Michael Schomaker (Autor)

Selektieren und Kombinieren von Modellen unter Berücksichtigung der Problematik fehlender Daten

https://cuvillier.de/de/shop/publications/774

2. Modelle in Wissenschaft und Statistik

Das Ziel der Wissenschaft wird in der allgemeinen Wissenschaftstheorie mit Erkenntnisgewinn, dem Erwerb von neuem Wissen[1], gleichgesetzt. Insbesondere will die Wissenschaft nicht nur Tatsachen feststellen, sondern auch Ursachen und Erklärungen für diese finden. Die Suche allgemeiner Strukturen und Beziehungen beschränkt sich dabei nicht nur auf wahrnehmbare, sondern auch auf nicht-wahrnehmbare Gegenstandsbereiche. Die Systematisierung dieser Bereiche, also die Reduzierung ihrer Vielfältigkeit auf einige wenige elementare Faktoren, ist dabei von größtem Interesse. Diese Form der Abstraktion wird in der Wissenschaft unter anderem durch die Konstruktion von *Modellen* erreicht.

Seien G_1 und G_2 Gegenstände[2], S_1, S_2 Sätze[3] und bedeute $M(A, B)$, dass A Modell für B ist; dann lassen sich folgende Formen von Modellen unterscheiden (vgl. auch Detel (2007, Seite 94)):

(i)	$M(G_1, G_2)$	G_1 ist ein *strukturelles Modell* für G_2 ,
(ii)	$M(S_1, G_1)$	S_1 ist ein *abstraktes* bzw. *idealisiertes Modell* für G_1 ,
(iii)	$M(G_1, S_1)$	G_1 ist ein *semantisches Modell* für S_1 ,
(iv)	$M(S_1, S_2)$	S_1 ist ein *theoretisches Modell* für S_2 .

Im Falle struktureller Modelle sind also Gegenstände Modelle anderer Gegenstände; so zum Beispiel das maßstabsgetreue Modell einer Brücke oder das Doppelhelix-Modell der DNA. Bei abstrakten und idealisierten Modellen werden hingegen Sätze als Modelle für Gegenstände verwendet. Sie versuchen die Komplexität von Phänomenen einzuschränken, nur ihre Kernelemente zu erfassen und sie unter dem Ziel der Verständlichkeit

[1] Die traditionelle Definition nach Plato bezeichnet „Wissen“ als *wahre, gerechtfertigte Meinung*. Auch wenn diese Definition nicht unumstritten ist (vgl. Gettier (1963)), so ist sie im Kontext von Modellbildung und Modellselektion völlig ausreichend.

[2] In der Philosophie bezeichnet ein Gegenstand eine Sache oder eine Entität, die Eigenschaften besitzen und Beziehungen zu anderen Gegenständen haben kann.

[3] Ein Satz ist eine im Sinne der Logik widerspruchsfreie Aussage (z.B. über einen oder mehrere Gegenstände), die mittels eines Beweises, das heißt aus Axiomen und bereits vorhandenen Sätzen, hergeleitet werden kann.

vereinfachend darzustellen. Damit bilden sie das Rückgrat aller empirischen Wissenschaften. Ein Beispiel dafür ist etwa das Modell idealer Gase, mit dem sich unter gewöhnlichen Bedingungen viele thermodynamische Prozesse von Gasen verstehen und beschreiben lassen, das für tiefe Temperaturen und hohen Druck jedoch keine adäquate Modellierung mehr bietet; auch die vereinfachende Annahme unabhängiger und identisch verteilter Beobachtungen einer Stichprobe, um Verfahren der statistischen Inferenz besser verwenden zu können, kann als idealisiertes Modell verstanden werden; oder das Gesetz von Henry Darcy, das die Wasserströmung in porösen Flüssigkeiten modelliert und ursprünglich durch Versuche in einem Sandbett entstanden ist, heutzutage jedoch vor allem als spezielle Lösung der Navier-Stokes-Gleichungen motiviert werden kann. Bei semantischen Modellen sind es Gegenstände, die Sätze wahr machen; beispielsweise die rationalen Zahlen $\mathbb{Q}$ ohne Null, mit der Multiplikation als Verknüpfung und der Eins als neutralem Element, für die abelschen Gruppen im Bereich der Gruppentheorie der Mathematik. Sind Sätze Modelle für andere Sätze, so spricht man von theoretischen Modellen - insbesondere dann, wenn wissenschaftliche Theorien vorteilhaft für andere Theorien verwendet werden können. In diesem Sinne ist die klassische Mechanik ein theoretisches Modell für die Quantenmechanik.

