

Analyse af blandet data

Af

KAARE BRANDT PETERSEN

Kaare Brandt Petersen er cand. scient. og har læst fysik ved Københavns Universitet. Både bachelorprojekt og speciale blev til i et samarbejde med en afdeling på IMM ved Danmarks Tekniske Universitet, hvor han nu er ved at afslutte sit ph.d. projekt om en teknik til dataanalyse der anvender metoder fra statistisk fysik.

Email: kbp@imm.dtu.dk

Fysik er ikke alene en videnskab om quarker, kometer og komplekse systemer. Det er også et arnested for metoder der finder anvendelser i alverdens forskellige sammenhænge som finansering, hjerneforskning og dataanalyse, der alle tre er eksempler på at fysikkens redskaber nu bliver fundet og genanvendt i langt bredere sammenhænge end de oprindeligt var udviklet til. I denne artikel skal vi se lidt nærmere på et eksempel hvor teknikker fra statistisk fysik har fået en hovedrolle i en metode til analyse af data – en analysemetode til separation af signaler men som også kan fortolkes som en generalisering af en af dataanalysens mest anvendte og hæderkronede redskaber, nemlig Principal Component Analysis (PCA).

I begyndelsen af 1990'erne kom en signalbehandlings teknik frem med betegnelsen "Independent Component Analysis", forkortet ICA. Det er en teknik til såkaldt blind signalseparation, dvs. opgaven om at skille blandede signaler ad uden at vide præcist hvordan de er blevet blandet. Grundstenen i ICA er at foretage denne separation alene ved at antage at de blandede signaler er statistisk uafhængige. ICA har vist sig både interessant som forskningsområde og nyttigt i anvendelser, og har derfor

fået massiv opmærksomhed de seneste ti år. Men inden vi fortæber os i detaljerne, så lad os se på den opgave som ICA er i stand til at løse.

Introduktion

Man skal forestille sig en opgave der lyder: Gæt tallene s_1 og s_2 hvor det eneste man ved er at den vægtede sum er 1.7

$$1.7 = \frac{2}{3}s_1 + \frac{4}{3}s_2$$

Det er naturligvis en umulig opgave hvis vi ikke ved noget mere, men hvis vi ved at tallene s_1, s_2 er trukket uafhængigt fra en fordeling, hvis sandsynlighedsmasse er koncentreret omkring 0, så er gæt hvor s_1, s_2 begge er små meget mere sandsynlige end at de er store og med modsat fortegn. Så selvom ligningen stadig har uendeligt mange løsninger, så gør ekstra statistisk viden at nogle gæt er bedre end andre. Vi kan udvide problemet på en måde der gør det både nemmere og sværere: I stedet for kun een ligning har vi nu to, men til gengæld kan vi forestille os at vi ikke kender koefficienterne for s_1 og s_2 , dvs opgaven er nu at bestemme både \mathbf{A} og \mathbf{s} i ligningen

$$\begin{bmatrix} 1.7 \\ 0.2 \end{bmatrix} = \mathbf{A}\mathbf{s}$$

hvor vi kun kender den fordeling $p(\mathbf{s})$ som vektoren \mathbf{s} er trukket fra. Selvom opgaven minder om den første er den meget vanskeligere fordi vi ikke kender \mathbf{A} . Men hvis vi sørger for at have mere data, dvs ikke bare en enkelt men mange ligninger som den ovenfor

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t \quad t = 1, \dots, N$$

så kan det faktisk lade sig gøre at estimere både matricen \mathbf{A} og vektorerne \mathbf{s}_t . Denne opgave er essentielt hvad ICA er konstrueret til at løse: Kun givet data vektorerne \mathbf{x}_t og fordelingen $p(\mathbf{s})$, finder ICA et godt estimat af \mathbf{A} og \mathbf{s}_t ved at antage at koordinaterne af \mathbf{s} er statistisk uafhængige af hinanden. Matricen \mathbf{A} kaldes *blandingsmatricen* og vektorerne \mathbf{s}_t 's komponenter kaldes for *signalerne*.

