07	0.052	0.045	0.034	0.036
	$q_n(0.05)$	2	0	0	0	0	0	0	0
	$q_n(0.95)$	7	4	1	1	1	0	0	0
% false-positive	mean	14.19	16.52	11.04	5.824	3.495	1.475	0.76	0.314
	$q_n(0.05)$	11.05	5	2	1	0	0	0	0
	$q_n(0.95)$	17	23	25	14	9	4	3	1
λ_p	mean	2.353	2.406	2.431	2.447	2.456	2.47	2.477	2.488
	$q_n(0.05)$	2.327	2.382	2.407	2.422	2.433	2.446	2.452	2.464
	$q_n(0.95)$	2.382	2.431	2.456	2.471	2.481	2.493	2.5	2.511

Tabel 5.2: Resultaten van de DS bij $p = 100$, $S = 10$ en $M = 1000$. De waarde van λ_p is per simulatierun geschat via Monte-Carlo simulaties (zie analyse).

		n							
		20	30	40	50	60	80	100	150
$\|\beta-\hat{\beta}_{\text{DS}}\|^2_{\ell_2}$	mean	462.5	94.84	4.058	0.2281	0.08424	0.03042	0.01746	0.009204
	$q_n(0.05)$	39.76	0.2496	0.04911	0.02593	0.01826	0.01141	0.007974	0.005468
	$q_n(0.95)$	1284	459.8	6.256	0.742	0.2382	0.07191	0.03589	0.01426
$\|\beta-\hat{\beta}_{Gauss-\text{DS}}\|^2_{\ell_2}$	mean	1565	223.1	3.365	0.01213	0.00479	0.002308	0.001573	0.0008911
	$q_n(0.05)$	38.5	0.01227	0.003819	0.002066	0.001331	0.0007867	0.0005614	0.0003337
	$q_n(0.95)$	1324	454.4	1.196	0.0251	0.01099	0.004927	0.003222	0.001702
% false-negative	mean	4.457	1.401	0.2	0.071	0.066	0.042	0.03	0.034
	$q_n(0.05)$	2	0	0	0	0	0	0	0
	$q_n(0.95)$	7	4	1	1	1	0	0	0
% false-positive	mean	14.03	16.44	10.61	6.225	3.43	1.3	0.547	0.139
	$q_n(0.05)$	11	6	2	1	0	0	0	0
	$q_n(0.95)$	17	23	23	15	9	4	2	1

Tabel 5.3: Resultaten van de DS bij $p = 100$, $S = 10$ en $M = 1000$. De waarde van λ_p is gelijk aan $\sqrt{2\log p} \approx 3.035$.

		n							
		20	30	40	50	60	80	100	150
$\|\beta-\hat{\beta}_{\text{DS}}\|^2_{\ell_2}$	mean	462.9	94.43	3.91	0.1948	0.07521	0.02903	0.01746	0.009904
	$q_n(0.05)$	37.03	0.1914	0.039	0.02291	0.01667	0.01087	0.007974	0.005875
	$q_n(0.95)$	1286	456.8	5.439	0.6157	0.2112	0.06876	0.03589	0.01516
$\|\beta-\hat{\beta}_{Gauss-\text{DS}}\|^2_{\ell_2}$	mean	981.3	94.65	3.328	0.01095	0.004854	0.002329	0.001573	0.0008821
	$q_n(0.05)$	37.24	0.01235	0.003745	0.002143	0.001372	0.0007818	0.0005614	0.0003319
	$q_n(0.95)$	1323	452.6	1.002	0.02419	0.01099	0.004945	0.003222	0.001694
% false-negative	mean	4.45	1.38	0.187	0.064	0.064	0.042	0.03	0.035
	$q_n(0.05)$	2	0	0	0	0	0	0	0
	$q_n(0.95)$	7	4	1	1	1	0	0	0
% false-positive	mean	14.07	16.52	10.39	6.134	3.435	1.319	0.547	0.118
	$q_n(0.05)$	11	6	2	1	0	0	0	0
	$q_n(0.95)$	17	23	23	15	9	4	2	1
λ_p		2.448	2.608	2.716	2.797	2.862	2.96	3.035	3.166

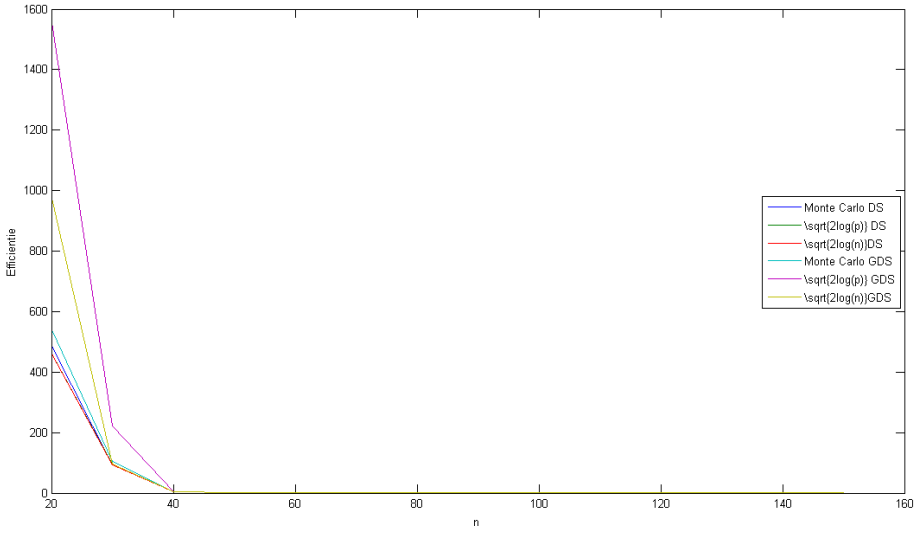
Tabel 5.4: Resultaten van de DS bij $p = 100$, $S = 10$ en $M = 1000$. De waarde van λ_p is gelijk aan $\sqrt{2\log n}$.

		n							
		20	30	40	50	60	80	100	150
$\|\beta-\hat{\beta}_{\text{DS}}\|^2_{\ell_2}$	mean	465.3	93.27	3.367	0.03927	0.03518	0.05732	7.183	0.02215
	$q_n(0.05)$	38.17	0.03611	0.01863	0.01673	0.01847	0.02974	0.09347	0.01453
	$q_n(0.95)$	1273	450.7	0.5006	0.08327	0.05955	0.1014	28.37	0.03217
$\|\beta-\hat{\beta}_{Gauss-\text{DS}}\|^2_{\ell_2}$	mean	1148	1175	3.504	0.0381	0.03495	0.05514	7.163	0.01877
	$q_n(0.05)$	35.47	0.02905	0.01601	0.01587	0.01679	0.02581	0.08749	0.01123
	$q_n(0.95)$	1461	524.1	0.4901	0.08047	0.06131	0.09811	29.29	0.02842
% false-negative	mean	4.43	1.31	0.109	0.022	0.03	0.015	0.004	0.008
	$q_n(0.05)$	2	0	0	0	0	0	0	0
	$q_n(0.95)$	7	4	1	0	0	0	0	0
% false-positive	mean	14.3	17.87	15.91	18.86	24.88	43.69	78.71	50.5
	$q_n(0.05)$	12	8	10	13	18	34	63	41
	$q_n(0.95)$	17	24	24	25	31	53	89	60

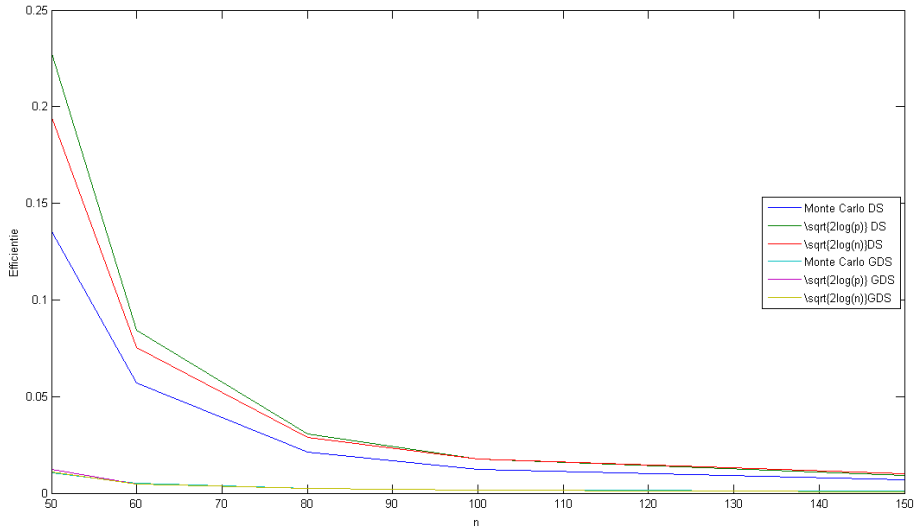
Tabel 5.5: Resultaten van de DS bij $p = 100$, $S = 10$ en $M = 1000$. De waarde van λ_p is gelijk aan 0.

5.4.2 Kwaliteit van de DS en Gauss-DS De kwaliteit van de DS en de Gauss-DS voor verschillende λ_p waarden zijn weergeven in figuur 5.1. Omdat $\|\hat{\beta}-\beta\|^2_{\ell_2}$ zeer groot is voor kleinere waarden van n, geeft figuur 5.2 hetzelfde weer vanaf $n = 50$.

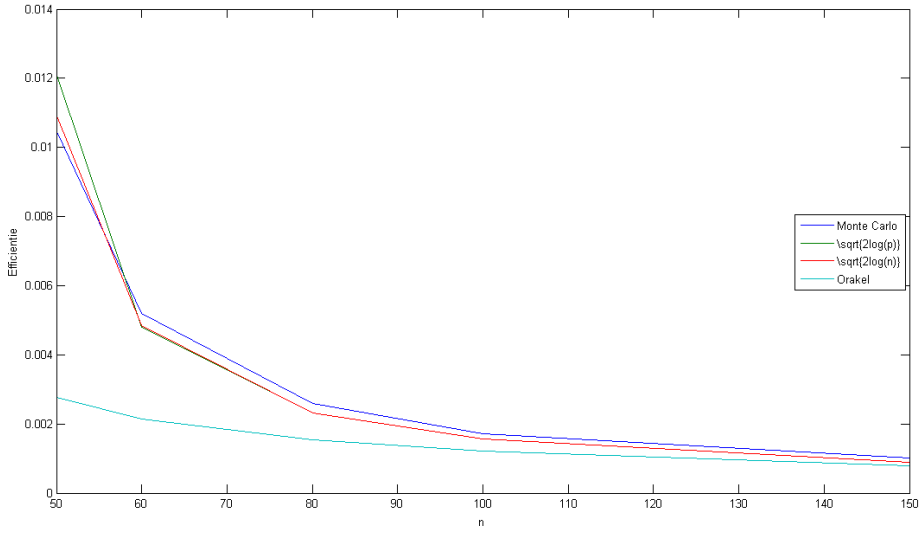
Omdat het verschil in $\|\hat{\beta}-\beta\|^2_{\ell_2}$ bij de Gauss-DS hier nog niet zichtbaar is, is dit in figuur 5.3 geplot. Om de Gauss-DS met de orakelschatter te vergelijken, zit de orakelschatter ook in deze figuur.



Figuur 5.1: De gemiddelde $\|\hat{\beta} - \beta\|_{\ell_2}^2$ van de DS en de Gauss-DS voor verschillende λ_p waarden voor alle n.



