Asymptotisches Verhalten des Erwartungswertes für den größten Wert bei n unabhängigen Beobachtungen einer normalverteilten Variablen

von

Nikolaus* und Rolf Dieter† Grigorieff

Nr. 647 1999

*Brandeis University, Waltham, MA 02454-9110, USA, e-mail: niko@brandeis.edu

†Technische Universität Berlin, 10623 Berlin, Germany, E-Mail: grigo@math.tu-berlin.de

Asymptotisches Verhalten des Erwartungswertes für den größten Wert bei n unabhängigen Beobachtungen einer normalverteilten Variablen

N. Grigorieff * und R.D. Grigorieff †

Einer der großen Fortschritte der gegenwärtigen Biophysik besteht in der Entwicklung ständig raffinierterer Meßverfahren, die es gestatten, in immer feinere Strukturen hineinzublicken. Eine der Methoden, die dabei verwendet werden, besteht in der Bestimmung der gesuchten Struktur mit Hilfe einer großen Zahl von Messungen. Jede einzelne für sich genommen, ist viel zu unscharf, etwa verrauscht oder anderweitig überlagert, als daß die gewünschten Daten herausgelesen werden könnten. Aber stehen mehrere Messungen zur Verfügung, so kann man versuchen, die in jeder einzelnen von ihnen ja doch enthaltenen Informationen über die zu bestimmende Struktur durch Ausrichtung der zweidimensionalen Meßfenster relativ zueinander zu verstärken und auf diese Weise aus einer gewissen Zahl unabhängiger Messungen zufällige Signale herauszufiltern und so das gesuchte Bild herauszupräparieren. In der modernen Strukturbiologie wird ein solches Verfahren angewendet, um die dreidimensionale Gestalt eines Proteinmoleküls oder -komplexes zu bestimmen, indem man Bilder einzelner Moleküle oder Komplexe im Elektronenmikroskop aufnimmt [1]. Da die Proteine im Elektronenstrahl sehr schnell zerfallen, kann man nur Bilder mit geringer Strahlendosis aufnehmen. Solche Bilder haben im allgemeinen ein sehr kleines Signal-Rausch-Verhältnis. Daher muß über mehrere Tausend zueinander ausgerichtete Einzelbilder gemittelt werden, bevor Details der Struktur sichtbar werden.

Mit dem Ziel, die Auflösung ständig weiter zu steigern, wird somit eine immer größere Zahl von Einzelmessungen getätigt, und der Prozeß des gegenseitigen Ausrichtens umfaßt Ensembles, bestehend aus in die Tausende gehenden Elementen. Wird die Methode in diesen extremen Bereich getrieben, so stellt sich naturgemäß die Frage, ob sie weiterhin verläßlich ist. Zum Beispiel stellt sich heraus, daß zwei Elemente, die ausschließlich durch eine Normalverteilung beschriebenes Rauschen enthalten, so ausgerichtet werden können, daß man einen von Null verschiedenen Korrelationskoeffizienten bekommt. Werden keine anderen Kriterien betrachtet, so könnte dieser von Null verschiedene Korrelationskoeffizient als

*Brandeis University, Waltham, MA 02454-9110, USA. E-mail: niko@brandeis.edu

†Technische Universität Berlin, MA 6-4, Straße d. 17. Juni 135, 10623 Berlin, Germany. E-Mail: grigo@math.tu-berlin.de

ein schwaches Signal aufgefaßt werden, welches sich bei der Aufsummierung von Tausenden von ausgerichteten Elementen immer weiter verstärkt. In der Strukturbiologie kann der beschriebene Effekt zu falschen Ergebnissen führen, wenn die Auflösung der rekonstruierten Proteinstruktur abgeschätzt werden soll. Verschiedene Verfahren sind in Gebrauch: Die Fourier Shell Correlation mißt die Korrelation zwischen zwei getrennt ausgerechneten Strukturen [4]. Ein zweites Verfahren, die Spectral Signal-To-Noise Ratio, schätzt das Verhältnis der Varianzen von Signal und Rauschen [5]. Eine detailliertere Erörterung dieser Fragestellungen und neue Ergebnisse bei der Strukturbestimmung von Proteinen findet man in [2].