Eksempler på anvendelser

Det kan lyde noget restriktivt, at signalerne antages at være statistisk uafhængige, men i overraskende mange anvendelser er det tilstrækkeligt sandt til at ICA virker.

Det klassiske anvendelseseksempel er et cocktailparty: Hver persons stemme betragtes som et signal og fordi der er mange mennesker der taler samtidig, så vil optagelser være en blanding af mange menneskers stemmer. Opgaven for ICA teknikken er derfor at behandle de optagede mikrofonsignaler, så der kommer en og kun en stemme på hver kanal. Man behøver ikke at være i efterretningsbranchen for at se det anvendelige i dette, for brugere af høreapparater klager ofte over at de blandede stemmer er vanskeligere at skille ad gennem et høreapparat end gennem den naturlige høreelse. Dette skyldes at lyden formes omkring hovedet på en ganske særlig måde der giver den naturlig høreelse muligheder for blandt andet bedre stedbestemmelse.

Et andet eksempel er fra hjerneforskning, hvor en person har fået placeret 38 sugekopper med sensorer på hovedet og derefter bliver bedt om at se eller høre en række udvalgte effekter. Det interessante for hjerneforskerne er hvilke dele af hjernen der er aktive under forskellige stimuli, men det er ikke direkte synligt med de optagede signaler. Grunden er, at hver sugekop optager ændringer i det elektromagnetiske felt, men denne samlede ændring ved sugekoppen er sammensat af aktiviteten overalt i hjernen, ligesom det vi hører ved cocktailpartyet er en blanding af mange menneskers stemmer. ICA anvendes derfor på signaler fra sugekopperne og giver forskerne de signaler, som man under modellens antagelser estimerer er hjernecentrenes aktiviteter.

Et sidste eksempel i denne samling er anvendelsen af ICA på et internet chatrum. På tv-stationen CNN's hjemmeside, kan man deltage i en online diskussion i et chatrum og derved give sin mening om et eller flere emner til kende. Andre mennesker svarer på det og der opstår en flertrådet diskussion om den gruppe af emner der interesserer chatrummets brugere. ICA er blevet anvendt på indlæggene på dette chatrum for at undersøge om det var muligt i stikord at bestemme de emner der blev diskuteret og resultatet var positivt. Selvom det ikke i samme grad som i det første

eksempel er klart hvad der er facit, stemmer resultatet af ICA analysen overens med menneskelig vurdering af antallet af emner.

Teknikken bag

De nærmere detaljer om hvordan ICA fungerer kan kun beskrives skitsevist når man ikke har mange sider til sin rådighed. Dette skyldes at ICA ikke er en enkelt teknik, men en samling af teknikker der benytter forskellige metoder men har den samme grundantagelse: At signalerne er statistisk uafhængige. Årsagen til at metoderne kan være så forskellige selvom grundprincippet er ens, er at princippet om statistisk uafhængighed ikke er entydigt kvantificerbart. Der er simpelthen ikke een rigtig måde at måle hvor meget eller lidt uafhængige to signaler er af hinanden og derfor har mange forskellige forskere foreslået mange forskellige modeller. Der er blandt andet forsøgt anvendt en minimering af gensidig information mellem signalerne, hvor information her skal forstås i en snæver informationsteoretisk forstand. Andre anvender entropi og atter andre maximum likelihood estimering af parametrene ud fra en generativ statistisk uafhængig model. Men der er flere forskelle end dette. For de forskellige teknikker benytter også forskellige grundlæggende antagelser om modellen

- **Støj.** De første modeller var uden støj, men det har vist sig nyttigt at antage gaussisk støj for at håndtere real-world data og særlig optagelsessituationer.
- **Tidslig korrelation.** Nogle af modellerne antager at signalerne er tidsligt korrelerede, mens andre har det som en essentiel del at de ikke er tidsligt korrelerede.
- **Completeness.** Forholdet mellem antal af optagelser (dimensioner i \mathbf{x}_t) og signaler (dimensioner i \mathbf{s}_t), spiller en kritisk rolle. Alle teknikker kan håndtere situationen med lige mange signaler som observationer, men kun en mindre del af teknikkerne generaliserer til mere krævende forhold.

