

Abschlussarbeiten bei der Arbeitsgruppe Metho(dolo)gische Grundlagen der Statistik und ihre Anwendungen

Prof. Dr. Thomas Augustin, Dr. Marco Cattaneo, Dr. Carolin Strobl,
Gero Walter, Andrea Wiencierz

4. Dezember 2009

1 Vorbemerkungen

1.1 Prinzipielle Typen von Abschlussarbeiten

- ausführlicher und strukturierter Literatur-Überblick über ein Thema
- Verallgemeinerung eines klassischen Verfahrens durch Adaption von verallgemeinerten Methoden (Intervallwahrscheinlichkeit, Messfehler)
- Implementierung und Untersuchung einer Methode, die bereits in der Theorie existiert, oder Vergleich zweier Methoden
- konzeptuelle oder explorative Weiterentwicklung von bisher nicht-etablierten Ansätzen
- Analyse eines komplexen Datenbeispiels mithilfe einer geeigneten Kombination etablierter Methoden

1.2 Bachelor's Thesis, Master's Thesis bzw. Diplomarbeit

Die meisten Themenbereiche eignen sich sowohl als Themen für Bachelorarbeiten als auch für Master- und Diplomarbeiten. Selbstverständlich wird bei der genauen Themenstellung auf die unterschiedlichen Bearbeitungszeiten und Vorkenntnisse Rücksicht genommen.

1.3 Kolloquium

Es soll ein regelmäßiges Kolloquium angeboten werden, um den Austausch zwischen den einzelnen Studierenden, die eine Abschlussarbeit schreiben, zu fördern und allgemeine Aspekte gemeinsam diskutieren zu können. Jede(r) Studierende soll dabei die Möglichkeit erhalten, die Arbeit etwa zum Zeitpunkt der Anmeldung und nach etwa zwei Drittel der Bearbeitungszeit kurz vorzustellen, um breites Feedback zu den Plänen bzw. Ergebnissen zu erhalten. In das Kolloquium können bei Bedarf auch allgemeine Tutorien (z.B. zur Literaturrecherche mit Datenbanken) integriert werden.

2 Themen

2.1 Fehler-in-den-Variablen

Fehler-in-den-Variablen-Modelle nehmen die Tatsache ernst, dass in den unterschiedlichsten Anwendungsgebieten häufig die eigentlich interessierenden Variablen nicht direkt und exakt beobachtbar sind; stattdessen stehen nur stark fehlerbehaftete Größen (Proxyvariablen) zur Verfügung. Typische Beispiele sind

- fehlerbehaftete technische Messungen i.e.S.,
- stark gerundete Daten durch eingeschränkte Messpräzision (Quantization Noise Problematik),
- durch Erinnerungseffekte verzerrte Daten,
- zur Anonymisierung bewusst kontaminierte Daten,
- Skalenindizes und andere Operationalisierungen komplexer Konstrukte (in den Wirtschafts- und Sozialwissenschaften, aber zum Beispiel auch in der Arbeitsepidemiologie).

Die meisten statistischen Verfahren setzen allerdings eigentlich vollständige und fehlerfrei beobachtete Daten voraus. Daher versucht man mithilfe von Fehler-in-den-Variablen-Modellen, den Fehlermechanismus selbst zu modellieren und damit geeignet in die statistische Analyse einzubinden bzw. den Einfluss der Fehler auf das Ergebnis der Analyse zu korrigieren, um gravierende Fehlschlüsse bei der Interpretation zu vermeiden.

Korrekturverfahren für Berkson-Fehler beim Cox-Modell

Fehlklassifikationsmodelle für vergrößerte Daten / Coarsening

Fehler-in-den-Variablen-Modelle und Intervalldaten

Anwendung bzw. Erweiterung von Fehler-in-den-Variablen-Verfahren auf intervallwertige Daten

Inverse probability weighted correction (IPWC)

Diese Methoden korrigieren Schätzungen aus fehlerbehafteten Datensätzen mithilfe einer Umgewichtung von Daten (ähnliche Idee wie Horvitz-Thompson-Schätzer in der Stichprobentheorie).

