Hertentamen – Voortgezette biostatistiek / Biomedische wiskunde

3 juni 2015; 18:30-20:30

NB. Geef een duidelijke toelichting bij de antwoorden. Na correctie liggen de tentamens ter inzage bij het onderwijsbureau. Het gebruik van een (ouderwetse) rekenmachine is toegestaan, maar niet dat van een programmeerbare danwel grafische rekenmachine of een mobiele telefoon. Veel succes!

Normering: 1a) 3, 1b) 3, 1c) 3, 1d) 3, 2a) 3, 2b) 3, 2c) 3, 3a) 3, 3b) 3, 3c) 3, 3d) 3.

Vraag 1 *(Markov modellen)*
Op de waddeneilanden wordt het weer elke dag geclassificeerd als zonnig, bewolkt of regenachtig. Het weer voor de volgende dag hangt alleen af van het weer van vandaag, maar niet van de dagen daarvoor. Het weer wordt derhalve met een 1^{ste} orde Markov proces gemodelleerd. Als het een bewolkte dag is, dan zijn de respectievelijke kansen voor de volgende dag: β (regen), 2β (bewolkt) en β (zon). Regent het echter, dan schijnt de daaropvolgende dag de zon niet maar is het met kans α wel bewolkt. En mocht vandaag de zon schijnen, dan is het morgen met kans γ bewolkt of regent het met kans $\gamma/4$.

Vraag 1a)
Geef de toestandsruimte en transitie-matrix van het boven beschreven Markov proces. Specificeer daarbij de restricties op de parameters. Teken ook het *state diagram* met daarin van elke overgang de bijhorende kans.

Vraag 1b)
Heeft dit 1^{ste} orde Markov proces een stationaire verdeling? Zo ja, geef deze.

Vraag 1c)
Is dit 1^{ste} orde Markov proces reversibel? Motiveer het antwoord. Belicht tevens of je reversibiliteit noodzakelijk vindt voor een zinvolle beschrijving van het weer op de Waddeneilanden.

Vraag 1d)
Gegeven is het volgende verloop van een willekeurige doch representatieve week: `regen`, `regen`, `bewolkt`, `zon`, `zon`, `regen`, `bewolkt`. Gebruik de maximum likelihood methode om de vrije parameters van het proces op basis van dit verloop te schatten.

Vraag 2 *(Hidden Markov model)*
Beschouw een ternaire DNA sequentie (bestaande uit elementen `A`, `B`, en `C`). De geobserveerde sequentie (van `A`'s, `B`'s, en `C`'s) kan worden verklaard vanuit de functionaliteit van het DNA. Drie functionele elementen worden onderscheiden: `promotor`, `intron`, en `exon`. De verdeling van `A`'s, `B`'s, en `C`'s kan worden gemodelleerd m.b.v. een hidden Markov model (HMM), waarbij de func-

tionele elementen de toestanden van de onderliggende Markov keten representeren. De parameters $(\boldsymbol{\pi}, \mathbf{P}, \mathbf{B})$ (resp. de startverdeling, de transitie- en emissie-matrix) van het HMM worden gegeven door $\boldsymbol{\pi} = (1, 0, 0)^T$,

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \text{en} \qquad \mathbf{B} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

De rijen van de matrices $\mathbf{P}$ en $\mathbf{B}$ representeren de functionele elementen (in de volgorde zoals boven gespecificeerd). De kolommen van $\mathbf{B}$ corresponderen met van `A`, `B`, en `C` (in deze volgorde).

Vraag 2a
Bereken $P((Y_1, Y_2, Y_3) = (\texttt{A}, \texttt{B}, \texttt{C}))$ en $P((Y_1, Y_2, Y_3) = (\texttt{B}, \texttt{B}, \texttt{B}))$.

Vraag 2b
Wat is de kans op $X_2 = \texttt{promotor}$ gegeven de geobserveerde sequentie $(Y_1, Y_2, Y_3) = (\texttt{A}, \texttt{B}, \texttt{C})$, i.e. $P(X_2 = \texttt{promotor} \,|\, (Y_1, Y_2, Y_3))$? Noem deze kans voor het vervolg $p_{\texttt{p}}$.

Analoog aan $p_{\texttt{p}}$ zou men $p_{\texttt{i}} = P(X_2 = \texttt{intron} \,|\, (Y_1, Y_2, Y_3))$ en $p_{\texttt{e}} = P(X_2 = \texttt{exon} \,|\, (Y_1, Y_2, Y_3))$ kunnen bepalen.

