Robert-Wichard-Pohl-Preis

Die Bayes'sche Variante

Zur Logik der Datenanalyse

Volker Dose

Physikerinnen und Physiker gewinnen ihre experimentellen Daten mit teurem Gerät und hohem Zeitaufwand. Doch beste Ausstattung und ausreichend Zeit müssen auch von entsprechend sorgfältiger Datenanalyse begleitet sein. Die von Bayes begründete und von Laplace ausgebaute Methode, von E. T. Jaynes als „The Logic of Science" bezeichnet, sollte dringend zum Standard werden und zwar sowohl in der Forschung als auch in der Lehre.

In einem Brief an Papst Urban VIII. zur Widerlegung des kopernikanischen Weltbildes heißt es: „Tiere, die sich bewegen, verfügen über Gliedmaßen und Muskeln. Die Erde besitzt weder Gliedmaßen noch Muskeln; also bewegt sie sich auch nicht". Diese Beweisführung erscheint uns heute natürlich grotesk. Zu ihrer Zeit war sie es jedoch nicht. Der inverse Schluss stellte ein ernsthaftes philosophisches Problem dar, dessen Lösung von dem englischen Geistlichen Thomas Bayes (1702–1761), stammt und in einer Arbeit mit dem Titel „An essay towards solving a problem in the doctrine of chances" 1763 posthum veröffentlicht wurde. Nach rund 250 Jahren sollte man meinen, dass die Bayes'schen Überlegungen Allgemeingut geworden sind. Das ist mitnichten so, weder im täglichen Leben noch in der Wissenschaft. Tendieren wir nicht alle dazu, im Bereich der Medizin z. B. das Testergebnis positiv/negativ im Umkehrschluss mit krank/gesund gleichzusetzen? Die Formulierung des Bayes'schen Theorems, wie es heutigen Tages in der Wissenschaft verwendet wird, verdanken wir keinem Geringeren als Pierre Simon de Laplace (1749–1827). Von ihm stammen auch erste Anwendungen auf Probleme der Himmelsmechanik, der medizinischen Statistik und der juristischen Beweiswürdigung. Leider stand einer weiteren Entwicklung der Theorie und ihrer Anwendungen ein häufig erheblicher Rechenaufwand entgegen. Angesichts der heute verfügbaren paradiesischen Rechenmöglichkeiten wird dieses Gebiet der Wissenschaft aber von einer zusehends wachsenden Gemeinde wieder entdeckt und bringt die während des 20. Jahrhunderts entwickelte Wahrscheinlichkeitstheorie, die auf der Identität Wahrscheinlichkeit = Häufigkeit gründet, zunehmend in Bedrängnis.

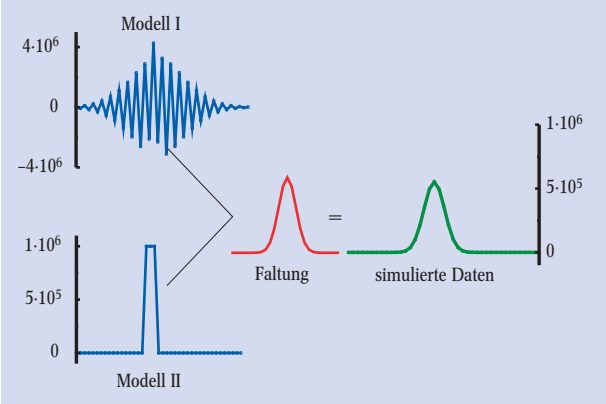

Abb. 1: Zwei völlig verschiedene Spektralfunktionen (blau) werden durch die Faltung mit einer gegebenen Apparatefunktion ununterscheidbar in den Datenraum abgebildet. Der inverse Schluss von den Daten auf die Spektralfunktion ist daher nicht ohne weitere Kenntnisse möglich.

Wo aber kommt die Physik ins Spiel? Der Weg von der Philosophie der Aufklärung zur Analyse von Daten aus physikalischen Experimenten oder astronomischen Beobachtungen scheint ja auf den ersten Blick nicht gerade sehr direkt zu verlaufen. Der Schein trügt! Traditionell wird in der Physik der Grad an Übereinstimmung zwischen experimentell erhobenen Daten und einer Modellrechnung zum Maßstab für die Güte der physikalischen Modellvorstellung angesehen. Je kleiner die Abweichung zwischen beiden, umso wahrscheinlicher das zugrunde gelegte Modell.

Ich gestehe gerne, dass auch ich lange Jahre von diesem Umkehrschluss überzeugt war, und möchte nun den Versuch unternehmen, auch Ihre (unterstellte) Überzeugung zu erschüttern. Natürlich muss die Modellrechnung die gegebenen Daten im Rahmen der experimentellen Fehler wiedergeben, damit das zugrunde gelegte Modell überhaupt als Kandidat für den physikalischen Sachverhalt in Frage kommt. Diese Bedingung ist notwendig. Sie ist aber nicht hinreichend, wie mit Hilfe der Abb. 1 gezeigt werden soll.

Wir gehen aus von der als Modell II bezeichneten trapezförmigen Spektralfunktion und nehmen an, dass ihre Messung mit einer Apparatur erfolgt, die durch die mit „Faltung" bezeichnete Auflösungsfunktion charakterisiert sei. Das Ergebnis dieser hypothetischen Messung wäre dann die als „simulierte Daten"

Kompakt

- Meist wird in der Physik der Grad an Übereinstimmung zwischen den Messdaten und einer Modellrechnung als Maßstab für die Güte eines physikalischen Modells angesehen.
- Dies ist jedoch nicht hinreichend, um entscheiden zu können, ob ein Modell wirklich für die Beschreibung eines physikalischen Sachverhalts in Frage kommt.
- Bei der Entscheidung dieser Frage hilft die Anwendung der von Bayes begründeten und von Laplace ausgebauten Methode.

Prof. Dr. Volker Dose, Max-Planck-Institut für Plasmaphysik, Boltzmannstr. 2, 85748 Garching – Preisträgerartikel anlässlich der Verleihung des Robert-Wichard-Pohl-Preises 2005 auf der 69. DPG-Jahrestagung in Berlin.

