Laboration 1: Irreducibla Markovkedjor

Benjamin Kjellson (tidigare version skriven av Anders Björkström)

2016-05-26

Viktigt: Innan ni läser vidare

Gör först följande:

1. I den övre menyn i RStudio, tryck på `Tools --> Global Options...`. I rutan som dyker upp, där det står "Default text encoding", tryck "Change" och välj "UTF-8".
2. Gå in på kurshemsidan och ladda ner mallen för denna laboration (döp den till något vettigt när ni sparar den på er dator). Öppna den i RStudio. Skriv er rapport i denna fil.

För dig som använder en av skolans datorer

Om du har sparat din `.Rmd`-fil t.ex. i mappen `Documents` eller mappar i den, och du får ett konstigt felmeddelande när du trycker på `Knit HTML`, så finns det ett par sätt att lösa detta problem. På kurshemsidan, under samma dag som denna labb, så finns det en länk VIKTIGT! Du som använder skolans laptop. Klicka på den för att läsa hur du löser problemet.

Krav för laboration 1

Tänk på följande **krav** på er rapport:

- Rapporten måste kunna läsas av någon som inte har läst labbinstruktionerna. Så ni måste skriva vad det är ni ska göra innan ni gör det, och berätta vad syftet är.
- Rapporten måste vara skriven i **R Markdown** eller knitr.
- All kod som används måste synas i labbrapporten, men ska inte beskrivas i detalj i rapporten.
- Alla **tabeller och diagram** måste förses med **numrering och beskrivande text**, och refereras till i rapportens vanliga text på rätt sätt. Diagram måste ha lämpliga rubriker på axlarna och tabeller lämpliga rubriker på kolumnerna.

Funktioner för laboration 1

Följande funktioner kan vara användbara i denna laboration. Definiera dem därför i ett kodstycke någon gång innan första uppgiften. **Funktionerna behöver bara definieras EN gång, INTE varje gång innan de används.**

```r
# Denna funktion räknar ut matrisen A upphöjt till n, enligt den iterativa
# definitionen A^n = A %*% A %*% ... %*% A  (n stycken A), med A^0 = I.
# Exempel: mpow(A, 3) == A %*% A %*% A
mpow <- function(A, n) {
  resultat <- diag(nrow(A))
  potens <- n
  while (potens > 0) {
    resultat <- A %*% resultat
    potens <- potens - 1
  }
  return(resultat)
}

# Låt A vara en matris innehållandes sannolikheter. Denna funktion testar om
# raderna i A är identiska upp till de d första decimalerna. Som ett exempel,
# talet 0.12309 är lika med 0.12301 upp till den fjärde decimalen, men avrundat
# till 4 decimaler är dessa tal ej lika.
# Funktionen returnerar TRUE om raderna är identiska; FALSE annars.
rows_equal <- function(A, d = 4) {
  A_new <- trunc(A * 10^d) # förstora talet och ta heltalsdelen
  for (k in 2:nrow(A_new)) {
    # Kolla om alla element i rad 1 är lika med motsvarande element i rad k
    if (!all(A_new[1, ] == A_new[k, ])) {
      # Om något element skiljer sig så är raderna ej lika
      return(FALSE)
    }
  }
  # Hamnar vi här så var alla rader lika
  return(TRUE)
}

# Låt A och B vara matriser innehållandes sannolikheter. Denna funktion testar
# om elementen A är identiska, upp till de d första decimalerna, med motsvarande
# element i matrisen B.
# Funktionen returnerar TRUE om matriserna är identiska; FALSE annars.
matrices_equal <- function(A, B, d = 4) {
  A_new <- trunc(A * 10^d)
  B_new <- trunc(B * 10^d)
  if (all(A_new == B_new)) {
    return(TRUE)
  } else {
    return(FALSE)
  }
}
```

Introduktion

När man använder Markovkedjor som matematiskt hjälpmedel för problem i verkliga livet, så är man oftast intresserad av att undersöka vad övergångssannolikheterna har för gränsvärden då antalet tidssteg n går mot oändligheten. Existerar dessa gränsvärden, och vad är de i så fall? Vi vet från Chapman-Kolmogorovs ekvationer att övergångsmatriserna i n steg ges av matrisen $\mathbf{P}^n$, där $\mathbf{P}$ är kedjans övergångsmatris. Sats 4.1 på sida 206 i Ross (upplaga 11; i upplaga 10 på sida 215, och i upplaga 9 på sida 205) säger att om kedjan är irreducibel (och ergodisk) så går följden $\mathbf{P}^n$ mot en matris, som vi kan beteckna $\mathbf{P}^\infty$, vars radvektorer alla är likadana. Om kedjan har N tillstånd, numrerade 1 till N, så kan vi skriva

$$\mathbf{P}^\infty = \begin{pmatrix} \pi_1 & \pi_2 & \dots & \pi_N \\ \pi_1 & \pi_2 & \dots & \pi_N \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 & \pi_2 & \dots & \pi_N \end{pmatrix}$$

Radvektorn $(\pi_1, \pi_2, \dots, \pi_N)$ beskriver kedjans asymptotiska fördelning. Den här laborationen går ut på att bekräfta Sats 4.1 genom att undersöka ett exempel.

