

Bachelorprojekt

Statistik og Sandsynlighedsteori

Preben Blæsild

24. november 2010

På bacheloruddannelsen i matematik skal du lave et bachelorprojekt, der har et omfang på 10 ECTS. Projektet evalueres efter 7-trinsskalaen, og der er ekstern censur.

Hvis du vil lave et projekt inden for statistik eller sandsynlighedsteori, kan du enten vælge at arbejde alene eller delvist sammen med andre studerende. I sidstnævnte tilfælde skal hver enkelt studerende dog selv udarbejde en rapport.

For at få ideer til emnet for projektet kan du kontakte Preben Blæsild eller en potentiel vejleder.

Nedenfor giver vi eksempler på emneområder. En potentiel vejleder er nævnt i parentes, og ofte er der en kort beskrivelse af emnet.

Der er også mulighed for at lave projekt inden for andre emner end de nævnte. Spørg, hvis du har specielle ønsker eller ideer.

- (1) *Simulering* (Jens Ledet Jensen). Basale principper i simulering studeres med udgangspunkt i noterne *Stochastic simulations: concepts and applications* (Jens Ledet Jensen), eller med udgangspunkt i en tilsvarende lærebog. Dette projekt er ikke relevant hvis man har fulgt et kursus i simulering.
- (2) *Lévy processer* (Svend-Erik Graversen eller Jan Pedersen). I kurset *Stokastiske processer* gives en kort introduktion til Lévy processer. I et bachelorprojekt kan man fordybe sig yderligere inden for dette emne, og f.eks. studere sammenhængen mellem foldningssemigrupper, uendeligt delelige fordelinger og udfaldsfunktionerne for en Lévy proces. Endvidere kan man se på simulation af Lévy processer.

- (3) *Centrale grænseværdisætninger; Berry-Esseen; Poisson convergence; evt. stabile fordelinger.* (Jan Pedersen eller Jens Ledet Jensen). Den centrale grænseværdisætning fortæller, at gennemsnittet af en iid-følge med endelig varians, passende normeret, konvergerer mod en normalfordeling. I dette projekt kan man studere konvergenshastigheden (Berry-Esseen). Desuden kan man studere situationer, hvor man får en Poissonfordeling i grænsen. Endelig kan man, som afslutning, evt. relatere til stabile fordelinger. Man er meget velkommen til at supplere med simulationer.
- (4) *Fødsels- og dødsprocesser; mere generelle Markovprocesser med kontinuert tid og tælleligt tilstandsrum.* (Jan Pedersen eller Jens Ledet Jensen). En fødselsproces, N_t , modellerer antal personer der fødes i en population i et interval $[0, t]$ for $t \geq 0$. Man er blandt interesseret i at undersøge, om der gælder, at N_t er endelig med sandsynlighed 1.
- I dette projekt studerer man fødselsprocesser og mere generelle Markovprocesser, f.eks. fødsels- og dødsprocesser. Man er meget velkommen til at supplere med simulationer.
- (5) *Stationære processer* (Jan Pedersen). Stationære processer er en særdeles vigtig modelklasse. I et bachelorprojekt kan man sætte sig ind i den basale teori for denne processtype og studere vigtige eksempler. Blandt nøgleordene er: spektralfordelingen; stokastisk integration; ergodesætningen; stationære Gaussiske processer.
- (6) *Elastiske deformationer og hjernebilleder* (Asger Hobolth) I billedanalyse kan man være interesseret i at undersøge hjerneaktiviteterne mellem to forskellige individer, men før en sådan undersøgelse kan begynde er det nødvendigt at transformere data så man sammenligner tilsvarende positioner. En mulig klasse af transformationer er de såkaldte elastiske deformationer, og de kan findes vha. finite element analysis. Finite element analysis (FEA) er et meget udbredt redskab i ingeniørverdenen. FEA involverer triangulering og teori fra den flerdimensionelle normalfordeling.
- (7) *Permutationstest* (Jens Ledet Jensen) Ved et permutationstest betragter man de værdier en teststørrelse kan opnå ved at permutere observationerne. Man opnår herved et eksakt test uden brug af fordelingsantagelser på observationerne.

Projektet tager udgangspunkt i bogen: *Permutation Methods: a Distance Function Approach* af Paul W. Mielke og Kenneth J. Berry. Kapitel 1 og 2 skal læses og dernæst valgfrit fra kapitlerne 4-7.