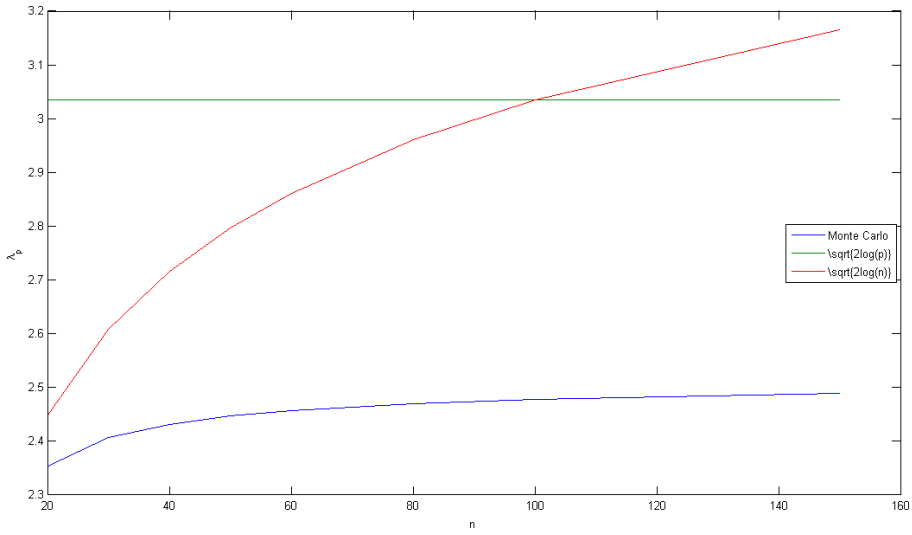
Figuur 5.2: De gemiddelde $\|\hat{\beta} - \beta\|_{\ell_2}^2$ van de DS en de Gauss-DS voor verschillende λ_p waarden voor $n = 50$ tot $n = 150$.



Figuur 5.3: De gemiddelde $\|\hat{\beta} - \beta\|_{\ell_2}^2$ van de Gauss-DS en de orakelschatter voor verschillende λ_p waarden voor $n = 50$ tot $n = 150$.

5.4.3 Het schatten van λ_p via de Monte-Carlo simulaties Zoals eerder is vermeld, zijn er in de simulaties verschillende waarden voor λ_p uitgeprobeerd: $\sqrt{2\log p}$, $\sqrt{2\log n}$, 0 en een λ_p verkregen via Monte-Carlo simulaties. Hoe de laatstgenoemde λ_p geschat wordt, is verteld in de analyse van de simulatie. Voor verschillende waarden van n is de gemiddelde λ_p van alle simulaties te vinden in tabel 5.2.

Figuur 5.4 toont de geschatte waarde van λ_p ten opzichte van $\sqrt{2\log p}$ (constante) en $\sqrt{2\log n}$. Het is duidelijk te zien, dat de 'Monte-Carlo' geschatte λ_p aanzienlijk kleiner is dan $\sqrt{2\log p}$ en $\sqrt{2\log n}$.



Figuur 5.4: De geschatte waarde van λ_p via MC-simulatie ten opzichte van $\sqrt{2\log p}$ en $\sqrt{2\log n}$.

5.5 Conclusie

De resultaten van de simulaties laten duidelijk zien dat de DS onder bepaalde omstandigheden in staat is om een sparse parametervector te schatten met weinig metingen. Door het uitvoeren van de simulaties heb ik meer inzicht gekregen in de werking van de DS; de verkregen conclusies worden in de komende subparagrafen verwerkt.

5.5.1 Verhoudingen $S:n:p$ Wanneer er te weinig metingen beschikbaar zijn, kan de DS de parametervector niet schatten. Dit is duidelijk te zien door te kijken naar $\|\beta-\hat{\beta}_{\text{DS}}\|_{\ell_2}^2$ in de kolommen waar $n=20$. Doordat dan ook de verkeerde parameters worden geselecteerd, kan de Gauss-DS dit ook niet verbeteren.

In de resultaten is duidelijk te zien dat β veel beter geschat wordt wanneer n toeneemt; dit was uiteraard ook te verwachten. In de setting van de simulatie lijkt $n=50$ de grens te zijn waarna β zeer goed geschat kan worden (met de Gauss-DS).

Bij $n=50$ lijkt de DS een degelijke schatter voor β te zijn. Dat het percentage false negatives dan zeer laag is, zorgt er voor dat de Gauss-DS schatter enorm goed is; in de resultaten is $\|\beta-\hat{\beta}_{Gauss-\text{DS}}\|_{\ell_2}^2$ ongeveer een factor 4 of 5 groter dan $\|\beta-\beta^\star\|_{\ell_2}^2$. Dit verschil, een factor 4 of 5, komt hoogstwaarschijnlijk omdat het percentage false positives nog ongeveer 5% is.

5.5.2 De keuze van λ_p Het is niet meteen duidelijk dat $\lambda_p = 0$ slechtere resultaten geeft dan andere keuzes voor λ_p. Pas bij grotere waarden van n ontstaat er duidelijk verschil. Vreemd is het gemiddelde van $\|\hat{\beta} - \beta\|_{\ell_2}^2$ bij $n = 100$; omdat het gemiddelde bij $n = 150$ weer lager is, kan ik verklaring geven waarom er bij $n = 100$ meerdere uitschieters zijn (de kwantielen geven aan dat er uitschieters zijn). Verder moet er gezegd worden dat het berekenen van de DS bij deze waarde van λ_p enorm lang duurde.

De keuze voor λ_p is zo, dat β in het toegelaten van de DS ligt. Hiervoor wordt door [8] de waarde $\sqrt{2\log(p)}$ voorgesteld om te gebruiken, maar wordt er bij de numerieke experimenten geopperd dat Monte Carlo-simulatie een betere λ_p zal geven. Zelf is $\lambda_p = \sqrt{2\log(n)}$ ook nog uitgeprobeerd.

Figuur 5.4 laat duidelijk zien dat de λ_p die telkens bepaald is via Monte Carlo-simulatie, kleiner is dan $\sqrt{\log(p)}$; voor kleine n lijkt het zelfs van n af te hangen. Dit is ook logisch, aangezien er bij een grotere n meer trekkingen uit de standaard-normale verdeling worden getrokken om de ruis te simuleren, het supremum zal dan ook hoger liggen.

Het is lastig om te zeggen, dat één bepaalde keuze voor λ_p kwalitatief betere resultaten geeft. Voor de DS geldt dat de λ_p van de Monte Carlo-simulaties betere resultaten voor $\|\hat{\beta} - \beta\|_{\ell_2}^2$ geeft, dit is goed te zien in figuur 5.2. Wanneer er echter gekeken wordt naar de Gauss-DS schatter, dan moet dezelfde λ_p (Monte Carlo) het onderspit delven. Na $n = 50$ geven de waarden $\sqrt{2\log(p)}$ en $\sqrt{2\log(n)}$ betere resultaten. Merk echter op dat $\|\hat{\beta}_{Gauss-\text{DS}} - \beta\|_{\ell_2}^2$ voor verschillende waarden van λ_p niet ver van de orakelschatter af ligt.

Doordat $\|\hat{\beta}_{Gauss-\text{DS}} - \beta\|_{\ell_2}^2$ groter is voor de grotere waarden van n bij de λ_p die bepaald is via Monte Carlo simulatie, komt de vraag naar boven of de Monte Carlo schatter voor λ_p wel de juiste schatter is. De afweging moet gemaakt worden of het ruimer nemen van λ_p gewenst is, zodat het selecteren van de juiste parameters beter verloopt.

5.5.3 Het schatten van I Duidelijk is dat het percentage false-negatives veel kleiner is dan het percentage false-positives; kijkende naar de Gauss-DS is dit echter niet erg. Bij een false-negative wordt een coëfficient van $\hat{\beta}_{Gauss-\text{DS}}$ op 0 gezet terwijl deze niet-nul is in β. Dit maakt $\|\beta - \hat{\beta}_{Gauss-\text{DS}}\|_{\ell_2}^2$ meteen groot. Bij een false-positive wordt er een element meegenomen, dit element moet bij de $\hat{\beta}_{Gauss-\text{DS}}$ nog geschat worden. Bij het schatten van $\hat{\beta}_{Gauss-\text{DS}}$ moeten weinig parameters geschat worden aan de hand van n metingen: de false-positive kan dus dicht bij 0 geschat worden. Hierdoor is de impact van een false-positive dus veel kleiner dan die van een false-negatives.

Tot slot kan nog het argument gegeven worden dat de niet-nullen van β getrokken worden uit een stochast die normaal verdeeld is met gemiddelde 0 en variantie 10000; het kan dus voorkomen dat een β_i dicht bij 0 getrokken wordt. Dit kan betekenen dat het percentage false-negatives hier iets hoger uitvalt.

*5.5.4 Gauss-*DS Wanneer er voldoende metingen zijn, kan de Gauss-DS enorm 'goed' werken. Omdat de Gauss-DS $\|\beta_{\hat{I}} - \hat{\beta}_{\hat{I}}\|_{\ell_2}^2$ wil minimaliseren, levert dit goede resultaten op voor $\|\beta - \hat{\beta}\|_{\ell_2}^2$. Het gebruiken van de DS als een voorspellende orakel kan dus een 'betere' schatter geven, waarbij opgemerkt moet worden dat beter afhankelijk is van de context; wanneer het gewenst is om $\|\beta - \hat{\beta}\|_{\ell_2}^2$ te minimaliseren, dan geeft dit een betere schatter.

Duidelijk is dat de Gauss-DS na ongeveer $n = 50$ metingen zeer lage waarden voor $\|\beta - \hat{\beta}\|_{\ell_2}^2$ geeft, doordat de DS dan de juiste parameters goed selecteert. De winst in $\|\beta - \hat{\beta}\|_{\ell_2}^2$ die in de gebruikte setting behaald is door de Gauss-DS, is bijna een factor 10 ten opzichte van de DS. Bij $n \geq 50$ is dit zelfs groter dan een factor 10 voor $\lambda_p = \sqrt{2\log(p)}$ en $\lambda_p = \sqrt{2\log(n)}$. Bij kleinere waarden van n zorgt de Gauss-DS niet voor een aanzienlijke winst, waarschijnlijk komt dit door een combinatie van te veel false positives en false negatives.

5.5.5 Discussie Veel conclusies die zojuist zijn gegeven, zijn afhankelijk van de geteste setting. Voor andere soorten designmatrices X zal λ_p veranderen. Dit zal effect op de efficiëntie van de gebruikte schatters en de verhoudingen $S : n : p$.

Toch is er veel behaald met de simulaties. Zo is er meer inzicht verkregen in de keuzes voor λ_p en is aangetoond dat de DS en de Gauss-DS zeker goede schatters voor β kunnen zijn wanneer $n < p$. Ook in het klassieke geval, $n > p$, kan de Dantzig Selector gebruikt worden als schatter voor β.

6

Conclusie

Wanneer gekeken wordt naar het lineaire model $y = X\beta + \varepsilon$ met n metingen en p parameters, dan is het is sommige gevallen mogelijk om de parametervector te schatten als $n < p$. De voorgaande zin is met enige voorzichtigheid geschreven, in veel gevallen is het schatten van de parametervector nog steeds onmogelijk.

In het bachelorproject is er gekeken naar de Dantzig Selector, een schatter voor β indien deze sparse is en $n < p$ kan zijn. De Dantzig Selector kijkt naar alle vectoren die 'bijna' voldoen aan de normaalvergelijkingen en kiest van die vectoren die vector uit, waarvan de ℓ_1-norm minimaal is. Het minimaliseren op de ℓ_1-norm zorgt er voor, dat coëfficiënten op 0 gezet worden; hierdoor worden de 'juiste' parameters geselecteerd.

Wanneer de designmatrix X voldoet aan de restricted isometrie eigenschap en wanneer λ_p goed gekozen is, dan werkt de Dantzig Selector. Stelling 9 laat zelfs zien dat de DS dan zelfs heel 'goed' werkt. Ten opzichte van de ideale schatter moet de DS maar een factor $2\log(p)$ inleveren[1].

Toch moeten er enige kanttekeningen geplaatst worden. Het kan heel goed voorkomen dat X al niet voldoet aan $\delta_2(X) < 1$, waardoor een unieke sparse representatie voor y niet gegarandeerd kan worden. In dat geval, is het onmogelijk om de parametervector te schatten. Verder is het lastig om de restricted isometrie constanten $\delta_S(X)$ te bepalen als p heel groot is; dit kan dan te rekenintensief zijn.

