AUSTRIAN JOURNAL OF STATISTICS
Volume 31 (2002), Number 1, 89-103

Neuere Entwicklungen in der Konzentrationsmessung

Helmut Strasser
Institut für Statistik, Wirtschaftsuniversität Wien

Gerhart Bruckmann zum 70. Geburtstag gewidmet

Abstract: A general approach for the quantization of statistical data and distributions is considered. The concept is closely related to the statistical measurement of concentration and to the mathematical theory of majorization. In particular, it is the theoretical basis of those compression algorithms which are analysed by Pötzelberger and Strasser (2001).

1 Einleitung

Bei der statistischen Analyse von hochdimensionalen Datensätzen besteht häufig die grundsätzliche Schwierigkeit, daß der Stichprobenumfang, dh. die Anzahl der Datensätze, zu gering ist, um eine Beschreibung durch ein parametrisches Modell tatsächlich statistisch auswerten zu können. Ein Beispiel soll diese Feststellung verdeutlichen.

Es sind im Bereich der Marktforschung Datensätze mit 15 bis 30 Variablen keine Seltenheit, wobei aber der Stichprobenumfang n dieser Datensätze in der Regel die Marke von einigen Hundert nicht übersteigt, sondern eher deutlich darunter bleibt. Wir nehmen der Einfachheit halber an, daß die Variablen dichotom sind, dh. nur die Werte 0 und 1 besitzen. Solche Datensätze entstehen aus Befragungen mit Alternativfragen. Obwohl die Datenvektoren nur endlich viele Werte annehmen können, ist an eine volle Modellierung nicht zu denken, da die Anzahl der möglichen Werte ($2^{15} = 32\,768$, $2^{30} = 1,1\,E\,10$) die Größe der realistischen Stichprobenumfänge bei weitem übersteigt. Es ist üblich, für solche Datenätze wesentlich kleinere Modelle zu verwenden, die nur die Wechselwirkung zwischen Paaren von Variablen beschreiben. Es sind dies beispielsweise autologistische Modelle, deren bedingte Verteilungen linearen Logitmodellen entsprechen. Die Anzahl der Modellparameter bei solchen Modellen beträgt $d(d+1)/2$, wenn d die Anzahl der Variablen bezeichnet. Hat die Befragung, aus der die Daten stammen, einen Vergleich von k Marken zum Ziel, dann handelt es sich statistisch um ein k-Stichprobenproblem und man hat es mit $kd(d+1)/2$ Parametern zu tun, denen n Datensätze gegenüberstehen. Ist beispielsweise $d = 20$ und $k = 8$, so haben wir es mit 1 680 Parametern zu tun, für deren Analyse einige Hundert Datensätze kaum ausreichen.

Ein Ausweg aus der beschriebenen Schwierigkeit besteht darin, das Skalenniveau der Datensätze so stark zu senken, daß eine Modellierung mit einer wesentlich geringeren Anzahl von Parametern möglich wird. Die Senkung des Skalenniveaus erfolgt dadurch, daß die Datenvektoren $x \in \mathbb{R}^d$ mit einer Abbildung $f : \mathbb{R}^d \to M$ kodiert werden. Bei univariaten quantitativen Datensätzen ist die Kodierung durch Rangzahlen das berühmteste und erfolgreichste Beispiel für eine Senkung des Skalenniveaus. Im Bereich der multivariaten Statistik ist die Senkung des Skalenniveaus ein ungleich schwierigeres Problem. Es sind vor allem zwei Möglichkeiten verbreitet.

Eine Möglichkeit besteht darin, als Bildbereich der Kodierung eine niedrigdimensionalen Raum, z.B. $M = \mathbb{R}$, $M = \mathbb{R}^2$ oder $M = \mathbb{R}^3$ zu wählen. Verwendet man eine lineare

Abbildung $f : \mathbb{R}^d \to M$, die einen möglichst großen Teil der ursprünglichen Datenstreuung in die kodierten Daten hinüberrettet, dann betreibt man Hauptkomponentenanalyse.

Eine andere Möglichkeit für die Senkung des Skalenniveaus besteht darin, als Bildmenge M eine endliche Teilmenge von $\mathbb{R}^d$ zu wählen. Diese zweite Möglichkeit wird als Quantisierung bezeichnet. Der Gegenstand der vorliegenden Arbeit ist eine Bericht über neuere Ansätze zur Quantisierung von multivariaten Datensätzen.

Wir werden im folgenden einen allgemeinen theoretischen Ansatz zur Behandlung des Quantisierungsproblems vorstellen. Es wird sich herausstellen, daß dieser Ansatz sowohl mit dem klassischen statistischen Konzept der Konzentrationsmessung als auch mit dem mathematischen Konzept der Majorisierung von Verteilungen zusammenhängt. Die Zielvorstellung einer optimalen Quantisierung wird sich durch diesen theoretischen Ansatz exakt fassen lassen, und es ist dadurch möglich, praktisch realisierbare Algorithmen zur Berechnung solcher Quantisierungen anzugeben. Die Klasse dieser Algorithmen umfaßt sowohl vertraute Methoden aus der statistischen Clusteranalyse als auch moderne Verfahren aus dem Bereich der künstliche neuronalen Netze.

Der von uns verwendete Quantisierungsansatz stammt für eindimensionale Daten von Bock (1992). Der Ansatz von Bock wurde durch Pötzelberger und Strasser auf den multivariaten Fall erweitert. Die theoretische Untersuchung der Algorithmen findet sich bei Pötzelberger und Strasser (2001). Experimentelle Untersuchungen zu den Algorithmen bietet Steiner (2000), Pötzelberger (2000a,b) untersucht die mathematische Grundlagen des Quantisierungskonzepts. Eine Übersicht über die erzielten Resultate findet sich in Strasser (2000a,c). Weitere entscheidungstheoretische Gesichtspunkte werden in Strasser (2000b) diskutiert.