- Literaturüberblick über verschiedene IPWC-Ansätze für zensierten Daten
- Anwendung der sog. Methode der doppelt-robusten Korrektur
- IPWC anwenden auf *unabhängige* Variable in Regression (bisher nur auf *abhängige* Variable angewendet)

Sensitivitätsanalyse von Fehlerkorrekturmodellen bezüglich ihrer Annahmen über das Messfehlermodell

- Simulationsstudien (fehlerbehaftete Daten simulieren, Messfehlerkorrektur unter Annahmen die nicht dem Fehlerprozess entsprechen)
- Weiterentwicklung von Messfehlermodellen auf allgemeinere Verteilungsannahmen. Die Messfehler sind i.d.R. als normalverteilt angenommen. Was passiert bei „nicht ganz“ normalverteilten Fehlern?
- Systematisierte Sensitivitätsanalyse im Sinne der Partiellen Identifikation

Anonymisierungsverfahren

Immer häufiger stellen amtliche Einrichtungen (wie das Statistische Bundesamt oder die Bundesanstalt für Arbeit) Daten zur wissenschaftlichen Analyse zur Verfügung, wobei zur Anonymisierung die Daten absichtlich mit einem bekannten Fehlerprozess kontaminiert werden.

Wichtig ist es hier, zu erkennen, wie empfindlich die Ergebnisse der gängigen statistischen Verfahren gegenüber den gängigen Anonymisierungsverfahren sind, und dann auch gegebenenfalls geeignete Korrekturverfahren zu entwickeln.

2.2 Intervallwahrscheinlichkeit und verwandte Ansätze

Viele Anwender bezweifeln die Relevanz statistischer Aussagen mit dem Argument, dass die entsprechenden Analysen „überpräzise“ und damit der Komplexität des Gegenstands nicht angemessen sind. Dies gilt insbesondere, wenn eine große Zahl von letztendlich nur mathematisch begründeten Zusatzannahmen zu treffen sind, um mit einem komplizierten statistischen Verfahren bei komplexen Daten überhaupt zu einem eindeutigen Ergebnis zu kommen.

Seit kurzem gewinnt eine prinzipiell andere Vorgehensweise mehr und mehr Anhänger: Man löst sich von dem idealisierten Präzisionsanspruch klassischer statistischer Verfahren und versucht, dafür glaubwürdigere und zuverlässigere Aussagen zu gewinnen, indem man die Information durch eine *Menge* von passenden Modellen beschreibt. Dabei reflektiert die „Größe“ der Menge die Ungenauigkeit der Information. Entsprechende Mengen von Wahrscheinlichkeiten (oft als *Credal Sets* bezeichnet) führen auf sog. *Intervallwahrscheinlichkeiten* (engl. imprecise probabilities, mit der sog. Dempster-Shafer-Theorie als Spezialfall); bei Mengen von Parameterschätzungen spricht man von *partiell identifizierten* Schätzungen. Die entsprechenden Methoden werden sehr erfolgreich in der Entscheidungstheorie eingesetzt, wo sie helfen, bekannte Paradoxien bei Modellierung ökonomischer Entscheidungen und unsicheren Expertenwissens zu überwinden, und liefern auch einen formalen Überbau über die robuste Statistik.

Dempster-Shafer-Theorie bei gerundeten Daten / Heaping

Intervallwahrscheinlichkeit und Intervalldaten

Modellierung von Intervallbeobachtungen auf Basis verallgemeinerter Wahrscheinlichkeitsmodelle aus dem Bereich der Intervallwahrscheinlichkeit

Partielle Identifikation

- Literaturüberblick
- Vergleich von Manskis Modell der Partiellen Identifikation (Ökonometrie) mit dem verwandten Ansatz der systematischen Sensitivitätsanalyse (Molenberghs u.a.) (Biostatistik)
- Partielle Identifikation bei Fehlklassifikation
- Partielle Identifikation bei Messfehlern
- Partielle Identifikation in Schätzgleichungen, insbesondere bei Rundung / Heaping

Implementierung von Algorithmen für die Entscheidungstheorie mit Intervallwahrscheinlichkeit

Die üblichen Entscheidungskriterien in der Intervallwahrscheinlichkeit sind mit linearer Optimierung darstellbar; eine Implementierung in R und ein umfassender Vergleich der Methoden steht noch aus.

Dynamische (In)kohärenz bei sequentiellen Entscheidungen

Arbeitet man mit einem allgemeineren Wahrscheinlichkeitsbegriff in Entscheidungsproblemen, so gelten unter Umständen einige „Selbstverständlichkeiten“ der sequentiellen Informationsverarbeitung (Hauptsatz der Bayes-Entscheidungstheorie, Bellman-Prinzip der dynamischen Optimierung) nicht mehr, wodurch man sich in einem gewissen Sinn dynamisch inkohärent (widersprüchlich) verhält.