Auch in der Statistik spielen Modelle eine herausragende Rolle. Das Sammeln, das Aufbereiten, die Analyse und die Interpretation von Daten eröffnet eine Fülle an Möglichkeiten statistischer Modellierung. Als weitgehend empirische Wissenschaft sind es insbesondere abstrakte und idealisierte Modelle, denen eine große Bedeutung zuzuschreiben ist. Aufgrund von Beobachtungen in der Form von Daten, sollen statistische Modelle Phänomene abstrahieren und beschreiben. Ein statistisches Modell in seiner allgemeinsten Form bezeichnet dabei eine parametrisierte Familie von Wahrscheinlichkeitsverteilungen

$$\mathcal{F} \;=\; \{f(y;\theta),\, \theta \in \Theta\}\,, \tag{2.1}$$

bei der die Beziehung einzelner Elemente eines Phänomens anhand einer Parametrisierung der Dichte von y durch θ in $\mathcal{F}$ beschrieben wird. Der Parameterraum Θ muss dabei nicht endlich-dimensional sein. Ein Beispiel ist die Menge aller Normalverteilungen

$$\mathcal{F} \;=\; \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left[\frac{y-\mu}{\sigma}\right]^2 \right) ; \theta = (\mu,\sigma) \in \mathbb{R}^p \times (0,\infty) \right\},$$

wobei μ den Erwartungswert und σ^2 die Varianz von y beschreibt. Häufig meint ein statistisches Modell auch die funktionale Beziehung zwischen einer Zielgröße y und potentiellen Einflussgrößen $X_1, \ldots, X_p$,

$$y = f(X_1, \ldots, X_p; \theta) + \epsilon\,, \tag{2.2}$$

wobei $f(\cdot)$ eine noch unbestimmte Funktion und ϵ eine Zufallskomponente bezeichnet. Ein typisches Beispiel ist das lineare Regressionsmodell

$$y \;=\; \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)\,,$$

bei dem die Zielgröße y über eine Linearkombination der $X_1, \ldots, X_p$ modelliert wird. In der Regel betrachtet man zur Modellierung dieser Abhängigkeitsbeziehung die bedingte, parametrisierte Dichte $f(y|X_1, \ldots, X_p; \theta)$, weswegen (2.2) als Spezialfall von (2.1) aufgefasst werden kann.

Ziel der statistischen Modellselektion ist es, aus einer Menge von Kandidatenmodellen $\mathcal{M} = \{M_1, \ldots, M_k\} \subset \mathcal{F}$ ein Modell M_{κ^*} auszuwählen, das die Daten – unter noch zu definierenden Gesichtspunkten – gut beschreibt. Dies betrifft insbesondere die Wahl geeigneter Regressoren bei Regressionsmodellen, die Anzahl von Kontrollpunkten in der Kontur- und Bildanalyse, die Ordnung autoregressiver Prozesse, die Anzahl von Faktoren in der Faktorenanalyse, die Wahl eines geeigneten Kerndichteschätzers und andere Problemstellungen der Statistik und verwandter Gebiete, vergleiche auch Linhart und Zucchini (1986) sowie Rao und Wu (2001).

Sparsamkeit

Um Phänomene beschreiben und verstehen zu können, sollte ein gewähltes Modell einerseits ein möglichst genaues Abbild der Realität liefern, andererseits nur die Komplexität in Anspruch nehmen, die nötig ist, um die wichtigsten Kausalitäten und Merkmale der Daten abzubilden. Da zudem insbesondere im statistischen Kontext die Varianz, etwa der Parameterschätzung von θ, mit der Komplexität steigt, der Bias in der Regel dagegen fällt, stellt sich die Frage nach einem geeigneten Kompromiss. Ist ein statistisches Modell zu komplex, so nennt man es überangepasst, ist es zu simpel, so nennt man es unterangepasst. In der Literatur wird häufig die Auffassung vertreten, dass bei gleicher Erklärungskraft das weniger komplexe Modell gewählt werden sollte. Dieses Prinzip ist

auch als *Prinzip der Sparsamkeit* bzw. als *Occam's Razor*[4] bekannt. Die Wahl eines solchen Modells ist jedoch keineswegs trivial, Burnham und Anderson (2002) bemerken hierzu:

