MÜNCHENER STOCHASTIK-TAGE 1998 **Sekt.-Nr.:3**

Datenanalyse und robuste Verfahren

Sektionsleiter: F. Pukelsheim, Augsburg; A. Unwin, Augsburg

Gallegos, Maria Teresa und Gunter Ritter
Fakultät für Mathematik und Informatik, Universität Passau

Bayessche Klassifikation und Selektion

In der klassischen Diskriminanzanalyse möchte man ein Objekt aufgrund seines Merkmalsvektors einer von n verschiedenen Klassen zuordnen. Die verschiedenen Klassen sind durch ihre *bekannten* Verteilungen charakterisiert. Es gibt ein zur Klassifikation *duales* Problem, das wir *Selektion* nennen: Wir beobachten verschiedene *Varianten* eines Objektes, d.h., es gibt nur ein Objekt aber Merkmale werden aus verschiedenen Varianten dieses Objektes extrahiert. Jede Variante ist durch einen Merkmalsvektor dargestellt. Unter anderem soll diejenige Variante geschätzt werden, die eine vorgegebene Klasse normal repräsentiert. Diese Variante nennen wir die *normale Variante.*
Die gemeinsame Verteilung des Tupels der Varianten ist unbekannt. Man kennt nur die Verteilung der *normalen* Variante. Diese Tatsache verhindert oft die Berechnung eines Bayes–Schätzers für die Position der normalen Variante. Es werden solche Schätzer für Klassifikation und Selektion mit und ohne Nebenbedingungen für diese Situation angegeben, die nur von der Verteilung der normalen Varianten abhängen und eine möglichst kleine Fehlerrate haben.

Hennig , Christian
Institut für Mathematische Stochastik , Universität Hamburg

Fixed Point Clusters: Using outlier identification for finding clusters

For a stochastical approach to cluster analysis, a cluster in a dataset is considered as a subset whose points are generated by a common distribution $P \in \mathcal{P}_0$. $\mathcal{P}_0$ denotes the class of „cluster reference distributions", i.e. the distributions generating datasets which are treated as „homogeneous". A mixture model

$$Q_0 = \sum_{i=1}^{s} \epsilon_i P_i, \ \sum_{i=1}^{s} \epsilon_i = 1, \ \epsilon_i > 0, P_i \in \mathcal{P}_0, i = 1, \ldots, s.$$

is often used as a model for the whole dataset. This model is a special case of the „contamination model"

$$Q_1 = (1-\epsilon)P_0 + \epsilon P^*, \ 0 \leq \epsilon < 1, \ P_0 \in \mathcal{P}_0,$$

where the points from P^* can be interpreted as „outliers" w.r.t. P_0. The aim of Fixed Point Clustering is to find subsets of the data which belong together in the sense that they contain no outliers, but all other points of the dataset are outliers w.r.t. the subset. If P^* is well seperated from P_0, this can be used to find the points generated by P_0 without the assumption of the mixture model that the rest of the dataset is partitioned into clusters of the same kind. Clusters can overlap and there can be points in the dataset which do not belong to any cluster. The concept is applied to the problem of linear regression clustering.

Henschel, Volkmar
Universität Rostock

Exakte Verteilungen von Teststatistiken für die Parameter der Pareto-Verteilung

Die Pareto-Verteilung beschreibt ein einfaches, aber gut handhabbares Modell für ökonomische Daten, z.B. für die Verteilung von Einkommen. In Muniruzzaman (1957) sind Teststatistiken zur Prüfung von Hypothesen über die Parameter der Pareto-Verteilung und deren Verteilung unter der Nullhypothese angegeben.
In diesem Vortrag werden exakte Verteilungen von Tests für die Parameter der Pareto-Verteilung sowohl unter der Nullhypothese als auch unter der Alternative angegeben. Dabei werden Typ II zensierte Daten zugelassen, wie in der Arbeit von Chen (1996).
Zur Herleitung der Resultate wird die enge Verbindung der Pareto-Verteilung zur Exponentialverteilung genutzt. Für jene wurde eine einfache Maßdarstellungsformel im n-dimensionalen Stichprobenraum entwickelt, s. Henschel und Richter (1997). Die dort entwickelten Methoden können auf die Pareto-Verteilung angewandt werden. Auf diesem Wege wird eine Verallgemeinerung der Pareto-Verteilung hergeleitet.

Literatur

Z. Chen. Joint confidence region for the parameters of pareto distribution. *Metrika*, 44(3):191–198, 1996.

V. Henschel und W.–D. Richter. Geometric Generalization of the Exponential Law. *Preprints aus dem Fachbereich Mathematik*, Universität Rostock, Preprint 97/14, 1997.

A.N.M. Muniruzzaman. On measures of location and dispersion and tests of hypotheses on a Pareto population. *Bulletin of the Calcutta Statistical Association.* 7:115–123, 1957.

