Hertentamen – Biostatistiek 3 / Biomedische wiskunde

2 juni 2014; 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, 1e) 3, 2a) 3, 2b) 3, 2c) 3, 2d) 3, 2e) 3.

Vraag 1 *(Markov modellen)*
Beschouw een oneindig lange, binaire DNA sequentie (bestaande uit `0`-en en `1`-en). De sequentie kan worden gemodelleerd m.b.v. een 1^{ste} orde Markov proces. De kans dat een `0` door een `1` wordt opgevolgd is gelijk aan α. De kans op een overgang in omgekeerde richting is gelijk aan $\alpha/2$.

Vraag 1a)
Geef de toestandsruimte, 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)
Een deel (ter lengte 20) van de DNA sequentie is bekend: `00110001010111101101`. Gebruik de maximum likelihood methode om α m.b.v. deze sequentie te schatten. Neem hierbij aan dat de startverdeling gegeven wordt door $\pi = (0.5, 0.5)^{\mathrm{T}}$.

In tegenstelling tot wat bij vraag 1c gesuggereerd wordt, kan de DNA sequentie van `0`-en en `1`-en niet direct worden geobserveerd. De sequentie wordt achterhaald middels de moleculen die zich aan het DNA hechten. Aan elk element van het DNA hecht zich een molecuul: `A`, `B` of `C`. De sequentie van deze moleculen kan worden gemodelleerd m.b.v. een hidden Markov model (HMM), waarbij de sequentie van `0`-en en `1`-en 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 zijn zoals als bovenstaand, en

$$\mathbf{B} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \end{pmatrix}.$$

De rijen van de matrices $\mathbf{P}$ en $\mathbf{B}$ representeren `0` en `1`. De kolommen van $\mathbf{B}$ corresponderen met drie moleculen (in alfabetisch volgorde).

Vraag 1d)
Bereken $P((Y_1, Y_2, Y_3) = (\mathtt{C}, \mathtt{B}, \mathtt{A}))$. Neem hierbij aan dat het antwoord op vraag 1c luidt $\hat{\alpha} = 0.70$.

Vraag 1e
Wat is de meest aannemelijk toestandssequentie die ten grondslag ligt aan de geobserveerde DNA sequentie $(Y_1, Y_2, Y_3) = (\mathtt{C}, \mathtt{B}, \mathtt{A})$? Veronderstel wederom $\hat{\alpha} = 0.70$.

Vraag 2 *(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:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 14 & 12 \\ 12 & 16 \end{pmatrix}\right)$$

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.

Vraag 2a)
Leg uit wat $Y_2 \perp\!\!\!\perp Y_3 \,|Y_1$ betekent, zowel wiskundig/statistisch maar illustreer ook de biologische relevantie (binnen de context van een 3 gen pathway).

Vraag 2b)
Wat impliceert de relatie $Y_2 \perp\!\!\!\perp Y_3 \,|Y_1$ voor coefficienten van de regressie-vergelijking $Y_3 = \beta_1 Y_1 + \beta_2 Y_2 + \varepsilon$ binnen dit 3 gen pathway?

Voor de resterende opgaven kan de relatie $Y_2 \perp\!\!\!\perp Y_3 \,|Y_1$ niet verondersteld worden. Wel geldt nu voor de regressie-coefficienten dat $\beta_1 = 1 = \beta_2$.

Vraag 2c)
Bepaal $\mathrm{Cov}(Y_2, Y_3)$ en $\mathrm{Var}(Y_3)$.

Vraag 2d)
Geef de correlatie tussen Y_1 en Y_3.

Vraag 2e)
Welke conclusies m.b.t. de conditionele onafhankelijkheden binnen het 3-gen pathway zijn gerechtvaardigd op basis van de bovenstaande informatie (vraag 2a en 2b dus niet meenemende)?

FORMULE BLAD
Bij het tentamen kunnen de volgende formules handig zijn.

De inverse van een 2×2 matrix $\mathbf{A}$ 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 $\mathbf{A}$ 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 $\det(\mathbf{A}) = a_{11}(a_{33}a_{22} - a_{32}a_{23}) - a_{21}(a_{33}a_{12} - a_{32}a_{13}) + a_{31}(a_{23}a_{12} - a_{22}a_{13})$.

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:

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

De oplossingen van de vergelijken $ax^2 + bx + c = 0$, $a, b, c \in \mathbb{R}$, worden gegeven door de zgn. abc-formule:

$$x = (-b \pm \sqrt{b^2 - 4ac})/(2a).$$

Antwoorden

Vraag 1a
De toestandsruimte $\mathcal{S}$ bestaat uit toestanden `0` en `1`. De transitie-matrix wordt gegeven door:

$$\mathbf{P} \;=\; \begin{pmatrix} 1-\alpha & \alpha \\ \frac{1}{2}\alpha & 1-\frac{1}{2}\alpha \end{pmatrix}$$

waarbij $0 \leq \alpha \leq 1$. Includeer ook *state diagram*.

