Algebraische Statistik – ein junges Forschungsgebiet

Dipl.-Math. Marcus Weber

Disputationsvortrag

15. Februar 2006

Gliederung

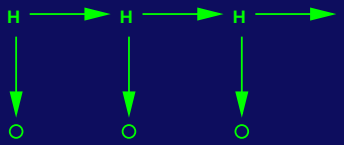
1. Statistische Modelle
2. Algebraische Interpretation statistischer Probleme
3. Der Buchberger-Algorithmus
4. Diskussion bisheriger Resultate

1. Statistische Modelle

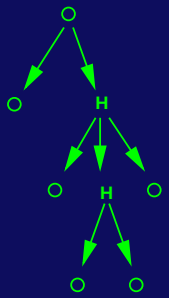
Statistisches Modell

Modellierung bedingter Wahrscheinlichkeiten versteckter (H) und beobachtbarer (O) Ereignisse

Markov Chain Model



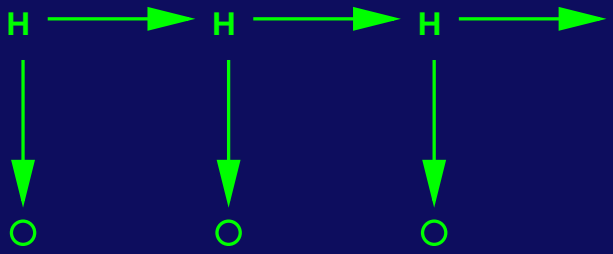
Hidden Markov Model



Tree Model

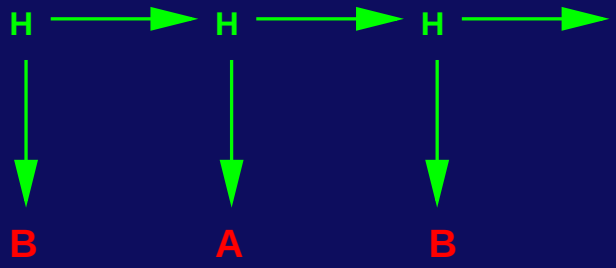
Beobachtung

Beispiel: Hidden Markov Model mit $O \in \{A, B, C\}$ und $H \in \{a, b\}$.



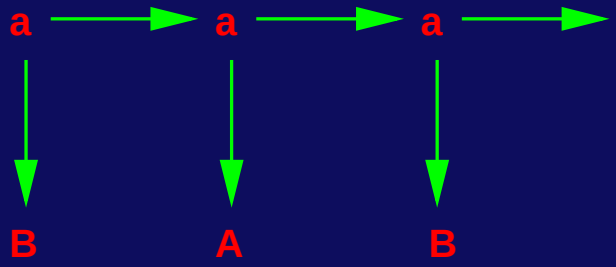
Beobachtung

Beispiel: Hidden Markov Model mit $O \in \{A, B, C\}$ und $H \in \{a, b\}$.



Beobachtung

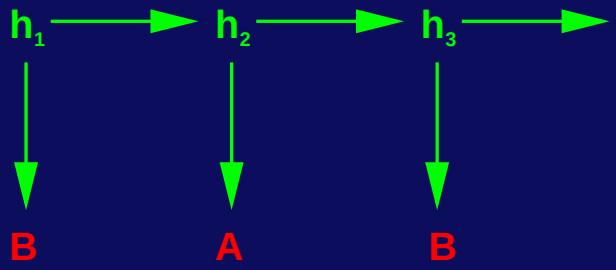
Beispiel: Hidden Markov Model mit $O \in \{A, B, C\}$ und $H \in \{a, b\}$.



$$w_{a \to B} \cdot w_{a \to a} \cdot w_{a \to A} \cdot w_{a \to a} \cdot w_{a \to B}$$

Beobachtung

Beispiel: Hidden Markov Model mit $O \in \{A, B, C\}$ und $H \in \{a, b\}$.



$$w_{BAB} = \sum_{h_1,h_2,h_3 \in \{a,b\}} w_{h_1 \to B} \cdot w_{h_1 \to h_2} \cdot w_{h_2 \to A} \cdot w_{h_2 \to h_3} \cdot w_{h_3 \to B}$$

Algebraisches Statistisches Modell

$$w_{BAB} = \sum_{h_1,h_2,h_3 \in \{a,b\}} w_{h_1 \to B} \cdot w_{h_1 \to h_2} \cdot w_{h_2 \to A} \cdot w_{h_2 \to h_3} \cdot w_{h_3 \to B}$$