Das alte ökonomische Konzept der Konzentrationsmessung wird in der modernen Statistik unter dem Namen Majorisierung behandelt. Wir beginnen bei unserer Darstellung mit jenem Spezialfall der Majorisierung, den wir als Quantisierung bezeichnen. Anschließend erklären wir das allgemeine Konzept der Majorisierung von Verteilungen. Wir zeigen schließlich, daß dieses Konzept die alte statistische Idee der Konzentrationsmessung umfaßt. Nach diesen Vorbereitungen wenden wir uns dem Thema der optimalen Quantisierung von Verteilungen zu.

Zum Abschluß dieses einleitenden Abschnitts erklären wir noch einige technische Begriffe. Unter empirischen Verteilungen verstehen wir hier zunächst Häufigkeitsverteilungen von Daten in $\mathbb{R}^d$, aber auch Inhaltsverteilungen auf endlichen Grundgesamtheiten wie zum Beipiel Vermögensverteilungen. Sämtliche dieser Konzepte spielen aber ebenfalls für Wahrscheinlichkeitsverteilungen, also für Modelle von empirischen Verteilungen eine Rolle. Um Modelle und empirische Verteilungen gemeinsam behandeln zu können, verwenden wir W-Maße auf der Borel-σ-Algebra von $\mathbb{R}^d$ als allgemeinen Oberbegriff.

Im folgenden werden immer wieder W-Maße und konvexe Funktionen auftreten. Zur Vereinfachung der Sprechweise vereinbaren wir, daß alle auftretenden W-Maße endliche erste Momente haben, und daß alle auftretenden konvexen Funktion durch eine affinlineare Funktion von unten dominiert sind. Diese beiden Bedingungen garantieren, daß konvexe Funktionen stets quasi-integrierbar sind.

2 Quantisierung von Verteilungen

Es seien P und Q zwei W-Maße auf $\mathcal{B}$ und es besitze Q eine endliche Trägermenge, also

$$Q = \sum_{j=1}^{m} \alpha_j \varepsilon_{y_j}.$$

Wir sagen, daß das W-Maß Q durch das W-Maß P majorisiert wird, wenn Q aus P durch eine bestimmten Vereinfachungsvorgang gewonnen wird. Diesen Vorgang nennen wir Quantisierung des W-Maßes P.

Der Vorgang der Quantisierung erfolgt in zwei Schritten und wird nun beschrieben.

1. Im ersten Schritt wird das W-Maß P in eine Mischung von endlich vielen anderen W-Maßen zerlegt. Damit ist folgendes gemeint: Wir suchen W-Maße $P_1, \ldots, P_m$, sodaß sich P als konvexe Linearkombination

$$P(A) = \sum_{j=1}^{m} \alpha_j P_j(A), \quad A \in \mathcal{B}, \tag{1}$$

dieser W-Maße darstellen läßt. Die Gewichte $\alpha_1, \ldots, \alpha_m$ der konvexen Linearkombination werden dann als Gewichte für das Maß Q verwendet.

2. Im zweiten Schritt bilden wir die ersten Momente (Baryzentren) der Maße $P_1, \ldots, P_m$, also

$$y_j = \int x \, P_j(dx), \quad j = 1, 2, \ldots, m, \tag{2}$$

und wählen diese Baryzentren $y_1, \ldots, y_m$ als Trägerpunkte des Maßes Q.

In wenigen Worten läßt sich der Übergang von P nach Q so zusammenfassen: Das W-Maß Q ensteht aus P durch Desintegration (Entmischung) von P in endliche viele W-Maße und anschließende Konzentration der Gesamtmasse von P auf die Baryzentren der Mischungskomponenten.

Wir gelangen so zum Begriff der Quantisierung.

Definition 2.1: *Es seien P und Q zwei W-Maße. Das W-Maß $Q = \sum_{j=1}^{m} \alpha_j \varepsilon_{y_j}$ mit endlicher Trägermenge wird von P majorisiert bzw. ist eine Quantisierung von P, in Zeichen $P \succ Q$, wenn es W-Maße $P_1, \ldots, P_m$ gibt, sodaß die Gleichungen (1) und (2) gelten.*

Ein W-Maß P läßt im allgemeinen zahlreiche unterschiedliche Quantisierungen zu. Ein besonderes wichtiger Spezialfall einer Quantisierung liegt vor, wenn das majorisierte W-Maß Q durch eine Zerlegung der Trägermenge von P erzeugt wird. Quantisierungen, die durch Zerlegungen erzeugt werden, spielen eine große Rolle in der statistischen Clusteranalyse und Klassifikation.

Ist $\mathcal{C} = (C_1, C_2, \ldots, C_m)$ eine Zerlegung der Grundmenge $\mathbb{R}^d$, dann ergibt sich daraus in natürlicher Weise eine Desintegration der Form (1) von P, indem man festsetzt

$$\alpha_1 := P(C_1), \ldots, \alpha_m := P(C_m), \tag{3}$$

und

$$P_1 := P(\cdot|C_1), \ldots, P_m := P(\cdot|C_m).$$

Das Besondere an dieser Desintegration von P besteht darin, daß die Komponenten P_1, $\ldots, P_m$ W-Maße mit disjunkten Trägermengen sind, was bei einer allgemeinen Desintegration (1) nicht verlangt ist.

Eine Zerlegung $\mathcal{C}$ führt also, wie wir gesehen haben, zu einer Desintegration der Form (1) von P, was dem ersten Schritt der Bildung einer Quantisierung Q entspricht. Um zu Q zu gelangen, müssen wir daher nur mehr den zweiten Schritt, also die Mittelwertbildung durchführen. Die Trägerpunkte von Q sind dann einfach die Zentroide der Mengen der Zerlegung $\mathcal{C}$, also

$$y_1 := \frac{1}{P(C_1)} \int_{C_1} x\, P(dx), \ldots, y_m := \frac{1}{P(C_m)} \int_{C_m} x\, P(dx). \tag{4}$$

3 Majorisierung von Verteilungen

Wir wenden uns nun dem allgemeinen Fall zu. Er besteht darin, daß das majorisierte W-Maß keine endliche Trägermenge haben muß. Um diese allgemeine Definition der Relation $P \succ Q$ vorzubereiten, wollen wir zunächst den speziellen Vorgang der Quantisierung ein wenig umformulieren.