Wenn die Elemente aus einer genügend großen Zahl von Einzelmessungen bestehen (z.B. Pixel in einem Bild), dann läßt sich der Korrelationskoeffizient zweier Rauschelemente in guter Näherung durch eine Normalverteilung beschreiben. Es wird angenommen, daß man beim Ausrichten der Rauschelemente n mögliche Transformationen hat, die n verschiedene Korrelationskoeffizienten liefern. Zur Ausrichtung wird dann die Transformation mit dem größten Korrelationskoeffizienten verwendet. Das Ziel dieser Note ist, eine wahrscheinlichkeitstheoretische Formel herzuleiten, die den Wert dieses größten Korrelationskoeffizienten angibt. Es wird eine normal verteilte Variable mit Mittelwert Null und Varianz σ betrachtet. Berechnet werden soll der Erwartungswert des größten Wertes r von n unabhängigen Beobachtungen der Variablen. Diese Größe läßt sich auch als Erwartungswert verstehen, daß n unabhängige Beobachtungen der Variablen nicht alle gleichzeitig kleiner als r sind. In der vorliegenden Untersuchung ist das Verhalten des Erwartungswertes für große n von Interesse.

Es wird der Fall, daß die Varianz $\sigma = 1$ ist, betrachtet. Der allgemeine Fall läßt sich durch eine Variablentransformation im Integral darauf zurückführen. Sei also

$$p(r) := \frac{1}{\sqrt{2\pi}} e^{-r^2/2} \quad \text{und} \quad P(r) := \int_{-\infty}^{r} p(s)ds.$$

Gesucht ist das asymptotische Verhalten für $n \to \infty$ des Erwartungswertes

$$\overline{r}_n := \int_{-\infty}^{\infty} r p_n(r) dr \tag{1}$$

mit

$$p_n(r) := \frac{dP^{n+1}(r)}{dr} = (n+1)P^n(r)p(r). \tag{2}$$

Der hier untersuchten Fragestellung verwandt ist die Bestimmung des asymptotischen Verhaltens für große n des Maximums $\tilde{r}_n$ der Verteilung $p_n(r)$ aus (2), die in [3] erfolgt ist. Das dort gefundene Ergebnis (siehe [3], Formel 4.2.3(11)) ist das gleiche wie für $\bar{r}_n$ in (4), nämlich

$$\tilde{r}_{n-1} = \sqrt{2\ln(0,4n)} \quad \text{für} \quad n \to \infty. \tag{3}$$

Der Beweis für (3) ist einfacher zu führen als für das Resultat (4).

Das Ergebnis dieser Note ist der folgende

Satz *Das asymptotische Verhalten des Erwartungswertes $\bar{r}_n$ aus (1) ist*

$$\bar{r}_n = \sqrt{2\ln n} \quad \text{für} \quad n \to \infty. \tag{4}$$

B e w e i s : Im folgenden wird von dem für $r > 0$ bestehenden Zusammenhang

$$1 - P(r) = \frac{1}{\sqrt{2\pi}} \int_r^\infty e^{-s^2/2} ds = \frac{p(r)}{r}(1 - R(r)) \tag{5}$$

mit

$$0 < R(r) < \frac{1}{r^2}$$

Gebrauch gemacht, den man durch einmalige partielle Integration des Integrals in (5) bestätigen kann.

Um die Notation in den folgenden Rechnungen übersichtlich zu halten, führen wir die Folge $a_n := \sqrt{\ln n}$ für $n \in N$ ein. Sie besitzt die Eigenschaften

$$\frac{np^2(a_n)}{a_n^2} \to 0 \quad \text{und} \quad \frac{np(a_n)}{a_n} \to \infty \quad \text{für} \quad n \to \infty. \tag{6}$$

Die Wahl von a_n ist essentiell, um das behauptete asymptotische Verhalten zu beweisen. Bei Verwendung von (5) und der Abkürzung $\delta(r) := 1 - R(r)$ ergibt sich die Darstellung

$$\frac{1}{n+1} \int_{a_n}^\infty r p_n(r) dr = \int_{a_n}^\infty r p(r) (1 - \delta(r) \frac{p(r)}{r})^n \, dr. \tag{7}$$