Förberedelser

Läs avsnitt 4.1 till 4.5 i Ross. Läs igenom den här instruktionen i dess helhelt. Läs också dokumentet Bruksanvisning för simulering av Markovkedja, som finns på kurshemsidan.

Bakgrund: Kaffeautomaten på Institutionen för Experimentell Ekonomi

På Institutionen för experimentell ekonomi finns en kaffeautomat, som inte alltid fungerar perfekt. Dess tillstånd kan beskrivas på en skala från 0 till 5, där 0 innebär att automaten inte fungerar alls, och 5 motsvarar en felfri maskin. Ju sämre tillståndet är, desto mer troligt är det att någon ringer efter en reparatör. Tyvärr är det inte säkert att reparatören lyckas återställa maskinen till tillstånd 5. Ibland gör hen ett dåligt jobb, och får endast upp automaten till tillstånd 3. Som resultat av detta kommer maskinens tillstånd att ändras från en dag till nästa på ett sätt som beskrivs av en Markovkedja vars övergångsmatris är (tillstånd 0 först och 5 sist):

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ 0.1 & 0.1 & 0 & 0.4 & 0 & 0.4 \\ 0 & 0.2 & 0.2 & 0.3 & 0 & 0.3 \\ 0 & 0 & 0.3 & 0.5 & 0 & 0.2 \\ 0 & 0 & 0 & 0.4 & 0.6 & 0 \\ 0 & 0 & 0 & 0 & 0.4 & 0.6 \end{pmatrix}$$

Du kan definiera matrisen $\mathbf{P}$ i R genom

```
P <- matrix(c(0,     0,   0, 0.5,   0, 0.5,
              0.1, 0.1,   0, 0.4,   0, 0.4,
              0,   0.2, 0.2, 0.3,   0, 0.3,
              0,     0, 0.3, 0.5,   0, 0.2,
              0,     0,   0, 0.4, 0.6,   0,
              0,     0,   0,   0, 0.4, 0.6),
            nrow = 6,
            ncol = 6,
            byrow = TRUE)
```

Uppgift 1

Besvara frågorna 1 a-d och 2 a-b genom att räkna ut matrisen **P** upphöjt till ett lämpligt tal, och läsa av på rätt ställe i matrisen. Observera att det som frågas efter är sannolikhetsfördelningar, alltså sannolikheterna för alla de olika tillstånd som kan komma ifråga. Skriv inte ut hela matrisen i dina svar. Frågor:

1. Om automaten fungerar perfekt idag, som är måndag, vilken sannolikhetsfördelning beskriver då dess tillstånd
   a. på onsdag?
   b. om en vecka (7 dagar)?
   c. om två veckor?
   d. om tre månader (90 dagar)?
2. Om automaten idag (måndag) är i tillstånd 3, vilken sannolikhetsfördelning beskriver då dess tillstånd
   a. på torsdag?
   b. om en vecka (7 dagar)?
   c. om två veckor?
   d. om tre månader (90 dagar)?

Ge svaren på dessa frågor i två tabeller (en för vardera fråga) som ser ut på följande vis (påhittade siffror):

Tid	Tillstånd 0	Tillstånd 1	Tillstånd 2	Tillstånd 3	Tillstånd 4	Tillstånd 5
På torsdag	0.000	0.067	0.133	0.200	0.267	0.333
Om en vecka	0.048	0.095	0.143	0.190	0.238	0.286
Om två veckor	0.074	0.111	0.148	0.185	0.222	0.259
Om tre månader	0.091	0.121	0.152	0.182	0.212	0.242

Du kan skapa en tabell som den ovan genom att skapa en `data.frame` på följande vis:

```
df <- data.frame(Tid = c("På torsdag", "Om en vecka",
                         "Om två veckor", "Om tre månader"))
df <- cbind(df, rbind(svar_a, svar_b, svar_c, svar_d))
names(df)[-1] <- paste0("Tillstånd ", 0:5)
```

där t.ex. `svar_a` är en vektor med sannolikhetsfördelningen i a-uppgiften. För att få detta som en tabell där sannolikheterna visas avrundade till 3 decimaler skriver du sedan

```
knitr::kable(df, digits = 3, caption = "Tabell X: ...")
```

Studera och kommentera de två tabellerna.