- (8) *Cox processer* (Eva Vedel). I dette projekt vil den studerende få lejlighed til at sætte sig ind i modeller for Cox processer, dvs. Poisson punkt processer med tilfældig intensitetsfunktion. Modellerne anvendes til at beskrive klumpede punkt mønstre og har en række anvendelser, bl.a. i beskrivelsen af plante populationer. Det er muligt analytisk at udlede en række egenskaber ved disse processer, herunder intensitetsfunktionen og den såkaldte par-korrelations funktion.
- (9) *Geometriske sandsynligheder* (Markus Kiderlen) I 1733 spurgte Compte du Buffon: "Hvis jeg kaster en nål tilfældigt på et stykke linjeret papir, hvad er så sandsynligheden for at nålen skærer en af linjerne?" I dette projekt vil studerende få lejlighed til at undersøge mere generelle tilfældige geometriske figurer i planen og svare på spørgsmål som: "Hvad er sandsynligheden for at en 'tilfældigt kastet' firkant rammer en trekant?" Svaret vil kræve definition af en naturlig fordeling af tilfældige to-dimensionale mængder. Det viser sig, at de omtalte sandsynligheder kan udtrykkes ved hjælp af fladeindhold og perimenter af de betragtede figurer.
- (10) *Poisson processer af linjer i planen* (Markus Kiderlen) Vi betragter en kollektion af tilfældige linjer, hvis hældninger er taget fra en endelig mængde. Hvis antallet af linjer, som skærer vilkårlige testmængder, er tilfældigt og Poisson-fordelt, så er kollektionen en såkaldt Poisson proces af linjer i planen. Skæringspunkterne af linjerne danner et tilfældigt punktmønster. Formålet med projektet er at beskrive sammenhængen mellem middelværdierne af antallet af linjer og antallet af punkter. Det viser sig, at man kan angive en retningsfordeling for linjerne, så det forventede punktantal bliver maksimalt.
- (11) *Randkorrektur i punktprocesteori* (Markus Kiderlen) En punktproces er en mængde af tilfældige punkter i planen. Punktprocesser bruges som modeller i anvendelser, fx. i biologi, hvor punkterne kan representere celler i et vævssnit. Vi antager at punktprocessen kun kan observeres i et kompakt vindue. Beregning af summariske estimatorer såsom fordelingen af afstanden mellem en punkt i punktprocessen til dens tætteste

nabo kan ikke beregnes ud fra observationen i vinduet alene: den tætteste nabo kunne ligge udenfor vinduet. I dette bachelorprojekt skal der beskrives forskellige metoder af randkorrektur, og vises, at de fører til middelværdirette estimatorer. Man er meget velkommen til at supplere med simulationer.

- (12) *Rumlige prædiktions - kriging* (Markus Kiderlen) Vi ser på et datasæt bestående af målinger af arsenikkoncentrationen i grundvand i brønde i USA. Ved hjælp af en lineær model kan arsenikkoncentrationen forudsiges på lokaliteter hvor der ikke foreligger målinger. Dette kræver en model af den rumlige kovariansfunktion. Valget af modellen er afgørende for prædiktionskvaliteten og skal analyseres i projektet.
- (13) *Stokastiske beviser* (Markus Kiderlen) En række sætninger kan vises ved brug af sandsynlighedsargumenter. For eksempel vides, at en tilfældigt drejet (konveks) sten kaster en skygge hvis areal i gennemsnit er lig med en fjerdedel af stenens overflade. En stokastisk variabel kan ikke altid være strengt mindre end middelværdien, og derfor kan stenen altså drejes således, at skyggens areal er mindst en fjerdedel af stenens overflade. Vi har udledt en geometrisk sætning ved hjælp af et sandsynlighedsargument. I bachelorprojektet ser vi på flere eksempler.
- (14) *Principal Curves and Surfaces* (Kristjana Yr Jonsdottir og Kim Mouridsen) er generaliseringer af principal komponenter til ikke-lineære under-rum. Vi vil undersøge, hvordan principal curves kan bruges til at finde strukturer i 2D punktmønstre og for eksempel anvende teknikkerne til finde læsionsområder på MRI-skanninger af hjernen.
- (15) *Analysis of Means - ANOM* (Jørgen Granfeldt) sammenligner de absolutte afvigelser af en række gennemsnit med deres samlede gennemsnit. Dette problem blev studeret i 1827 af Laplace. Selve navnet *Analysis of Means* er fra 1967. *Analysis of Means* er et test af nulhypotesen: *ingen forskel i middelværdi*, men den kan illustreres grafisk og giver umiddelbart oplysning om hvilke gennemsnit, der skiller sig ud. *Analysis of Means* har mange anvendelser i kvalitetsforbedring.