$w_{H \to H}$	a	b
a	θ_1	θ_2
b	θ_3	θ_4

$w_{H \to O}$	A	B	C
a	θ_5	θ_6	θ_7
b	θ_8	θ_9	θ_{10}

$$\mathbf{p} : \Delta_{2\times 2,2\times 3} \subset I\!R^{10} \to \Delta_{27} \subset I\!R^{27}$$

Algebraisches Statistisches Modell

$$w_{BAB} = \sum_{h_1,h_2,h_3 \in \{a,b\}} w_{h_1 \to B} \cdot w_{h_1 \to h_2} \cdot w_{h_2 \to A} \cdot w_{h_2 \to h_3} \cdot w_{h_3 \to B}$$

$w_{H \to H}$	a	b
a	θ_1	θ_2
b	θ_3	θ_4

$w_{H \to O}$	A	B	C
a	θ_5	θ_6	θ_7
b	θ_8	θ_9	θ_{10}

$\mathbf{p} : \mathbb{C}^{10} \to \mathbb{C}^{27}$

$\mathbf{p}_i(\Theta)$ ist ein Polynom in $\Theta = [\theta_1, \ldots, \theta_{10}]$.

2. Algebraische Interpretation statistischer Probleme

Wahrscheinlichste Pfade

$$w_{BAB} = \sum_{h_1,h_2,h_3 \in \{a,b\}} w_{h_1 \to B} \cdot w_{h_1 \to h_2} \cdot w_{h_2 \to A} \cdot w_{h_2 \to h_3} \cdot w_{h_3 \to B}$$

Wahrscheinlichste Pfade

$$w_{BAB} = \sum_{h_1,h_2,h_3 \in \{a,b\}} w_{h_1 \to B} \cdot w_{h_1 \to h_2} \cdot w_{h_2 \to A} \cdot w_{h_2 \to h_3} \cdot w_{h_3 \to B}$$

$$w^*_{BAB} = \max_{h_1,h_2,h_3 \in \{a,b\}} w_{h_1 \to B} \cdot w_{h_1 \to h_2} \cdot w_{h_2 \to A} \cdot w_{h_2 \to h_3} \cdot w_{h_3 \to B}$$

Wahrscheinlichste Pfade

$$w_{BAB} = \sum_{h_1,h_2,h_3 \in \{a,b\}} w_{h_1 \to B} \cdot w_{h_1 \to h_2} \cdot w_{h_2 \to A} \cdot w_{h_2 \to h_3} \cdot w_{h_3 \to B}$$

$$w^*_{BAB} = \max_{h_1,h_2,h_3 \in \{a,b\}} w_{h_1 \to B} \cdot w_{h_1 \to h_2} \cdot w_{h_2 \to A} \cdot w_{h_2 \to h_3} \cdot w_{h_3 \to B}$$

$$l^*_{BAB} = \max_{h_1,h_2,h_3 \in \{a,b\}} l_{h_1 \to B} + l_{h_1 \to h_2} + l_{h_2 \to A} + l_{h_2 \to h_3} + l_{h_3 \to B}$$

Wahrscheinlichste Pfade

$$w_{BAB} = \sum_{h_1,h_2,h_3 \in \{a,b\}} w_{h_1 \to B} \cdot w_{h_1 \to h_2} \cdot w_{h_2 \to A} \cdot w_{h_2 \to h_3} \cdot w_{h_3 \to B}$$

$$w^*_{BAB} = \max_{h_1,h_2,h_3 \in \{a,b\}} w_{h_1 \to B} \cdot w_{h_1 \to h_2} \cdot w_{h_2 \to A} \cdot w_{h_2 \to h_3} \cdot w_{h_3 \to B}$$

$$l^*_{BAB} = \max_{h_1,h_2,h_3 \in \{a,b\}} l_{h_1 \to B} + l_{h_1 \to h_2} + l_{h_2 \to A} + l_{h_2 \to h_3} + l_{h_3 \to B}$$

$$l^*_{BAB} = \bigoplus_{h_1,h_2,h_3 \in \{a,b\}} l_{h_1 \to B} \odot l_{h_1 \to h_2} \odot l_{h_2 \to A} \odot l_{h_2 \to h_3} \odot l_{h_3 \to B}$$

Wahrscheinlichste Pfade

$$w_{BAB} = \sum_{h_1,h_2,h_3 \in \{a,b\}} w_{h_1 \to B} \cdot w_{h_1 \to h_2} \cdot w_{h_2 \to A} \cdot w_{h_2 \to h_3} \cdot w_{h_3 \to B}$$

