Tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats:	14 april, 2025, fm, Johanneberg.
Hjälpmedel:	Physics Handbook, Beta Mathematics Handbook, typgodkänd kalkylator.
Examinator:	Christian Forssén.
Jourhavande lärare:	Christian Forssén (031–772 3261).

Tentamen består av sex uppgifter som kan ge maximalt 60 poäng totalt. För att bli godkänd med betyg 3 krävs 24 poäng, för betyg 4 krävs 36 poäng och för betyg 5 krävs 48 poäng.

Rättningsprinciper: Alla svar skall motiveras (om ej annat anges) och införda storheter skall förklaras. Lösningarna förväntas vara välstrukturerade och begripligt presenterade. Skriv och rita tydligt. Vid tentamensrättning gäller följande allmänna principer:

- För maximal (10) poäng krävs fullständigt korrekt och välmotiverad lösning.
- Mindre fel ger 1-3 poängs avdrag. Gäller även mindre brister i presentationen.
- Lösningar som inte går att följa (t.ex. avsaknad av relevant figur, ej definierade variabler, svårtydd, etc) renderar poängavdrag även om svaret verkar vara korrekt.
- Allvarliga principiella fel ger fullt poängavdrag.
- Allvarliga fel som leder till orimliga resultat kan ge lägre poängavdrag om orimligheten pekas ut.
- Även skisserade lösningar kan ge delpoäng.

Lycka till!

---

1. Svara på följande delfrågor:

   (a) Använd produktregeln för sannolikheter för att härleda Bayes teorem
   $$\mathbb{P}\left(M \mid \mathcal{D}, I\right) = \frac{\mathbb{P}\left(\mathcal{D} \mid M, I\right) \mathbb{P}\left(M \mid I\right)}{\mathbb{P}\left(\mathcal{D} \mid I\right)}$$
   (2 poäng)

   (b) Parameterestimering innebär att utföra inferens av modellparametrarna $\boldsymbol{\theta}$ för en modell $M(\boldsymbol{\theta})$ givet data $\mathcal{D}$. Hur formuleras Bayes teorem vid parameterestimering och vilken (eller vilka) av fördelningarna som ingår i Bayes teorem är normaliserad till 1 vid marginalisering över parameterdomänen $\boldsymbol{\Omega}$ (dvs $\boldsymbol{\theta} \in \boldsymbol{\Omega}$)? *(2 poäng)*

(c) Antag att $p\left(x \mid I\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ beskriver fördelningen av slumpvariabeln X. Finn den transformation $Z = Z(X)$ som ger en standard normalfördelning för Z. Observera att transformationen av fördelningen skall visas explicit. *(6 poäng)*

2. Prestandan för en maskininlärningsmodell uppskattas ofta med en felfunktion (error function) som vi allmänt betecknar $E(\hat{y}, y)$, där $\hat{y}$ är en modellförutsägelse och y motsvarande instans måldata.

   (a) Antag att det finns tillgång till träningsdata ($\mathcal{D}_{\text{train}}$) och valideringsdata ($\mathcal{D}_{\text{val}}$). Introducera (definiera) E_{train} och E_{val} och förklara vid vilket steg av träningen som dessa mått beräknas.

   (b) Vid stokastisk optimering genomförs upprepade gradientsteg av modellparametrarna med avseende på slumpmässigt utvalda delmängder av träningsdatan $\mathcal{D}_{\text{train},i} \subset \mathcal{D}_{\text{train}}$. Är det möjligt för måttet E_{train} att öka efter ett sådant optimeringssteg? (Motivera ditt svar.)

   (c) Ge tre exempel på användbara felfunktioner där minst en är specifik för klassificeringsalgoritmer. Dina tre exempel skall vara olika på så sätt att de inte garanteras att ha samma optimum vid träning mot en given uppsättning träningsdata.

   (3 poäng per deluppgift, plus 1 poäng om alla är helt korrekta.)