Unter Beachtung der ersten Beziehung in (6) und von

$$(1 - \epsilon)^n = e^{-n\epsilon}(1 + O(n\epsilon^2)) \quad \text{für} \quad n\epsilon^2 \to 0$$

erhält man aus (7)

$$\frac{1}{n+1} \int_{a_n}^\infty r \rho_n(r) dr \;\simeq\; \int_{a_n}^\infty r \rho(r) e^{-n\delta(r)\rho(r)/r} \, dr,$$

wobei "$\simeq$" Gleichheit bis auf Größen der Ordnung $o(1)$ für $n \to \infty$ bedeutet. Mit der Substitution $p(r) = s$ und der Abkürzung

$$\Delta(s) := \delta(\sqrt{-2\ln(\sqrt{2\pi}s)}\,)$$

ergibt sich weiter

$$\frac{1}{n+1} \int_{a_n}^\infty r p_n(r) dr \;\simeq\; \int_0^{p(a_n)} e^{-ns\Delta(s)/\sqrt{-2\ln(\sqrt{2\pi}s)}} \, ds. \tag{8}$$

Im letzten Integral substituiert man $ns = t\sqrt{2\ln n}$ und gelangt zu der Beziehung

$$\frac{1}{n+1}\int_{a_n}^{\infty} rp_n(r)dr \;\simeq\; \frac{\sqrt{2\ln n}}{n}\int_0^{n\,p(a_n)/\sqrt{2\ln n}} e^{-td(t,n)/\sqrt{w(t,n)}}dt \tag{9}$$

mit

$$d(t,n) := \Delta\left(\frac{t\sqrt{2\ln n}}{n}\right) \quad \text{und} \quad w(t,n) := 1 - \frac{\ln(4\pi t) + \ln(\ln n)}{2\ln n}\,.$$

Es konvergiert punktweise $d(t,n) \to 1$ und $w(t,n) \to 1$ für $n \to \infty$ und die Funktion $g(t) := 1$ für $0 \le t < 1$ sowie $g(t) := e^{-t/2}$ für $1 \le t < \infty$ ist für genügend große n eine integrierbare Majorante des Integranden. Nach dem Satz von Fubini geht dann das Integral in (9) gegen

$$\int_0^{\infty} e^{-t}dt = 1. \tag{10}$$

Es wird der verbliebene Teil des Integrals in (1) abgeschätzt. Da $P(r)$ monoton wachsend ist, gilt

$$\begin{aligned} |\int_{-\infty}^{0} rp_n(r)dr| &\le (n+1)P^n(0)\int_0^{\infty} rp(r)dr \\ &= \frac{n+1}{\sqrt{2\pi}}(\frac{1}{2})^n \to 0 \quad \text{für} \quad n \to \infty. \end{aligned} \tag{11}$$

Weiter hat man unter Berücksichtigung von (5), der zweiten Beziehung in (6) und $p(a_n)/a_n \to 0$ für $n \to \infty$

$$\begin{aligned} \int_0^{a_n} rp_n(r)dr &\le (n+1)P^n(a_n)\int_0^{\infty} rp(r)dr \le \frac{n+1}{\sqrt{2\pi}}(1-\frac{p(a_n)}{2a_n})^n \\ &= \frac{n+1}{\sqrt{2\pi}}[(1-\frac{p(a_n)}{2a_n})^{2a_n/p(a_n)}]^{np(a_n)/2a_n} \to 0 \quad \text{für} \quad n \to \infty. \end{aligned}$$

Insgesamt ist damit das behauptete asymptotische Verhalten (3) bewiesen. □

Literatur

[1] J. Frank: *Three-Dimensional Electron Microscopy of Macromolecular Assemblies.* San Diego: Academic Press 1996

[2] N. Grigorieff: *Resolution measurement in structures derived from single particles.* In preparation.

[3] E. J. Gumbel: *Statistics of Extremes*. New York: Columbia Univ. Press 1958

[4] W. O. Saxton und W. Baumeister: *The correlation averaging of a regularly arranged bacterial cell envelope protein.* J. Microsc. **127**, 127-138 (1982)

[5] M. Unser, B. L. Trus und A. C. Steven: *A new resolution criterion based on spectral signal-to-noise ratios.* Ultramicroscopy **23**, 39-52 (1987)