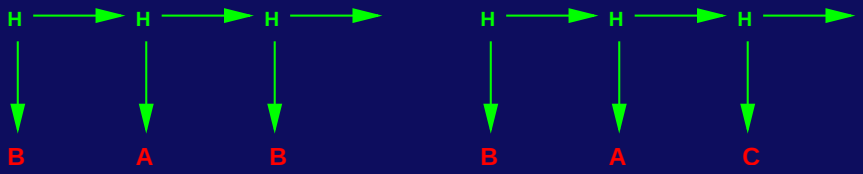
$$w^*_{BAB} = \max_{h_1,h_2,h_3 \in \{a,b\}} w_{h_1 \to B} \cdot w_{h_1 \to h_2} \cdot w_{h_2 \to A} \cdot w_{h_2 \to h_3} \cdot w_{h_3 \to B}$$

$$l^*_{BAB} = \max_{h_1,h_2,h_3 \in \{a,b\}} l_{h_1 \to B} + l_{h_1 \to h_2} + l_{h_2 \to A} + l_{h_2 \to h_3} + l_{h_3 \to B}$$

$$l^*_{BAB} = \bigoplus_{h_1,h_2,h_3 \in \{a,b\}} l_{h_1 \to B} \odot l_{h_1 \to h_2} \odot l_{h_2 \to A} \odot l_{h_2 \to h_3} \odot l_{h_3 \to B}$$

Geschicktes Ausrechnen (Distributiv-Gesetz, Horner-Schema) dieses "Polynoms": Viterbi-Algorithmus (Hidden Markov Models), Needlemann-Wunsch-Algorithmus (Sequenz-Alignment-Graph)

Modell-Invarianten



Zusammenhänge zwischen $w_{BAB}, w_{BAC}, \ldots$, die unabhängig von der Wahl der Parameter Θ sind, sondern nur das Statistische Modell betreffen? Z.B.:

$$w_{AAA} - w_{BAC} = w_{CCC} - w_{CAB}$$

Modell-Invarianten

$$\mathbf{p} : \mathbb{C}^{10} \to \mathbb{C}^{27}$$

$$\begin{aligned} w_{AAA} &= \mathbf{p}_1(\Theta) \\ w_{AAB} &= \mathbf{p}_2(\Theta) \\ &\vdots \\ w_{CCC} &= \mathbf{p}_{27}(\Theta) \end{aligned}$$

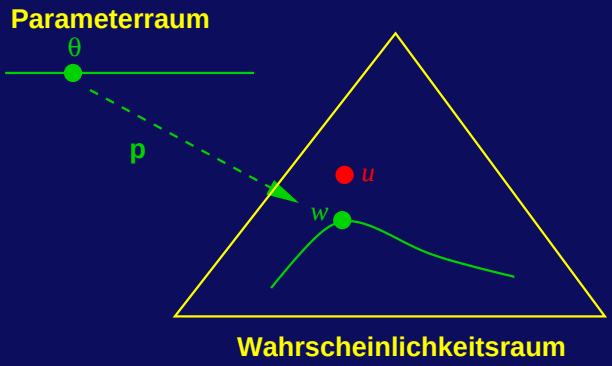
Modell-Invarianten

$$\mathbf{p} : \mathbb{C}^{10} \to \mathbb{C}^{27}$$

$$\begin{aligned} 0 &= w_{AAA} - \mathbf{p}_1(\Theta) \\ 0 &= w_{AAB} - \mathbf{p}_2(\Theta) \\ &\vdots \\ 0 &= w_{CCC} - \mathbf{p}_{27}(\Theta) \end{aligned}$$

27 algebraische Gleichungen mit 37=27+10 "Unbekannten". Varietät $\mathcal{V}$.
Verfahren: Elimination der 10 Parameter ("Eliminationstheorie").