Vraag 1b
Ja, heeft stationaire verdeling: irreducibel en aperiodiek. Gebruik dan: $\boldsymbol{\varphi}^T\mathbf{P} = \boldsymbol{\varphi}^T$ en $\varphi_0+\varphi_1 = 1$. Dit geeft het volgende stelsel van vergelijkingen:

$$\begin{aligned} (1-\alpha)\varphi_0 + \frac{1}{2}\alpha\varphi_1 &= \varphi_0, \\ \frac{1}{2}\alpha\varphi_0 + (1-\frac{1}{2}\alpha)\varphi_1 &= \varphi_1. \end{aligned}$$

De eerste vergelijking levert: $\varphi_0 = \frac{1}{2}\varphi_1$. Daar ze tesaam tot een moeten sommeren: $\varphi_0 = \frac{1}{3}$ en $\varphi_1 = \frac{2}{3}$.

Vraag 1c
De likelihood van deze sequentie wordt gegeven door: $P(X_1 = \mathtt{0})\prod_{t=2}^{20} P(X_t \,|\, X_{t-1})$. Ofwel: $1/2 \times (1-\alpha)\times\alpha\times(1-\alpha/2)\times\alpha/2\times(1-\alpha)^2\times\alpha\times\alpha/2\times\alpha\times\alpha/2\times\alpha\times(1-\alpha/2)^3\times\alpha/2\times\alpha\times(1-\alpha/2)\times\alpha/2\times\alpha$. Versimpeld: $1/2 \times \alpha^6 \times (1-\alpha)^3 \times (\alpha/2)^5 \times (1-\alpha/2)^5$. Neem de logaritme, deez is proportioneel aan: $6\log(\alpha) + 3\log(1-\alpha) + 5\log(\alpha) + 5\log(1-\alpha/2)$. Stel eerste orde afgeleide gelijk aan nul: $11/\alpha - 3/(1-\alpha) - 2.5/(1-\alpha/2) = 0$. Of: $11(1-\alpha/2)(1-\alpha) - 3\alpha(1-\alpha/2) - 2.5\alpha(1-\alpha) = 0$. Of: $11 - 16.5\alpha + 5.5\alpha^2 - 3\alpha + 1.5\alpha^2 - 2.5\alpha + 2.5\alpha^2 = 0$. Of: $11 - 22\alpha + 9.5\alpha^2 = 0$. Nu de abc-regel toepassen geeft $\hat{\alpha} = 22/19 \pm \sqrt{22^2 - 38*11}/19 \approx 0.796$.

Vraag 1d
Uit de parametrizatie van het HMM volgt dat slechts twee onderliggende sequenties de geobserveerde kunnen genereren: $\mathtt{1} \to \mathtt{1} \to \mathtt{0}$ en $\mathtt{1} \to \mathtt{0} \to \mathtt{0}$. Verder:

$$\begin{aligned} &P(Y_1 = \mathtt{C}, Y_2 = \mathtt{B}, Y_3 = \mathtt{A}) \\ &\quad= \sum_{X_1,X_2,X_3} P(Y_1 = \mathtt{C}, Y_2 = \mathtt{B}, Y_3 = \mathtt{A} \,|\, X_1, X_2, X_3)P(X_1, X_2, X_3) \\ &\quad= P(Y_1 = \mathtt{C}, Y_2 = \mathtt{B}, Y_3 = \mathtt{A} \,|\, X_1 = \mathtt{1}, X_2 = \mathtt{0}, X_3 = \mathtt{0})P(X_1 = \mathtt{1}, X_2 = \mathtt{0}, X_3 = \mathtt{0}) \\ &\qquad+ P(Y_1 = \mathtt{C}, Y_2 = \mathtt{B}, Y_3 = \mathtt{A} \,|\, X_1 = \mathtt{1}, X_2 = \mathtt{1}, X_3 = \mathtt{0})P(X_1 = \mathtt{1}, X_2 = \mathtt{1}, X_3 = \mathtt{0}) \\ &\quad= \frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\alpha(1-\alpha) + \frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}(1-\frac{1}{2}\alpha)\frac{1}{2}\alpha. \end{aligned}$$

Substitueer nu $\alpha = 0.7$.

Vraag 1e
Gevraag wordt:

$$\begin{aligned} &\max_{(X_1,X_2,X_3)} P(X_1, X_2, X_3 \,|\, Y_1 = \mathtt{C}, Y_2 = \mathtt{B}, Y_3 = \mathtt{A}) \\ &\qquad= \max_{(X_1,X_2,X_3)} \frac{P(Y_1 = \mathtt{C}, Y_2 = \mathtt{B}, Y_3 = \mathtt{A} \,|\, X_1, X_2, X_3)P(X_1, X_2, X_3)}{P(Y_1 = \mathtt{C}, Y_2 = \mathtt{B}, Y_3 = \mathtt{A})}, \end{aligned}$$

waar de Bayes regel gebruikt is. Delen door de likelihood levert geen verschil voor het maximum, dus het volstaat naar de teller van de bovenstaande breuk te kijken.

