Tentamen – Biostatistiek 3 / Biomedische wiskunde

25 maart 2014; 12:00-14:00

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)*
Om vast te stellen of een huidtumor goed- of kwaadaardig is, wordt een biopt afgenomen. Het biopt bevat enkel tumorcellen en wordt middenin in een petrischaal vol normale cellen geplaatst. De verspreiding van de tumorcellen door de petrischaal kan worden gemodelleerd m.b.v. een 1^{ste} orde Markov proces. Hiertoe is de petrischaal in een grid van vierkante millimeters (mm^2) opgedeeld. Op een mm^2 bevinden zich altijd vele cellen. Voor een mm^2 op enige afstand van het midden van de petrischaal wordt elk uur gekeken welke celtypen erin voorkomen. Indien de mm^2 door slechts één celtype bevolkt wordt, dan wordt het niet binnen één uur volledig overgenomen door het andere celtype. Wel is er dan een kans van 0.10 dat op het volgende tijdspunt beide celtypen in de mm^2 aanwezig zijn. Tenslotte, als een mm^2 het ene uur beide celtype bevat, dan is er een kans α (of β) dat bij de volgende observatie er enkel normale (of tumor) cellen in deze mm^2 huizen.

Vraag 1a)
Geef de toestandsruimte, startverdeling 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)
Een biopt wordt als kwaardaardig bestempeld als na lange tijd er meer dan 80% kans is dat een mm^2 enkel tumorcellen bevat. Voor welke α en β geschiedt dit?

Vraag 1d)
Wat is de kans dat op uur t een mm^2 enkel tumorcellen bevat gegeven dat het 2 uur later verschoond is van dit celtype? Neem hierbij aan dat het proces stationair is.

Vraag 1e)
Voor een mm^2 in de petrischaal is gedurende 20 uur de aanwezige celtypen gemeten:
`NNNNBBTTBBNBBBTTTTTT`,
waar `N` staat voor normaal, `T` voor tumor en `B` voor beide. Gebruik de maximum likelihood methode om α en β m.b.v. de gerapporteerde data te schatten.

Vraag 2 *(Netwerken)*
Beschouw een pathway van 3 genen. Expressie niveaus van de drie genen is gemeten in honderd individuen.

Vraag 2a)
Geef het volledige, enkelvoudige lineare regressie-model dat de expressie niveau's van gen 1 verklaard m.b.v. die van gen 2. Neem geen intercept in het model op.

Het regressie-model van vraag 2a is gefit op de data. De R-output staat hieronder:

```
          Coefficients:
                       Estimate  Std.Error  t-value   Pr(>|t|)
          Gen2          0.73044    0.05872    12.44  <2e-16  ***

---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1

Residual standard error:  2.471 on 99 degrees of freedom
Multiple R-squared:  0.6098, Adjusted R-squared:  0.6059
F-statistic:  154.7 on 1 and 99 DF, p-value:  < 2.2e-16
```

Vraag 2b)
Welke conclusies m.b.t. de conditionele onafhankelijkheden binnen het 3-gen pathway zijn gerechtvaardigd op basis van bovenstaande R-output? Motiveer!

Vraag 2c)
Geef aan hoe het antwoord op vraag 2b verandert als ook bekend is dat de correlatie tussen de expressie niveau's van gen 1 en 3 nul is? Motiveer!

Vraag 2d)
Geef aan hoe het antwoord op vraag 2b verandert als ook bekend is dat de partiële correlatie tussen de expressie niveau's van gen 1 en 3, geconditioneerd op die van gen 2, nul is? Motiveer!

Vraag 2e)
De expressie niveaus van de drie genen volgen een trivariate normale verdeling:

$$\begin{pmatrix} Y_{1,i} \\ Y_{2,i} \\ Y_{3,i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 14 & 12 & 3 \\ 12 & 16 & 4 \\ 3 & 4 & 3\frac{1}{2} \end{pmatrix} \right)$$

Welke conclusies m.b.t. de conditionele onafhankelijkheden binnen het 3-gen pathway zijn nu gerechtvaardigd?

