Wolfgang K. Seiler

Algebraische Statistik

Vorlesung im Frühjahrssemester 2024
an der Universität Mannheim

Dieses Skriptum entstand parallel zur Vorlesung und sollte mit möglichst geringer Verzögerung erscheinen. Dieses Jahr wird es parallel zur Vorlesung korrigiert, sofern mir Fehler oder Verbesserungsmöglichkeiten auffallen. Trotzdem ist die Qualität idieses Skriptums auf keinen Fall mit der eines. Lehrbuch zu vergleichen; insbesondere sind Fehler bei dieser Entstehensweise nicht nur möglich, sondern **sicher**. Dabei handelt es sich wohl leider nicht immer nur um harmlose Tippfehler, sondern auch um Fehler bei den mathematischen Aussagen. Da mehrere Teile aus anderen Skripten für Hörerkreise der verschiedensten Niveaus übernommen sind, ist die Präsentation auch teilweise ziemlich inhomogen.

Das Skriptum sollte daher mit Sorgfalt und einem gewissen Mißtrauen gegen seinen Inhalt gelesen werden. Falls Sie Fehler finden, teilen Sie mir dies bitte persönlich oder per e-mail (seiler@math.uni-mannheim.de) mit. Auch wenn Sie Teile des Skriptums unverständlich finden, bin ich für entsprechende Hinweise dankbar. In der online Version werde ich alle bekannten Fehler korrigieren.

Kapitel 0
Einführung

Verglichen mit den klassischen Teilgebieten der Mathematik ist die Algebraische Statistik sehr jung: Die ersten Arbeiten dazu erschienen erst in den letzten Jahren des vorigen Jahrhunderts, die meisten erst in diesem. Von daher ist auch noch nicht so ganz klar, womit sich die Algebraische Statistik alles beschäftigt: Grundsätzlich geht es um die Anwendung algebraischer Methoden auf statistische Fragestellungen, aber man ist noch weit davon entfernt genau zu wissen, welche statistischen Fragestellungen einer algebraischen Behandlung zugänglich sind und welche nicht. In dieser Vorlesung wird es um zwei Themengebiete gehen, über die in den vergangenen zwei Jahrzehnten erfolgreich gearbeitet wurde, die aber bei weitem nicht das gesamte Gebiet der Algebraischen Statistik umfassen.

Als erstes geht es darum, zu einer vorgegebenen (oder gezielt gewählten) Stichprobe statistische Modelle zu finden, deren Parameter auf Grund der vorliegenden Daten bestimmt werden können.

Das zweite Thema sind Kontingenztafeln, wie sie etwa bei Vierfeldertests auftreten. Unter geeigneten Bedingungen wie hinreichender Größe der Stichprobe und Annahmen über die zu Grunde liegende Verteilungsfunktion kann man hier Hypothesen durch χ^2-Tests überprüfen. Oft sind aber wegen eines zu hohen Aufwands oder zu hoher Kosten nur kleine Stichproben möglich, und über die Verteilungsfunktionen weiß man auch nicht immer Bescheid. Schon lange, bevor irgend jemand von Algebraischer Statistik redete, gab es auch dazu bereits alternative Verfahren wie beispielsweise FISHERs exakten Test, jedoch sind diese sehr aufwendig. Mit den sogenannten MARKOV-Basen liefert

die Algebraische Statistik eine weitere Methode, die iterativ und mit geringerem Aufwand als exakte Verfahren in solchen Fällen deutlich zuverlässigere Ergebnisse liefert als χ^2-Tests.

§1: Statistische Modelle

Bei manchen Fragestellungen ist auf Grund von Naturgesetzen bekannt, welche Form der Zusammenhang zwischen zwei oder mehreren Größen hat; unbekannt sind nur die Parameter. Mißt man beispielsweise für einen festen Leiter Spannung und Stromstärke, so ist nach dem OHMschen Gesetz die Spannung U stets gleich dem Widerstand R des Leiters mal der Stromstärke I. Gesucht ist der Wert des Widerstands R.

Bei wirtschafts- und sozialwissenschaftlichen Fragestellungen ist oft kein Gesetz bekannt; hier muß man versuchen, eine möglichst einfache Funktion zu finden, die die beobachteten Daten möglichst gut beschreibt.

Allgemein betrachten wir als Modelle Funktionen

$$f\colon \Theta \times \mathbb{R}^n \to \mathbb{R}^m ,$$

die jedem n-tupel $x = (x_1, \ldots, x_n)$ von Eingangsgrößen ein m-tupel $y = (y_1, \ldots, y_m)$ von Ausgangsgrößen zuordnen in Abhängigkeit von r Parametern $(\theta_1, \ldots, \theta_r) \in \Theta \subseteq \mathbb{R}^r$. Wir beschränken uns hier auf den häufigsten Fall $m = 1$ einer einzigen Ausgangsgröße y. Außerdem betrachten wir nur Modelle, die in den *Parametern* $\theta_1, \ldots, \theta_r$ linear sind. In den Eingangsgrößen $x_1, \ldots, x_n$ sollen die Modelle in der Algebraischen Statistik polynomial sein; wir lassen also keine transzendenten Funktionen zu. (Tatsächlich kann man mit den Methoden der Algebraischen Statistik auch gewisse transzendente Funktionen wie Exponentialfunktionen und trigonometrische Funktionen behandeln, indem man Funktionen der Art $e^{\lambda x}$ als Variablen betrachtet.)

Zur Klärung der Begriffe seien zunächst einige grundlegende Definitionen aus der Algebra kurz wiederholt:

Definition: *a)* Ein Ring ist eine Menge R zusammen mit zwei Rechenoperationen „+“ und „·“ von $R \times R$ nach R, für die gilt:

1.) R ist bezüglich „+“ eine abelsche Gruppe, d.h. für die Addition gilt das Kommutativgesetz $f + g = g + f$ sowie das Assoziativgesetz

$(f + g) + h = f + (g + h)$ für alle $f, g, h \in R$, es gibt ein Element $0 \in R$, so daß $0 + f = f + 0 = f$ für alle $f \in R$, und zu jedem $f \in R$ gibt es ein Element $-f \in R$, so daß $f + (-f) = 0$ ist.

2.) Die Verknüpfung „$\cdot$“: $R \times R \to R$ erfüllt das Assoziativgesetz $f(gh) = (fg)h$, und es gibt ein Element $1 \in R$, so daß $1f = f1 = f$.

3.) „+“ und „$\cdot$“ erfüllen die Distributivgesetze $f(g + h) = fg + fh$ und $(f + g)h = fh + gh$.

b) Ein Ring heißt *kommutativ,* falls zusätzlich noch das Kommutativgesetz $fg = gf$ der Multiplikation gilt.
c) Ein Ring heißt *nullteilerfrei* wenn gilt: Falls ein Produkt $fg = 0$ verschwindet, muß mindestens einer der beiden Faktoren f, g gleich Null sein. Ein nullteilerfreier kommutativer Ring heißt *Integritätsbereich.*

Natürlich ist jeder Körper ein Ring; für einen Körper werden schließlich genau dieselben Eigenschaften gefordert und zusätzlich auch noch die Kommutativität der Multiplikation sowie die Existenz multiplikativer Inverser. Ein Körper ist somit insbesondere auch ein Integritätsbereich.

Das bekannteste Beispiel eines Rings, der kein Körper ist, sind die ganzen Zahlen; auch sie bilden einen Integritätsbereich.

Für die Betrachtung nichtlinearer Gleichungssysteme interessieren uns allerdings vor allem Polynomringe. Da auch diese kommutativ sind, vereinbaren wir:

Wenn nicht explizit etwas anderes gesagt wird, soll *Ring* in dieser Vorlesung stets für einen *kommutativen Ring* stehen.

Definition: R sei ein Ring, und $X_1, \ldots, X_n$ seien n Symbole, die nicht in R liegen.
a) Ein *Monom* ist ein Produkt $X_1^{\alpha_1} \cdots X_n^{\alpha_n}$ mit nichtnegativen ganzen Zahlen $\alpha_1, \ldots, \alpha_n$. Die Summe der α_i bezeichnen wir als den *Grad* des Monoms. Oft schreiben wir kurz X^α mit $\alpha = (\alpha_1, \ldots, \alpha_n)$.
b) Ein *Polynom* über R in den Variablen $X_1, \ldots, X_n$ ist eine endliche Linearkombination f von Monomen mit Koeffizienten aus R. Falls diese nicht Null ist, bezeichnen wir den größten Grad eines in f vorkommenden Monoms als den *Grad* $\deg f$ von f. Für das Polynom $f = 0$

definieren wir keinen Grad.
c) Die Menge aller Polynome über R in den Variablen $X_1, \ldots, X_n$ bezeichnen wir als den *Polynomring* $R[X_1, \ldots, X_n]$ über R in den Variablen $X_1, \ldots, X_n$.

Es ist klar, daß $R[X_1, \ldots, X_n]$ mit der offensichtlichen Addition und Multiplikation ein Ring ist. Wir nehmen dabei natürlich an, daß die X_i untereinander kommutieren.

Wir interessieren uns vor allem für Polynomringe über Körpern; für Induktionsbeweise ist es aber oft nützlich, beispielsweise den Polynomring $\mathbb{Q}[X, Y]$ aufzufassen als den Polynomring in Y über dem Ring $R = \mathbb{Q}[X]$; daher die allgemeinere Definition.

In der Statistik arbeitet man meist über dem Körper $\mathbb{R}$ der reellen Zahlen. In der Algebraischen Statistik wollen wir allerdings (soweit möglich) exakt algebraisch rechnen, und da gibt es im Falle von $\mathbb{R}$ Probleme: Da dieser Körper überabzählbar ist, können wir nur die Elemente einer Teilmenge vom Maß Null überhaupt durch endliche Formelausdrücke oder sonstige Beschreibungen angeben, und selbst für diese hat D. RICHARDSON 1968 gezeigt, daß es keinen Algorithmus geben kann, der entscheidet, ob zwei solche Beschreibungen dieselbe Zahl definieren oder nicht – selbst wenn man für die Beschreibungen nur sehr einfache Formelausdrücke zuläßt.

Wenn wir es mit nur endlich vielen Daten zu tun haben (und das ist beim konkreten Rechnen natürlich immer der Fall), gibt es allerdings stets einen abzählbaren Teilkörper von $\mathbb{R}$, der alle diese Zahlen enthält; oft wird uns sogar der Körper $\mathbb{Q}$ der rationalen Zahlen genügen. Wir legen uns daher nicht auf einen Körper fest, sondern arbeiten über einem beliebigen Körper k, der in dieser Vorlesung freilich immer ein relativ „kleiner“ Teilkörper von $\mathbb{R}$ oder (in ganz seltenen Fällen) eventuell auch $\mathbb{C}$ sein wird.

Modelle, die in den Eingangsgrößen polynomial sind, können aufgefaßt werden als Polynome, deren Koeffizienten Funktionen der Paramter θ_ν sind. Für jedes unserer Modelle gibt es also eine endliche Menge M von Monomen $M_\nu = X^{\alpha^{(\nu)}} \in k[X_1, \ldots, X_n]$, $\nu = 1, \ldots, s$, wobei

$\alpha^{(\nu)} = \left(\alpha_1^{(\nu)}, \ldots, \alpha_n^{(\nu)}\right)$ in $\mathbb{N}_0^n$ liegt, und für jedes ν haben wir eine Funktion $a_\nu \colon \Theta \to k$ derart, daß

$$f(\theta_1, \ldots, \theta_r, x_1, \ldots, x_n) = \sum_{\nu=1}^{s} a_\nu(\theta_1, \ldots, \theta_r) x_1^{\alpha_1^{(\nu)}} \cdots x_n^{\alpha_n^{(\nu)}}$$

ist. Wir wollen nur Modelle betrachten, die in $\theta_1, \ldots, \theta_r$ linear sind, und das auch noch auf die einfachst mögliche Weise: Für uns ist $r = s$ und $\Theta = \mathbb{R}^r$, also

$$f(\theta_1, \ldots, \theta_r, x_1, \ldots, x_n) = \sum_{\nu=1}^{r} \theta_\nu x_1^{\alpha_1^{(\nu)}} \cdots x_n^{\alpha_n^{(\nu)}} .$$

Tatsächlich machen wir noch eine weitere Einschränkung: Um möglichst kleine Monome zu bekommen, verlangen wir, daß die Menge der auf der rechten Seite auftretenden Monome ein *Ordnungsideal* ist im Sinne der folgenden

Definition: Eine endliche Menge von Monomen heißt *Ordnungsideal,* falls sie mit jedem Element auch dessen sämtliche Teiler enthält.

Liegt also das Monom X^2Y in einem Ordnungsideal, so muß dieses auch die Monome $1, X, Y, X^2$ und XY enthalten.

Die Bezeichnung *Ordnungsideal* ist leider etwas unglücklich, denn Ordnungsideale habe mit den Idealen in Ringen, die wir in Kürze betrachten werden, nicht das Geringste zu tun. Da sich die Bezeichnung aber eingebürgert hat, müssen wir damit leben.

Unter einer *Stichprobe* verstehen wir eine endliche Teilmenge $\mathcal{S}$ von k^{n+1} mit Elementen der Form $(x_1, \ldots, x_n, y)$, wobei y die beobachtete Ausgangsgröße zu den Eingangsgrößen $x_1, \ldots, x_n$ ist. Die dazu gehörige Menge der n-tupel $(x_1, \ldots, x_n) \in \mathbb{R}^n$ bezeichnen wir als ein *Design,* da diese Teilmenge oft *a priori* so gewählt wird, daß die zugehörigen y-Werte möglichst viel Information liefern. Es gibt verschiedene Methoden solche Designs zu konstruieren, darunter insbesondere auch viele algebraische, allerdings wird die Designtheorie (oder Theorie der optimalen Versuchsplanung) üblicherweise nicht zur Algebraischen Statistik gezählt, und sie ist auch deutlich älter als diese. Sie wird in dieser Vorlesung keine Rolle spielen.

Ausgehend von einem Design $D \subset \mathbb{R}^n$ können wir eine Stichprobe auch auffassen als eine Funktion $D \to \mathbb{R}$, die als Funktion auf einer endlichen Menge einfach durch die Tabelle der Funktionswerte gegeben ist und keine speziellen mathematischen Eigenschaften haben muß.

Wenn wir eine Stichprobe $\mathcal{S} \subset \mathbb{R}^{n+1}$ und ein Modell $f\colon \Theta \times \mathbb{R}^n \to \mathbb{R}$ haben, wollen wir die Parameter $(\theta_1, \ldots, \theta_r) \in \Theta$ so bestimmen, daß im Idealfall gilt

$$f(\theta_1, \ldots, \theta_r, x_1, \ldots, x_n) = y \quad \text{für alle } (x_1, \ldots, x_n, y) \in \mathcal{S}\,.$$

Das wird im allgemeinen nicht möglich sein, denn selbst wenn der Zusammenhang zwischen Eingangs- und Ausgangsgrößen wie beim OHMschen Gesetz feststeht, wird es auf Grund von Meßfehlern und Zufallsschwankungen (etwa der Temperaturabhängigkeit des Widerstands) fast nie ein $\theta \in \Theta$ geben, für das diese Gleichung für jedes Element der Stichprobe gilt. Wir müssen uns daher begnügen, ein θ zu finden, so daß

$$f(\theta_1, \ldots, \theta_r, x_1, \ldots, x_n) \approx y \quad \text{für alle } (x_1, \ldots, x_n, y) \in \mathcal{S}\,.$$

Der seit GAUSS übliche Ansatz zur Definition der ungefähren Gleichheit in diesem Zusammenhang ist die *Methode der kleinsten Quadrate:* Wir suchen ein $\theta \in \Theta$ derart, daß die Summe der quadratischen Abweichungen zwischen linker und rechter Seite minimal wird.

§2: Bestimmung der Parameter durch ein lineares Gleichungssystem

Wir interessieren uns in dieser Vorlesung nur für Modelle, die linear in den Variablen $\theta_1, \ldots, \theta_r$ sind. Für jedes Element $(x_1^{(j)}, \ldots, x_n^{(j)}, y^{(j)})$ der Stichprobe $\mathcal{S}$ gibt es daher Elemente $a_{j1}, \ldots, a_{jr}$ aus einem geeigneten Körper $k \subset \mathbb{R}$ derart, daß

$$f(\theta_1, \ldots, \theta_r, x_1^{(j)}, \ldots, x_n^{(j)}) = \sum_{\nu=1}^{r} a_{j\nu}\theta_\nu$$

ist. Bezeichnet θ den Spaltenvektor mit Komponenten $\theta_1, \ldots, \theta_r$ und y den mit Komponenten $y_1, \ldots, y_s$, sollte mit der Matrix $A = (a_{j\nu})$ aus $k^{s\times r}$ also gelten

$$A\theta = y\,.$$

Tatsächlich wird dieses lineare Gleichungssystem für θ meist überbestimmt und unlösbar sein. Nach GAUSS suchen wir stattdessen nach einem Vektor θ derart, daß der EUKLIDische Abstand zwischen den beiden Vektoren $A\theta$ und y minimal wird.

Falls das Gleichungssystem lösbar ist, ist dieser Abstand für jeden Lösungsvektor gleich Null, und wir sind fertig – kürzer kann kein Abstand sein.

Andernfalls ist für jeden Vektor θ, und damit auch für den, den wir suchen, das Produkt $A\theta$ von y verschieden; für ein optimales θ sei es etwa gleich $\bar{y}$. Dann ist $\bar{y}$ ein Vektor, der sich in der Form $A\theta$ darstellen läßt, und unter allen solchen Vektoren ist er derjenige, für den die Länge des Differenzvektors $\bar{y} - y$ minimal ist.

Die Matrix $A \in k^{s\times r}$ definiert eine lineare Abbildung

$$\varphi\colon k^r \to k^s; \quad \theta \mapsto A\theta\,;$$

deren Bildraum sei U. Falls die rechte Seite y in U liegt, ist das Gleichungssystem lösbar; andernfalls suchen wir einen Vektor $\theta \in k^r$, für den die Länge des Vektors $A\theta - y$ minimal wird. Da die Vektoren, die sich in der Form $A\theta$ darstellen lassen, genau die Vektoren aus U sind, ist somit $A\theta = \pi_U(y)$ die orthogonale Projektion von y nach U. Diese könnten wir *im Prinzip* bestimmen, indem wir die QR-Zerlegung von A berechnen, denn dann bilden die ersten Spalten von Q eine Orthogonalasis von U, die durch die weiteren Spalten zu einer Orthogonalbasis von ganz k^s ergänzt wird. Sobald wir $\pi_U(y)$ kennen, können wir das lineare Gleichungssystem $A\theta = \pi_U(y)$ lösen und somit θ bestimmen.

Wir wollen uns überlegen, wie wir θ auch ohne die rechnerisch aufwendige QR-Zerlegung bestimmen können.

Für den gesuchten Vektor θ (oder für die gesuchten Vektoren θ) ist $A\theta = \varphi_U(y)$. Da $A\theta$ bereits in U liegt, ist $\pi_U(A\theta) = A\theta$, also ist die Gleichung $A\theta = \pi_U(y)$ äquivalent zu

$$\pi_U(A\theta) = \pi_U(y) \quad \text{oder} \quad A\theta - y \in \operatorname{Kern} \pi_U = U^\perp\,.$$

Das orthogonale Komplement $U^\perp$ von U besteht aus allen Vektoren $y \in k^n$, die senkrecht stehen auf U, für die also gilt

$$\langle A\theta, y\rangle = 0 \quad \text{für alle } \theta \in k^m\,.$$

Wie aus der Linearen Algebra bekannt, gilt für Skalarprodukte die Gleichung

$$\langle A\theta, y\rangle = \left\langle \theta, A^T y\right\rangle ,$$

wobei A^T die zu A transponierte Matrix bezeichnet. (Falls wir über einem nichtreellen Teilkörper der komplexen Zahlen arbeiten, gilt Entsprechendes, wenn wir an Stelle der transponierten Matrix deren komplex konjugierte Matrix nehmen, also die adjungierte Matrix $A^* = \overline{A^T}$.) y liegt also genau dann in $U^\perp$, wenn $A^T y$ senkrecht steht auf allen Vektoren $\theta \in k^r$. Ein solcher Vektor aus k^r ist insbesondere $A^T y$ selbst; wegen der positiven Definitheit des Skalarprodukts ist also $A^T y = 0$. Ist umgekehrt $A^T y = 0$, verschwindet $\left\langle \theta, A^T y\right\rangle$ für alle $\theta \in k^r$, also ist

$$U^\perp = \left\{ y \in k^s \;\middle|\; A^T y = 0 \right\} .$$

$A\theta - y$ liegt daher genau dann im Kern von π_U, wenn $A^T(A\theta - y)$ verschwindet oder, anders ausgedrückt, wenn θ eine Lösung des linearen Gleichungssystems

$$(A^T A)\,\theta = A^T y$$

ist. Dieses Gleichungssystem läßt sich schnell aufstellen und dann, falls es Lösungen gibt, nach GAUSS lösen.

Als Beispiel betrachten wir den einfachsten Fall der linearen Regression, die Bestimmung einer Ausgleichsgeraden. Hier erwarten wir einen linearen Zusammenhangs $\theta_1 x + \theta_2 = y$ zu N Wertepaaren $(x_i, y_i) \in \mathbb{R}^2$, wobei N sinnvollerweise größer als zwei sein sollte. Wir haben dann N Gleichungen

$$\theta_1 x_i + \theta_2 = y_i$$

mit unbekannten Parametern θ_1 und θ_2 und bekannten Zahlen x_i und y_i. Wir haben also ein lineares Geichungssystem mit N Gleichungen in den beiden Variablen θ_1 und θ_2.

Fassen wir die Werte x_i zusammen zu einem Vektor $x \in \mathbb{R}^N$ und die y_i zu einem Vektor $y \in \mathbb{R}^N$, so läßt sich dieses Gleichungssystem kurz schreiben als

$$\theta_1 x + \theta_2 1 = y ,$$

wobei $1 \in \mathbb{R}^N$ jenen Vektor bezeichnet, dessen sämtliche Komponenten gleich eins sind.

Die Matrix des Gleichungssystems ist somit die $N \times 2$-Matrix A mit Spalten x und 1 und A^T ist die $2 \times N$-Matrix, in deren erster Zeile die x_i stehen, während in der zweiten lauter Einser stehen. Somit ist

$$A^T A = \begin{pmatrix} \langle x, x \rangle & \langle x, 1 \rangle \\ \langle x, 1 \rangle & \langle 1, 1 \rangle \end{pmatrix} \quad \text{und} \quad A^T y = \begin{pmatrix} \langle x, y \rangle \\ \langle 1, y \rangle \end{pmatrix} .$$

Das Gleichungssystem wird also zu

$$\langle x, x \rangle \, \theta_1 + \langle x, 1 \rangle \, \theta_2 = \langle x, y \rangle \quad \text{und} \quad \langle x, 1 \rangle \, \theta_1 + N\theta_2 = \langle 1, y \rangle \ .$$

Seine Matrix ist genau dann singulär, wenn ihre Determinante verschwindet, wenn also $N \cdot \langle x, x \rangle = \langle x, 1 \rangle^2$ ist. Nach der CAUCHY-SCHWARZschen Ungleichung ist

$$|\langle 1, x \rangle| \leq |1| \cdot |x| = \sqrt{N} \, |x| \, , \quad \text{also} \quad |\langle 1, x \rangle|^2 \leq N \cdot \langle x, x \rangle \ .$$

Ausgeschrieben ist das die Ungleichung

$$\left(\sum_{i=1}^{N} x_i \right)^2 \leq N \sum_{i=1}^{N} x_i^2 \, ,$$

und bekanntlich wird die CAUCHY-SCHWAZsche Ungleichung genau dann zur Gleichung, wenn die beiden Vektoren linear abhängig sind. Im Falle von x und 1 ist das genau dann der Fall, wenn alle x_i denselben Wert haben, was in praktischen Anwendungen fast nie der Fall sein dürfte. In diesem Fall ist die erste Gleichung ein Vielfaches der zweiten, und es gibt unendlich viele Lösungen. In allen anderen Fällen ist die Matrix invertierbar, so daß es eine eindeutige Lösung gibt.

Führen wir die in der Ausgleichsrechnung traditionell benutzten Abkürzungen

$$[x^r] = \sum_{i=1}^{N} x_i^r \, , \quad [y^r] = \sum_{i=1}^{N} x_i^r \quad \text{und} \quad [x^r y^s] = \sum_{i=1}^{N} x_i^r y_i^s$$

ein, so erhält das Gleichungssystem die übersichtlichere Gestalt

$$[x^2]\theta_1 + [x]\theta_2 = [xy] \quad \text{und} \quad [x]\theta_1 + N\theta_2 = [y] \, .$$

Subtraktion von $[x]/[x^2]$ mal der ersten Gleichung von der zweiten führt auf

$$\left(N - \frac{[x]^2}{[x^2]}\right)\theta_2 = [y] - \frac{[x]}{[x^2]}[xy]$$

oder $(N[x^2] - [x]^2)\theta_2 = [y][x^2] - [x][xy]$, d.h.

$$\theta_2 = \frac{[y][x^2] - [x][xy]}{N[x^2] - [x]^2} \,.$$

(Man beachte, daß im Falle der eindeutigen Lösbarkeit sowohl $[x^2] > 0$ als auch $N[x^2] - [x]^2 > 0$ ist.)

