Hertentamen – Voortgezette biostatistiek / Biomedische wiskunde

1 juni 2016; 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, mobiele telefoon of laptop. Succes!

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

Vraag 1 *(Markov modellen)*
Wanneer met de juiste tussenpozen gekeken wordt, laat de verandering in het aantal ($n = 0, \ldots, n_{max}$) exemplaren van een bepaald mRNA molecuul in de cel over de tijd ($t = 1, 2, 3, \ldots$) zich beschrijven door een 1^{ste} orde Markov keten. Tussen twee opeenvolgende tijdspunten wordt er met kans α één (en nooit meer) nieuw examplaar van het mRNA molecuul van het DNA afgeschreven. Gelijktijdig en onafhankelijk (!) wordt er met kans β één (en nooit meer) bestaand exemplaar van het mRNA molecuul afgebroken. Als $n = 0$ of $n = n_{max}$ wordt er geen exemplaar van het mRNA molecuul afgebroken danwel afgeschreven, respectievelijk, maar kan er natuurlijk wel worden afgeschreven ($n = 0$) danwel afgebroken ($n = n_{max}$) met voornoemde kansen.

Vraag 1a)
Zij $n_{max} = 3$. 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:* bereken eerst de kans dat het aantal toeneemt, afneemt en gelijk blijft, alvorens de transitie-matrix op te stellen.

Ga bij de resterende onderdelen van deze vraag uit van $n_{max} = 2$ en de volgende transitie-matrix :

$$\mathbf{P} = \begin{pmatrix} \gamma & 1-\gamma & 0 \\ \gamma & 1-\gamma-\delta & \delta \\ 0 & \delta & 1-\delta \end{pmatrix},$$

waarbij rijen- en kolommen-volgorde correspondeert met toestanden `0`, `1` en `2` mRNA moleculen in de cel. Merk op: deze transitie-matrix kent een andere parameterizatie (wat de betekenis van de elementen niet verandert).

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

Vraag 1c)
Gegeven is het volgende aantal examplaren van het mRNA molecuul in de cel over tien tijdspunten: `1, 2, 2, 2, 1, 2, 2, 1, 0, 1`. Gebruik de maximum likelihood methode om de parameters γ en δ 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
Twee hedendaagse organismen met een binaire genetische code (bestaande uit nullen en éénen) 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 α. De t-staps transitie-matrix is derhalve:

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

Dit substitutie-proces heeft een uniforme stationaire verdeling.

Vraag 2a)
Drie generaties scheidt de hedendaagse organismen van hun gemeenschappelijk voorouder. Wat is de kans dat een locus in beide hedendaagse organismen een één bevat (neem hierbij wederom stationariteit aan)? Neem aan dat $\alpha = 0.10$.

Vraag 2b)
Drie generatie scheidt de hedendaagse organismen van hun gemeenschappelijk voorouder. Gegeven dat een locus van één van de hedendaagse organismen een één bevat, wat is dan de kans, onder stationariteit, dat die locus in de gemeenschappelijke voorouder ook door een één bezet wordt (vergeet hierbij het bestaan van het andere hedendaagse organisme)?

Vraag 2c)
Drie generatie scheidt de hedendaagse organismen van hun gemeenschappelijk voorouder. Gegeven dat een locus van één van de hedendaagse organismen een één bevat terwijl diezelfde locus in de gemeenschappelijke voorouder door een nul bezet wordt (en vergeet wederom het andere hedendaagse organisme), wat is de kans dat die locus in de opa van dit hedendaagse organisme ook een één bevat?

Vraag 3
Beschouw een pathway van 3 genen. De expressie-niveau's van de drie genen, gerepresenteerd door de random variabelen Y_1, Y_2 en Y_3, zijn verdeeld als $(Y_1, Y_2, Y_3)^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ met $\boldsymbol{\mu} = (0,0,0)^\top$.