De Dantzig Selector leent zich goed voor speciale toepassingen, waarbij de designmatrix X zelf in te stellen is (calibreren) en waarbij zeker aangenomen mag worden, dat de parametervector sparse is.

Een resultaat uit de simulatie dat voor mij verrassend was, was dat $n = 50$ metingen schijnbaar genoeg is om β te kunnen reconstrueren met de Gauss-DS. Dit terwijl het aantal parameters $p = 100$ en $S = 10$.

Al met al is de Dantzig Selector een goede schatter voor β wanneer X voldoet aan de restricted isometrie eigenschap en β sparse is. Wanneer er voldoende metingen zijn, kan zelfs geprobeerd worden om de Dantzig Selector te gebruiken als orakel voor de Gauss-DS schatter.

[1] Hierbij is $\lambda_p = \sqrt{2\log(p)}$ gekozen, deze grens kan wellicht nog scherper gemaakt worden.

A

Voorkennis

A.1 Norm

In het gehele verslag wordt vaak gebruik gemaakt van verschillende normen. Een herhaling van het begrip norm kan daarom geen kwaad.

Definitie 11 (Norm)**.** Een *norm* op een vectorruimte V over een lichaam K is een afbeelding $\|\| : V \to \mathbb{R}_{\geq 0}$ zodanig dat

1. $\|v\| > 0$ voor alle $v \in V$ en $v \neq 0$;
2. $\|\lambda v\| = |\lambda| \|v\|$ voor alle $v \in V$ en $\lambda \in \mathbb{R}$;
3. $\|v + w\| \leq \|v\| + \|w\|$ voor alle $v, w \in V$ (driehoeksongelijkheid).

De veelgebruikte normen in dit verslag zijn $\|\|_{\ell_\infty}$ en $\|\beta\|_{\ell_p} = \sqrt[p]{\sum_i (\beta_i)^p}$:

$$\|\beta\|_{\ell_1} = \sum_i |\beta_i|$$
$$\|\beta\|_{\ell_2} = \sqrt{\sum_i (\beta_i)^2}$$
$$\|\beta\|_{\ell_\infty} = \sup_i \{\beta_i\}$$

In veel artikelen wordt er ook gerefereerd naar de ℓ_0-'norm'. Merk op dat dit geen norm is omdat niet voldaan wordt aan de tweede eis: $\|\lambda v\| = |\lambda| \|v\|$ voor alle $v \in V$ en $\lambda \in \mathbb{R}$.

A.2 Normale matrix

Definitie 12 (Normale matrix)**.** Een $n \times n$ matrix A heet normaal als $AA^* = A^* A$.

Een reële matrix is dus normaal als $AA^T = A^T A$.

A.3 Rayleigh quotiënt

Definitie 13 (Rayleigh quotiënt)**.** Zij A een normale matrix. Het quotiënt

$$R_A(x) = \frac{\langle x, Ax\rangle}{\langle x, x\rangle}$$

dat voor $x \neq \mathbf{0}$ gedefinieerd is heet het *Rayleigh quotiënt* van A.

Merk op dat het bereik van het Rayleigh quotient het reële interval $[\lambda_{\min}, \lambda_{\max}]$ is, waarbij $\lambda_{\min} \leq \ldots \leq \lambda_{\max}$ de eigenwaarden zijn van A.

A.4 Ongelijkheid van Cauchy

Stelling 14 (Ongelijkheid van Cauchy)**.** *Laat $n \in \mathbb{N}_{\geq 1}$, en laat $a_1, a_2, \ldots, a_n$ en $b_1, b_2, \ldots, b_n$ twee n-tallen reële getallen zijn. Dan geldt*

$$\sqrt{a_1^2 + a_2^2 + \ldots + a_n^2} \cdot \sqrt{b_1^2 + b_2^2 + \ldots + b_n^2} \geq |a_1 b_1| + |a_2 b_2| + \ldots |a_n b_n|.$$

Gevolg. *Laat $a \in \mathbb{R}^n$. Dan geldt*

$$\|a\|_{\ell_1} \leq \sqrt{n}\, \|a\|_{\ell_2}.$$

Bewijs. Neem $b = \mathbf{1} \in \mathbb{R}^n$. De ongelijkheid van Cauchy geeft

$$\|a\|_{\ell_1} = \sum_{i=1}^{n} |a_i \cdot 1| \leq \sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} 1^2} = \sqrt{n}\, \|a\|_{\ell_2}.$$

□

B

Bewijzen

B.1 Bewijs LP-DS

Stelling 15 (LP-DS)**.** *Neem aan dat de* DS *uniek is. Laat* $(u^*, \hat{\beta}^*)$ *de oplossing van het onderstaande lineaire programma zijn.*

$$f = \min \sum_i u_i \qquad odv \tag{B.1}$$

$$-u \leq \tilde{\beta} \leq u \tag{B.2}$$

$$-\lambda_p \sigma \mathbf{1} \leq X^T (y - X\tilde{\beta}) \leq \lambda_p \sigma \mathbf{1} \tag{B.3}$$

$$u \geq \mathbf{0} \tag{B.4}$$

Dan geldt dat $\hat{\beta}^*$ *de Dantzig Selector* (DS) *is. Hierbij is* $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^p$*. Verder geldt voor alle vectoren* $a, b \in \mathbb{R}^p$*,* $a = \left(a_1, \ldots, a_p\right)^T$ *en* $b = \left(b_1, \ldots, b_p\right)^T$*, dat* $a \leq b$ *dan en slechts dan als* $a_i \leq b_i$ *voor alle* $i \in \mathbb{N}$*,* $1 \leq i \leq p$*.*

Bewijs. Allereerst zal opgemerkt worden dat beide toegelaten gebieden dezelfde verzameling vectoren toelaten (dezelfde $\tilde{\beta}$). Vervolgens wordt aangetoond dat voor elk oplossingspaar $(u^*, \hat{\beta}^*)$ geldt dat $u_i = |\hat{\beta}_i|$ voor alle i. Tot slot wordt aangetoond dat de oplossing van het lineair programma de DS moet zijn.

Laat $(\tilde{u}, \tilde{\beta})$ in het toegelaten gebied van het lineair programma liggen. Voorwaarde (B.3) geeft $-\lambda_p \sigma \leq (X^T (y - X\tilde{\beta}))_i \leq \lambda_p \sigma$ voor elke $i \in \{1, \ldots, p\}$. Omdat dit geldt voor elke i, betekent dit dat

$$\left\| X^T (y - X\tilde{\beta}) \right\|_{\ell_\infty} = \sup_i |(X^T (y - X\tilde{\beta}))_i| \leq \lambda_p \sigma.$$

De vector $\tilde{\beta}$ ligt dus ook in het toegelaten gebied van de DS. Laat de vector $\tilde{\beta}$ nu in het toegelaten gebied van de DS liggen. De randvoorwaarden (B.2) en (B.4) leggen geen grens op voor $\tilde{\beta}$. Dit betekent dat $\tilde{\beta}$ alleen moet voldoen aan (B.3); in de vorige paragraaf is al aangetoond dat dit zo is. Het lineair programma en de DS kijken dus naar dezelfde verzameling van vectoren $\tilde{\beta}$.

Laat $(u^*, \hat{\beta}^*)$ de oplossing zijn van het lineair programma. Randvoorwaarde (B.2) geeft dat $u_i^* \geq |\hat{\beta}_i^*|$ voor alle i. Er zal aangetoond worden dat $u_i^* > |\hat{\beta}_i^*|$ geen mogelijke

oplossing voor het lineair programma geeft. Stel dat er een $j \in \mathbb{N}$, $1 \leq j \leq p$ bestaat met $u_j^* > |\hat{\beta}_j^*|$.

Bekijk het paar $(u^{**}, \hat{\beta}^*)$ gedefinieerd door

$$u_i^{**} = \begin{cases} u_i^* & i \neq j \\ |\beta_i^*| & i = j \end{cases}.$$

Merk op dat $(u^{**}, \hat{\beta}^*)$ in het toegelaten gebied van het lineair programma ligt. Er geldt

$$\sum_i u_i^{**} = u_j^{**} + \sum_{i \neq j} u_i^{**} = |\beta|_j^* + \sum_{i \neq j} u_i^* < u_j^* + \sum_{i \neq j} u_i^* = \sum_i u_i^*. \tag{B.5}$$

Maar dan moet $(u^{**}, \hat{\beta}^*)$ de oplossing zijn van het lineair programma. Tegenspraak: er bestaat geen $j \in \mathbb{N}$, $1 \leq j \leq p$ met $u_j^* > |\hat{\beta}_j^*|$. Dus geldt voor de oplossing van het lineair programma dat $u_j^* = |\hat{\beta}_j^*|$ alle $j \in \mathbb{N}$, $1 \leq j \leq p$. Voor de oplossing van het lineair programma geldt dus ook dat $\|u^*\|_{\ell_1} = \|\hat{\beta}^*\|_{\ell_1}$.

Stel dat $\|\hat{\beta}^*\|_{\ell_1} < \|\hat{\beta}\|_{\ell_1}$. Dan kan $\hat{\beta}$ per definitie niet de DS zijn, tegenspraak. Stel dat $\|\hat{\beta}^*\|_{\ell_1} > \|\hat{\beta}\|_{\ell_1}$. Bekijk het paar $(\hat{u}, \hat{\beta})$ waarbij $\hat{u}_i = \hat{\beta}_i$ voor alle i. Dan geldt

$$\sum_i \hat{u}_i = \sum_i |\hat{\beta}_i| = \left\|\hat{\beta}\right\|_{\ell_1} < \left\|\hat{\beta}^*\right\|_{\ell_1} = \sum_i |\hat{\beta}_i^*| = \sum_i u_i.$$

Dan zou $(\hat{u}, \hat{\beta})$ de oplossing van het lineair programma moeten zijn, tegenspraak! Dit betekent dat $\|\hat{\beta}^*\|_{\ell_1} = \|\hat{\beta}\|_{\ell_1}$.

Neem aan dat $\hat{\beta}^*$ niet de Dantzig Selector $\hat{\beta}$ is. De aanname dat de DS uniek is geeft met $\|\hat{\beta}^*\|_{\ell_1} = \|\hat{\beta}\|_{\ell_1}$ echter een tegenspraak, $\hat{\beta}^*$ is dus de Dantzig Selector. □

B.2 Equivalentie LASSO en BPDN

In deze paragraaf zal bewezen worden dat voor een gegeven $\lambda \geq 0$ er een $s \geq 0$ bestaat zodanig dat de oplossingen van (LASSO) en (BPDN) gelijk zijn aan elkaar.

B.2.1 *'⇒'* Kies een $\lambda \geq 0$. Laat

$$\hat{\beta}_1 = \arg\min_{\tilde{\beta}} \left\|Y - X\tilde{\beta}\right\|_{\ell_2}^2 + \lambda \left\|\tilde{\beta}\right\|_{\ell_1}. \tag{B.6}$$

Merk op dat λ, $\left\|\hat{\beta}_1\right\|_{\ell_1}$ en $\left\|Y - X\hat{\beta}_1\right\|_{\ell_2}^2$ nu gegeven zijn.