Einsetzen von θ_2 in die erste Gleichung ergibt dann auch

$$\theta_1 = \frac{[xy] - [x]\theta_2}{[x^2]} \,.$$

Zurück zu unserem Ausgangsproblem: Wir haben eine Stichprobe

$$\mathcal{S} = \left\{(x_1^{(j)}, \ldots, x_n^{(j)}, y^{(j)}) \;\middle|\; j = 1, \ldots, m\right\} \subset \mathbb{R}^{n+1}$$

und suchen dazu Modelle, deren Parameter sich auf Grund der Stichprobe schätzen lassen. Unsere Strategie dabei wird folgende sein: Wir bestimmen zunächst alle durch Ordnungsideale gegebenen Modelle, deren Parameter sich an Hand der Stichprobe berechnen lassen. Diese Modelle werden, da wir exakte Gleichheit fordern, meist viel zu viele Parameter enthalten. Nach Berechnung dieser Parameter für jedes der erhaltenen Modellen sehen wir hoffentlich, welche Monome jeweils eine wichtige Rolle spielen, d.h. also Koeffizienten haben, die sich deutlich von der Null unterschreiden, und welche nicht. Je nach Anwendung liefert uns auch irgendeine Theorie Aussagen darüber, welche Monome wichtig sein sollten. Wenn wir uns auf die wichtigen Monome beschränken, erhalten wir jeweils ein Modell, für das wir die Parameter nicht mehr so bestimmen können, daß Gleichheit herrscht, aber darauf können wir die gerade betrachtete Methode der kleinsten Quadrate anwenden. Wenn wir das für verschiedene Modelle durchführen, können

wir die EUKLIDischen Abstände zwischen den linken und rechten Seiten vergleichen und erhalten auch daraus Hinweise, wie gut die einzelnen Modelle zur Beschreibung des Zusammenhangs zwischen den Eingangs- und den Ausgangsgrößen geeignet sind.

Für den ersten Schritt, das Auffinden aller geeigneter Ordnungsideale zu einer gegebenen Stichprobe, verwendet die Algebraische Statistik die 1966 von BRUNO BUCHBERGER in seiner Dissertation eingeführten und nach seinem Lehrer benannten GRÖBNER-Basen, mit denen wir uns als nächstes beschäftigen werden.

Kapitel 1
Gröbner-Basen

Die klassische Aufgabe der Algebra besteht in der Lösung von Gleichungen und Gleichungssystemen. Im Falle eines Systems von Polynomgleichungen in mehreren Veränderlichen kann die Lösungsmenge sehr kompliziert sein und, sofern sie unendlich ist, möglicherweise nicht einmal explizit angebbar: Im Gegensatz zum Fall linearer Gleichungen können wir hier im allgemeinen keine endliche Menge von Lösungen finden, durch die sich alle anderen Lösungen ausdrücken lassen. GRÖBNER-Basen liefern zu einem gegebenen nichtlinearen Gleichungssystem ein einfacheres System mit der gleichen Lösungsmenge; es ist eine Art Verallgemeinerung der Treppengestalt, die der GAUSS-Algorithmus liefert. Zumindest bei endlichen Lösungsmengen lassen sich diese auch konkret angeben – sofern wir die Nullstellen von Polynomen einer Veränderlichen explizit berechnen können.

§1: Algebraische Vorbereitungen

Wenn wir lineare Gleichungssysteme mit dem GAUSS-Algorithmus lösen, verändern wir das Gleichungssystem sukzessive, indem wir Gleichungen so durch Linearkombinationen mit anderen Gleichungen ersetzen, daß sich an der Lösungsmenge nichts ändert. Indem wir eine lineare Gleichung

$$a_1X_1 + \cdots + a_nX_n = b$$

über einem Körper k mit dem $(n+1)$-Tupel $(a_1, \ldots, a_n, b) \in k^{n+1}$ identifizieren, sehen wir leicht, daß die sämtlichen linearen Gleichungen in n Unbekannten über einem Körper k einen $(n+1)$-dimensionalen Vektorraum bilden; die Gleichungen eines konkreten linearen Gleichungssystems erzeugen darin einen Untervektorraum. Dieser besteht aus allen

Linearkombinationen der gegebenen Gleichungen, und das sind gleichzeitig alle linearen Gleichungen, die auf der Lösungsmenge des linearen Gleichungssystems verschwinden. Zwei lineare Gleichungssysteme haben somit genau dann die gleiche Lösungsmenge, wenn sie den gleichen Untervektorraum erzeugen.

Bei der Lösung nichtlinearer Gleichungssysteme können wir versuchen, ähnlich vorzugehen. Angenommen, wir suchen die Menge der gemeinsamen Nullstellen der beiden Polyonome

$$f = X^2Y^2 + 2X^3 - 3X^2 - X \quad \text{und} \quad g = Y^2 + X - 3\,,$$

also die Menge

$$\mathcal{L} = \left\{(x,y) \in \mathbb{R} \;\middle|\; f(x,y) = g(x,y) = 0\right\}.$$

Wir könnten den Term X aus beiden Gleichungen eliminieren, indem wir die Summe $f + g$ betrachten, aber offensichtlich hilft uns das nicht wirklich weiter. Nützlicher wäre es wahrscheinlich, einen der „größeren“ Terme zu eliminieren. Linearkombinationen mit Skalaren als Koeffizienten helfen uns dabei nicht weiter. Wenn wir aber g mit X^2 multiplizieren erhalten wir das Polynom $X^2Y^2 + X^3 - 3X^2$, das genau wie f den Term X^2Y^2 enthält, und

$$f - X^2 g = X^3 - X = X(X+1)(X-1)$$

ist in der Tat einfacher als die Ausgangspolynome f und g: Die Differenz hängt nur noch von X ab und verschwindet bei $x = 0$ und $x = \pm 1$. Setzen wir diese drei Werte nacheinander in die beiden Polynome ein, erhalten wir für $x = 0$

$$f(0,y) = 0 \quad \text{und} \quad g(0,y) = y^2 - 3 \quad \text{mit Lösung} \quad y = \pm\sqrt{3}\,,$$

Für $x = 1$ erhalten wir

$$f(1,y) = y^2 - 2 \quad \text{und} \quad g(0,y) = y^2 - 2 \quad \text{mit Lösung} \quad y = \pm\sqrt{2}\,,$$

und für $x = -1$ schließlich erhalten wir

$$f(-1,y) = y^2 - 4 \quad \text{und} \quad g(0,y) = y^2 - 4 \quad \text{mit Lösung} \quad y = \pm 2\,.$$

Die Lösungsmenge ist somit

$$\mathcal{L} = \left\{\left(0,\sqrt{3}\right), \left(0,-\sqrt{3}\right), \left(1,\sqrt{2}\right), \left(1,-\sqrt{2}\right), (-1,2), (-1,-2)\right\}.$$

Im Gegensatz zum Fall der linearen Gleichungssysteme sollten wir uns im nichlinearen Fall also nicht darauf beschränken, Gleichungen mit Skalaren zu multiplizieren und entsprechende Linearkombinationen zu betrachten, sondern wir sollten die Multiplikation mit *beliebigen* Polynomen zulassen. Dies führt auf den Begriff des *Ideals* in einem Ring:

Definition: Eine nichtleere Teilmenge I eines Rings R heißt *Ideal,* in Zeichen $I \lhd R$, wenn gilt:
1.) Für je zwei Elemente $f, g \in I$ ist auch $f + g \in I$
2.) Für jedes $f \in I$ und jedes $r \in R$ liegt auch rf in I.

Bei den Produkten verlangen wir also, daß sie bereits dann in I liegen, wenn nur *ein* Faktor in I liegt.

Die Bedingung, daß ein Ideal mindestens ein Element enthalten muß, können wir auch ersetzen durch die Bedingung, daß es die Null von R enthalten muß, denn wenn es irgendein Element $f \in R$ enthält, muß es gemäß der zweiten Bedingung auch $0 \cdot f = 0$ enthalten.

Um mit dem Idealbegriff vertraut zu werden, betrachten wir zunächst Ideale im Ring der ganzen Zahlen:

Lemma: Zu jedem Ideal $I \lhd \mathbb{Z}$ gibt es eine ganze Zahl $n \in \mathbb{Z}$, so daß $I = \{nq \mid q \in \mathbb{Z}\}$.

Beweis: I ist nach Definition nicht leer, enthält also mindestens ein Element. Falls I nur aus der Null besteht, können wir $n = 0$ setzen und sind fertig. Wenn es ein Element $m \neq 0$ gibt, enthält das Ideal auch dessen sämtliche ganzzahlige Vielfachen, insbesondere also gibt es in I dann positive Zahlen. Die kleinste dieser Zahlen sei n. Wir wollen uns überlegen, daß I genau aus den ganzzahligen Vielfachen von n besteht.

Dazu sei $m \in I$ ein beliebiges Element von I. Wir dividieren m mit Rest durch q; das Ergebnis sei

$$m : n = q \quad \text{Rest } r \quad \text{mit} \quad 0 \leq r < n \,.$$

Dann liegt mit m und n auch $r = m - qn$ in I und ist echt kleiner als n. Da n die kleinste positive Zahl in I ist, muß daher $r = 0$ sein, d.h. $m = qn$ ist ein ganzzahliges Vielfaches von n.

■

Definition: *a)* Ist R ein Ring und $f \in R$ so bezeichnen wir

$$(f) \underset{\text{def}}{=} \{rf \mid r \in R\}$$

als das von f erzeugte *Hauptideal.*
b) R heißt *Hauptidealring,* wenn jedes Ideal von R ein Hauptideal ist.

Das gerade bewiesene Lemma zeigt also, daß $\mathbb{Z}$ ein Hauptidealring ist.

Allgemeiner definieren wir

Definition: Ist R ein Ring und ist $M \subset R$ eine Teilmenge von R, so ist das *von M erzeugte Ideal* (M) das kleinste Ideal von R, das M enthält, d.h. den Durchschnitt aller Ideale, die M enthalten. Für eine endliche Menge $M = \{f_1, \ldots, f_m\}$ schreiben wir (M) kurz als $(f_1, \ldots, f_m)$. Die Menge M bezeichnen wir als ein *Erzeugendensystem* des Ideals I.

Diese Definition macht nicht wirklich klar, wie das von M erzeugte Ideal aussieht. Da uns für das praktische Rechnen nur endlich erzeugte Ideale interessieren, möchte ich mich auf diesen Fall beschränken; die Verallgemeinerung auf beliebige Mengen M sollte für jeden, der den nachfolgenden Beweis verstanden hat, offensichtlich sein.

Lemma: $\left(f_1, \ldots, f_m\right) = \left\{ \sum_{j=1}^{m} r_j f_j \;\middle|\; r_j \in R \right\}$

Beweis: Da jedes Ideal, das $f_1, \ldots, f_m$ enthält, auch für beliebige Elemente $r_1, \ldots, r_m \in R$ die Produkte $r_j f_j$ enthält und damit auch deren Summe, ist klar, daß die rechte Seite in jedem Ideal enthalten ist, das die f_j enthält. Außerdem ist die rechtsstehende Menge selbst ein Ideal: Da sie die f_j enthält, ist sie nicht leer. Die Summe zweier Elemente ist offensichtlich wieder ein Element, da wir einfach die Koeffizienten der einzelnen f_j addieren müssen. Wenn wir schließlich ein Element der rechten Seite mit einem beliebigen Element $r \in R$ multiplizieren, werden einfach alle Koeffizienten mit r multipliziert. Somit ist die rechte Seite in der Tat das kleinste Ideal, das alle f_j enthält.

■

Sei nun $R = k[X_1, \ldots, X_n]$ der Polynomring in n Variablen über einem Körper k, und seien $f_1, \ldots, f_m \in R$ Polynome. Wir interessieren uns

für die Lösungsmenge des durch die f_j gegebenen Gleichungssystems, also die Menge aller $(x_1, \ldots, x_n) \in k^n$, für die alle f_j verschwinden. Wir definieren gleich allgemein

Definition: Die Nullstellenmenge einer Teilmenge $M \subseteq k[X_1, \ldots, X_n]$ ist

$$V(M) \underset{\text{def}}{=} \left\{(x_1, \ldots, x_n) \in k^n \;\middle|\; f(x_1, \ldots, x_n) = 0 \text{ für alle } f \in M\right\}.$$

Im Falle eine endlichen Menge $M = \{f_1, \ldots, f_m\}$ schreiben wir kurz $V(f_1, \ldots, f_m)$. Falls wir uns für Lösungen aus einem größeren Körper $K > k$ interessieren, schreiben wir entsprechend $V_K(I)$ und $V_K(f_1, \ldots, f_m)$.

(In der algebraischen Geometrie bezeichnet man Mengen dieser Art als Varietäten; daher der Buchstabe V.)

Lemma: Ist $I = (f_1, \ldots, f_m)$ das von den f_j erzeugte Ideal, so ist

$$V(I) = V(f_1, \ldots, f_m).$$

Beweis: Da alle f_j in I liegen, ist natürlich $V(I) \subseteq V(f_1, \ldots, f_m)$. Umgekehrt sei $(x_1, \ldots, x_n)$ ein Element von $V(f_1, \ldots, f_m)$ und g irgendein Element von I. Nach dem vorigen Lemma gibt es Polynome $r_j \in R$; so daß $g = \sum_{i=1}^{m} r_j f_j$ ist. Damit ist auch

$$g(x_1, \ldots, x_n) = \sum_{j=1}^{m} r_j(x_1, \ldots, x_m) f_j(x_1, \ldots, x_m) = 0,$$

so daß $(x_1, \ldots, x_n)$ in $V(I)$ liegt. Damit ist das Lemma bewiesen. ∎

Dieses Lemma zeigt, daß zwei Gleichungssysteme

$$f_1(x_1, \ldots, x_n) = 0, \quad \ldots, \quad f_m(x_1, \ldots, x_n) = 0$$

und

$$g_1(x_1, \ldots, x_n) = 0, \quad \ldots, \quad g_r(x_1, \ldots, x_n) = 0$$

die gleiche Lösungsmenge haben, wenn die Ideale $(f_1, \ldots, f_m)$ und $(g_1, \ldots, g_r)$ übereinstimmen.

Die Umkehrung dieser Aussage ist allerdings falsch. Einfache Gegenbeispiele können wir schon bei nur einer Gleichung in einer Variablen finden:

Die Polynome $X^2+1, X^2+2, X^2+3, \ldots \in \mathbb{R}[X]$ haben allesamt keine reellen Nullstellen. Trotzdem sind die Ideale

$$\left(X^2+1\right), \left(X^2+2\right), \left(X^2+3\right), \ldots$$

natürlich allesamt verschieden. Dieses Problem verschwindet allerdings, wenn wir uns nicht mehr nur auf reelle Nullstellen beschränken, sondern auch komplexe zulassen.

Anders ist das bei den Polynomen $X, X^2, X^3, \ldots$. Egal über welchem Körper wir arbeiten: Jedes dieser Polynome hat die Null als einzige Nullstelle, aber trotzdem sind die Ideale $\left(X^d\right) \lhd k[X]$ für verschiedene Werte von d verschieden.

Für das Problem, zu einer gegebenen Stichprobe die damit identifizierbaren Modelle zu finden, betrachtet die Algebraische Statistik das zugehörige Design als Lösungsmenge eines nichtlinearen Gleichungssystems. Es ist daher wichtig, den Zusammenhang zwischen Gleichungssystemen (oder Idealen) und deren Lösungsmengen zu kennen. Damit beschäftigt sich der Rest dieses Paragraphen.

Als erstes definieren wir die Summe und das Produkt zweier Ideale:

Definition: *a)* Die Summe $I + J$ zweier Ideale I, J eines Rings R ist das kleinste Ideal, das sowohl I als auch J enthält.
b) Das Produkt IJ dieser Ideale ist das kleinste Ideal, das alle Produkte fg mit $f \in I$ und $g \in J$ enthält.

Man überlegt sich leicht (mit dem gleichen Argument, mit dem wir das Ideal $(f_1, \ldots, f_m)$ oben explizit bestimmt haben), daß $I + J$ gerade die Menge aller $f + g$ mit $f \in I$ und $g \in J$ ist. IJ dagegen enthält im allgemeinen auch Elemente, die sich *nicht* in der Form fg mit $f \in I$ und $g \in J$ darstellen lassen: Ist etwa $I = J = (X, Y) \lhd \mathbb{R}[X, Y]$, so enthält IJ mit $X^2 = X \cdot X$ und $Y^2 = Y \cdot Y$ auch deren Summe $X^2 + Y^2$, die sich nicht als Produkt zweier Polynome aus $\mathbb{R}[X, Y]$ schreiben läßt. Wenn wir $\mathbb{R}$ durch $\mathbb{C}$ ersetzen, läßt sich $X^2 + Y^2$ zwar zerlegen als

$(X + iY)(X - iY)$, aber auch im Komplexen gibt es Gegenbeispiele: So ist etwa das Polynom $X^2 + Y^2 - Z^2 \in \mathbb{C}[X, Y, Z]$ irreduzibel, also nicht als Produkt zweier nichtkonstanter Faktoren darstellbar, aber es liegt trotzdem im Produkt des Ideals $I = (X, Y, Z)$ mit sich selbst. Im Produkt IJ liegen daher auch alle (endlichen) Summen der Form $\sum f_j g_j$ mit $f_j \in I$ und $g_j \in J$. Da diese (analog zum obigen Argument) ein Ideal bilden, besteht IJ genau aus diesen Summen.

Satz: Für zwei Ideale I, J im Polynomring $R = k[X_1, \ldots, X_n]$ gilt

a) Ist $I \subseteq J$, so ist $V(J) \subseteq V(I)$

b) $V(I + J) = V(I) \cap V(J)$

c) $V(IJ) = V(I) \cup V(J)$

Beweis: a) Sei $(x_1, \ldots, x_n) \in V(J)$. Dann verschwindet $f(x_1, \ldots, x_n)$ für alle $f \in J$. Da I eine Teilmenge von J ist, verschwindet $f(x_1, \ldots, x_n)$ dann erst recht für alle $f \in I$. Somit liegt. $(x_1, \ldots, x_n)$ in $V(I)$.

b) Da $I + J$ das kleinste Ideal ist, das sowohl I als auch J enthält, liegt $V(I + J)$ nach *a)* sowohl in $V(I)$ als auch in $V(J)$, also auch in deren Durchschnitt. Liegt umgekehrt ein Punkt $(x_1, \ldots, x_n)$ sowohl in $V(I)$ als auch in $V(J)$, so liegt er auch in $V(I + J)$, denn wie wir gerade gesehen haben, läßt sich jedes Element von $I + J$ schreiben als $f + g$ mit $f \in I$ und $g \in J$, und sowohl f als auch g verschwinden im Punkt $(x_1, \ldots, x_n)$.

c) Da IJ erzeugt wird von den Produkten fg mit $f \in I$ und $g \in J$ und jedes dieser Produkte sowohl in I als auch in J liegt, ist IJ eine Teilmenge sowohl von I als auch von J. Nach *a)* liegt $V(I) \cup V(J)$ daher in $V(IJ)$.

Umgekehrt sei $(x_1, \ldots, x_n) \in V(IJ)$, liege aber nicht in $V(I)$. Dann gibt es ein $f \in I$ mit $f(x_1, \ldots, x_n) \neq 0$. Für jedes $g \in J$ liegt aber fg in IJ, so daß das Produkt $f(x_1, \ldots, x_n)g(x_1, \ldots, x_n)$ verschwinden muß. Da die Funktionswerte im Körper k liegen und der Faktor $f(x_1, \ldots, x_n)$ nicht verschwindet, muß $g(x_1, \ldots, x_n) = 0$ sein für alle $g \in J$; der Punkt liegt also in $V(J)$. Somit liegt er in jedem Fall in $V(I) \cup V(J)$. ∎

§2: Gauß und Euklid

Zur (exakten) Lösung eines linearen Gleichungssystems in mehreren Veränderlichen verwenden wir üblicherweise den GAUSS-Algorithmus. Für die Lösung eines System von Polynomgleichungen höheren Grades in nur einer Veränderlichen können wir den EUKLIDischen Algorithmus verwenden, denn die gemeinsamen Nullstellen zweier Polynome in einer Veränderlichen sind gerade die Nullstellen ihres größten gemeinsamen Teilers, so daß wir das System durch mehrfache Anwendung des EUKLIDischen Algorithmus reduzieren können auf eine einzige Polynomgleichung, die zudem noch den praktischen Vorteil eines im allgemeinen deutlich kleineren Grads aufweist.

BUCHBERGERs Ansatz zur Lösung nichtlinearer Gleichungssysteme in mehreren Veränderlichen kann als eine Kombination von Ideen hinter dem GAUSSschen Eliminationsverfahren und dem EUKLIDischen Algorithmus aufgefaßt werden; er hat Anwendungen, die weit über das Problem der Lösung nichtlinearer Gleichungssysteme hinausgehen. In der Tat wurde die Grundidee des Verfahrens bereits knapp vor BUCHBERGER, und ohne daß dieser davon wußte, von dem japanischen Mathematiker HEISUKE HIRONAKA entdeckt, der es für ein klassisches Problem der algebraischen Geometrie entwickelte: Für die damit bewiesene sogenannte Auflösung der Singularitäten einer algebraischen Varietät über einem Körper der Charakteristik Null erhielt HIRONAKA 1970 die Fields-Medaille, die damals höchste Auszeichnung der Mathematik. (Seit 2003 vergibt die Norwegische Akademie der Wissenschaften den wie einen Nobelpreis dotierten Abelpreis.)

Wenn wir ein lineares Gleichungssystem durch GAUSS-Elimination lösen, bringen wir es zunächst auf eine Treppengestalt, indem wir die erste vorkommende Variable aus allen Gleichungen außer der ersten eliminieren, die zweite aus allen Gleichungen außer den ersten beiden, uns so weiter, bis wir schließlich Gleichungen haben, deren letzte entweder nur eine Variable enthält oder aber eine Relation zwischen Variablen, für die es sonst keine weiteren Bedingungen mehr gibt. Konkret sieht ein Eliminationsschritt folgendermaßen aus: Wenn wir im Falle der beiden Gleichungen

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = u \quad \text{mit} \quad a_1 \neq 0 \tag{1}$$

$$b_1x_1 + b_2x_2 + \cdots + b_nx_n = v \tag{2}$$

die Variable X_1 mit Hilfe von (1) aus (2) eliminieren wollen, ersetzen wir die zweite Gleichung durch ihre Summe mit $-b_1/a_1$ mal der ersten. Die theoretische Rechtfertigung für diese Umformung besteht darin, daß das Gleichungssystem bestehend aus (1) und (2) sowie das neue Gleichungssystem dieselbe Lösungsmenge haben, und daran ändert sich auch dann nichts, wenn noch weitere Gleichungen dazukommen.

Ähnlich können wir vorgehen, wenn wir ein nichtlineares Gleichungssystem in nur einer Variablen betrachten: Am schwersten sind natürlich die Gleichungen vom höchsten Grad, also versuchen wir, die zu reduzieren auf Polynome niedrigeren Grades. Das kanonische Verfahren dazu ist die Polynomdivision: Haben wir zwei Polynome

$$f = a_dX^d + a_{d-1}X^{d-1} + \cdots + a_1X + a_0 \quad \text{und}$$
$$g = b_eX^e + b_{e-1}X^{e-1} + \cdots + b_1X + b_0$$

mit $e \leq d$, so dividieren wir f durch g, d.h. wir berechnen einen Quotienten q und einen Rest r derart, daß $f = qg + r$ ist und r entweder verschwindet oder kleineren Grad als g hat. Konkret: Bei jedem Divisionsschritt haben wir ein Polynom

$$f = c_\delta X^\delta + c_{\delta-1}X^{\delta-1} + \cdots + c_1X + c_0 \quad \text{mit} \quad c_\delta \neq 0\,,$$

das wir für $\delta \geq e$ mit Hilfe des Divisors

$$g = b_eX^e + b_{e-1}X^{e-1} + \cdots + b_1X + b_0$$

reduzieren, indem wir es ersetzen durch

$$f - \frac{b_e}{c_\delta}X^{\delta-e}g\,.$$

Das führen wir so lange fort, bis f auf Null oder ein Polynom von kleinerem Grad als e reduziert ist: Das ist dann der Divisionsrest r. Auch hier ist klar, daß sich nichts an der Lösungsmenge ändert, wenn man die beiden Gleichungen f, g ersetzt durch g, r, denn

$$f = qg + r \quad \text{und} \quad r = f - qg\,,$$

d.h. f und g verschwinden genau dann für einen Wert x, wenn g und r an der Stelle x verschwinden.

In beiden Fällen ist die Vorgehensweise sehr ähnlich: Wir vereinfachen das Gleichungssystem schrittweise, indem wir eine Gleichung ersetzen durch ihre Summe mit einem geeigneter Vielfachen einer anderen Gleichung.

Dieselbe Strategie wollen wir auch anwenden Systeme von Polynomgleichungen in mehreren Veränderlichen. Erstes Problem dabei ist, daß wir nicht wissen, wie wir die Monome eines Polynoms anordnen sollen und damit, was der führende Term ist. Dazu gibt es eine ganze Reihe verschiedener Strategien, von denen je nach Anwendung mal die eine, mal die andere vorteilhaft ist.

§3: Monomordnungen und der Divisionsalgorithmus

Wir betrachten Polynome in n Variablen $X_1, \ldots, X_n$ über einem Körper k und setzen zur Abkürzung

$$X^\alpha = X_1^{\alpha_1} \cdots X_n^{\alpha_n} \quad \text{mit} \quad \alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0 \,.$$

Terme der Form X^α haben wir in §1 als Monome bezeichnet und ihnen die Summe der α_i als Grad zugeordnet.

Eine Anordnung der Monome ist offensichtlich äquivalent zu einer Anordnung auf $\mathbb{N}_0^n$, und es gibt sehr viele Möglichkeiten, diese Menge anzuordnen. Für uns sind allerdings nur Anordnungen interessant, die einigermaßen kompatibel sind mit der algebraischen Struktur des Polynomrings $k[X_1, \ldots, X_n]$; beispielsweise wollen wir sicherstellen, daß der führende Term des Produkts zweier Polynome das Produkt der führenden Terme der Faktoren ist – wie wir es auch vom Eindimensionalen her gewohnt sind. Daher definieren wir

Definition: *a)* Eine Monomordnung ist eine Ordnungsrelation „$<$“ auf $\mathbb{N}_0^n$, für die gilt

1. „$<$“ ist eine Linear- oder Totalordnung, d.h. für zwei Elemente $\alpha, \beta \in \mathbb{N}_0^n$ ist entweder $\alpha < \beta$ oder $\beta < \alpha$ oder $\alpha = \beta$.
2. Für $\alpha, \beta, \gamma \in \mathbb{N}_0^n$ gilt: $\alpha < \beta \Longrightarrow \alpha + \gamma < \beta + \gamma$.
3. „$<$“ ist eine Wohlordnung, d.h. jede Teilmenge $M \subseteq \mathbb{N}_0^n$ hat ein kleinstes Element.

b) Für ein Polynom $f = \sum_{\alpha \in M} c_\alpha X^\alpha \in k[X_1, \ldots, X_n]$ mit $c_\alpha \neq 0$ für alle α aus der endlichen Teilmenge $M \subset \mathbb{N}_0^n$ sei γ das größte Element von M bezüglich einer fest gewählten Monomordnung. Dann bezeichnen wir bezüglich dieser Monomordnung

- γ = multideg f als Multigrad von f
- X^γ = FM(f) als führendes Monom von f
- c_γ = FK(f) als führenden Koeffizienten von f
- $c_\gamma X^\gamma$ = FT(f) als führenden Term von f

Den Grad deg f von f hatten wir in Kapitel 0 als den höchsten Grad eines Monoms von f definiert; je nach gewählter Monomordnung muß das nicht unbedingt der Grad des führenden Monoms sein.