 1617-9439/05/0808-67

bezeichnete Kurve. Natürlich lässt sich dieser Weg auch rückwärts durchlaufen, d. h. aus den Daten kann man durch eine Entfaltungsoperation die Spektralfunktion zurückerhalten. Modifizieren wir jetzt für eine zweite Rechnung die Daten um eine Winzigkeit, nämlich um ein Zählereignis in dem Kanal, in dem das Signal maximal ist, – das entspricht einer relativen Änderung von 2×10^{-6} – und führen den Entfaltungsprozess mit den modifizierten Daten erneut durch, so erhalten wir das mit Modell I bezeichnete Resultat. Es oszilliert heftig, ist nicht positiv definit, was wir von einer Spektraldichte erwarten, und hat eine um den Faktor vier höhere Amplitude als die Funktion Modell II. Gleichwohl gibt es die Daten wieder mit einer Genauigkeit, die in einem Zählexperiment erst mit einer Ereigniszahl von $2{,}5 \times 10^{11}$ im Maximum erreicht würde. Die Übereinstimmung einer Modellrechnung mit gemessenen Daten ist daher sicher eine notwendige, aber keineswegs eine hinreichende Voraussetzung für die Gültigkeit der zugrunde gelegten physikalischen Vorstellung.

Das Bayes'sche Theorem

Wir verdanken nicht nur diese Einsicht dem Reverend Thomas Bayes, sondern auch die Formulierung des Auswegs aus dem Dilemma. Wir bezeichnen mit M das Modell, mit $\vec{\theta}$ die zum Modell gehörigen Parameter und mit $\vec{d}$ die Daten, die wir erklären möchten. Dann gilt, entgegen landläufiger Annahme, dass die Wahrscheinlichkeitsdichte p der Parameter $\vec{\theta}$, $p(\vec{\theta}\,|\,\vec{d}, M, I)$, gegeben ein Modell M und Daten $\vec{d}$ im Allgemeinen etwas ganz anderes ist als die Wahrscheinlichkeitsdichte der Daten, gegeben das Modell und seine Parameter. Wir benutzen hier die übliche Schreibweise für bedingte Wahrscheinlichkeiten. Danach ist die Größe vor dem senkrechten Strich die Variable und die Größen danach sind die Bedingungen für die Formulierung der Funktion. Es gilt also

$$p(\vec{\theta}|\vec{d}, M, I) \neq p(\vec{d}|\vec{\theta}, M, I)\ . \tag{1}$$

In I sind alle weiteren impliziten Informationen, über die wir verfügen, subsummiert, z. B. die Gründe, die uns zur Wahl des Modells M veranlasst haben. Aus (1) erhält man das Bayes'sche Theorem, indem die rechte Seite um einen Faktor ergänzt wird.

$$p\left(\vec{\theta}\,|\,\vec{d}, M, I\right) = \frac{p\left(\vec{\theta}\,|\,M, I\right)}{p\left(\vec{d}\,|\,M, I\right)} \cdot p\left(\vec{d}\,|\,\vec{\theta}, M, I\right). \tag{2}$$

Formal folgt es aus der Produktregel für bedingte Wahrscheinlichkeiten, mit deren Hilfe die Funktion $p(\vec{\theta}, \vec{d}|M, I)$ auf einfachere Elemente zurückgeführt werden kann durch Vergleich der beiden äquivalenten Entwicklungen.

$$\begin{aligned} p(\vec{\theta}|M, I) \cdot p(\vec{d}|\vec{\theta}, M, I) &= p(\vec{\theta}, \vec{d}|M, I) \\ &= p(\vec{d}|M, I) \cdot p(\vec{\theta}|\vec{d}, M, I) \end{aligned} \tag{3}$$

Natürlich wollen wir annehmen, dass die Funktion $p(\vec{\theta}|\vec{d}, M, I)$ auf 1 normiert sei. Die linke Seite in (2) bezeichnet man auch als Posterior-Wahrscheinlichkeitsdichte von $\vec{\theta}$. Entsprechend heißt $p(\vec{\theta}|M, I)$ Prior-Dichte. $p(\vec{d}|\vec{\theta}, M, I)$ wird im Folgenden als Funktion von $\vec{\theta}$ betrachtet und heißt dann „likelihood". $p(\vec{d}|M, I)$ bezeichnet man als Evidenz. Ihre Bedeutung wird sogleich klar werden. Aus (2) folgt durch Integration beider Seiten über $\vec{\theta}$

$$p(\vec{d}|M, I) = \int p(\vec{\theta}|M, I) \cdot p(\vec{d}|\vec{\theta}, M, I) d\vec{\theta}\ , \tag{4}$$

und unter Benutzung der Produktregel (3) die wichtige Marginalisierungsregel

$$p(\vec{d}|M, I) = \int p(\vec{d}, \vec{\theta}|M, I) d\vec{\theta}\ , \tag{5}$$

für die es in der auf dem Häufigkeitsbegriff aufbauenden traditionellen Wahrscheinlichkeitstheorie kein Äquivalent gibt. Aus (4) ersehen wir auch die Bedeutung der Größe $p(\vec{d}|M, I)$. Sie stellt die Wahrscheinlichkeit der Daten unter der Annahme des Modells M dar, wenn über alle möglichen Zahlenwerte der das Modell beschreibenden Parameter summiert (integriert) wird. $p(\vec{d}|M, I)$ ist damit eine Schlüsselgröße, aus der mit Hilfe des Bayes'schen Theorems die hoch interessante Frage nach der Wahrscheinlichkeit des Modells $p(M|\vec{d}, I)$ im Lichte der Daten beantwortet werden kann:

$$p\left(M|\,\vec{d}, I\right) = \frac{p(M|I)}{p\left(\vec{d}|\,I\right)} \cdot p\left(\vec{d}|\,M, I\right). \tag{6}$$

