

NB. Geef een duidelijke toelichting bij de antwoorden. Het gebruik van een rekenmachine is toegestaan, maar geen programmeerbare/grafische rekenmachine, mobiele telefoon of laptop. Succes!

Normering: Drie punten per vraagonderdeel.

Vraag 1 (*Markov modellen*)

Het dag-tot-dag verloop van een depressie wordt gemodelleerd met een 1^{ste} orde Markov keten. Depressie wordt gediagnosticeerd in vier niveau's met oplopende ernst aan de hand van de hoeveelheid aanwezige symptomen: 0 (geen), 1 (licht), $2^{t/m} 4$ (mild), $5^{t/m} 9$ (zwaar). In het algemeen kan met kans α de ernst van de depressie één niveau toe nemen, maar de 'mild→zwaar'-overgang is twee keer zo aannemelijk. Andersom neemt de ernst met kans β één niveau af, ongeacht tussen welke twee aangrenzende niveau's. Vanuit de twee minst ernstige van de vier niveau's is het mogelijk met een kans γ in een zware depressie te geraken. Natuurlijk kan de ernst ook onveranderd blijven, maar andere niet vermelde overgangen tussen de verschillende niveau's worden niet geobserveerd.

- a) Geef de toestandsruimte en transitie-matrix van het boven beschreven Markov proces met daarbij de parameter-restricties. Teken ook het toestandsdiagram.

Voor de resterende onderdelen van deze vraag worden de twee lichtste niveau's samengevoegd en dient uitgegaan te worden van de volgende transitie-matrix :

$$\mathbf{P} = \begin{pmatrix} \frac{1}{10} & \omega & \tau \\ \omega & \frac{1}{10} & \tau \\ \omega & \tau & \frac{1}{10} \end{pmatrix}.$$

waarbij rijen- en kolommen-volgorde overeenkomen met de niveau's in oplopende volgorde.

- b) Heeft dit 1^{ste} orde Markov proces een stationaire verdeling? Zo ja, geef deze.
c) Is dit 1^{ste} orde Markov proces reversibel? Neem hierbij aan dat de stationaire verdeling gegeven wordt door $\varphi = \frac{1}{3}(1, 1, 1)^\top$ (met toestanden in dezelfde volgorde als de rijen van $\mathbf{P}$).
d) De niveau's van een patient zijn gedurende tien dagen geobserveerd: **mild, mild, zwaar, zwaar, zwaar, mild, mild, geen/licht, geen/licht, mild**. Gebruik de maximum likelihood methode om de parameters ω en τ op basis van deze observaties te schatten. Veronderstel hierbij de kans om bij deze patient op de eerste dag het **mild**-niveau te zien gelijk aan 1.

Vraag 2 (*Hidden Markov model*)

Het aantal depressie-symptomen over de dagen, $\{Y_t\}_{t=1}^\infty$ met Y_t het aantal symptomen op dag t , wordt nu gemodelleerd met een hidden Markov model. Daarin wordt aangenomen dat iemand depressief is, of niet. Deze toestand wordt niet direct geobserveerd. Een resultante daarvan, het aantal symptomen (welk op dezelfde wijze als aan het begin van Vraag 1 in vier groepen gecategoriseerd worden, nu zonder de voornoemde interpretatie), wel. Het hidden Markov model heeft toestandsruimte $\mathcal{S} = \{\text{niet depressief}, \text{depressief}\}$ en alfabet $\mathcal{A}$ bestaande uit de vier groepen met een oplopend symptoom aantal (in die volgorde). De bijbehorende startverdeling, transitie-

en emissie-matrix zijn (respectievelijk) $\pi = (0.7, 0.3)^\top$,

$$\mathbf{P} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}, \text{ en } \mathbf{B} = \begin{pmatrix} 0.6 & 0.4 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.4 & 0.5 \end{pmatrix}.$$

- a) Bereken de volgende transitie-kans $P(Y_t = 1 \mid Y_{t+1} = 2 \text{ \% } 4)$ waarbij t groot genoeg zodat het onderliggende proces stationair verondersteld mag worden.
- b) Wat is de meest aannemelijke toestandssequentie die ten grondslag ligt aan de geobserveerde sequentie $P((Y_1, Y_2, Y_3) = (1, 2 \text{ \% } 4, 5 \text{ \% } 9))$?

