Statistical integration of diverse omics data

Bouhaddani, S. el

Samenvatting

De afgelopen decennia hebben de biomedische meettechnieken een sterke ontwikkeling doorgemaakt, waardoor we in staat zijn om op grote schaal verschillende 'omics' (biologische datasets) datatypes te meten. Een simultane statistische analyse van alle data stelt ons in staat het onderliggende biologische mechanisme beter te begrijpen. Echter, deze simultane aanpak is niet zonder statistische uitdagingen. De belangrijkste uitdagingen zijn ten eerste een complexe correlatiestructuur binnen en tussen elk datatype, en ten tweede dat deze data vaak veel meer variabelen bevat dan het aantal afgenomen monsters. Door deze uitdagingen kunnen traditionele methoden, zoals het multivariate lineair model, niet worden toegepast. Andere uitdagingen zijn systematische verschillen in de grootte, schaal, verdeling en type meettechniek tussen de datasets. Dit proefschrift gaat over omics data-integratie als een holistische benadering om meerdere omics datasets gezamenlijk te analyseren. In het eerste deel bestuderen we data-integratiemethoden en hun toepasbaarheid op meerdere omics data gemeten in cohortstudies. In het tweede deel stellen we een probabilistisch kader voor om zowel de relaties als de systematische verschillen tussen twee omics datasets te modelleren.

Het eerste hoofdstuk geeft een inleiding tot de huidige data-integratiemethodiek en beschrijft enige uitdagingen bij het schatten van relaties tussen datasets. Eerst presenteren we kenmerken die gemeenschappelijk en specifiek zijn voor de omics data. Vervolgens worden verschillende omics data-integratie benaderingen geïntroduceerd, deze worden verenigd in het Structural Equation Modeling (SEM) kader. In dit kader kunnen eigenschappen zoals sterke correlaties en hoge dimensionaliteit (zeer groot aantal variabelen) worden omvat. Standaard modelleren SEM's geen heterogeniteit tussen omics datasets, daarom kijken we naar uitbreidingen van het SEM raamwerk die ook dataspecifieke delen in het model bevatten. De huidige data-integratiebenaderingen verschillen vooral in de parametrisering van de residuele variantietermen. Strategieën om de parameters in een SEM te schatten, vallen (grofweg) in twee categorieën: Least Squares (kleinste kwadraten) en Maximum Likelihood (maximale waarschijnlijkheid). De laatste categorie heeft de voorkeur, want het vermindert het risico op overfitten en kan makkelijker worden uitgebreid om complexe afhankelijkheidsstructuren en *supervised* omics data-integratie mee te nemen in het formuleren van het model (zoals in hoofdstuk 6). In beide strategieën vormen hoog-dimensionale data een uitdaging voor de schattingsmethode. We beschrijven identificeerbaarheid van de parameters van het SEM-model en maken hierbij natuurlijke aannames. Ten slotte is op het gebied van omics data-integratie statistische software nodig om het gebruik en de ontwikkeling van nieuwe methodologie te vergemakkelijken. Het hoofdstuk eindigt met uitdagingen van de ontwikkeling van open-source software.

Hoofdstuk 2 is het eerste van twee hoofdstukken die als doel hebben de bestaande data-integratiemethoden en hun toepasbaarheid op omics data gemeten in cohortstudies te bestuderen. Hiertoe worden transcriptomische en metabolomische data van de DILGOM[2] studie geïntegreerd. Een bepaalde methode die gebaseerd is op Least Squares, Two-way Orthogonal Partial Least Squares (O2PLS), is toegepast om zowel de gezamenlijke als de specifieke delen te schatten. De componenten in deze delen worden stapsgewijs berekend met behulp van een singuliere waarde decompositie in elke stap. Om het optimale aantal componenten te bepalen, stellen we een snel alternatief voor standaard kruis-validatie voor. Via een simulatiestudie wordt de prestatie