Gehen wir nochmals von der in den Gleichungen (1) und (2) beschriebenen Situation aus. Wir definieren einen Markoffkern $D : (y, A) \mapsto D(y, A)$ durch

$$D(y, A) = P_j(A) \quad \text{wenn } y = y_j,\ j = 1, 2, \ldots, m,$$

oder in anderen Worten

$$D(y, A) := \sum_{j=1}^{m} P_j(A) 1_{\{y_j\}}(y),\ A \in \mathcal{B}, y \in \mathbb{R}^d.$$

Damit ist D für Q-fast alle $y \in \mathbb{R}^d$ definiert. Mit diesem Markoffkern können wir die Gleichungen (1) und (2) neu anschreiben. Es gilt

$$P(A) = \sum_{j=1}^{m} \alpha_j P_j(A) \quad \Leftrightarrow \quad P(A) = \int D(y, A) Q(dy)$$

und

$$y_j = \int x\, P_j(dx),\ j = 1, 2, \ldots, m, \quad \Leftrightarrow \quad y = \int x D(y, dx)\ Q\text{-f.ü.}$$

Ein Markoffkern D mit den beschriebenen Eigenschaften wird (mit einer aus der Potentialtheorie kommenden Bezeichnung) als Dilation von Q nach P bezeichnet.

Die folgende Definition überträgt den beschriebenen Gedanken auf beliebige W-Maße.

Definition 3.1: *Es seien P und Q W-Maße auf $\mathbb{R}^d$. Das W-Maß P ist eine Majorisierung des W-Maßes Q (in Zeichen $P \succ Q$), wenn es eine Dilation von Q nach P gibt, dh. wenn es einen Markoffkern $D : \mathcal{B} \times \mathbb{R}^d \to \mathbb{R}$ mit den Eigenschaften*

$$P(A) = \int D(A, y) Q(dy), \quad A \in \mathcal{B}, \tag{5}$$

und

$$y = \int xD(dx, y) \quad Q - \text{f.ü.} \tag{6}$$

gibt.

Man kann leicht nachprüfen, daß für W-Maße Q mit endlicher Trägermenge die Definitionen (2.1) und (3.1) äquivalent sind.

Genauso wie endliche Zerlegungen $\mathcal{C}$ in natürlicher Weise zu majorisierten W-Maßen mit endlicher Trägermenge führen, lassen sich auch durch Unter-σ-Algebren von $\mathcal{B}$ majorisierte W-Maße konstruieren. Um dies zu sehen, betrachten wir den bedingten E-Wert

$$E(X|\mathcal{C}) := \sum_{j=1}^{m} \frac{1}{P(C_j)} \int_{C_j} xP_j(dx) \cdot 1_{C_j},$$

wobei $X : x \mapsto x$ die Identität bezeichnet, und überzeugen uns davon, daß in der Konstruktion (3) und (4) das W-Maß Q einfach das Bildmaß von P unter $E(X|\mathcal{C})$ ist, also $Q = P * E(X|\mathcal{C})$.

Die eben beschriebene Sichtweise läßt sich nun wesentlich allgemeiner fassen.

Theorem 3.2: *Es sei $(\Omega, \mathcal{A}, W)$ ein W-Raum und $Z : \Omega \to E$ eine meßbare Abbildung. Weiters seien $\mathcal{C}_1$ und $\mathcal{C}_2$ Unter-σ-Algebren von $\mathcal{B}$ und*

$$P := W * E(Z|\mathcal{C}_1), \quad Q := W * E(Z|\mathcal{C}_2).$$

Dann gilt $\mathcal{C}_1 \supseteq \mathcal{C}_2 \;\Rightarrow\; P \succ Q$.

Mathematisch gesprochen besagt dieser Satz, daß für W-Maße, die als Bildmaße von bedingten Erwartungswerten gebildet werden, die Inklusionshalbordnung für σ-Algebren in der Majorisierungshalbordnung für W-Maße enthalten ist. Dieser abstrakte Sachverhalt hat eine informationstheoretische Interpretation: Wenn die Verteilungen P und Q durch Restriktion eines W-Maßes W auf unterschiedliche Informationsmengen $\mathcal{C}_1$ und $\mathcal{C}_2$ entstehen, dann ist der Informationsvergleich (die Inklusionshalbordnung) in der Majorisierungsrelation enthalten.

Diese Tatsache ist auch deshalb von besonders großem Interesse, weil man zeigen kann, daß die Relation $P \succ Q$ immer auf diese Art und Weise auf die Inklusionshalbordnung für σ-Algebren zurückgeführt werden kann. Wir werden diese Umkehrung des vorangehenden Satzes im nächsten Unterabschnitt näher erläutern und beweisen.

Der Vollständigkeit halber führen wir nun den Beweis des Satzes (3.2).

Beweis: Zur Vereinfachung der Bezeichnung sei $X := E(Z|\mathcal{C}_1)$ und $Y := E(Z|\mathcal{C}_2)$. Als Vorbereitung für den eigentlichen Beweis zeigen wir zunächst, daß $E(X|Y) = Y$ W-f.ü., oder in anderen Worten

$$E(E(Z|\mathcal{C}_1)|E(Z|\mathcal{C}_2)) = E(Z|\mathcal{C}_2) \quad W - \text{f.ü.} \tag{7}$$

Da für jede Borelmenge B die Beziehung $\{E(Z|\mathcal{C}_2) \in B\} \in \mathcal{C}_2 \subseteq \mathcal{C}_1$ gilt, erhält man

$$\int_{E(Z|\mathcal{C}_2)\in B} E(Z|\mathcal{C}_1)dW = \int_{E(Z|\mathcal{C}_2)\in B} ZdW = \int_{E(Z|\mathcal{C}_2)\in B} E(Z|\mathcal{C}_2)dW,$$

woraus die behauptete Gleichung (7) folgt.

