Advanced non-homogeneous dynamic Bayesian network models for statistical analyses of time series data

Shafiee Kamalabad, Mahdi

Samenvatting

Eén van de statistische uitdagingen in veel onderzoeksgebieden is het uit tijdseriedata afleiden van de topologie van netwerken van op elkaar inwerkende eenheden. Een klasse van statistische modellen die veel toegepast wordt om met deze opgave om te gaan, is de klasse van dynamische Bayesiaanse netwerkmodellen (DBN's). De onderliggende aanname van conventionele DBN's is dat het onderliggende proces een homogeen Markov proces is, zodat voor DBN's de netwerkparameters niet mogen veranderen in de tijd. Daardoor kunnen DBN's niet omgaan met niet-homogene en niet-stationaire regelgevende processen, die voorkomen in veel belangrijke toepassingen in de werkelijkheid.

Onlangs zijn niet-homogene dynamische Bayesiaanse netwerkmodellen (NH-DBN's) geïntroduceerd, welke een belangrijk statistisch hulpmiddel zijn geworden om deze beperkende veronderstelling te omzeilen. NH-DBN's zijn geïmplementeerd met verschillende toewijzingsmodellen om de tijdseriedata te verdelen in afzonderlijke dataverzamelingen. Deze modellen leiden de datasegmentatie, de gezamenlijke netwerkstructuur en de segment- of componentspecifieke interactieparameters uit de data af.

In dit proefschrift hebben we ons gefocust op het verbeteren van changepoint (CPS) verdeelde NH-DBN's; dit zijn de NH-DBN's die het meest toegepast worden om complexe systemen te modelleren. Deze modellen leiden changepoints af die de data verdelen in afzonderlijke segmenten en de segmentspecifieke netwerkparameters worden voor elk segment apart geleerd. In veel toepassingen in de werkelijkheid verdelen deze NH-DBN's een tijdserie in nog kortere segmenten. Het leren van de netwerkparameters voor elk segment afzonderlijk ('ontkoppelde' NH-DBN-modellen) leidt tot te grote flexibiliteit en opgedreven inferentieonzekerheden. Bovendien, deze modellen bevatten niet de redelijke aanname vooraf dat naburige segmenten vaak meer kans hebben op vergelijkbare netwerkinteractieparameters dan segmenten ver van elkaar. Om deze knelpunten aan te pakken zijn Bayesiaanse modellen met koppelingsmechanismen tussen de segmentspecifieke parameters voorgesteld.

Bayesiaanse modellen met parameterkoppeling kunnen leiden tot significant verbeterde nauwkeurigheden van netwerkreconstructies wanneer de segmentspecifieke parameters vergelijkbaar zijn. Onlangs hebben we echter ontdekt dat koppeling contraproductief kan worden wanneer de segmentspecifieke parame-

ters verschillend zijn. De reden daarvoor is dat zowel het sequentiële als het globale koppelingsschema geen effectief mechanisme voor ontkoppeling heeft. Dit is voor veel echte toepassingen een beperking. We hebben deze knelpunten in dit proefschrift aangepakt door vier nieuwe NH-DBN's te introduceren.

Een ander scenario, dat overeenkomt met veel werkelijke toepassingen, komt voor wanneer tijdseriedata vaak wordt verzameld onder verschillende experimentele omstandigheden. Hierbij zijn er, in plaats van één enkele tijdserie die verdeeld kan worden in segmenten met een normale tijdsorde, K (korte) tijdseries zonder normale orde. Dit zijn inwisselbare eenheden en er is geen noodzaak om de segmentatie af te leiden. In deze situatie is het a priori vaak onduidelijk of de netwerkparameters werkelijk componentspecifiek zijn of dat zij constant zijn voor alle componenten. In echte toepassingen daarentegen, kunnen beide soorten parameters tegelijkertijd voorkomen. We hebben daarom dit probleem aangepakt door nieuwe gedeeltelijke NH-DBN's te introduceren, gebaseerd op Bayesiaanse regressiemodellen met partitieontwerpmatrix.