Beispiele von Monomordnungen sind

a) Die lexikographische Ordnung: Hier ist $\alpha < \beta$ genau dann, wenn für den ersten Index i, in dem sich α und β unterscheiden, $\alpha_i < \beta_i$ ist. Betrachtet man Monome X^α als Worte über dem (geordneten) Alphabet $X_1, \ldots, X_n$, kommt hier ein Monom X^α genau dann vor X^β, wenn die entsprechenden Worte im Lexikon in dieser Reihenfolge gelistet werden. Die ersten beiden Forderungen an eine Monomordnung sind klar, und auch die Wohlordnung macht keine großen Probleme: Man betrachtet zunächst die Teilmenge aller Exponenten $\alpha \in I$ mit kleinstmöglichem α_1, unter diesen die Teilmenge derer mit kleinstmöglichem α_2, *usw.,* bis man bei α_n angelangt ist. Spätestens hier ist die verbleibende Teilmenge einelementig, und ihr einziges Element ist das gesuchte kleinste Element von I.

b) Die graduierte lexikographische Ordnung: Hier ist der Grad eines Monoms erstes Ordnungskriterium: Ist deg $X^\alpha <$ deg X^β, so ist $\alpha < \beta$. Falls beide Monome gleichen Grad haben, soll $\alpha < \beta$ genau dann gelten, wenn α im lexikographischen Sinne kleiner als β ist. Auch hier sind offensichtlich alle drei Forderungen erfüllt.

c) Die inverse lexikographische Ordnung: Hier ist $\alpha < \beta$ genau dann, wenn $\alpha_i < \beta_i$ für den *letzten* Index i, in dem sich α und β unterscheiden. Das entspricht offensichtlich gerade der lexikographischen Anord-

nung bezüglich des rückwärts gelesenen Alphabets $X_n, \ldots, X_1$. Entsprechend läßt sich natürlich auch bezüglich jeder anderen Permutation des Alphabets eine Monomordnung definieren, so daß diese Ordnung nicht sonderlich interessant ist – außer als Bestandteil der als nächstes betrachteten Monomordnung:

d) Die graduierte inverse lexikographische Ordnung: Wie bei der graduierten lexikographischen Ordnung ist hier der Grad eines Monoms erstes Ordnungskriterium: Falls $\deg X^\alpha < \deg X^\beta$, ist $\alpha < \beta$, und nur falls beide Monome gleichen Grad haben, soll $\alpha < \beta$ genau dann gelten, wenn α im Sinne der inversen lexikographischen Ordnung *größer* ist als β. Man beachte, daß wir hier also nicht nur die Reihenfolge der Variablen umkehren, sondern auch die Ordnungsrelation im Fall gleicher Grade. Es ist nicht schwer zu sehen, daß auch damit eine Monomordnung definiert wird: Mit den ersten beiden Forderungen gibt es wie üblich keine Probleme, und wenn wir eine Menge M von Monomen haben, gibt es darin eine Teilmenge bestehend aus den Monomen kleinsten Grades. Da es für jeden Grad nur endlich viele Monome gibt, ist diese Menge endlich, hat also bezüglich der inversen lexikographischen Ordnung nicht nur ein kleinstes, sondern auch ein größtes Element. Dieses ist das kleinste Element von M bezüglich der graduierten inversen lexikographischen Ordnung.

Diese vier Beispiele spielen bei vielen klassischen Anwendungen von GRÖBNER-Basen eine Hauptrolle, es gibt aber noch unendlich viele weitere mögliche Monomordnungen. Für unsere Zwecke in der Algebraischen Statistik brauchen wir *alle* diese Monomordnungen. Zum Glück wird sich aber zeigen, daß stets eine endliche Teilmenge davon ausreicht, um alle identifizierbaren Modelle zu bestimmen.

Für das folgende werden wir noch einige Eigenschaften einer (beliebigen) Monomordnung benötigen, die in der Definition nicht erwähnt sind.

Als erstes wollen wir uns überlegen, daß bezüglich jeder Monomordnung auf $\mathbb{N}_0^n$ kein Element kleiner sein kann als $(0, \ldots, 0)$: Wäre nämlich

$\alpha < (0, \ldots, 0)$, so wäre wegen der zweiten Eigenschaft auch

$$2\alpha = \alpha + \alpha < \alpha + (0, \ldots, 0) = \alpha$$

und so weiter, so daß wir eine unendliche Folge

$$\alpha > 2\alpha > 3\alpha > \cdots$$

hätten, im Widerspruch zur dritten Forderung.

Daraus folgt nun sofort, daß das Produkt zweier Monome nicht kleiner sein kann als die beiden Faktoren und damit auch, daß ein echter Teiler eines Monoms bezüglich jeder Monomordnung kleiner sein muß als dieses. Außerdem folgt, daß für ein Produkt von Polynomen stets $\mathrm{FM}(fg) = \mathrm{FM}(f) \cdot \mathrm{FM}(g)$ ist.

Die Eliminationsschritte beim GAUSS-Algorithmus können auch als Divisionen mit Rest verstanden werden, und beim EUKLIDischen Algorithmus ist ohnehin alles Division mit Rest. Für ein Verallgemeinerung der beiden Algorithmen auf Systeme nichtlinearer Gleichungssysteme brauchen wir also auch einen Divisionsalgorithmus für Polynome in mehreren Veränderlichen, der die eindimensionale Polynomdivision mit Rest und die Eliminationsschritte beim GAUSS-Algorithmus verallgemeinert.

Beim GAUSS-Algorithmus brauchen wir im allgemeinen mehr als nur einen Eliminationsschritt, bis wir eine Gleichung auf eine Variable reduziert haben; entsprechend wollen wir auch hier einen Divisionsalgorithmus betrachten, der im Unterschied zum klassischen Fall auch mehrere Divisoren gleichzeitig behandeln kann.

Wir gehen also aus von einem Polynom f aus dem Polynomring $R = f \in k[X_1, \ldots, X_n]$, wobei k irgendein Körper ist, in dem wir rechnen können. Meistens ist $k = \mathbb{Q}$ oder eine endliche Erweiterung davon. Dieses Polynom wollen wir dividieren durch m weitere Polynome $f_1, \ldots, f_m \in R$, d.h. wir suchen Polynome $a_1, \ldots, a_m, r \in R$, so daß

$$f = a_1 f_1 + \cdots + a_m f_m + r$$

ist, wobei r in irgendeiner noch zu präzisierenden Weise kleiner als die f_j sein soll.

Da es sowohl bei GAUSS als auch bei EUKLID auf die Anordnung der Terme ankommt, legen wir als erstes eine Monomordnung fest; wenn im folgenden von führenden Termen *etc.* die Rede ist, soll es sich stets um die führenden Terme *etc.* bezüglich dieser Ordnung handeln.

Mit dieser Konvention geht der Algorithmus dann folgendermaßen:

Gegeben sind $f, f_1, \ldots, f_m \in R$ und eine Monomordnung auf R.
Berechnet werden $a_1, \ldots, a_m, r \in R$ mit $f = a_1 f_1 + \cdots + a_m f_m + r$, wobei r kein Monom enthält, das durch das führende Monom eines der f_j teilbar ist.

1. Schritt (Initialisierung): Setze $a_1 = \cdots = a_m = r = 0$ und $p = f$.

2. Schritt (Endebedingung): Im Falle $p = 0$ endet der Algorithmus.

3. Schritt (Divisionsschritt): Falls keiner der führenden Terme FT f_j den führenden Term FT p teilt, wird p ersetzt durch $p - \mathrm{FT}\, p$ und r durch $r + \mathrm{FT}\, p$. Andernfalls sei j der kleinste Index, für den FT f_j Teiler von FT p ist; der Quotient sei q. Dann wird a_j ersetzt durch $a_j + q$ und p durch $p - q f_j$. Danach geht es zurück zum 2. Schritt.

Offensichtlich ist die Bedingung $f - p = a_1 f_1 + \cdots + a_m f_m + r$ nach der Initialisierung im ersten Schritt erfüllt, und sie bleibt auch bei jeder Anwendung des Divisionsschritts erfüllt. Außerdem endet der Algorithmus nach endlich vielen Schritten: Bei jedem Divisionsschritt wird der führende Term von p eliminiert, und alle Monome, die eventuell neu dazukommen, sind kleiner oder gleich dem führenden Monom von f_j. Da letzteres das (alte) führende Monom von p teilt, kann es nicht größer sein als dieses, d.h. der führende Term des neuen p ist kleiner als der des alten. Wegen der Wohlordnungseigenschaft einer Monomordnung kann es keine unendliche absteigende Kette von Monomen geben; daher muß der Algorithmus nach endlich vielen Schritten abbrechen.

Bei der klassischen Polynomdivision für Polynome in einer Variablen über einem Körper wissen wir, daß der Rest kleineren Grad hat als der Divisor. Das muß hier nicht der Fall sein; wir können nur sagen, daß der Rest keine Monome enthält, die durch den führenden Term eines der Divisoren f_j teilbar sind.

Um den Algorithmus besser zu verstehen, betrachten wir zunächst zwei Beispiele:

Als erstes dividieren wir $f = X^2Y + XY^2 + Y^2$ durch $f_1 = XY - 1$ und $f_2 = Y^2 - 1$.

Zur Initialisierung setzen wir $a_1 = a_2 = r = 0$ und $p = f$. Wir verwenden die lexikographische Ordnung; bezüglich derer ist der führende Term von p gleich X^2Y und der von f_1 gleich XY. Letzterer teilt X^2Y, wir setzen also

$$p \leftarrow p - Xf_1 = XY^2 + X + Y^2 \quad \text{und} \quad a_1 \leftarrow a_1 + X = X\,.$$

Neuer führender Term von p ist XY^2; auch das ist ein Vielfaches von XY, also setzen wir

$$p \leftarrow p - Yf_1 = X + Y^2 + Y \quad \text{und} \quad a_1 \leftarrow a_1 + Y = X + Y\,.$$

Nun ist X der führende Term von p, und der ist weder durch XY noch durch Y^2 teilbar, also kommt er in den Rest:

$$p \leftarrow p - X = Y^2 + Y \quad \text{und} \quad r \leftarrow r + X = X\,.$$

Der nun führende Term Y^2 von p ist gleichzeitig der führende Term von f_2 und nicht teilbar durch XY, also wird

$$p \leftarrow p - f_2 = Y + 1 \quad \text{und} \quad a_2 \leftarrow a_2 + 1 = 1\,.$$

Die verbleibenden Terme von p sind weder durch XY noch durch Y^2 teilbar, kommen also in den Rest, so daß wir als Ergebnis erhalten

$$f = a_1f_1 + a_2f_2 + r \;\text{ mit }\; a_1 = X + Y,\;\; a_2 = 1 \;\text{ und }\; r = X + Y + 1\,.$$

Wenn wir statt durch das Paar (f_1, f_2) durch (f_2, f_1) dividiert hätten, hätten wir im ersten Schritt zwar ebenfalls X^2Y durch XY dividiert, denn durch Y^2 ist es nicht teilbar. Der neue führende Term XY^2 ist aber durch beides teilbar, und wenn f_2 an erster Stelle steht, nehmen wir im Zweifelsfall dessen führenden Term. Man rechnet leicht nach, daß man hier mit folgendem Ergebnis endet:

$$f = a_1f_1 + a_2f_2 + r \quad \text{mit} \quad a_1 = X + 1, \quad a_2 = X \quad \text{und} \quad r = X + 1\,.$$

Wie wir sehen, sind also sowohl die „Quotienten“ a_j als auch der „Rest“ r von der Reihenfolge der f_j abhängig. Sie hängen natürlich im allgemeinen auch von der verwendeten Monomordnung ab.

Als zweites Beispiel wollen wir $f = XY^2 - X$ durch die beiden Polynome $f_1 = XY + 1$ und $f_2 = Y^2 - 1$ dividieren. Im ersten Schritt dividieren wir XY^2 durch XY mit Ergebnis Y, ersetzen also f durch $-X - Y$. Diese beiden Terme sind weder durch XY noch durch Y^2 teilbar, also ist unser Endergebnis

$$f = a_1 f_1 + a_2 f_2 + r \quad \text{mit} \quad a_1 = Y, \quad a_2 = 0 \quad \text{und} \quad r = -X - Y\,.$$

Hätten wir stattdessen durch (f_2, f_1) dividiert, hätten wir als erstes XY^2 durch Y^2 dividiert mit Ergebnis X; da $f = X f_2$ ist, geht die Division hier ohne Rest auf. Der Divisionsalgorithmus erlaubt uns also nicht einmal die sichere Feststellung, ob f als Linearkombination der f_j darstellbar ist oder nicht. Als alleiniges Hilfsmittel zur Lösung nichtlinearer Gleichungssysteme reicht er offenbar nicht aus. Daher müssen wir in den folgenden Paragraphen noch weitere Werkzeuge betrachten.

§4: Der Hilbertsche Basissatz

Die Grundidee des Algorithmus von BUCHBERGER besteht darin, das Gleichungssystem so abzuändern, daß möglichst viele seiner Eigenschaften bereits an den führenden Termen der Gleichungen ablesbar sind.

Angenommen, wir haben ein nichtlineares Gleichungssystem

$$f_1(x_1, \ldots, x_n) = \cdots = f_m(x_1, \ldots, x_n) = 0$$

mit $f_j \in R = k[X_1, \ldots, X_n]$; seine Lösungsmenge sei $\mathcal{L} \subseteq k^n$.

Wie wir aus §1 wissen, hängt $\mathcal{L}$ nur ab von dem Ideal $I = (f_1, \ldots, f_m)$; zur Lösung des Systems sollten wir daher versuchen, ein möglichst „einfaches“ Erzeugendensystem für dieses Ideal zu finden.

Ganz besonders einfach (wenn auch selten ausreichend) sind Ideale, die von Monomen erzeugt werden:

Definition: Ein Ideal $I \lhd R = k[X_1, \ldots, X_n]$ heißt *monomial,* wenn es eine (nicht notwendigerweise endliche) Menge von Monomen gibt, die I erzeugt.

Nehmen wir an, I werde erzeugt von den Monomen X^α mit α aus einer Indexmenge A. Ist dann X^β irgendein Monom aus I, kann es als endliche Linearkombination

$$X^\beta = \sum_{j=1}^{r} f_j X^{\alpha^{(j)}} \quad \text{mit} \quad \alpha^{(j)} \in A$$

geschrieben werden, wobei die f_j irgendwelche Polynome aus R sind.

Da sich jedes Polynom als Summe von Monomen schreiben läßt, können wir die f_j als k-Linearkombinationen von Monomen X^γ schreiben und bekommen damit eine neue Darstellung von X^β als Summe von Termen der Form $cX^\gamma X^\alpha$ mit $\alpha \in A, \gamma \in \mathbb{N}_0^n$ und $c \in k$. Sortieren wir diese Summanden nach den resultierenden Monomen $X^{\gamma+\alpha}$ und fassen alle Summanden mit gleichem Monom zusammen, so entsteht eine k-Linearkombination verschiedener Monome, die insgesamt gleich X^β ist. Das ist aber nur möglich, wenn diese Summe aus dem einen Summanden X^β besteht, d.h. β läßt sich schreiben in der Form $\beta = \alpha + \gamma$ mit einem $\alpha \in A$ und einem $\gamma \in \mathbb{N}_0^n$.

Dies zeigt, daß ein Monom X^β genau dann in I liegt, wenn $\beta = \alpha + \gamma$ ist mit einem $\alpha \in A$ und einem $\gamma \in \mathbb{N}_0^n$, d.h. X^β ist das Produkt eines der erzeugenden Monome mit *irgendeinem* Monom. Das gesamte Ideal I besteht daher genau aus den Polynomen f, die sich als k-Linearkombinationen solcher Monome schreiben lassen.

Insbesondere liegt ein Polynom f genau dann in einem monomialen Ideal I, wenn jedes seiner Monome dort liegt.

Der HILBERTschen Basissatz, um den es in diesem Paragraphen geht, sagt aus, daß jedes Ideal im Polynomring $k[X_1, \ldots, X_n]$ über einem Körper von endlich vielen seiner Elemente erzeugt werden kann. HILBERT bewies diesen Satz ohne Umweg über monomiale Ideale. Für uns werden aber in diesem Kapitel monomiale Ideale eine herausragende Rolle spielen. Deshalb betrachten wir zunächst das (historisch jüngere)

Lemma von DICKSON, wonach ein monomiales Ideal von endlich vielen Monomen erzeugt werden kann, und folgern daraus den HILBERTschen Basissatz.

Lemma von Dickson: Jedes monomiale Ideal in $R = k[X_1, \ldots, X_n]$ kann von endlich vielen Monomen erzeugt werden.

Tatsächlich gilt sogar eine leicht schärfere Aussage: Ist $A \subseteq \mathbb{N}_0^n$ die Menge aller Exponenten von Monomen aus I, so ist $I = (X^\alpha | \alpha \in A)$, und die Behauptung folgt aus dem nächsten

Lemma: Ist A eine beliebige Teilmenge von $\mathbb{N}_0^n$ und I das von den Monomen X^α mit $\alpha \in A$ erzeugte Ideal von $R = k[X_1, \ldots, X_n]$, so gibt es eine endliche Teilmenge $A' \subseteq A$ derart, daß I bereits von den Monomen X^α mit $\alpha \in A'$ erzeugt wird.

Der *Beweis* wird durch vollständige Induktion nach n geführt. Im Fall $n = 1$ ist alles klar, denn da sind die Monome gerade die Potenzen der einzigen Variable, d.h. A ist eine Teilmenge von $\mathbb{N}_0$, und wir können für A' die einelementige Menge bestehend aus der kleinsten Zahl in A nehmen.

Im Fall $n > 1$ und $\alpha \in \mathbb{N}_0^n$ setzen wir $X'^\alpha = X_1^{\alpha_1} \cdots X_{n-1}^{\alpha_{n-1}}$ und betrachten zunächst das Ideal $J \lhd k[X_1, \ldots, X_{n-1}]$, das von den X'^α mit $\alpha \in A$ erzeugt wird. Nach Induktionsvoraussetzung gibt es eine endliche Teilmenge $A'' \subseteq A$, so daß J bereits von den Monomen X'^α mit $\alpha \in A''$ erzeugt wird.

Die Monome X^α mit $\alpha \in A''$ erzeugen zumindest ein Teilideal $I' \subseteq I$ von I. Ist r der größte Exponent α_n der in einem der n-tupel $\alpha \in A''$ auftritt, so enthält I' für jedes Monom $X^\beta \in J$ mit $\beta \in \mathbb{N}_0^{n-1}$ das Monom $X^\beta X_n^r$, denn X^β ist ja ein Vielfaches eines der Monome X'^α mit $\alpha \in A''$.

Es gibt aber natürlich auch noch Monome in I, in denen X_n mit einem kleineren Exponenten als r auftritt. Um auch diese Elemente zu erfassen, betrachten wir für jedes $s < r$ das Ideal $J_s \lhd k[X_1, \ldots, X_{n-1}]$, das von allen jeden Monomen X'^α mit $\alpha \in A$ erzeugt wird, für die $X'^\alpha X_n^s$

in I liegt. Nach Induktionsannahme gibt es auch zu jedem der Ideale J_s eine endliche Teilmenge A_s von A, so daß J_s erzeugt wird von den Monomen X'^{α} mit $\alpha \in A_s$. Zusammen mit den Monomen der Form $X'^{\alpha} X_n^s$ mit $\alpha \in A_s$ erzeugen dann die Monome X^{α} mit $\alpha \in A''$ ganz I. Die Monome $X'^{\alpha} X^s$ mit $\alpha \in A_s$ liegen in I, also gibt es zu jedem dieser Monome ein $\alpha \in A$, das es teilt. $A'_s \subseteq A$ sei die Menge aller so erhaltener $\alpha \in A$. Dann ist $A' = A'' \cup A'_0 \cup \cdots A'_{r-1}$ eine endliche Teilmenge von A mit der Eigenschaft, daß die X^{α} mit $\alpha \in A'$ das Ideal I erzeugen.

∎

LEONARD EUGENE DICKSON (1874–1954) wurde in Iowa geboren, wuchs aber in Texas auf. Seine Bachelor- und Mastergrade bekam er von der University of Texas, dann wurde er mit seiner 1896 eingereichten Dissertation *Analytic Representation of Substitutions on a Power of a Prime Number of Letters with a Discussion of the Linear Group* der erste an die University of Chicago promovierte Mathematiker. Auch die weiteren seiner 275 wissenschaftlichen Arbeiten, darunter acht Bücher, beschäftigen sich vor allem mit der Algebra und Zahlentheorie. Den größten Teil seines Berufslebens verbrachte er als Professor an der University of Chicago, besuchte aber auch regelmäßig die University of California in Berkeley.

Beliebige Ideale sind im allgemeinen nicht monomial; schon das von $X + 1$ erzeugte Ideal in $k[X]$ ist ein Gegenbeispiel, denn es enthält weder das Monom X noch das Monom 1, im Widerspruch zu der oben gezeigten Eigenschaft eines monomialen Ideals, zu jedem seiner Elemente auch dessen sämtliche Monome zu enthalten.

Um monomiale Ideale auch für die Untersuchung solcher Ideale nützlich zu machen, wählen wir eine Monomordnung auf R und definieren für ein beliebiges Ideal $I \lhd R \underset{\text{def}}{=} k[X_1, \ldots, X_n]$ das monomiale Ideal

$$\mathrm{FM}(I) = \Big(\mathrm{FM}(f) \;\Big|\; f \in I \setminus \{0\}\Big),$$

das von den führenden Monomen *aller* Elemente von I erzeugt wird – außer natürlich dem nicht existierenden führenden Monom der Null.

Hilbertsche Basissatz: Jedes Ideal $I \lhd R = k[X_1, \ldots, X_n]$ hat ein endliches Erzeugendensystem.

Beweis: Nach dem Lemma von DICKSON ist FM(I) erzeugt von endlich vielen Monomen. Jedes dieser Monome ist, wie wir bereits wissen, ein Vielfaches eines der erzeugenden Monome, also eines führenden Monoms eines Elements von I. Es gibt somit endlich viele Polynome $f_1, \ldots, f_m \in I$ mit der Eigenschaft, daß FM(f_1), …, FM(f_m) das Ideal FM(I) erzeugen. Wir wollen zeigen, daß die Polynome $f_1, \ldots, f_m$ das Ideal I erzeugen.

Dazu sei f ein beliebiges Element von I. Wir wenden wir den Divisionsalgorithmus an auf f und die f_j und erhalten eine Darstellung

$$f = a_1 f_1 + \cdots + a_m f_m + r$$

mit Polynomen $a_1, \ldots, a_m$ und r aus R. Wir müssen zeigen, daß der Divisionsrest r verschwindet.

Falls r *nicht* verschwindet, garantiert der Divisionsalgorithmus, daß das führende Monom FM(r) von r durch kein führendes Monom FM(f_j) eines der Divisoren f_j teilbar ist. Andererseits ist aber

$$r = f - (a_1 f_1 + \cdots + a_m f_m)$$

ein Element von I, und damit liegt FM(r) im von den FM(f_j) erzeugten Ideal FM(I). Somit muß FM(r) Vielfaches eines FM(f_j) sein, ein Widerspruch. Also ist $r = 0$.

■

DAVID HILBERT (1862–1943) wurde in Königsberg geboren, wo er auch zur Schule und zur Universität ging. Er promovierte dort 1885 mit einem Thema aus der Invariantentheorie, habilitierte sich 1886 und bekam 1893 einen Lehrstuhl. 1895 wechselte er an das damals auch international führende Zentrum der deutschen Mathematik, die Universität Göttingen, wo er bis zu seiner Emeritierung im Jahre 1930 lehrte. Seine Arbeiten umfassen ein riesiges Spektrum aus unter anderem Invariantentheorie, Zahlentheorie, Geometrie, Funktionalanalysis, Logik und Grundlagen der Mathematik sowie auch zur Relativitätstheorie. Er gilt als einer der Väter der modernen Algebra.

§5: Gröbner-Basen und der Buchberger-Algorithmus

Angesichts der Rolle der führenden Monome im obigen Beweis bietet sich folgende Definition an für eine Idealbasis, bezüglich derer möglichst viele Eigenschaften bereits an den führenden Monomen abgelesen werden können:

Definition: Eine endliche Teilmenge $G = \{g_1, \ldots, g_m\} \subset I$ eines Ideals $I \lhd R = k[X_1, \ldots, X_n]$ heißt Standardbasis oder GRÖBNER-Basis von I, falls die Monome FM(g_j) das Ideal FM(I) erzeugen.

WOLFGANG GRÖBNER wurde 1899 im damals noch österreichischen Südtirol geboren. Nach Ende des ersten Weltkriegs, in dem er an der italienischen Front kämpfte, studierte er zunächst an der TU Graz Maschinenbau, beendete dieses Studium aber nicht, sondern begann 1929 an der Universität Wien ein Mathematikstudium. Nach seiner Promotion ging er zu EMMY NOETHER nach Göttingen, um dort Algebra zu lernen. Aus materiellen Gründen mußte er schon bald nach Österreich zurück, konnte aber auch dort zunächst keine Anstellung finden, so daß er Kleinkraftwerke baute und im Hotel seines Vaters aushalf. Ein italienischen Mathematiker, der dort seinen Urlaub verbrachte, vermittelte ihm eine Stelle an der Universität Rom, die er 1939 wieder verlassen mußte, nachdem er sich beim Anschluß Südtirols an Italien für die deutsche Staatsbürgerschaft entschieden hatte. Während des zweiten Weltkriegs arbeitete er größtenteils an einem Forschungsinstitut der Luftwaffe, nach Kriegsende als Extraordinarius in Wien, dann als Ordinarius in Innsbruck, wo er 1980 starb. Seine Arbeiten beschäftigen sich mit der Algebra und algebraischen Geometrie sowie mit Methoden der Computeralgebra zur Lösung von Differentialgleichungen.