3. Betrakta en prediktion $\hat{y}^{\odot} \equiv \hat{y}(\boldsymbol{x}^{\odot})$ med en maskininlärningsmodell som har tränats med en uppsättning träningsdata. Denna prediktion skall jämföras med en framtida observation $y^{\odot} \equiv y(\boldsymbol{x}^{\odot})$. Alla observationer är behäftade med ett irreducibelt fel, dvs $y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon$ där $f(\boldsymbol{x})$ är den underliggande datagenererande processen ("sanningen"). Vi kan anta att observationsfelen är oberoende och identiskt fördelade med $\mathbb{E}[\epsilon] = 0$ och $\text{Var}[\epsilon] = \sigma_\epsilon^2$.

   Visa att (kvadraten av) det förväntade prediktionsfelet kan skrivas som en summa av tre termer

   $$\mathbb{E}\left[\left(y^{\odot} - \hat{y}^{\odot}\right)^2\right] = \left(f^{\odot} - \mathbb{E}\left[\hat{y}^{\odot}\right]\right)^2 + \text{Var}\left[\hat{y}^{\odot}\right] + \sigma_\epsilon^2$$

   där $f^{\odot} = f(\boldsymbol{x}^{\odot})$. Förklara kort innebörden av varje term. *(10 poäng)*

4. Betrakta ett neuralt nätverk med p noder i insignalslagret, ett gömt lager med L noder, samt ett utsignalslager med en enda nod. Samtliga noder i det gömda lagret använder samma aktiveringsfunktion

$f(z) = 1/(1 + e^{-z})$. Utsignalsnoden använder en linjär aktiveringsfunktion $f(z) = z$. Samtliga aktiveringar z inkluderar en biasvikt.

(a) Hur många parametrar har vi totalt i vårt neurala nätverk?

(b) Betrakta små signaler och vikter så att utsignalerna från noderna i det gömda lagret kan approximeras med linjära funktioner av insignalen $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$. Teckna ett explicit uttryck för utsignalen från en av dessa gömda noder.

(c) Visa att den slutliga utsignalen också blir en linjär funktion av insignalen. Relatera koefficienterna i detta linjära samband till vikterna i nätverket.

(3 poäng per deluppgift, plus 1 poäng om alla korrekta.)

5. Avgör om Markovkedjan som beskrivs av övergångsmatrisen

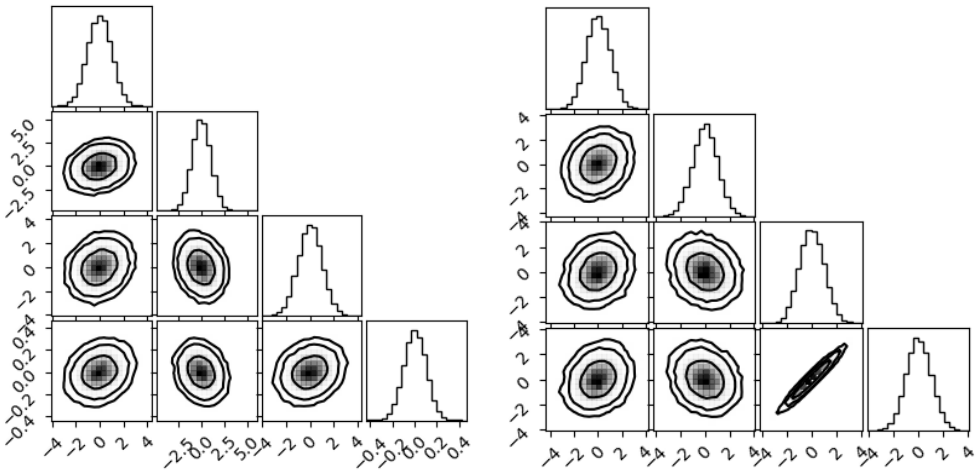
$$T = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{bmatrix}$$

är reversibel. Du skall först verifiera att detta är en giltig övergångsmatris.

(10 poäng)

6. A posteriori-fördelningar från en Bayesiansk parameterinferens för två olika modeller, fall (a) och fall (b), illustreras i figurerna nedan. Figurerna visar samtliga bivariata och univariata marginalfördelningar från respektive inferens. Figurerna inkluderar axelskalor men ej axelrubriker. Konturnivåerna motsvarar 68%, 95% och 99% sannolikhetsmassa.