Nach (6) wird auch klar, warum der Informationshintergrund I mitgeführt werden muss, die Evidenz hängt nämlich davon ab. Es gehört zu den herausragenden Leistungen der Bayes'schen Theorie, dass sie für einen gegebenen Datensatz die Wahrscheinlichkeiten für eine Menge alternativer Modelle $\{M_i\}$ zu berechnen gestattet. Sowohl der inverse Schluss nach (2) als auch die Marginalisierung nach (4, 5) erfordern die Einführung von Prior-Wahrscheinlichkeiten $p(\vec{\theta}|M, I)$ bzw. $p(M\,|\,I)$, an denen sich immer wieder Kritik an der Bayes'schen Logik entzündet. Sie gipfelt in dem Einwurf, dass die Einführung dieser Prior-Wahrscheinlichkeiten das Ergebnis präjudiziere. $p(\vec{\theta}|M, I)$ und $p(M|I)$ sind Wahrscheinlichkeitsdichten, die alle Kenntnisse über $\vec{\theta}$ bzw. M beinhalten, bevor die neuen Daten $\vec{d}$ in Betracht gezogen werden. Der Faktor $p(\vec{d}|\vec{\theta}, M, I)$, der die Dateninformation enthält, modifiziert dann diesen Kenntnisstand je nachdem, wie informativ er ist. Enthält $p(\vec{d}|\vec{\theta}, M, I)$ wenig oder keine Information über die Parameter $\vec{\theta}$, so ist es nur logisch, dass $p(\vec{\theta}|\vec{d}, M, I) = p(\vec{\theta}|M, I)$ bleibt. Ist andererseits der Datenfaktor sehr informativ, so spielt die Struktur in $p(\vec{\theta}|M, I)$ eine kleine bis vernachlässigbare Rolle, und $p(\vec{\theta}|\vec{d}, M, I)$ hat dieselbe Form wie $p(\vec{d}|\vec{\theta}, M, I)$ als Funktion von $\vec{\theta}$ bei gegebenen Daten. Mit einer geeigneten multiplikativen Konstanten auf der rechten Seite von (1) wird aus der Ungleichung eine Gleichung. Unter diesen Voraussetzungen führt also der konzeptionell falsche traditionelle Schluss des Physikers zum richtigen Ergebnis. Die Tatsache, dass in einer Vielzahl von Fällen durch hinreichenden Fleiß bei den Messungen der Informationsgehalt der likelihood ad libitum gesteigert werden kann, mag auch mitverantwortlich sein für die zögerliche Akzeptanz, der die Bayes'sche Logik bei der Behandlung experimenteller Daten in der Physik begegnet. Dass auch in der Physik der Fleiß zuweilen kaum überwindbare Hürden vorfinden kann, lässt sich aus der Diskussion zu Abb. 1 erahnen.

Als Einführung in die Bayes'sche Theorie empfehle ich die Bücher von D. S. Sivia [1], M. Tribus [2] und E. T. Jaynes [3]. Die Schönheit der Bayes'schen Theorie, die mit (2, 3) und (4, 5) vollständig formuliert ist, liegt für mich in ihrer Einfachheit. Einen Eindruck von der Leistungsfähigkeit der Theorie sollen die im Folgenden beschriebenen drei Beispiele vermitteln.

Die Inversion der Faltung

In dem Klassiker „Physik der Sternatmosphären“ von 1955 schreibt A. Unsöld: „Nachdem wir uns klargemacht haben, wie man das Apparateprofil, d.h. das vom Spektrographen und Mikrophotometer erzeugte Profil einer monochromatischen Linie, ermitteln kann, wenden wir uns dem praktisch sehr wichtigen Problem der Entzerrung gemessener Linienprofile zu.“ Die Faltung, von der Unsöld hier spricht, ist wohl die am häufigsten in der Datenanalyse anzutreffende Integralgleichung. In vielen physikalischen Problemen ist die gesuchte Funktion in einer Integralgleichung der Form

$$D(E)=\int_{-\infty}^{\infty} K(E,E')S(E')dE', \qquad (7)$$

verpackt, die es zu invertieren gilt, um an die gesuchte Information $S(E')$ zu kommen. Der Kern $K(E,E')$ führt zur Faltung, wenn er von der Form $K(E,E')=K(E-E')$ ist. Das im Folgenden diskutierte Beispiel ist von allgemeinerer Form und beschreibt die Analyse eines Experiments zur spinpolarisierten inversen Photoemission an Ni(110) [4]. Spinpolarisierte inverse Photoemission ist eine probate Methode zur Untersuchung der elektronischen Bandstruktur von Ferromagneten. Bei Ferromagneten ist jedes Energieband aufgespalten in zwei spinabhängige Subbänder, die zu den zwei möglichen Orientierungen des Elektronenspins gehören. Zum Phänomen des Magnetismus tragen jedoch nur diejenigen Bänder bei, bei denen nicht beide Subbänder entweder voll besetzt oder vollständig unbesetzt sind. Da die Energieaufspaltung der Subbänder $\Delta E_{ex} \leq 2$ eV ist, sind das Bänder in der Umgebung der Fermi-Energie. Für die Theorie interessant ist die genaue Größe der Austauschaufspaltung ΔE_{ex} und ihre Abhängigkeit von der Temperatur. Zu ihrer Bestimmung sind im Allgemeinen zwei Messungen nötig, spinpolarisierte Photoemission für das besetzte Subband und spinpolarisierte inverse Photoemission für das unbesetzte Subband. Bei endlicher Temperatur werden jedoch zunehmend Löcher unterhalb der Fermi-Energie verfügbar, die dann natürlich auch für die inverse Photoemission zugänglich sind. Einen besonders günstigen Fall bietet Ni(110) auf der X-Z-W-Hochsymmetrielinie der Brillouin-Zone. Das besetzte (Majoritäts)-Band liegt hier Bandstrukturrechnungen zur Folge nur etwa 50 meV unterhalb der Fermi-Energie und ist bereits bei Zimmertemperatur ($kT \approx 25$ meV) merklich entvölkert. Das dazugehörige unbesetzte (Minoritäts)-Band liegt etwa 0,5 eV oberhalb der Fermi-Energie. Der Kern der Integralgleichung (7) für das Experiment ist daher ein Produkt aus der Fermi-Funktion $f(E_F,E',T)$ und der Apparatefunktion $A(E-E')$. Da die Genauigkeit, mit der die Spektralfunktion $S(E')$ aus den Daten $d(E)$ bestimmbar ist, natürlich davon abhängt, wie genau der Kern $K(E,E')$ bekannt ist, kommt der präzisen Bestimmung der Apparatefunktion $A(E-E')$ besondere Bedeutung zu. Für das vorliegende Experiment lässt sie sich glücklicherweise mit genügender Genauigkeit bestimmen. Vor Metalloberflächen existieren für ein sich näherndes Elektron nämlich unbesetzte elektronische Zustände im Bildkraftpotential. Sie sind der inversen Photoemission zugänglich und besitzen eine so lange Lebensdauer, dass ihre Linienbreite sehr gut vernachlässigt werden kann und das gemessene Signal daher in derselben Näherung direkt die Apparatefunktion abbildet. Abb. 2b) und d) zeigen experimentelle Daten zur spinpolarisierten inversen Photoemission für den Übergang $Z_4 \rightarrow Z_2$ in Ni.