Es sei nun D ein Kern mit der Eigenschaft

$$D(y, B) = W(X \in B|Y = y) \text{ für } B \in \mathcal{B},\ Q - \text{f.ü.}$$

Dann gelten die Aussagen

$$\int D(y, B)Q(dy) = \int W(X \in B|Y)dW = W(X \in B) = P(B)$$

und

$$\int xD(Y, dx) = E(X|Y) = Y \quad W\text{-f.ü..}$$

Aus der zweiten Gleichung folgt

$$\int xD(y, dx) = y \quad Q\text{-f.ü.}$$

Daher sind die Bedingungen für $P \succ Q$ erfüllt. □

Es gibt mehrere äquivalente Definitionen der Majorisierungsrelation. Wir diskutieren im folgenden eine Version, die in Zusammenhang mit der Informationstheorie steht.

Sind P und Q zwei W-Maße auf $\mathbb{R}^d$, dann gibt es im allgemeinen zahlreiche W-Maße R auf $\mathbb{R}^d \times \mathbb{R}^d$, deren Randverteilungen mit P bzw. Q übereinstimmen. Man kann nun die Gültigkeit von $P \succ Q$ dadurch charakterisieren, daß ein solches W-Maß R mit einer ganz bestimmten Eigenschaft existiert.

Es seien X und Y die Projektionen von $\mathbb{R}^d \times \mathbb{R}^d$ auf die Komponenten, und zwar sei $X : (x, y) \mapsto x$ und $Y : (x, y) \mapsto y$.

Theorem 3.3: *Folgende Aussagen sind äquivalent:*

1. $P \succ Q$.

2. Es gibt ein W-Maß $R|\mathcal{B}^2$, sodaß $R * X = P$, $R * Y = Q$, und $E_R(X|Y) = Y$.

Die informationstheoretische Interpretation lautet folgendermaßen. Die Zufallsgrößen X und Y werden als Output und Input eines Informationskanals angesehen. Der Informationskanal stört das Inputsignal Y, wodurch das Outputsignal X stochastisch verfälscht erscheint. Die Bedingung 2 des Satzes besagt nun, daß die Störung keine systematische Verfälschung bewirkt, sondern erwartungstreu erfolgt. Mit dieser Interpretation läßt sich der Satz in folgender Weise umformulieren:

Eine Verteilung P majorisiert eine Verteilung Q genau dann, wenn es einen erwartungstreuen Informationskanal gibt, der Q als Inputverteilung und P als Outputverteilung besitzt.

Beweis:

$1. \Rightarrow 2.$: Es sei D eine Dilation von Q nach P. Wir definieren

$$R(A \times B) = \int_B D(y, A)Q(dy), \quad A, B \in \mathcal{B}.$$

Offensichtlich ist $R * X = P$ und $R * Y = Q$. Außerdem gilt

$$\int_B X dR = \int_B \int x D(y, dx) Q(dy) = \int_B Y dR,$$

woraus sich die Gleichung $E_R(X|Y) = Y$ ergibt.

$2. \Rightarrow 1.$: Es sei R ein W-Maß mit den geforderten Eigenschaften. Wir definieren D als einen Kern mit der Eigenschaft

$$D(y, A) = R(X \in A|Y = y), \quad A \in \mathcal{B},\ R - \text{f.ü.}$$

Dann ist zunächst klar, daß die Bedingung (5) erfüllt ist. Die Bedingung (6) ergibt sich aus der Gleichung $E_R(X|Y) = Y$. □

Als Folgerung erhalten wir die angekündigte Umkehrung von Satz (3.2).

Korollar 3.4: *Folgende Aussagen sind äquivalent:*

1. $P \succ Q$.

2. Es gibt einen W-Raum $(\Omega, \mathcal{A}, W)$*, eine meßbare Abbildung* $Z : \Omega \to \mathbb{R}^d$ *sowie Unter-*σ*-Algebren* $\mathcal{C}_1 \supseteq \mathcal{C}_2$ *von* $\mathcal{A}$*, sodaß*

$$P := W * E(Z|\mathcal{C}_1), \quad Q := W * E(Z|\mathcal{C}_2).$$

Beweis: Die Implikation $2. \Rightarrow 1.$ ist der Inhalt des Satzes (3.2). Für den Beweis der Umkehrung $1. \Rightarrow 2.$ verwenden wir den Satz (3.3).

Wir definieren $\Omega := (\mathbb{R}^d)^2$, $\mathcal{A} := \mathcal{C}_1 := \mathcal{B}(\mathbb{R}^d)^2$, $\mathcal{C}_2 := Y^{-1}(\mathcal{B})$, $W := R$ und $Z := X$. Dann ist leicht nachzurechnen, daß die unter 2. beschriebene Situation vorliegt. □

Die Majorisierungsrelation ist eine Halbordnung auf der Menge aller W-Maße, d.h. es gelten die Aussagen

$$P \succ P,$$
$$P \succ Q, Q \succ R \Longrightarrow P \succ R.$$

Aus der Definition (3.1) ist außerdem ersichtlich, daß W-Maße, die hinsichtlich der Majorisierungsrelation vergleichbar sind, gleiche Mittelwerte haben müssen.

Aus der Ungleichung von Jensen ergibt sich, daß

$$P \succ Q \Longrightarrow \int f\, dP \geq \int f\, dQ \tag{8}$$

für jede konvexe Funktion $f : \mathbb{R}^d \to \mathbb{R}$. Daraus folgt unter anderem, daß die Halbordnung der Majorisierung sogar identifizierend ist.

