Tentamen – Voortgezette biostatistiek / Biomedische wiskunde

27 maart 2015; 15:15-17:15

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)*
Bechouw het verloop van de besmetting met een mogelijk ziek-makend doch niet dodelijk virus. Het verloop (inclusief screening) over de dagen heen kan met een 1^{ste} orde Markov keten beschreven worden. Op tijdstip nul is iedereen virus vrij. Daarna raakt dagelijks $100\alpha\%$ van de virus-vrije populatie met het virus besmet. Vanaf de dag van besmetting ruimt elke dag bij $100\beta\%$ van de besmette mensen het lichaam het virus zelfstandig op. Om het virus versneld uit te roeien is een screeningsprogramma opgezet. Dagelijks wordt $100\gamma\%$ van de virus-vrije en een evenzo groot percentage van de besmette populatie gescreend op aanwezigheid van het virus. Screening gebeurt in een steriele omgeving waar geen besmetting kan plaatsvinden. Bij positieve uitslag volgt behandeling waarna iemand virus vrij is.

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. *Hint:* modelleer 'screening' als een aparte toestand.

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 voornoemde verloop.

Vraag 1d)
Gegeven is het volgende verloop van een willekeurig individu (representief voor de populatie): `virus-vrij`, `virus-vrij`, `besmet`, `besmet`, `besmet`, `screening`, `virus-vrij`. Neem aan dat $\alpha = \beta$. Gebruik de maximum likelihood methode om de parameters α en γ op basis van dit verloop te schatten.

Vraag 2 *(Phylogenetische bomen)*
Twee hedendaagse organismen met een binaire genetische code (bestaande uit nullen (`0`-en) en

enen (1-en) hebben een gemeenschappelijke voorouder. Het substitutie-proces in een willekeurige locus volgt een 1^{ste} orde Markov proces. De kans op een substitutie is gelijk aan $\alpha = 0.25$. De t-staps transitie-matrix (rijen en kolommen corresponderende met volgorde van de toestandsruimte $\mathcal{S} = \{0, 1\}$) is derhalve:

$$\mathbf{P}^{(t)} = \mathbf{P}^t = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} + (1-2\alpha)^t \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}.$$

De stationaire verdeling van dit substitutie-proces is uniform.

Vraag 2a)
De hedendaagse soorten worden door t generaties gescheiden van hun laatst gemeenschappelijk voorouder. Bereken de kans dat alle organismen een nul (0) op de betreffende locus hebben.

Vraag 2b)
Herhaal vraag 2a, nu gegeven dat er geen substitutie heeft plaatsgevonden.

Vraag 2c)
Wederom scheiden t generaties de hedendaagse organismen van hun gemeenschappelijke voorouder. Gegeven dat de locus van één van de hedendaagse soorten door een nul bezet wordt, wat is dan de kans (onder stationariteit) dat de locus van de andere hedendaagse soort ook door een nul bezet wordt?

Vraag 3 *(Netwerken)*
Beschouw een pathway van drie genen dat het signaal van gen A verwerkt. De andere twee genen (B en C) reageren verschillend op het signaal, getuige het volgende systeem van lineaire regressie vergelijkingen dat hun expressie niveaus beschrijft als een functie van dat van de andere *twee* genen in het pathway (maar enkel dat van A speelt een rol): $Y_b = -Y_a + \gamma\varepsilon_b$ en $Y_c = Y_a + \gamma\varepsilon_c$, waar γ een positieve constante en Y_a, ε_b en ε_c onafhankelijk en alle standaard normaal verdeeld ($\mathcal{N}(0,1)$) zijn.

Vraag 3a)
Bepaal de partiële correlatie tussen de (expressie niveau's van de) genen B en C direct (dat is, zonder de inversie van de covariantie matrix).

Vraag 3b)
Bepaal de partiële correlatie tussen de (expressie niveau's van de) genen A en C (nu m.b.v. de inversie van de covariantie matrix).

Vraag 3c)
Geef de definitie van conditionele onafhankelijkheid van de expressie niveaus van twee genen gegeven de derde binnen het pathway. En geef alle (conditionele) (on)afhankelijkheidsrelaties tussen de drie genen die door bovenstaande informatie gelegitimeerd worden.