Vraag 2c
De meest aannemelijk toestands*sequentie* die ten grondslag ligt aan een geobserveerde DNA sequentie is m.b.v. het Viterbi algoritme bepaald. Komt de meest aannemelijke toestand op positie $t = 2$ uit deze sequentie altijd overeen met de toestand welk $p_{\texttt{i}}$, $p_{\texttt{e}}$ en $p_{\texttt{p}}$ maximalizeert? Motiveer.

Vraag 3 *(Netwerken)*
Beschouw een pathway van drie genen dat het signaal van gen A verwerkt. Gesteld dat de onafhankelijkheidsrelaties tussen de expressie van drie genen (gerepresenteerd door random variabelen Y_a, Y_b, Y_c), wordt gegeven door de volgende graaf: $\mathcal{G} = (\mathcal{V} = \{a, b, c\}, \mathcal{E} = \{(a, b), (b, c)\})$.

Vraag 3a)
Gesteld dat het regressie model $Y_a = \beta_b Y_b + \beta_c Y_c + \varepsilon_a$ wordt gefit. Wat voor waarde voor de schatter van β_c verwacht je dan?

Vergeet de bovenvermelde onafhankelijkheidsrelaties. Veronderstel daarentegen dat de verdelingen van de expressies van de genen in het pathway gegeven worden door: $Y_a \sim \mathcal{N}(0, 1)$, $Y_c \sim \mathcal{N}(0, 1)$ en $Y_b | \{Y_a, Y_c\} \sim \mathcal{N}(Y_a + 2Y_c, 1)$, waarbij Y_a, Y_c en de fout in Y_b, onafhankelijk zijn.

Vraag 3b)
Bereken de correlatie tussen de (expressie niveau's van de) genen B en C.

Vraag 3c)
Bepaal de partiële correlatie tussen de (expressie niveau's van de) genen B en C.

Vraag 3d)
Het blijkt experimenteel mogelijk de expressie van gen A te controleren. Voor welke waarde van Y_a wordt de resulterende variantie in de andere twee genen geminimalizeerd? *Hint:* Conditioneer binnen een trivariate normale verdeling!

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:

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

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

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

met $\boldsymbol{\mu}$ en $\boldsymbol{\Sigma}$ de mean en covariance 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:

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

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

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

Antwoorden
Antwoord op vraag 1
Vraag 1a
De toestandsruimte $\mathcal{S}$ bestaat uit toestanden `zon`, `bewolkt` en `regen`. De transitie-matrix (in de volgorde van de voornoemde toestanden) wordt gegeven door:

$$\mathbf{P} \;=\; \begin{pmatrix} 1-5\gamma/4 & \gamma & \gamma/4 \\ \beta & 2\beta & \beta \\ 0 & \alpha & 1-\alpha \end{pmatrix}$$

waarbij $0 \leq \alpha \leq 1$, $\beta = 1/4$ en $0 \leq \gamma \leq 4/5$. Includeer ook *state diagram*.

Vraag 1b
Ja, heeft stationaire verdeling want het process is irreducibel (alle toestanden worden met kans groter dan nul aangedaan indien genoeg tijd verstrijkt) en aperiodiek (de transitiematrix impliceert stochasticiteit: geen vaste volgorde waarin de toestanden aangedaan worden).

Voor het bepalen van de stationaire verdeling, gebruik: $\boldsymbol{\varphi}^{\mathrm{T}}\mathbf{P} = \boldsymbol{\varphi}^{\mathrm{T}}$ en $\varphi_z + \varphi_b + \varphi_r = 1$. Dit geeft het volgende stelsel van vergelijkingen:

$$\begin{aligned} (1-5\gamma/4)\varphi_z + \beta\varphi_b &= \varphi_z, \\ \gamma\varphi_z + 2\beta\varphi_b + \alpha\varphi_r &= \varphi_b, \\ \gamma\varphi_z/4 + \beta\varphi_b + (1-\alpha)\varphi_r &= \varphi_r, \\ \varphi_z + \varphi_b + \varphi_r &= 1. \end{aligned}$$

De eerste vergelijking levert: $\beta\varphi_b = \frac{5\gamma}{4}\varphi_z$. Oftewel (gebruikende $\beta = 1/4$): $\varphi_b = 5\gamma\varphi_z$. Stop dit in de één-na-laatste vergelijking: $\alpha\varphi_r = \frac{3\gamma}{2}\varphi_z$. Daar de stationaire kansen tesamen tot één moeten sommeren:

$$1 \;=\; \varphi_z + 5\gamma\varphi_z + \frac{3\gamma}{2\alpha}\varphi_z \;=\; \varphi_z(1 + 5\gamma + \frac{3\gamma}{2\alpha}) \;=\; \varphi_z \frac{2\alpha + 10\alpha\gamma + 3\gamma}{2\alpha}.$$

Dus:

$$(\varphi_z, \varphi_b, \varphi_r) \;=\; \frac{1}{2\alpha + 10\alpha\gamma + 3\gamma}\,(2\alpha, 10\alpha\gamma, 3\gamma)\,.$$

Deze kansen liggen in het interval $(0, 1)$ en sommeren tot één.