Vraag 3 (Netwerken)

Beschouw een pathway van 3 genen. De expressie niveaus van genen 1 en 2, aangeduid met random variabelen Y_1 en Y_2 , volgen een bivariate verdeling:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 & 2 \\ 2 & 2 \end{pmatrix}\right)$$

De expressie van het derde gen, Y_3 , wordt gegeven door de regressie-vergelijking: $Y_3 = \beta_1 Y_1 + \beta_2 Y_2 + \varepsilon$. De random variable ε , welk de ruis representeert, volgt een standaard normale verdeling en is onafhankelijk van zowel Y_1 als Y_2 .

- a) Wat impliceert de relatie $Y_2 \not\perp Y_3 \mid Y_1$ voor coëfficiënten van de regressie-vergelijking $Y_3 = \beta_1 Y_1 + \beta_2 Y_2 + \varepsilon$ binnen dit 3 gen pathway?
- b) Neem aan dat $\beta_1 = 2$ en $\beta_2 = 1$. Bepaal $\text{Cov}(Y_1, Y_3)$ en $\text{Var}(Y_3)$.
- c) Neem aan dat $\beta_1 = 2$ en $\beta_2 = 1$. Geef de correlatie tussen Y_2 en Y_3 .
- d) Vergeet de antwoorden op onderdelen b) en c). Neem dat $\text{Cov}(Y_1, Y_3) = 0$, $\text{Cov}(Y_2, Y_3) = -1$, en $\text{Cov}(Y_3, Y_3) = 3$. Welke conclusies m.b.t. de conditionele onafhankelijkheden binnen het 3-gen pathway zijn gerechtvaardigd op basis van de bovenstaande informatie?

FORMULE BLAD

Bij het tentamen kunnen de volgende formules handig zijn.

De inverse van een 2×2 matrix $\mathbf{A}$ met elementen $a_{j_1, j_2} = (\mathbf{A})_{j_1, j_2}$ is:

$$\mathbf{A}^{-1} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} = [\det(\mathbf{A})]^{-1} \begin{pmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{pmatrix}$$

met $\det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21}$.

De inverse van een 3×3 matrix $\mathbf{A}$ met elementen $a_{j_1, j_2} = (\mathbf{A})_{j_1, j_2}$ is:

$$\mathbf{A}^{-1} = [\det(\mathbf{A})]^{-1} \begin{pmatrix} a_{33}a_{22} - a_{32}a_{23} & -(a_{33}a_{12} - a_{32}a_{13}) & a_{23}a_{12} - a_{22}a_{13} \\ -(a_{33}a_{21} - a_{31}a_{23}) & a_{33}a_{11} - a_{31}a_{13} & -(a_{23}a_{11} - a_{21}a_{13}) \\ a_{32}a_{21} - a_{31}a_{22} & -(a_{32}a_{11} - a_{31}a_{12}) & a_{22}a_{11} - a_{21}a_{12} \end{pmatrix}$$

met $\det(\mathbf{A}) = a_{11}(a_{33}a_{22} - a_{32}a_{23}) - a_{21}(a_{33}a_{12} - a_{32}a_{13}) + a_{31}(a_{23}a_{12} - a_{22}a_{13})$.

De dichtheidsfunctie van de multivariaat normale verdeling van random variable $\mathbf{Y}$ is

$$f(\mathbf{Y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp[-(\mathbf{Y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu}) / 2],$$

met $\boldsymbol{\mu}$ en $\boldsymbol{\Sigma}$ de verwachting en covariantie parameters, respectievelijk.