Die Theorie der GRÖBNER-Basen wurde von seinem Studenten BRUNO BUCHBERGER in dessen Dissertation entwickelt. BUCHBERGER wurde 1942 in Innsbruck geboren, wo er auch Mathematik studierte und 1966 bei GRÖBNER promovierte mit der Arbeit *Ein Algorithmus zum Auffinden der Basiselemente des Restklassenrings nach einem nulldimensionalen Polynomideal.* Er arbeitete zunächst als Assistent, nach seiner Habilitation als Dozent an der Universität Innsbruck, bis er 1974 einen Ruf auf den Lehrstuhl für Computermathematik an der Universität Linz erhielt. Dort gründete er 1987 das Research Institute for Symbolic Computation (RISC), dessen Direktor er bis 1999 war. 1989 initiierte er in Hagenberg (etwa 20 km nordöstlich von Linz) die Gründung eines Softwareparks mit angeschlossener Fachhochschule; er hat mittlerweile fast Tausend Mitarbeiter. Außer mit Computeralgebra beschäftigt er sich auch im Rahmen des Theorema-Projekts mit dem automatischen Beweisen mathematischer Aussagen.

Im Beweis des HILBERTschen Basissatzes haben wir gesehen, daß Elemente $f_1, \ldots, f_m$ eines Ideals I, deren führende Monome das Ideal FM(I) erzeugen, einn Erzeugendensystem von I bilden. Daher er-

zeugt eine GRÖBNER-Basis von I das Ideal. Außerdem hat jedes Ideal I im Polynomring eine GRÖBNER-Basis, denn nach dem Lemma von DICKSON hat das Ideal der führenden Monome ein endliches Erzeugendensystem, und jedes Monom aus diesem Erzeugendensystem ist führendes Monom eines Polynoms $f_j \in I$. Die Menge der Polynome f_j ist offensichtlich eine GRÖBNER-Basis im Sinne der obigen Definition.

Bevor wir uns damit beschäftigen, wie man diese berechnen kann, wollen wir zunächst eine wichtige Eigenschaft betrachten.

$\{g_1, \ldots, g_m\}$ sei eine GRÖBNER-Basis eines Ideals $I \lhd R$. Wir wollen ein beliebiges Element $f \in R$ durch $g_1, \ldots, g_m$ dividieren. Dies liefert als Ergebnis

$$f = a_1 g_1 + \cdots + a_m g_m + r \,,$$

wobei kein Monom von r durch eines der Monome FM(g_j) teilbar ist. Wie wir wissen, sind allerdings bei der Polynomdivision im allgemeinen weder der Divisionsrest r noch die Koeffizienten a_j auch nur im entferntesten eindeutig. Wir wollen untersuchen, wie sich das hier verhält.

Angenommen. wir haben zwei Darstellungen

$$f = a_1 g_1 + \cdots + a_m g_m + r = b_1 g_1 + \cdots + b_m g_m + s$$

der obigen Form. Dann ist

$$(a_1 - b_1) g_1 + \cdots + (a_m - b_m) g_m = s - r \,.$$

Links steht ein Element von I, also auch rechts. Andererseits enthält aber weder r noch s ein Monom, das durch eines der Monome FM(g_j) teilbar ist, d.h. $r - s = 0$, da die FM(g_j) ja das Ideal FM(I) erzeugen. Somit ist bei der Division durch die Elemente einer GRÖBNER-Basis der Divisionsrest eindeutig bestimmt. Insbesondere ist f genau dann ein Element von I, wenn der Divisionsrest verschwindet. Wenn wir eine GRÖBNER-Basis haben, können wir also leicht entscheiden, ob ein gegebenes Element $f \in R$ im Ideal I liegt.

Nachdem im Fall einer GRÖBNER-Basis der Divisionsrest nicht von der Reihenfolge der Basiselemente abhängt, können wir ihn durch ein Symbol bezeichnen, das nur von der *Menge* $G = \{g_1, \ldots, g_m\}$ abhängt; wir schreiben $\overline{f}^G$.

Als nächstes wollen wir uns mit der Frage beschäftigen, wie wir für ein vorgegebenes Ideal I eine GRÖBNER-Basis bestimmen können.

Dazu müssen wir uns als erstes überlegen, *wie* das Ideal gegeben sein soll. Wenn wir dkonkrett rechnen wollen, müssen wir irgendeine Art von endlicher Information haben; was sich anbietet ist natürlich ein endliches Erzeugendensystem.

Wir gehen also aus von einem Ideal $I = (f_1, \ldots, f_m)$ und suchen eine GRÖBNER-Basis. Das Problem ist, daß die Monome $\mathrm{FM}(f_j)$ im allgemeinen nicht ausreichen, um das monomiale Ideal $\mathrm{FM}(I)$ zu erzeugen: Für $I = (f_1, f_2)$ mit $f_1 = X^2 + X$ und $f_2 = X^2 + Y$ etwa ist bezüglich der lexikographischen Ordnung $\mathrm{FM}(f_1) = \mathrm{FM}(f_2) = X^2$, da aber beispielsweise auch $f_1 - f_2 = X - Y$ in I liegt, enthält $\mathrm{FM}(I)$ zumindest auch noch dessen führenden Term X, so daß X^2 allein nicht zur Erzeugung von $\mathrm{FM}(I)$ ausreicht. Wir müssen daher Linearkombinationen der f_j finden, deren führende Monome durch kein führendes Monom eines f_j teilbar sind, und zumindest diese noch mit dazu nehmen.

BUCHBERGERs Idee dazu orientiere sich an den Eliminationsschritten im GAUSS-Algorithmus, die er mit den sogenannten S-Polynomen verallgemeinerte auf Linearkombinationen mit Polynomen als Koeffizienten: Für zwei Polynome $f, g \in R$ mit $\mathrm{FM}(f) = X^\alpha$ und $\mathrm{FM}(g) = X^\beta$ sei X^γ das kgV von X^α und X^β, d.h. $\gamma_i = \max(\alpha_i, \beta_i)$ für $i = 1, \ldots, n$. Das S-Polynom von f und g ist

$$S(f,g) \underset{\text{def}}{=} \frac{X^\gamma}{\mathrm{FT}(f)} \cdot f - \frac{X^\gamma}{\mathrm{FT}(g)} \cdot g\,.$$

Da $\frac{X^\gamma}{\mathrm{FT}(f)} \cdot f$ und $\frac{X^\gamma}{\mathrm{FT}(g)} \cdot g$ beide nicht nur dasselbe führende Monom X^γ haben, sondern es wegen der Division durch den führenden *Term* statt nur das führende Monom auch beide mit Koeffizient eins enthalten, fällt es bei der Bildung von $S(f,g)$ weg. Daher ist das führende Monom von $S(f,g)$ kleiner als X^γ. Das folgende Lemma ist der Kern des Beweises, daß S-Polynome alles sind, was wir brauchen, um GRÖBNER-Basen zu berechnen:

Lemma: Für die Polynome $f_1, \ldots, f_m \in R$ sei

$$S = \sum_{j=1}^{m} \lambda_j X^{\alpha^{(j)}} f_j \quad \text{mit} \quad \lambda_j \in k \quad \text{und} \quad \alpha^{(j)} \in \mathbb{N}_0^n$$

eine Linearkombination, zu der es ein $\delta \in \mathbb{N}_0^n$ gebe, so daß alle Summanden X^δ als führendes Monom haben, d.h. $\alpha^{(j)} + \text{multideg } f_j = \delta$ für $j = 1, \ldots, m$. Falls multideg $S < \delta$ ist, gibt es Elemente $\lambda_{ij} \in k$, so daß

$$S = \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_{ij} X^{\delta - \gamma^{(ij)}} S(f_i, f_j)$$

ist mit $X^{\gamma^{(ij)}} = \text{kgV}\big(\text{FM}(f_i), \text{FM}(f_j)\big)$.

Beweis: Der führende Koeffizient von f_j sei μ_j. Dann ist $\lambda_j \mu_j$ der führende Koeffizient von $\lambda_j X^{\alpha^{(j)}} f_j$. Somit ist multideg S genau dann kleiner als δ, wenn $\sum_{i=j}^{m} \lambda_i \mu_j$ verschwindet. Wir normieren alle $X^{\alpha^{(j)}} f_j$ auf führenden Koeffizienten eins, indem wir durch μ_j dividieren und somit $p_j = X^{\alpha^{(j)}} f_j / \mu_j$ betrachten. Dann ist

$$\begin{aligned} S = \sum_{j=1}^{m} \lambda_j \mu_j p_j = {} & \lambda_1 \mu_1 (p_1 - p_2) + (\lambda_1 \mu_1 + \lambda_2 \mu_2)(p_2 - p_3) + \cdots \\ & + (\lambda_1 \mu_1 + \cdots + \lambda_{m-1} \mu_{m-1})(p_{m-1} - p_m) \\ & + (\lambda_1 \mu_1 + \cdots + \lambda_m \mu_m) p_m \,, \end{aligned}$$

wobei der Summand in der letzten Zeile genau im Falle multideg $S < \delta$ verschwindet.

Da alle p_i denselben Multigrad δ und denselben führenden Koeffizienten eins haben, kürzen sich in den Differenzen $p_i - p_j$ die führenden Terme weg, genau wie in den S-Polynomen. In der Tat: Bezeichnen wir den Multigrad von $\text{kgV}\big(\text{FM}(f_i), \text{FM}(f_j)\big)$ mit $\gamma^{(ij)}$, so ist

$$p_i - p_j = X^{\delta - \gamma^{(ij)}} S(f_i, f_j) \,.$$

Damit hat die obige Summendarstellung von S die gewünschte Form. ∎

Daraus folgt ziemlich unmittelbar das Kriterium von BUCHBERGER, wonach ein Erzeugendensystem eines Ideals genau dann eine GRÖBNER-Basis ist, wenn jedes S-Polynom zweier Erzeugender bei der Division durch das Erzeugendensystem den Rest Null ergibt. Da wir gesehen haben, daß es durchaus von der Reihenfolge der Divisoren abhängen kann, ob wir einen verschwindenden oder nichtverschwindenden Rest erhalten, wollen wir die Bedingung „Rest Null“ etwas abschwächen:

Definition: Ein Polynom f reduziert auf Null bezüglich der Polynome $f_1, \ldots, f_m$, wenn es Polynome $h_1, \ldots, h_m$ gibt, für die gilt
1.) $f = h_1 f_1 + \cdots + h_m f_m$
2.) Bezüglich der betrachteten Monomordnung ist FM($h_j f_j$) für jedes $j = 1, \ldots, m$ kleiner oder gleich FM(f).

Wenn der Divisionsalgorithmus Rest Null liefert, ist diese Bedingung offensichtlich erfüllt, aber sie ist etwas allgemeiner, und sie reicht aus für BUCHBERGERs Kriterium:

Satz: Ein Erzeugendensystem $f_1, \ldots, f_m$ eines Ideals I im Polynomring $R = k[X_1, \ldots, X_n]$ ist genau dann eine GRÖBNER-Basis, wenn jedes S-Polynom $S(f_i, f_j)$ bezüglich $f_1, \ldots, f_m$ auf Null reduziert.

Beweis: Als R-Linearkombination von f_i und f_j liegt das S-Polynom $S(f_i, f_j)$ im Ideal I; falls $f_1, \ldots, f_m$ eine GRÖBNER-Basis von I ist, hat es also Rest Null bei der Division durch $f_1, \ldots, f_m$.

Umgekehrt sei $f_1, \ldots, f_m$ ein Erzeugendensystem von $I \lhd R$ mit der Eigenschaft, daß alle $S(f_i, f_j)$ bezüglich $f_1, \ldots, f_m$ auf Null reduzieren. Wir wollen zeigen, daß $f_1, \ldots, f_m$ dann eine GRÖBNER-Basis ist, daß also die führenden Monome FM(f_1), …, FM(f_m) das Ideal FM(I) erzeugen.

Dazu sei $f \in I$ ein beliebiges Element; wir müssen zeigen, daß FM(f) im von den FM(f_j) erzeugten Ideal liegt.

Da f in I liegt, gibt es eine Darstellung

$$f = h_1 f_1 + \cdots + h_m f_m \quad \text{mit} \quad h_j \in R\,.$$

Falls sich hier bei den führenden Termen nichts wegkürzt, ist das führende Monom von f gleich dem führenden Monom mindestens eines Produkts $h_j f_j$ und somit ein Vielfaches des führenden Monoms von FM(f_j), so daß FM(f) im von den FM(f_j) erzeugten Ideal liegt.

Falls sich die maximalen unter den führenden Termen FT($h_j f_j$) gegenseitig wegkürzen, läßt sich die entsprechende Teilsumme der $h_j f_j$ nach dem vorigen Lemma auch als eine Summe von S-Polynomen schreiben. Diese wiederum lassen sich nach Definition als Linearkombinationen der f_j darstellen, wobei das führende Monom eines jeden Summanden in dieser Linearkombination höchstens so groß ist wie das führende Monom des jeweiligen S-Polynoms. Damit erhalten wir eine neue Darstellung

$$f = \widetilde{h}_1 f_1 + \cdots + \widetilde{h}_m f_m \quad \text{mit} \quad \widetilde{h}_j \in R\,,$$

in der der maximale Multigrad eines Summanden echt kleiner ist als in der obigen Darstellung, denn in der Darstellung als Summe von S-Polynomen sind die Terme mit dem maximalem Multigrad verschwunden.

Mit dieser Darstellung können wir wie oben argumentieren: Falls sich bei den führenden Termen nichts wegkürzt, haben wir FM(f) als Element des von den FM(f_j) erzeugten Ideals dargestellt, andernfalls erhalten wir wieder via S-Polynome und deren Reduktion eine neue Darstellung von f als Linearkombination der f_j mit noch kleinerem maximalem Multigrad der Summanden, und so weiter. Das Verfahren muß schließlich mit einer Summe ohne Kürzungen bei den führenden Termen enden, da es nach der Wohlordnungseigenschaft einer Monomordnung keine unendliche absteigende Folge von Multigraden geben kann. ∎

Der BUCHBERGER-Algorithmus in seiner einfachsten Form macht aus diesem Satz ein Verfahren zur Berechnung einer GRÖBNER-Basis aus einem vorgegebenen Erzeugendensystem eines Ideals:

Gegeben sind m Elemente $f_1, \ldots, f_m \in R = k[X_1, \ldots, X_n]$.

Berechnet wird eine GRÖBNER-Basis $g_1, \ldots, g_r$ des davon erzeugten Ideals $I = (f_1, \ldots, f_m)$ mit $g_j = f_j$ für $j \leq m$.

1. Schritt (Initialisierung): Setze $g_j = f_j$ für $j = 1, \ldots, m$; die Menge $\{g_1, \ldots, g_m\}$ werde mit G bezeichnet.

2. Schritt: Setze $G' = G$ und teste für jedes Paar $(f, g) \in G' \times G'$ mit $f \neq g$, ob der Rest r bei der Division von $S(f, g)$ durch die Elemente von G' (in irgendeiner Reihenfolge angeordnet) verschwindet. Falls nicht, wird G ersetzt durch $G \cup \{r\}$.

3. Schritt: Ist $G = G'$, so endet der Algorithmus mit G als Ergebnis; andernfalls geht es zurück zum zweiten Schritt.

Wenn der Algorithmus im dritten Schritt endet, ist der Rest bei der Division von $S(f, g)$ durch die Elemente von G stets das Nullpolynom; nach dem gerade bewiesenen Satz ist G daher eine GRÖBNER-Basis. Da sowohl die S-Polynome als auch ihre Divisionsreste in I liegen und G ein Erzeugendensystem von I enthält, ist auch klar, daß es sich dabei um eine GRÖBNER-Basis von I handelt. Wir müssen uns daher nur noch überlegen, daß der Algorithmus nach endlich vielen Iterationen abbricht.

Wenn im zweiten Schritt ein nichtverschwindender Divisionsrest r auftaucht, ist dessen führendes Monom durch kein führendes Monom eines Polynoms $g \in G$ teilbar. Das von den führenden Monomen der $g \in G$ erzeugte Ideal von R wird daher größer, nachdem G um r erweitert wurde. Wenn dies unbeschränkt möglich wäre, erhielten wir eine unendliche aufsteigende Folge von monomialen Idealen J_i, von denen jedes echt größer wäre als sein Vorgänger:

$$J_1 \subset J_2 \subset \cdots \subset J_i \subset J_{i+1} \subset \cdots .$$

Natürlich ist auch die Vereinigung J aller J_i ein monomiales Ideal, hat also nach dem Lemma von DICKSON ein endliches Erzeugendensystem $\{M_1, \ldots, M_q\}$. Da jedes M_j in einem J_i und damit auch in allen folgenden liegen muß, gibt es ein m, so daß alle M_j in J_m liegen. Damit ist $J = (M_1, \ldots, M_q) \subseteq J_m$, im Widerspruch zur Annahme, daß J_{m+1} und damit auch J echt größer als J_m ist.

Der Algorithmus kann auf mehrere offensichtliche Weisen optimiert werden: Beispielsweise muß man beim wiederholten Durchlaufen des zweiten Schritts keine S-Polynome mehr betrachten, die schon beim

vorigen Durchlauf berechnet wurden: Falls ein solches S-Polynom modulo der dort betrachteten Menge G auf Divisionsrest Null hatte, reduziert es erst recht bezüglich der nun betrachteten größeren Menge G auf Null, und wenn nicht, ist der alte Divisionsrest ein Element der neuen Menge G, so daß es modulo dieser Menge auf Null reduziert.

Der Hauptaufwand beim BUCHBERGER-Algorithmus besteht in der Berechnung und Reduktion der S-Polynome. Falls wir daher überflüssige Anwendungen des Divisionsalgorithmus im Voraus erkennen können, spart das viel Aufwand. Ein einfaches Kriterium dazu ist

Lemma: Sind $\mathrm{FM}(f)$ und $\mathrm{FM}(g)$ teilerfremd, so reduziert $S(f,g)$ modulo f, g auf Null.

Beweis: Aus der Definition eines S-Polynoms folgt sofort, daß sich an $S(f,g)$ nichts ändert, wenn wir f und/oder g mit einer Konstanten multiplizieren: Da wir in beiden Termen durch den führenden Köffizienten dividieren, ändert sich nichts am jeweiligen Produkt. Daher können wir o.B.d.A. davon ausgehen, daß sowohl f als auch g den führenden Koeffizienten eins haben.

Wenn die führenden Monome teilerfremd sind, ist ihr kleinstes gemeinsames Vielfaches gleich ihrem Produkt. Schreiben wir $f = \mathrm{FM}(f) + p$ und $g = \mathrm{FM}(g) + q$, ist daher

$$\begin{aligned} S(f,g) &= \frac{\mathrm{FM}(f)\,\mathrm{FM}(g)}{\mathrm{FT}(f)} \cdot f - \frac{\mathrm{FM}(f)\,\mathrm{FM}(g)}{\mathrm{FT}(g)} \cdot g \\ &= \mathrm{FM}(g) \cdot f - \mathrm{FM}(f) \cdot g \\ &= (g-q) \cdot f - (f-p) \cdot g \\ &= gf - qf - fg + pg = -qf + pg\,. \end{aligned}$$

Damit ist $S(f,g)$ als Linearkombination von f und g dargestellt; um zu sehen, daß es modulo f und g auf Null reduziert, müssen wir noch zeigen, daß sich in der Summe $-qf + pg$ die führenden Terme nicht wegheben können. Dazu reicht es zu zeigen, daß die führenden Monome der beiden Summanden verschieden sind.

Wäre dies nicht der Fall, so wäre $\mathrm{FM}(q)\,\mathrm{FM}(f) = \mathrm{FM}(p)\,\mathrm{FM}(g)$. Da $\mathrm{FM}(f)$ und $\mathrm{FM}(g)$ teilerfremd sind, müßte also $\mathrm{FM}(g)$ ein Teiler von

FM(q) sein und FM(f) einer von FM(p). Beides ist aber unmöglich, denn FM(f) ist echt größer als FM(p) und FM(g) echt größer als FM(q), während ein Teiler eines Monoms bezüglich jeder Monomordnung kleiner ist als dieses. Damit ist das Lemma bewiesen. ∎

Teilerfremdheit der führenden Monome ist natürlich äquivalent dazu, daß es keine Variable gibt, die in *beiden* führenden Monomen vorkommt.

Es gibt inzwischen zahlreiche weniger offensichtliche Verbesserungen und Optimierungen des BUCHBERGER-Algorithmus. Wir wollen uns aber mit dem Prinzip begnügen, um dafür mehr Zeit für Themen zu bekommen, die für die algebraische Statistik relevanter sind.

Der BUCHBERGER-Algorithmus hat den Nachteil, daß er das vorgegebene Erzeugendensystem in jedem Schritt größer macht, ohne je ein Element zu streichen. Dies ist weder beim GAUSS-Algorithmus noch beim EUKLIDischen Algorithmus der Fall, bei denen jeweils eine Gleichung durch eine andere *ersetzt* wird. Obwohl wir sowohl die Eliminationsschritte des GAUSS-Algorithmus als auch die einzelnen Schritte der Polynomdivisionen beim EUKLIDischen Algorithmus durch S-Polynome ausdrücken können, *müssen* wir im allgemeinen Fall zusätzlich zu g und $S(f,g)$ auch noch das Polynom f beibehalten; andernfalls kann sich die Lösungsmenge ändern:

Als Beispiel können wir das Gleichungssystem

$$f(X,Y) = X^2Y + XY^2 + 1 = 0 \quad \text{und} \quad g(X,Y) = X^3 - XY - Y = 0$$

betrachten. Wenn wir mit der lexikographischen Ordnung arbeiten, sind hier die einzelnen Monome bereits der Größe nach geordnet, insbesondere stehen also die führenden Monome an erster Stelle und

$$S(f,g) = Xf(X,Y) - Yg(X,Y) = X^2Y^2 + XY^2 + X + Y^2 .$$

Der führende Term X^2Y^2 ist durch den führenden Term X^2Y von f teilbar; subtrahieren wir Yf vom S-Polynom, erhalten wir das nicht weiter reduzierbare Polynom

$$h(X,Y) = -XY^3 + XY^2 + X + Y^2 - Y .$$

Sowohl $g(X,Y)$ als auch $h(X,Y)$ verschwinden im Punkt $(0,0)$; dieser ist aber keine Lösung des Ausgangssystems, da $f(0,0) = 1$ nicht verschwindet.

Aus diesem Grund werden die nach dem BUCHBERGER-Algorithmus berechneten GRÖBNER-Basen oft sehr groß und unhandlich. Betrachten wir dazu als Beispiel das System aus den beiden Gleichungen

$$f_1 = X^3 - 2XY \quad \text{und} \quad f_2 = X^2Y - 2Y^2 + X$$

und berechnen eine GRÖBNER-Basis bezüglich der graduiert lexikographischen Ordnung.

$$S(f_1, f_2) = Yf_1 - Xf_2 = -X^2$$

ist weder durch den führenden Term von f_1 noch den von f_2 teilbar, muß also als neues Element f_3 in die Basis aufgenommen werden.

$$S(f_1, f_3) = f_1 + Xf_3 = -2XY$$

kann wieder mit keinem der f_j reduziert werden, muß also als neues Element f_4 in die Basis. Genauso ist es mit

$$f_5 = S(f_2, f_3) = f_2 + Yf_3 = -2Y^2 + X\,.$$

Für das so erweiterte Erzeugendensystem, bestehend aus den Polynomen

$$f_1 = X^3 - 2XY, \quad f_2 = X^2Y - 2Y^2 + X, \quad f_3 = -X^2,$$
$$f_4 = -2XY \quad \text{und} \quad f_5 = -2Y^2 + X\,,$$

sind die S-Polynome

$$S(f_1, f_2) = f_3, \quad S(f_1, f_3) = f_4 \quad \text{und} \quad S(f_2, f_3) = f_5$$

trivialerweise auf Null reduzierbar, die anderen Kombinationen müssen wir nachrechnen:

$$S(f_1, f_4) = Yf_1 + \frac{X^2}{2}f_4 = -2XY^2 = Yf_4$$

$$S(f_1, f_5) = Y^2f_1 + \frac{X^3}{2}f_5 = -2XY^3 + \frac{X^4}{2} = \frac{X}{2}f_1 + f_2 + Y^2f_4 - f_5$$

$$S(f_2, f_4) = f_2 + \frac{X}{2}f_4 = -2Y^2 + X = f_5$$

$$S(f_2, f_5) = Yf_2 + \frac{X^2}{2}f_5 = \frac{X^3}{2} + XY - 2Y^3 = \frac{1}{2}f_1 - \frac{1}{2}f_4 + Yf_5$$

$$S(f_3, f_4) = -Yf_3 - \frac{X}{2}f_4 = 0$$

$$S(f_3, f_5) = -Y^2 f_3 - \frac{X^2}{2}f_5 = \frac{1}{2}f_1 - \frac{1}{2}f_4$$

$$S(f_4, f_5) = -\frac{Y}{2}f_4 - \frac{X}{2}f_5 = \frac{X^2}{2} = -\frac{1}{2}f_3$$

Somit bilden diese fünf Polynome eine GRÖBNER-Basis des von f_1 und f_2 erzeugten Ideals.

Zum Glück brauchen wir aber nicht alle fünf Polynome. Das folgende Lemma gibt ein Kriterium, wann man auf ein Erzeugendes verzichten kann, und illustriert gleichzeitig das allgemeine Prinzip, wonach bei einer GRÖBNER-Basis alle wichtigen Eigenschaften anhand der führenden Termen ablesbar sein sollten:

Lemma: G sei eine GRÖBNER-Basis des Ideals $I \lhd k[X_1, \ldots, X_n]$, und $g \in G$ sei ein Polynom, dessen führendes Monom im von den führenden Monomen der restlichen Basiselemente erzeugten monomialen Ideal liegt. Dann ist auch $G \setminus \{g\}$ eine GRÖBNER-Basis von I.