Tips:

- Använd funktionen `mpow` som du bör ha definierat i din rapport innan denna uppgift.
- För att komma åt kolumn j i en matris `A` så kan du skriva `A[, j]`. Detta ger dig kolumn j som en vektor. Motsvarande för rad i är `A[i, ]`. Du kommer åt elementet på rad i och kolumn j genom `A[i, j]`.
- Avrunda gärna dina sannolikheter när du svarar. Fyra decimaler räcker för denna fråga.

Uppgift 2

a. Om man räknar ut $\mathbf{P}$ upphöjt till större och större tal n, hur stort behöver n vara innan följden har konvergerat? Vi kan anse att den har konvergerat när $\mathbf{P}^n \approx \mathbf{P}^{n+1}$ med fyra decimalers noggrannhet, och dessutom alla rader i $\mathbf{P}^n$ är likadana. Dvs, likhet råder då de fyra första decimalerna är exakt likadana för elementen i $\mathbf{P}^n$ och motsvarande element i $\mathbf{P}^{n+1}$, och detsamma för raderna i $\mathbf{P}^n$. Hur lyder en rad i $\mathbf{P}^n$ för det värdet på n? Svara med fyra decimalers noggrannhet.

   - Tips: skriv en loop för att jämföra $\mathbf{P}^n$ med $\mathbf{P}^{n+1}$, där n ändras för varje steg genom loopen. Använd funktionerna `mpow`, `rows_equal` och `matrices_equal` som du redan har definierat.

b. Räkna ut Markovkedjans stationära fördelning genom att skriva kod i R som löser ekvationssystemet tillhörande Sats 4.1 i Ross. Skiljer sig resultatet från radvektorn i a-uppgiften?

   - I din rapport, skriv upp systemet i matris- och vektorform (skriv inte ut hela $\mathbf{P}$ med alla dess element; symbolen $\mathbf{P}$ räcker) och visa eventuella steg du använder för att skriva om systemet innan du låter koden ta över.
   - Använd dina kunskaper om linjär algebra (som kanske behöver fräschas upp). I filen Laboration 0: Del 2 - Matriser på kurshemsidan hittar du några R-funktioner som kan vara användbara. Ni uppmanas också att Googla för att hitta andra funktioner, eller sätt att använda funktioner ni redan känner till. Se till att läsa funktionernas hjälpsidor, så att ni förstår dem innan ni använder dem.

Uppgift 3

Om man för statistik över vilka tillstånd maskinen har befunnit sig i varje dag under ett års tid, kommer man då att få se en fördelning som liknar den stationära fördelningen i uppgift 2? Det kanske låter som en självklarhet, men det är svårt att *bevisa* att det alltid måste bli så. Undersök saken genom att simulera en utveckling som startar i tillstånd 5 och pågår i 1000 dagar. Illustrera resultatet med hjälp av ett stapeldiagram (använd funktionen `barplot` som i filen Bruksanvisning för simulering av Markovkedja, och kom ihåg att ha lämpliga rubriker på axlarna).

- Här kan koden i dokumentet Bruksanvisning för simulering av Markovkedja vara användbar (särskilt de mer generella funktionerna).

Uppgift 4

Om maskinen idag är i tillstånd 5, vad är då sannolikheten att den igår befann sig i tillstånd 1? Vi ska undersöka denna fråga på två sätt:

a. Besvara frågan rent empiriskt, genom att gå igenom resultatet från simuleringen i uppgift 3. Om vi tittar på alla dagar då maskinen har varit i tillstånd 5, hur stor andel av dessa dagar föregås av en dag då maskinen är i tillstånd 1? För att få en andel och inte ett antal så behöver du dividera med något. Använd definitionen av betingad sannolikhet för att klura ut vad (och skriv ned denna definition med samma beteckningar som i tipset nedan).

b. Man kan också besvara frågan mer teoretiskt genom att använda den stationära fördelningen från uppgift 2 och Bayes formel. Gör det! Stämmer resultaten överens?

Tips: Om du låter X_n beteckna tillståndet under dag n, och X_{n-1} tillståndet under dag $n-1$, vilken är då sannolikheten du vill beräkna? Skriv ned denna innan du svarar på (a) och (b) ovan!

Redovisning

Laborationen skall utföras i grupper om högst två personer. Redovisning sker genom skriftliga svar på uppgifterna 1–4, som ska vara inlämnade senast det datum som anges på schemat. De som har skrivit sitt namn på redogörelsen ska vara beredda på att besvara muntliga följdfrågor som examinatorn kan vilja ställa. Var noga med att följa kraven som ställdes i början av denna instruktion!