Das zentrale Resultat der Theorie der Majorisierung ist das sogenannte Dilationskriterium, das auch als Satz von Blackwell-Sherman-Stein bekannt ist. Die Aussage dieses Satzes besteht darin, daß die Gültigkeit der Ungleichungen (8) nicht nur notwendig, sondern sogar hinreichend für $P \succ Q$ ist.

Theorem 3.5: *Folgende Aussagen sind äquivalent:*

1. $P \succ Q$.
2. $\int f\,dP \geq \int f\,dQ$ *für alle konvexen Funktionen* $f : \mathbb{E} \to \mathbb{R}$.

Der Beweis der Implikation (2)$\Longrightarrow$(1) hat eine lange Geschichte, die mit Hardy, Littlewood und Polya (1929) beginnt. Die von uns zitierte Version des Satzes findet sich zum Beispiel in Torgersen (1991), Theorem 7.2.17, p. 347, (wobei dort unter (ii) die Gleichung $T = DS$ durch $S = DT$ zu ersetzen ist).

4 Majorisierung und Konzentrationsmessung

Wir beginnen daher mit einem Überblick über die Geschichte dieses Begriffs.

Das Konzept der Majorisierung hat eine lange Geschichte und zahlreiche verschiedene Anwendungen. Der Bogen der Anwendungen spannt sich von der statistischen Konzentrationsmessung über innermathematische Anwendungen in Zusammenhang mit Ungleichungen bis hin zur Informationstheorie und zur statistischen Entscheidungstheorie. Das Buch von Marshall und Olkin (1979) gibt Auskunft über die vielfältigen innermathematischen Anwendungen des Majorisierungsbegriffs. Blackwell (1951) entdeckte einen Zusammenhang des Majorisierungsbegriffs zur statistischen Entscheidungstheorie. Die Anwendung des Majorisierungsbegriffs in der statistischen Entscheidungstheorie erhielt entscheidende Impulse durch LeCam (1964) und Torgersen (1970). Eine ausführliche Darstellung dieser Resultate und eine Erweiterung vieler der im Rahmen der statistischen Entscheidungstheorie gewonnenen Resultate auf das ursprüngliche allgemeine Konzept der Majorisierung finden sich bei Torgersen (1991).

Die klassische Anwendung der Majorisierungsrelation ist die Messung der Vermögenskonzentration in einer Bevölkerung. Historisch gesehen ging das Konzept der Majorisierung als mathematische Abstraktion aus der Konzentrationsmessung hervor. Details dazu finden sich in Marshall und Olkin (1979). Wir gehen den umgekehrten Weg und werden die ökonomischen Interpretationen aus den allgemeinen Konzepten gewinnen. Auf diesem Weg ist es zweckmäßig, den Spezialfall von diskreten Maßen etwas genauer anzusehen.

Es seien

$$P = \sum_{i=1}^{m} \alpha_i \epsilon_{x_i} \quad \text{und} \quad Q = \sum_{j=1}^{n} \beta_j \epsilon_{y_j}$$

zwei W-Maße auf $\mathbb{R}^d$. Wir lassen ausdrücklich zu, daß die Trägerpunkte x_i bzw y_j nicht alle paarweise verschieden sind.

Eine Dilation D wird durch eine spaltenstochastische $m \times n$-Matrix $W = (w_{ij})$ dargestellt, indem

$$w_{ij} := D(\{x_i\}, y_j)$$

definiert wird. Die Darstellung von D lautet dann

$$D(A, y) = \sum_{j=1}^{n} \sum_{i=1}^{m} w_{ij} \epsilon_{x_i}(A) 1_{\{y_j\}}(y).$$

Es gelte nun $P \succ Q$. Übersetzt man nun die Gleichungen (1) und (2), so ergibt sich

$$\begin{aligned}\alpha_i = P(\{x_i\}) &= \int D(\{x_i\}, y) Q(dy) \\ &= \sum_{j=1}^{n} D(\{x_i\}, y_j)\beta_j = \sum_{j=1}^{n} w_{ij}\beta_j.\end{aligned}$$

und

$$\begin{aligned}y_j &= \int x D(dx, y_j) \\ &= \sum_{i=1}^{m} x_i D(\{x_i\}, y_j) = \sum_{i=1}^{m} x_i w_{ij}.\end{aligned}$$

Die Beziehung $P \succ Q$ kann für diskrete Maße in der Sprache der linearen Algebra ausgedrückt werden. Zu diesem Zweck fassen wir die Gewichte zu Spaltenvektoren zusammen, also

$$a = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}, \quad b = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix},$$

und die Trägerpunkte $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}^d$ ordnen wir als Spaltenvektoren von Matrizen an,also

$$A = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{m1} \\ x_{12} & x_{22} & \dots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1d} & x_{2d} & \dots & x_{md} \end{pmatrix},$$

und

$$B = \begin{pmatrix} y_{11} & y_{21} & \dots & y_{n1} \\ y_{12} & y_{22} & \dots & y_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1d} & y_{2d} & \dots & y_{nd} \end{pmatrix}.$$

Mit diesen Bezeichnungen gilt dann der folgende Satz.

Theorem 4.1: *Es gilt genau dann $P \succ Q$, wenn es eine spaltenstochastische Matrix W gibt, sodaß $a = Wb$ und $AW = B$.*

Die Gleichung $AW = B$ besagt, daß die Trägerpunkte des W-Maßes Q Konvexkombinationen der Trägerpunkte des W-Maßes P sind. Die Gleichung $a = Wb$ kann dahingehend interpretiert werden, daß bei der Bildung der Trägerpunkte von Q aus den Trägerpunkten von P die Gewichtsbilanz ausgeglichen ist. Diese Interpretation steht in Zusammenhang mit der klassischen ökonomischen Deutung der Majorisierungsrelation.