Voor alle $\tilde{\beta} \in \mathbb{R}^p$ geldt

$$\left\|Y - X\hat{\beta}_1\right\|_{\ell_2}^2 + \lambda \left\|\hat{\beta}_1\right\|_{\ell_1} \leq \left\|Y - X\tilde{\beta}\right\|_{\ell_2}^2 + \lambda \left\|\tilde{\beta}\right\|_{\ell_1}. \tag{B.7}$$

De term $\lambda \left\| \tilde{\beta} \right\|_{\ell_1}$ van (B.7) aftrekken en herschrijven geeft

$$\left\| Y - X\hat{\beta}_1 \right\|_{\ell_2}^2 + \lambda \left(\left\| \hat{\beta}_1 \right\|_{\ell_1} - \left\| \tilde{\beta} \right\|_{\ell_1} \right) \leq \left\| Y - X\tilde{\beta} \right\|_{\ell_2}^2 . \tag{B.8}$$

Kies $s = \left\| \hat{\beta}_1 \right\|_{\ell_1}$ en kijk nu naar alle $\tilde{\beta}$ met $\|\tilde{\beta}\|_{\ell_1} \leq s$. Dan geldt voor alle $\tilde{\beta}$ met $\|\tilde{\beta}\|_{\ell_1} \leq s$ dat $\left\| \hat{\beta}_1 \right\|_{\ell_1} - \left\| \tilde{\beta} \right\|_{\ell_1} \geq 0$. Omdat $\lambda \geq 0$ geldt $\lambda \left(\left\| \hat{\beta}_1 \right\|_{\ell_1} - \left\| \tilde{\beta} \right\|_{\ell_1} \right) \geq 0$. Dit toepassen op (B.8) geeft

$$\left\| Y - X\hat{\beta}_1 \right\|_{\ell_2}^2 \leq \left\| Y - X\hat{\beta}_1 \right\|_{\ell_2}^2 + \lambda \left(\left\| \hat{\beta}_1 \right\|_{\ell_1} - \left\| \tilde{\beta} \right\|_{\ell_1} \right) \leq \left\| Y - X\tilde{\beta} \right\|_{\ell_2}^2 . \tag{B.9}$$

Voor alle $\tilde{\beta}$ met $\tilde{\beta} \leq s$ geldt $\left\| Y - X\hat{\beta}_1 \right\|_{\ell_2}^2 \leq \left\| Y - X\tilde{\beta} \right\|_{\ell_2}^2$. Als $s = \left\| \hat{\beta}_1 \right\|_{\ell_1}$ is $\hat{\beta}_1$ dus ook de LASSO schatter.

B.3 Rayleigh quotiënt

Stelling (Rayleigh quotiënt, stelling 2 pagina 22)**.** *Zij X een $n \times p$ matrix en $A \subset J$ met $|A| \leq S$. Laat $\lambda_{\min} \leq \ldots \leq \lambda_{\max}$ de eigenwaarden van $X_A^T X_A$ zijn. Dan geldt*

$$\lambda_{\min} \|c\|^2 \leq \|X_A c\|^2 \leq \lambda_{\max} \|c\|^2 .$$

Bewijs. Laat $\|\|$ de ℓ_2-norm zijn. Merk op dat $X_A^T X_A = (X_A)^T \left(X_A^T\right)^T = \left(X_A^T X_A\right)^T$, de submatrix $X_A^T X_A$ is dus een normale matrix. Dit betekent, dat het Rayleigh quotiënt voor $R_{X_A^T X_A}(c)$ voor $X_A^T X_A$ gedefinieerd is. Voor het Rayleigh quotiënt $R_{X_A^T X_A}(c)$ geldt, dat voor elke vector $c \in \mathbb{R}^n$ geldt, dat

$$R_{X^T X}(c) = \frac{\langle c, X_A^T X_A c \rangle}{\langle c, c \rangle} \in [\lambda_{\min}, \lambda_{\max}],$$

waarbij $\lambda_{\min} \leq \ldots \leq \lambda_{\max}$ de eigenwaarden van $X_A^T X_A$ zijn. Uitschrijven geeft

$$\begin{aligned}
\lambda_{\min} \leq \frac{\langle c, X_A^T X_A c \rangle}{\langle c, c \rangle} &\leq \lambda_{\max} \\
\langle c, c \rangle \lambda_{\min} \leq \langle c, X_A^T X_A c \rangle &\leq \lambda_{\max} \langle c, c \rangle \\
\lambda_{\min} \|c\|^2 \leq c^T X_A^T X_A c &\leq \lambda_{\max} \|c\|^2 \\
\lambda_{\min} \|c\|^2 \leq (X_A c)^T X_A c &\leq \lambda_{\max} \|c\|^2 \\
\lambda_{\min} \|c\|^2 \leq \|X_A c\|^2 &\leq \lambda_{\max} \|c\|^2 .
\end{aligned}$$

□

B.4 Verband RIC en ROC

Stelling (Restricted isometrie constante – restricted orthogonaliteitsconstante, stelling 4 pagina 24)**.** *Voor alle $S, S' \in \mathbb{N}$ met $S + S' \leq p$ geldt $\theta_{S,S'}(X) \leq \delta_{S+S'}(X)$.*

Bewijs. Zij $S, S' \in \mathbb{N}$ met $S + S' \leq p$ willekeurig gegeven. Zij $A \subseteq J$ en $A' \subseteq J$ twee disjuncte verzamelingen met $|A| \leq S$ en $|A'| \leq S'$. Vanwege (3.3) hoeft er alleen gekeken te worden naar alle eenheidsvectoren $e \in \mathbb{R}^{|A|}$ en $e' \in \mathbb{R}^{|A'|}$ met $\|e\|_{\ell_2} = \|e'\|_{\ell_2} = 1$.

Zij $u \in \mathbb{R}^p$, $\|u\|_{\ell_2} = 1$, willekeurig gegeven met als support A. Voor alle $i \in A$ geldt dus $u_i \neq 0$ en voor alle $i \in A^c$ geldt $u_i = 0$. Zij $u' \in \mathbb{R}^p$, $\|u'\|_{\ell_2} = 1$, willekeurig gegeven met als support A'.

Merk op dat $u_A \in \mathbb{R}^{|A|}$ en $u_{A'} \in \mathbb{R}^{|A'|}$ eenheidsvectoren zijn. Laat $B = A \cup A'$. Beschouw de vectoren

$$\begin{aligned} X(u+u') &= X_A u_A + X_{A'} u'_{A'} = X_B (u+u')_B \qquad \text{en} \\ X(u-u') &= X_A u_A - X_{A'} u'_{A'} = X_B (u-u')_B. \end{aligned}$$

Aangezien A en A' disjunct zijn, geldt $|B| = |A| + |A'| \leq S + S'$ en geldt $\|u\|_{\ell_2}^2 + \|u'\|_{\ell_2}^2 = \|u+u'\|_{\ell_2}^2 = \|u-u'\|_{\ell_2}^2 = \|(u+u')_B\|_{\ell_2}^2 = \|(u-u')_B\|_{\ell_2}^2 = 2$. We verkrijgen

$$\begin{aligned} (1-\delta_{S+S'}(X)) \left\|(u+u')_B\right\|_{\ell_2}^2 &\leq \left\|X_B(u+u')_B\right\|_{\ell_2}^2 \leq (1+\delta_{S+S'}(X)) \left\|(u+u')_B\right\|_{\ell_2}^2; \\ (1-\delta_{S+S'}(X)) \left\|(u-u')_B\right\|_{\ell_2}^2 &\leq \left\|X_B(u-u')_B\right\|_{\ell_2}^2 \leq (1+\delta_{S+S'}(X)) \left\|(u-u')_B\right\|_{\ell_2}^2; \\ 2(1-\delta_{S+S'}(X)) &\leq \left\|X_B(u+u')_B\right\|_{\ell_2}^2 \leq 2(1+\delta_{S+S'}(X)); \qquad \text{(B.10)} \\ 2(1-\delta_{S+S'}(X)) &\leq \left\|X_B(u-u')_B\right\|_{\ell_2}^2 \leq 2(1+\delta_{S+S'}(X)). \qquad \text{(B.11)} \end{aligned}$$

Laat f en g twee vectoren zijn. De eigenschappen van het inproduct geven ons

$$\begin{aligned} \langle f+g, f+g\rangle - \langle f-g, f-g\rangle &= \langle f,f\rangle + 2\langle f,g\rangle + \langle g,g\rangle - \langle f,f\rangle + 2\langle f,g\rangle - \langle g,g\rangle \\ &= 4\langle f,g\rangle; \\ \|f+g\|_{\ell_2}^2 - \|f-g\|_{\ell_2}^2 &= 4\langle f,g\rangle. \qquad \text{(B.12)} \end{aligned}$$

Laat $a, b, c \in \mathbb{R}$ met

$$\begin{aligned} 2-a < b < 2+a; \\ 2-a < c < 2+a. \end{aligned}$$

Dan geldt $|b-c| < 2a$. Doordat $|b-c| = \max\{b-c, c-b\}$ kunnen we zonder beperking

der algemeenheid kijken naar $b-c$. Er geldt

$$\begin{aligned} -2+a &> -c > -2-a; \\ -2-a &< -c < -2+a; \\ 2-a &< b < 2+a; \\ -2a &< b+-c < 2a; \\ |b-c| &< 2a \end{aligned} \tag{B.13}$$

Het toepassen van (B.12) en (B.13) op (B.11) geeft

$$\begin{aligned} \left| \left\| X_B(u+u')_B \right\|_{\ell_2}^2 - \left\| X_B((u-u')_B) \right\|_{\ell_2}^2 \right| &\leq 4\delta_{S+S'}(X); \\ \left| \left\| X_A u_A + X_{A'} u'_{A'} \right\|_{\ell_2}^2 - \left\| X_A u_A - X_{A'} u'_{A'} \right\|_{\ell_2}^2 \right| &\leq 4\delta_{S+S'}(X); \qquad \text{en dus geldt} \\ \left| \left\langle X_A u, X_{A'} u' \right\rangle \right| &\leq \delta_{S+S'}(X) \end{aligned} \tag{B.14}$$

Aangezien A, A', u en u' willekeurig gekozen waren, geldt voor alle disjuncte verzamelingen $A, A' \subseteq J$ met $|A| \leq S$ en $|A'| \leq S'$ en alle vectoren $u_A \in \mathbb{R}^{|A|}$ en $u'_{A'} \in \mathbb{R}^{|A'|}$ met $\|u_A\|_{\ell_2} = \|u'_{A'}\|_{\ell_2} = 1$ dat

$$\left| \left\langle X_A u_A, X_{A'} u'_{A'} \right\rangle \right| \leq \delta_{S+S'}(X).$$

De definitie van de restricted orthogonaliteitsconstante geeft nu $\theta_{S,S'}(X) \leq \delta_{S+S'}(X)$. □

C

Bewijs: Kwaliteit Dantzig Selector

C.1 Kwaliteitsgarantie DS

Stelling (Kwaliteitsgarantie DS, stelling 9 pagina 28)**.** *Zij $\beta \in \mathbb{R}^p$ een S-sparse vector en X een matrix die voldoet aan $\delta(X)_{2S} + \theta(X)_{S,2S} < 1$. Laat $\sigma = 1$ en $\hat{\beta} = \hat{\beta}_{\text{DS}}$ de Dantzig Selector zijn. Onder de orthogonaliteitsconditie geldt*

$$\left\|\beta - \hat{\beta}_{\text{DS}}\right\|_{\ell_2}^2 \leq \left(\frac{4}{1 - \delta(X)_S - \theta(X)_{S,2S}}\right)^2 \lambda_p^2 \cdot S\sigma^2.$$

C.1.1 Belangrijke ongelijkheid 1: afleiding van (C.4) Neem de orthogonaliteitsconditie aan. Laat h de vector zijn met $\hat{\beta} = \beta + h$. Door het bewijs uit [13] over uniciteit bij ℓ_1 optimalisatie te volgen, zal afgeleid worden dat $\|h_{I^c}\|_{\ell_1} \leq \|h_I\|_{\ell_1}$.