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:

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

Antwoorden

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

$$\mathbf{P} = \begin{pmatrix} 0.9 & 0 & 0.1 \\ 0 & 0.9 & 0.1 \\ \alpha & \beta & 1-\alpha-\beta \end{pmatrix}$$

waarbij $0 \leq \alpha, \beta, \alpha + \beta \leq 1$. De bijbehorende startverdeling is $\boldsymbol{\pi} = (1, 0, 0)^\top$. Includeer ook het toestandsdiagram.

Vraag 1b
Gebruik: $\boldsymbol{\varphi}^T \mathbf{P} = \boldsymbol{\varphi}^T$ en $\varphi_n + \varphi_t + \varphi_b = 1$. Dit geeft het volgende stelsel van vergelijkingen:

$$\begin{aligned} 0.9\varphi_n + \alpha\varphi_b &= \varphi_n, \\ 0.9\varphi_t + \beta\varphi_b &= \varphi_t, \\ \varphi_n + \varphi_t + \varphi_b &= 1. \end{aligned}$$

De eerste twee vergelijkingen leveren: $\varphi_n = 10\alpha\varphi_b$ en $\varphi_t = 10\beta\varphi_b$. Substitutie hiervan in laatste vergelijking geeft:

$$(10\alpha + 10\beta + 1)\varphi_b = 1,$$

waaruit φ_b opgelost kan worden. Dit geeft ook meteen φ_n en φ_t:

$$\begin{aligned} \varphi_n &= 10\alpha/(10\alpha + 10\beta + 1), \\ \varphi_t &= 10\beta/(10\alpha + 10\beta + 1), \\ \varphi_b &= 1/(10\alpha + 10\beta + 1). \end{aligned}$$

Duidelijk, $\varphi_n + \varphi_t + \varphi_b = 1$.

Vraag 1c
Wanneer geldt: $10\beta/(10\alpha + 10\beta + 1) > 0.8$. Simpelweg oplossen van deze vergelijking levert:

$$10\beta > 0.8 * (10\alpha + 10\beta + 1).$$

Ofwel:

$$2\beta > 8\alpha + 0.8.$$

Vraag 1d
Gevraagd: $P(X_t = \mathtt{T} \,|\, X_{t+2} = \mathtt{N})$. De transitie $\mathtt{T} \to \mathtt{N}$ kan niet in één tijdstap plaatsvinden, de toestand `B` moet altijd tussenliggend aangedaan worden. Kortom, de enige mogelijk sequentie is $\mathtt{T} \to \mathtt{B} \to \mathtt{N}$. Dus:

$$\begin{aligned} P(X_t = \mathtt{T} \,|\, X_{t+2} = \mathtt{N}) &= \sum_{x_{t+1}} P(X_t = \mathtt{T}, X_{t+1} = x_{t+1} \,|\, X_{t+2} = \mathtt{N}) \\ &= P(X_t = \mathtt{T}, X_{t+1} = \mathtt{B} \,|\, X_{t+2} = \mathtt{N}). \end{aligned}$$

Nu herhaaldelijk de definitie van de conditionele kans toepassend:

$$\begin{aligned} P(X_t = \mathtt{T}, X_{t+1} = \mathtt{B} \,|\, X_{t+2} = \mathtt{N}) &= P(X_t = \mathtt{T}, X_{t+1} = \mathtt{B}, X_{t+2} = \mathtt{N})/P(X_{t+2} = \mathtt{N}) \\ &= P(X_{t+2} = \mathtt{N} \,|\, X_t = \mathtt{T}, X_{t+1} = \mathtt{B}) \\ &\quad \times P(X_{t+1} = \mathtt{B} \,|\, X_t = \mathtt{T}) P(X_t = \mathtt{T})/P(X_{t+2} = \mathtt{N}) \\ &= P(X_{t+2} = \mathtt{N} \,|\, X_{t+1} = \mathtt{B}) \\ &\quad \times P(X_{t+1} = \mathtt{B} \,|\, X_t = \mathtt{T}) P(X_t = \mathtt{T})/P(X_{t+2} = \mathtt{N}) \\ &= \alpha \times 0.1 \times \varphi_T/\varphi_N, \end{aligned}$$

waarin de 1ste Markov eigenschap is gebruikt alsmede de aanname van stationariteit.