Fall (a) Fall (b)



(a) Hur många modellparametrar har respektive modell?

(b) Fördelningarna har tagis fram genom Metropolissampling med en standard normalfördelning som förslagsfunktion. I bägge fallen visade sig samplingen vara väldigt ineffektiv (acceptanskvoten var lägre än 10%). Förklara varför (det finns en huvudanledning i respektive fall).

(c) Resultaten visar att den ena modellen skulle kunna förenklas. Vilken, hur och varför?

(2 poäng för (a) och 4 poäng vardera för (b) och (c))

Tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats:	14 april, 2025, fm, Johanneberg.
Hjälpmedel:	Physics Handbook, Beta Mathematics Handbook, typgodkänd kalkylator.
Lösningsskiss:	Christian Forssén.

Tentamen består av sex uppgifter som kan ge maximalt 60 poäng totalt. För att bli godkänd med betyg 3 krävs 24 poäng, för betyg 4 krävs 36 poäng och för betyg 5 krävs 48 poäng.

Rättningsprinciper: Alla svar skall motiveras (om ej annat anges) och införda storheter skall förklaras. Lösningarna förväntas vara välstrukturerade och begripligt presenterade. Skriv och rita tydligt. Vid tentamensrättning gäller följande allmänna principer:

- För maximal (10) poäng krävs fullständigt korrekt och välmotiverad lösning.
- Mindre fel ger 1-3 poängs avdrag. Gäller även mindre brister i presentationen.
- Lösningar som inte går att följa (t.ex. avsaknad av relevant figur, ej definierade variabler, svårtydd, etc) renderar poängavdrag även om svaret verkar vara korrekt.
- Allvarliga principiella fel ger fullt poängavdrag.
- Allvarliga fel som leder till orimliga resultat kan ge lägre poängavdrag om orimligheten pekas ut.
- Även skisserade lösningar kan ge delpoäng.

Detta är enbart en skiss av den fullständiga lösningen. Det kan innebära att vissa mellansteg i uträkningarna, som egentligen är nödvändiga för en komplett lösning, inte redovisas.

---

1. Svara på följande delfrågor:

   (a) Använd produktregeln för sannolikheter för att härleda Bayes teorem

   $$\mathbb{P}\left(M \mid \mathcal{D}, I\right) = \frac{\mathbb{P}\left(\mathcal{D} \mid M, I\right)\mathbb{P}\left(M \mid I\right)}{\mathbb{P}\left(\mathcal{D} \mid I\right)}$$

   (2 poäng)

   (b) Parameterestimering innebär att utföra inferens av modellparametrarna $\boldsymbol{\theta}$ för en modell $M(\boldsymbol{\theta})$ givet data $\mathcal{D}$. Hur formuleras Bayes teorem vid parameterestimering och vilken (eller vilka) av fördelningarna som ingår i Bayes teorem är normaliserad till 1

vid marginalisering över parameterdomänen $\boldsymbol{\Omega}$ (dvs $\boldsymbol{\theta} \in \boldsymbol{\Omega}$)? *(2 poäng)*

(c) Antag att $p\left(x \mid I\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ beskriver fördelningen av slumpvariabeln X. Finn den transformation $Z = Z(X)$ som ger en standard normalfördelning för Z. Observera att transformationen av fördelningen skall visas explicit. *(6 poäng)*

Lösning:

(a) Se till exempel ekvation 15.1 och 15.3 i kurskompendiet.

(b) Den relevanta formuleringen är

$$p\left(\boldsymbol{\theta} \mid \mathcal{D}, I\right) = \frac{p\left(\mathcal{D} \mid \boldsymbol{\theta}, I\right) p\left(\boldsymbol{\theta} \mid I\right)}{p\left(\mathcal{D} \mid I\right)},$$

där man bör påpeka att sambandet gäller för (betingade) sannolikhetstätheter, $p(\ldots \mid \ldots)$, pga det kontinuerliga utfallsrummet. Här kan vi notera att a posteriorifördelningen $p\left(\boldsymbol{\theta} \mid \mathcal{D}, I\right)$ och a priorifördelningen $p\left(\boldsymbol{\theta} \mid I\right)$ är sannolikhetstätheter för modellparametrarna och därmed normaliserade till 1 vid marginalisering över parameterdomänen $\boldsymbol{\Omega}$.