Vraag 1c
Reversibiliteit toetst men met behulp van de detailed balance equations:

$$\varphi_z(\mathbf{P})_{z,r} \;=\; \frac{2\alpha}{2\alpha + 10\alpha\gamma + 3\gamma} \cdot \gamma/4 \;\neq\; \frac{3\gamma}{2\alpha + 10\alpha\gamma + 3\gamma} \cdot 0 \;=\; \varphi_r(\mathbf{P})_{r,z}.$$

Het model is derhalve niet reversibel. Reversibiliteit is hier niet noodzakelijk omdat het proces in de tijd gevolgd kan worden.

Vraag 1d
De likelihood van deze sequentie wordt gegeven door: $P(X_1 = \mathtt{r}) \prod_{t=2}^{7} P(X_t \,|\, X_{t-1})$. Ofwel: $\pi_{\mathtt{r}} \times (1-\alpha) \times \alpha \times \beta \times (1-5\gamma/4) \times \gamma/4 \times \alpha$. Neem de logaritme, deze is proportioneel aan: $\log(\pi_{\mathtt{r}}) + \log(1-\alpha) + 2\log(\alpha) + \log(\beta) + \log(1-5\gamma/4) + \log(\gamma/4)$. Stel eerste orde afgeleiden (naar

α en γ) gelijk aan nul: $2/\alpha - 1/(1-\alpha) = 0$ en $1/\gamma - 5/4(1-5\gamma/4) = 0$. Los beide vergelijkingen op: $\hat{\alpha} = 2/3$ en $\hat{\gamma} = 2/5$.

Antwoord op vraag 2

Vraag 2a

Voor de eerste drie posities zijn er drie mogelijke onderliggende sequenties: `PPP`, `PPI` en `PIE`. Enkel de eerste kan leiden tot de geobserveerde sequentie `BBB`, anderzijds kan enkel de middelste leiden tot `ABC`. De kansen op deze onderliggende sequenties zijn: $P((\texttt{PPP})) = 1 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ en $P((\texttt{PPI})) = 1 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Bepaal nu de emissie-kansen: $P((\texttt{BBB}) \,|\, (\texttt{PPP})) = \frac{1}{8}$ en $P(\texttt{ABC}) \,|\, (\texttt{PPI})) = \frac{1}{8}$. Tesaam levert dit:

$$\begin{aligned} P((\texttt{BBB})) &= P((\texttt{BBB}) \,|\, (\texttt{PPP}))P(\texttt{PPP}) \\ &= \frac{1}{4} \times \frac{1}{8} = \frac{1}{32} \end{aligned}$$

en (net zo) $P((\texttt{ABC})) = P((\texttt{ABC}) \,|\, (\texttt{PPI}))P(\texttt{PPI}) = \frac{1}{32}$.

Vraag 2b

Gevraagd is $P(X_2 = \texttt{P} \,|\, (\texttt{ABC}))$. Gebruik makende van het feit dat de onderliggende sequentie altijd in een `P` moet beginnen, is de gevraagde kans gelijk aan:

$$P(X_2 = \texttt{P} \,|\, (\texttt{ABC})) \quad = \quad P((X_1, X_2) = (\texttt{P, P}) \,|\, (\texttt{ABC})).$$

M.b.v. de regel van Bayes wordt dit:

$$P((\texttt{ABC}) \,|\, (X_1, X_2) = (\texttt{P, P}))\, P((X_1, X_2) = (\texttt{P, P}))\, /\, P((\texttt{A , B, C})).$$

Substitueer in deze uitdrukking de kansen reeds verkregen bij onder *a)* van deze vraag:

$$P((\texttt{ABC}) \,|\, (X_1, X_2) = (\texttt{P, I})) \cdot \frac{1}{2} \cdot \big(\frac{1}{16}\big)^{-1}.$$