> *„Parsimony lies between the evil of under- and overfitting"*

und Forster (1998) fragt:

> *„How much better must the complex model fit before we say that the extra parameter is necessary? Or, when should the better fit of complex models be 'explained away' as arising from the greater tendency of complex models to fit noise? How do we trade off fit with simplicity?"*

Um sich in der Statistik, auch unter Beachtung der Sparsamkeit, zwischen mehreren konkurrierenden Modellen für ein bestes entscheiden zu können, sind Verfahren und Kriterien notwendig. Statistische Modellselektion umfasst dabei häufig

(i) eine risikobasierte Entscheidung durch Optimierung eines Selektionskriteriums,

(ii) das sukzessive Testen von Hypothesen,

(iii) oder ein ad-hoc Vorgehen.

Punkt (i) beinhaltet insbesondere Selektionskriterien auf Basis von Vorhersagefehlern, im Rahmen der Informationstheorie und bayesianischer Natur. Ausgewählte Verfahren und Methoden zu diesen drei, wie auch einigen anderen Punkten sollen in den Abschnitten 3.1–3.5 vorgestellt und motiviert werden. Dies geschieht, wenn möglich, in allgemeinster Form; stets jedoch im Hinblick auf die Wahl geeigneter Regressoren in linearen und generalisierten linearen Regressionsmodellen.

[4] William Ockham (1285–1349) formulierte als erster ein Prinzip der Sparsamkeit (häufig wiedergegeben als „*entia non sunt multiplicanda praeter necessitatem*"), das weit über die Statistik hinaus in Biologie, Medizin und Philosophie bekannt ist. Heutzutage existieren im Detail viele verschiedene Fassungen und Versionen dieses Prinzips; prinzipiell lässt es sich jedoch so verstehen, dass für zwei wissenschaftliche Theorien bzw. Erklärungen unter festen Bedingungen diejenige zu bevorzugen ist, die einfacher ist. Die Rechtfertigung dieses Prinzips wird in der Wissenschaftstheorie diskutiert, vergleiche etwa Sober (1981) und Forster und Sober (1994). Die Begründungen sind dabei sowohl pragmatisch motiviert, wie etwa der oben angedeutete Punkt, dass Phänomene auf diese Weise besser verstanden werden können, als auch wissenschaftstheoretisch; so das Argument, dass das Ziel der Wissenschaft in der Approximation der Wahrheit besteht (vgl. auch die untenstehende Diskussion) und dieses Ziel nicht ohne die Berücksichtigung von Sparsamkeit erreicht werden kann. Kritische Stimmen sprechen dem Prinzip keine allgemeine Gültigkeit zu. So meint selbst Sober (2002): „*It may turn out, that simplicity has no global justification – that its justification varies from problem to problem*".

Die meisten der vorgestellten Methoden beinhalten dabei das Sparsamkeitsprinzip mehr oder weniger explizit. Es stellen sich in diesem Zusammenhang jedoch die grundlegende Fragen: Was ist Sparsamkeit? Wie lässt sich Sparsamkeit messen und konstruieren? Ist Sparsamkeit eindeutig? Diesen Fragen entspringt ein natürlicher Diskurs, ob Sparsamkeit Teil eines empirischen Vorgehens sein kann oder ob es ein künstliches, insofern rationalistisches, Konzept ist. Prinzipiell setzt Empirismus voraus, dass jede Form von Wissen über Erfahrung, beispielsweise über Daten, gewonnen wird, während unter einer rationalistischen Denkweise Erkenntnis in erster Linie unabhängig von unseren Sinneseindrücken entsteht. Deswegen wird in der Literatur teilweise argumentiert, dass das Sparsamkeitsprinzip ein extraempirisches Element ist, das primär in der Statistik, aus pragmatischen Gründen, herangezogen wird und somit als rationalistisch angesehen werden muss; so schreiben Forster und Sober (1994):

> *„Giving weight to simplicity thus seems to embody a kind of rationalism“* .