Uit de orthogonaliteitsconditie volgt $\left\|\hat{\beta}\right\|_{\ell_1} \leq \|\beta\|_{\ell_1}$. Merk op, dat uit $\left\|\hat{\beta}\right\|_{\ell_1} \leq \|\beta\|_{\ell_1}$ en $\hat{\beta} = \beta + h$ gesteld kan worden, dat

$$0 \leq \|\beta\|_{\ell_1} - \|\beta + h\|_{\ell_1}. \tag{C.1}$$

Gebruik maken van de definitie van de ℓ_1-norm, het herschikken van de sommatie en gebruik maken van het feit dat $\beta_i = 0$ voor $i \in I^c$, geeft

$$\begin{aligned}
\|\beta\|_{\ell_1} - \|\beta + h\|_{\ell_1} &= \sum_{i=1}^{p} |\beta_i| - \sum_{i=1}^{p} |\beta_i + h_i| \\
&= \sum_{i=1}^{p} \left(|\beta_i| - |\beta_i + h_i|\right) \\
&= \sum_{i \in I} \left(|\beta_i| - |\beta_i + h_i|\right) + \sum_{i \in I^c} \left(|\beta_i| - |\beta_i + h_i|\right) \\
&= \sum_{i \in I} \left(|\beta_i| - |\beta_i + h_i|\right) + \sum_{i \in I^c} \left(|0| - |0 + h_i|\right) \\
&= \sum_{i \in I} \left(|\beta_i| - |\beta_i + h_i|\right) + \sum_{i \in I^c} -|h_i| \\
&= \sum_{i \in I} \left(|\beta_i| - |\beta_i + h_i|\right) - \sum_{i \in I^c} |h_i|. \qquad \text{(C.2)}
\end{aligned}$$

Het toepassen van de driehoeksongelijkheid op $|\beta_i| = |\beta_i + h_i - h_i|$ geeft de ongelijkheden

$$\begin{aligned} |\beta_i| = |\beta_i + h_i - h_i| \leq |\beta_i + h_i| + |-h_i| = |\beta_i + h_i| + |h_i|, \\ |\beta_i| - |\beta_i + h_i| \leq |h_i|. \end{aligned} \tag{C.3}$$

Het toepassen van ongelijkheid (C.3) op de eerste sommatie-term in (C.2) geeft

$$\begin{aligned} \|\beta\|_{\ell_1} - \|\beta + h\|_{\ell_1} &= \sum_{i \in I} \left(|\beta_i| - |\beta_i + h_i|\right) - \sum_{i \in I^c} |h_i| \\ &\leq \sum_{i \in I} |h_i| - \sum_{i \in I^c} |h_i|. \end{aligned}$$

Door nu gebruik te maken van ongelijkheid (C.1), geeft dit

$$\begin{aligned} 0 \leq \|\beta\|_{\ell_1} - \|\beta + h\|_{\ell_1} \leq \sum_{i \in I} |h_i| - \sum_{i \in I^c} |h_i|, \\ \sum_{i \in I^c} |h_i| \leq \sum_{i \in I} |h_i|. \end{aligned}$$

Merk op dat dit herschreven kan worden naar

$$\|h_{I^c}\|_{\ell_1} \leq \|h_I\|_{\ell_1}. \tag{C.4}$$

C.1.2 Belangrijke ongelijkheid 2: afleiding van (C.5) Voor alle $\tilde{\beta}$ uit het toegestane gebied van de DS geldt $\left\|X^T\left(y - X\tilde{\beta}\right)\right\|_{\ell_\infty} \leq \lambda_p$. Onder de orthogonaliteitsconditie geldt dat β en $\hat{\beta}$ in het toegestane gebied van de DS liggen. Merk op dat

$$\begin{aligned} X^T X h &= X^T X\left(\hat{\beta} - \beta\right) \\ &= X^T\left(X\hat{\beta} - X\beta\right) \\ &= X^T\left(X\hat{\beta} - y + y - X\beta\right) \\ &= X^T\left(\left(X\hat{\beta} - y\right) + \left(y - X\beta\right)\right) \\ &= -X^T\left(y - X\hat{\beta}\right) + X^T\left(y - X\beta\right) \end{aligned}$$

De driehoeksongelijkheid hierop toepassen levert

$$\begin{aligned} \left\|X^T X h\right\|_{\ell_\infty} &= \left\|-X^T\left(y - X\hat{\beta}\right) + X^T\left(y - X\beta\right)\right\|_{\ell_\infty} \\ &\leq \left\|X^T\left(y - X\hat{\beta}\right)\right\|_{\ell_\infty} + \left\|X^T\left(y - X\beta\right)\right\|_{\ell_\infty} \\ &\leq \lambda_p + \lambda_p = 2\lambda_p, \\ \left\|X^T X h\right\|_{\ell_\infty} &\leq 2\lambda_p. \end{aligned} \tag{C.5}$$

C.1.3 Gebruik lemma Het laatste gedeelte van het bewijs maakt gebruik van het onderstaande lemma. In deze subparagraaf zal het bewijs afgemaakt worden, het lemma wordt bewezen in paragraaf C.2.

Lemma. *Zij I een verzameling met $|I| = S$ en $\delta(X)_{2S} + \theta(X)_{S,2S} < 1$. Laat I_1 de index-verzameling zijn die correspondeert met de S grootste posities van h buiten I. Laat $I_2 = I \cup I_1$. Dan geldt*

$$\left\|h_{I_2}\right\|_{\ell_2} \leq \frac{1}{1-\delta_{2S}(X)} \left\|X_{I_2}^T X h\right\|_{\ell_2} + \frac{\theta_{S,2S}(X)}{(1-\delta_{2S}(X))\sqrt{S}} \|h_{I^c}\|_{\ell_1} \tag{C.6}$$

en

$$\|h\|_{\ell_2}^2 \leq \left\|h_{I_2}\right\|_{\ell_2}^2 + \frac{1}{S}\|h_{I^c}\|_{\ell_1}^2. \tag{C.7}$$

Ongelijkheid (C.4) gecombineerd met de ongelijkheid van Cauchy geeft

$$\|h_{I^c}\|_{\ell_1} \leq \|h_I\|_{\ell_1} \leq \sqrt{S}\|h_I\|_{\ell_2} \leq \sqrt{S}\left\|h_{I_2}\right\|_{\ell_2}. \tag{C.8}$$

Merk op dat een bovengrens van

$$\left\|X_{I_2}^T X h\right\|_{\ell_2} = \sqrt{\sum_{i\in I_2} \left(X^T X h\right)_i^2}$$

gevonden kan worden door gebruik te maken van (C.5). We verkrijgen

$$\begin{aligned}
\left\|X_{I_2}^T X h\right\|_{\ell_2} &= \sqrt{\sum_{i\in I_2} \left(X^T X h\right)_i^2} \\
&\leq \sqrt{\sum_{i\in I_2} \left(\left\|X^T X h\right\|_{\ell_\infty}\right)^2} \\
&\leq \sqrt{2S\cdot\left(\left\|X^T X h\right\|_{\ell_\infty}\right)^2} \\
&\leq \sqrt{2S}\cdot\left\|X^T X h\right\|_{\ell_\infty} \\
&\leq \sqrt{2S}\cdot 2\lambda_p.
\end{aligned} \tag{C.9}$$

Door gebruik te maken van (C.6), (C.8) en (C.9) kan via recursie een bovengrens aan $\|h_{I_2}\|_{\ell_2}$ gegeven worden. Het combineren van (C.6), (C.8) en (C.9) geeft

$$\left\|h_{I_2}\right\|_{\ell_2} \leq \frac{1}{1-\delta_S(X)}\sqrt{2S}\cdot 2\lambda_p + \frac{\theta(X)_{S,2S}}{(1-\delta_S(X))\sqrt{S}}\sqrt{S}\left\|h_{I_2}\right\|_{\ell_2}. \tag{C.10}$$

Herschikken van (C.10) geeft

$$\left(1 - \frac{\theta(X)_{S,2S}}{(1-\delta_S(X))}\right)\left\|h_{I_2}\right\|_{\ell_2} \leq \frac{1}{1-\delta_S(X)}\sqrt{2S}\cdot 2\lambda_p. \tag{C.11}$$

Aangezien $\delta_S(X) + \theta(X)_{S,2S}(X) < 1$ geldt $1-\delta_S(X) - \theta(X)_{S,2S}(X) > 0$ en ook $1-\delta_S(X) > 0$. Dit geeft

$$1 - \frac{\theta(X)_{S,2S}}{1-\delta_S(X)} = \frac{1-\delta_S(X)}{1-\delta_S(X)} - \frac{\theta(X)_{S,2S}}{1-\delta_S(X)} = \frac{1-\delta_S(X)-\theta(X)_{S,2S}(X)}{1-\delta_S(X)} > 0 \tag{C.12}$$

Het toepassen van (C.12) in (C.11) geeft

$$\left\|h_{I_2}\right\|_{\ell_2} \leq \frac{1}{1-\delta_S(X)-\theta(X)_{S,2S}(X)}\sqrt{2S}\cdot 2\lambda_p. \tag{C.13}$$

Merk op dat een bovengrens gezocht wordt voor $\left\|\hat{\beta}-\beta\right\|_{\ell_2}^2 = \|h\|_{\ell_2}^2$. Ongelijkheid (C.7) uit het lemma kan in dit geval gebruikt worden voor het stellen van een bovengrens. Ongelijkheid (C.8) geeft

$$0 \leq \|h_{I^c}\|_{\ell_1} \leq \sqrt{S}\left\|h_{I_2}\right\|_{\ell_2}, \tag{C.8}$$

$$\frac{1}{S}\|h_{I^c}\|_{\ell_1}^2 \leq \left\|h_{I_2}\right\|_{\ell_2}^2. \tag{C.14}$$

Substitutie van (C.14) in (C.7) geeft $\|h\|_{\ell_2}^2 \leq 2\left\|h_{I_2}\right\|_{\ell_2}^2$. Invullen van (C.13) laat zien, dat nu voldaan is aan dat wat te bewijzen was:

$$\begin{aligned}\|h\|_{\ell_2}^2 &\leq 2\left\|h_{I_2}\right\|_{\ell_2}^2 \\ &\leq 2\cdot\left(\frac{1^2}{\left(1-\delta_S(X)-\theta(X)_{S,2S}\right)^2}\cdot 2S\cdot 2^2\lambda_p^2\right), \\ &\leq \left(\frac{4}{\left(1-\delta_S(X)-\theta(X)_{S,2S}\right)}\right)^2\cdot S\cdot\lambda_p^2.\end{aligned}$$

Tot slot de substitutie $h=\hat{\beta}-\beta$. Merk op dat $\|\hat{\beta}-\beta\|_{\ell_2}^2 = \|\beta-\hat{\beta}\|_{\ell_2}^2$,

$$\left\|\beta-\hat{\beta}\right\|_{\ell_2}^2 \leq \left(\frac{4}{\left(1-\delta_S(X)-\theta(X)_{S,2S}\right)}\right)^2\cdot S\cdot\lambda_p^2.$$

De stelling is bewezen.

C.2 Bewijs van het lemma

Lemma. *Zij I een verzameling met $|I|=S$ en $\delta(X)_{2S}+\theta(X)_{S,2S}<1$. Laat I_1 de index-verzameling zijn die correspondeert met de S grootste posities van h buiten I. Laat $I_2 = I\cup I_1$. Dan geldt*

$$\left\|h_{I_2}\right\|_{\ell_2} \leq \frac{1}{1-\delta_{2S}(X)}\left\|X_{I_2}^T X h\right\|_{\ell_2} + \frac{\theta_{S,2S}(X)}{(1-\delta_{2S}(X))\sqrt{S}}\|h_{I^c}\|_{\ell_1} \tag{C.6}$$

en

$$\|h\|_{\ell_2}^2 \leq \left\|h_{I_2}\right\|_{\ell_2}^2 + \frac{1}{S}\|h_{I^c}\|_{\ell_1}^2. \tag{C.7}$$

Het lemma wordt bewezen zoals dit in [8] is gedaan. Deze paragraaf werkt het bewijs echter iets meer uit dan dat het in [8] gedaan is, verder is de notatie gewijzigd naar de in dit verslag gebruikte notatie.