Wir betrachten eine Population, bei der das Vermögen eines Mitglieds durch einen Vektor $x \in \mathbb{R}^d$ repräsentiert wird. Es gibt m verschiedene Vermögensstrukturen $x_1, x_2, \dots,$

x_m, die mit den relativen Häufigkeiten $\alpha_1, \alpha_2, \ldots, \alpha_m$ auftreten. Die Verteilung der Vermögen wird also durch die Häufigkeitstabelle

x_1	α_1
x_2	α_2
$\vdots$	$\vdots$
x_m	α_m

,

angegeben, also durch das diskrete Maß

$$P = \sum_{i=1}^{m} \alpha_i \epsilon_{x_i}.$$

Unter einem Vermögenstransfer oder einer Umverteilung versteht man die Bildung neuer Vermögensstrukturen mit neuen Häufigkeiten, dh.

y_1	β_1
y_2	β_2
$\vdots$	$\vdots$
y_n	β_n

, bzw. $Q = \sum_{j=1}^{n} \beta_j \epsilon_{y_j}$,

nach folgender Methode: Jedes Vermögen x_i der ursprünglichen Vermögensverteilung wird mit dem Anteil w_{ij} zur Bildung eines Vermögens y_j der neuen Vermögensverteilung herangezogen. Das bedeutet, daß

$$y_j = \sum_{i=1}^{m} x_i w_{ij} \text{ für } 1 \leq j \leq n, \tag{9}$$

daß also die Gleichung $AW = B$ aus Satz (4.1) erfüllt ist. Zur Bildung der neuen Häufigkeitsverteilung werden

$$w_{i1}\beta_1 + w_{i2}\beta_2 + \cdots + w_{in}\beta_n$$

Vermögen mit der Struktur x_i benötigt. Daher muß

$$\alpha_i = \sum_{j=1}^{n} w_{ij}\beta_j \text{ für } 1 \leq i \leq m,$$

sein, es muß also auch die Gleichungen $a = Wb$ aus Satz (4.1) erfüllt sein.

Die durch $W = (w_{ij})$ definierte Umverteilung ist nivellierend, wenn die Linearkombinationen (9) sogar Konvexkombinationen sind. In anderen Worten heißt das, daß W eine spaltenstochastische Matrix ist.

Damit haben wir eine Deutung der Majorisierungsrelation erhalten: Für die Vermögensverteilungen P und Q gilt genau dann $P \succ Q$, wenn Q aus P durch eine nivellierende Umverteilung hervorgeht.

Das Convex-Function Kriterium (3.5) besagt in diesem Zusammenhang, daß bei progressiven Steuertarifen eine nivellierende Umverteilung das Steueraufkommen mindert.

Der einfachste Spezialfall der Majorisierungsrelation tritt liegt dann vor, wenn $d = 1$, $m = n$, und $\alpha_i = \beta_j = 1/n$ für alle i und j. In diesem Fall werden die W-Maße P und Q durch die Vektoren $x \in \mathbb{R}^n$ und $y \in \mathbb{R}^n$ ihrer Trägerpunkte repräsentiert. Es handelt sich dann bei P und Q einfach um empirische Verteilungen von Datenlisten der Länge n in $\mathbb{R}$, also

$$P = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{x_i}, \quad Q = \frac{1}{n}\sum_{j=1}^{n} \varepsilon_{y_j}.$$

Da in diesem Fall die Gewichte α_i und β_j alle gleich groß sind, bedeutet die Gleichung $a = Wb$ aus Satz (4.1), daß die Matrix W auch zeilenstochastisch ist. Wir erhalten auf diese Weise:

Korollar 4.2: *Es seien P und Q die empirischen Verteilungen zweier Vektoren x und y in $\mathbb{R}^n$. Es gilt $P \succ Q$ genau dann, wenn es eine doppelt-stochastische Matrix W gibt, sodaß $x'W = y'$.*

Das in dieser Aussage verwendete Kriterium ist, historisch gesehen, eines der frühesten Definitionen der Majorisierungsrelation. Allerdings verwendet das ursprüngliche Konzept der Konzentrationsmessung, welches auf Dalton (1920) zurückgeht, eine noch einfachere Formulierung. Wenn die Datenliste y aus der Datenliste x dadurch hervorgeht, daß für zwei bestimmte Indizes i und j

$$\begin{aligned} y_i &= \alpha x_i + (1-\alpha)x_j,\ 0 \le \alpha \le 1, \\ y_j &= \beta x_i + (1-\beta)y_j,\ 0 \le \beta \le 1, \end{aligned}$$

so spricht man von einem sogenannten Transfer. Jeder Transfer kann durch eine geeignete doppeltstochastische Matrix W und die Operation $AW = B$ dargestellt werden. Die ursprüngliche Definition von $P \succ Q$ bestand nun darin, daß Q aus P durch eine Folge von Transfers hervorgeht. Bemerkenswert ist nun die Tatsache, daß diese Definition mit der in Korollar (4.2) verlangten Bedingung sogar äquivalent ist. Für den eindimensionalen Fall $d = 1$ stammt dieses Resultat von Hardy, Littlewood und Polya (1929). Die Übertragung des Resultats auf den mehrdimensionalen Fall $d > 1$ findet sich zum Beispiel bei Marshall und Olkin (1979). Für weitere historische Informationen verweisen wir eberfalls auf Marshall und Olkin (1979).

