

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/67526> holds various files of this Leiden University dissertation.

**Author:** Kpogbezan, G.B.

**Title:** Prior information and variational Bayes in high dimensional statistical network inference

**Issue Date:** 2018-12-10

# Samenvatting

In dit proefschrift hebben we statistische methoden voor de analyse van hoog-dimensionele data ontwikkeld. In het bijzonder besteden we aandacht aan reconstructie van hoog-dimensionele netwerken. In de genetica is de identificatie van gen-regulerende netwerken cruciaal voor het begrijpen van genfunctie, en derhalve belangrijk zowel in behandeling als voorspelling van ziekten. Hoog-dimensionele netwerkreconstructie is een zeer uitdagende taak aangezien het aantal mogelijke grafen exponentieel groeit met het aantal variabelen (e.g. genen). Echter, een aantal van de verbanden tussen deze variabelen zijn mogelijk bekend vanuit de literatuur. Zo zijn de huidige overtuigingen op het gebied van interacties tussen genen bijvoorbeeld geconcentreerd in kennisbanken als KEGG en Reactome. We introduceren een raamwerk waarin het opnemen van dergelijke *a-priori* informatie in de reconstructie mogelijk wordt op een zachte manier: het informeert de analyse wanneer correct, maar kan gecompenseerd worden indien volledig incompatibel met de data. We behandelen ook de onderwerpen genetische associatie studies (*eQTL mapping*) en data integratie.

In hoofdstuk 1 introduceren we een nieuwe *global-local shrinkage prior* van het type *ridge* voor niet-gerichte netwerk reconstructie gebaseerd op *SEMs* met *a-posteriori* selectie van zijden. De voorgestelde aanpak is computationeel snel en doet het beter dan bekende concurrenten zoals de *graphical lasso*.

In hoofdstuk 2 breiden we hoofdstuk 1 uit door prior informatie toe te voegen in de reconstructie van niet-gerichte netwerken. Er wordt rekening gehouden met deze prior kennis op een zachte manier die de beschikbare data toelaat om de prior informatie te compenseren indien deze niet relevant is. Bovendien is de voorgestelde methode in staat om de overeenstemming tussen de beschikbare data en de prior informatie expliciet te schatten, hetgeen een noviteit is bij het toevoegen van prior informatie in netwerk inferentie.

In hoofdstuk 3 introduceren we een raamwerk voor het simultaan analyseren van meerdere gerelateerde hoog-dimensionele en complexe datasets. Zulke analyses omvatten onder andere gen-regulerende netwerk reconstructies, genetische associatie studies (e.g. *eQTL mapping*) en data integratie in genomica. Om de analyse van kleine  $n$  relatief tot grote  $p$  mogelijk te maken introduceren we de *hoefijzer* prior die *sparsity* toelaat; een gewenste eigenschap voor de analyse van zulke data. We illustreren de procedure met twee toepassingen, namelijk: de reconstructie van gen-regulerende netwerken en *eQTL mapping*.

In hoofdstuk 4 verkennen we diverse methoden die gen-regulerende netwerken reconstrueren door combinatie van observationele (*in vivo*) en temporele cellijn gen-expressie (*in vitro*) data. De dynamiek van de humane cel wordt aangenomen een eerste-orde vector autoregressie VAR(1) model te volgen en er wordt onderzocht hoe de onderliggende model parameters efficient geleerd kunnen worden door gebruik van de twee soorten datasets. We zien in een toepassing op echte data dat de reconstructie

van de conditionele onafhankelijkheidsgraaf significant verbeterd door informatie van de cellijn te lenen. Bovendien lijken onze voorgestelde strategieën om de VAR(1) model parameters te leren consistent in onze data-gedreven analyse en tonen behoud van transcriptionele dynamiek tussen de *in vivo* en *in vitro* omgevingen.