Indien een p -variate normaal verdeelde random variabele $\mathbf{Z}$ als volgt gepartitioneerd kan worden:

$$\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{pmatrix} \right),$$

dan wordt de conditionele verdeling van $\mathbf{Y}|\mathbf{X}$ gegeven door:

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} (\mathbf{X} - \boldsymbol{\mu}_X), \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY}).$$

Zij A en B twee gebeurtenissen. De regel van Bayes zegt dan: $P(A|B) = P(B|A) \cdot P(A) / P(B)$.

Zij A en $B_1, \dots, B_n$ gebeurtenissen z.d.d. $\sum_{i=1}^n P(B_i) = 1$. De *total probability law* zegt dan:

$$P(A) = \sum_{i=1}^n P(A, B_i).$$

Zij $\mathbf{W}$, $\mathbf{X}$, $\mathbf{Y}$ en $\mathbf{Z}$ random vectoren, $\mathbf{A}$ en $\mathbf{B}$ non-random matrices van geschikte dimensies, en c een constante. Dan geldt: $\text{Cov}(c, \mathbf{Y}) = \mathbf{0}$, $\text{Cov}(\mathbf{Y}, \mathbf{Y}) = \text{Var}(\mathbf{Y})$, $\text{Cov}(\mathbf{Y}, \mathbf{Z}) = \mathbf{0}$ als $\mathbf{Y}$ en $\mathbf{Z}$ onafhankelijk zijn, $\text{Cov}(\mathbf{A}\mathbf{X}, \mathbf{B}\mathbf{Y}) = \mathbf{A}\text{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}^\top$, en

$$\text{Cov}(\mathbf{W} + \mathbf{X}, \mathbf{Y} + \mathbf{Z}) = \text{Cov}(\mathbf{W}, \mathbf{Y}) + \text{Cov}(\mathbf{W}, \mathbf{Z}) + \text{Cov}(\mathbf{X}, \mathbf{Y}) + \text{Cov}(\mathbf{X}, \mathbf{Z}).$$

Antwoorden

Antwoord op vraag 1

Vraag 1a

De toestandsruimte is $\mathcal{S} = \{\text{geen, licht, mild, zwaar}\}$. Gebruiken we dat de rijen van een transitie-matrix tot een sommen, dan levert dat:

$$\mathbf{P} = \begin{pmatrix} 1 - (\alpha + \gamma) & \alpha & 0 & \gamma \\ \beta & 1 - (\alpha + \beta + \gamma) & \alpha & \gamma \\ 0 & \beta & 1 - (2\alpha + \beta) & 2\alpha \\ 0 & 0 & \beta & 1 - \beta \end{pmatrix}.$$

Alles kansen dienen in het interval $[0, 1]$ te liggen. De vrije parameter α , β en γ dienen derhalve te voldoen aan: $0 \leq \alpha \leq 0.5$, $0 \leq \beta, \gamma \leq 1$ zodanig dat ook $2\alpha + \beta \leq 1$ en $\alpha + \beta + \gamma \leq 1$. Includeer ook het *state diagram*.

Vraag 1b

Ja, heeft stationaire verdeling: irreducibel (elke toestand kan in eindige tijd bereikt worden vanuit elke andere) en aperiodiek (overgangskansen zijn niet gelijk aan 1, dus toestanden worden niet met een vast ritme aangedaan). Gebruik dan: $\boldsymbol{\varphi}^\top \mathbf{P} = \boldsymbol{\varphi}^\top$ en $\varphi_G + \varphi_M + \varphi_Z = 1$. Met weglating van de eerste vergelijking geeft het volgende stelsel van vergelijkingen:

$$\begin{aligned} \omega \varphi_G + \frac{1}{10} \varphi_M + \tau \varphi_Z &= \varphi_M, \\ \tau \varphi_G + \tau \varphi_M + \frac{1}{10} \varphi_Z &= \varphi_Z, \\ \varphi_G + \varphi_M + \varphi_Z &= 1. \end{aligned}$$