Bewijs. Bekijk de afbeelding $f : R^{|I_2|} \to \mathbb{R}^n$ gegeven door $f(c) = X_{I_2} c = \sum_{j \in I_2} c_j X_j$ (restrictie). Noem $W \subseteq \mathbb{R}^n$ de kolomruimte van X_{I_2}, dan is $\mathrm{im}(X_{I_2}) = \ker(X_{I_2}^T) = W^\perp$, waarbij $W \oplus W^\perp = \mathbb{R}^n$. Aangezien per definitie geldt dat $\delta_{2S}(X) \geq 0$ en $\theta_{S,2S}(X) \geq 0$, geeft het gegeven $\delta_{2S}(X) + \theta_{S,2S}(X) < 1$ dat $\delta_{2S}(X) < 1$. Voor alle vectoren $c \in \mathbb{R}^{|I_2|}$ geldt

$$0 < (1 - \delta_S(X)) \|c\|_{\ell_2}^2 \leq \left\| X_{I_2} c \right\|_{\ell_2}^2 \leq (1 + \delta_S(X)) \|c\|_{\ell_2}^2 .$$

Een gevolg hiervan is dat voor alle vectoren $c \in \mathbb{R}^{|I_2|}$ ook het onderstaande geldt

$$\sqrt{1 - \delta_S(X)} \|c\|_{\ell_2} \leq \left\| X_{I_2} c \right\|_{\ell_2} \leq \sqrt{1 + \delta_S(X)} \|c\|_{\ell_2} . \tag{C.15}$$

Laat de afbeelding $P_W : \mathbb{R}^n \to \mathbb{R}^n$ de orthogonale projectie van $\mathbb{R}^n$ op W zijn. Aangezien $\ker(X_{I_2}^T) = W^\perp$ geldt voor alle $z \in \mathbb{R}^n$ dat $X_{I_2}^T z = X_{I_2}^T P_W z$. Immers, als $z \in W$ dan geldt $P_W z = z$; als $z \in W^\perp$ dan geldt dat $P_W z = \mathbf{0} \in \ker(X_{I_2}^T)$, zowel z als $P_W z$ liggen dan in $\ker(X_{I_2}^T)$. Dit betekent dat voor alle $z \in \mathbb{R}^n$ geldt

$$\begin{aligned} \sqrt{1 - \delta_S(X)} \|P_W z\|_{\ell_2} \leq \left\| X_{I_2}^T P_W z \right\|_{\ell_2} \leq \sqrt{1 + \delta_S(X)} \|P_W z\|_{\ell_2} ; \\ \sqrt{1 - \delta_S(X)} \|P_W z\|_{\ell_2} \leq \left\| X_{I_2}^T z \right\|_{\ell_2} \leq \sqrt{1 + \delta_S(X)} \|P_W z\|_{\ell_2} . \end{aligned} \tag{C.16}$$

Aangezien C.16 geldt voor alle vectoren $z \in \mathbb{R}^n$, geldt dit ook voor Xh. Merk op dat we geïnteresseerd zijn in een ondergrens voor $X_{I_2}^T Xh$. Aangezien $1 - \delta_S(X) > 0$ is gegeven, dat

$$\|P_W Xh\|_{\ell_2} \leq \frac{\left\| X_{I_2}^T Xh \right\|_{\ell_2}}{\sqrt{1 - \delta_S(X)}} = (1 - \delta_S(X))^{-1/2} \left\| X_{I_2}^T Xh \right\|_{\ell_2} . \tag{C.17}$$

Nu zal een ondergrens voor $\|P_W Xh\|_{\ell_2}$ gegeven worden. Dit wordt gedaan door de verzameling I^c op te splitsen in verzamelingen A_j van grootte S.

Sorteer de elementen van h_{I^c} van groot naar klein genummerd $k_1, \ldots, k_{p-|I|}$; het grootste element van h_{I^c} is dus h_{k_1}. Laat $I = A_0$ en laat

$$A_j = \{k_i \in \mathbb{N} : (j-1)S + 1 \leq i \leq jS \text{ en } 1 \leq i \leq p - |I|\}.$$

Merk op dat $A_1 = I_1$ de verzameling indices is van de S grootste coëfficienten van h buiten $I = A_0$. De verzameling A_2 bevat de indices van de S grootste coëfficienten van h buiten A_0 en A_1; A_2 bevat dus de indices van de S grootste coëfficienten van h buiten $I_2 = A_0 \cup A_1$, enzovoort.

Splits $P_W Xh$ nu op als

$$\begin{aligned} P_W Xh &= P_W X_{A_0} h_{A_0} + P_W X_{A_1} h_{A_1} + \sum_{j \geq 2} P_W X_{A_j} h_{A_j} \\ &= P_W X_{I_2} h_{I_2} + \sum_{j \geq 2} P_W X_{A_j} h_{A_j} . \end{aligned} \tag{C.18}$$

Per definitie geldt $X_{I_2} h_{I_2} \in W$, dit geeft $P_W X_{I_2} h_{I_2} = X_{I_2} h_{I_2}$. Omdat $\delta_{2S}(X) < 1$

$$\begin{aligned}(1-\delta_{2S}(X)) \left\| h_{I_2} \right\|_{\ell_2}^2 &\leq \| P_W X_{I_2} h_{I_2} \|_{\ell_2}^2 = \| X_{I_2} h_{I_2} \|_{\ell_2}^2; \\ \sqrt{1-\delta_{2S}(X)} \left\| h_{I_2} \right\|_{\ell_2} &\leq \| P_W X_{I_2} h_{I_2} \|_{\ell_2}. \end{aligned} \tag{C.19}$$

De lineaire afbeelding P_W is een orthogonale projectie dan en slechts dan als $P_W P_W = P_W^T = P_W$. Dit betekent, dat

$$\begin{aligned}\left\| P_W X_{A_j} h_{A_j} \right\|_{\ell_2}^2 &= \left\langle P_W X_{A_j} h_{A_j}, P_W X_{A_j} h_{A_j} \right\rangle; \\ &= \left\langle P_W^T P_W X_{A_j} h_{A_j}, X_{A_j} h_{A_j} \right\rangle; \\ &= \left\langle P_W X_{A_j} h_{A_j}, X_{A_j} h_{A_j} \right\rangle. \end{aligned} \tag{C.20}$$

Omdat $P_W X_{A_j} h_{A_j}$ de orthogonale projectie is op de kolomruimte van X_{I_2}, bestaat er een lineaire combinatie $\sum_{j \in I_2} c_j X_j$ zodanig dat $P_W X_{A_j} h_{A_j} = \sum_{j \in I_2} c_j X_j$. Er bestaat dus een vector $c \in \mathbb{R}^{|I_2|}$ zodanig dat $P_W X_{A_j} h_{A_j} = X_{I_2} c$. Het gegeven $\theta_{S,2S}(X)$ gebruiken geeft

$$\left\langle P_W X_{A_j} h_{A_j}, X_{A_j} h_{A_j} \right\rangle = \left\langle X_{I_2} c, X_{A_j} h_{A_j} \right\rangle \leq \theta_{S,2S}(X) \| c \|_{\ell_2} \| h_{A_j} \|_{\ell_2}. \tag{C.21}$$

Het toepassen van (C.20) en de restricted isometrie op (C.21) geeft

$$\left\| P_W X_{A_j} h_{A_j} \right\|_{\ell_2}^2 \leq \frac{\theta_{S,2S}(X)}{\sqrt{1-\delta_S(X)}} \left\| P_W X_{A_j} h_{A_j} \right\|_{\ell_2} \left\| h_{A_j} \right\|_{\ell_2}. \tag{C.22}$$

Omdat $\| P_W X_{A_j} \| \geq 0$ geldt

$$\left\| P_W X_{A_j} h_{A_j} \right\|_{\ell_2} \leq \frac{\theta_{S,2S}(X)}{\sqrt{1-\delta_S(X)}} \left\| h_{A_j} \right\|_{\ell_2}. \tag{C.23}$$

Merk op, dat

$$\| \sum_{j \geq 2} P_W X_{A_j} \|_{\ell_2} \leq \sum_{j \geq 2} \| P_W X_{A_j} \|_{\ell_2} \leq \frac{\theta_{S,2S}(X)}{\sqrt{1-\delta_S(X)}} \sum_{j \geq 2} \left\| h_{A_j} \right\|_{\ell_2}. \tag{C.24}$$

Door een bovengrens te geven voor $\| h_{A_j} \|_{\ell_2}$, kan een bovengrens voor $\| P_W X_{A_j} h_{A_j} \|_{\ell_2}$ gegeven worden; dit wordt nu gedaan.

Merk op dat voor alle elementen uit $h_{A_{j+1}}$ geldt, dat ze kleiner of gelijk zijn aan de elementen van h_{A_j}. Ze zijn dus ook kleiner of gelijk aan het gemiddelde van de elementen uit h_{A_j}. Op de laatste verzameling na geldt voor alle verzamelingen A_j dat deze per definitie S elementen heeft. Dit betekent, dat voor elke i in A_{j+1} geldt

$$|h_i| \leq \frac{\| h_{A_j} \|_{\ell_1}}{S}.$$

Het toepassen van dit gegeven geeft

$$0 \leq \|h_{A_{j+1}}\|_{\ell_2}^2 = \sum_{i \in A_{j+1}} h_i^2 \leq S \cdot \frac{\|h_{A_j}\|_{\ell_1}^2}{S^2} = \frac{\|h_{A_j}\|_{\ell_1}^2}{S};$$

$$\|h_{A_{j+1}}\|_{\ell_2} \leq \frac{\|h_{A_j}\|_{\ell_1}}{\sqrt{S}}. \tag{C.25}$$

Substitutie van (C.25) in (C.24) geeft

$$\begin{aligned} \|\sum_{j\geq 2} P_W X_{A_j}\|_{\ell_2} &\leq \frac{\theta_{S,2S}(X)}{\sqrt{1-\delta_S(X)}} \sum_{j\geq 2} \left\|h_{A_j}\right\|_{\ell_2} \\ &\leq \frac{\theta_{S,2S}(X)}{\sqrt{1-\delta_S(X)}} \sum_{j\geq 1} \left\|h_{A_j+1}\right\|_{\ell_2} \\ &\leq \frac{\theta_{S,2S}(X)}{\sqrt{1-\delta_S(X)}} \sum_{j\geq 1} \frac{\|h_{A_j}\|_{\ell_1}}{\sqrt{S}}. \end{aligned} \tag{C.26}$$

Merk op, dat $\sum_{j\geq 1} \|h_{A_j}\|_{\ell_1}$ een sommatie is over alle elementen van I^c. Dit betekent, dat (C.26) omgeschreven kan worden naar

$$\|\sum_{j\geq 2} P_W X_{A_j}\|_{\ell_2} \leq \frac{\theta_{S,2S}(X)}{\sqrt{S}\cdot\sqrt{1-\delta_S(X)}} \|h_{I^c}\|_{\ell_1}. \tag{C.27}$$

Het toepassen van (C.19) en (C.17) op (C.27) geeft

$$\begin{aligned} &\sqrt{1-\delta_{2S}(X)}\left\|h_{I_2}\right\|_{\ell_2} - \frac{\theta_{S,2S}(X)}{\sqrt{S}\cdot\sqrt{1-\delta_S(X)}}\|h_{I^c}\|_{\ell_1} \leq \|P_W X h\|_{\ell_2} \leq (1-\delta_S(X))^{-1/2}\left\|X_{I_2}^T X h\right\|_{\ell_2}; \\ &\sqrt{1-\delta_{2S}(X)}\left\|h_{I_2}\right\|_{\ell_2} - \frac{\theta_{S,2S}(X)}{\sqrt{S}\cdot\sqrt{1-\delta_S(X)}}\|h_{I^c}\|_{\ell_1} \leq (1-\delta_S(X))^{-1/2}\left\|X_{I_2}^T X h\right\|_{\ell_2}. \end{aligned} \tag{C.28}$$

Herschikken en een vermenigvuldiging met $0 < (1-\delta_{2S}(X))^{-1/2}$ bewijst de eerste ongelijkheid uit het lemma:

$$\left\|h_{I_2}\right\|_{\ell_2} \leq \frac{1}{1-\delta_S(X)}\left\|X_{I_2}^T X h\right\|_{\ell_2} + \frac{\theta_{S,2S}(X)}{\sqrt{S}\cdot(1-\delta_S(X))}\|h_{I^c}\|_{\ell_1}. \tag{C.29}$$

De afleiding van de tweede ongelijkheid is een stuk korter. Merk op dat

$$\|h\|_{\ell_2}^2 \leq \|h_{I_2}\|_{\ell_2}^2 + \|h_{I_2^c}\|_{\ell_2}^2. \tag{C.30}$$

Wanneer de gevraagde bovengrens voor $\|h_{I_2^c}\|_{\ell_2}^2$ is gevonden, is het lemma bewezen.