Offene Kreise gehören zum Minoritätsband und volle Kreise zum Majoritätsband. Die beiden Teilbilder unterscheiden sich durch die Probentemperatur. Zum Teilbild (b) gehört eine Temperatur von 0,72 T_c und zum Teilbild (d) die Temperatur 0,82 T_c, wobei T_c die Curie-Temperatur der Volumenmagnetisierung von Ni ist. Der erwartete Intensitätsanstieg im Signal des Majoritätsbandes mit steigender Temperatur ist deutlich

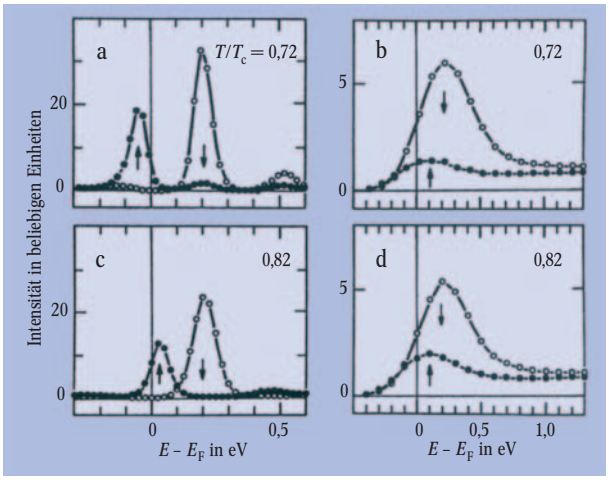

Abb. 2: **Spin-abhängige Quasiteilchen-Spektraldichten (a, c) und experimentelle Daten der inversen Photoemission (b, d) für den Übergang $Z_4 \rightarrow Z_2$ in Nickel bei zwei verschiedenen Temperaturen.**

zu sehen. Aus diesen Daten gilt es nun, die unter dem Integral versteckte spektrale Dichte S zu berechnen. Dazu wurde im vorliegenden Fall das Modell „punktweise Rekonstruktion auf einem äquidistanten Gitter“ gewählt mit den Parametern $\vec{\theta}=\{S_k\}=$„Funktionswerte an den Gitterpunkten“. Das Integral (7) wird damit zu einer Summe mit dem Wert $D_i=D(E_i)$. Nun ist der zugehörige Messpunkt d_i aber nicht exakt, sondern lediglich bis auf einen Fehler σ_i bekannt. Für ein Zählexperiment ist eine gute Schätzung für den Messfehler $\sigma_i=(d_i)^{1/2}$. Bei hinreichend hohen Ereigniszahlen lässt sich die einem Zählexperiment zugrunde liegende Poisson-Verteilung annähern durch die Normalverteilung

$$p\left(d_i|\vec{S},\sigma_i,M,I\right)=\frac{1}{\sigma_i\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma_i^2}(d_i-D_i)^2\right\}. \qquad (8)$$

Ein Element für die Inversion, nämlich die likelihood, ist damit spezifiziert. Es fehlt noch die Prior-Dichte $p(\vec{S}|M,I)$. Um Katastrophen, wie Modell 1 in Abb. 1 auszuschließen, die wir nämlich als unphysikalische Lösung des Problems zurückweisen würden, muss $p(\vec{S}|M,I)$ sicher die Eigenschaft haben, nur positive $\{S_i\}$ zu erlauben. Falls darüber hinaus keine detaillierten Vorkenntnisse über Form, Lage oder Amplitude der gesuchten Spektraldichten verfügbar sind, sollte $p(\vec{S}|M,I)$ von sich aus keine Struktur in das Inversionsergebnis einbringen. Beide Forderungen werden erfüllt von der Dichte

$$p\left(\vec{S}|M,m,\alpha,I\right)=\exp\left\{-\alpha\sum S_i\ln\frac{S_i}{m}\right\}/Z(\alpha), \qquad (9)$$

mit der Normierungskonstante $Z(\alpha)$. Diese Dichte wird maximal für $S_i=m$. Wählt man also für m eine hinreichend kleine Zahl, so wird die Rekonstruktion von S_i kleiner m ausfallen, wenn die Information in der likelihood nichts anderes verlangt. Gleichung (9) enthält darüber hinaus noch einen neuen (Hyper-) Parameter α. Er reguliert, mit welcher Stärke die Prior-Information (9) im Vergleich zur Dateninformation (8) in die Posterior-Dichte eingeht. Er wurde so gewählt, dass die Summe der normierten quadratischen Abwei-

chungen $\Sigma_i (d_i - D_i)^2 / \sigma_i^2 = N$ wird, wobei N die Zahl der Datenpunkte ist. Eine Diskussion über die rigorose Bayes'sche Behandlung von α würde den Rahmen dieses Artikels sprengen.