Ein Thema hierzu könnte darin bestehen, die diesbezügliche Diskussion in der Ökonomie und im Operations Research systematisch zusammenzufassen. Eher theoretische Arbeiten sollten diese Problematik für verschiedene verallgemeinerte Entscheidungskriterien untersuchen und die Konsequenzen für die statistische Datenanalyse, die ja auch als (sequentielles) Entscheidungsproblem formalisiert werden kann, untersuchen.

Hables Minimum-Distanz-Schätzer

Hable hat in seiner Dissertation (2008) an unserem Institut ein Minimum-Distanz-Verfahren entwickelt, mit dem parametrisierte Intervallwahrscheinlichkeitsmodelle geschätzt werden können. Die Verfahren sollen weiter untersucht werden und dann auf weitere Modelle ausgedehnt werden.

Ansätze zur Regression mit Intervallwahrscheinlichkeit

Es gibt verschiedene Vorgehensweisen, Regressionsmodelle auf Intervallwahrscheinlichkeiten zu verallgemeinern. Diese sollen implementiert, angewendet, verglichen und weiterentwickelt werden.

Credal Maximum Likelihood

Dieser Ansatz verspricht eine zuverlässigere Modellierung unbeobachteter Heterogenität. Die entsprechenden Methoden sollen implementiert, evaluiert und weiterentwickelt werden, um sie dann auch mit klassischen gemischten Modellen vergleichen zu können.

Untere / obere Schätzgleichungen

Die Theorie der unverzerrten Schätzgleichungen (Scorefunktionen) stellt einen formalen Rahmen zur Gewinnung konsistenter Schätzer bereit. Es soll eine konsequente Verallgemeinerung auf Intervallwahrscheinlichkeit untersucht und weiterentwickelt werden.

Intervallwahrscheinlichkeitsmodelle für kategoriale Daten

Das wohl bekannteste (und einfachste) Intervallwahrscheinlichkeitsmodell ist das sog. Imprecise-Dirichlet-Modell (IDM, Walley, 1996, JRSSB). Interessant wäre es, im Rahmen einer Literaturarbeit die mittlerweile sehr zahlreichen Anwendungen in verschiedensten Bereichen zu sichten und zu systematisieren.

Als weitere Themen bieten sich Vergleiche des IDMs mit einem alternativen Ansatz (Coolen & Augustin, 2009, IJAR) in verschiedenen Situationen an.

Graphische Modelle

Graphische Modelle (insbesondere Bayes-Netze und Markov-Ketten) sind ein wichtiges Forschungsgebiet im Bereich der Intervallwahrscheinlichkeit.

- **Lernen von Wahrscheinlichkeiten in graphischen Modellen**
Vergleich von verschiedenen Ansätzen
- **Algorithmen für graphische Modelle**
Verallgemeinerung der Algorithmen für Bayes-Netze
- **Klassische und verallgemeinerte Markov-Ketten**
Markov-Ketten bilden das einfachste dynamische Modell. Es sollen verschiedene Anwendungen, etwa im Marketing, der Soziologie oder den Ingenieurwissenschaften, gesichtet werden und untersucht werden, inwiefern sich die Analyse durch verallgemeinerte Modelle verbessern lassen.

Eine weitere Arbeit soll untersuchen, inwieweit sich die Verfahren zur Analyse des ifo-Konjunkturindex einsetzen lassen.

Lineare partielle Information

Isoliert von den Hauptsträngen der Theorie der Intervallwahrscheinlichkeit hat sich in der ökonomischen Entscheidungstheorie der Ansatz der linearen partiellen Information entwickelt. Eine Arbeit soll diese Entwicklung systematisch aufbereiten und die zugehörigen Anwendungen sichten.

Robuste Bayes-Verfahren

Das Generalized iLUCK-model (Walter & Augustin, 2009, JSTP) ist ein Modell für die bayesianische Datenanalyse, bei der anstatt mit einer einzelnen Priori-Verteilung mit Mengen von Priori-Verteilungen gearbeitet wird.

