

Hertentamen – Voortgezette biostatistiek / Biomedische wiskunde

3 juni 2015; 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, 2a) 3, 2b) 3, 2c) 3, 3a) 3, 3b) 3, 3c) 3, 3d) 3.

Vraag 1 (*Markov modellen*)

Op de Waddeneilanden wordt het weer elke dag geclassificeerd als zonnig, bewolkt of regenachtig. Het weer voor de volgende dag hangt alleen af van het weer van vandaag, maar niet van de dagen daarvoor. Het weer wordt derhalve met een 1^{ste} orde Markov proces gemodelleerd. Als het een bewolkte dag is, dan zijn de respectievelijke kansen voor de volgende dag: β (regen), 2β (bewolkt) en β (zon). Regent het echter, dan schijnt de daaropvolgende dag de zon niet maar is het met kans α wel bewolkt. En mocht vandaag de zon schijnen, dan is het morgen met kans γ bewolkt of regent het met kans $\gamma/4$.

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.

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 weer op de Waddeneilanden.

Vraag 1d)

Gegeven is het volgende verloop van een willekeurige doch representatieve week: **regen, regen, bewolkt, zon, zon, regen, bewolkt**. Gebruik de maximum likelihood methode om de vrije parameters van het proces op basis van dit verloop te schatten.

Vraag 2 (*Hidden Markov model*)

Beschouw een ternaire DNA sequentie (bestaande uit elementen A, B, en C). De geobserveerde sequentie (van A's, B's, en C's) kan worden verklaard vanuit de functionaliteit van het DNA. Drie functionele elementen worden onderscheiden: **promotor**, **intron**, en **exon**. De verdeling van A's, B's, en C's kan worden gemodelleerd m.b.v. een hidden Markov model (HMM), waarbij de func-

tionele elementen 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 worden gegeven door $\boldsymbol{\pi} = (1, 0, 0)^T$,

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \text{en} \quad \mathbf{B} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

De rijen van de matrices $\mathbf{P}$ en $\mathbf{B}$ representeren de functionele elementen (in de volgorde zoals boven gespecificeerd). De kolommen van $\mathbf{B}$ corresponderen met van A, B, en C (in deze volgorde).

Vraag 2a

Bereken $P((Y_1, Y_2, Y_3) = (\mathbf{A}, \mathbf{B}, \mathbf{C}))$ en $P((Y_1, Y_2, Y_3) = (\mathbf{B}, \mathbf{B}, \mathbf{B}))$.

Vraag 2b

Wat is de kans op $X_2 = \text{intron}$ gegeven de geobserveerde sequentie $(Y_1, Y_2, Y_3) = (\mathbf{A}, \mathbf{B}, \mathbf{C})$, i.e. $P(X_2 = \text{intron} | (Y_1, Y_2, Y_3))$? Noem deze kans voor het vervolg $p_{\mathbf{i}}$.

Analoog aan $p_{\mathbf{i}}$ zou men $p_{\mathbf{e}} = P(X_2 = \text{exon} | (Y_1, Y_2, Y_3))$ en $p_{\mathbf{p}} = P(X_2 = \text{promotor} | (Y_1, Y_2, Y_3))$ kunnen bepalen.

Vraag 2c (bonusvraag!)

De meest aannemelijk toestandssequentie die ten grondslag ligt aan een geobserveerde DNA sequentie is m.b.v. het Viterbi algoritme bepaald. Komt de meest aannemelijke toestand op positie $t = 2$ uit deze sequentie altijd overeen met de toestand welk $p_{\mathbf{i}}$, $p_{\mathbf{e}}$ en $p_{\mathbf{p}}$ maximalizeert? Motiveer.

Vraag 3 (Netwerken)

Beschouw een pathway van drie genen dat het signaal van gen A verwerkt. Gesteld dat de onafhankelijkheidsrelaties tussen de expressie van drie genen (gerepresenteerd door random variabelen Y_a, Y_b, Y_c), wordt gegeven door de graaf $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ met $\mathcal{V} = \{a, b, c\}$ en $\mathcal{E} = \{(a, b), (b, c)\}$.

Vraag 3a)

Gesteld dat het regressie model $Y_a = \beta_b Y_b + \beta_c Y_c + \varepsilon_a$ wordt gefit. Wat voor waarde voor de schatter van β_c verwacht je dan?

Vergeet de bovenvermelde onafhankelijkheidsrelaties. Veronderstel daarentegen dat de verdelingen van de expressies van de genen in het pathway gegeven worden door: $Y_a \sim \mathcal{N}(0, \sigma_a^2)$, $Y_c \sim \mathcal{N}(0, \sigma_c^2)$ en $Y_b | \{Y_a, Y_c\} \sim \mathcal{N}(Y_a + 2Y_c, \sigma_c^2)$, waarbij Y_a, Y_c en de fout in Y_b , onafhankelijk zijn.

Vraag 3b)

Bereken de correlatie tussen de (expressie niveau's van de) genen B en C.

Vraag 3c)

Bepaal de partiële correlatie tussen de (expressie niveau's van de) genen B en C.

Vraag 3d)

Het blijkt experimenteel mogelijk de expressie van gen A te controleren. Voor welke waarde van Y_a wordt de resulterende variantie in de andere twee genen geminimaliseerd? *Hint*: Conditioneer binnen een trivariate normale verdeling!

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 dichtheidsfunctie van de multivariaat normale verdeling van random variable $\mathbf{Y}$ is

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

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:

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

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 $B_1, \dots, B_n$ gebeurtenissen z.d.d. $\sum_{i=1}^n P(B_i) = 1$. De *total probability law* zegt dan:

$$P(A) = \sum_{i=1}^n P(A, B_i).$$