Beweis: $G \setminus \{g\}$ ist nach Definition genau dann eine GRÖBNER-Basis von I, wenn die führenden Terme der Basiselemente das Ideal FM(I) erzeugen. Da G eine GRÖBNER-Basis von I ist und die führenden Terme egal ob mit oder ohne FT(g) dasselbe monomiale Ideal erzeugen, ist das klar. ∎

Man beachte, daß sich dieses Lemma nur anwenden läßt, wenn G eine GRÖBNER-Basis von I ist; wir können nicht schon während des Rechengangs im BUCHBERGER-Algorithmus Elemente streichen. Im obigen Beispiel etwa wird das Ideal $I = (f_1, f_2)$ natürlich auch erzeugt von f_1, f_2 und f_3; dabei ist FM(f_1) = X^3, FM(f_2) = X^2Y, und FM(f_3) = X^2 teilt beide dieser Monome. Wenn das Lemma auf die Basis f_1, f_2, f_3 anwendbar wäre, könnten wir also f_1 und f_2 streichen und f_3 wäre für sich allein eine GRÖBNER-Basis von I. Natürlich ist aber $I \neq (-X^2)$, denn weder f_1 noch f_2 sind Vielfache von X^2.

Von der Menge $\{f_1, f_2, f_3, f_4, f_5\}$ haben wir mit Hilfe des Kriteriums von BUCHBERGER verifiziert, daß sie eine GRÖBNER-Basis von I ist; deshalb können wir das Lemma darauf anwenden und f_1, f_2 streichen. Wir können das aber erst jetzt tun, denn im Verlauf der Berechnungen wurden f_1 und f_2 noch gebraucht um $f_4 = S(f_1, f_3)$ und $f_5 = S(f_2, f_3)$ zu konstruieren. Somit ist $I = (f_3, f_4, f_5)$, und darauf können wir das Lemma nicht weiter anwenden, denn

$$\mathrm{FM}(f_3) = X^2, \quad \mathrm{FM}(f_4) = XY \quad \text{und} \quad \mathrm{FM}(f_5) = Y^2,$$

wobei keines dieser drei Monome Vielfaches eines der anderen ist.

Zur weiteren Normierung können wir noch durch die führenden Koeffizienten teilen und erhalten dann eine *minimale* GRÖBNER-Basis mit

$$g_1 = X^2, \quad g_2 = XY \quad \text{und} \quad g_3 = Y^2 - \frac{X}{2}.$$

Definition: Eine *minimale* GRÖBNER-Basis von I ist eine GRÖBNER-Basis von I mit folgenden Eigenschaften:

1.) Alle $g \in G$ haben den führenden Koeffizienten eins

2.) Für kein $g \in G$ liegt $\mathrm{FM}(g)$ im von den führenden Monomen der übrigen Elemente erzeugten Ideal.

Da ein Monom X^α genau dann im von einer Menge M von Monomen erzeugten Ideal liegt, wenn es durch eines dieser Monome teilbar ist, können wir die zweite Bedingung auch so ausdrücken, daß es keine zwei Elemente $g \neq g'$ in G geben darf, für die $\mathrm{FM}(g)$ ein Teiler von $\mathrm{FM}(g')$ ist.

Es ist klar, daß jede GRÖBNER-Basis zu einer minimalen GRÖBNER-Basis verkleinert werden kann: Durch Division können wir alle führenden Koeffizienten zu eins machen ohne etwas an der Erzeugung zu ändern, und nach obigem Lemma können wir nacheinander alle Elemente eliminieren, die die zweite Bedingung verletzen.

Wir können aber noch mehr erreichen: Wenn nicht das führende, sondern einfach *irgendein* Monom eines Polynoms $g \in G$ im von den führenden Termen der übrigen Elemente erzeugten Ideal liegt, ist dieses Monom teilbar durch das führende Monom eines anderen Polynoms $h \in G$. Wir

können den Term mit diesem Monom daher zum Verschwinden bringen, indem wir g ersetzen durch g minus ein Vielfaches von h. Da sich dabei nichts an den führenden Termen der Elemente von G ändert, bleibt G eine GRÖBNER-Basis. Wir können somit aus den Elementen einer minimalen GRÖBNER-Basis Terme eliminieren, die durch den führenden Term eines anderen Elements teilbar sind. Was dabei schließlich entstehen sollte, ist eine *reduzierte* GRÖBNER-Basis:

Definition: Eine reduzierte GRÖBNER-Basis von I ist eine GRÖBNER-Basis von I mit folgenden Eigenschaften:
1.) Alle $g \in G$ haben den führenden Koeffizienten eins
2.) Für kein $g \in G$ liegt ein Monom von g im von den führenden Monomen der übrigen Elemente erzeugten Ideal.

Die minimale Basis im obigen Beispiel ist offenbar schon reduziert, denn außer g_3 bestehen alle Basispolynome nur aus dem führendem Term, und bei g_3 ist der zusätzliche Term linear, kann also nicht durch die quadratischen führenden Monome der anderen Polynome teilbar sein.

Reduzierte GRÖBNER-Basis haben eine für das praktische Rechnen mit Idealen sehr wichtige zusätzliche Eigenschaft:

Satz: Jedes Ideal $I \lhd k[X_1, \ldots, X_n]$ hat (bei vorgegebener Monomordnung) eine eindeutig bestimmte reduzierte GRÖBNER-Basis.

Beweis: Wir gehen aus von einer minimalen GRÖBNER-Basis G und ersetzen nacheinander jedes Element $g \in G$ durch seinen Rest bei der Polynomdivision durch $G \setminus \{g\}$. Da bei einer minimalen GRÖBNER-Basis kein führendes Monom eines Element das führende Monom eines anderen teilen kann, ändert sich dabei nichts an den führenden Termen, G ist also auch nach der Ersetzung eine minimale GRÖBNER-Basis. In der schließlich entstehenden Basis hat kein $g \in G$ mehr einen Term, der durch den führenden Term eines Elements von $G \setminus \{g\}$ teilbar wäre, denn auch wenn wir bei der Reduktion der einzelnen Elemente durch eine eventuell andere Menge geteilt haben, hat sich doch an den führenden Termen der Basiselemente nichts geändert. Also gibt es eine reduzierte GRÖBNER-Basis.

Nun seien G und G' zwei reduzierte GRÖBNER-Basen von I. Jedes Element $f \in G'$ liegt insbesondere in I, also ist $\overline{f}^G = 0$. Insbesondere muß der führende Term von f durch den führenden Term eines $g \in G$ teilbar sein. Umgekehrt ist aber auch $\overline{g}^{G'} = 0$, d.h. der führende Term von g muß durch den führenden Term eines Elements von $f' \in G'$ teilbar sein. Dieser führende Term teilt dann insbesondere den führenden Term von f, und da G' als reduzierte GRÖBNER-Basis minimal ist, muß $f' = f$ sein. Somit gibt es zu jedem $g \in G$ genau ein $f \in G'$ mit FM(f) = FM(g). Insbesondere haben G und G' dieselbe Elementanzahl. Tatsächlich muß sogar $f = g$ sein, denn $f - g$ liegt in I, enthält aber keine Term, der durch den führenden Term irgendeines Elements von G teilbar wäre. Also ist $f - g = 0$.

■

Bemerkung: Die Forderung in den Definitionen von minimalen und reduzierten GRÖBNER-Basen, daß alle führenden Koeffizienten eins sein müssen, ist zwar nützlich für theoretische Diskussionen, führt aber im Falle von Polynomen mit rationalen Koeffizienten oft dazu, daß die Koeffizienten Nenner haben. Computeralgebrasysteme können zwar mit rationalen Zahlen rechnen, indem sie diese durch Paare teilerfremder ganzer Zahlen darstellen, aber diese Rechnungen sind erheblich aufwendiger als solche mit ganzen Zahlen. Daher liefern einige Computeralgebrasysteme beim Kommando zur Berechnung einer reduzierten GRÖBNER-Basis anstelle von Polynomen mit führendem Koeffizienten eins solche mit teilerfremden ganzzahligen Koeffizienten.

Maxima etwa hat ein Paket zur Berechnung von GRÖBNER-Basen, das man mit `load(grobner)` laden kann. Für die Polynome $f_1 = 2XY - 1$ und $f_2 = 3Y^2 - 1$ aus $\mathbb{Z}[X, Y]$ liefert

```
poly_reduced_grobner([2*X*Y-1, 3*Y^2 - 1], [X, Y]);
```

die (im Sinne von Maxima) reduzierte GRÖBNER-Basis

$$[3Y^2 - 1, 2X - 3Y]$$

bezüglich der lexikographischen Ordnung. Auch beim Divisionsalgorithmus haben alle Ergebnisse ganzzahlige Koeffizienten, dafür wird noch eine Konstante c ausgegeben, mit der der Dividend multipliziert

werden muß, damit das Ergebnis stimmt. (Alternativ kann man natürlich auch die Quotienten und den Rest durch diese Konstante dividieren.) Bei der Division von $f = X^2Y + XY^2 + Y^2$ durch f_1 und f_2 etwa führt der Befehl

```
poly_pseudo_divide(X^2*Y + X*Y^2 + Y^2,
    [2*X*Y-1, 3*Y^2], [X,Y]);
```

zur Ausgabe

$$[[3Y + 3X, 2], 3Y + 3X + 2, 6, 3]$$

mit der Bedeutung, daß

$$6f = (3Y + 3X)f_1 + 2f_2 + (3Y + 3X + 2)$$

ist, wobei drei Reduktionsschritte durchgeführt wurden.

Kapitel 2
Systeme von nichtlinearen Polynomgleichungen

GRÖBNER-Basen haben eine Vielzahl von Anwendungen in der Algebra; wir wollen uns hier vor allem damit beschäftigen, wie sie direkt oder im Zusammenspiel mit anderen Methoden zur expliziten Lösung nichtlinearer Gleichungssysteme führen können. Explizit angebbar sind die Lösungen meist nur, wenn die Lösungsmenge endlich ist; daher werden wir uns meist auf solche Systeme beschränken und interessieren uns daher auch für Kriterien, wie wir einem Gleichungssystem die Endlichkeit seiner Lösungsmenge ansehen können.

§1: Gröbner-Basen für nichtlineare Gleichungssysteme

Wir gehen aus von m Polynomgleichungen

$$f_j(x_1, \ldots, x_n) = 0 \quad \text{mit} \quad f_j \in k[X_1, \ldots, X_n] \quad \text{für} \quad j = 1, \ldots, m$$

und suchen die im vorigen Kapitel mit $V(f_1, \ldots, f_m)$ bezeichnete Lösungsmenge

$$\left\{(x_1, \ldots, x_n) \in k^n \;\middle|\; f_j(x_1, \ldots, x_n) = 0 \text{ für } j = 1, \ldots, m\right\}.$$

Diese wird allerdings oft leer sein; für $f_1 = X^2 - 2$ und $f_2 = Y^2 - 3$ aus $\mathbb{Q}[X, Y]$ etwa ist diese Menge leer, da die Lösungen $(\pm\sqrt{2}, \pm\sqrt{3})$ nicht in $\mathbb{Q}^2$ liegen. Wir betrachten daher meist noch einen zweiten Körper K, der k enthält, und interessieren uns allgemeiner für die Lösungsmenge $V_K(f_1, \ldots, f_m) \subseteq K^n$.

Der Körper k sollte dabei möglichst klein sein, denn mit den Elementen dieses Körpers müssen wir rechnen, und je größer der Körper, desto aufwendiger sind seine Rechenoperationen. In konkreten Beispielen

werden wir uns meist auf $k = \mathbb{Q}$ beschränken und – soweit möglich – sogar versuchen, unsere Konstruktionen in $\mathbb{Z}[X]$ durchzuführen.

Der Körper K hingegen sollte so groß sein, daß er für ein Gleichungssystem, daß in irgendeinem Körper eine nichtleere endliche Lösungsmenge hat, diese Lösungsmenge enthält. Wir werden meist $K = \mathbb{C}$ betrachten.

Wie wir bereits aus §1 des vorigen Kapitels wissen, hängt die Lösungsmenge des Gleichungssystems nur ab vom Ideal $I = (f_1, \ldots, f_m)$; wir suchen ein Erzeugendensystem $\{g_1, \ldots, g_r\}$ dieses Ideals, aus dem wir mehr über die Mengen

$$V_K(I) = V_K(f_1, \ldots, f_m) = V_K(g_1, \ldots, g_r)$$

ablesen können. Wir erwarten natürlich, daß wir hier vor allem im Falle einer geeigneten GRÖBNER-Basis $\{g_1, \ldots, g_r\}$ eventuell Erfolg haben.

Viele Lösungsansätze für Gleichungssysteme in mehreren Veränderlichen beruhen auf der Elimination von Variablen: Im ℓ-ten Schritt suchen wir nach Bedingungen, die ein $(n - \ell)$-Tupel $(x_{\ell+1}, \ldots, x_n)$ erfüllen muß, wenn es ein ℓ-Tupel $(x_1, \ldots, x_\ell)$ gibt, so daß $(x_1, \ldots, x_n)$ in $V(I)$ liegt. Eine solche Bedingung ist trivial: Für jedes Polynom $f \in I$, in dem die Variablen $X_1, \ldots, X_\ell$ nicht vorkommen, muß $f(x_{\ell+1}, \ldots, x_n) = 0$ sein.

Definition: *a)* Das ℓ-te *Eliminationsideal* eines Ideal $I \lhd k[X_1, \ldots, X_n]$ ist $I_\ell = I \cap k[X_{\ell+1}, \ldots, X_n]$.
b) Eine Monomordnung $<$ heißt *Eliminationsordnung* für $X_1, \ldots, X_\ell$, wenn jedes Monom, das mindestens eine der Variablen $X_1, \ldots, X_\ell$ enthält, größer ist als alle Monome, die nur $X_{\ell+1}, \ldots, X_n$ enthalten.

Die lexikographische Ordnung mit $X_1 > X_2 > \cdots > X_{n-1} > X_n$ ist offensichtlich für jedes ℓ eine Eliminationsordnung für $X_1, \ldots, X_\ell$, die graduiert lexikographische aber nicht, da bezüglich dieser beispielsweise $X_1 < X_n^2$ ist.

Satz: Ist G eine GRÖBNER-Basis von I bezüglich einer Eliminationsordnung für $X_1, \ldots, X_\ell$, so ist $G \cap I_\ell$ eine GRÖBNER-Basis von I_ℓ.

Beweis: Die Elemente von $G = \{g_1, \ldots, g_m\}$ seien so angeordnet, daß $G \cap I_\ell = \{g_1, \ldots, g_r\}$ ist. Wir müssen zeigen, daß sich jedes $f \in I_\ell$ als Linearkombination von $g_1, \ldots, g_r$ mit Koeffizienten aus $k[X_{\ell+1}, \ldots, X_n]$ darstellen läßt.

Der Divisionsalgorithmus bezüglich der gewählten Ordnung gibt uns eine Darstellung $f = h_1 g_1 + \cdots + h_m g_m$ von f als Element von I. Die Polynome $g_{r+1}, \ldots, g_m$ enthalten jeweils mindestens eine der Variablen $X_1, \ldots, X_\ell$, und da wir eine Eliminationsordnung verwenden, muß auch das führende Monom eine dieser Variablen enthalten. Da kein Monom von f eine dieser Variablen enthält, kann im Divisionsalgorithmus das führende Monom eines dieser Polynome nie Teiler des führenden Monoms des jeweils betrachteten Polynoms p sein, Somit ist $h_{r+1} = \cdots = h_m = 0$, und in keinem der Polynome $h_1, \ldots, h_r$ kann eine der Variablen $X_1, \ldots, X_\ell$ auftreten. Dies zeigt, daß f im von $g_1, \ldots, g_r$ erzeugten Ideal von $k[X_{\ell+1}, \ldots, X_n]$ liegt, d.h. dieses Ideal wird von $g_1, \ldots, g_r$ erzeugt.

Um zu zeigen, daß es sich dabei sogar um eine GRÖBNER-Basis handelt, können wir zum Beispiel zeigen, daß alle $S(g_i, g_j)$ mit $i, j \leq r$ ohne Rest durch $g_1, \ldots, g_r$ teilbar sind. Da G nach Voraussetzung eine GRÖBNER-Basis ist, sind sie auf jeden Fall ohne Rest durch G teilbar, und wieder kann bei der Division nie der führende Term eines Dividenden durch den eines g_i mit $i > r$ teilbar sein, d.h. $S(g_i, g_j)$ ist als Linearkombination von $g_1, \ldots, g_r$ mit Koeffizienten aus $k[g_1, \ldots, g_r]$ darstellbar. ∎

Daraus ergibt sich eine Strategie zur Lösung nichtlinearer Gleichungssysteme nach Art des GAUSS-Algorithmus: Wir gehen aus von der lexikographischen Ordnung, die ja für jedes ℓ eine Eliminationsordnung für $X_1, \ldots, X_\ell$ ist, und bestimmen eine (reduzierte) GRÖBNER-Basis für das von den Gleichungen erzeugte Ideal des Polynomrings $k[X_1, \ldots, X_n]$. Dann betrachten als erstes das Eliminationsideal I_{n-1}. Dieses besteht nur aus Polynomen in X_n; falls wir mit einer reduzierten GRÖBNER-Basis arbeiten, gibt es darin höchstens ein solches Polynom.

Falls es ein solches Polynom gibt, muß jede Lösung des Gleichungssystem als letzte Komponente eine von dessen Nullstellen haben. Wir

bestimmen daher diese Nullstellen (in K) und setzen sie nacheinander in das restliche Gleichungssystem ein. Dadurch erhalten wir Gleichungssysteme in $n-1$ Unbekannten, wo wir nach Gleichungen nur in X_{n-1} suchen können. Diese erhalten wir, indem wir bei allen Erzeugenden des Eliminationsideals I_{n-2} für X_n nacheinander die Werte aus $V_K(I_{n-1}) \subset k$ einsetzen. Nachdem wir so $V_K(I_{n-2}) \subset K^2$ bestimmt haben, können wir analog die Mengen $V_K(I_{n-3}) \subset K^3$ und so weiter bis $V_K(I) \subset K^n$ bestimmen.

Betrachten wir noch einmal das Beispiel gegen Ende von §5 des vorigen Kapitels mit

$$f_1 = X^3 - 2XY \quad \text{und} \quad f_2 = X^2Y - 2Y^2 + X\,.$$

Dort hatten wir die reduzierte GRÖBNER-Basis bezüglich der graduiert lexikographischen Ordnung berechnet; sie besteht aus

$$g_1 = X^2, \quad g_2 = XY \quad \text{und} \quad g_3 = Y^2 - \frac{X}{2}\,.$$

Da die graduiert lexikographische Ordnung keine Eliminationsordnung für X ist, können wir nicht erwarten, daß $\{g_1, g_2, g_3\} \cap k[Y]$ ein Erzeugendensystem des Eliminationsideals $(f_1, f_2) \cap k[Y]$ liefert, und in der Tat liegt keines der g_i in $k[Y]$. Zufälligerweise liegt aber $g_1 = X^2$ in $k[X]$, wir wissen also, daß für jede Lösung (x, y) des Gleichungssystems $x = 0$ sein muß. $g_2 = XY$ verschwindet für alle solche Punkte automatisch, und $g_3 = Y^2 - X/2$ verschwindet genau dann, wenn auch $y = 0$ ist. Somit ist $V_K(f_1, f_2) = \{(0,0)\}$ für jeden Erweiterungskörper K von k.

Wenn wir das Gleichungssystem mit dem hier vorgestellten Verfahren lösen wollen. können wir zum Beispiel mit der lexikographischen Ordnung arbeiten. Da die führenden Terme von f_1 und f_2 bei beiden Ordnungen gleich sind und viele der zu berechnenden S-Polynome nur aus einem Term bestehen, ändert sich zunächst nichts: Wie bei der graduiert lexikographischen Ordnung kommen wir auf

$$f_3 = S(f_1, f_2) = -X^2, \quad f_4 = S(f_1, f_3) = -2XY \quad \text{und}$$
$$f_5 = S(f_2, f_3) = X - 2Y^2\,.$$

Auch $S(f_1, f_4) = -2XY^2 = Yf_4$ kann wie dort auf Null reduziert werden, bei der Berechnung von $S(f_1, f_5)$ ist jetzt aber nicht mehr Y^2, sondern X das führende Monom. Somit ist

$$S(f_1, f_5) = f_1 - X^2 f_5 = 2X^2Y^2 - 2XY = 2Yf_2 + 2f_4 + 4Y^3 ,$$

das S-Polynom läßt sich also modulo $\{f_1, f_2, f_3, f_4, f_5\}$ nicht auf Null reduzieren und wir müssen $f_6 = 4Y^3$ als neues Element in die Basis aufnehmen. Erst jetzt zeigt eine mühsame Rechnung, die man am besten seinem Computer überläßt, daß $S(f_i, f_j)$ für alle $1 \leq i < j \leq 6$ modulo $\{f_1, f_2, f_3, f_4, f_5, f_6\}$ auf Null reduziert werden kann, womit wir eine GRÖBNER-Basis gefunden haben.

Die führenden Monome der sechs Basiselemente bezüglich der lexikographischen Ordnung sind

$$\begin{array}{lll} \mathrm{FM}(f_1) = X^3 , & \mathrm{FM}(f_2) = X^2Y , & \mathrm{FM}(f_3) = -X^2 , \\ \mathrm{FM}(f_4) = -2XY , & \mathrm{FM}(f_5) = X, , & \mathrm{FM}(f_6) = 4Y^3 ; \end{array}$$

wir können also f_1 bis f_4 eliminieren. Die reduzierte GRÖBNER-Basis bedeutet besteht somit aus $g_1 = X - 2Y^2$ und $g_2 = Y^3$.

Das Eliminationsideal I_1 wird daher erzeugt von $g_2 = Y^3$, d.h. für jede Lösung (x, y) muß y verschwinden. Setzen wir $y = 0$ in g_1 ein, so sehen wir, daß auch x verschwinden muß, der Nullpunkt ist also die einzige Lösung.

Es war ein Zufall, daß wir dieses Ergebnis auch der GRÖBNER-Basis bezüglich der graduiert lexikographischen Ordnung ansehen konnten; bei komplizierteren Systemen wird dort oft jedes Basiselement alle Variablen enthalten, so daß wir nichts sehen können. Trotzdem kann die graduiert lexikographische Ordnung zur Lösung nichtlinearer Gleichungssysteme nützlich sein: 1993 publizierten J.C. FAUGÈRE, P. GIANNI, D. LAZARD und T. MORA einen heute nach ihren Anfangsbuchstaben als FGLM benannten Algorithmus, der für ein Ideal I mit endlicher Nullstellenmenge $V(I)$ effizient eine GRÖBNER-Basis bezüglich der lexikographischen Ordnung bestimmt auf dem Umweg über die graduiert lexikographische Ordnung. Wir werden später sehen, daß wir im Falle einer endlichen Lösungsmenge diese auch ausgehend von einer beliebigen GRÖBNER-Basis mit alternativen Techniken bestimmen können.

Nun kann es beim obigen Verfahren für nichtlineare Gleichungssysteme natürlich vorkommen, daß I_{n-1} das Nullideal ist; falls unter den Lösungen des Systems unendlich viele Werte für die letzte Variable vorkommen, muß das sogar so sein. Es kann sogar vorkommen, daß *alle* Eliminationsideale außer $I_0 = I$ das Nullideal sind. In diesem Fall führt die gerade skizzierte Vorgehensweise zu nichts.

Bevor wir uns darüber wundern, sollten wir uns überlegen, was wir überhaupt unter der Lösung eines nichtlinearen Gleichungssystems verstehen wollen. Im Falle einer endlichen Lösungsmenge ist das klar: Dann wollen wir eine Auflistung der sämtlichen Lösungstupel. Bei einer unendlichen Lösungsmenge ist das aber nicht mehr möglich. Im Falle eines linearen Gleichungssystems wissen wir, daß die Lösungsmenge ein affiner Raum ist; wir können sie daher auch wenn sie unendlich sein sollte durch endlich viele Daten eindeutig beschreiben, zum Beispiel durch eine spezielle Lösung und eine Basis des Lösungsraums des zugehörigen homogenen Gleichungssystems.

Bei nichtlinearen Gleichungssystemen gibt es im allgemeinen keine solche Beschreibung unendlicher Lösungsmengen: Die Lösungsmenge des Gleichungssystems

$$X^2 + 2Y^2 + 3Z^2 = 100 \quad \text{und} \quad 2X^2 + 3Y^2 - Z^2 = 0$$

etwa ist die Schnittmenge eines Ellipsoids mit einem elliptischen Kegel; sie besteht aus zwei ovalen Kurven höherer Ordnung. Die GRÖBNER-Basis besteht in diesem Fall aus den beiden Polynomen

$$X^2 - 11Z^2 + 300 \quad \text{und} \quad Y^2 + 7Z^2 - 200\,,$$

stellt uns dieselbe Menge also dar als Schnitt eines hyperbolischen und eines elliptischen Zylinders. Eine explizitere Beschreibung der Lösungsmenge ist schwer vorstellbar.

Die semialgebraische Geometrie hat Methoden entwickelt, wie man auch allgemeinere Lösungsmengen nichtlinearer Gleichungssysteme über $\mathbb{R}$ oder einem Teilkörper davon durch eine sogenannte zylindrische Zerlegung qualitativ beschreiben kann. Dazu wird der $\mathbb{R}^n$ in Teilmengen zerlegt, in denen die Lösungsmenge entweder ein einfaches qualitatives

Verhalten hat oder aber leeren Durchschnitt mit der Teilmenge. Dadurch kann man insbesondere feststellen, in welchen Regionen des $\mathbb{R}^n$ Lösungen zu finden sind.

In manchen Fällen lassen sich Lösungsmengen parametrisieren; wie man mit Methoden der algebraischen Geometrie zeigen kann, ist das aber im allgemeinen nur bei Gleichungen kleinen Grades der Fall und kommt daher für allgemeine Lösungsalgorithmen nicht in Frage.