[2]DIeet, Levensstijl en Genetische determinanten van Obesitas en Metabool syndroom

van O2PLS geëvalueerd. Hieruit volgde dat een hoog ruisniveau in de data negatieve invloed heeft op de schatting van de specifieke delen, terwijl de gezamenlijke delen grotendeels onveranderd blijven. Het toepassen van O2PLS op de omics datasets in de DILGOM-studie toonde aan dat O2PLS de gezamenlijke delen structureerde in interpreteerbare genetische en metabole componenten. De top genen waren betrokken bij anti-ontstekingsprocessen, en de metabole componenten lieten een clustering van verschillende lipoproteïne-subtypes zien. Deze bevindingen onderschrijven daarnaast de resultaten van eerder uitgevoerde standaardanalyses op deze data.

In het derde hoofdstuk wordt een gratis en open-source softwarepakket, OmicsPLS, geïntroduceerd dat O2PLS in R implementeert samen met verschillende tools, zoals kruis-validatie. Het geheugenefficiënte algoritme kan omgaan met zowel laag- als hoogdimensionale data. Om het gebruik van de software te stimuleren worden de interpretatie van het model en de parameters besproken. Verder worden, om de toepassing van omics-data-integratie te vergemakkelijken, verschillende post-analysefuncties geïmplementeerd: plotfuncties met behulp van het ggplot2-pakket, een textuele samenvatting van de O2PLS *fit* en wrapperfuncties om gemakkelijk toegang te krijgen tot de schattingen. Er is een simulatiestudie uitgevoerd om OmicsPLS te vergelijken met een alternatieve methode wat betreft accuraatheid van de schattingen en snelheid van het algoritme. OmicsPLS was veel sneller en het alternatief convergeerde vaak niet naar een oplossing, wanneer er dataspecifieke delen aanwezig waren. OmicsPLS is toegepast op genetische en glycomische gegevens van het Kroatische Korcula cohort. De gezamenlijke componenten lieten een duidelijke clustering van de moleculaire structuur van de glycanen zien, en de top genen waren betrokken bij de manipulatie en de locatie van de glycoproteïnen.

In de hoofdstukken 4 en 5 wordt een probabilistisch kader voor (hoogdimensionale) omics data-integratie voorgesteld en bestudeerd. Hoofdstuk 4 presenteert het Probabilistisch PLS (PPLS) model voor de relatie tussen twee omics datasets. Het model is gebaseerd op het standaard SEM-kader, waarbij gebruik wordt gemaakt van een normale verdeling voor de latente variabelen. Identificeerbaarheid van het model wordt aangetoond met behulp van natuurlijke restricties over de parameters. Deze parameters worden met Maximum Likelihood geschat met behulp van een EM-algoritme. De natuurlijke restricties worden meegenomen in de M-stap. De EM-stappen kunnen afzonderlijk worden berekend voor de parameters die horen bij de eerste en de tweede dataset. Het EM-algoritme kan geheugenefficiënt worden geïmplementeerd. Een expliciete formule voor de standaardfouten kan worden verkregen door de geobserveerde Fisher-informatiematrix te berekenen. Er is een uitgebreide simulatiestudie uitgevoerd om de prestaties van PPLS in termen van bias en variantie van de schatters te evalueren en te vergelijken met andere methoden. De PPLS schattingen waren erg robuust tegen afwijkingen van de normaliteit en presteerden goed in zowel laag- als hoogdimensionale gegevens. Het PPLS model werd gebruikt om de IgG1- en IgG2-delen van de glycoom-datasets van twee Kroatische cohorten, Korcula en Vis, te integreren. De gezamenlijke componenten in het Korcula cohort lieten clusters van glycaan subtypen zien; deze componenten werden gerepliceerd in het Vis cohort.