Gebruik uit de laatste vergelijking dat $\varphi_G + \varphi_M = 1 - \varphi_Z$ en substitueer dit in de tweede vergelijking. Los nu op voor φ_Z (gebruik makende van $\tau + \omega + \frac{1}{10} = 1$):

$$\varphi_Z = \tau \left(\tau + \frac{9}{10} \right)^{-1} = \tau (2\tau + \omega)^{-1},$$

waarmee het ook duidelijk is dat $\varphi_Z \in [0, 1]$. Substitueer deze oplossing in de laatste vergelijking en isoleer φ_G :

$$\varphi_G = 1 - \varphi_M - \tau (2\tau + \omega)^{-1}.$$

Gebruik dit in de eerste vergelijking:

$$\omega (1 - \varphi_M - \varphi_Z) + (1 - \tau - \omega) \varphi_M + \tau \varphi_Z = \varphi_M.$$

Herschreven:

$$\omega (1 - \varphi_Z) + \tau \varphi_Z = (2\omega + \tau) \varphi_M.$$

Ofwel:

$$\varphi_M = (2\omega + \tau)^{-1} [\omega (1 - \varphi_Z) + \tau \varphi_Z].$$

Daar $\omega, \tau, \varphi_Z \in [0, 1]$ volgt $\varphi_M \in [0, 1]$. Gebruikmakende van het bovenstaande verkrijgen we als laatste:

$$\varphi_G = 1 - (2\omega + \tau)^{-1} [\omega (1 - \varphi_Z) + \tau \varphi_Z] - \tau (2\tau + \omega)^{-1}.$$

Uit de constructie van de laatste kans volgt: $\varphi_G + \varphi_M + \varphi_Z = 1$. Rest te verifiëren dat $\varphi_G \in [0, 1]$. Hiertoe:

$$\begin{aligned}(2\omega + \tau)(2\tau + \omega)\varphi_G &= (2\omega + \tau)(2\tau + \omega) - (2\tau + \omega)[\omega(1 - \varphi_Z) + \tau\varphi_Z] - \tau(2\omega + \tau) \\ &= 5\tau\omega + 2\tau^2 + 2\omega^2 - \omega^2 - \omega\tau - \tau^2 - 2\tau\omega - \tau^2 \\ &= 2\tau\omega + \omega^2 = \omega(2\tau + \omega).\end{aligned}$$

And thus $\varphi_G = \omega(2\omega + \tau)^{-1} \in [0, 1]$,

Vraag 1c

Reversibiliteit toetst men met behulp van de detailed balance equations. De pijn zit hem in:

$$\varphi_{G/L}(\mathbf{P})_{G/L,Z} = \frac{1}{3}\tau \neq \frac{1}{3}\omega = \varphi_{Z}(\mathbf{P})_{Z,G/L}.$$

Het model is derhalve irreversibel als $\tau \neq \omega$ en reversibel als $\tau = \omega$.

Vraag 1d

De likelihood van deze sequentie wordt gegeven door: $P(X_1 = \mathbb{M}) \prod_{t=2}^{10} P(X_t | X_{t-1})$. Ofwel: $1 \times \frac{1}{10} \times \tau \times \frac{1}{10} \times \frac{1}{10} \times \tau \times \frac{1}{10} \times \omega \times \frac{1}{10} \times \omega = (\frac{1}{10})^5 \times \tau^2 \times \omega^2 = (\frac{1}{10})^5 \times \tau^2 \times (\frac{9}{10} - \tau)^2$. Neem de logaritme (dit verandert de lokatie van het maximum niet), deze is proportioneel aan: $2 \log(\tau) + 2 \log(\frac{9}{10} - \tau)$. Stel de afgeleide naar τ gelijk aan nul: $2\tau^{-1} - 2(9/10 - \tau)^{-1} = 0$. Oplossen levert: $\hat{\tau} = 9/20$. En daarmee $\hat{\omega} = 9/10 - \hat{\tau} = 9/10 - 9/20 = 9/20$.