Vraag 3d)
Beschrijf het effect van γ op het conditionele onafhankelijkheidsnetwerk (welk bepaald is op basis van een vaste cut-off op de absolute partiële correlaties), en verklaar het effect aan de hand van het regressie model.

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 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:

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

Antwoorden
Antwoord op vraag 1
Vraag 1a
De toestandsruimte $\mathcal{S}$ bestaat uit toestanden `virus-vrij`, `besmet` en `screening`. De transitiematrix (in de volgorde van de voornoemde toestanden) wordt gegeven door:

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

waarbij $0 \leq \alpha, \beta, \gamma \leq 1$, $\alpha + \gamma \leq 1$ 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 $\varphi_{vv} + \varphi_b + \varphi_s = 1$. Dit geeft het volgende stelsel van vergelijkingen:

$$\begin{aligned} (1-\alpha-\gamma)\varphi_{vv} + \beta\varphi_b + \varphi_s &= \varphi_{vv}, \\ \alpha\varphi_{vv} + (1-\beta-\gamma)\varphi_b &= \varphi_b, \\ \gamma\varphi_{vv} + \gamma\varphi_b &= \varphi_s. \end{aligned}$$

De tweede vergelijking levert: $\varphi_b = \frac{\alpha}{\beta+\gamma}\varphi_{vv}$. Stop dit in de laatste vergelijking: $\varphi_s = (\gamma + \frac{\alpha\gamma}{\beta+\gamma})\varphi_{vv}$. Daar de stationaire kansen tesaam tot een moeten sommeren:

$$\begin{aligned} 1 &= \varphi_{vv} + \frac{\alpha}{\beta+\gamma}\varphi_{vv} + (\gamma + \frac{\alpha\gamma}{\beta+\gamma})\varphi_{vv} \\ &= \varphi_{vv}(1 + \frac{\alpha+\alpha\gamma}{\beta+\gamma} + \gamma) \;=\; \varphi_{vv}\frac{(1+\gamma)(\alpha+\beta+\gamma)}{\beta+\gamma}. \end{aligned}$$

Dus:

$$(\varphi_{vv}, \varphi_b, \varphi_s) \;=\; \frac{\beta+\gamma}{(1+\gamma)(\alpha+\beta+\gamma)}\left(1, \frac{\alpha}{\beta+\gamma}, \gamma + \frac{\alpha\gamma}{\beta+\gamma}\right).$$

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

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

$$\begin{aligned} \varphi_b(\mathbf{P})_{b,s} &= \frac{\beta+\gamma}{(1+\gamma)(\alpha+\beta+\gamma)}\frac{\alpha}{\beta+\gamma}\cdot\gamma \\ &\neq \frac{\beta+\gamma}{(1+\gamma)(\alpha+\beta+\gamma)}\left(\gamma + \frac{\alpha\gamma}{\beta+\gamma}\right)\cdot 0 \;=\; \varphi_s(\mathbf{P})_{s,b}. \end{aligned}$$

Het model is derhalve niet reversibel.

Vraag 1d
De likelihood van deze sequentie wordt gegeven door: $P(X_0 = \mathtt{vv})\prod_{t=2}^{6} P(X_t \,|\, X_{t-1})$. Ofwel: $1 \times (1-\alpha-\gamma) \times \alpha \times (1-\beta-\gamma)^2 \times \gamma \times 1$. Neem de logaritme, deze is proportioneel aan: $\log(\alpha) + \log(1-\alpha-\gamma) + 2\log(1-\beta-\gamma) + \log(\gamma)$. Substitueer nu $\beta = \alpha$ en stel eerste orde afgeleiden (naar α en γ) gelijk aan nul: $1/\alpha - 3/(1-\alpha-\gamma) = 0$ en $1/\gamma - 3/(1-\alpha-\gamma) = 0$. Trek de twee vergelijkingen van elkaar af en concludeer: $\alpha = \gamma$. Substitueer deze zojuist verkregen kennis in een van de vergelijkingen: $1/\gamma = 3/(1-2\alpha) = 0$. Ofwel: $3\alpha = (1-2\alpha)$. Dus: $\hat{\alpha} = \hat{\gamma} = 1/5$.