Vraag 3a)
Data van de expressie-niveau's zijn beschikbaar. Twee regressie-modellen, die de expressie niveaus's van Y_1 in termen van Y_2 danwel Y_2 en Y_3 verklaren, worden m.b.v. deze data gefit, nl.: $Y_1 = \beta_2 Y_2 + \varepsilon$ en $Y_1 = \tilde{\beta}_2 Y_2 + \tilde{\beta}_3 Y_3 + \tilde{\varepsilon}$. De `R`-output voor deze regressie fits is (respectievelijk)

```
    Estimate  Std.  Error  t value  Pr(>|t|)
Y2  -0.01021       0.10532   -0.097     0.923
```

en

```
    Estimate  Std.  Error  t value  Pr(>|t|)
Y2  -0.45198       0.08467   -5.338  6.08e-07
Y3   0.87230       0.08423   10.357   < 2e-16
```

Verklaar waarom het effect van Y_2 op Y_1 in het ene model wel en in het andere niet significant is.

Vraag 3b)

Neem nu aan dat:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix},$$

Reken de partiële correlatie tussen Y_1 en Y_2 gegeven Y_3 uit.

Vraag 3c)
Bepaal de verdeling van Y_1 conditioneel op Y_2 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

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

met $\boldsymbol{\mu}$ en $\boldsymbol{\Sigma}$ de verwachting en covariantie 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:

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

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

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

Antwoorden

Antwoord op vraag 1

Vraag 1a

De toestandsruimte $\mathcal{S}$ bestaat uit toestanden $\{0, 1, 2, 3\}$. De kans op een toename is gelijk aan de kans op afschrijven en afbraak: $P(\,+\,\mathtt{1}) = \alpha(1-\beta)$. Evenzo, de kans op een afname: $P(-\mathtt{1}) = (1-\alpha)\beta$. En, de kans op geen verandering in aantal $P(\mathtt{0}) = (1-\alpha)(1-\beta)+\alpha\beta$. De transitie-matrix (in de volgorde van de toestanden) wordt gegeven door:

$$\mathbf{P} \quad = \quad \begin{pmatrix} 1-\alpha & \alpha & 0 & 0 \\ (1-\alpha)\beta & (1-\alpha)(1-\beta)+\alpha\beta & \alpha(1-\beta) & 0 \\ 0 & (1-\alpha)\beta & (1-\alpha)(1-\beta)+\alpha\beta & \alpha(1-\beta) \\ 0 & 0 & \beta & (1-\beta) \end{pmatrix}.$$

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

Vraag 1b

Ja, heeft stationaire verdeling: irreducibel en aperiodiek. Gebruik dan: $\boldsymbol{\varphi}^\top \mathbf{P} = \boldsymbol{\varphi}^\top$ en $\varphi_0 + \varphi_1 + \varphi_2 = 1$. Dit geeft het volgende stelsel van vergelijkingen:

$$\begin{aligned} \gamma\varphi_0 + \gamma\varphi_1 &= \varphi_0, \\ (1-\gamma)\varphi_0 + (1-\gamma-\delta)\varphi_1 + \delta\varphi_2 &= \varphi_1, \\ \delta\varphi_1 + (1-\delta)\varphi_2 &= \varphi_2. \end{aligned}$$

De eerste vergelijking levert: $\gamma\varphi_1 = (1-\gamma)\varphi_0$. Net zo de derde vergelijking: $\delta\varphi_1 = \delta\varphi_2$. Ofwel, $\varphi_1 = \varphi_2$. Daar de stationaire kansen tesamen tot een moeten sommeren:

$$1 \quad = \quad \frac{\gamma}{1-\gamma}\varphi_1 + \varphi_1 + 1 \quad = \quad \frac{2-\gamma}{1-\gamma}\varphi_1.$$

Dus: $(\varphi_0, \varphi_1, \varphi_2) = (\gamma, 1-\gamma, 1-\gamma)/(2-\gamma)$. Deze kansen liggen in het interval $(0, 1)$ en sommeren tot een.