Die folgenden Kapitel werden zeigen, dass dieses Argument nicht korrekt ist. Obgleich Bestandteil fast jedes statistischen Verfahrens bzw. Kriteriums, kann die Interpretation der Aufsplittung in einen Anpassungs- und einen Sparsamkeitsterm als Kompromiss zwischen Bias und Varianz in der Regel nur a posteriori erfolgen; a priori liegt den Methoden meist ein grundlegend anderes Prinzip zugrunde, etwa die Approximation von Wahrheit, die Verringerung von Vorhersagefehlern oder die Maximierung von posteriori-Wahrscheinlichkeiten – die Sparsamkeitsterme entstehen dabei gewissermaßen als „Abfallprodukt“ bei der Evaluierung der eigentlichen Zielsetzung. Es zeigt sich daher, dass die Konzepte statistischer Modellselektion fast ausschließlich datengestützt und empirisch motiviert sind und keine rationalistische Rechtfertigung benötigen.

Eine Ausnahme bildet die von Jorma Rissanen begründete Theorie der Minimum Description Length. Hierbei bildet der Kompromiss zwischen Anpassung und Sparsamkeit das Fundament aller von Rissanen (1978) erarbeiteten Verfahren. Konzepte aus der Informationstheorie helfen dabei, die Länge wissenschaftlicher Theorien, im Speziellen statistischer Modelle, zu beschreiben und dadurch Modelle zu wählen, die Wissen am besten „verschlüsseln“ können und dennoch anschließend dieses Wissen am besten zu reproduzieren vermögen. Dadurch wird nicht nur das Prinzip der Sparsamkeit direkt bei der Konstruktion von Modellwahlkriterien verwendet, sondern auch erstmals eine konkrete Ausarbeitung davon präsentiert, wie Sparsamkeit zu messen ist. Eine überschaubare Einführung bietet Abschnitt 3.5.1; die dort angegebenen Referenzen erlauben darüber hinaus einen tieferen Einblick in die Thematik, die in den folgenden Kapiteln keine zentrale Rolle einnehmen wird.

Wahrheit

Ein Vergleich und eine Beurteilung der einzelnen Verfahren, insbesondere der Selektionskriterien, erfolgt meist durch die Betrachtung asymptotischer Güteeigenschaften, wie der Konsistenz und der Effizienz. Abschnitt 3.6 beschäftigt sich hiermit ausführlich. Es zeigt sich, dass eine der entscheidenden Voraussetzungen zur Optimalität eines Modellwahlkriteriums in der Tatsache liegt, ob in der Menge der Kandidatenmodelle $\mathcal{M} = \{M_1, \ldots, M_k\}$ das wahre, datengenerierende Modell M_κ^* enthalten ist oder nicht. Es zeigt sich ferner, dass das Konzept eines wahren Modells als solches problematisch ist. Obgleich Voraussetzung in der Konstruktion vieler populärer Modellwahlkriterien, ist seine Existenz nicht unumstritten. Burnham und Anderson (2002) bemerken

> *„The words 'true model' represent an oxymoron"*

und de Leuuw (1988) meint in diesem Zusammenhang lapidar

> *„Truth is elusive"*

Die Diskussion dieses Aspekts verlagert sich in der statistischen Literatur jedoch meist in Richtung der nahezu gleichwertigen Frage der Dimensionalität eines wahren Modells. Ist es von unendlicher Dimension, so ist es implizit nicht in der Menge der Kandidatenmodelle enthalten; ist es von endlicher Dimension, so kann es durchaus Bestandteil dieser Menge sein. Abschnitt 3.6 präsentiert wichtige Resultate zur Optimalität von Modellwahlkriterien und diskutiert ihre Nützlichkeit anhand ausgewählter Aspekte.