Antwoord op vraag 2

Vraag 2a

Notatie: random variabelen $X_t^{(1)}$, $X_t^{(2)}$ en $X_0^{(ca)}$ representeren de binaire nucleotide in organisme 1, 2 en de laatst gemeenschappelijke voorouder (ca: common ancestor). Het subscript geeft de generatie aan. Gevraagd:

$$\begin{aligned} & P(X_t^{(1)} = 0, X_t^{(2)} = 0, X_0^{(ca)} = 0) \\ &= P(X_t^{(1)} = 0, X_t^{(2)} = 0 \,|\, X_0^{(ca)} = 0) P(X_0^{(ca)} = 0) \\ &= \frac{1}{2} P(X_t^{(1)} = 0 \,|\, X_0^{(ca)} = 0) P(X_t^{(2)} = 0 \,|\, X_0^{(ca)} = 0) \;=\; \frac{1}{2}\left[\frac{1}{2} + \frac{1}{2}(1-2\alpha)^t\right]^2, \end{aligned}$$

waar gebruik is gemaakt van de definitie van conditionele kans, de conditionele onafhankelijkheid van de hedendaagse organismen gegeven hun laatst gemeenschappelijk voorouder, en de gegeven t-staps transitiematrix. Voor volledigheid substitueer $\alpha = 0.25$.

Vraag 2b

Gevraagd is zelfde als bij 2a, maar tussentijds heeft er geen substitutie plaatsgevonden:

$$\begin{aligned} & P(X_t^{(1)} = 0, X_t^{(2)} = 0, X_{t-1}^{(1)} = 0, X_{t-1}^{(2)} = 0, \ldots, X_1^{(1)} = 0, X_1^{(2)} = 0, X_0^{(ca)} = 0) \\ &= P(X_t^{(1)} = 0, X_t^{(2)} = 0, X_{t-1}^{(1)} = 0, X_{t-1}^{(2)} = 0, \ldots, X_1^{(1)} = 0, X_1^{(2)} = 0 \,|\, X_0^{(ca)} = 0) P(X_0^{(ca)} = 0) \\ &= P(X_t^{(1)} = 0, X_{t-1}^{(1)} = 0, \ldots, X_1^{(1)} = 0 \,|\, X_0^{(ca)} = 0) \\ &\quad P(X_t^{(2)} = 0, X_{t-1}^{(2)} = 0, \ldots, X_1^{(2)} = 0, \,|\, X_0^{(ca)} = 0) P(X_0^{(ca)} = 0) \\ &= P(X_t^{(1)} = 0 \,|\, X_{t-1}^{(1)} = 0) \cdot \ldots \cdot P(X_1^{(1)} = 0 \,|\, X_0^{(ca)} = 0) \\ &\quad P(X_t^{(2)} = 0 \,|\, X_{t-1}^{(2)} = 0) \cdot \ldots \cdot P(X_1^{(2)} = 0, \,|\, X_0^{(ca)} = 0) P(X_0^{(ca)} = 0) \;=\; \frac{1}{2}(1-\alpha)^{2t}. \end{aligned}$$

waar gebruik is gemaakt van de definitie van conditionele kans, de conditionele onafhankelijkheid van de hedendaagse organismen gegeven hun laatst gemeenschappelijk voorouder, en de gegeven 1-staps transitiematrix. Voor volledigheid substitueer $\alpha = 0.25$.

Vraag 2c

Gevraagd is:

$$\begin{aligned} P(X_t^{(1)} = 0 \,|\, X_t^{(2)} = 0) &= P(X_t^{(1)} = 0, X_t^{(2)} = 0) / P(X_t^{(2)} = 0) \\ &= \sum_{x \in \{0,1\}} P(X_t^{(1)} = 0, X_t^{(2)} = 0, X_0^{(ca)} = x) / P(X_t^{(2)} = 0) \\ &= P(X_t^{(1)} = 0, X_t^{(2)} = 0, X_0^{(ca)} = 0) / P(X_t^{(2)} = 0) \\ &\quad + P(X_t^{(1)} = 0, X_t^{(2)} = 0, X_0^{(ca)} = 1) / P(X_t^{(2)} = 0) \\ &= P(X_t^{(1)} = 0, X_t^{(2)} = 0 \,|\, X_0^{(ca)} = 0) \frac{P(X_0^{(ca)} = 0)}{P(X_t^{(2)} = 0)} \\ &\quad + P(X_t^{(1)} = 0, X_t^{(2)} = 0 \,|\, X_0^{(ca)} = 1) \frac{P(X_0^{(ca)} = 1)}{P(X_t^{(2)} = 0)} \\ &= \{\tfrac{1}{2}\left[\tfrac{1}{2} + \tfrac{1}{2}(1-2\alpha)^t\right]^2 + \tfrac{1}{2}\left[\tfrac{1}{2} - \tfrac{1}{2}(1-2\alpha)^t\right]^2\} / P(X_t^{(2)} = 0), \end{aligned}$$