In **hoofdstuk 1** hebben we een introductie en een overzicht van de bestaande netwerkmodellen gegeven. Ook hebben we in hoofdlijnen aangegeven waar dit proefschrift over gaat.

In **hoofdstuk 2** hebben we twee nieuwe modellen geboden, gebaseerd op stuksgewijs Bayesiaanse regressiemodellen, namelijk het gedeeltelijk segmentgewijs gekoppeld NH-DBN-model en het gegeneraliseerde, volledig sequentieel gekoppelde model. Onze empirische resultaten laten zien dat het gedeeltelijk gekoppelde model leidt tot een verbeterde nauwkeurigheid van netwerkreconstructies. Voor het gegeneraliseerde gekoppelde model hebben we geen consequente verbeteringen gezien ten opzichte van het volledig gekoppelde NH-DBN-model.

In **hoofdstuk 3** hebben we daarom het gegeneraliseerde, volledig sequentieel gekoppelde model verfijnd. Voor het verfijnde model met een hyperprior distributie op de tweede hyperparameter van de a priori-distributie van de koppelingsparameter zien we een verbetering in de nauwkeurigheid van netwerkreconstructies.

In **hoofdstuk 4** hebben we een nieuw NH-DBN-model gepresenteerd met gedeeltelijk zijdegewijs gekoppelde en segment-specifieke netwerkparameters. Dit model werkt op de individuele zijden. In plaats van *alle* randen te forceren gekoppeld te zijn, werkt ons model zijdegewijs en leidt het voor elke afzonderlijke zijde uit de data af of de bijbehorende parameters gekoppeld moeten worden of juist in alle segmenten ontkoppeld moeten blijven. Dit nieuwe model heeft ook het ongekoppelde en het gekoppelde NH-DBN als limietgevallen. We hebben empirisch aangetoond op gist-genexpressie tijdseries dat het nieuwe model een hoogste nauwkeurigheid van netwerkreconstructie behaald. Voor Arabidopsis thaliana-genexpressiedata laten we zien dat ons nieuwe model niet alleen een netwerkvoorspelling geeft, maar ook onderscheid kan maken tussen zijden waarvan de regulerende effecten gelijk blijven in de tijd en zijden waarvan de regulerende effecten meer substantiële veranderingen in de tijd ondergaan.

In **hoofdstuk 5** hebben we een gedeeltelijk NH-DBN-model geïntroduceerd, dat in feite een Bayesiaans regressiemodel is met partitieontwerpmatrix. Het

nieuwe model beoogt de beste afweging te maken tussen een homogeen model en een niet-homogeen model. Voor elke netwerkinteractie is er een parameter en het nieuwe model leidt uit de data af of deze parameter constant is of tussen segmenten varieert. Daarnaast stellen we voor om een Gaussisch proces gebaseerde benadering te gebruiken om niet-equidistante metingen te kunnen verwerken.

Door dit nieuwe model toe te passen op gistdata hebben we aangetoond dat het de nauwkeurigheid van netwerkreconstructies verbetert. We hebben het nieuwe model gebruikt om de topologieën van de mTORC1-data te reconstrueren. De afgeleide topologieën van netwerken vertoonden kenmerken die overeenkomen met de biologie-literatuur.

Hoofdstuk 6 ging over een vergelijkend evaluatieonderzoek naar populaire niet-homogene Poisson-modellen voor count data. Voor dit onderzoek zijn het standaard homogene Poisson model (HOM) en drie niet-homogene varianten, namelijk een Poisson changepoint-model (CPS), een Poisson free-mixture-model (MIX) en een Poisson hidden-Markov-model (HMM), geïmplementeerd in zowel een frequentistisch als een Bayesiaans kader. Het eerste hoofddoel is het onafhankelijk vergelijken van de prestaties van de vier bovengenoemde modellen voor beide modelleringskaders (Bayesiaans en frequentistisch). Daarna voeren we een paarsgewijze vergelijking uit tussen de vier Bayesiaanse en de vier frequentistische modellen om op te helderen in hoeverre de resultaten van de twee paradigma's ('Bayesiaans versus frequentistisch') verschillen.