(c) Transformationen $z = f(x) = (x-\mu)/\sigma$ ger inversen $x = f^{-1}(z) = \sigma z + \mu$ och Jacobianen $|dx/dz| = \sigma$.
Därför får vi att $p_Z(z|I) = p_X(x = \sigma z + \mu|I)\sigma$ vilket, med den givna formen för $p_X(x|I)$, ger

$$p_Z(z|I) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

Detta är en standard normalfördelning (medelvärdet noll och variansen ett).

---

2. Prestandan för en maskininlärningsmodell uppskattas ofta med en felfunktion (error function) som vi allmänt betecknar $E(\hat{y}, y)$, där $\hat{y}$ är en modellförutsägelse och y motsvarande instans måldata.

(a) Antag att det finns tillgång till träningsdata ($\mathcal{D}_{\text{train}}$) och valideringsdata ($\mathcal{D}_{\text{val}}$). Introducera (definiera) E_{train} och E_{val} och förklara vid vilket steg av träningen som dessa mått beräknas.

(b) Vid stokastisk optimering genomförs upprepade gradientsteg av modellparametrarna med avseende på slumpmässigt utvalda delmängder av träningsdatan $\mathcal{D}_{\text{train},i} \subset \mathcal{D}_{\text{train}}$. Är det möjligt för måttet E_{train} att öka efter ett sådant optimeringssteg? (Motivera ditt svar.)

(c) Ge tre exempel på användbara felfunktioner där minst en är specifik för klassificeringsalgoritmer. Dina tre exempel skall vara olika på så sätt att de inte garanteras att ha samma optimum vid träning mot en given uppsättning träningsdata.

(3 poäng per deluppgift, plus 1 poäng om alla är helt korrekta.)

Lösning:

(a) Vi introducerar felfunktionen för tränings- respektive valideringsdata

$$E_{\text{train}} = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} E(\hat{y}(x_i), y_i), \quad \text{för } (x_i, y_i) \in \mathcal{D}_{\text{train}},$$

$$E_{\text{val}} = \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} E(\hat{y}(x_i), y_i), \quad \text{för } (x_i, y_i) \in \mathcal{D}_{\text{val}}.$$

Vid slutet av varje träningsiteration (eller varje epok) beräknas dessa mått för den aktuella versionen av modellen. Notera alltså att de inte används för gradientstegsoptimering (det är kostnadsfunktionen som används till det).

(b) Vi inför delsumman $E_{\text{train},i} = \sum_j E(\hat{y}(x_j), y_j)$, för $(x_j, y_j) \in \mathcal{D}_{\text{train},i}$, och konstaterar att denna delsumma garanterat minskar vid det enskilda gradientsteget. Däremot medför inte detta något villkor på den återstående delen av E_{train}, dvs $E_{\text{train},\bar{i}} \equiv E_{\text{train}} - E_{\text{train},i}$ är inte med i bestämningen av gradientsteget och kan öka. Den delen skulle faktiskt kunna öka såpass mycket att den totala summan E_{train} ökar.

(c) För regression används ofta kvadratfelet $E(\hat{y}, y) = (\hat{y} - y)^2$ eller absolutfelet $E(\hat{y}, y) = |\hat{y} - y|$. För klassificering används ofta felklassificering som ett mått $E(\hat{y}, y) = 1 - \delta_{\hat{y},y}$.

---

3. Betrakta en prediktion $\hat{y}^{\odot} \equiv \hat{y}(\boldsymbol{x}^{\odot})$ med en maskininlärningsmodell som har tränats med en uppsättning träningsdata. Denna prediktion

skall jämföras med en framtida observation $y^{\odot} \equiv y(\boldsymbol{x}^{\odot})$. Alla observationer är behäftade med ett irreducibelt fel, dvs $y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon$ där $f(\boldsymbol{x})$ är den underliggande datagenererande processen ("sanningen"). Vi kan anta att observationsfelen är oberoende och identiskt fördelade med $\mathbb{E}[\epsilon] = 0$ och $\mathrm{Var}[\epsilon] = \sigma_\epsilon^2$.

