

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 $\pi = (0.5, 0.5)^T$.

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 $(\pi, \mathbf{P}, \mathbf{B})$ (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 $P((Y_1, Y_2, Y_3) = (\mathbf{C}, \mathbf{B}, \mathbf{A}))$. 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 $(Y_1, Y_2, Y_3) = (\text{C}, \text{B}, \text{A})$? 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 de regressie-vergelijking $Y_1 = \beta_1 Y_2 + \beta_2 Y_3 + \varepsilon$ 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-coëfficiënten dat $\beta_1 = 1 = \beta_2$.

Vraag 2c)

Bepaal $\text{Cov}(Y_2, Y_3)$ en $\text{Var}(Y_3)$.

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:

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

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

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