Antwoord op vraag 2

Vraag 2a

Gebruik eerst de definitie van de conditionele kans:

$$P(Y_t = 1 | Y_{t+1} = 2 \text{ v m } 4) = P(Y_t = 1, Y_{t+1} = 2 \text{ v m } 4) / P(Y_{t+1} = 2 \text{ v m } 4).$$

Vanuit de gegeven parameters van het hidden Markov model valt verder af te leiden dat er is slechts twee onderliggende sequentie aan deze transitie-kans ten grondslag kunnen liggen $(X_t, X_{t+1}) = (0, 1)$ en $(X_t, X_{t+1}) = (1, 1)$, waar 0 en 1 voor niet- en depressief staan. Gebruik nu de totale kans wet:

$$\begin{aligned}P(Y_t = 1, Y_{t+1} = 2 \text{ v m } 4) &= P(Y_t = 1, Y_{t+1} = 2 \text{ v m } 4 | (X_t, X_{t+1}) = (0, 1)) P((X_t, X_{t+1}) = (0, 1)) \\ &\quad + P(Y_t = 1, Y_{t+1} = 2 \text{ v m } 4 | (X_t, X_{t+1}) = (1, 1)) P((X_t, X_{t+1}) = (1, 1)) \\ &= P(Y_t = 1 | X_t = 0) P(Y_{t+1} = 2 \text{ v m } 4 | X_{t+1} = 1) P(X_{t+1} = 1 | X_t = 0) P(X_t = 0) \\ &\quad + P(Y_t = 1 | X_t = 1) P(Y_{t+1} = 2 \text{ v m } 4 | X_{t+1} = 1) P(X_{t+1} = 1 | X_t = 1) P(X_t = 1) \\ &= 0.4 \times 0.4 \times 0.1 \times \varphi_{X_t=0} + 0.1 \times 0.4 \times 0.9 \times \varphi_{X_t=1}.\end{aligned}$$

Tot slot, de transitie matrix is symmetrisch en moet derhalve een uniforme stationaire distributie hebben. Deze kans is dus gelijk aan 0.026. Rest nog:

$$P(Y_{t+1} = 2 \text{ v m } 4) = P(Y_{t+1} = 2 \text{ v m } 4 | X_{t+1} = 1) P(X_{t+1} = 1) = 0.4 \times 0.5.$$

Samengevoegd levert dat: $P(Y_t = 1 | Y_{t+1} = 2 \text{ v m } 4) = 0.026/0.2 = 0.13$.

Vraag 2b

Concludeer op basis van de gegeven parameters dat slechts twee onderliggende sequenties deze

observatie kunnen genereren: $(X_1, X_2, X_3) = (0, 1, 1)$ en $(X_1, X_2, X_3) = (1, 1, 1)$. Welk van deze onderliggende sequenties maximalizeert de kans $P((X_1, X_2, X_3) | (Y_1, Y_2, Y_3))$. Bereken hiertoe voor elke van deze sequenties:

$$\begin{aligned} P((X_1, X_2, X_3) = (x_1, x_2, x_3) | (Y_1, Y_2, Y_3) = (1, 1 \text{ \% } 4, 5 \text{ \% } 9)) \\ = P((Y_1, Y_2, Y_3) = (1, 1 \text{ \% } 4, 5 \text{ \% } 9) | (X_1, X_2, X_3) = (x_1, x_2, x_3)) \\ \times P((X_1, X_2, X_3) = (x_1, x_2, x_3)) / P((Y_1, Y_2, Y_3) = (1, 1 \text{ \% } 4, 5 \text{ \% } 9)). \end{aligned}$$