Hoofdstuk 5 presenteert een probabilistische herformulering van O2PLS en een uitbreiding van PPLS om gezamenlijke en specifieke kenmerken in twee sets van variabelen te modelleren. Het voorgestelde model, Probabilistic O2PLS (PO2PLS), bevat ook dataspecifieke componenten. Net als bij PPLS zijn de parameters identificeer-

baar door gebruik te maken van natuurlijke aannames. Ze worden met Maximum Likelihood geschat met behulp van een geheugenefficiënt EM-algoritme. Daarnaast worden de standaardfouten berekend. De prestaties van PO2PLS werden geëvalueerd met een uitgebreide simulatiestudie, waarbij de nadruk lag op predictieprestaties en interpreteerbaarheid. Vooral in de scenario's waar de data heterogeen of hoogdimensionaal zijn, presteerde PO2PLS beter dan alternatieve methoden. Met name overfitting was aanzienlijk minder dan met O2PLS. Het PO2PLS model werd toegepast op omics data van twee studies: transcriptomische en metabolomische data van het DILGOM-cohort (zoals in hoofdstuk 2), en genetica en glycomische data van het Korcula-cohort (zoals in hoofdstuk 3). In de DILGOM-studie werden de top genen in elke PO2PLS gezamenlijke component geclusterd in verschillende genetische *pathways* (routes) zonder overlap tussen de componenten, in lijn met de huidige biologische kennis. De pathways behorende bij de top O2PLS genen werden echter gevonden in beide gezamenlijke O2PLS componenten. Verder gebruikten we het Korcula-cohort als training- en het Vis-cohort als testdata. De predictiefout van PO2PLS was kleiner dan van O2PLS.

Het zesde hoofdstuk bespreekt uitbreidingen van het probabilistische data-integratie raamwerk om complexe onderzoeksopzetten op te nemen in het model. In het bijzonder wordt een *supervised* kader in overweging genomen om een uitkomst in omics data-integratie te incorporeren. Ter illustratie wordt in het eerste deel van het hoofdstuk O2PLS toegepast op genetische en epigenetische data uit de Genetic Analysis Workshop (GAW). Deze data werd gemeten met behulp van een longitudinale familiestudieopzet. Het doel is om het effect van deze datasets op de triglycerideniveaus te modelleren. Een aanpak bestaande uit drie stappen is gebruikt: eerst zijn met O2PLS de tijdsafhankelijke delen in de epigenetische data opgevangen door middel van de dataspecifieke componenten. In de tweede stap werden gezamenlijke genetische en tijdsafhankelijke componenten geschat met O2PLS. Tot slot werd de uitkomst gerelateerd aan de componenten door middel van *lineair mixed models* waarbij rekening werd gehouden met correlaties binnen families. De top genen in de gezamenlijke delen werden gekoppeld aan immunologische pathways, terwijl in de genetisch-specifieke delen clustering van families werd waargenomen. Het tweede deel van dit hoofdstuk schetst een nieuw supervised probabilistisch model, als uitbreiding van PO2PLS, om tegelijkertijd de relatie tussen de genetische, epigenetische en triglyceride-data te modelleren. Dit model omvat epigenetische correlaties over tijdspunten en familiaire correlaties tussen de individuen. Elke deel van de totale likelihood kan opnieuw worden ontkoppeld in verschillende delen, waardoor afzonderlijke optimalisaties mogelijk worden voor de parameters in het genetische, epigenetische en triglyceride deel.

Al met al kan het probabilistische modelleringskader worden uitgebreid om de omics data gemeten met verschillende typen studieopzetten te analyseren, en verschillende soorten uitkomsten te integreren. Er is echter meer onderzoek nodig in deze richting, waarbij niet-lineaire relaties of niet-normale uitkomsten in het gezamenlijke model worden opgenomen. Ook kan het probabilistisch kader worden uitgebreid om de relatie tussen meer dan twee datasets tegelijkertijd te modelleren om meer dan twee omics datatypen te analyseren. Samen met gratis en open-source software zal dit de ontwikkeling van nieuwe methoden op het gebied van omics data-integratie bevorderen.