Visa att (kvadraten av) det förväntade prediktionsfelet kan skrivas som en summa av tre termer

$$\mathbb{E}\left[\left(y^{\odot} - \hat{y}^{\odot}\right)^2\right] = \left(f^{\odot} - \mathbb{E}\left[\hat{y}^{\odot}\right]\right)^2 + \mathrm{Var}\left[\hat{y}^{\odot}\right] + \sigma_\epsilon^2$$

där $f^{\odot} = f(\boldsymbol{x}^{\odot})$. Förklara kort innebörden av varje term. *(10 poäng)*

Lösning:
Detta är bevis 1 från bevislistan (se kurshemsida och hänvisning till kompendiet).

---

4. Betrakta ett neuralt nätverk med p noder i insignalslagret, ett gömt lager med L noder, samt ett utsignalslager med en enda nod. Samtliga noder i det gömda lagret använder samma aktiveringsfunktion $f(z) = 1/(1 + e^{-z})$. Utsignalsnoden använder en linjär aktiveringsfunktion $f(z) = z$. Samtliga aktiveringar z inkluderar en biasvikt.

   (a) Hur många parametrar har vi totalt i vårt neurala nätverk?

   (b) Betrakta små signaler och vikter så att utsignalerna från noderna i det gömda lagret kan approximeras med linjära funktioner av insignalen $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$. Teckna ett explicit uttryck för utsignalen från en av dessa gömda noder.

   (c) Visa att den slutliga utsignalen också blir en linjär funktion av insignalen. Relatera koefficienterna i detta linjära samband till vikterna i nätverket.

   (3 poäng per deluppgift, plus 1 poäng om alla korrekta.)

   Lösning:

   (a) Antalet parametrar är $(p+1) \cdot L + (L+1) \cdot 1 = (p+2) \cdot L + 1$.

   (b) Utsignalen från nod 1 härleds genom Taylorutveckling:

$$y^1(z^1) = f(z^1)\big|_{z^1=0} + f'(z^1)\big|_{z^1=0}\, z^1 + \mathcal{O}((z^1)^2) = \frac{1}{2}\left(1 + \frac{z^1}{2} + \mathcal{O}((z^1)^2)\right).$$

Med $z^1 = w^1_{01} + \sum_{i=1}^{p} w^1_{i1} x_i$ finner vi att

$$y^1 \approx \frac{1}{2} + \frac{1}{4}\left(w^1_{01} + \sum_{i=1}^{p} w^1_{i1} x_i \right).$$

(c) Den slutliga utsignalen är $y = w^2_{01} + \sum_{l=1}^{L} w^2_{l1} y^1_l$. Signalerna y^1_l från det gömda lagret ges av (b). Därigenom fås den linjära utsignalen

$$y = k_0 + k_1 x_1 + k_2 x_2 + \ldots + k_p x_p$$

med

$$k_0 = w^2_{01} + \frac{1}{4} \sum_{l=1}^{L} w^2_{l1} (2 + w^1_{0l}),$$

$$k_i = \frac{1}{4} \sum_{l=1}^{L} w^2_{l1} w^1_{il}.$$

---

5. Avgör om Markovkedjan som beskrivs av övergångsmatrisen

$$T = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{bmatrix}$$

är reversibel. Du skall först verifiera att detta är en giltig övergångsmatris.