Das Convex-Function Kriterium lautet nun:

Korollar 4.3: *(Hardy, Littlewood und Polya, 1929) Es seien P und Q die empirischen Verteilungen zweier Vektoren x und y in $\mathbb{R}^n$. Es gilt $P \succ Q$ genau dann, wenn*

$$P \succ Q \iff \sum_{i=1}^{n} f(x_i) \ge \sum_{i=1}^{n} f(y_i) \quad \textit{für allen konvexen Funktionen } f.$$

Der Zusammenhang zwischen der Majorisierungsrelation auf $\mathbb{R}$ und der mathematischen Theorie der Ungleichungen führte zu einer großen Zahl neuer Ergebnisse. Das Buch von Marshall und Olkin (1979) gibt Auskunft über die vielfältigen innermathematischen Anwendungen des Majorisierungsbegriffs.

5 Optimale Quantisierung

Es ist klar, daß eine Verteilung P viele verschiedene Quantisierungen besitzen kann. Selbst wenn man sich auf Quantisierungen mit einer festen Anzahl m von Trägerpunkten beschränkt, ist die Anzahl der möglichen Quantisierungen unübersehbar groß. Dies liegt schon alleine daran, daß jede denkbare Zerlegung eines Datensatzes in m Teilmenge auf die schon früher beschriebene Weise zu einer Quantisierung führt.

Durch die Quantisierung eines Datensatzes entsteht zwangsläufig ein Informationsverlust. Angesichts der Tatsache, daß es sehr viele Möglichkeiten gibt, einen gegebenen Datensatz zu quantisieren, stellt sich die Frage nach den Gesichtspunkten und den Methoden, die bei der Herstellung einer Quantisierung angewandt werden sollen.

Hier gilt es zwei Gesichtspunkte zu berücksichtigen. Jede Quantisierung einer Verteilung führt einerseits zu einer Vereinfachung, aber andererseits zu einem Informationsverlust. Eine plausible Strategie bei der Konstruktion einer Quantisierung kann nun darin bestehen, unter der Bedingung einer bestimmten gewünschten Einfachheit den Informationsverlust der Quantisierung möglichst gering zu halten. Sehen wir uns nun an, wie sich ein solcher programmatischer Gedanke präzisisieren und möglicherweise realisieren läßt.

Die Einfachheit einer Quantisierung wird am besten durch die Anzahl der Trägerpunkte beschrieben. Je weniger Trägerpunkte, desto einfacher ist die Quantisierung. Die einfachste Quantisierung besteht in einer Einpunktverteilung, die ihre Masse im Mittelpunkt der Verteilung P konzentriert. Bei dieser Quantisierung ist der Informationsverlust am größten. Lassen wir komplexere Quantisierungen zu, dann können wir deren Komplexität durch eine obere Schranke für die Anzahl der Trägerpunkte in Grenzen halten. Es sei also

$$\mathcal{D}_m(P) := \{Q : P \succ Q, \, \mathrm{supp}(Q) \leq m\}$$

die Menge aller Quantisierungen von P mit höchstens m Trägerpunkten. Die Aufgabe besteht nun darin, in der Menge $\mathcal{D}_m(P)$ solche Quantisierungen zu finden, für die der Informationsverlust gemessen an P möglichst klein ist.

Wie soll aber der Informationsverlust einer Quantisierung gemessen werden? Ein naheliegender Gedanke besteht darin, ein skalarwertiges Informationsmaß zu verwenden. Dies ist eine Möglichkeit, auf die wir später noch zurückkommen werden. Zunächst gehen wir aber von der Majorisierungshalbordnung selbst aus.

Es seien Q_1und Q_2 zwei Quantisierungen in $\mathcal{D}_m(P)$. Dann kann es sein, daß eine der beiden Quantisierungen die andere majorisiert, also zum Beispiel $Q_1 \succ Q_2$. Das bedeutet, daß die Quantisierung Q_2 aus der Quantisierung Q_1 durch Aufteilung und anschließende Mittelwertbildung hervorgeht. Dieser Vorgang kann bei Q_2 auf keinen Fall zusätzliche Information über P erzeugen, die nicht schon in der Quantisierung Q_1 enthalten wäre. Wenn also der Fall $Q_1 \succ Q_2$ vorliegt, dann ist die Quantisierung Q_1 vorzuziehen.

Da die Majorisierungsrelation nur eine Halbordnung ist, kann man nicht erwarten, daß $\mathcal{D}_m(P)$ ein Maximum besitzt. Es ist nur realistisch, maximale, dh. unübertreffliche Quantisierungen zu suchen. Daraus ergibt sich ein einfacher Gesichtspunkt für die Auswahl von Quantisierungen in der Menge $\mathcal{D}_m(P)$. Wir sind nur an solchen Quantisierungen interessiert, die durch keine andere Quantisierung majorisiert werden.

Definition 5.1: *Eine Quantisierung $Q^* \in \mathcal{D}_m(P)$ ist zulässig oder maximal, wenn sie von keiner anderen Quantisierung majorisiert wird, also wenn*

$$Q \succ Q^*, Q \in \mathcal{D}_m(P) \Longrightarrow Q = Q^*.$$

Bevor man sich mit der Frage befaßt, wie man zulässige Quantisierungen findet, ist es nötig, einige theoretische Absicherungen vorzunehmen. Resultate in dieser Richtung wurden von Pötzelberger (2000a,b) erzielt. In gewissem Sinn wurde damit eine vollständige theoretische Lösung des Problem für den Fall erzielt, bei dem das W-Maß P stetig ist (eine Lebesgue-Dichte besitzt). Die Beweise dieser Resultate sind mathematisch sehr anspruchsvoll, und wir können daher hier nur eine Übersicht anbieten.

Das erste Resultat sichert die Existenz von zulässigen Quantisierungen.

Theorem 5.2: *(Pötzelberger, 2000b) Es sei P ein stetiges W-Maß. Für jede Quantisierung $Q \in \mathcal{D}_m(P)$ gibt es eine bessere und zugleich zulässige Quantisierung $Q^* \in \mathcal{D}_m(P)$, sodaß also $Q \prec Q^*$.*

Das nächste Resultat ist erstaunlich und überaus wichtig. Es zeigt nämlich, daß zulässige Quantisierungen immer auf Zerlegungen beruhen, also auf einer sehr speziellen Konstruktion von Quantisierungen. Darüber hinaus sind die dabei in betracht zu ziehenden Zerlegungen von einer sehr speziellen und einfachen Struktur.