Abb. 2a) und c) zeigen das Resultat der Inversion. Die Energieskala ist hier gegebenüber dem rechten Teilbild um einen Faktor zwei gedehnt. Die Analyse ergibt zwei sehr schön getrennte Spektraldichten, die zu den jeweiligen Spinorientierungen gehören, und eine Auflösungsverbesserung um etwa einen Faktor fünf. Interessant ist auch, dass die Zahl der rekonstruierten Werte von $\hat{S}$ die Zahl der Datenpunkte übersteigen kann, hier um einen Faktor zwei. Der Grund liegt in der Konvexität von (9). Für die Physik des Magnetismus ist interessant, dass sich die Majoritätsspektraldichte mit steigender Temperatur auf die Minoritätsdichte zu bewegt. Aus den Messungen bei sechs verschiedenen Temperaturen folgte, dass sich die Austauschaufspaltung mit einer reskalierten Brillouin-Funktion mit der Curie-Temperatur des Nickel-Volumens beschreiben lässt. Die Datenanalyse hat in diesem Fall damit physikalische Schlussfolgerungen ermöglicht, die aus den Rohdaten allein kaum hätten gezogen werden können. Es ist vielleicht von Interesse, hier anzumerken, dass die traditionelle Inversion des Problems ähnlich wie in Abb. 1 eine mit der Amplitude 10^6 oszillierende und damit völlig unsinnige Lösung ergibt [5].

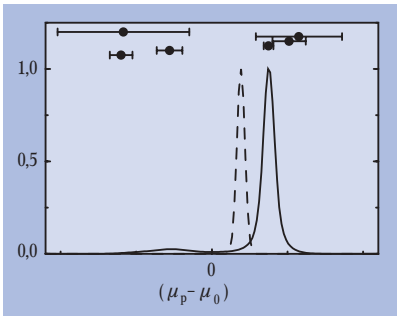

Abb. 3:
Posterior-Dichte für das magnetische Moment μ_p des Protons bezüglich eines willkürlich gewählten Ursprungs μ_0. Die gestrichelte Kurve folgt, wenn die Streuung der Daten als mit den angegebenen Standardabweichungen vereinbar angenommen wird. Die durchgezogene Kurve ist das Resultat des Mischungsmodells.

Im zweiten Beispiel sollen Situationen betrachtet werden, in denen ein Modell allein nicht ausreicht, um die vorhandenen Daten befriedigend zu erklären. Das Problem lässt sich einfach verstehen an Hand von Abb. 3. Die Daten mit Fehlerbalken, die hier aufgetragen sind, repräsentieren Messungen des magnetischen Moments des Protons und stammen aus der Codata-Auswertung 1986 [7] unserer physikalischen Konstanten und Konversionsfaktoren, wie sie jedes Jahr im Augustheft von Physics Today veröffentlicht werden. Das Modell für die Auswertung dieser Daten ist, dass sie einen bestimmten Wert μ_p des Moments approximieren. Die likelihood wäre dann wieder Gleichung (8), wenn D_i durch die Konstante μ_p ersetzt wird. Als Prior-Dichte wählen wir in diesem Fall eine Konstante $1/\Delta$, wobei Δ ein Bereich von μ_p-Werten ist, der den vermuteten wahren Wert sicher enthält. Die Posterior-Dichte für μ_p ist dann das Produkt über i von Termen gemäß Gleichung (8). Diese Dichte ist als Funktion von μ_p in Abb. 3 gestrichelt wiedergegeben. Ihr Maximum ist übrigens gegeben durch das gewichtete Mittel der Daten, errechnet mit den Formeln, die jeder Student im Anfängerpraktikum lernen muss. Das Resultat ist in diesem Falle mehr als kurios. Der Fehler des Resultats, numerisch berechnet aus der Varianz der gestrichelten Verteilung, ist kleiner als der kleinste Fehler einer Einzelmessung. Gleichwohl liegt das errechnete Mittel an einer Stelle, die von keinem der Fehlerbalken der Einzelmessungen überstrichen wird. Die sture Anwendung der Formeln aus dem Anfängerpraktikum, die man übrigens mit entsprechend unsinnigen Resultaten auch in manchen Publikationen finden kann, übergeht eine wichtige Voraussetzung. Diese lautet, dass die Streuung der Daten mit den Fehlerangaben verträglich sein muss. Tatsächlich ändert sich die gestrichelte Verteilung nicht, wenn wir bei konstant gehaltenen Fehlern die Daten bezüglich ihres Mittelwertes der Streckung um einen beliebigen Faktor unterwerfen. Das ist natürlich völlig absurd. Die Beschreibung der Daten – nichts anderes ist die likelihood – muss daher erweitert werden. In diesem Fall nehmen wir an, dass ein Datenpunkt mit einer Wahrscheinlichkeit β mit dem angegebenen Fehler korrekt beschrieben wird, während mit einer Wahrscheinlichkeit $1-\beta$ der angegebene Fehler um einen Faktor y_i modifiziert werden muss. Es gilt $y_i > 1$ für den Fall, dass die Streuung angesichts der spezifizierten Fehler zu groß ist, andernfalls $y_i < 1$. Der Faktor y ist mit i zu indizieren, er hängt von Lage und Fehler eines individuellen Datenpunktes ab. Die Wahrscheinlichkeit β trägt keinen Index, da sich ihre Bedeutung erst aus der Menge aller Datenpunkte mit zugehörigen Fehlern ergibt. Die (Hyper-)Parameter y_i und β gehören zwar zum Problem, sind aber für sich kaum von Interesse. Die Marginalisierungsregel der Bayes'schen Theorie (4, 5) erlaubt, sie aus der Rechnung zu entfernen, wenn man geeignete Verteilungen für y_i und β spezifiziert. Erfreulicherweise ist diese Spezifikation wenig kritisch. So wurde für β eine uniforme Verteilung im Bereich $0 \le \beta \le 1$ gewählt. Im Sinne des Entropieprinzips ist das auf diesem Intervall die Verteilung mit minimaler Information. Für y_i wurde die Verteilung $1/y_i^2$ im Intervall $1 \le y_i < \infty$ gewählt. Sie entspricht der Annahme, dass die spezifizierten Fehler nicht mit der in diesem Fall beobachteten Datenstreuung verträglich sind und mit einem Faktor $y_i > 1$ mit abnehmendem Gewicht zu multiplizieren sind. Die Posterior-Verteilung für μ_p ist jetzt komplizierter, sie entsteht nämlich als Produkt aus Faktoren, die ihrerseits die Summe von zwei Termen sind. Als Funktion von μ_p ist die Posterior-Dichte in Abb. 3 durchgezogen dargestellt. Ihr Maximum liegt nun im Einklang mit dem gesunden Menschenverstand im Bereich des mit Abstand genauesten Datenpunkts. Der Fuß der Verteilung ist breiter geworden und es erscheint ein sekundäres Maximum. Während man die gestrichelte Kurve angemessen durch Mittelwert und Varianz charakterisieren kann und damit auf ihre explizite Darstellung verzichten könnte, gilt dies ganz offensichtlich nicht für die durchgezogene Kurve. Wenn wir nicht damit leben wollen, dass die Messungen zum magnetischen Moment des Protons auch unter dem erweiterten Modell unverträglich bleiben, führt an neuen Präzisionsmessungen kein Weg vorbei, denn eine weitere Komplizierung des Modells, etwa durch Betrachtung von Mischungen mit mehr als zwei Komponenten, lässt die geringe Zahl von Datenpunkten kaum zu.