- **Signalfordeling.** Den fordeling $p(\mathbf{s})$ som man antager at signalerne er trukket fra, varierer i større eller mindre grad blandt teknikkerne.

Alt i alt er der så stor forskel på ICA teknikkerne at man ikke kan beskrive dem mere præcist end ved de grundlæggende antagelser om signalseparation under statistisk uafhængighed. Men visse teknikker er mere interessante og varierede end andre, og her præsenteres som et eksempel den ICA teknik der har betegnelsen *Mean Field ICA*.

Et eksempel: Mean Field ICA

Som navnet fortæller anvendes der i Mean Field ICA teknikker fra statistisk fysik, men den grundliggende model er en gaussisk støjmodel uden tidsskorrelation i signalerne, dvs

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \mathbf{n}_t, \quad \mathbf{n}_t \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

Costfunktionen i Mean Field ICA, dvs den centrale funktion der forsøges optimeret, er log likelihooden, dvs sandsynligheden for det observerede datasæt, givet et bestemt valg af parametre. I en notation, hvor \mathbf{X} betyder det samlede datasæt og \mathbf{S} den samlede mængde signaldata, er log likelihood'en \mathcal{L} defineret ved

$$\mathcal{L}(\mathbf{A}, \Sigma) = \ln p(\mathbf{X}|\mathbf{A}, \Sigma) = \ln \int p(\mathbf{X}|\mathbf{A}, \Sigma, \mathbf{S})p(\mathbf{S})d\mathbf{S}$$

Likelihooden $\mathcal{L}(\mathbf{A}, \Sigma)$ kan altså ses som (logaritmen til) sandsynligheden for at man, givet et sæt parametre $\{\mathbf{A}, \Sigma\}$ har få observerede datasæt \mathbf{X} . Når man maksimerer likelihooden med hensyn til parametrene så skruer man altså på parametrene indtil man får den størst mulige sandsynlighed for det datasæt man nu en gang har. I den forstand er maximum likelihood estimering det samme som at vælge de mest sandsynlige parametre. Inde i integralet forekommer to andre fordelinger, nemlig $p(\mathbf{X}|\mathbf{A}, \Sigma, \mathbf{S})$ og $p(\mathbf{S})$. Den første er støjfordelingen, dvs en gauss fordeling. Den anden fordeling $p(\mathbf{S})$ er signalernes såkaldte *prior* fordeling, dvs. den fordeling af signalernes værdier \mathbf{S} man må forvente *inden* man har observeret sit datasæt \mathbf{X} . Denne fordeling kan være meget forskellig afhængig af om

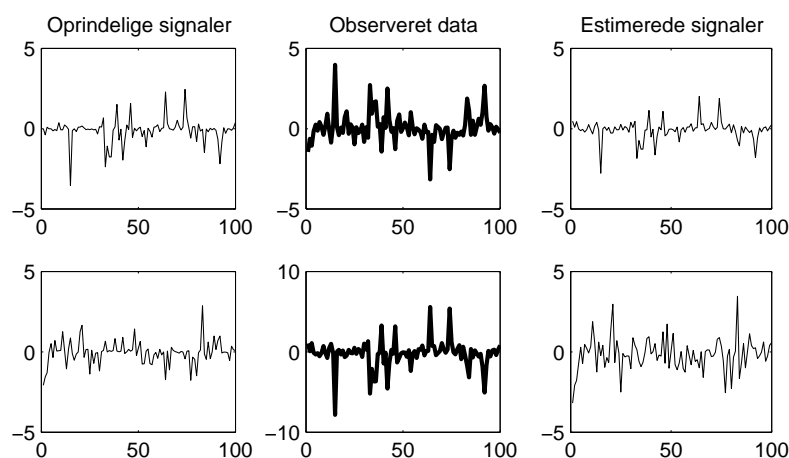
man arbejder med talesignaler, supersoniske vibrationer eller hvad det nu kan være.

