Tentamen – Voortgezette biostatistiek / Biomedische wiskunde

22 maart 2016; 08:45-10:45

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)*
Wanneer met de juiste tussenpozen gekeken wordt, laten de wisselingen (i.e. elke overgang, ook als de toestand niet veranderd) van de aktiviteiten tussen twee momentopnamen in het leven van een cel zich beschrijven door een 1^{ste} orde Markov keten. Het merendeel van de tijd vertoont de cel reguliere aktiviteiten (e.g. signaalverwerking). In een klein percentage van de wisselingen ($100\alpha\%$) begint de cel te delen. Dat doet de cel totdat het delingsproces voltooid is (bij elke wisseling is daar een kans β op), waarna hij terugkeert tot de reguliere (en niets anders!) orde van de dag. Sporadisch (met kans $\alpha/5$) staakt de cel zijn reguliere aktiviteiten om in een lange slaap ('senescence') te vervallen. Met kans $\beta\gamma$ ontwaakt hij hieruit enkel om de reguliere aktiviteiten weer te hervatten.

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.

Ga bij de resterende onderdelen van deze vraag uit van de volgende transitie-matrix:

$$\mathbf{P} = \begin{pmatrix} 1-\beta & \beta & 0 \\ \alpha & 1-2\alpha & \alpha \\ 0 & \beta & 1-\beta \end{pmatrix},$$

waarbij rijen- en kolommen-volgorde correspondeert met toestanden '`deling`', '`regulier`' en '`senescence`'.

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

Vraag 1c)
Gegeven is het volgende aktiviteiten-verloop van een cel: `regulier`, `regulier`, `deling`, `deling`, `regulier`, `regulier`, `senescence`, `senescence`, `senescence`, `regulier`. Gebruik de maximum likelihood methode om de parameters α en β op basis van dit verloop te schatten. Veronderstel hierbij de startverdeling $\boldsymbol{\pi} = (0, 1, 0)^\top$, welk dezelfde toestandsvolgorde als (de rijen van) $\mathbf{P}$ kent.

Vraag 2 *(Hidden Markov model)*
Beschouw een binaire DNA sequentie (bestaande uit nullen en enen). De geobserveerde sequentie $\{Y_t\}_{t=1}^T$ bestaande uit `0`-en en `1`-en kan worden verklaard vanuit de functionaliteit van het DNA. Twee functionele elementen worden onderscheiden: `intron` en `exon`. De verdeling van `0`-en en `1` kan worden gemodelleerd m.b.v. een hidden Markov model (HMM), waarbij de functionele elementen de toestanden van de onderliggende 1^{ste} order 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 $\boldsymbol{\pi} = (0,1)^\top$,

$$\mathbf{P} = \begin{pmatrix} 0.5 & 0.5 \\ 0.9 & 0.1 \end{pmatrix}, \qquad \text{en} \qquad \mathbf{B} = \begin{pmatrix} 1.0 & 0.0 \\ 0.2 & 0.8 \end{pmatrix}$$

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 `0` en `1` (in deze volgorde).

Vraag 2a
Bereken $P((Y_1, Y_2, Y_3) = (\texttt{1}, \texttt{1}, \texttt{0}))$.

Vraag 2b
Laat random variabelen X_t de onderliggende functionele elementen gemodelleerd door de Markov keten representeren. Bereken $P(X_t = \texttt{exon} | Y_t = \texttt{0}, Y_{t-1} = \texttt{1})$. Neem hierbij aan dat het onderliggende Markov proces stationair is met stationaire verdeling: $(\varphi_{\texttt{intron}}, \varphi_{\texttt{exon}})^\top = (9,5)^\top/14$.

Vraag 2c
Wat is de meest aannemelijk toestandssequentie die ten grondslag ligt aan de geobserveerde DNA sequentie $(Y_1, Y_2) = (\texttt{0}, \texttt{0})$?

Vraag 3
Beschouw een pathway van 3 genen. De expressie-niveau's van de drie genen zijn als volgt verdeeld: $Y_1 \sim \mathcal{N}(0,1)$, $Y_3 \sim \mathcal{N}(0,1)$ en $Y_2 | \{Y_1, Y_3\} \sim \mathcal{N}(Y_1 + Y_3, 1)$, waarbij Y_1, Y_3 en de fout in Y_2 (i.e. het deel in Y_2 niet verklaard door Y_1 en Y_3) onafhankelijk zijn.

Vraag 3a)
Geef de gezamenlijke, multivariate verdeling van (Y_1, Y_2, Y_3).