Vraag 1c

De likelihood van deze sequentie wordt gegeven door: $P(X_1 = \mathtt{1}) \prod_{t=2}^{10} P(X_t \,|\, X_{t-1})$. Ofwel: $1 \times (1-\delta) \times \delta \times (1-\delta) \times \delta \times \delta \times (1-\delta) \times \delta \times \gamma \times (1-\gamma) = \delta^4(1-\delta)^3\gamma(1-\gamma)$. Neem de logaritme, deze is proportioneel aan: $4\log(\delta) + 3\log(1-\delta) + \log(\gamma) + \log(1-\gamma)$. Stel eerste orde afgeleiden (naar γ en δ) gelijk aan nul: $4/\delta - 3/(1-\delta) = 0$ en $1/\gamma - 1/(1-\gamma) = 0$. Oplossen levert: $\hat{\delta} = 4/7$ en $\hat{\gamma} = 1/2$.

Antwoord op vraag 2

Vraag 2a

Gebruik de total probability law:

$$\begin{aligned} P(X_{t=3}^{(1)} = 1, X_{t=3}^{(2)} = 1) \quad &= \quad P(X_{t=3}^{(1)} = 0, X_{t=3}^{(2)} = 1 \,|\, X_{t=0}^{(ca)} = 0) P(X_{t=0}^{(ca)} = 0) \\ &\qquad + P(X_{t=3}^{(1)} = 1, X_{t=3}^{(2)} = 1 \,|\, X_{t=0}^{(ca)} = 1) P(X_{t=0}^{(ca)} = 1) \\ &= \quad P(X_{t=3}^{(1)} = 1 \,|\, X_{t=0}^{(ca)} = 0) P(X_{t=3}^{(2)} = 1 \,|\, X_{t=0}^{(ca)} = 0) P(X_{t=0}^{(ca)} = 0) \\ &\qquad + P(X_{t=3}^{(1)} = 1 \,|\, X_{t=0}^{(ca)} = 1) P(X_{t=3}^{(2)} = 1 \,|\, X_{t=0}^{(ca)} = 1) P(X_{t=0}^{(ca)} = 1) \\ &= \quad (0.244)^2 \frac{1}{2} + (0.756)^2 \frac{1}{2} \\ &= \quad 0.315536. \end{aligned}$$

Vraag 2b
Gebruik de regel van Bayes danwel definitie van conditionele kans:

$$\begin{aligned} P(X_{t=0}^{(ca)} = 1 \,|X_{t=3}^{(1)} = 1) &= P(X_{t=3}^{(1)} = 1 \,|X_{t=0}^{(ca)} = 1) P(X_{t=0}^{(ca)} = 1) / P(X_{t=3}^{(1)} = 1) \\ &= 0.756 \frac{1}{2} / P(X_{t=3}^{(1)} = 1), \end{aligned}$$

waar

$$\begin{aligned} P(X_{t=3}^{(1)} = 1) &= P(X_{t=3}^{(1)} = 1 \,|\, X_{t=0}^{(ca)} = 0) P(X_{t=0}^{(ca)} = 0) + P(X_{t=3}^{(1)} = 1 \,|\, X_{t=0}^{(ca)} = 1) P(X_{t=0}^{(ca)} = 1) \\ &= (0.244)/2 + (0.756)/2 \;=\; 1/2. \end{aligned}$$

De kans is dus 0.756.

Vraag 2c
De gevraagde kans (gebruik makende van definitie van de conditionele kans):

$$P(X_{t=1}^{(3)} = 1 \,|\, X_{t=3} = 1, X_{t=0}^{(ca)} = 0) \;=\; P(X_{t=3} = 1, X_{t=1}^{(3)} = 1, X_{t=0}^{(ca)} = 0) / P(X_{t=3} = 1, X_{t=0}^{(ca)} = 0)$$

in welke (gebruikmakende van de total probability law en de Markov eigenschap)

$$\begin{aligned} P(X_{t=3} = 1, X_{t=1}^{(3)} = 1, X_{t=0}^{(ca)} = 0) &= P(X_{t=3} = 1, X_{t=2}^{(4)} = 0, X_{t=1}^{(3)} = 1, X_{t=0}^{(ca)} = 0) \\ &\quad + P(X_{t=3} = 1, X_{t=2}^{(4)} = 1, X_{t=1}^{(3)} = 1, X_{t=0}^{(ca)} = 0) \\ &= 1/2 \times \alpha \times \alpha \times \alpha + 1/2 \times \alpha \times (1 - \alpha) \times (1 - \alpha) \end{aligned}$$

en

$$\begin{aligned} P(X_{t=3} = 1, X_{t=0}^{(ca)} = 0) &= P(X_{t=3} = 1 \,|\, X_{t=0}^{(ca)} = 0) P(X_{t=0}^{(ca)} = 0) \\ &= 0.244/2. \end{aligned}$$