- einfache Datenbeispiele mit dem Generalized iLUCK-model, evtl. Programmieren (Erweiterung des luck-Package)
- Definition von prior-data conflict im Rahmen des Generalized iLUCK-model vergleichen mit der Definition von Evans & Moshonov
- Vergleich des Generalized iLUCK-model mit Aufsätzen von Whitcomb, Pericchi & Walley, insbesondere bezüglich der Reaktion auf prior-data conflict
- Behandlung von mehrdimensionalen Parametern im Generalized iLUCK-model, insbesondere die Definition von prior-data conflict im Mehrdimensionalen
- Das Generalized iLUCK-model kann auch für die Schätzung der Parameter einer linearen Regression verwendet werden. Themen in diesem Zusammenhang sind:
 - Datenbeispiele
 - Verallgemeinerung auf nicht-normalverteilte Zielgrößen über latente-Hilfsgröße-Methode, z.B. auf binäre Zielvariablen
- MCMC für robuste Bayes-Verfahren

Klassifikation und Intervallwahrscheinlichkeit

Es gibt verschiedene Ideen, Klassifikations- und Regressionsbäume und darauf aufbauende Methoden durch Betrachtung von Intervallwahrscheinlichkeitsmodellen zu stabilisieren und zu robustifizieren. Diese sollen implementiert, evaluiert, verglichen und weiterentwickelt werden.

2.3 Likelihood-Methoden

Viele der bekanntesten statistischen Verfahren basieren auf der Likelihood-Funktion (z.B. Maximum-Likelihood-Schätzer oder Likelihood-Ratio-Test) und die Bedeutung von Likelihood-Methoden in den Anwendungen nimmt ständig zu.

Robuste Likelihood-Methoden

Die Likelihood-Methoden sind i.A. nicht robust gegenüber kleinen Änderungen in den Verteilungsannahmen. Abhilfe können allgemeinere Modelle oder modifizierte Likelihood-Funktionen schaffen.

- **Robustifizierung der Likelihood-Funktion**
Vergleich von verschiedenen Ansätzen
- **Lokationsschätzung mit Penalisierung von Ausreißern**
Vergleich von verschiedenen robusten Schätzern

- **Regression mit Penalisierung von Ausreißern**
Entwicklung / Implementierung von Approximationsalgorithmen für ML-Regression

Fuzzy Daten

Unschärfe Daten können mithilfe von Likelihood-Funktionen repräsentiert werden. Somit können Likelihood-Methoden für deren Analyse verwendet werden.

- **Verwandtschaft zu Fehler-in-den-Variablen-Methoden**
Variable selbst ist präzise und nur unscharf beobachtet
- **Fuzzy-Modellierung intervallwertiger Beobachtungen**
Modellierung intervallwertiger Beobachtungen als Fuzzy-Daten
- **Likelihood-Methoden mit Fuzzy Daten**
Likelihood-Funktion und ML-Schätzer mit Fuzzy Daten
- **Fuzzifizierung als Robustifizierung**
Lohnt es sich manchmal Daten zu verunschärfen?

Likelihood-basierte Entscheidungstheorie

Entscheidungen können direkt auf Basis der Likelihood-Funktion getroffen werden. Obwohl die Likelihood-basierte Entscheidungstheorie bei den speziellen Entscheidungsproblemen der Statistik (Schätzungen und Tests) sehr erfolgreich ist, wurde die allgemeine Likelihood-basierte Entscheidungstheorie bisher nicht viel erforscht.

- **Entscheidung unter komplexer Unsicherheit**
Vergleich von verschiedenen Ansätzen
- **Diskrete Entscheidungsprobleme**
Entwicklung / Implementierung von Algorithmen für Likelihood-basierte Entscheidungen
- **Stetige Schätzprobleme**
Entwicklung / Implementierung von Algorithmen für verallgemeinerte ML-Schätzer

2.4 Statistische Verfahren für die Psychologie

Für alle Themen müssen vorhandene R-Funktionen angewendet und z.T. selbst erweitert, und Simulationsstudien in R durchgeführt werden!

Rasch Modell

Das Rasch Modell wird zur Testkonstruktion und Messung von Fähigkeiten, Einstellungen und Persönlichkeitsmerkmalen in der Psychologie verwendet; bekannt wurde es vor allem durch seine Anwendung in der PISA-Studie. Das Rasch Modell und seine Erweiterungen enthalten Parameter für die Eigenschaften der Personen und der Aufgaben, die mit verschiedenen Likelihood-basierten Ansätzen

geschätzt werden. Um sicherzustellen, dass ein Test die z.T. sehr strengen Anforderungen des Modells erfüllt, wurde eine Vielzahl von statistischen Modelltests entwickelt. Diese Modelltests sind nötig um z.B. auszuschließen, dass ein Mathematik-Test Probanden mit schlechteren Sprachkenntnissen benachteiligt.