In gewisser Weise ist diese Diskussion auch Bestandteil einer alten Realismus-Debatte: Die Annahme einer denkunabhängigen Realität, einer Realität die sich erkennen und erfassen lässt und damit auch letztlich zu Wissen unabhängig von menschlichen Theorien und Konventionen führt, kann als Position für einen erkenntnistheoretischen, wissenschaftlichen Realismus verstanden werden.[5] Eine solche Sichtweise impliziert, dass eine Wirklichkeit, eine Wahrheit existiert und wir diese erfahren und beschreiben können und dass die Annäherung an diese Wahrheit insofern auch Ziel der Wissenschaft ist. Im Gegensatz dazu existieren viele nicht-realistische Positionen, die sich in philosophischen Denkweisen, wie etwa dem Relativismus, dem Instrumentalismus oder dem konstruktiven Empirismus äußern. Letzterer geht auf van Fraaasen (1980) zurück und verneint das

[5] Tatsächlich umfasst der Begriff des Realismus eine Vielzahl philosophischer Positionen, die sich auf unterschiedliche Gegenstandsbereiche beziehen. Diese sind im Kontext dieser Arbeit jedoch nicht weiter relevant; der Kerngedanke und das Stichwort einer „denkunabhängigen Realität" genügt für die folgende Diskussion.

Ziel der Wissenschaft als Approximation von Wahrheit. Die Doktrin eines konstruktiven Empirismus sieht vor, die Wissenschaft als reinen Beobachter zu betrachten, der wahre Aussagen über beobachtbare Phänomene und Experimente machen kann, ausdrücklich aber nicht über unbeobachtbare Phänomene und damit über eine den Beobachtungen zugrundeliegende Wahrheit. Es ist fragwürdig, ob solche oder andere (beispielsweise relativistische) Standpunkt hilfreich sind. Sober (2000) quittiert die Diskussion mit den Worten:

> *„Realism says that the goal of science is to discover which theories are true; [constructive] empiricism maintains that the goal is to discover theories that are empirically adequate [...] In both cases, truth is the property that matters“*

Tatsächlich ist die Realismus-Debatte im Kontext statistischer Modellselektion nicht entscheidend. Auch wenn aufgrund ihrer Konzeption viele der in Kapitel 3 vorgestellten Methoden eine zugrundeliegende Wahrheit, zumindest in Form eines datengenerierenden Prozesses, benötigen und damit auch eine prinzipiell realistische Sichtweise angenommen wird, so steht vor allem die Identifizierung relevanter Effekte eines Phänomens im Vordergrund. Ob Aussagen über eine Wahrheit und ihre Existenz getroffen werden müssen, ist fraglich. Dies macht auch Abschnitt 3.6 deutlich: Ob die oben erwähnten Qualitätsmerkmale von Konsistenz und Effizienz im Kontext statistischer Modellwahl sinnvoll sind, wird dort diskutiert.

Grenzen des Wissens

Die Suche geeigneter Modelle zur Charakterisierung von Phänomenen unterliegt häufig gewissen Beschränkungen, insbesondere solchen, die sich aus den Grenzen empirischen Wissens ergeben. Dies betrifft vor allem die mangelnde Verfügbarkeit und das Fehlen von Daten: Die Herausforderung, Modelle zu bilden, zu wählen und zu schätzen, auch unter Beachtung der oben diskutierten Sachverhalte der Sparsamkeit und Optimalität, erfährt im Kontext fehlender Werte eine zusätzliche Dimension. Die Diskussion und Evaluierung von statistischen Methoden zur Modellselektion unter Berücksichtigung dieser Problematik ist Schwerpunkt dieser Arbeit und wird weitgehend in den Kapiteln 5-7 erörtert. Zusätzlich berücksichtigt werden dabei auch die Grenzen statistischer Modellselektionsverfahren. Die Unsicherheit bezüglich der Wahl eines geeigneten Modells ist ein weiterer Faktor, dem besondere Beachtung geschenkt wird. Die Kombination verschiedener Konzepte aus verschiedenen Teilgebieten der Statistik sollen helfen, ein relativ allgemein angelegtes Sammelsurium an Methoden zur Bewältigung dieser Probleme zu liefern. Die Illustration dieser Methoden beschränkt sich aus Gründen der Übersichtlichkeit dabei auf

lineare und logistische Regressionsmodelle, wie auch exemplarisch auf die Faktorenanalyse. Wie aus der obigen Diskussion bereits zu erkennen, wird dabei stets eine zu einem gewissem Maße realistische Grundhaltung angenommen, da stets eine Annäherung an eine zugrundeliegende Wahrheit, zumindest in Form eines datengenerierenden Prozesses, vorausgesetzt wird. Alle vorgestellten Methoden, mit Ausnahme der MDL-Methodik, sind dabei rein empirisch zu motivieren und bedürfen keiner zusätzlich rationalistischen Rechtfertigung.