Ausreißer, Signal und Untergrund

Das gerade vorgestellte Mischungsmodell hat Anwendung auf „Ausreißer"-Situationen im weitesten Sinne. So ist es zunächst naheliegend an den Fall zu denken, dass die Konstante μ_p des obigen Beispiels durch eine Funktion von einer zum Datenpunkt d_i gehörigen unabhängigen Variablen E_i und einen Satz von Parametern $\vec{\theta}$ ersetzt wird [8]. Für Messungen, die aus nicht erkennbaren Gründen an häufigen „Ausreißern" leiden, kann dann ein Mischungsmodell sehr hilfreich für die Schätzung der Parameter sein. Wir denken dabei nicht an die Situation, dass ein Messwert so weit abseits des Haupttrends liegt, dass er offensichtlich zu einer Fehlmessung gehört. Gemeint sind vielmehr

diejenigen Fälle, bei denen der Übergang vom Haupttrend zum manifesten „Ausreißer“ mehr oder weniger fließend wird. Es soll auch daran erinnert werden, dass der Maßstab für die Abweichung vom Haupttrend der Fehler des betrachteten Messpunktes ist. Gegen stark abweichende Messpunkte mit entsprechend sehr großem (unabhängig bestimmten!) Messfehler ist nichts einzuwenden. Sie können sehr wohl zu den „guten“ Messungen gehören. Komplementäres gilt für wenig abweichende Punkte mit sehr geringem Messfehler. Eine Anwendung des Mischungsmodells auf diese Situation findet sich in [6]. Die Bedeutung der Messfehler für die Analyse experimenteller Daten deutet sich hier bereits an und wird weiter unten noch stärker untermauert.

Eine weitere bedeutende Anwendung des Mischungsmodells ergibt sich bei dem häufig auftretenden Problem der Trennung von Signal und Untergrund. Ein Beispiel dafür ist die Elementanalyse mit Hilfe der durch Teilchenbeschuss erzeugten charakteristischen Röntgenstrahlung. Die charakteristischen Linien sitzen immer auf einem Bremsstrahlungsuntergrund. Dieser ist zwar verhältnismäßig schwach, wenn man von der Anregung durch Elektronen zur Anregung mit Protonen oder Heliumionen übergeht. Trotzdem begrenzt er die Nachweisempfindlichkeit für Spurenelemente und es kann der Wunsch entstehen, der Nachweisgrenze durch Trennung von Signal vom Untergrund näherzukommen. Dabei muss man voraussetzen, dass wir alle Strukturen, die langsam mit der Energie variieren, als Untergrund betrachten. Diese langsam variierenden Signalanteile werden dann durch geeignete glatte Funktionen modelliert. Als sehr gut geeignet und hinreichend flexibel haben sich kubische Splines erwiesen, deren Stützstellen man auf einem im Vergleich zu den Datenpunkten groben Gitter adaptiv verteilt. Adaptiv bedeutet dabei, dass man den lokalen Stützstellenabstand von der Variation der anzupassenden Datenregion abhängig macht. Weite Intervalle also bei schwacher Variation und engere an steileren Stellen. Die beiden Modelle, die dann in die Mischung eingehen, sind die Hypothesen „der Datenpunkt enthält nur Untergrund“ und „der Datenpunkt setzt sich zusammen aus Untergrund und Signal“. Der Signalanteil wird dann wieder gemäß (4) marginalisiert.

Natürlich ist die Trennung von Signal und Untergrund im strengen Sinne gar nicht möglich. Jeder Datenpunkt enthält einen Anteil Untergrund und einen Anteil Signal, wie klein letzterer auch immer sein mag. Das Problem muss daher streng genommen umformuliert werden. Man sollte eigentlich die Wahrscheinlichkeiten dafür, dass ein Datenpunkt nur Untergrund enthält, vergleichen mit der Wahrscheinlichkeit, dass er sich aus Untergrund und Signal zusammensetzt. So wurden die ROSAT-Daten im linken Teil der Abb. 4 bearbeitet. Der Untergrund wurde hier modelliert durch 25 Elemente der exakten Lösung des Spline-Problems in zwei Dimensionen auf unendlichem Definitionsbereich, angesichts von 1500 Bildpunkten eine kleine Zahl. Das rechte Teilbild zeigt lediglich diejenigen Pixel, für die die Wahrscheinlichkeit, dass sie Signal und Untergrund enthalten, größer ist als die Wahrscheinlichkeit, dass sie nur Untergrund enthalten. Diese Bildanteile werden dann als Röntgenquellen angesehen. In den ausgedehnten schwarzen Bereichen überwiegt dagegen die Wahrscheinlichkeit für „nur Untergrund“.