Vraag 3b)
Laat zien dat de partiële correlatie tussen Y_1 en Y_3 ongelijk is aan nul.

Vraag 3c)
Middels een experiment zijn de expressie niveau's van de drie genen in n samples gemeten. De expressie niveau's van het eerste gen worden via een lineair regressiemodel verklaard in termen van die van de andere twee genen: $Y_{i,1} = \beta_2 Y_{i,2} + \beta_3 Y_{i,3} + \varepsilon_i$ voor $i = 1, \ldots, n$. Is de verwachtte schatting van de regressiecoefficient β_3 gelijk of ongelijk aan nul? Motiveer je antwoord. In beide gevallen, verklaar hoe het antwoord rijmt vraag 3b en met de onafhankelijkheid van Y_1 en Y_3.

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 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:

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

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: $\mathrm{Cov}(c, \mathbf{Y}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{Y}, \mathbf{Y}) = \mathrm{Var}(\mathbf{Y})$, $\mathrm{Cov}(\mathbf{Y}, \mathbf{Z}) = \mathbf{0}$ als $\mathbf{Y}$ en $\mathbf{Z}$ onafhankelijk zijn, $\mathrm{Cov}(\mathbf{AX}, \mathbf{BY}) = \mathbf{A}\mathrm{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}^{\top}$, en

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

Antwoorden

Antwoord op vraag 1

Vraag 1a

De toestandsruimte $\mathcal{S}$ bestaat uit toestanden `deling`, `regulier` en `senescence`. De transitiematrix (in de volgorde van de voornoemde toestanden) wordt gegeven door:

$$\mathbf{P} \quad = \quad \begin{pmatrix} 1-\beta & \beta & 0 \\ \alpha & 1-6\alpha/5 & \alpha/5 \\ 0 & \beta\gamma & 1-\beta\gamma \end{pmatrix}$$

waarbij $0 \leq \alpha, \beta \leq 1$, $6\alpha \leq 5$, $0 \leq \gamma$, en $\beta\gamma \leq 1$. Includeer ook *state diagram*.

Vraag 1b

Ja, heeft stationaire verdeling: irreducibel en aperiodiek. Gebruik dan: $\boldsymbol{\varphi}^{\mathrm{T}}\mathbf{P} = \boldsymbol{\varphi}^{\mathrm{T}}$ en $\varphi_d + \varphi_r + \varphi_s = 1$. Dit geeft het volgende stelsel van vergelijkingen:

$$\begin{aligned} (1-\beta)\varphi_d + \alpha\varphi_r &= \varphi_d, \\ \beta\varphi_d + (1-2\alpha)\varphi_r + \beta\varphi_s &= \varphi_r, \\ \alpha\varphi_d + (1-\beta)\varphi_s &= \varphi_s. \end{aligned}$$

De eerste vergelijking levert: $\alpha\varphi_r = \beta\varphi_d$. Net de derde vergelijking: $\alpha\varphi_r = \beta\varphi_s$. Daar de stationaire kansen tesamen tot een moeten sommeren:

$$1 \quad = \quad (\alpha/\beta)\varphi_r + \varphi_r + (\alpha/\beta)\varphi_r \quad = \quad \varphi_r(\beta + 2\alpha)/\beta.$$

Dus: $(\varphi_d, \varphi_r, \varphi_s) = (\alpha, \beta, \alpha)/(\beta + 2\alpha)$. Deze kansen liggen in het interval $(0, 1)$ en sommeren tot een.

Vraag 1c

De likelihood van deze sequentie wordt gegeven door: $P(X_1 = \mathtt{r}) \prod_{t=2}^{10} P(X_t \,|\, X_{t-1})$. Ofwel: $1 \times (1-2\alpha) \times \alpha \times (1-\beta) \times \beta \times (1-2\alpha) \times \alpha \times (1-\beta)^2 \times \beta = (1-2\alpha)^2 \times \alpha^2 \times (1-\beta)^3 \times \beta^2$. Neem de logaritme, deze is proportioneel aan: $2\log(\alpha) + 2\log(1-2\alpha) + 3\log(1-\beta) + 2\log(\beta)$. Stel eerste orde afgeleiden (naar α en β) gelijk aan nul: $-2/(1-2\alpha) + 1/\alpha = 0$ en $-3/(1-\beta) + 2/\beta = 0$. Oplossen levert: $\hat{\alpha} = 1/4$ en $\hat{\beta} = 2/5$.