Stets möglich ist das umgekehrte Problem, d.h. die Beschreibung einer parametrisch gegebenen Menge in impliziter Form. Hier gehen wir aus von Gleichungen der Form

$$x_1 = \varphi_1(t_1, \ldots, t_m), \quad \ldots, \quad x_n = \varphi_n(t_1, \ldots, t_m)\,,$$

und wir suchen Polynome $f_1, \ldots, f_r$ aus $k[X_1, \ldots, X_n]$, die auf der Menge aller jener $(x_1, \ldots, x_n)$ verschwinden, für die es eine solche Darstellung gibt (und eventuell noch auf Grenzwerten davon).

Dazu wählen wir eine lexikographische Ordnung auf dem Polynomring $k[T_1, \ldots, T_m, X_1, \ldots, X_n]$, bei der alle T_i größer sind als die X_j, und bestimmen eine GRÖBNER-Basis für das von den Polynomen $X_i - \varphi_i(T_1, \ldots, T_m)$ erzeugte Ideal. Dessen Schnitt mit $k[X_1, \ldots, X_n]$ ist ein Eliminationsideal, hat also als Basis genau die Polynome aus der GRÖBNER-Basis, in denen keine T_i vorkommen.

Fast genauso können wir auch zu einer vorgegebenen endlichen Menge von Punkten ein Gleichungssystem konstruieren, das genau diese Menge als Lösungsmenge hat; dies spielt beispielsweise in der algebraischen Statistik eine Rolle, wenn zu einem vorgegebenen Design die damit schätzbaren Modelle identifiziert werden sollen.

Wir gehen aus von r Punkten

$$P_i = \left(x_1^{(i)}, \ldots, x_n^{(i)}\right) \in k^n, \quad i = 1, \ldots, r,$$

und suchen ein Ideal $I \lhd k[X_1, \ldots, X_n]$, dessen Elemente genau in den Punkten P_i verschwinden. Im Falle nur eines Punktes P_i können wir einfach das Ideal

$$I_i = \left(X_1 - x_1^{(i)}, \ldots, X_n - x_n^{(i)}\right)$$

nehmen; bei mehreren Punkten brauchen wir den Durchschnitt der Ideale I_1 bis I_r, für den wir kein offensichtliches Erzeugendensystem haben.

Betrachten wir stattdessen die Punkte

$$Q_i = \left(t_1^{(i)}, \ldots, t_r^{(i)}, x_1^{(i)}, \ldots, x_n^{(i)}\right) \in k^{r+n} \quad \text{mit} \quad t_j^{(i)} = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{sonst} \end{cases},$$

so erzeugen die Polynome

$$\left(X_j - x_j^{(i)}\right) T_i \in k[T_1, \ldots, T_r, X_1, \ldots, X_n]$$

für $i = 1, \ldots, n$ und $j = 1, \ldots, r$ zusammen mit dem Polynom $T_1 + \cdots + T_r - 1$ ein Ideal J, das alle Punkte Q_i als Nullstellen hat: Die Polynome $\left(X_j - x_j^{(i)}\right) T_i$ verschwinden in Q_i, da $x_j^{(i)}$ die j-te Koordinate von Q_i ist, und für $\ell \neq i$ verschwindet $\left(X_j - x_j^{(i)}\right) T_\ell$, da $t_\ell^{(i)}$ verschwindet.

Ist umgekehrt $Q = (t_1, \ldots, t_r, x_1, \ldots, x_n) \in k^{r+n}$ keiner der Punkte Q_i, so gibt es für jedes i mindestens eine Koordinate, in der sich Q von Q_i unterscheidet. Ist dies etwa die X_j-Koordinate, so ist $X_j - x_j^{(i)}$ in Q in Q von Null verschieden; $\left(X_j - x_j^{(i)}\right) T_i$ kann daher nur verschwinden, wenn $t_i = 0$ ist. Dies kann aber nicht für alle i der Fall sein, denn die Summe der t_i ist eins, da $T_1 + \cdots + T_r - 1$ verschwindet. Somit liegt Q nicht in $V(J)$.

Damit haben wir ein Ideal $J \lhd k[T_1, \ldots, T_r, X_1, \ldots, X_n]$ gefunden, dessen Nullstellen genau die Punkte $Q_1, \ldots, Q_r \in k^{r+n}$ sind. Die Punkte $P_1, \ldots, P_r$ sind die Projektionen der Q_i von k^{n+r} nach k^n; deshalb ist klar, daß alle Polynome aus

$$I \underset{\text{def}}{=} J \cap k[X_1, \ldots, X_n]$$

in den Punkten P_i verschwinden. Wir erhalten ein Erzeugendensystem dieses Ideals, indem wir bezüglich einer Eliminationsordnung für $T_1, \ldots, T_r$ eine GRÖBNER-Basis von J berechnen und davon nur die Polynome betrachten, die keine der Variablen T_i enthalten.

§2: Der Hilbertsche Nullstellensatz

Wie wir wissen, stimmen die Lösungsmengen zweier Gleichungssysteme

$$f_1(x_1, \ldots, x_n) = \cdots = f_m(x_1, \ldots, x_n) = 0$$

und

$$g_1(x_1, \ldots, x_n) = \cdots = g_p(x_1, \ldots, x_n) = 0$$

überein, wenn die Ideale $(f_1, \ldots, f_m)$ und $(g_1, \ldots, g_p)$ übereinstimmen. Umgekehrt folgt aber nicht aus der Gleichheit der Lösungsmengen, daß auch die Ideale gleich sein müssen. In diesem Paragraphen wollen wir genauer untersuchen, was hier gilt.

Dazu betrachten wir als erstes den Fall von Polynomen in nur einer Veränderlichen X. Hier können wir uns mit einer einzigen Gleichung begnügen, denn es gilt

Lemma: Der Polynomring $R = k[X]$ über einem Körper k ist ein Hauptidealring.

Beweis: Wir müssen zeigen, daß jedes Ideal I von R ein Hauptideal ist, also von einem einzigen Polynom f erzeugt werden kann. Sei also I ein beliebiges Ideal in R. Falls I nur aus der Null besteht, ist es das von der Null erzeugte Hauptideal, andernfalls wählen wir ein Polynom f minimalen Grades aus I und wollen zeigen, daß $I = (f)$ ist. Dazu sei g ein beliebiges Polynom aus I. Wir dividieren es durch f:

$$g : f = q \text{ Rest } r \quad \text{oder} \quad g = qf + r\,,$$

wobei entweder $r = 0$ ist oder $\deg r < \deg f$. Da mit f und g auch $r = g - qf$ in I liegt, kann letzteres nicht sein: Nach Konstruktion enthält I kein Polynom vom Grad kleiner $\deg f$. Also ist $r = 0$ und $g = qf$ liegt in (f).

∎

Sind also I und J zwei Ideale in $k[X]$, so gibt es Polynome $f, g \in k[X]$, für die $I = (f)$ und $J = (g)$ ist. Da für jedes $c \in k \setminus \{0\}$ die Polynome f und cf dasselbe Ideal erzeugen, können wir dabei annehmen, daß sowohl f als auch g den führenden Koeffizienten eins haben. Dann ist $I = J$ genau dann, wenn $f = g$ ist.

Um zu sehen, was es bedeutet, daß $V_K(I) = V_K(J)$ ist, betrachten wir zunächst den Fall, daß $k = K = \mathbb{Q}$ ist. Offensichtlich ist dann

$$V_{\mathbb{Q}}(X^2 - 2) = V_{\mathbb{Q}}(X^2 - 3) = V_{\mathbb{Q}}(X^2 + 1) = V_{\mathbb{Q}}(X^2 + 5) = \emptyset\,,$$

ohne daß irgendwelche zwei der betrachteten Polynome gleich wären. Es ist dabei unerheblich, daß die Nullstellenmengen jeweils leer sind: Hätten wir jedes der vier betrachteten Polynome noch mit $X - 1$ multipliziert, hätten wir vier neue Polynome erhalten, die allesamt die Eins als einzige rationale Nullstelle haben und trotzdem verschieden sind. Über einem hinreichend großen Körper, etwa dem der kompklexen Zahlen, haben freilich alle betrachteten Polynome verschiedene Nullstellenmengen.

Aber auch über $\mathbb{C}$ gilt nicht, daß aus $V_{\mathbb{C}}(f) = V_{\mathbb{C}}(g)$ die Gleichheit der Ideale (f) und (g) folgt: Beispielsweise ist

$$V_{\mathbb{C}}\big(X(X-1)^2\big) = V_{\mathbb{C}}\big(X^2(X-1)\big) = \{0,1\}\,,$$

aber keines der beiden Polynome liegt auch nur im vom anderen erzeugten Ideal. Allgemein ist offenbar für r komplexe Zahlen $z_1, \ldots, z_r$ und r natürliche Zahlen $e_1, \ldots, e_r$ stets

$$V_{\mathbb{C}}\Big((X-z_1)^{e_1} \cdots (X-z_r)^{e^r}\Big) = V_{\mathbb{C}}\big((X-z_1)\cdots(X-z_r)\big)\,,$$

und sofern nicht alle $e_i = 1$ sind, erzeugen die beiden Polynome verschiedene Ideale.

Nach dem sogenannten *Fundamentalsatz der Algebra* hat jedes nichtkonstante Polynom f mit komplexen Koeffizienten mindestens eine komplexe Nullstelle. Da wir aber algorithmisch rechnen wollen, eignet sich der Körper der komplexen Zahlen aber nicht für uns, denn er ist überabzählbar, und wie bereits in der Einführung erwähnt, zeigte RICHARDSON, daß schon in $\mathbb{R}$, erst recht also in $\mathbb{C}$, algorithmisch nicht entschieden werden kann, ob zwei Ausdrücke dieselbe Zahl bezeichnen. Deshalb müssen wir auch kleinere Körper betrachten und definieren:

Definition: Ein Körper K heißt *algebraisch abgeschlossen,* wenn jedes nichtkonstante Polynom $f \in K[X]$ mindestens eine Nullstelle in K hat.

Durch Polynomdivision folgt leicht induktiv:

Lemma: Ist K algebraisch abgeschlossen, so läßt sich jedes Polynom vom Grad d aus $K[X]$ schreiben als

$$f = c(X-x_1)\cdots(X-x_d) \quad \text{mit} \quad c \in K \setminus \{0\} \ \text{ und } \ x_1, \ldots, x_d \in K\,.$$

Die x_i müssen dabei nicht notwendigerweise verschieden sein. ∎

Im folgenden betrachten wir Polynome über einem beliebigen Körper k; in den Beispielen wird das fast immer der Körper $\mathbb{Q}$ sein. Zusätzlich betrachten wir einen Körper K, der k enthält. Im Falle von $\mathbb{Q}$ kann man zeigen, daß es einen abzählbaren algebraisch abgeschlossenen Körper $\overline{\mathbb{Q}} \subset \mathbb{C}$ gibt, der $\mathbb{Q}$ enthält; er besteht aus allen *algebraischen* Zahlen, d.h. allen komplexen Zahlen z, für die es ein Polynom $0 \neq f \in \mathbb{Q}[X]$ gibt mit $f(z) = 0$. Die semialgebraische Geometrie hat Algorithmen entwickelt, mit denen man in diesem Körper algorithmisch rechnen und entscheiden kann, wann zwei Ausdrücke gleich sind.

Um die Beweise der folgenden Sätze etwas zu vereinfachen, wollen wir dort allerdings zusätzlich annehmen, daß der Körper K überabzählbar viele Elemente enthält. Diese Annahme ist aber nicht notwendig für die Gültigkeit der Sätze; mit etwas größerem Aufwand lassen sich die Beweise so führen, daß sie für beliebige algebraisch abgeschlossene Körper gelten.

Sei also für den Rest dieses Paragraphen k irgendein Körper, und K sei ein algebraisch abgeschlossener Körper mit überabzählbar vielen Elementen, der k enthält.

Als erstes wollen wir uns mit der Frage beschäftigen, für welche Ideale $I \lhd k[X_1, \ldots, X_n]$ die Lösungsmenge $V_K(I)$ in K^n leer ist. Ein Beispiel ist offensichtlich: Natürlich ist $I = k[X_1, \ldots, X_n]$ ein Ideal, und da es insbesondere die Konstante eins enthält, ist $V_K(I) = \emptyset$. Eine (schwache) Form des HILBERTschen Nullstellensatzes besagt, daß dies das einzige Beispiel ist. Zur Vorbereitung des Beweises definieren wir

Definition: R sei ein Ring.
a) $I \lhd R$ ist ein *echtes* Ideal, falls $I \neq R$.
b) Ein echtes Ideal $\mathfrak{m} \lhd R$ heißt *maximales* Ideal, wenn R das einzige Ideal ist, das $\mathfrak{m}$ als echte Teilmenge enthält.
c) Ein echtes Ideal $\mathfrak{p} \lhd R$ heißt *Primideal,* wenn gilt: Liegt für zwei Elemente $f, g \in R$ das Produkt fg in $\mathfrak{p}$, so liegt mindestens einer der Faktoren f, g in $\mathfrak{p}$.

Wie aus der Zahlentheorie bekannt, teilt eine Primzahl p genau dann das Produkt zweier Zahlen a, b, wenn sie mindestens einen der beiden Faktoren teilt; in $\mathbb{Z}$ sind also die von den Primzahlen erzeugten Hauptideale Primideale. Dazu kommt wegen der Nullteilerfreiheit auch noch das Nullideal.

Durch vollständige Induktion beweist man leicht

Lemma: Ist $\mathfrak{p}$ ein Primideal und liegt ein Produkt $f_1 \cdots f_n$ von Elementen $f_i \in R$ in $\mathfrak{p}$, so liegt mindestens einer der Faktoren f_i in $\mathfrak{p}$. ∎

Lemma: Jedes maximale Ideal $\mathfrak{m} \lhd R$ ist ein Primideal.

Beweis: Das Produkt fg zweier Elemente $f, g \in R$ liege in $\mathfrak{m}$. Wir müssen zeigen, daß mindestens einer der Faktoren f, g in $\mathfrak{m}$ liegt. Im Falle $f \in \mathfrak{m}$ sind wir fertig; andernfalls ist $\mathfrak{m} + (f) = R$ wegen der Maximalität von $\mathfrak{m}$. Es gibt daher Elemente $m \in \mathfrak{m}$ und $h \in R$, so daß $m + hf = 1$ ist. Damit ist $g = mg + hfg \in \mathfrak{m}$, denn $m \in \mathfrak{m}$ und $fg \in \mathfrak{m}$. ∎

Lemma: Jedes echte Ideal $I \lhd k[X_1, \ldots, X_n]$ liegt in einem maximalen Ideal $\mathfrak{m} \lhd k[X_1, \ldots, X_n]$.

Beweis: Falls I selbst maximal ist, sind wir fertig; andernfalls gibt es ein echtes Ideal I_1, das I als echte Teilmenge enthält. Auch wenn I_1 ein maximales Ideal ist, sind wir fertig; andernfalls gibt es ein echtes Ideal I_2, das I_1 als echte Teilmenge enthält, und so weiter. Wenn dieses Verfahren nach endlich vielen Schritten abbricht, haben wir ein maximales Ideal gefunden, das I enthält. Andernfalls gibt es eine unendliche aufsteigende Folge von Idealen $I \subset I_1 \subset I_2 \subset \cdots$. Die Vereinigung aller I_j ist selbst ein Ideal in $k[X_1, \ldots, X_n]$ und hat damit nach dem HILBERTschen Basissatz ein endliches Erzeugendensystem $\{f_1, \ldots, f_m\}$. Jedes f_i liegt in einem der Ideale I_j und damit auch in allen I_ℓ mit $\ell > j$. Wegen der Endlichkeit des Erzeugendensystems gibt es daher einen Index r derart, daß alle f_i in I_r liegen. Dann ist aber $I = I_r = I_{r+1} = \cdots$, im Widerspruch zu der Annahme, daß jedes I_j echte Teilmenge von I_{j+1} ist. Somit bricht das Verfahren nach endlich vielen Schritten ab und liefert ein maximales Ideal $\mathfrak{m}$, in dem I enthalten ist. ∎

(Tatsächlich gilt auch dieses Lemma für beliebige Ringe; da dort der HILBERTsche Basissatz nicht gelten muß, beweist man es für den allgemeinen Fall mit Hilfe des ZORNschen Lemmas.)

Für ein Ideal I eines Rings R können wir eine Äquivalenzrelation $\sim$ auf R definieren durch

$$f \sim g \Longleftrightarrow f - g \in I\,.$$

Die Menge der Äquivalenzhlassen bezeichnen wir als den *Faktorring* R/I, und die Äquivalenzklasse eines Elements $f \in R$ mit $f + I$. Man überlegt sich leicht, daß R/I wirklich ein Ring ist, daß also insbesondere im Fall $f+I = f'+I$ und $g+I = g'+I$ auch $(f+g)+I = (f'+g')+I$ ist und $fg + I = f'g' + I$.

Speziell für $R = k[X_1, \ldots, X_n]$ ist R auch ein k-Vektorraum, und auch jedes Ideal $I \lhd R$ ist ein solcher. In diesem Fall ist R/I als Vektorraum einfach der Faktorraum dieser beiden Vektorräume. Wir können dann also insbesondere auch von der k-Dimension $\dim_k R/I$ reden. Diese wird im folgenden häufiger eine Rolle spielen.

Ist etwa $R = k[X]$ und $I = (f)$ für ein Polynom f vom Grad $d > 0$, so ist X^d modulo f äquivalent zu einem Polynom vom Grad höchstens $d - 1$ in X, so daß die Restklassen der Eins und der Potenzen X^e mit $1 \leq e < d$ eine k-Vektorraumbasis von R/I bilden. Somit ist hier $\dim_k R/I = d$.

Für ein Ideal I des Polynomrings $R = k[X_1, \ldots, X_n]$ definieren die Elemente von R/I als Funktionen $V_K(I) \to K$, die jedem Punkt $(x_1, \ldots, x_n) \in V_K(I)$ das Körperelement $f(x_1, \ldots, x_n) \in K$ zuordnen, wobei f irgendein Element der Restklasse ist. Ist nämlich g ein anderes Element derselben Restklasse, so liegt $f - g$ in I, verschwindet also auf allen Elementen von $V_K(I)$, so daß die Werte von f und von g dort übereinstimmen. Da das Ideal I durch $V_K(I)$ nicht eindeutig bestimmt ist, wissen wir allerdings noch nicht, unter welchen Bedingungen wir wir R/I mit dem Ring aller (mengentheoretischer) Abbildungen $V_K(I) \to I$ identifizieren können. Einen ersten Schritt in diese Richtung geben die folgenden Sätze, die HILBERT in seiner 1893 erschienenen Arbeit *Ueber die vollen Invariantensysteme* (Mathematische Annalen **36**, S. 313–373) veröffentlicht hat, und die auch für viele andere Fragen

fundamental sind. Sie alle werden unter dem Namen *Hilbertscher Nullstellensatz* zusammengefaßt; eine erste Version ist die folgende:

Schwache Form des Hilbertschen Nullstellensatzes: Für ein echtes Ideal $I \lhd k[X_1, \ldots, X_n]$ ist $V_K(I) \neq \emptyset$.

Beweis: Nach dem HILBERTschen Basissatz hat jedes Ideal I ein endliches Erzeugendensystem $\{f_1, \ldots, f_m\}$. Wir betrachten das von den f_i erzeugte Ideal $\bar{I}$ in $K[X_1, \ldots, X_n]$. Da eine Basis des k-Vektorraums $k[X_1, \ldots, X_n]/I$ auch Basis des K-Vektorraums $K[X_1, \ldots, X_n]/\bar{I}$ ist, muß auch $\bar{I}$ ein echtes Ideal von $K[X_1, \ldots, X_n]$ sein und liegt somit in einem maximalen Ideal $\mathfrak{m} \lhd K[X_1, \ldots, K_n]$. Der Satz folgt daher aus der folgenden alternativen Version des HILBERTschen Nullstellensatzes:

Satz: Die maximalen Ideale $\mathfrak{m} \lhd K[X_1, \ldots, X_n]$ sind genau die Ideale

$$\mathfrak{m} = (X_1 - x_1, \ldots, X_n - x_n) \quad \text{mit} \quad (x_1, \ldots, x_n) \in K^n \, .$$

Beweis: $\mathfrak{m} = (X_1 - x_1, \ldots, X_n - x_n)$ ist der Kern der Abbildung

$$\begin{cases} K[X_1, \ldots, X_n] \to K \\ \qquad\qquad f \mapsto f(x_1, \ldots, x_n) \end{cases} .$$

Ist daher I ein Ideal, das $\mathfrak{m}$ echt enthält, so muß der Vektorraum $K[X_1, \ldots, X_n]/I$ ein echter Untervektorraum von $K[X_1, \ldots, X_n]/\mathfrak{m}$ sein. Da letzterer nach dem Homomorphiesatz isomorph zum eindimensionalen Vektorraum K ist, muß dies der Nullraum sein. Somit ist $I = K[X_1, \ldots, X_m]$, d.h. $\mathfrak{m}$ ist ein maximales Ideal.

Umgekehrt sei $\mathfrak{m}$ ein maximales Ideal. Wenn wir zeigen können, daß es Elemente $x_1, \ldots, x_n$ gibt, für die $X_i - x_i$ in $\mathfrak{m}$ liegt, ist $(X_1 - x_1, \ldots, X_n - x_n) \subseteq \mathfrak{m}$, und da links ein maximales Ideal steht, müssen beide Seiten gleich sein.

Angenommen, es gibt ein $i \in \{1, \ldots, n\}$, für das $X_i - x$ für kein $x \in K$ im Ideal $\mathfrak{m}$ liegt. Wegen der Maximalität von $\mathfrak{m}$ ist dann

$$\mathfrak{m} + (X_i - x) = K[X_1, \ldots, X_n] \quad \text{für alle } x \in K \, .$$

Somit gibt es für jedes $x \in K$ ein Polynom $f_x \in \mathfrak{m}$ sowie ein Polynom $h_x \in K[X_1, \ldots, X_n]$ derart, daß $f_x + h_x \cdot (X_i - x) = 1$ ist. Da 1 nicht

in $\mathfrak{m}$ liegt, ist $h_x \neq 0$. Wir wählen für jedes $x \in K$ ein festes Polynom h_x (und damit auch f_x), das obige Gleichung erfüllt, und setzen $K_d = \{x \in K \mid \deg h_x = d\}$ für jedes $d \in \mathbb{N}_0$. Da K nach Voraussetzung überabzählbar viele Elemente enthält und K die Vereinigung der K_d ist, muß mindestens eine der Mengen K_d unendlich viele Elemente enthalten. (Nur an dieser Stelle geht die Voraussetzung der Überabzählbarkeit ein, und wie bereits erwähnt, gibt es alternative Beweise, die ohne diese Voraussetzung auskommen.)

Wir wählen eine solche Menge K_d und betrachten den Vektorraum $K[X_1, \ldots, K_n]_d$ aller Polynome vom Grad höchstens d. Da es nur endlich viele Monome vom Grad höchstens d gibt, ist dies ein endlichdimensionaler K-Vektorraum. Wir wählen eine natürliche Zahl r, die größer ist als seine Dimension, und dazu r Elemente $x^{(1)}, \ldots, x^{(r)} \in K$ mit $h_{x^{(i)}} \in k[X_1, \ldots, X_n]_d$. Zwischen diesen Polynomen muß dann eine kineare Abhängigkeit bestehen. Es gibt daher Elemente $\lambda_1, \ldots, \lambda_r \in K$, die nicht allesamt verschwinden, so daß

$$\lambda_1 h_{x^{(1)}} + \cdots + \lambda_r h_{x^{(r)}} = 0$$

ist. Dazu definieren wir

$$g = \sum_{j=1}^{r} \lambda_j \prod_{\ell \neq j} \left(X_i - x^{(\ell)}\right) \in K[X_i]\,.$$

Dieses Polynom liegt auch in $\mathfrak{m}$, denn wegen

$$1 = f_{x^{(j)}} + h_{x^{(j)}} \left(X_i - x^{(j)}\right) \quad \text{für } j = 1, \ldots, r$$

ist

$$\begin{aligned} g &= \sum_{j=1}^{r} \lambda_j \left(f_{x^{(j)}} + h_{x^{(j)}} \left(X_i - x^{(j)}\right)\right) \prod_{\ell \neq j} \left(X_i - x^{(\ell)}\right) \in K[X_i] \\ &= \sum_{j=1}^{r} \lambda_j f_{x^{(j)}} \prod_{\ell \neq j} \left(X_i - x^{(\ell)}\right) + \left(\sum_{j=1}^{r} \lambda_j h_{x^{(j)}}\right) \prod_{\ell=1}^{n} \left(X_i - x^{(\ell)}\right) \\ &= \sum_{j=1}^{r} \lambda_j \prod_{\ell \neq j} \left(X_i - x^{(\ell)}\right) f_{x^{(j)}} \in \mathfrak{m}\,, \end{aligned}$$

da $\sum_{j=1}^{r} \lambda_j h_{x^{(j)}}$ verschwindet und alle $f_{x^{(j)}}$ in $\mathfrak{m}$ liegen.

g ist nicht das Nullpolynom, denn für jeden Index ν ist

$$g\big(x^{(\nu)}\big) = \sum_{j=1}^{r} \lambda_j \prod_{\ell \neq j} \big(x^{(\nu)} - x^{(\ell)}\big) = \lambda_\nu \prod_{\ell \neq \nu} \big(x^{(\nu)} - x^{(\ell)}\big) \,.$$

Da die $x^{(\ell)}$ paarweise verschieden sind und mindestens ein λ_ν nicht verschwindet, muß mindestens einer dieser Werte von Null verschieden sein.