Theorem 5.3: *(Pötzelberger, 2000b) Es sei P ein stetiges W-Maß. Für jede zulässige Quantisierung $Q^* \in \mathcal{D}_m(P)$ wird von einer Zerlegung $\mathcal{C} = (C_1, C_2, \ldots, C_m)$ bestehend aus m konvexen Polytopen erzeugt.*

Damit sind die Quantisierungen, die für die praktische Anwendung in Betracht kommen, bereits sehr stark eingeschränkt. Allerdings ist die Vereinfachung noch nicht geeignet, Algorithmen zur Berechnung von zulässigen Quantisierungen anzugeben. Dies ist erst auf Grund des nächsten Resultates möglich.

Es geht also um die Frage, wie man zulässige Quantisierungen konstruieren kann. Für eine beliebig gewählte konvexe Funktion f definieren wir die Menge

$$\mathcal{O}_m(P, f) := \Big\{ Q^* \in \mathcal{D}_m(P) : \int f \, dQ^* = \sup_{Q \in \mathcal{D}_m(P)} \int f \, dQ \Big\}.$$

Diese Menge $\mathcal{O}_m(P, f)$ enthält Quantisierungen, die durch spezielle Optimierungsaufgabe mit einer skalarwertigen Zielfunktion definiert sind. Wir werden im nächsten und letzten Abschnitt dieser Arbeit darüber berichten, daß es für die numerische Lösung dieser Optimierungsprobleme überaus wirksame Algorithmen gibt, die durch Bock (1992) und Pötzelberger und Strasser (2001) definiert und genau untersucht wurden.

Das dritte Resultat besagt, daß mit diesen praktisch lösbaren Optimierungsprobleme ausschließlich zulässige Quantisierungen gewonnen werden.

Theorem 5.4: *(Pötzelberger, 2000b) Es sei P ein stetiges W-Maß und f eine konvexe Funktion, welche nicht das Maximum von weniger als m affin-linearen Funktionen ist. Dann sind alle Quantisierungen in $\mathcal{O}_m(P, f)$ zulässig in $\mathcal{D}_m(P)$.*

Das letzte theoretische Resultat betrifft die Frage, ob man durch die beschriebenen Optimierungsprobleme im wesentlichen alle zulässigen Quantisierungen erhalten kann. Dies ist in einem approximativen Sinn tatsächlich der Fall.

Theorem 5.5: *(Pötzelberger, 2000b) Es sei P ein stetiges W-Maß und $Q \in \mathcal{D}_m(P)$ sei so daß $|supp(Q)| = m$. Diese Quantisierungen Q ist genau dann zulässig wenn es eine Folge von konvexen Funktionen f_n gibt, von denen keine das Maximum von weniger als $m-1$ affin-linearen Functionen ist, sodaß*

$$Q = \lim_{n\to\infty} Q_n \text{ schwach,wobei } Q_n \in \mathcal{O}_m(P, f_n).$$

Literatur

D. Blackwell. Comparison of experiments. In L. LeCam and J. Neyman, editors, *Proc. 2nd Berkeley Symp. Math. Statistics Prob.*, pages 93–102, 1951.

H.H. Bock. A clustering technique for maximizing ϕ-divergence, noncentrality and discriminating power. In M. Schader, editor, *Analyzing and Modeling Data and Knowledge*, pages 19–36, Berlin Heidelberg New York, 1992. Springer Verlag.

H. Dalton. The measurement of the inequality of incomes. *Econom. J.*, 30:348–361, 1920.

G.H. Hardy, J.E. Littlewood, and G. Polya. Some simple inequalities satisfied by convex functions. *Messenger Math.*, 58:145–152, 1929.

L. LeCam. Sufficiency and approximate sufficiency. *Ann. Math. Statist.*, 35:1419–1455, 1964.

A.W. Marshall and I. Olkin. *Inequalities: Theory of majorization and its applications*. Academic Press, 1979.

K. Pötzelberger. Admissible unbiased quantizations: Distributions with linear components. Technical report, Department of Statistics, Vienna University of Economics and Business Administration, 2000a.

K. Pötzelberger. Admissible unbiased quantizations: Distributions without linear components. Technical report, Department of Statistics, Vienna University of Economics and Business Administration, 2000b.

K. Pötzelberger and H. Strasser. Clustering and quantization by msp-partitions. *Statistics and Decisions*, 19:331–371, 2001.

G. Steiner. *Statistical Data Compression by Optimal Segmentation*. Service Fachverlag, Wien, 2000.

H. Strasser. Data compression and statistical inference. In T. Kollo, editor, *Proceedings of the 6th Tartu Conference on Multivariate Statistics (Satellitemeeting of ISI 52nd Session)*, 2000a.

H. Strasser. Reduction of complexity. In J. Mazanec and H. Strasser, editors, *A Nonparametric Approach to Perceptions-Based Market Segmentation: Foundations*. Springer, Berlin, 2000b.

H. Strasser. Towards a statistical theory of optimal quantization. Technical report, Department of Statistics, Vienna University of Economics and Business Administration, 2000c.

E.N. Torgersen. Comparison of experiments when the parameter space is finite. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 16:219–249, 1970.

E.N. Torgersen. *Comparison of statistical experiments*. Cambridge Univ. Press, 1991.

Adresse des Autors:

o.Univ.Prof. Dr. Helmut Strasser
Abteilung für experimentelle Mathematik und Statistik
Wirtschaftsuniversität Wien
UZA II, 5. Stock
Augasse 2-6
A-1090 Wien

Tel. +43 1 31336 5051
Fax +43 1 31336 734
E-Mail: Helmut.Strasser@wu-wien.ac.at