waar gebruik is gemaakt van de *total probability law*, de definitie van conditionele kans, de conditionele onafhankelijkheid van de hedendaagse organismen gegeven hun laatst gemeenschappelijk voorouder, en de gegeven t-staps transitiematrix. Rest dus $P(X_t^{(2)} = \mathtt{0})$ te berekenen:

$$\begin{aligned} P(X_t^{(2)} = \mathtt{0}) &= P(X_t^{(2)} = \mathtt{0} \,|\, X_0^{(ca)} = \mathtt{0})P(X_0^{(ca)} = \mathtt{0}) + P(X_t^{(2)} = \mathtt{0} \,|\, X_0^{(ca)} = \mathtt{1})P(X_0^{(ca)} = \mathtt{1}) \\ &= \tfrac{1}{2}\left[\tfrac{1}{2} + \tfrac{1}{2}(1-2\alpha)^t\right] + \tfrac{1}{2}\left[\tfrac{1}{2} - \tfrac{1}{2}(1-2\alpha)^t\right]. \end{aligned}$$

Voeg samen en – voor de volledigheid – substitueer $\alpha = 0.25$.

Antwoord op vraag 3
Vraag 3a
De partiele correlatie:

$$\begin{aligned} \rho(Y_b, Y_c \,|\, Y_a) &= \rho(-Y_a + \gamma\varepsilon_b, Y_a + \gamma\varepsilon_c \,|\, Y_a) \\ &= \rho(\gamma\varepsilon_b, \gamma\varepsilon_c) = \gamma^2\rho(\varepsilon_b, \varepsilon_c) = 0. \end{aligned}$$

Vraag 3b
De covariantie matrix van het systeem is:

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1+\gamma^2 & -1 \\ 1 & -1 & 1+\gamma^2 \end{pmatrix}.$$

Dan:

$$\begin{aligned} \rho(Y_a, Y_c \,|\, Y_b) &= \frac{-(\Sigma^{-1})_{a,c}}{\sqrt{(\Sigma^{-1})_{a,a}}\sqrt{(\Sigma^{-1})_{c,c}}} = \frac{-[-(1+\gamma^2)+1]}{\sqrt{(1+\gamma^2)^2-1}\sqrt{(1+\gamma^2)-1}} \\ &= \frac{\gamma^2}{\sqrt{(1+\gamma^2)^2-1}\sqrt{\gamma^2}} = \frac{1}{\sqrt{2+\gamma^2}}. \end{aligned}$$

Vraag 3c
Definitie: factorizatie van de gezamelijke dichtheidsfunctie zodanig dat in het gewenste variabelen paar niet gezamenlijk in een factor van de factorizatie opduikt. Gevraagde relaties: $Y_a \not\perp\!\!\!\perp Y_b$, $Y_a \not\perp\!\!\!\perp Y_c$, $Y_b \not\perp\!\!\!\perp Y_b$, $Y_a \not\perp\!\!\!\perp Y_b \,|\, Y_c$, $Y_a \not\perp\!\!\!\perp Y_c \,|\, Y_b$, $Y_b \perp\!\!\!\perp Y_c \,|\, Y_a$.

Vraag 3d
Van het antwoord op de vorige deelvraag (3b) moge het duidelijk zijn dat de partiele correlatie naar nul gaat als γ erg groot wordt. Dit kan begrepen worden vanuit het regressie model: het signaal (de relatie tussen de genen) verdrinkt in de ruis.