Als drittes Beispiel soll die Analyse von Massenspektren dienen, die mit dem weit verbreiteten Quadrupol-Spektrometer gemessen wurden. Quadrupol-Massenspektrometer werden ja nicht nur in der Forschung, sondern auch in Vakuumtechnik und Prozesskontrolle im weitesten Sinne als Monitore eingesetzt. Zu einem Forschungsinstrument per se werden sie eigentlich erst im Zusammenhang mit Bayes’scher Datenanalyse. Das Grundproblem der Interpretation von Spektren aus Massenspektrometern mit Elektronenstoß-Ionisation ist die Fragmentierung. Sie führt dazu, dass die Spektren verschiedener Spezies sich unter Umständen stark überlappen und daher aus den beobachteten Spektren nicht ohne weiteres die erzeugenden Spezies folgen. Einen besonders ungünstigen Fall stellen in dieser Hinsicht die Massenspektren von Kohlenwasserstoffen dar. An ihnen besteht andererseits besonderes Interesse bei der plasmagestützten Deposition von Diamant- und amorphen wasserstoffhaltigen Kohlenstoffschichten. Bei Quadrupol-Spektrometern kommt darüber hinaus noch erschwerend hinzu, dass die beobachteten Fragmentierungsmuster eines Moleküls empfindlich von den dem Benutzer zugänglichen Einstellungen abhängen. Der Rückgriff auf in der Literatur verfügbare Fragmentierungsmuster kann daher allenfalls als grobe Leitlinie angesehen werden und eignet sich keinesfalls für eine quantitative Analyse der Spektren nach der Fußgängermethode. Diese setzt daher immer eine vollständige Zahl von Eichmessungen an reinen Gasen voraus und geht dennoch in sehr vielen Fällen fehl. Das nicht nur, weil die Fehler bei dieser sukzessiven Subtraktion kumulieren, sondern häufig ist auch die Voraussetzung, dass mindestens ein Massenkanal in jeder Stufe des Verfahrens nur von einer Spezies herrührt, nicht gegeben. Weiter ist keineswegs gewährleistet, dass man keine negativen Werte für die Konzentration gewisser Spezies erhält. Schließlich ist die stillschweigende Annahme, dass die Eichmessungen fehlerfrei seien, natürlich ungerechtfertigt, denn sie wurden ja mit demselben Instrument gemessen, das die zu analysierenden Spektren produzierte.

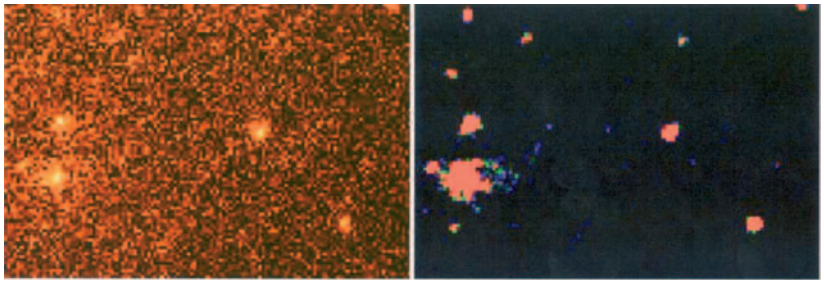

Abb. 4:
Das linke Teilbild zeigt die im ROSAT All Sky X-Ray Survey gemessene Röntgenstrahlungsintensität für einen kleinen Beobachtungsausschnitt. Das rechte Teilbild zeigt diejenigen Bereiche, die als Ergebnis einer Bayes‘schen Analyse besser durch ein Modell „Quelle + Untergrund“ als durch „Untergrund“ allein beschrieben werden. (Quelle: ESA)

Einen Ausweg aus all diesen Schwierigkeiten bietet die Bayes’sche Vorgehensweise. Sie erlaubt die zu analysierenden Spektren auf gleicher Stufe mit den Eichmessungen zu behandeln [9] und außerdem die in den existierenden Tabellenwerten enthaltene Information über die Fragmentierungsmuster mit dem experimentellen Material zu kombinieren. In diesem Fall ist man auch weder auf einen vollständigen Satz von Eichmessungen angewiesen, noch muss vorausgesetzt werden, welche Spezies mit Sicherheit in der Gasprobe vorhanden sind. Vielmehr ermöglicht Gleichung (6) ja, verschiedene Zusammensetzungen (= Modelle) zur Analyse des vorliegenden Datensatzes heranzuziehen und ihre jeweilige Wahrscheinlichkeit zu berechnen.

Daraus kann man quantitativ schließen, wie wichtig die Vernachlässigung oder Hinzunahme einer speziellen Spezies ist. Gleichung (6) begrenzt darüber hinaus automatisch die Modellkomplexität. Während die Summe der quadratischen Abweichungen zwischen Daten und Modellrechnung natürlich mit zunehmender Modellkomplexität monoton abnimmt, durchläuft die Modellwahrscheinlichkeit nach (6) ein Maximum. Die Größe,

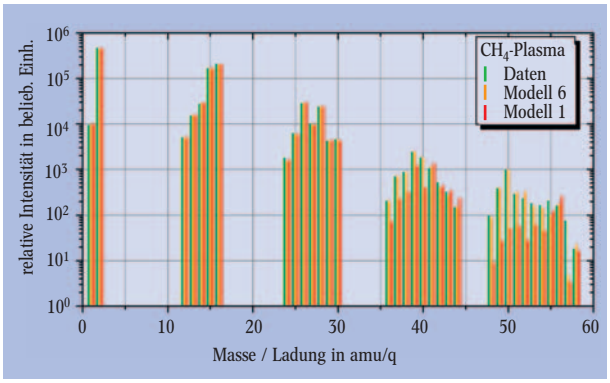

Abb. 5: **Das Massenspektrum von Neutralteilchen aus einer mit Methan betriebenen, induktiv angeregten Plasmaquelle (grün) im Vergleich mit Rechnungen nach dem besten (Modell 6) und dem schlechtesten (Modell 1) aus insgesamt sieben untersuchten Modellen.**

die den notwendigen gegenläufigen Effekt enthält, ist $p(\vec{d}\,|\,I)$ in (6). Sie führt dazu, dass (6) Ockhams Prinzip genügt, nach dem eine Sache auf möglichst einfache Weise die befriedigendste Beschreibung erfährt. Jenseits des Maximums von $p(M\,|\,\vec{d},I)$ wird nur noch Rauschen angepasst [6, 9].