Parameterschätzung, Log-Likelihood



Parameterschätzung, Log-Likelihood

Maximiere bezüglich Θ die Log-Likelihood-Funktion:

$$L_u(\Theta) = \sum_{i=1}^{27} u_i \, \log(\mathbf{p}_i(\Theta))$$

Algebraischer Ansatz: Berechne *alle* kritischen Punkte von L_u, auch alle komplexwertigen.

$$0 \quad = \quad \frac{\partial L_u(\Theta)}{\partial \theta_j} = \sum_{i=1}^{27} \frac{u_i}{\mathbf{p}_i(\Theta)} \frac{\partial \mathbf{p}_i(\Theta)}{\partial \theta_j}, \quad j = 1, \ldots, 10.$$

Parameterschätzung, Log-Likelihood

Maximiere bezüglich Θ die Log-Likelihood-Funktion:

$$L_u(\Theta) = \sum_{i=1}^{27} u_i \, \log(\mathbf{p}_i(\Theta))$$

Algebraischer Ansatz: Berechne *alle* kritischen Punkte von L_u, auch alle komplexwertigen.

$$0 = \frac{\partial L_u(\Theta)}{\partial \theta_j} = \sum_{i=1}^{27} u_i \, z_i \, \frac{\partial \mathbf{p}_i(\Theta)}{\partial \theta_j}, \qquad j = 1, \ldots, 10,$$

$$0 = z_i \mathbf{p}_i(\Theta) - 1, \qquad i = 1, \ldots, 27.$$

3. Der Buchberger-Algorithmus

Varietäten

$$\begin{aligned} 0 &= p_1(x_1, x_2, \ldots, x_m) \\ 0 &= p_2(x_1, x_2, \ldots, x_m) \\ &\vdots \\ 0 &= p_n(x_1, x_2, \ldots, x_m) \end{aligned}$$

$$\mathcal{V}(p_1, \ldots, p_n) \subseteq \mathbb{C}^m$$

Idee der Gauß-Elimination:
Umformung eines *linearen* Gleichungssystems in ein äquivalentes *lineares* System "einfacherer" Gestalt.

$$
\begin{array}{rrrrcrrrr}
0 = & 2x_1 & -5x_2 & +3x_3 & -4 & & 0 = & x_1 & -\frac{5}{2}x_2 & +\frac{3}{2}x_3 & -2 \\
0 = & 4x_1 & +x_2 & +x_3 & +1 & \Leftrightarrow & 0 = & & x_2 & -\frac{8}{11}x_3 & +\frac{9}{11} \\
0 = & x_1 & -x_2 & -x_3 & -6 & & 0 = & & & x_3 & +\frac{23}{8}
\end{array}
$$

Idee des Buchberger-Algorithmus

Idee des Buchberger-Algorithmus (Buchberger, 1965):
Umformung eines *algebraischen* Gleichungssystems in ein äquivalentes *algebraisches* System "einfacherer" Gestalt.

Idee des Buchberger-Algorithmus

Idee des Buchberger-Algorithmus (Buchberger, 1965):
Umformung eines *algebraischen* Gleichungssystems in ein äquivalentes *algebraisches* System "einfacherer" Gestalt.

$$
\begin{aligned}
0 = p(x),\ 0 = q(x) \quad &\Rightarrow \quad 0 = p(x) \cdot r(x), \quad \forall r \in \mathbf{K}[X_1, X_2, \ldots] \\
&\Rightarrow \quad 0 = p(x) + q(x).
\end{aligned}
$$

Idee des Buchberger-Algorithmus

Idee des Buchberger-Algorithmus (Buchberger, 1965):
Umformung eines *algebraischen* Gleichungssystems in ein äquivalentes *algebraisches* System "einfacherer" Gestalt.

$$\begin{array}{rcl} & \Rightarrow & 0 = p(x) \cdot r(x), \quad \forall r \in \mathbf{K}[X_1, X_2, \ldots] \\ 0 = p(x),\ 0 = q(x) & & \\ & \Rightarrow & 0 = p(x) + q(x). \end{array}$$

Varietät eines Ideals:

$$\mathcal{V}(\mathcal{F}) = \mathcal{V}(\langle \mathcal{F} \rangle) = \mathcal{V}(\langle \mathcal{G} \rangle) = \mathcal{V}(\mathcal{G})$$

$$\mathcal{V}(\langle 2x^3 - 5x^2 - 8x + 20,\ x^3 + x^2 - 4x - 4 \rangle) = \mathcal{V}(\langle x^2 - 4 \rangle) = \{2, -2\}$$

Euklidische Ringe

Beispiel: Ideal einer Familie $\mathcal{F}$ ganzer Zahlen.

$$\langle 198, 108\rangle = \langle ggT(198, 108)\rangle = \langle 18\rangle.$$

Euklidischer Algorithmus: Größter gemeinsamer Teiler kann durch sukzessive Division mit Rest (wobei der "Grad" des Restgliedes abnimmt) bestimmt werden.

$$\begin{aligned} 198 &= 108 \cdot 1 + 90 \\ 108 &= 90 \cdot 1 + 18 \\ 90 &= 18 \cdot 5 + 0 \end{aligned}$$

Hauptidealringe: *Euklidische Ringe sind Hauptidealringe.* Beipiel: Polynomring in einer Variablen mit "Grad"=höchster vorkommender Exponent des Polynoms.

$$\mathcal{V}(\langle 2x^3 - 5x^2 - 8x + 20, x^3 + x^2 - 4x - 4\rangle) = \mathcal{V}(\langle x^2 - 4\rangle) = \{2, -2\}$$

Polynomringe mehrerer Variablen

Polynomringe in *mehreren* Variablen sind *keine* euklidischen Ringe. Sprich: Nicht jedes Ideal wird von einem einzigen Polynom erzeugt.