Gebruik nu het feit dat de observaties conditioneel (op de onderliggende sequentie) onafhankelijk zijn:

$$8 \cdot P(Y_1 = \texttt{A} \,|\, X_1 = \texttt{P}) \cdot P(Y_2 = \texttt{B} \,|\, X_2 = \texttt{P}) \cdot P(Y_3 = \texttt{B} \,|\, X_2 = \texttt{P})$$

De eerste twee kansen worden gegeven in de emissie matrix $\mathbf{B}$ en de resterende kans kan m.b.v. de *total probability law* achterhaald worden:

$$8 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \sum_{x_3 \in \{\texttt{P, I, E}\}} P(\texttt{C} \,|\, X_3 = x_3) \cdot P(X_3 = x_3 \,|\, X_2 = \texttt{P}).$$

Substitueer nu de ontbrekende kansen uit de transitie en emissie matrix and verkrijg: $P(X_2 = \texttt{P} \,|\, (\texttt{ABC})) = 1$.

Vraag 2c

Gebruik makende van de *total probability law* kunnen we schrijven:

$$P(X_2 \,|\, (Y_1, Y_2, Y_3)) \quad = \quad \sum_{X_1} \sum_{X_3} P((X_1, X_2, X_3) \,|\, (Y_1, Y_2, Y_3))$$

Dit is niet noodzakelijk gelijk aan:

$$\max_{X_1,X_2,X_3} P((X_1,X_2,X_3)\,|\,(Y_1,Y_2,Y_3)),$$

de grootheid welk Viterbi maximilizeert.

Antwoord op vraag 3

Vraag 3a

Verwacht wordt $\hat{\beta}_c = 0$. Uit de conditionele onafhankelijkheids relaties weten we dat $Y_a \perp\!\!\!\perp Y_c \,|\, Y_b$. Dat betekent dat Y_c geen extra informatie bevat t.o.v. Y_b ten aanzien van het gedrag van Y_a. In de regressie analyse verwacht je derhalve dat Y_c geen additionele (t.o.v. Y_b) variatie van Y_a verklaard.

Vraag 3b

M.b.t. de mean vector: $E(Y_a) = 0 = E(Y_c)$ en dus: $E(Y_b) = E(Y_a + 2Y_c + \varepsilon_c) = E(Y_a) + 2E(Y_c) + E(\varepsilon_b) = 0{+}0{+}0$. M.b.t. de covariantie, daar Y_a, Y_c, en ε_b onafhankelijk zijn, hebben we: $\mathrm{Var}(Y_b) = \mathrm{Var}(Y_a{+}2Y_c{+}\varepsilon_b) = \mathrm{Var}(Y_a){+}4\mathrm{Var}(Y_c){+}\mathrm{Var}(\varepsilon_b) = 1{+}4{+}1 = 6$. Evenzo, weer de onafhankelijkheid gebruikende: $\mathrm{Cov}(Y_b, Y_a) = \mathrm{Cov}(Y_a + 2Y_c + \varepsilon_2, Y_a) = \mathrm{Cov}(Y_a, Y_a) + 2\mathrm{Cov}(Y_c, Y_a) + \mathrm{Cov}(\varepsilon_b, Y_a) = 1$. Op zelfde wijze vindt men: $\mathrm{Cov}(Y_b, Y_c) = 2$. Samenvattend:

$$\mathbf{Y} \;\sim\; \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 0 \\ 1 & 6 & 2 \\ 0 & 2 & 1 \end{pmatrix}\right).$$

De gevraagde correlatie is dus $\mathrm{Cor}(Y_b, Y_c) = \mathrm{Cov}(Y_b, Y_c)/[\mathrm{Var}(Y_b)\mathrm{Var}(Y_c)]^{1/2} = 2/\sqrt{6}$.

Vraag 3c

Bepaal de inverse van de covariantie matrix:

$$\begin{pmatrix} 2 & -1 & 2 \\ -1 & 1 & -2 \\ 2 & -2 & 5 \end{pmatrix}.$$

Herschaal nu zodat de diagonaal enkel 1-en bevat en vermenigvuldig vervolgens de niet-diagonaal elementen met -1. Dit is exact de partiële correlatie matrix. De gevraagde partiële correlatie is dan: $2/\sqrt{5}$.

Vraag 3d

Het formuleblad geeft de conditionele verdeling van (Y_b, Y_c) gegeven Y_a. De variantie van deze conditionele verdeling hangt niet af van Y_a. Kortom, de varianties van genen B en C worden door elke waarden van Y_a geminimalizeerd.