Her er en række emner inden for sandsynlighedsteori (Jørgen Hoffmann-Jørgensen, Svend-Erik Graversen og Steen Thorbjørnsen):

- (16) *Stokastiske matricer* (Steen Thorbjørnsen) En stokastisk $n \times n$ matrix er en $n \times n$ matrix $X = (x_{ij})_{1 \leq i, j \leq n}$, hvor indgangene x_{ij} er reelle eller komplekse stokastiske variable defineret på et sandsynlighedsfelt (Ω, \mathcal{F}, P) . Hvis X er selvadjungeret, kan man man betragte egenværdierne

$$\lambda_1(X) \leq \lambda_2(X) \leq \dots \leq \lambda_n(X),$$

som reelle stokastiske variable, og man kan studere deres simultane fordeling eller f.eks. fordelingen af den største eller mindste af egenværdierne. Det kanoniske eksempel på en selvadjungeret stokastisk matrix er det såkaldte "Gaussian Unitary Ensemble", hvor indgangene er kompleks normalfordelte og uafhængige (over diagonalen). I dette tilfælde konvergerer egenværdifordelingen i passende forstand mod Wigner's semi-cirkel fordeling $\frac{1}{2\pi} \sqrt{4 - x^2} 1_{[-2, 2]}(x) dx$ for $n \rightarrow \infty$ (under passende normalisering).

- (17) *Stokastiske processer som er "exchangeable" eller "spreadable"* (Steen Thorbjørnsen) En følge $\xi = (\xi_1, \xi_2, \dots)$ af (reelle) stokastiske variable kaldes "exchangeable", hvis $(\xi_{\pi(1)}, \xi_{\pi(2)}, \xi_{\pi(3)}, \dots)$ har samme fordeling som ξ for enhver permutation π af \mathbb{N} , der opfylder, at $\pi(n) = n$ for alle pånær endeligt mange n i \mathbb{N} . Følgen ξ kaldes "spreadable" hvis $(\xi_{l_1}, \xi_{l_2}, \xi_{l_3}, \dots)$ har samme fordeling som ξ for enhver voksende følge $l_1 < l_2 < l_3 < \dots$ af naturlige tal. Et berømt resultat af de Finetti og Ryll-Nardzewski udtrykker, at både "exchangeability" og "spreadability" af ξ er ækvivalent med, at ξ er betinget i.i.d. med hensyn til en bestemt del- σ -algebra i det betragtede sandsynlighedsfelt.
- (18) *Selv-similære stokastiske processer* (Steen Thorbjørnsen) En stokastisk proces $(X_t)_{t \geq 0}$ kaldes selv-similær, hvis der for enhver konstant a i $(0, \infty)$ findes en konstant $c(a)$ i $(0, \infty)$, således at

$$(X_{at})_{t \geq 0} \stackrel{d}{=} (c(a)X_t)_{t \geq 0}.$$

Den Brownske bevægelse er et eksempel på en selv-similær proces, og man kan vise, at de selv-similære Lévy processer præcis er de Lévy processer, der har strengt stabile marginalfordelinger.

- (19) *Én-dimensionale diffusionsprocesser* (Svend-Erik Graversen) Én-dimensionale diffusionsprocesser er kontinuerte stærke Markov processer med

reelle værdier. Dette bachelorprojekt kan skrives i forlængelse af kurset *Stokastiske Processer*. Der kan for eksempel fokuseres på karakterisering via *scale function* og *speed measure* eller på forbindelsen til 2.-ordens differentialoperatorer.

- (20) *Random Walks*. En "random walk" er følgen af partialsummer af uafhængige og identisk fordelte variable eller vektorer. Tilfældet hvor de stokastiske variable kun antager værdierne -1 og $+1$ er kendt fra *Matematisk Modellering 1*. Specielle emner kan være rekurrens/transiens, loven om den itererede logaritme og arcussinus-loven.
- (21) *Konvergens i fordeling af stokastiske processer*. F.eks. den Brownske bevægelse som grænseværdi af "random walks".
- (22) *Markovkæder med generelt tilstandsrum*. F.eks. rekurrens/transiens og potentialteori.
- (23) *Optimale stoptider i Markov kæder*. F.eks. den superharmoniske indhyldnings forbindelse til Snell strategien, med "kassespillet" som ledende eksempel.
- (24) *Centrale grænseværdisætninger for martingaler*.