(10 poäng)

Lösning:

- Övergångsmatrisen T har icke-negativa element $T(i,j) \geq 0$ för alla i, j samt normaliserade radsummor, dvs $\sum_j T_{ij} = 1$ för alla i.
- Lös ekvationssystemet

$$\sum_i \pi_i T_{ij} = \pi_j, \quad \forall j$$

under tvångsvillkoret $\pi_1 + \pi_2 + \pi_3 = 1$ för att finna stationärfördelningen. Lösningen är (mellansteg eller verifiering av lösningen skall redovisas)

$$\pi = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right).$$

Kontrollera detaljerad balans

$$\pi_i T_{ij} = \pi_j T_{ji}, \quad \forall i, j$$

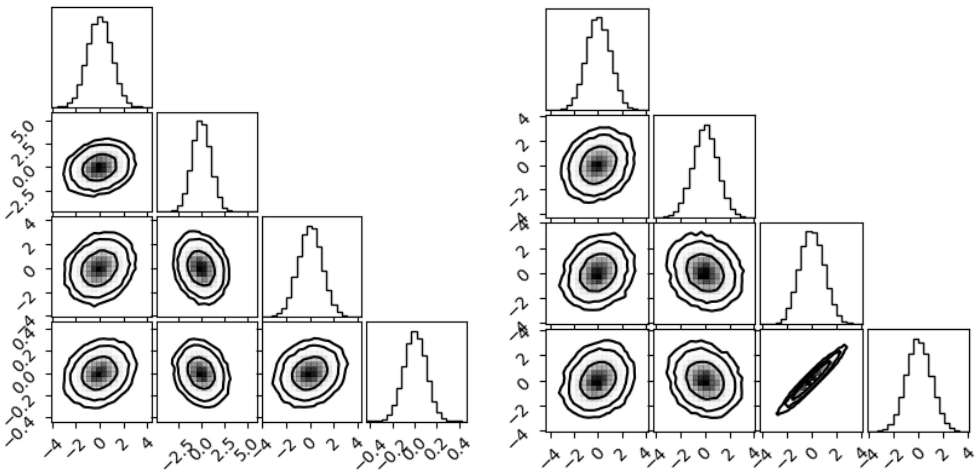
Samtliga dessa likheter är inte uppfyllda (se t.ex. $\pi_1 T_{12} = 1/6 \neq \pi_2 T_{21} = 0$) och kedjan är alltså inte reversibel.

En insikt (som dock inte krävs för full poäng) är följande: Man kan notera att ovan övergångsmatris beskriver uppdateringar som alltid innebär ett oförändrat tillstånd eller en övergång i en given riktning. En betraktelse av denna kedja med omvänd tidsriktning skulle alltså inte se likadan ut utan övergångarna skulle ske åt andra hållet.

---

6. A posteriori-fördelningar från en Bayesiansk parameterinferens för två olika modeller, fall (a) och fall (b), illustreras i figurerna nedan. Figurerna visar samtliga bivariata och univariata marginalfördelningar från respektive inferens. Figurerna inkluderar axelskalor men ej axelrubriker. Konturnivåerna motsvarar 68%, 95% och 99% sannolikhetsmassa.

Fall (a) Fall (b)



(a) Hur många modellparametrar har respektive modell?

(b) Fördelningarna har tagis fram genom Metropolissampling med en standard normalfördelning som förslagsfunktion. I bägge fallen visade sig samplingen vara väldigt ineffektiv (acceptanskvoten var lägre än 10%). Förklara varför (det finns en huvudanledning i respektive fall).

(c) Resultaten visar att den ena modellen skulle kunna förenklas. Vilken, hur och varför?

(2 poäng för (a) och 4 poäng vardera för (b) och (c))

Lösning:

(a) Fyra. (Lika många som antalet univariata marginalfördelningar.)

(b) Den sista modellparametern i fall (a) har en betydlgt kortare längdskala än de övriga. Det betyder att den genomsnittliga steglängden i Metropolisalgoritmen är alldeles för stor i den riktningen och vi kommer ofta att missa det intressanta området.

Vad gäller fall (b) så gar vi en väldigt stark korrelation mellan de två sista parametrarna. Eftersom stegförslagen är isotropa (likformigt oberoende av riktning) så kommer vi ofta att lämna det intressanta området.

(c) Modellen från fall (b) kan förenklas genom att reducera antalet fria modellparametrar. Den observerade starka korrelationen mellan de två sista parametrarna visar nämligen att dessa kan ersättas med en enda obekant.