Når man differentierer log likelihood med hensyn til parametrene får man at udtrykkene afhænger af middelværdierne $\langle \mathbf{S} \rangle$ og $\langle \mathbf{S}\mathbf{S}^T \rangle$, der er middelværdier med hensyn til fordelingen $p(\mathbf{S}|\mathbf{X})$, kaldet signalernes *posterior* fordeling. Med mindre man gør sig meget restriktive antagelser om fordelingen af signalerne, er denne fordeling ret kompliceret og middelværdierne er derfor tilsvarende vanskelige at bestemme.

Men her kommer ældre hæderkronede teknikker fra statistisk fysik til hjælp. Man kan nemlig med fordel bestemme disse middelværdier ved hjælp af mean field teknikker af varierende kompleksitet. Til laveste orden kan man bruge såkaldt naiv middelfeltsteori, hvor koblingerne erstattes af et ydre felt, og denne approximation kan skærpes ved brug af lineær respons teori eller TAP varianter. Det er også muligt at inddrage stadig mere avancerede teknikker som Bethe approximation der inddrager parvise koblinger eller Kikuchi der inddrager endnu flere højere ordens effekter. Ikke altid er afvejningen mellem den øgede kompleksitet og præcision i estimatet tilfredsstillende, og man må derfor lede efter den rigtige grænse i en given problemstilling. Med Mean Field teknikken til at bestemme middelværdierne, kan vi kombinere delene til en samlet algoritme der kører i to trin:

1. Estimer $\langle \mathbf{S} \rangle$ og $\langle \mathbf{S}\mathbf{S}^T \rangle$ med mean field metoder, givet et fastholdt foreløbigt estimat af \mathbf{A} og Σ
2. Maksimer log likelihooden med hensyn til \mathbf{A} og Σ , givet det seneste estimat på $\langle \mathbf{S} \rangle$ og $\langle \mathbf{S}\mathbf{S}^T \rangle$.

Denne trinvis algoritme har betegnelsen Expectation-Maximization (EM) og har attraktive egenskaber med hensyn til implementering af analytiske udtryk, men er til gengæld ofte meget længe om at konvergere. Den samlede algoritme fungerer under gode forhold, som man kan se i eksemplet.



Opsummering

Det *kan* faktisk lade sig gøre med ICA at estimere både blandingsforholdene og signalerne alene på baggrund af antagelsen om statistisk uafhængighed. Der er mange ICA teknikker på banen og nogle af dem er begyndt at trække på erfaringerne inden for statistisk fysik til at løse de stadig mere komplicerede opgaver der ligger i vejen for en tilfredsstillende løsning.

Fronten for ICA forskningen går nu ved udvidelser til signaler med tidsligt ekko og ikkelineære blandinger og det skal blive interessant at se hvilke redskaber fra etablerede forskningsområder der bliver brug for og, hvilke muligheder der vil blive udviklet til fremtidens analyser af kompliceret data .

Litteratur

- [1] P. Comon, "Independent component analysis, a new concept?", *Signal Processing*, **36**:287-314, 1994.
- [2] A. Bell, T. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution", *Neural Computation*, **7**:1129-1159, 1995.
- [3] T-W. Lee, "Independent Component Analysis - Theory and Applications", Kluwer, 1998.
- [4] A. Hyvarinen, J. Karhunen, E. Oja, "Independent Component Analysis", John Wiley & Sons, 2001.
- [5] P. Hoejen-Soerensen, O. Winther, L. K. Hansen, "Mean Field Approaches to independent Component Analysis", *Neural Computation*, **14**:889-918, 2002.