Al reeds geconcludeerd dat er slechts twee onderliggende sequentie zijn. De vraag vertaald dus naar welk van deze twee de posterior kans maximalizeert:

$$P(Y_1 = \mathtt{C}, Y_2 = \mathtt{B}, Y_3 = \mathtt{A} \,|\, X_1 = \mathtt{1}, X_2 = \mathtt{0}, X_3 = \mathtt{0}) P(X_1 = \mathtt{1}, X_2 = \mathtt{0}, X_3 = \mathtt{0}) \;=\; \frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\alpha(1-\alpha)$$

en

$$P(Y_1 = \mathtt{C}, Y_2 = \mathtt{B}, Y_3 = \mathtt{A} \,|\, X_1 = \mathtt{1}, X_2 = \mathtt{1}, X_3 = \mathtt{0}) P(X_1 = \mathtt{1}, X_2 = \mathtt{1}, X_3 = \mathtt{0}) \;=\; \frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}(1-\frac{1}{2}\alpha)\frac{1}{2}\alpha.$$

Daar $1 - \alpha = 0.3$ en $1 - \alpha/2 = 0.65$. Kortom, de tweede sequentie $X_1 = \mathtt{1}, X_2 = \mathtt{1}, X_3 = \mathtt{0}$ is derhalve de meest waarschijnlijke.

Antwoord op vraag 2

Vraag 2a
De gegeven relatie is een conditioneel onafhankelijkheidsstatement: de expressie niveaus van gen 2 en 3 (conditioneel op die van gen 1) zijn onafhankelijk. Dit betekent dat een factorizatie van de conditionele verdelingsfunctie in $f_{Y_2,Y_3 \,|\, Y_1}(y_1, y_2, y_3) = g_{Y_2 \,|\, Y_1}(y_1, y_2) h_{Y_3 \,|\, Y_1}(y_1, y_3)$ mogelijk is. De biologische relavantie van het conditioneel onafhankelijkheidsstatement laat zich als volgt illustreren (zie ook college). Stel dat, voor het goed functioneren van de cel, de expressie van gen 3 binnen een bepaalde bandbreedte dient te zijn. En dat gen 2 een verstorende rol op de expressie van gen 3 heeft. Dan kan het effect van gen 2 op gen 3 geneutraliseerd worden door de expressie van gen 1 te controleren.

Vraag 2b
Dit conditioneel onafhankelijkheidsstatement impliceert dat de corresponderende partiële correlatie nul is. Daar de partiële correlaties en de regressie-coefficienten een één-op-één relatie hebben is de corresponderende regressie-coefficient (hier β_2) ook gelijk aan nul.

Vraag 2c
De gevraagde covariantie:

$$\begin{aligned}
\mathrm{Cov}(Y_2, Y_3) &= \mathrm{Cov}(Y_2, Y_1 + Y_2 + \varepsilon_1) \\
&= \mathrm{Cov}(Y_2, Y_1) + \mathrm{Cov}(Y_2, Y_2) + \mathrm{Cov}(Y_2, \varepsilon_1) \\
&= \mathrm{Cov}(Y_2, Y_1) + \mathrm{Var}(Y_2, Y_2) \\
&= 12 + 16 \;=\; 28.
\end{aligned}$$

en de variantie:

$$\begin{aligned}
\mathrm{Var}(Y_3) &= \mathrm{Cov}(Y_1 + Y_2 + \varepsilon_1, Y_1 + Y_2 + \varepsilon_1) \\
&= \mathrm{Var}(Y_1, Y_1) + \mathrm{Var}(Y_2, Y_2) + \mathrm{Var}(\varepsilon_1, \varepsilon_1) + 2\mathrm{Cov}(Y_1, Y_2) + 2\mathrm{Cov}(Y_1, \varepsilon_1) + 2\mathrm{Cov}(Y_2, \varepsilon_1) \\
&= 14 + 16 + 1 + 2 \times 12 \;=\; 55.
\end{aligned}$$

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

Vraag 2d
De correlatie wordt nu gegeven door:

$$\mathrm{Cor}(Y_1, Y_3) \;=\; \frac{\mathrm{Cov}(Y_1, Y_3)}{\sqrt{\mathrm{Var}(Y_1)}\sqrt{\mathrm{Var}(Y_3)}} \;=\; \frac{26}{\sqrt{14}\sqrt{55}} \;\approx\; 0.937.$$

Vraag 2e
Uit de opgave en bovenstaande opgaven volgt:

$$\boldsymbol{\Sigma} \;=\; \begin{pmatrix} 14 & 12 & 26 \\ 12 & 16 & 28 \\ 26 & 28 & 55 \end{pmatrix}.$$

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

$$\boldsymbol{\Sigma}^{-1} \;=\; \begin{pmatrix} 1.200 & -0.850 & -1.000 \\ -0.850 & 1.175 & -1.000 \\ -1.000 & -1.000 & 1.000 \end{pmatrix}$$

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