Verallgemeinerte Division mit Rest: Sei $\mathcal{F} = \{p_1, p_2, \ldots, p_n\} \subset \mathbf{K}[X_1, X_2, \ldots]$ gegeben, dann existiert zu jedem $p \in \mathbf{K}[X_1, X_2, \ldots]$ eine Zerlegung

$$p = a_1 p_1 + a_2 p_2 + \ldots + a_n p_n + r$$

mit Polynomen $a_1, \ldots, a_n, r \in \mathbf{K}[X_1, X_2, \ldots]$.
Und: $a_i p_i \neq 0 \Rightarrow \mathrm{mulgrad}(p) \geq \mathrm{mulgrad}(a_i p_i)$.

Gröbner-Basen

$p = xy^2 - x, \quad \mathcal{F} = \{xy + 1, y^2 - 1\}$

$$xy^2 - x = y(xy + 1) + 0(y^2 - 1) \quad + (-x - y)$$

$$xy^2 - x = x(y^2 - 1) + 0(xy + 1) \quad + 0$$

Gibt es eine Basis $\mathcal{G}$ des Ideals $\langle \mathcal{F} \rangle$, so dass der Rest bei verallgemeinerter Division eindeutig bestimmt ist?

Antwort ja: "Gröbner-Basen". Buchberger-Algorithmus (verallgemeinert Gauß-Elimination und Euklidischen Algorithmus)

Buchberger-Algorithmus

0. $\mathcal{F}$ Erzeugendensystem eines Ideals. $\mathcal{G} := \mathcal{F}$.
1. Erweiterung auf eine Gröbner-Basis:
   a) Konstruiere Testpolynome aus $\langle \mathcal{F} \rangle$ (siehe Gauß-Eliminationsschritt) und führe unter $\mathcal{G}$ die verallgemeinerte Division durch.
   b) Ist der Rest jeweils 0, dann ist die Gröbner-Basis ausreichend, gehe zu 2. Andernfalls füge die Nicht-Null-Reste zu $\mathcal{G}$ hinzu und wiederhole 1.
2. Normiere die Polynome aus $\mathcal{G}$ und reduziere (auf bestimmte Weise) die Gröbner-Basis.

4. Diskussion bisheriger Resultate

Einige algebraische Resultate

- **Colin Dewey, Peter Huggins, Kevin Woods, Bernd Sturmfels, Lior Pratcher:** Sensitivitätsanalyse des parametrisierten Sequenz-Alignments durch Polytop-Algebra.
- **Colin Dewey, Kevin Woods:** Biologisch korrektes Alignment, das für keine Parameterwahl des statistischen Modells (Sequenz-Alignment-Graph) optimal ist.
- **Marta Casanellas, Luis David Garcia, Seth Sullivant:** Katalog von Eigenschaften "kleiner" Bäume im Web, Auswahl des richtigen statistischen Modells für phylogenetische Bäume anhand dieser Daten möglich.
- **Sergi Elizade:** "Few inference functions theorem".

Zusammenfassung

Statistisches Problem	Algebraische Formulierung	Lösung
wahrsch. Pfade	(max,+)-Algebra	geschicktes Ausklammern
Modell-Invarianten	Eliminationstheorie	Buchberger-Algorithmus
Max-Likelihood	Varietäten	Buchberger-Algorithmus

Verwendete Literatur

Lior Patcher, Bernd Sturmfels (eds.): *Algebraic Statistics for Computational Biology,* Cambridge University Press, Oktober 2005.

Hans J. Stetter: *Numerical Polynomial Algebra,* Society for Industrial and Applied Mathematics, Philadelhia, 2004.

David Cox, John Little, and Donal O'Shea: *Ideals, Varieties and Algorithms. An Introduction to Computational Algebraic Geometry And Commutative Algebra*, 2nd ed., Undergraduated Texts in Mathematics, Springer-Verlag, New York, 1997.