- Vergleich von verschiedenen Likelihood-basierten Schätzverfahren
- Methoden zur Identifikation von Personen-Gruppen, die der Test “unterschiedlich behandelt” (anhand von Kovariablen und Mischverteilungsmodellen)

Bradley-Terry-Luce (BTL) Modell

Das BTL Modell wird verwendet, um aus Paarvergleichs-Daten die Präferenzen von Personen abzuleiten. Auch hier unterscheiden sich verschiedene Gruppen von Personen in den Parametern, die z.B. ihre Präferenzen für bestimmte Produkte beschreiben. Diese Verfahren werden in der Wahrnehmungsforschung, aber auch im Marketing, eingesetzt.

- Methoden zur Identifikation von Probanden-Gruppen, die der Test “unterschiedlich behandelt” (anhand von Kovariablen und Mischverteilungsmodellen)
- Ansätze zur Berücksichtigung von Eigenschaften der Produkte

Propensity Scores

In experimentellen Studien werden die Versuchspersonen zufällig den Versuchsbedingungen (z.B. Medikamenten- und Kontrollgruppe) zugeordnet. Dadurch unterscheiden sich die Gruppen nicht systematisch in ihren sonstigen Eigenschaften (z.B. Geschlecht, Alter, Ernährungsgewohnheiten), sondern nur in der Versuchsbedingung. Unterschiede zwischen den Gruppen (z.B. niedrigerer Blutdruck in der Medikamentengruppe) können damit kausal der Versuchsbedingung zugeschrieben werden.

In nicht-experimentellen Studien hingegen kann man nie ganz ausschließen, dass Unterschiede zwischen den Versuchsbedingungen durch eine dritte Variable hervorgerufen worden sind (z.B. nur jüngere Patienten mit guten Heilungschancen kriegen das teure Medikament verschrieben – also kein Wunder dass es dieser Gruppe besser geht als der Kontrollgruppe). Die Kontrolle durch Propensity Scores ist ein Versuch, den Einfluss von Drittvariablen zu kontrollieren, indem nur Personen mit der gleichen Wahrscheinlichkeit, z.B. in der Medikamentengruppe zu landen, verglichen werden.

- Simulationsstudien zur Illustration von Situationen, in denen Propensity Scores ungeeignet sind, um Kausalaussagen zu stützen
- Schätzung von Propensity Scores mit Random Forests, so dass z.B. auch nichtlineare Zusammenhänge zwischen Drittvariablen und Versuchsbedingung berücksichtigt werden

Evaluationsforschung

- Literaturrecherche und Vergleich von Evaluationskonzepten für Unternehmen und Hochschulen
- empirischer Vergleich von Frageformaten

2.5 Wirtschafts- und Sozialstatistik

Hedonische Preismessung

Konzentrations- und Armutsmessung

- Überblick über die Vielzahl unterschiedlicher Konzentrationsmaße
- Anwendung von verschiedenen Konzentrationsmaßen auf diverse Datensätze (z.B. anonymisierte Steuerdaten, beispielsweise zur Untersuchung der Verteilung von Einkommen, Vermögen, ...)
- Methodologische Probleme der Armutsmessung (Operationalisierung von Armut / Reichtum; ein kritischer statistischer Blick auf den Armuts- und Reichtumsbericht)
- Armutsmessung (in) für Entwicklungsländer(n)

Fortgeschrittene deskriptive Statistik im Sport

- Konzentrationsmessung, z.B. auf Fußballdaten anwenden
- Zeitreihen-Methoden
- Dynamische Rankings

Sozialindikatoren

Es gibt eine Vielzahl verschiedenster Sozialindikatoren; diese sollten aus statistischer Sicht aufbereitet, systematisiert und kritisch untersucht werden.

2.6 Sonstiges

Bibliometrie

Die Bibliometrie versucht, den „wissenschaftlichen Output“ von Forschern zu messen, insbesondere durch Analyse von Daten aus Publikations- und Zitationsdatenbanken (z.B. ISI Web of Knowledge).

- Systematische Reanalyse bibliometrischer Daten, z.B. nach Geschlecht, Wanderungsbewegungen, Netzwerken (Zitierkartelle)
- kritische Evaluation der Maßzahlen zur Messung des wissenschaftlichen Outputs (Output-Indikatoren, Impact-Maße für wissenschaftliche Zeitschriften, ...)

Die Geschichte des Instituts für Statistik in München

Unser Institut hat eine sehr interessante Geschichte. Sie soll, auch mit Unterstützung des Wissenschaftshistorikers Rudolf Seising (gegenwärtig Lehrstuhl für Geschichte der Naturwissenschaften), systematisch aufgearbeitet werden.