De rankschikking $k_1, \ldots, k_{p-|I|}$ van de elementen van h_{I^c} zorgt ervoor dat voor elk element geldt, dat deze kleiner of gelijk is aan het gemiddelde van zijn voorgangers en zichzelf. Voor elke $i \in \mathbb{N}, 1 \leq i \leq p-|S|$ geldt dus

$$\begin{aligned} |h_{k_i}| &\leq \frac{1}{i}\sum_{i\geq k}|h_{k_i}| \leq \frac{1}{i}\sum_{i\in I^c}|h_{k_i}| = \frac{\|h_{I^c}\|_{\ell_1}}{i}; \\ h_{k_i}^2 &\leq \frac{1}{i^2}\|h_{I^c}\|_{\ell_1}^2. \end{aligned} \tag{C.31}$$

Dit betekent, dat

$$\begin{aligned}
\|h_{I_2^c}\|_{\ell_2}^2 = \sum_{i\in I_2^c} h_{k_i}^2 &\leq \|h_{I^c}\|_{\ell_1}^2 \sum_{i\in I_2^c} \frac{1}{i^2} \\
&\leq \|h_{I^c}\|_{\ell_1}^2 \sum_{i\geq S+1} \frac{1}{i^2} \\
&\leq \|h_{I^c}\|_{\ell_1}^2 \int_S^\infty \frac{1}{x^2}\mathrm{d}x \\
&\leq \|h_{I^c}\|_{\ell_1}^2 \lim_{m\to\infty} \left[\frac{-1}{x}\right]_{x=S}^{x=m} \\
&\leq \|h_{I^c}\|_{\ell_1}^2 \cdot \frac{1}{S}. \qquad \text{(C.32)}
\end{aligned}$$

Het toepassen van (C.32) op (C.30) geeft

$$\|h\|_{\ell_2}^2 \leq \|h_{I_2}\|_{\ell_2}^2 + \frac{\|h_{I^c}\|_{\ell_1}^2}{S},$$

hetgeen te bewijzen was. □

D

Bewijs: Kwaliteit orakelschatter

Stelling (Kwaliteitsgarantie orakelschatter, stelling 10 pagina 32)**.** *Neem aan dat I, de support van β, gegeven is. Neem verder aan dat $\delta_{2S}(X) < 1$ en $|I| = S$. Laat $\beta^\star$ de orakelschatter zijn, $\beta^\star$ is de schatter voor β zijn wanneer I gegeven is. Dan geldt*

$$\mathbb{E}\left[\left\|\beta^\star - \beta\right\|_{\ell_2}^2\right] \geq \frac{S\sigma^2}{1+\delta_{2S}(X)}.$$

Bewijs. Allereerst moet opgemerkt worden, dat $X_I^T X_I$ inverteerbaar is. Omdat $|I| = S < 2S$, geeft de stelling over het Rayleigh quotiënt ons, dat de eigenwaarden van $X_I^T X_I$ in het interval $[1-\delta_{2S}(X), 1+\delta_{2S}(X)$ liggen. Dit betekent dat 0 geen eigenwaarde is van $X_I^T X_I$; de matrix is dus inverteerbaar.

Gebruikmakende van (4.6), $\beta_I^\star = \left(X_I^T X_I\right)^{-1} X_I^T y$, en (4.5), $y = X_I\beta_I + \varepsilon$, en een paar elementaire eigenschappen uit de lineaire algebra verkrijgen we

$$\begin{aligned}
\beta_I^\star &= \left(X_I^T X_I\right)^{-1} X_I^T y \\
&= \left(X_I^T X_I\right)^{-1} X_I^T \left(X_I\beta_I + \varepsilon\right) \\
&= \left(X_I^T X_I\right)^{-1} \left(X_I^T X_I\beta_I + X_I^T \varepsilon\right) \\
&= \left(X_I^T X_I\right)^{-1} \left(X_I^T X_I\right)\beta_I + \left(X_I^T X_I\right)^{-1} X_I^T \varepsilon \\
&= \beta_I + \left(X_I^T X_I\right)^{-1} X_I^T \varepsilon. \qquad \text{(D.1)}
\end{aligned}$$

Vergelijking (D.1) geeft $\beta_I^\star - \beta_I = \left(X_I^T X_I\right)^{-1} X_I^T \varepsilon$. Om $\mathbb{E}\left[\left\|\beta^\star - \beta\right\|_{\ell_2}^2\right]$ anders uit te drukken, is het handig om de ℓ_2 norm uit te schrijven.

$$\begin{aligned}
\mathbb{E}\left[\left\|\beta^\star - \beta\right\|_{\ell_2}^2\right] &= \mathbb{E}\left[\left(\sqrt{\sum_{i=1}^{p}\left(\beta^\star - \beta\right)^2}\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{p}\left(\beta^\star - \beta\right)^2\right] \qquad \text{(D.2)}
\end{aligned}$$

Door op te merken dat $\beta_i^\star = \beta_i = 0$ voor $i \in I^c$ geldt

$$\begin{aligned}\mathbb{E}\left[\left\|\beta^\star - \beta\right\|_{\ell_2}^2\right] &= \mathbb{E}\left[\sum_{i\in I}\left(\beta_i^\star - \beta_i\right)^2 + \sum_{i\in I^c}\left(\beta_i^\star - \beta_i\right)^2\right] \\ &= \mathbb{E}\left[\sum_{i\in I}\left(\beta_i^\star - \beta_i\right)^2\right] \\ &= \mathbb{E}\left[\left\|\beta_I^\star - \beta_I\right\|_{\ell_2}^2\right]. \end{aligned} \tag{D.3}$$

Merk op dat $\left\|\beta_I^\star - \beta_I\right\|_{\ell_2}^2 = \left\langle \beta_I^\star - \beta_I, \beta_I^\star - \beta_I \right\rangle$. Door (D.1), (D.3) en het zojuist genoemde te gebruiken, verkrijgen we

$$\begin{aligned}\mathbb{E}\left[\left\|\beta^\star - \beta\right\|_{\ell_2}^2\right] &= \mathbb{E}\left[\left\langle \beta_I^\star - \beta_I, \beta_I^\star - \beta_I \right\rangle\right] \\ &= \mathbb{E}\left[\left(\left(X_I^T X_I\right)^{-1} X_I^T \varepsilon\right)^T \left(X_I^T X_I\right)^{-1} X_I^T \varepsilon\right] \\ &= \mathbb{E}\left[\varepsilon^T X_I \left(\left(X_I^T X_I\right)^{-1}\right)^T \left(X_I^T X_I\right)^{-1} X_I^T \varepsilon\right] \end{aligned} \tag{D.4}$$

Merk op dat $\left\|\beta^\star - \beta\right\|_{\ell_2}^2$ een getal is. Door gebruik te maken van de eigenschappen van het spoor van een matrix, zal de laatste stap van het bewijs gegeven worden. Het spoor van een getal is het getal zelf, $\mathrm{Tr}\left(\left\|\beta^\star - \beta\right\|_{\ell_2}^2\right) = \left\|\beta^\star - \beta\right\|_{\ell_2}^2$; dit geeft

$$\mathbb{E}\left[\left\|\beta^\star - \beta\right\|_{\ell_2}^2\right] = \mathbb{E}\left[\mathrm{Tr}\left(\left\|\beta^\star - \beta\right\|_{\ell_2}^2\right)\right]. \tag{D.5}$$

Laat A een $n \times S$ matrix zijn en B een $S \times n$ matrix. Merk op dat

$$\mathrm{Tr}\left(AB\right) = \sum_{i=1}^{n}\sum_{j=1}^{S} A_{ij}B_{ji} = \mathrm{Tr}\left(BA\right). \tag{D.6}$$

Substitutie van (D.4) in (D.5) met het toepassen van (D.6) geeft

$$\begin{aligned}\mathbb{E}\left[\left\|\beta^\star - \beta\right\|_{\ell_2}^2\right] &= \mathbb{E}\left[\mathrm{Tr}\left(\varepsilon^T X_I \left(\left(X_I^T X_I\right)^{-1}\right)^T \left(X_I^T X_I\right)^{-1} X_I^T \varepsilon\right)\right] \\ &= \mathbb{E}\left[\mathrm{Tr}\left(\left(X_I^T X_I\right)^{-1} X_I^T \varepsilon\varepsilon^T X_I \left(\left(X_I^T X_I\right)^{-1}\right)^T\right)\right] \end{aligned} \tag{D.7}$$

Lineariteit van de verwachting geeft

$$\begin{aligned}\mathbb{E}\left[\mathrm{Tr}\left(\left(X_I^T X_I\right)^{-1} X_I^T \varepsilon\varepsilon^T X_I \left(\left(X_I^T X_I\right)^{-1}\right)^T\right)\right] &= \mathrm{Tr}\left(\left(X_I^T X_I\right)^{-1} X_I^T \mathbb{E}\left[\varepsilon\varepsilon^T\right] X_I \left(\left(X_I^T X_I\right)^{-1}\right)^T\right) \\ &= \mathrm{Tr}\left(\left(X_I^T X_I\right)^{-1} X_I^T \sigma^2 I_n X_I \left(\left(X_I^T X_I\right)^{-1}\right)^T\right) \\ &= \sigma^2 \mathrm{Tr}\left(\left(X_I^T X_I\right)^{-1} X_I^T X_I \left(\left(X_I^T X_I\right)^{-1}\right)^T\right) \\ &= \sigma^2 \mathrm{Tr}\left(\left(\left(X_I^T X_I\right)^{-1}\right)^T\right) \end{aligned} \tag{D.8}$$

Aangezien $\mathrm{Tr}(A) = \mathrm{Tr}\left(A^T\right)$, geeft D.8

$$\mathbb{E}\left[\left\|\beta^\star - \beta\right\|_{\ell_2}^2\right] = \sigma^2 \mathrm{Tr}\left(\left(X_I^T X_I\right)^{-1}\right). \tag{D.9}$$

De definitie van de restricted isometrie constante geeft dat de eigenwaarden van $X_I^T X_I$ liggen in $[1-\delta_{2S}(X), 1+\delta_{2S}(X)]$. De matrix $X_I^T X_I$ is dus een symmetrische matrix met positieve eigenwaarden $(1-\delta_{2S}(X) > 0)$. Dit betekent echter dat de eigenwaarden van $(X_I^T X_I)^{-1}$ in het interval $[\frac{1}{1+\delta_{2S}(X)}, \frac{1}{1-\delta_{2S}(X)}]$ liggen. Aangezien het spoor van een matrix gelijk is aan de som van zijn eigenwaarden en $X_I^T X_I$ een $S \times S$ matrix is, geeft dit

$$\mathrm{Tr}\left(\left(X_I^T X_I\right)^{-1}\right) \geq S \cdot \lambda_{\min} = \frac{S}{1+\delta_{2S}(X)}. \tag{D.10}$$

Ongelijkheid (D.10) toepassen op (D.9) geeft

$$\mathbb{E}\left[\left\|\beta^\star - \beta\right\|_{\ell_2}^2\right] \geq \frac{S\sigma^2}{1+\delta_{2S}(X)}.$$