Ein schon relativ schwieriges Beispiel für die Bayes'sche Zerlegung von Spektren eines Quadrupol-Instruments zeigt Abb. 5. Die zugrunde liegenden Messungen stammen vom Neutralgas aus einem induktiv angeregten Niedertemperaturplasma mit Methan als Arbeitsgas. Die Eingangsdaten setzten sich zusammen als Signal und Fehler von 34 Massenkanälen, die für 27 verschiedene Plasmaarbeitspunkte gemessen wurden. Dazu kamen 11 Eichmessungen für die Gase H_2, CH_4, C_2H_2, C_2H_4, C_2H_6, C_3H_6, C_3H_8, 1–C_4H_8, I–C_4H_8, I–C_4H_{10} und n–C_4H_{10}. Die untersuchten Modelle erhielten darüber hinaus noch die Moleküle C_3H_4, C_4H_2 und C_4H_6. Für alle Moleküle existieren Fragmentierungsmuster in Tabellen, die als Prior-Information mitbenutzt werden. Resultat der Analyse war zunächst die Modellwahrscheinlichkeit. Sie war am höchsten bei Vernachlässigung von I–C_4H_8 und I–C_4H_{10} und am niedrigsten bei Vernachlässigung von C_3H_4, C_4H_2 und C_4H_6. Weitere Resultate der Analyse sind die Fragmentierungsmuster der involvierten Spezies und deren Fehler für den gesamten Satz von Messungen sowie die Gaszusammensetzung für jedes einzelne Spektrum. Der Vergleich von berechneten Spektren nach bestem und schlechtestem Modell (aus insgesamt sieben) mit Messdaten eines herausgegriffenen Spektrums in Abb. 5 zeigt eine sehr gute Übereinstimmung für das beste Modell. Lediglich in Kanal 57 ergibt sich eine größere Abweichung. Sie ist sehr wahrscheinlich darauf zurückzuführen, dass die Messung C_5H_x und C_6H_y Kohlenwasserstoffe nicht mehr erfasst hat. Für Hexan hat nämlich das intensivste Fragment-Ion die Massenzahl 57.

Zwei Bemerkungen sollten an dieser Stelle gemacht werden. Meiner Erfahrung nach ist es sehr schwierig, just von den Leuten, die ihrerseits junge Studenten im Anfängerpraktikum mit der Fehlerrechnung getriezt haben, Fehlerangaben zu ihren Daten zu bekommen. Man wird gerne mit „x %“ abgespeist. Für die Abb. 5 zugrunde liegende likelihood von der Form (8) würde ein konstanter relativer Fehler aber bedeuten, dass jeder Massenkanal einen identischen Beitrag zur likelihood liefert. Das wiederum hieße, die offensichtlich vorhandene Intensitätsstruktur des Spektrums, die sich über fünf Größenordnungen in der Signalstärke erstreckt, als Information wegzubügeln. Das ist ein gutes Beispiel dafür, dass bei einer Bayes'schen Analyse die Fehler genauso wichtig sind wie die Daten, und möge auch als Warnung davor verstanden werden, Fehlerangaben auf die leichte Schulter zu nehmen.

Die zweite Bemerkung betrifft den Aufwand, der in Abb. 5 steckt. Die Zahl der Variablen des Problems ist 27 × Anzahl der Spezies plus Anzahl der Cracking-Koeffizienten aller Spezies. Im vorliegenden Fall ergibt das mehr als 400. Entsprechende Dimension haben die Integrale (4). Die Rechnungen sind daher durchaus nicht trivial und zeitaufwändig.

Diese drei Beispiele für die Anwendung der Bayes'schen Theorie müssen hier genügen. Sollte es mir gelungen sein, Ihr Interesse zu wecken, so finden Sie weitere in [6] und [10]. Das Schlusswort möchte ich dem prominenten Wahrscheinlichkeitstheoretiker George E. P. Box überlassen. Er soll gesagt haben [3]: „I believe, for instance, that it would be very difficult to persuade an intelligent physicist that current statistical practice was sensible, but that there would be much less difficulty with an approach via likelihood and Bayes' theorem.“ Diese Vermutung beschreibt sehr gut die Erfahrung, die ich selbst in den vergangenen Jahren gemacht habe.

Foto: J. Röhl

Der Autor

Volker Dose (rechts, bei der Preisverleihung mit DPG-Präsident Knut Urban) promovierte und habilitierte an der Universität Zürich. Anschließend war er Professor in Würzburg, bevor er 1985 als Direktor an das MPI für Plasmaphysik nach Garching ging. Seit 1991 lehrt er außerdem an der Universität Bayreuth. Volker Dose prägte das Gebiet der inversen Photoemission und beschäftigt sich seit den 90er-Jahren mit der Bayes'schen Statistik, die er auf ein breites Spektrum unterschiedlicher Probleme angewandt hat.

Literatur

[1] *D. S. Sivia*, Data Analysis: A Bayesian Tutorial, Clarendon Press, Oxford (1996)
[2] *M. Tribus*, Rational descriptions, decisions, and design, Pergamon (1969); Neuauflage bei Expira AB (2000)
[3] *E. T. Jaynes*, Probability Theory: The Logic of Science, Cambridge University Press, Cambridge (2003)
[4] *W. von der Linden, M. Donath und V. Dose*, Phys. Rev. Lett. **71**, 899 (1993)
[5] *W. von der Linden*, Appl. Phys. **A 60**, 155 (1995)
[6] *V. Dose*, Rep. Prog. Phys. **66**, 1421 (2003)
[7] *E. Cohen und B. Taylor*, Rev. Mod. Phys. **59**, 112 (1987)
[8] *V. Dose* et al., Nuclear Fusion **41**, 1671 (2001)
[9] *T. Schwarz-Selinger* et al., J. Mass Spectrom. **36**, 866 (2001); *H. D. Kang* et al., J. Mass Spectrom. **37**, 748 (2002); *H. D. Kang und V. Dose*, J. Vac. Sci. Technol. **A 21**, 1978 (2003)
[10] www.ipp.mpg.de/OP/Datenanalyse/index.php