De laatste term, $P((Y_1, Y_2, Y_3) = (1, 1 \text{ \% } 4, 5 \text{ \% } 9))$, kan hierbij genegeerd worden, daar ie in elk van deze kansen voorkomt. :

$$\begin{aligned} P((X_1, X_2, X_3) = (0, 1, 1) | (Y_1, Y_2, Y_3) = (1, 1 \text{ \% } 4, 5 \text{ \% } 9)) &\propto 0.7 \times 0.1 \times 0.9 \times 0.4 \times 0.4 \times 0.5, \\ P((X_1, X_2, X_3) = (1, 1, 1) | (Y_1, Y_2, Y_3) = (1, 1 \text{ \% } 4, 5 \text{ \% } 9)) &\propto 0.3 \times 0.9 \times 0.9 \times 0.1 \times 0.4 \times 0.5. \end{aligned}$$

Dus, $(X_1, X_2, X_3) = (0, 1, 1)$ is de meest aannemelijke onderliggende sequentie.

Antwoord op vraag 3

Vraag 3a

Dit conditioneel afhankelijkheidsstatement impliceert dat de corresponderende partiële correlatie ongelijk aan nul is. Daar de partiële correlaties en de regressie-coëfficiënten een één-op-één relatie hebben is de corresponderende regressie-coëfficiënt (hier β_2) ook ongelijk aan nul.

Vraag 3b

De gevraagde covariantie:

$$\begin{aligned} \text{Cov}(Y_1, Y_3) &= \text{Cov}(Y_1, 2Y_1 + Y_2 + \varepsilon_1) \\ &= \text{Cov}(Y_1, 2Y_1) + \text{Cov}(Y_1, Y_2) + \text{Cov}(Y_1, \varepsilon_1) \\ &= 2\text{Cov}(Y_1, Y_1) + \text{Var}(Y_2, Y_2) \\ &= 6 + 2 = 8. \end{aligned}$$

en de variantie:

$$\begin{aligned} \text{Var}(Y_3) &= \text{Cov}(2Y_1 + Y_2 + \varepsilon_1, 2Y_1 + Y_2 + \varepsilon_1) \\ &= \text{Var}(2Y_1, 2Y_1) + \text{Var}(Y_2, Y_2) + \text{Var}(\varepsilon_1, \varepsilon_1) + 2\text{Cov}(2Y_1, Y_2) + 2\text{Cov}(2Y_1, \varepsilon_1) + 2\text{Cov}(Y_2, \varepsilon_1) \\ &= 4 \times 3 + 2 + 1 + 4 \times 2 = 23. \end{aligned}$$

Bovenstaande afleidingen gebruiken de onafhankelijkheid tussen ε_1 en Y_1 alswel Y_2 .

Vraag 3c

De correlatie wordt nu gegeven door:

$$\text{Cor}(Y_2, Y_3) = \frac{\text{Cov}(Y_2, Y_3)}{\sqrt{\text{Var}(Y_2)}\sqrt{\text{Var}(Y_3)}} = \frac{6}{\sqrt{2}\sqrt{23}} \approx 0.8846.$$

Vraag 3d

Uit de opgave volgt:

$$\Sigma = \begin{pmatrix} 3 & 2 & 0 \\ 2 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix}.$$

Eenvoudig rekenwerk (welk beperkt kan worden tot de boven-diagonaal) levert:

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{3} \begin{pmatrix} 5 & -6 & -2 \\ -6 & 9 & 3 \\ -2 & 3 & 2 \end{pmatrix}$$

Daar de inverse covariante matrix 1-op-1 relateert aan de partiele correlaties corresponderen de conditionele onafhankelijkheden met nullen in $\boldsymbol{\Sigma}^{-1}$. Dus, $Y_1 \not\perp Y_3 \mid Y_2$, $Y_1 \not\perp Y_3 \mid Y_2$, en $Y_2 \not\perp Y_3 \mid Y_1$. Ofwel, er zijn geen conditionele onafhankelijkheden.