Antwoord op vraag 2

Vraag 2a

Notatie: random variabelen X_t representeren de onderliggende functionele elementen gemodelleerd door de Markov keten. Gevraagd:

$$\begin{aligned} P((Y_1, Y_2, Y_3) = (\mathtt{1}, \mathtt{1}, \mathtt{0})) &= P((Y_1, Y_2, Y_3) = (\mathtt{1}, \mathtt{1}, \mathtt{0}) | (X_1, X_2, X_3) = (\mathtt{E}, \mathtt{E}, \mathtt{E}))P((X_1, X_2, X_3) = (\mathtt{E}, \mathtt{E}, \mathtt{E})) \\ &\quad + P((Y_1, Y_2, Y_3) = (\mathtt{1}, \mathtt{1}, \mathtt{0}) | (X_1, X_2, X_3) = (\mathtt{E}, \mathtt{E}, \mathtt{I}))P((X_1, X_2, X_3) = (\mathtt{E}, \mathtt{E}, \mathtt{I})) \\ &= (0.8 \times 0.8 \times 0.2) \times (1 \times 0.1 \times 0.1) + (0.8 \times 0.8 \times 1) \times (1 \times 0.1 \times 0.9) \\ &= 0.00128 + 0.0576. \end{aligned}$$

Vraag 2b

Merk op dat X_{t-1} ook een exon moet zijn, dat is immers de enige die een 1 kan uitspugen. Uitwerken van de gevraagde kans levert dan:

$$\begin{aligned}
&P(X_t = \mathtt{E} | Y_t = \mathtt{0}, Y_{t-1} = \mathtt{1}) \\
&= P(X_t = \mathtt{E}, Y_t = \mathtt{0}, Y_{t-1} = \mathtt{1}) / P(Y_t = \mathtt{0}, Y_{t-1} = \mathtt{1}) \\
&= P(X_t = \mathtt{E}, X_{t-1} = \mathtt{E}, Y_t = \mathtt{0}, Y_{t-1} = \mathtt{0}) / P(Y_t = \mathtt{0}, Y_{t-1} = \mathtt{1}) \\
&= P(Y_t = \mathtt{0}, Y_{t-1} = \mathtt{1} \,|\, X_t = \mathtt{E}, X_{t-1} = \mathtt{E}) P(X_t = \mathtt{E}, X_{t-1} = \mathtt{E}) / P(Y_t = \mathtt{0}, Y_{t-1} = \mathtt{1}) \\
&= P(Y_t = \mathtt{0} \,|\, X_t = \mathtt{E}) P(Y_{t-1} = \mathtt{1} \,|\, X_{t-1} = \mathtt{E}) P(X_t = \mathtt{E} \,|\, X_{t-1} = \mathtt{E}) P(X_{t-1} = \mathtt{E}) / P(Y_t = \mathtt{0}, Y_{t-1} = \mathtt{1}) \\
&= 0.2 \times 0.8 \times 0.1 \times (5/14) / P(Y_t = \mathtt{0}, Y_{t-1} = \mathtt{1})
\end{aligned}$$

en

$$\begin{aligned}
P(Y_t = \mathtt{0}, Y_{t-1} = \mathtt{1}) &= P(Y_t = \mathtt{0}, Y_{t-1} = \mathtt{1} \,|\, X_t = \mathtt{E}, X_{t-1} = \mathtt{E}) P(X_t = \mathtt{E}, X_{t-1} = \mathtt{E}) \\
&\quad + P(Y_t = \mathtt{0}, Y_{t-1} = \mathtt{1} \,|\, X_t = \mathtt{I}, X_{t-1} = \mathtt{E}) P(X_t = \mathtt{I}, X_{t-1} = \mathtt{E}) \\
&= 0.2 \times 0.8 \times 0.1 \times (5/14) + 1.0 \times 0.8 \times 0.9 \times (5/14).
\end{aligned}$$

Vraag 2c
Mogelijk sequenties: $(\mathtt{E}, \mathtt{E})$ en $(\mathtt{E}, \mathtt{I})$. Welk van deze sequenties maximalizeert $P((X_1, X_2) \,|\, (Y_1, Y_2) = (\mathtt{0}, \mathtt{0}))$? Hiertoe bereken:

$$P((X_1, X_2) \,|\, (Y_1, Y_2) = (\mathtt{0}, \mathtt{0})) = P((Y_1, Y_2) = (\mathtt{0}, \mathtt{0}) \,|\, (X_1, X_2))\, P((X_1, X_2)) / P((Y_1, Y_2) = (\mathtt{0}, \mathtt{0})).$$