Müller , Christine
Georg-August-Universität

Schätzung von Varianzkomponenten mit hohem Bruchpunkt und hoher Effizienz

Für das einfach-klassifizierte Beobachtungsmodell mit zufälligen Effekten werden drei neue Varianzkomponenten-Schätzungen betrachtet. Die neuen Schätzungen basieren auf verallgemeinerten L-Statistiken und wurden vor kurzem von Uhlig (1995) vorgeschlagen. Für diese Schätzungen werden die Blockbruchpunkte, die Messungsbruchpunkte und die Effizienzen im Vergleich zu den ANOVA-Schätzungen vorgestellt. Es wird gezeigt, daß die Blockbruchpunkte ungefähr 50 % betragen und daß die Messungsbruchpunkte nicht kleiner als 25 % sind. Durch Herleitung der asymptotischen Normalverteilung der Schätzungen kann auch die asymptotische Effizienz bestimmt werden. Diese ist bei zwei von den Schätzungen größer als 82 % und erreicht in bestimmten Fällen sogar 95 %. Eine Simulationsstudie bestätigt die hohe Effizienz auch für kleine Stichproben. Auf einen Datensatz aus einem Ringversuch zur Überprüfung von Laboren werden die neuen Schätzungen angewendet.

Olbricht , Walter
Universität Bayreuth

Zur Erstellung intern konsistenter thermodynamischer Daten von Mineralen

Sogenannte intern konsistente thermodynamische Datenbasen sind von grundlegender Bedeutung in der Mineralogie, z. B. für die Berechnung von Phasendiagrammen. In dem Vortrag wird das zugrundeliegende Problem kurz erläutert und einschließlich der bisher dafür hauptsächlich verwendeten Methoden (Regressionsansatz und mathematische Programmierung) aus statistischer Sicht analysiert. Weiterhin wird ein Bayes–Ansatz vorgeschlagen, um die vorhandene Information aus kalorimetrischen Messungen und aus Reaktionsumkehrversuchen miteinander zu verbinden. Dieses Verfahren ist konzeptionell übersichtlich, aber rechentechnisch aufwendig. Zu seiner Realisierung wird daher Gibbs–Sampling benutzt. Die bisherigen Ergebnisse werden kurz skizziert. Der Vortrag basiert auf gemeinsamer Arbeit mit N. D. Chatterjee, K. Miller und R. Krüger.

Rieder und Ruckdeschel , Helmut, bzw. Peter
Universiät Bayreuth

Robuste Kalman–Filter

Ausgehend von einer robusten Variante des L_2-Abstands und robusten Influenzkurven für Regression schlagen wir zwei neue Filter vor, die gegen additive Ausreißer robust sind. Ein dritter Filter ist optimal und basiert auf einem Analogon von Hampels Lemma 5 für das Zustandsraummodell. Zur Behandlung von Innovationsausreißern wird die Rekursivität aufgegeben und ein Zusammenhang zwischen dem Filter von Boncelet und posterior Modus hergestellt.

Stuetzle , Werner **Hauptvortrag**
University of Washington

Estimation and Modeling Problems in Computer Graphics

3D photography is an emerging technology for capturing, viewing, and manipulating shape and visual appearance of 3D objects. It can be regarded as the inverse of computer aided manufacturing (CAM). The goal of CAM is to make physical objects from computer (CAD) models. The goal of 3D photography is to generate CAD models from physical objects. This requires sensors (shape cameras) and algorithms to convert the raw data produced by the sensor (typically colored 3D points on the surface of the object) into surface models. These models can then be viewed (rotated, etc) on displays, transmitted over the Internet, edited, etc. In the first part of my talk I will sketch the basic components of the technology and discuss potential applications. In the second part I will review statistical and computational problems, such as surface estimation, mesh simplification and multiresolution analysis of meshes.

Wilhelm , Adalbert
George Mason University, Fairfax VA and Universität Augsburg

An algebraic framework for selection and linking in high-interaction graphics

Selecting subsets in one graph and propagating this selection to other graphic displays (linking) are key features of high interaction graphics. In many statistical software packages these are only available in the form of scatterplot brushing, but other plot types can be succesfully linked in a similar way.
Based on the formal partition of a graph into four levels - frame, plot, variables and sample population - a taxonomy of selection procedures is given. Information of the various graph levels that is

relevant with respect to selection can be combined in algebraic partitions of the sample population.
Linking can then be seen as a function between two partitions corresponding to different graphs.
If the image partition is a refinement of the domain partition, linking can be done by attributes.
Otherwise, we use overlaying to visualize subset selection.

Refinement conditions are presented that allow proper visualization of m-nary selections ($m > 2$), multiple selections and fuzzy selections.

Modifications of the involved functions and partitions lead to immediate extensions of selecting and linking, as to linking different sample populations, linking scales, linking variables or linking position in frames.

The main purpose of this formal description is to develop a paradigm of linking views, to find new, useful, linked graphs, and to detect inconsistent or meaningless graphs.