□

E

Gebruikte code

E.1 Dantzig Selector

Het onderstaande scriptje is een functie die bij een gegeven y, X, $\lambda_p\sigma$ en p de Dantzig Selector uitrekent. Het scriptje maakt gebruik van het MATLAB® programma `linprog`.

```
function beta_hat = DS(X,y,p,grens)

f = zeros(1,2*p);
f(p+(1:p)) = 1;

A = zeros(4*p,2*p);
A(1:p,1:p) = X'*X;
A(p+(1:p),1:p) = -X'*X;
A(2*p+(1:p),1:p) = eye(p);
A(2*p+(1:p),p+(1:p)) = -1*eye(p);
A(3*p+(1:p),1:p) = -1*eye(p);
A(3*p+(1:p),p+(1:p)) = -1*eye(p);

bovengrens = zeros(4*p,1);
bovengrens(1:p) = grens + X'*y;
bovengrens((1:p)+p) = grens - X'*y;
bovengrens((1:p)+2*p) = 0;
bovengrens((1:p)+3*p) = 0;

% oplossing = linprog(f,A,bovengrens);
options = optimset('TolFun',1e-4,'MaxIter',500);
oplossing = linprog(f,A,bovengrens,[],[],[],[],[],
    options);
beta_hat = oplossing(1:p);

% options = optimset('LargeScale', 'off', 'Simplex', '
    on');
```

```
% options = optimset('Display','iter','TolFun',1e-2);
% options = optimset('Display','iter','MaxIter',7);
% options = optimset('Display','iter','TolFun',1e-8,'
    MaxIter',20);
% oplossing = linprog(f,A,bovengrens,[],[],[],[],[],
    options);
% options = optimset('Display','iter','TolFun',1e-4,'
    MaxIter',7);
```

E.2 Kleinste kwadratenmethode

De onderstaande functie wordt gebruikt om de kleinste kwadratenschatter uit te rekenen. Deze functie wordt ook aangeroepen om de orakelschatter $\beta^\star$ en de Gauss-DS schatter uit te rekenen.

```
function beta_hat = least_squares(X,y)
beta_hat = (X'*X)\(X'*y);
```

E.3 Simulatie

Dit bestand bevat de code die gebruikt is om de simulatie uit te voeren. Enkele regels voor het schrijven van matrices naar een tekstbestand zijn uit deze code gehaald, zodat de code overzichtelijk blijft. De namen van de variabelen komen overeen met de gebruikte namen in het bachelorproject, zo is S de cardinaliteit van de support I. De enige uitzondering op de naamgeving is `SSd`, waarmee $\|\beta - \hat{\beta}\|_{\ell_2}^2$ wordt bedoeld. De variabele `pad` dient ook voor eigen gemak en het automatiseren van het opslaan van verschillende resultaten.

```
clc
clear all
close all

tic
%  %  %  PARAMETERS
p=100; n=50; grootteS = 10; %n = 20# 30# 40# 50# 60#
    80#  100#  150#
N = 1000;

for n = [20, 30, 40, 50, 60, 80, 100, 150]
pad = ['/home/wikash/Documenten/Bachelorproject/v2/
    Matlab/Resultaten'];
```

```
12 latexpad = [pad '/LaTeX/simulatie_N-p-n-S_' num2str(N)
       '-' num2str(p) '-' num2str(n) '-' num2str(grootteS
   ) '_ldp=sqrt2logn_' '_' date '.mat'];
13
14 SSd = zeros(N,3);
15 false_negative_I_DS = zeros(N,1);
16 false_positive_I_DS = zeros(N,1);
17 ldp = zeros(N,1);
18
19 for i = 1:N
20         i
21         %  %  %  2. Genereer X en normaliseer kolommen
22         sigma = (1/3)*sqrt(grootteS/n);
23         X = normrnd(0,sigma,n,p);
24         for j = 1:p
25                 X(:,j) = X(:,j) / norm(X(:,j));
26         end
27
28         %  %  %  5. Genereren beta_origineel en beta
29         beta_origineel = zeros(p,1);
30         beta_origineel(1:grootteS) = 100*randn(
              grootteS,1);
31         beta_origineel = beta_origineel(randperm(
              length(beta_origineel))); % shuffle
32
33         I_origineel = beta_origineel~=0;
34         S_origineel = sum(I_origineel);
35
36         %  %  %  6. Bereken y
37         y = X*beta_origineel + normrnd(0,sigma,n,1);
38
39         % % % Orakel voorspelling
40         beta_hat_orakel =  zeros(p,1);
41         beta_hat_orakel(I_origineel) = least_squares(X
              (:,I_origineel),y);
42
43         %  %  %  4. Bepalen labda_p
44         labda_p = sqrt(2*log(n));
45 %       labda_p = 2*log(p);
46 %       labda_p = 0;
47 %       for j = 1:1000
48 %               mc_z = normrnd(0,1,n,1);
49 %               winnaars(j) = max(X'*mc_z);
50 %       end
51 %       labda_p = mean(winnaars);
52         ldp(i) = labda_p;
```

```
53          grens = labda_p*sigma;
54
55          % % % 7. Bepaal beta_hat via de DS
56          beta_hat_DS = DS(X,y,p,grens);
57          I_hat_DS = abs(beta_hat_DS) > sigma;
58          S_hat_DS = sum(I_hat_DS);
59
60          % % % 8. Bepaal de Gauss-DS
61          beta_hat_gaussDS = zeros(p,1);
62          beta_hat_gaussDS(I_hat_DS) = least_squares(X
                (:,I_hat_DS),y);
63          I_hat_gaussDS = abs(beta_hat_gaussDS) > sigma;
64          S_hat_gaussDS = sum(I_hat_gaussDS);
65
66          % % % RESULTATEN
67          SSd_orakel = sum((beta_hat_orakel -
                beta_origineel).^2)/p;
68          SSd_DS = sum((beta_hat_DS - beta_origineel)
                .^2)/p;
69          SSd_gaussDS = sum((beta_hat_gaussDS -
                beta_origineel).^2)/p;
70          SSd(i,:) = [SSd_orakel SSd_DS SSd_gaussDS];
71
72          false_negative_I_DS(i) =  (sum(I_origineel(
                I_hat_DS == 0))/p)*100;
73          false_positive_I_DS(i) = (sum(I_hat_DS(
                I_origineel == 0))/p)* 100;
74  end
75  toc
76
77  % % % Verwerking van de resultaten
78  mean_SSd = mean(SSd);
79  std_SSd = std(SSd);
80
81  result_SSd = zeros(3,3);
82  breedte_96_conf_interval_SSd = std_SSd*1.96/(sqrt(N));
        % breedte_96_conf_interval = std(SSd)*norminv
        (0.975)/(sqrt(N));
83  result_SSd(1,:) = mean_SSd -
        breedte_96_conf_interval_SSd;
84  result_SSd(2,:) = mean_SSd;
85  result_SSd(3,:) = mean_SSd +
        breedte_96_conf_interval_SSd;
86
87  mean_false_negative_I_DS = mean(false_negative_I_DS);
88  std_false_negative_I_DS = std(false_negative_I_DS);
```

```
89  breedte_96_conf_interval_false_negative_I_DS =
       std_false_negative_I_DS*1.96/(sqrt(N));
90  result_false_negative_I_DS(1) =
       mean_false_negative_I_DS -
       breedte_96_conf_interval_false_negative_I_DS;
91  result_false_negative_I_DS(2) =
       mean_false_negative_I_DS;
92  result_false_negative_I_DS(3) =
       mean_false_negative_I_DS +
       breedte_96_conf_interval_false_negative_I_DS;
93
94  mean_false_positive_I_DS = mean(false_positive_I_DS);
95  std_false_positive_I_DS = std(false_positive_I_DS);
96  breedte_96_conf_interval_false_positive_I_DS =
       std_false_positive_I_DS*1.96/(sqrt(N));
97  result_false_positive_I_DS(1) =
       mean_false_positive_I_DS -
       breedte_96_conf_interval_false_positive_I_DS;
98  result_false_positive_I_DS(2) =
       mean_false_positive_I_DS;
99  result_false_positive_I_DS(3) =
       mean_false_positive_I_DS +
       breedte_96_conf_interval_false_positive_I_DS;
100
101 mean_ldp = mean(ldp);
102 std_ldp = std(ldp);
103 breedte_96_conf_interval_ldp = std_ldp*1.96/(sqrt(N));
104 result_ldp(1) = mean_ldp -
       breedte_96_conf_interval_ldp;
105 result_ldp(2) = mean_ldp;
106 result_ldp(3) = mean_ldp +
       breedte_96_conf_interval_ldp;
107
108 %  %  %  Opslaan van de resultaten
109 latex_result = zeros(19,1);
110 latex_result(1) = n;
111 latex_result(1 + 0*3 + (1:3)) = result_SSd(:,1);
112 latex_result(1 + 1*3 + (1:3)) = result_SSd(:,2);
113 latex_result(1 + 2*3 + (1:3)) = result_SSd(:,3);
114 latex_result(1 + 3*3 + (1:3)) =
       result_false_negative_I_DS;
115 latex_result(1 + 4*3 + (1:3)) =
       result_false_positive_I_DS;
116 latex_result(1 + 5*3 + (1:3)) = result_ldp;
117 save(latexpad,'latex_result')
118 end
```

Bibliografie

[1] BICKEL, P. J. Discussion: The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics 35*, 6 (December 2007), 2352 – 2357.

[2] CANDÈS, E. J. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique 346*, 9 – 10 (May 2008), 589 – 592.

[3] CANDÈS, E. J., AND PLAN, Y. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics 37*, 5A (June 2009), 2145–2177.

[4] CANDÈS, E. J., ROMBERG, J., AND TAO, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory 52*, 2 (February 2006), 489 – 509.

[5] CANDÈS, E. J., ROMBERG, J., AND TAO, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics 59*, 8 (August 2006), 1207–1223.

[6] CANDÈS, E. J., AND TAO, T. Decoding by linear programming. *IEEE Transactions on Information Theory 51*, 12 (December 2005), 4203 – 4215.

[7] CANDÈS, E. J., AND TAO, T. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory 52*, 12 (December 2006), 5406–5425.

[8] CANDÈS, E. J., AND TAO, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics 35*, 6 (December 2007), 2313 – 2351.

[9] CANDÈS, E. J., AND TAO, T. Rejoinder: The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics 35*, 6 (December 2007), 2392 – 2404.

[10] CANDÈS, E. J., AND WAKIN, M. B. An introduction to compressive sampling. *IEEE Signal Processing Magazine 25*, 2 (March 2008), 21 – 30.

[11] CHEN, S. S., DONOHO, D. L., AND SAUNDERS, M. A. Atomic decomposition by basis pursuit. *Society for Industrial and Applied Mathematics 43*, 1 (2001), 129–159.

[12] DONOHO, D. L. For most large underdetermined systems of linear equations, the minimal ℓ_1 norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics 59*, 6 (June 2006), 797–829.

[13] DONOHO, D. L., AND HUO, X. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory 47*, 47 (2001), 2845 – 2862.

[14] EFRON, B., HASTIE, T., JOHNSTONE, I., AND TIBSHIRANI, R. Least angle regression. *Annals of Statistics 32*, 2 (April 2004), 407–499.

[15] HESTERBERG, T., CHOI, N. H., MEIER, L., AND FRALEY, C. Least angle and ℓ_1 penalized regression: A review. *Statistics Surveys 2* (February 2008), 61 – 93.

[16] LANGE, N. Statistical procedures for functional MRI. *Medical Radiology – Diagnostic Imaging and Radiation Oncology : Functional MRI*, 27 (1999), 301–335.

[17] LUSTIG, M., DONOHO, D. L., AND PAULY, J. M. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine 58*, 6 (October 2007), 1182 – 1195.

[18] OSBORNE, M. R., PRESNELL, B., AND TURLACH, B. A. On the LASSO and its dual. *Journal of Computational and Graphical Statistics 9*, 2 (June 2000), 319–337.

[19] RITOV, Y. Discussion: The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics 35*, 6 (December 2007), 2370 – 2372.

[20] TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 1 (1996), 267 – 288.

[21] ZHANG, Y. On theory of compressive sensing via ℓ_1-minimization: Simple derivations and extensions. CAAM Technical Report TR08-11, Rice University, Rice University, Houston, Texas, 77005., September 2008 2008.