De noemer (de likelihood) is dezelfde voor beide kansen en kan derhalve genegeerd worden. Rest:

$$\begin{aligned}
P((Y_1, Y_2) = (\mathtt{0}, \mathtt{0}) \,|\, (X_1, X_2) = (\mathtt{E}, \mathtt{E}))\, P((X_1, X_2) = (\mathtt{E}, \mathtt{E})) &= (0.2 \times 0.2) \times (1 \times 0.1) \\
P((Y_1, Y_2) = (\mathtt{0}, \mathtt{0}) \,|\, (X_1, X_2) = (\mathtt{E}, \mathtt{I}))\, P((X_1, X_2) = (\mathtt{E}, \mathtt{I})) &= (0.2 \times 1.0) \times (1 \times 0.9)
\end{aligned}$$

De sequentie $(\mathtt{E}, \mathtt{I})$ is dus meer aannemelijk.

Antwoord op vraag 3
Vraag 3a
M.b.t. de mean vector: $E(Y_1) = 0 = E(Y_3)$ en dus: $E(Y_2) = E(Y_1 + Y_3 + \varepsilon_2) = E(Y_1) + E(Y_3) + E(\varepsilon_2) = 0 + 0 + 0$. M.b.t. de covariantie, daar Y_1, Y_3, en ε_2 onafhankelijk zijn, hebben we: $\mathrm{Var}(Y_2) = \mathrm{Var}(Y_1 + Y_3 + \varepsilon_2) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_3) + \mathrm{Var}(\varepsilon_2) = 1 + 1 + 1 = 3$. Evenzo, weer de onafhankelijkheid gebruikende: $\mathrm{Cov}(Y_2, Y_1) = \mathrm{Cov}(Y_1 + Y_3 + \varepsilon_2, Y_1) = \mathrm{Cov}(Y_1, Y_1) + \mathrm{Cov}(Y_3, Y_1) + \mathrm{Cov}(\varepsilon_2, Y_1) = 1 = \mathrm{Cov}(Y_2, Y_3)$. Samenvattend:

$$\mathbf{Y} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 1 \end{pmatrix} \right).$$

Een alternatief antwoord vereist toepassing van de stelling uit Koller en Friedman, zoals vermeldt in de lecture slides.

Vraag 3b
Bepaal de inverse van de covariantie matrix:

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

Herschaal nu zodat de diagonaal enkel 1-en bevat. Dit is exact de partiele correlatie matrix. De gevraagde partiele correlatie is dan: $\frac{1}{2}$. Inderdaad ongelijk aan nul.

Ook zonder het antwoord op vraag 3a kan dit aangetoond worden. Hiertoe merk op dat een niet-nul partiële correlatie conditionele afhankelijkheid impliceert en derhalve dat de verdelingsfunctie niet factoriseert (ten opzichte van het corresponderende paar random variabelen). De gezamenlijke verdelingsfunctie is:

$$\begin{aligned} f_{(Y_1,Y_2,Y_3)}(y_1,y_2,y_3) &= f_{(Y_1|Y_2,Y_3)}(y_1,y_2,y_3) f_{(Y_2,Y_3)}(y_2,y_3) \\ &= f_{(Y_1|Y_2,Y_3)}(y_1,y_2,y_3) f_{Y_2}(y_2) f_{Y_3}(y_3). \end{aligned}$$

Het volstaat dus om aan te tonen dat $f_{(Y_1|Y_2,Y_3)}(y_1,y_2,y_3)$ niet factoriseert. Maar:

$$\begin{aligned} f_{(Y_1|Y_2,Y_3)}(y_1,y_2,y_3) &= C\exp\{-[y_2-(y_1+y_3)]^2/2\} \\ &= C\exp\{-[y_2^2-2y_1y_2-2y_1y_3-y_1^2-y_3^2-2y_1y_3]^2/2\}. \end{aligned}$$

Daar y_1 en y_3 niet te scheiden zijn, moet $\rho(Y_1, Y_3 \,|\, Y_2)$ ongelijk aan nul zijn.

Vraag 3c
Ongelijk aan nul. De partiële correlatie tussen Y_1 en Y_3 is $1-1$ gerelateerd aan de regressiecoefficient β_3: of ze zijn beide nul, of ze zijn beide ongelijk aan nul. Daar $\rho(Y_1, Y_3 \,|\, Y_2) \neq 0$, is $\beta_3 \neq 0$. Dit is in lijn met het antwoord op vraag 3b. Anderzijds, Y_1 en Y_3 zijn marginaal onafhankelijk (gegeven in de vraag). Dit lijkt in tegenspraak met hun conditionele afhankelijkheid. Deze laatste houdt echter rekening met Y_2, welk marginaal hun samenhang verbloemt.