Da g in $\mathfrak{m}$ liegt, kann g auch keine von Null verschiedene Konstante sein, hat also einen positiven Grad e. Über dem algebraisch abgeschlossenen Körper K zerfällt g daher in Linearfaktoren:

$$g = c\,(X_i - z_1)\ldots(X_i - z_e) \quad \text{mit} \quad c \in K \setminus \{0\},\ z_1, \ldots z_e \in K \,.$$

g liegt in $\mathfrak{m}$, aber nach Voraussetzung liegt keiner der Faktoren $X_i - z_j$ in $\mathfrak{m}$, und die Konstante $c \neq 0$ natürlich auch nicht. Dies ist ein Widerspruch, denn als maximales Ideal ist $\mathfrak{m}$ insbesondere ein Primideal. ∎

Somit hat also jedes echte Ideal $I \lhd k[X_1, \ldots, X_n]$ zumindest in einem Erweiterungskörper K von k mindestens eine Nullstelle. Damit folgt umgekehrt

Satz: Das Gleichungssystem

$$f_1(x_1, \ldots, x_n) = \cdots = f_m(x_1, \ldots, x_n) = 0$$

mit $f_1, \ldots, f_m \in k[X_1, \ldots, X_n]$ ist genau dann in jedem Erweiterungskörper K von k unlösbar, wenn es Polynome $h_1, \ldots, h_m$ in $X_1, \ldots, X_n$ gibt, so daß $h_1 f_1 + \cdots + h_m f_m = 1$ ist.

Beweis: Im Falle der Unlösbarkeit ist das von $f_1, \ldots, f_m$ erzeugte Ideal der ganze Polynomring, enthält also insbesondere die Eins. Da

$$(f_1, \ldots, f_m) = \big\{ h_1 f_1 + \cdots + h_m f_m \;\big|\; h_1, \ldots, h_m \in k[X_1, \ldots, X_n] \big\} \,,$$

hat auch die Eins eine Darstellung der verlangten Form.

Ist umgekehrt $h_1 f_1 + \cdots + h_m f_m = 1$ für irgendwelche Polynome $h_1, \ldots, h_m$, so ist für jeden Erweiterungskörper K von k und jedes n-Tupel $(x_1, \ldots, x_n) \in K^n$

$$h_1(x_1, \ldots, x_n) f_1(x_1, \ldots, x_n) + \cdots + h_m(x_1, \ldots, x_n) f_m(x_1, \ldots, x_n) = 1 \,,$$

so daß nicht alle $f_j(x_1, \ldots, x_n)$ verschwinden können.

■

Wenn wir eine GRÖBNER-Basis eines Ideals I kennen, ist es einfach zu entscheiden, ob $I = k[X_1, \ldots, X_n]$ ist (oder äquivalent, ob $1 \in I$): Da der führende Term eines jeden Polynoms aus I durch den führenden Term eines Elements der GRÖBNER-Basis teilbar sein muß, enthält diese im Falle eines Ideals, das die Eins enthält, ein Polynom, dessen führendes Monom die Eins ist. Da diese bezüglich jeder Monomordnung das kleinste Monom ist, muß somit die GRÖBNER-Basis eine Konstante enthalten. Die zugehörige minimale und erst recht die reduzierte GRÖBNER-Basis besteht in diesem Fall nur aus der Eins.

Aus dem gerade bewiesenen Satz folgt mit einem 1929 von J.L. RABINOWITSCH gefundenen Trick die von HILBERT 1893 ab Seite 320 unten der zitierten Arbeit bereits anders bewiesene

Starke Form des Hilbertschen Nullstellensatzes: k sei ein beliebiger Körper und K ein überabzählbarer algebraisch abgeschlossener Erweiterungskörper von k. Falls für ein Ideal $I \lhd k[X_1, \ldots, X_n]$ ein Polynom $f \in k[X_1, \ldots, X_n]$ auf ganz $V_K(I)$ verschwindet, gibt es ein $q \in \mathbb{N}$, so daß f^q in I liegt.

Beweis: Wir erweitern den Polynomring $k[X_1, \ldots, X_n]$ mit einer neuen Variablen X_{n+1} zu $k[X_1, \ldots, X_{n+1}]$ und betrachten dort für ein Erzeugendensystem $\{f_1, \ldots, f_m\}$ von I das Gleichungssystem

$$f_1(x_1, \ldots, x_n) = \cdots = f_m(x_1, \ldots, x_n) = 1 - x_{n+1} f(x_1, \ldots, x_n) = 0 \,.$$

Für jeden Punkt $(x_1, \ldots, x_n, x_{n+1}) \in K^{n+1}$, für den die $f_j(x_1, \ldots, x_n)$ verschwinden, verschwindet auch $f(x_1, \ldots, x_n)$, d.h.

$$1 - x_{n+1} f(x_1, \ldots, x_n) = 1 \,.$$

Somit haben diese $n+1$ Gleichungen keine gemeinsame Nullstelle. Wie wir gerade gesehen haben, gibt es dann Polynome $h_1, \ldots, h_{m+1} \in k[X_1, \ldots, X_{n+1}]$ derart, daß

$$h_1 f_1 + \cdots + h_m f_m + h_{m+1}(1 - X_{n+1} f) = 1$$

ist. Diese Gleichung bleibt gültig, wenn wir überall für X_{n+1} ein Polynom oder eine rationale Funktion in $X_1, \ldots, X_n$ einsetzen; wir setzen $X_{n+1} = 1/f$. Die h_j werden dann zu rationalen Funktionen in $X_1, \ldots, X_n$, wobei alle Nenner Potenzen von f sind. Ist f^q die höchste dieser Potenzen, so erhalten wir nach Multiplikation mit f^q eine Gleichung der Form

$$\widetilde{h}_1 f_1 + \cdots + \widetilde{h}_m f_m = f^q$$

mit $\widetilde{h}_j = f^q h_j\big(X_1, \ldots, X_n, 1/f\big) \in k[X_1, \ldots, X_n]$. Dies zeigt, daß f^q in $I = (f_1, \ldots, f_m)$ liegt.

∎

Definition: R sei ein Ring und $I \lhd R$ ein Ideal von R. Das *Radikal* von I ist die Menge

$$\sqrt{I} \underset{\text{def}}{=} \left\{ f \in R \;\middle|\; \exists q \in \mathbb{N} : f^q \in I \right\}.$$

Ist $I = \sqrt{I}$, so bezeichnen wir I als ein *Radikalideal.*

Das Radikal besteht also aus allen Ringelementen, die eine Potenz in I haben. Es ist selbst ein Ideal, denn sind $f, g \in \sqrt{I}$ zwei Elemente mit $f^p \in I$ und $g^q \in I$, so sind in

$$(f+g)^{p+q} = \sum_{\ell=0}^{p+q} \binom{p+q}{\ell} f^{p+q-\ell} g^{\ell}$$

die ersten q Summanden Vielfache von f^p, und die restlichen p sind Vielfache von g^q. Somit liegt jeder Summand in I, also auch die Summe. Für ein beliebiges $r \in R$ liegt natürlich auch rf in $\sqrt{I}$, denn seine q-te Potenz $(rf)^q = r^q f^q$ liegt in I, sobald f^q in I liegt.

Mit diesem neuen Begriff können wir den obigen Satz umformulieren:

Satz: Ein Polynom $f \in k[X_1, \ldots, X_n]$ verschwindet genau dann auf $V_K(I)$, wenn $f \in \sqrt{I}$.

∎

Anders ausgedrückt heißt dies

Satz: Für zwei Ideale $I, J \lhd k[X_1, \ldots, X_n]$ ist $V_K(I) = V_K(J)$ genau dann, wenn $\sqrt{I} = \sqrt{J}$ ist.

■

Falls ein Ideal mit seinem Radikal übereinstimmt, enthält es *alle* Polynome, die auf $V_K(I)$ verschwinden; zwei Polynome nehmen genau dann in jedem Punkt von $V_K(I)$ denselben Wert an, wenn ihre Differenz in I liegt, wenn sie also modulo I dieselbe Restklasse definieren.

Wenn das Ideal I nicht mit seinem Radikal übereinstimmt, gilt zwar nicht mehr *genau dann,* aber wir können trotzdem die Elemente des Faktorvektorraums $A = k[X_1, \ldots, X_n]/I$ auffassen als Funktionen von $V_K(I)$ nach K: Für jede Restklasse und jeden Punkt aus $V_K(I)$ nehmen wir einfach irgendein Polynom aus der Restklasse und setzen die Koordinaten des Punkts ein. Da die Differenz zweier Polynome aus derselben Restklasse in I liegt, wird sie nach Einsetzen des Punktes zu Null, der Wert hängt also nicht ab von der Wahl des Polynoms. Auch Polynome aus $K[X_1, \ldots, X_n]$ definieren in dieser Weise Funktionen $V_K(I) \to K$; hinreichend (aber nicht notwendig) dafür, daß zwei Polynome dieselbe Funktion definieren ist, daß ihre Differenz im von I erzeugten Ideal $\bar{I} \lhd K[X_1, \ldots, X_n]$ liegt.

Im Falle von Polynomen einer Veränderlichen ist jedes Ideal von $k[X]$ ein Hauptideal. Ist $I = (f)$ mit einem Polynom $f \neq 0$ vom Grad d, so können wir die Restklassen repräsentieren durch die Polynome vom Grad höchstens $d - 1$, denn jedes Polynom $g \in k[X]$ hat dieselbe Restklasse wie sein Divisionsrest bei der Polynomdivision durch f. Somit ist $A = k[X]/I$ in diesem Fall ein d-dimensionaler Vektorraum. Da $V_K(I)$ gerade aus den Nullstellen von f in K besteht, von denen es höchstens d verschiedene gibt, liefert die Dimension von A eine obere Schranke für die Elementanzahl von $V_K(I)$; wenn wir die Nullstellen mit ihrer Vielfachheit zählen, ist die Dimension von A sogar *gleich* der Gesamtzahl der Nullstellen. Im nächsten Paragraphen wollen wir uns überlegen, wie man ähnliche Ergebnisse auch für Systeme von Polynomgleichungen in mehreren Veränderlichen finden kann.

§3: Gleichungssysteme mit endlicher Lösungsmenge

Auch hier gehen wir wieder aus von einem beliebigen Körper k sowie einem algebraisch abgeschlossenen Erweiterungskörper K mit überabzählbar vielen Elementen. Letztere Bedingung ist nur notwendig, weil wir sie im Beweis des HILBERTschen Nullstellensatzes verwendet haben; wie bereits dort erwähnt, gibt es auch Beweise für den Fall, daß K ein beliebiger algebraisch abgeschlossener Körper ist, so daß alle Sätze dieses Paragraphen tatsächlich auch ohne die Voraussetzung der Überabzählbarkeit von K gelten.

Satz: I sei ein Ideal im Polynomring $k[X_1, \ldots, X_n]$ über dem Körper k, und K sei ein überabzählbarer algebraisch abgeschlossener Körper, in dem k enthalten sei. Dann gilt: $V_K(I)$ ist genau dann endlich, wenn der Faktorring $A = k[X_1, \ldots, X_n]/I$ ein endlichdimensionaler k-Vektorraum ist. In diesem Fall ist die Dimension von A eine obere Schranke für die Elementanzahl von $V_K(I)$.

Den recht umfangreichen *Beweis* führen wir in mehreren Schritten:

1. Schritt: Wenn der Vektorraum A endliche Dimension hat, ist $V_K(I)$ endlich.

Ist $d = \dim_k A$ so sind für jedes i die Potenzen $1, X_i, \ldots, X_i^d$ modulo I linear abhängig. Für jedes i liegt daher ein Polynom f_i vom Grad kleiner d aus $k[X_i]$ in I. Für jeden Punkt $x \in V_K(I)$ muß daher die i-te Koordinate x_i eine Nullstelle von f_i sein. Damit kann die i-te Koordinate höchstens $d-1$ verschiedene Werte annehmen, so daß $V_K(I)$ höchstens $(d-1)^n$ Elemente hat.

2. Schritt: $\bar{I}$ sei das von I in $K[X_1, \ldots, X_n]$ erzeugte Ideal. Wenn $V_K(I)$ endlich ist, hat der K-Vektorraum $\bar{A} = K[X_1, \ldots, X_n]/\bar{I}$ endliche Dimension.

Besteht $V_K(I)$ nur aus endlich vielen Punkten, so nimmt jede der Koordinatenfunktionen $X_1, \ldots, X_n$ auf $V_K(I)$ nur endlich viele Werte an; es gibt also für jedes i ein Polynom aus $K[X_i]$, das auf ganz $V_K(I)$ verschwindet. Nach dem HILBERTschen Nullstellensatz muß eine Potenz dieses Polynoms in $\bar{I}$ liegen, es gibt also auch in $\bar{I}$ für jedes i ein Polynom nur in X_i. Somit gibt es einen Grad d_i derart, daß sich X_i^e für $e \geq d_i$

modulo $\bar{I}$ durch die endlich vielen X_i-Potenzen $1, X_i, \ldots, X_i^{d_i-1}$ ausdrücken läßt. Damit läßt sich auch jedes Monom aus $K[X_1, \ldots, X_n]$ modulo $\bar{I}$ durch jene Monome ausdrücken, bei denen jede Variable X_i höchstens mit Exponent $d_i - 1$ auftritt. Da es nur endlich viele solche Monome gibt, ist $K[X_1, \ldots, X_n]/\bar{I}$ ein endlichdimensionaler K-Vektorraum.

3. Schritt: A ist genau dann endlichdimensional, wenn $\bar{A}$ endlichdimensional ist; in diesem Fall haben beide dieselbe Dimension.

Ist A endlichdimensional, so wählen wir eine Basis $\{b_1, \ldots, b_r\}$ und zu jedem Basiselement b_i ein Polynom $B_i \in k[X_1, \ldots, X_n]$, das modulo I gleich b_i ist. Zusammen mit einer Basis von I als k-Vektorraum bilden die B_i dann eine k-Vektorraumbasis von $k[X_1, \ldots, X_n]$. Über K wird die Basis von I zu einer K-Vektorraumbasis von $\bar{I}$, da sich jedes Element von $\bar{I}$ als eine K-Linearkombination von Elementen aus I schreiben läßt. Zusammen mit den B_i, die wir auch als Elemente von $K[X_1, \ldots, X_n]$ auffassen können, erhalten wir sowohl über k als auch über K eine Vektorraumbasis des ganzen jeweiligen Polynomrings, und damit ist klar, daß die Restklassen der B_i modulo $\bar{I}$ den Faktorring $\bar{A}$ erzeugen. Somit ist dieser als K-Vektorraum endlichdimensional.

Die Gleichheit von $\dim_k A$ und $\dim_K \bar{A}$ folgt, falls wir zeigen können, daß die Restklassen der B_i modulo $\bar{I}$ linear unabhängig sind.

Dazu zeigen wir die folgende, etwas allgemeinere Aussage: Sind $B_1, \ldots, B_r$ Polynome aus $k[X_1, \ldots, X_n]$ mit Restklassen $b_1, \ldots, b_r$ modulo I und Restklassen $\bar{b}_1, \ldots, \bar{b}_r$ modulo $\bar{I}$, so sind die b_i genau dann linear abhängig, wenn es die $\bar{b}_i$ sind.

Die eine Richtung ist einfach: Falls die b_i linear abhängig sind, gibt es Skalare $\lambda_i \in k$, die nicht alle verschwinden, so daß $\lambda_1 b_1 + \cdots + \lambda_r b_r$ der Nullvektor aus A ist. $\lambda_1 B_1 + \cdots + \lambda_r B_r$ liegt daher in I, also erst recht in $\bar{I}$, so daß auch $\lambda_1 \bar{b}_1 + \cdots + \lambda_r \bar{b}_r$ der Nullvektor aus $\bar{A}$ ist.

Wenn die $\bar{b}_i$ linear abhängig sind, gibt es $\lambda_i \in K$, so daß $\lambda_1 \bar{b}_1 + \cdots + \lambda_r \bar{b}_r$ der Nullvektor aus $\bar{A}$ ist, d.h. $\lambda_1 B_1 + \cdots + \lambda_r B_r$ liegt in $\bar{I}$. Da die λ_i nicht in k liegen müssen, nützt und das noch nichts, um etwas über die b_i auszusagen.

Um trotzdem deren lineare Abhängigkeit zu beweisen, wählen wir ein endliches Erzeugendensystem $f_1, \ldots, f_m$ des Ideals I. Wir wissen dann, daß es Polynome $g_1, \ldots, g_m$ aus $K[X_1, \ldots, X_n]$ gibt mit

$$\lambda_1 B_1 + \cdots + \lambda_r B_r = g_1 f_1 + \cdots + g_m f_m \,.$$

Die Polynome g_j sind K-Linearkombinationen von Monomen $M_{j\ell}$ in den Variablen X_i. Die obige Gleichung ist also äquivalent zu einer Gleichung der Form

$$\lambda_1 B_1 + \cdots + \lambda_r B_r - \sum_{j=1}^{m} \sum_{\ell=1}^{r_j} \mu_{j\ell} M_{j\ell} f_j = 0$$

mit Elementen $\mu_{j\ell} \in K$, die von den g_j abhängen. Sortieren wir diese Gleichung nach Monomen, können wir dies so interpretieren, daß ein (recht großes) lineares Gleichungssystem in den Variablen λ_i und $\mu_{j\ell}$ eine nichttriviale Lösung hat. Da die B_i und die f_j Polynome mit Koeffizienten aus k sind, ist dies ein homogenes lineares Gleichungssystem mit Koeffizienten aus k. Seine Lösungsmenge über k ist ein k-Vektorraum, für den uns der GAUSS-Algorithmus eine Basis liefert. Da der GAUSS-Algorithmus nirgends aus dem Körper hinausführt, in dem die Koeffizienten liegen, ist dies auch eine Basis des Lösungsraums über K; die beiden Vektorräume haben also dieselbe Dimension. Da wir wissen, daß es über K eine nichttriviale Lösung gibt, muß es daher auch über k eine geben.

Es gibt somit Elemente $\lambda'_i \in k$ und $\mu'_{j\ell} \in k$, die das Gleichungssystem lösen. Damit ist dann

$$\lambda'_1 B_1 + \cdots + \lambda'_r B_r = g'_1 f_1 + \cdots + g'_m f_m$$

mit Polynomen $g'_j \in k[X_1, \ldots, X_n]$, die linke Seite liegt also im Ideal I. Somit ist $\lambda'_1 b_1 + \cdots + \lambda'_r b_r$ der Nullvektor in A. Die λ'_i können nicht allesamt verschwinden, denn ansonsten müßte mindestens ein $\mu_{j\ell} \neq 0$ sein, Null wäre also gleich einer nichttrivialen Linearkombination von verschiedenen Monomen, was absurd ist. Also sind auch die b_i linear abhängig.

Bleibt noch zu zeigen, daß A endlichdimensional ist, wenn $\bar{A}$ endlichdimensional ist. Das folgt sofort aus der gerade gezeigten Äquivalenz

der linearen Abhängigkeit über k und über K: Hat $\bar{A}$ die endliche Dimension d, so ist jede Teilmenge von $\bar{A}$ mit mehr als d Elementen linear abhängig. Damit ist, wie wir gerade gesehen haben, auch jede Teilmenge von mehr als d Elementen aus A linear abhängig über k, also ist A endlichdimensional.

Im nächsten Schritt wollen wir das Zählen der Lösungen zurückführen auf das Zählen von Nullstellen eines Polynoms einer Veränderlichen.

Definition: Ein Polynom $u \in K[X_1, \ldots, X_n]$ heißt *separierend,* bezüglich $V_K(I)$, wenn es für keine zwei Elemente von $V_K(I)$ denselben Wert annimmt.

4. Schritt: Falls $V_K(I)$ endlich ist, gibt es ein separierendes homogenes lineares Polynom $u = c_1X_1 + \cdots c_nX_n$. Wir können dabei für u eines der speziellen Polynome

$$u_a = X_1 + aX_2 + a^2X_3 + \cdots + a^{n-1}X_n$$

wählen. Bezeichnet s die Elementanzahl von $V_K(I)$, so können wir a aus einer beliebig vorgebbaren Teilmenge von K mit mehr als $(n-1)\binom{s}{2} = \frac{1}{2}s(s-1)(n-1)$ Elementen liegt.

Für je zwei verschiedene Punkte $z, w \in V_K(I)$ ist $u_a(z) = u_a(w)$ genau dann, wenn

$$(z_1 - w_1) + (z_2 - w_2)a + (z_3 - w_3)a^2 + \cdots + (z_n - w_n)a^{n-1}$$

verschwindet. Die Koordinaten z_i, w_i von z und w sind Elemente von K; die $a \in K$, für die $u_a(z) = u_a(w)$ ist, sind also die Nullstellen eines Polynoms in einer Veränderlichen über K vom Grad höchstens $n-1$. Daher gibt es höchstens $n-1$ Werte $a \in K$, für die $u_a(z) = u_a(w)$ ist. Ist $s = \#V_K(I)$ endlich, so gibt es $\binom{s}{2}$ Paare aus voneinander verschiedenen Elementen; somit gibt es höchstens $(n-1)\binom{s}{2}$ Elemente $a \in K$, für die $u_a(z) = u_a(w)$ ist für *irgendwelche* voneinander verschiedene Elemente $z, w \in V_K(I)$.

(Hier haben wir benutzt, daß jeder algebraisch abgeschlossene Körper unendlich ist. Falls bereits k unendlich ist, etwa $k = \mathbb{Q}$, können wir sogar ein $a \in k$ finden, für das u_a separierend ist. Im hier meistens betrachteten

Fall $k = \mathbb{Q}$ können wir etwa eine ganze Zahl a mit $0 \le a \le (n-1)\binom{s}{2}$ wählen.)

5. Schritt: Die Elementanzahl s von $V_K(I)$ ist höchstens gleich der Dimension von A.

Da wir im 3. Schritt gesehen haben, daß $\dim_k A = \dim_K \bar{A}$ ist, können wir auch mit dieser Dimension argumentieren. Aus dem 4. Schritt wissen wir, daß es ein Polynom $u \in K[X_1, \ldots, X_n]$ gibt, das für jedes Element von $V_K(I)$ einen anderen Wert annimmt. Wir ersetzen u durch seine Restklasse $\tilde{u}$ modulo $\bar{I}$ in $\bar{A}$ und wollen uns überlegen, daß die Elemente $1, \tilde{u}, \ldots, \tilde{u}^{s-1} \in \bar{A}$ linear unabhängig sind: Angenommen, es gibt eine Relation der Form $\sum_{\ell=0}^{s-1} \lambda_\ell \tilde{u}^\ell = 0$ mit $\lambda_\ell \in K$. Das Polynom $\sum_{\ell=0}^{s-1} \lambda_\ell u^\ell \in K[X_1, \ldots, X_n]$ liegt dann in $\bar{I}$, verschwindet also für jedes der s Elemente von $V_K(I)$. Da u für jedes dieser Elemente einen anderen Wert annimmt, hat das Polynom $\sum_{\ell=0}^{s-1} \lambda_\ell U^\ell \in k[U]$ einerseits mindestens s verschiedene Nullstellen in K, andererseits ist sein Grad kleiner als s. Das ist nur für das Nullpolynom möglich; somit verschwinden alle Koeffizienten λ_ℓ, was die behauptete lineare Unabhängigkeit beweist. Somit enthält $\bar{A}$ mindestens s linear unabhängige Elemente, d.h. $r = \dim_K \bar{A} \ge s = \#V_K(I)$. Damit ist die Behauptung und auch der gesamte Satz bewiesen.

∎

Betrachten wir als Beispiel das von $f = X^2 + Y^2 - 1$ und $g = X - Y$ erzeugte Ideal $I \lhd \mathbb{Q}[X, Y]$. Seine Lösungsmenge ist, geometrisch gesehen, der Schnitt des Einheitskreises mit der ersten Winkelhalbierenden, besteht also aus den beiden Punkten $\left(\frac{1}{2}\sqrt{2}, \frac{1}{2}\sqrt{2}\right)$ und $\left(-\frac{1}{2}\sqrt{2}, -\frac{1}{2}\sqrt{2}\right)$.

Der Polynomring $\mathbb{Q}[X, Y]$ hat als $\mathbb{Q}$-Vektorraum eine Basis bestehend aus allen Monomen $X^a Y^b$ mit $a, b \in \mathbb{N}_0$. Modulo I sind X und Y äquivalent, und damit ist $X^a Y^b \sim X^{a+b}$. Außerdem ist $2X^2$ äquivalent zu $X^2 + Y^2$, und das wiederum ist wegen f äquivalent zu 1, d.h. $X^2 \sim \frac{1}{2}$. Daher ist jedes Monom äquivalent entweder zu einer Konstanten (falls $a+b$ gerade) oder einem skalaren Vielfachen von X. Da I kein Polynom der Form $\lambda X + \mu$ enthält, sind X und 1 modulo I linear unabhängig; somit bilden ihre Restklassen eine Basis des Vektorraums $\mathbb{Q}[X, Y]/I$.

Ersetzen wir in diesem Beispiel g durch $X^2 - Y^2 = (X + Y)(X - Y)$, so schneiden wir den Kreis mit beiden Winkelhalbierenden und haben nun eine vierelementige Lösungsmenge

$$V_{\mathbb{C}}(I) = \left\{ \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), \left(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right), \left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right) \right\}.$$

Modulo dem neuen Ideal I sind X und Y nicht mehr äquivalent, sondern nur noch X^2 und Y^2. Jedes Monom ist somit äquivalent entweder zu einer X-Potenz oder zu einem Monom der Form $X^a Y$. Da auch hier $X^2 \sim \frac{1}{2}$, ist es somit äquivalent zu einem skalaren Vielfachen eines der Monome $1, X, Y$ oder XY. Da keine Linearkombination dieser vier Monome in I liegt, bilden ihre Restklassen eine Basis von $\mathbb{Q}[X, Y]/I$.

In diesen beiden Beispielen waren sowohl die Lösungsmengen als auch Basen der Faktorringe einfach zu finden; im allgemeinen ist das eher nicht der Fall. Wenn wir aber eine GRÖBNER-Basis des Ideals I kennen, können wir leicht eine Vektorraumbasis des Faktorrings konstruieren:

Definition: $I \lhd k[X_1, \ldots, X_n]$ sei ein Ideal und G sei eine GRÖBNER-Basis bezüglich irgendeiner Monomordnung auf $k[X_1, \ldots, X_n]$. Ein Monom in $X_1, \ldots, X_n$ heißt *Standardmonom* (bezüglich G), wenn es für kein $g \in G$ durch das führende Monom von g teilbar ist.