Antwoord op vraag 3
Vraag 3a
In het eerste model wordt Y_1 enkel verklaard door Y_2. Hier wordt dus geen rekening gehouden met Y_3. Het afwezigheid van een effect van Y_2 op Y_1 kan mogelijk toegeschreven worden aan het feit dat dat verbloemd wordt door Y_3. Immers, als Y_2 en Y_3 gecorreleerd zijn en een tegengesteld effect op Y_1 hebben, kan dat het effect van Y_2 in het univariabele regressie model doen ondersneeuwen: geen significantie. Het multivariabele regressie model poogt daarentegen de variatie in Y_1 te verklaren in termen van Y_2 en Y_3 tesamen. De gecorreleerde, tegengestelde bijdragen van Y_2 en Y_3 kan nu ook ontrafeld worden. Dit heeft tot gevolg dat het effect van Y_2 op Y_1 goed geschat kan worden, en indien duidelijk ongelijk aan nul als significant aangemerkt wordt.

Vraag 3b
Voor de partiële correlatie matrix, neem eerst de inverse:

$$\mathbf{\Sigma}^{-1} \;=\; \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}^{-1} \;=\; \frac{1}{4} \begin{pmatrix} 3 & 1 & -2 \\ 1 & 3 & -2 \\ -2 & -2 & 4 \end{pmatrix}$$

Standardizeer deze matrix tot:

$$\begin{pmatrix} 1 & 1/3 & -2/\sqrt{12} \\ 1/3 & 1 & -2/\sqrt{12} \\ -2/\sqrt{12} & -2/\sqrt{12} & 1 \end{pmatrix}.$$

Rest nog de 'off-diagonal' elementen met -1 te vermenigvuldigen en de gevraagde partiële correlatie af te lezen: $\rho(Y_1, Y_2 \,|\, Y_3) = -1/3$.

Vraag 3c
Het formuleblad geeft deze verdeling in zijn algemene vorm. Vervang daar in $\mathbf{Y}$ en $\mathbf{X}$ door Y_1 en $(Y_2, Y_3)^\top$, respectievelijk. Dit geeft:

$$Y_1|(Y_2, Y_2)^\top \sim \mathcal{N}\{\mu_1 + \mathbf{\Sigma}_{1,23}\mathbf{\Sigma}_{23,23}^{-1}[(Y_2, Y_3)^\top - (\mu_2, \mu_3)^\top], \mathbf{\Sigma}_{1,1} - \mathbf{\Sigma}_{1,23}\mathbf{\Sigma}_{23,23}^{-1}\mathbf{\Sigma}_{23,1}\}.$$

Rest de termen in deze verdeling uit te werken. Voor de mean:

$$\begin{aligned} \mu_1 + \mathbf{\Sigma}_{1,23}\mathbf{\Sigma}_{23,23}^{-1}[(Y_2, Y_3)^\top - (\mu_2, \mu_3)^\top] &= 0 + (0,1)\left[\tfrac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\right]\left[\begin{pmatrix} Y_2 \\ Y_3 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right] \\ &= \frac{1}{3}(0,1)\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} Y_2 \\ Y_3 \end{pmatrix} \\ &= -\tfrac{1}{3}Y_2 + \tfrac{2}{3}Y_3. \end{aligned}$$

De variatie wordt:

$$\begin{aligned} \mathbf{\Sigma}_{1,1} - \mathbf{\Sigma}_{1,23}\mathbf{\Sigma}_{23,23}^{-1}\mathbf{\Sigma}_{23,1} &= 1 - (0,1)\left[\tfrac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\right]\begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= 1 - \tfrac{2}{3}. \end{aligned}$$

Hence, $Y_1|(Y_2, Y_2)^\top \sim \mathcal{N}(-\frac{1}{3}Y_2 + \frac{2}{3}Y_3, \frac{1}{3})$.