Vraag 1e
De volgende sequentie van toestanden is geobserveerd: NNNNBBTTBBNBBBTTTTTT. Dan wordt de likelihood voor deze sequentie gegeven door: $P(X_1 = \mathtt{N}) \prod_{t=2}^{20} P(X_t \,|\, X_{t-1})$. Ofwel: $1 \times 0.9^3 \times 0.1 \times (1-\alpha-\beta) \times \beta \times 0.9 \times 0.1 \times (1-\alpha-\beta) \times \alpha \times 0.1 \times (1-\alpha-\beta)^2 \times \beta \times 0.9^5$. Versimpeld: $0.9^9 \times \alpha \times 0.1^3 \times (1-\alpha-\beta)^4 \times \beta^2$. Neem de logaritme en stel eerste orde afgeleide gelijk aan nul: $1/\alpha - 4/(1-\alpha-\beta) = 0$ (voor α) en $2/\beta - 4/(1-\alpha-\beta) = 0$ (voor β). Oftewel: $(1-\alpha-\beta) - 4\alpha = 0$ en $2(1-\alpha-\beta) - 4\beta = 0$. Dit heeft als oplossing $\hat{\alpha} = 1/7$ en $\hat{\beta} = 2/7$.

Answer to question 2

Vraag 2a
Het eenvoudige lineaire regressie model:

$$Y_{i,1} = \beta_{1,2} Y_{i,2} + \varepsilon_{i,1},$$

met $\varepsilon_{i,1} \sim \mathcal{N}(0, \sigma_1^2)$ en $\mathrm{Cov}(\varepsilon_{i_1,1}, \varepsilon_{i_2,1}) = 0$ als $i_1 \neq i_2$.

Vraag 2b
De uitgevoerde regressie modelleert enkel de marginale afhankelijkheid tussen de expressie niveaus van de twee genen. Op basis hiervan kan niet worden uitgesloten dat het derde gen verantwoordelijk is voor de geobserveerde associatie. Kortom, er kan geen uitspraak worden gedaan over conditionele (on)afhankelijkheid van (bijv.) gen 1 en 2.

Vraag 2c
De additionele informatie betreft wederom een marginaal verband (of de afwezigheid daarvan). Op basis hiervan, tesaam met de gegeven regressie, kan nog steeds geen uitspraak worden gedaan over conditionele (on)afhankelijkheid van (bijv.) gen 1 en 2.

Vraag 2d
De gegeven partiële correlatie zegt dat de expressie niveaus van gen 1 en 3 (conditioneel op die van gen 2) onafhankelijk zijn. Ergo, dit impliceert dat de corresponderende regressie coefficienten ook nul zijn. Kortom, inclusie van de expressie van gen 3 in de gegeven regressie veranderd de conclusie m.b.t. het effect van gen 2 niet. De volgende uitspraken zijn nu gerechtvaardigd: $Y_1 \perp\!\!\!\perp Y_3 \,|\, Y_2$ en $Y_1 \not\perp\!\!\!\perp Y_2 \,|\, Y_3$.

Vraag 2e

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

$$\boldsymbol{\Sigma}^{-1} \quad = \quad \begin{pmatrix} 0.20 & -0.15 & 0.00 \\ -0.15 & 0.20 & -0.10 \\ 0.00 & -0.10 & 0.40 \end{pmatrix}$$

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