(Tatsächlich sollte man von Standardmonomen bezüglich einer Monomordnung reden, denn eine Menge G kann durchaus GRÖBNER-Basis bezüglich zweier verschiedener Monomordnungen sein, und zumindest einige ihrer Elemente können bezüglich dieser Monomordnungen verschiedene führende Monome haben, so daß ein Standardmonom bezüglich der einen Monomordnung keines bezüglich der anderen sein muß.)

Satz: Für jede GRÖBNER-Basis G eines Ideals $I \lhd k[X_1, \ldots, X_n]$ bilden die Restklassen der Standardmonome eine Vektorraumbasis von $k[X_1, \ldots, X_n]/I$.

Beweis: Zunächst sind diese Restklassen linear unabhängig, denn jede nichttriviale Linearkombination der Null entspräche einem Polynom h

aus I, dessen sämtliche Monome Standardmonome sind. Da die führenden Monome der Elemente von G das Ideal FM(I) erzeugen, müßte daher FM(h) Vielfaches eines FM(g) mit $g \in G$ sein, was der Definition eines Standardmonoms widerspricht.

Für ein beliebiges $f \in k[X_1, \ldots, X_n]$ liefert uns der Divisionsalgorithmus eine Darstellung

$$f = \sum_{g \in G} a_g g + r \quad \text{mit} \quad a_g, r \in k[X_1, \ldots, X_n] \,,$$

wobei r eine k-Linearkombination von Standardmonomen ist. Da die Summe der $a_g g$ in I liegt, ist f also äquivalent zu einer k-Linearkombination von Standardmonomen, so daß seine Restklasse die entsprechende Linearkombination von deren Restklassen ist. ∎

Dieser Satz gilt unabhängig davon, ob $k[X_1, \ldots, X_n]/I$ als Vektorraum endlichdimensional ist; er liefert uns auch ein einfaches Kriterium dafür, wann er endliche Dimension hat und wann somit die Lösungsmenge $V_K(I)$ endlich ist:

Lemma: G sei eine GRÖBNER-Basis eines Ideals $I \lhd k[X_1, \ldots, X_n]$ bezüglich irgendeiner Monomordnung. $V_K(I)$ ist genau dann endlich, wenn G für jedes i ein Polynom enthält, dessen führendes Monom eine X_i-Potenz ist.

Beweis: Falls die GRÖBNER-Basis für jedes i ein Polynom mit führendem Monom $X_i^{d_i}$ enthält, ist jedes Monom, in dem ein X_i mit einem Exponenten größer oder gleich d_i vorkommt, durch das führende Monom eines Elements der GRÖBNER-Basis teilbar. Die Monome, für die das nicht der Fall ist, haben für jedes i einen Exponenten echt kleiner d_i; es gibt also nur endlich viele Standardmonome. Somit hat A endliche Dimension, und $V_K(I)$ ist endlich.

Ist umgekehrt $V_K(I)$ endlich, so enthält $\bar{I}$ für jedes i ein Polynom aus $K[X_i]$ – siehe Schritt 2 im Beweis des obigen Satzes. Da die GRÖBNER-Basis von I gleichzeitig eine GRÖBNER-Basis von $\bar{I}$ ist, muß das führende Monom eines ihrer Elemente die höchste X_i-Potenz in diesem Polynom teilen und damit selbst eine Potenz von X_i sein. ∎

Für den Fall, daß $V_K(I)$ endlich ist, läßt der obige Satz noch wie folgt verschärfen:

Satz: Ist $D = V_K(I)$ endlich und τ eine Monomordnung, so gibt es zu jeder Funktion $\varphi\colon D \to K$ eine K-Linearkombination f von Standardmonomen bezüglich τ derart, daß $f(x) = \varphi(x)$ für alle $x \in D$. Insbesondere ist die Dimension von $k[X_1, \ldots, X_n]/I$ größer oder gleich der Elementanzahl von D. Die beiden Zahlen sind genau dann gleich, wenn I das Ideal $I(D)$ aller auf D verschwindender Polynome ist, was wiederum dazu äquivalent ist, daß I ein Radikalideal ist.

Beweis: Zunächst sollten wir uns überlegen, daß es überhaupt ein Polynom $\widetilde{f} \in K[X_1, \ldots, X_n]$ gibt mit $\widetilde{f}(x) = \varphi(x)$ für alle $x \in D$. Im Eindimensionalen können wir $\widetilde{f}$ nach LAGRANGE oder NEWTON als Interpolationspolynom konstruieren, und den allgemeinen Fall können wir wie folgt darauf zurückführen: Wie wir im vierten Schritt des Beweises des ersten Satzes in diesem Paragraphen gesehen haben, gibt es ein homogenes lineares Polynom $\ell \in k[X_1, \ldots, X_n]$, das auf den verschiedenen Punkten von D verschiedene Werte annimmt. Dazu betrachten wir das Interpolationspolynom aus $K[T]$, das für jeden Punkt $x \in D$ an der Stelle $t = \ell(x)$ den Wert $\varphi(x)$ annimmt. Setzen wir ℓ in dieses Polynom ein, erhalten wir ein Polynom $\widetilde{f}$ aus $k[X_1, \ldots, X_n]$, das für jedes $x \in D$ an der Stelle x den Wert $\varphi(x)$ annimmt. Da die Restklassen der Standardmonome eine Basis des Restklassenrings bilden, gibt es dazu eine Linearkombination f von Standardmonomen, die sich nur durch ein Polynom aus $\bar{I}$ von $\widetilde{f}$ unterscheidet, d.h. $f(x) = \widetilde{f}(x) = \varphi(x)$ für alle $x \in D$.

Die Funktionen $\varphi\colon D \to K$ bilden offensichtlich einen K-Vektorraum, den wir für $D = \{x^{(1)}, \ldots, x^{(r)}\}$ identifizieren können mit dem Vektorraum aller Tupel $\big(\varphi(x^{(1)}, \ldots, \varphi(x^{(r)})\big)$, also mit K^r. Die Dimension des Vektorraums aller dieser Funktionen ist somit gleich der Elementanzahl von D.

Diese Dimensionen ist genau dann gleich der Vektorraumdimension des Faktorrings, wenn die obige Linearkombination f durch φ eindeutig bestimmt ist. Sind f_1 und f_2 zwei verschiedene solche Linearkombina-

tionen, so verschwindet $f_1 - f_2$ auf ganz D, liegt also im Ideal $I(D)$. Genau dann, wenn dieses mit I übereinstimmt, können wir daraus folgern, daß $f_1 = f_2$ ist, und das ist nach dem HILBERTschen Nullstellensatz genau dann der Fall, wenn I ein Radikalideal ist. ∎

In §1 haben wir gesehen, wie man zu jeder endlichen Teilmenge $D \subset k^n$ ein Ideal $I \lhd k[X_1, \ldots, X_n]$ finden kann, für das $D = V_K(I)$ ist. Mit dem gerade bewiesenen Satz können wir nun sehen, daß das dort konstruierte Ideal gleich $I(D)$ ist:

Für $D = \{x^{(1)}, \ldots, x^{(r)}\} \subset k^n$ mit $x^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$ hatten wir die Punkte

$$y^{(i)} = (0, \ldots, 0, 1, 0, \ldots, 0, x_1^{(i)}, \ldots, x_n^{(i)}) \in k^{r+n}$$

betrachtet, wobei die Eins bei $y^{(i)}$ an der i-ten Stelle steht; die Menge dieser Punkte sei $\widetilde{D}$. Wie wir gesehen hatten, ist $\widetilde{D}$ die Nullstellenmenge jenes Ideals $J \lhd k[T_1, \ldots, T_r, X_1, \ldots, X_n]$, das erzeugt wird von den Polynomen $f_{ij} = T_i(X_j - x_j^{(i)})$ und dem Polynom $g = T_1 + \cdots + T_r - 1$. Wir wollen uns als erstes überlegen, daß J das Ideal *aller* auf $\widetilde{D}$ verschwindenden Funktionen ist: Da $f_{ij} \in J$, ist jedes Monom $T_i X_j$ modulo J äquivalent zu einem skalaren Vielfachen von T_i. Induktiv folgt, daß für jedes nichtkonstante Monom M in den X_j das Monom $T_i M$ äquivalent ist zu einem skalaren Vielfachen von T_i. Da g in J liegt, ist M selbst äquivalent zu $T_1 M + T_2 M + \cdots + T_r M$ und damit zu einem linearen Polynom in den T_i. Somit ist jedes Polynom aus $k[T_1, \ldots, T_r, X_1, \ldots, X_n]$ äquivalent zu einem Polynom nur in den T_i.

Für zwei verschiedene Punkte $x^{(i)}$ und $x^{(\ell)}$ aus D gibt es mindestens einen Index j, für den $x_j^{(i)} \neq x_j^{(\ell)}$ ist. Mit f_{ij} und $f_{\ell j}$ enthält J auch das Polynom

$$T_\ell f_{ij} - T_i f_{\ell j} = T_\ell T_i X_j - T_\ell T_i x_j^{(i)} - T_i T_\ell X_j + T_i T_\ell x^{(\ell)} = T_i T_\ell \left(x_j^{(\ell)} - x_j^{(i)}\right)$$

und damit das Produkt $T_i T_\ell$, so daß jedes Monom, das zwei verschiedene T_i enthält, modulo J verschwindet. Außerdem liegt für jedes T_i auch das Polynom $T_i g = T_i T_1 + \cdots + T_i T_r - T_i$ in J, d.h. modulo J ist T_i äquivalent zu $T_i T_1 + \cdots + T_i T_r$. Da alle $T_i T_\ell$ mit $\ell \neq i$ in J liegen,

ist T_i damit auch äquivalent zu T_i^2 und damit auch zu jeder höheren T_i-Potenz. Somit ist jedes Polynom äquivalent zu einem linearen Polynom in den T_i, wobei wir dieses homogen wählen können, da 1 äquivalent ist zur Summe der T_i.

Dies zeigt, daß der Restklassenring modulo J als k-Vektorraum höchstens die Dimension r hat. Diese Dimension hat auch der Vektorraum aller Funktionen $\widetilde{D} \to K$. Damit folgt aus dem gerade bewiesenen Satz, daß J ein Radikalideal sein muß. Dann ist aber auch $I = J \cap k[X_1, \ldots, X_n]$ ein Radikalideal, d.h. $I = I(D)$.

Kapitel 3
Anwendung auf Designs

Eine der Grundaufgaben der Statistik besteht darin, ausgehend von Stichproben Modelle zu entwickeln und deren Parameter zu schätzen. Wir gehen aus von einem Körper k, den wir meist als Teilkörper der reellen Zahlen auffassen werden. Da wir allerdings im Körper k exakt rechnen wollen, sollte k so klein wie möglich sein, etwa $k = \mathbb{Q}$.

Definition: Ein *Design* D in k^n ist eine endliche Teilmenge von k^n.

Oftmals haben die n Koordinaten der Punkte aus D inhaltliche Interpretationen, indem sie die Ausprägungen verschiedener Faktoren kodieren. Eine Stichprobe dient dann dazu, den Einfluß verschiedener der Faktoren auf ein Ergebnis abzuschätzen. Beispiele sind etwa der Ertrag eines landwirtschaftlichen Produkts in Abhängigkeit von Düngung, Bewässerung, Bodenbeschaffenheit *usw.,* oder der Preis, den ein Kunde für ein Auto zu zahlen bereit ist in Abhängigkeit von verschiedenen Ausstattungsmerkmalen. Oft werden für diese Faktoren nur endlich viele Stufen betrachtet. Diese können auf einer reinen Nominalskala liegen, etwa $\{$*rot*, *blau*, *gelb*$\}$ oder $\{$*vorhanden*, *nicht vorhanden*$\}$, weshalb man sich in der Literatur im Falle von ℓ Stufen oft mit der Kodierung durch die Zahlen von Null bis $\ell - 1$ begnügt. Wenn es die Darstellung erleichtert, werde ich dies im Folgenden auch gelegentlich tun, aber inhaltlich ändert sich dadurch nichts.

Definition: *a)* Ein volles faktorielles Design für n Faktoren ist ein kartesisches Produkt von n endlichen Teilmengen $S_i \subset k$. Falls alle M_i jeweils ℓ Elemente haben, sprechen wir von einem ℓ^n-Design.

b) Ein (fraktionelles) faktorielles Design für n Faktoren ist eine Teilmenge eines vollen faktoriellen Designs für diese Faktoren.

Offensichtlich kann jedes Design $D \subset k^n$ als fraktionelles faktorielles Design angesehen werden, etwa für S_i gleich Menge aller möglicher Werte, die die Variable x_i auf D annehmen kann.

§1: Allgemeine lineare Modelle

Wir haben uns bereits in Kapitel 0 mit statistischen Modellen beschäftigt; hier seien noch einmal kurz die im folgenden relevanten Bezeichnungen rekapituliert.

Wir starten mit einem Design $D = \{x^{(1)}, \ldots, x^{(r)}\} \subset k^n$. Typischerweise ist für jeden Punkt $x^{(j)} \in D$ ein Wert $y_j \in k$ gegeben; gesucht ist eine Funktion $f\colon k^n \to k$ mit $f(x^{(j)}) = y_j$ oder zumindest $f(x^{(j)}) \approx y_j$ für alle j.

Für den allgemeinsten (linearen) Ansatz zum Auffinden solcher Funktionen wählen wir eine endliche Menge $\mathbb{F} = \{f_1, \ldots, f_s\}$ von Funktionen $f_i\colon k^n \to k$ und betrachten Funktionen f der Form $f = \sum_{i=1}^{s} \theta_i f_i$ mit $\theta_1, \ldots, \theta_s \in k$.

Definition: Die Z-Matrix zu $\mathbb{F}$ und D ist die $s \times r$-Matrix Z mit Einträgen $z_{ij} = f_i(x^{(j)})$.

Für $f = \sum_{i=1}^{s} \theta_i f_i$ sollte dann idealerweise gelten

$$y_j = \sum_{i=1}^{s} \theta_i f_i(x^{(j)}) = \sum_{i=1}^{s} \theta_i z_{ij} \; ;$$

ausgedrückt mit den Vektoren

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix} \quad \text{und} \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_s \end{pmatrix}$$

wird dies zu $y^T = \theta^T Z \quad$ oder $\quad Z^T \theta = y$.

Falls $r < s$ ist, kann dieses Gleichungssystem keine eindeutig bestimmte Lösung haben. Für $r = s$ gibt es genau dann eine, wenn die Z-Matrix invertierbar ist; dann ist $\theta = (Z^T)^{-1} y$.

In der Statistik ist meist $r > s$; in diesem Fall wird es im allgemeinen keine Lösung geben. Dann müssen wir uns begnügen mit einem Vektor θ derart, daß der Fehler $\varepsilon = y - Z^T\theta$ möglichst klein wird. Wie wir in Kapitel 0 gesehen haben, führt dies auf das lineare Gleichungssystem

$$ZZ^T\theta = Zy\,.$$

ZZ^T ist eine quadratische Matrix; es gibt daher genau dann eine eindeutig bestimmte Lösung, wenn diese Matrix nichtsingulär ist. Das wiederum ist äquivalent dazu, daß Z (und damit auch Z^T) den Rang s hat: In diesem Fall ist nämlich die Abbildung von k^r nach k^s, die einem Vektor x den Vektor Zx zuordnet, surjektiv, und die Abbildung von k^s nach k^r, die θ auf $Z^T\theta$ abbildet, hat ein s-dimensionales Bild. Somit hat die Abbildung von k^r nach k^r, die einem Vektor $x \in k^r$ den Vektor $ZZ^Tx = Z(Z^Tx)$ zuordnet, ein s-dimensionales Bild, d.h. der Rang von ZZ^T ist s. Ist umgekehrt der Rang von Z kleiner als s, so erst recht der von ZZ^T, so daß ZZ^T in der Tat genau dann nichtsingulär ist, wenn der Rang von Z gleich s ist.

Damit haben wir gezeigt

Satz: Falls der Rang von Z gleich s ist, gibt es genau einen Vektor $\theta \in k^s$, für den die Differenz $Z^T\theta - y$ minimale Länge hat, nämlich $\theta = (ZZ^T)^{-1}y$.

■

§2: Polynomiale lineare Modelle

In dieser Vorlesung soll es in erster Linie um Modelle gehen, bei denen die Funktionen aus $\mathbb{F}$ Monome sind. Schon lange vor dem Aufkommen der algebraischen Statistik betrachtete man im Sinne möglichst einfacher Modelle vorzugsweise Mengen $\mathbb{F}$, die mit jedem Monom auch dessen sämtliche Teiler enthalten; diese hatten wir in Kapitel 0 als Ordnungsideale bezeichnet. Um auch andere Modelle nicht ganz auszuschließen, fassen wir die folgende Definition etwas allgemeiner:

Definition: k sei ein Körper, $R = k[X_1, \ldots, X_n]$ ein Polynomring über k, und $\mathbb{T}$ sei die Menge aller Monome $X_1^{e_1} \cdots X_n^{e_n}$ mit $e_i \in \mathbb{N}_0$.
a) Ein lineares Modell mit $\mathbb{F} \subset \mathbb{T}$ heißt *monomiales lineares Modell;*

die Menge $\mathbb{F}$ wird als der *Träger* des Modells bezeichnet.
b) Eine nichtleere Teilmenge $\mathcal{O}$ von $\mathbb{T}$ heißt *Ordnungsideal,* wenn für jedes Monom aus $\mathcal{O}$ auch dessen sämtliche Teiler in $\mathcal{O}$ liegen.
c) Das Modell heißt *vollständig,* wenn $\mathbb{F}$ ein Ordnungsideal ist.

Das Ergebnis am Ende des letzten Kapitels zeigt, wie wir zu jedem Design D ein vollständiges lineares Modell finden können: Wir fassen D zunächst auf als Nullstellenmenge eines Ideals I des Polynomrings R. Wie man ein solches Ideal I finden kann, haben wir am Ende von §1 des vorigen Kapitels gesehen, und ganz am Ende des Kapitels haben wir uns überlegt, daß es das volle Designideal $I(D)$ ist. Danach wählen wir eine Monomordnung τ auf R und berechnen dazu eine GRÖBNER-Basis von I.

Definition: *a)* $\mathrm{Est}_\tau(D)$ ist die Menge aller Standardmonome zu einer GRÖBNER-Basis von $I(D)$ bezüglich der Monomordnung τ.
b) Die Menge aller Mengen $\mathrm{Est}_\tau(D)$, wobei τ die sämtlichen Monomordnungen von R durchläuft, heißt der (GRÖBNER-)*Fächer* von D.

Mit unseren bisherigen Mitteln ist die Konstruktion des Fächers meist recht aufwendig, da wir dazu GRÖBNER-Basen des Designideals zu jeder der unendlich vielen möglichen Monomordnungen kennen und damit auch berechnen müssen.

Im Falle eines vollen faktoriellen Design $D = S_1 \times \cdots \times S_n$ haben wir damit allerdings keinerlei Schwierigkeiten: Offensichtlich bilden die Polynome

$$g_i = \prod_{x \in S_i} (X_i - x)$$

ein Erzeugendensystem von $I(D)$, und dieses Erzeugendensystem ist nach dem Kriterium von BUCHBERGER bezüglich jeder Monomordnung eine GRÖBNER-Basis, denn g_i ist ein Polynom nur in X_i, so daß sein führendes Monom eine Potenz von X_i ist. Somit sind die führenden Monome zweier verschiedener g_i teilerfremd. Wie wir bei der Diskussion von BUCHBERGERs Kriterium gesehen haben reduziert daher $S(g_i, g_j)$ für $i \neq j$ bezüglich jeder Monomordnung auf Null.

Damit ist klar, daß die Standardmonome für ein volles faktorielles Design $D = S_1 \times \cdots \times S_n$ bezüglich jeder Monomordnung genau diejenigen Monome $X_1^{e_1} \cdots X_n^{e_n}$ sind, für die jedes e_i kleiner ist als die Elementanzahl der entsprechenden Menge S_i.

§3: Designs mit minimalem Fächer

Der Fächer eines vollständigen faktoriellen Designs besteht aus genau einem Blatt; kleiner kann er nicht werden. Deshalb definieren wir

Definition: Ein Design D hat minimalen Fächer, wenn alle Monomordnungen τ auf die gleiche Menge $\mathrm{Est}_\tau(D)$ führen.

Das ist insbesondere dann der Fall, wenn alle Monomordnungen zur gleichen GRÖBNER-Basis führen; betrachten wir also diesen Fall etwas genauer:

Definition: *a)* Eine Teilmenge $G = \{g_1, \ldots, g_m\}$ eines Ideals I von $k[X_1, \ldots, X_n]$ heißt *universelle* GRÖBNER-Basis, wenn sie bezüglich jeder Monomordnung auf $k[X_1, \ldots, X_n]$ eine GRÖBNER-Basis ist.
b) G heißt *Super-G-Basis* bezüglich einer Monomordnung τ, wenn jede Teilmenge von G bezüglich τ eine GRÖBNER-Basis des von ihr erzeugten Ideals ist.

Die beiden hier definierten Konzepte haben eigentlich nichts miteinander zu tun; bei den Designs, die wir hier betrachten wollen, ist aber ihr Zusammenspiel nützlich. Wir beginnen mit einem Kriterium zur Charakterisierung von Super-G-Basen:

Lemma: $G \subset I$ ist genau dann eine Super-G-Basis, wenn für je zwei Elemente $f, g \in G$ gilt:

$$\mathrm{FM}_\tau\big(\mathrm{ggT}(f,g)\big) = \mathrm{ggT}\big(\mathrm{FM}_\tau(f), \mathrm{FM}_\tau(g)\big)\,. \qquad (*)$$

Beweis durch Induktion nach $m = \#G$:

Der Fall $m = 1$ ist trivial, denn dann ist $I = (f)$ ein Hauptideal, d.h. zu jedem $g \in I$ gibt es ein Polynom h, so daß $g = fh$ ist. Damit ist auch

$\mathrm{FM}_\tau(g) = \mathrm{FM}_\tau(f)\,\mathrm{FM}_\tau(h)$, d.h. $\mathrm{FM}_\tau(I)$ wird von $\mathrm{FM}_\tau(f)$ erzeugt, so daß jedes Erzeugende von I eine GRÖBNER-Basis von I definiert.

Für $m = 2$ ist $G = \{f, g\}$, und da $\{f\}$ und $\{g\}$ GRÖBNER-Basen der Hauptideale (f) und (g) sind, ist $\{f, g\}$ genau dann eine Super-G-Basis, wenn es eine GRÖBNER-Basis ist. Wir müssen daher zeigen, daß $\{f, g\}$ genau dann eine GRÖBNER-Basis von I ist, wenn gilt

$$\mathrm{FM}_\tau\big(\mathrm{ggT}(f, g)\big) = \mathrm{ggT}\big(\mathrm{FM}_\tau(f), \mathrm{FM}_\tau(g)\big)\,.$$

Mit $h = \mathrm{ggT}\big(\mathrm{FM}_\tau(f), \mathrm{FM}_\tau(g)\big)$ ist das kgV der führenden Monome gleich $\mathrm{FM}_\tau(f)\,\mathrm{FM}_\tau(g)/h$, also

$$S(f, g) = \frac{\mathrm{FM}_\tau(g)}{\mathrm{FK}_\tau(f)h} f - \frac{\mathrm{FM}_\tau(f)}{\mathrm{FK}_\tau(g)h} g\,.$$

Nach dem Kriterium von BUCHBERGER ist G genau dann eine GRÖBNER-Basis, wenn dieses Polynom modulo $\{f, g\}$ auf Null reduziert, wenn es also Polynome p, q gibt, so daß $S(f, g) = pf + qg$ ist, wobei weder $\mathrm{FM}_\tau(pf)$ noch $\mathrm{FM}_\tau(qg)$ bezüglich τ größer als das führende Monom von $S(f, g)$ sein darf. Angenommen, das ist der Fall. Mit

$$\widetilde{g} = \frac{\mathrm{FM}_\tau(g)}{\mathrm{FK}_\tau(f)h} - p \quad \text{und} \quad \widetilde{f} = \frac{\mathrm{FM}_\tau(f)}{\mathrm{FK}_\tau(g)h} - q$$

ist dann

$$\widetilde{g}f - \widetilde{f}g = \frac{\mathrm{FM}_\tau(g)}{\mathrm{FK}_\tau(f)h} f - pf - \frac{\mathrm{FM}_\tau(f)}{\mathrm{FK}_\tau(g)h} g + qg = S(f, g) - (fp + qg) = 0\,.$$

In der Definition von $\widetilde{g}$ *bzw.* $\widetilde{f}$ ist der jeweils erste Summand der führende Term, denn wäre ein Term aus p *bzw.* q größer, so wäre das führende Monom von pf *bzw.* qg größer als das von $S(f, g)$.

Da h der ggT der führenden Monome von f und g ist, sind die führenden Monome von $\widetilde{f}$ und $\widetilde{g}$ teilerfremd, und damit sind auch $\widetilde{f}$ und $\widetilde{g}$ teilerfremd. Wegen $\widetilde{g}f = \widetilde{f}g$ muß daher $\widetilde{f}$ ein Teiler von f sein und $\widetilde{g}$ einer von g. Da Polynomringe faktoriell sind, gibt es somit ein Polynom $\widetilde{h}$ derart, daß $f = \widetilde{h}\widetilde{g}$ und $g = \widetilde{h}\widetilde{f}$ ist. Wegen der Teilerfremdheit von $\widetilde{f}$ und $\widetilde{g}$ folgt daraus, daß $\widetilde{h} = \mathrm{ggT}(f, g)$ sein muß, und wenn wir die führenden Monome betrachten, sehen wir, daß h das führende Monom von $\widetilde{h}$ ist. Falls G eine GRÖBNER-Basis ist, gilt also $(*)$.

Für die Umkehrung nehmen wir der Einfachheit halber an, daß f und g jeweils führenden Koeffizienten eins haben; dadurch ändert sich weder das Ideal noch das S-Polynom. Ist dann $(*)$ erfüllt und $\widetilde{h} = \mathrm{ggT}(f, g)$, so gibt es teilerfremde Polynome $\widetilde{f}, \widetilde{g}$ derart, daß $f = \widetilde{h} \cdot \widetilde{f}$ und $g = \widetilde{h} \cdot \widetilde{g}$ ist. Wir können annehmen, daß auch diese Polynome alle die Eins als höchsten Koeffizienten haben, und natürlich ist $\widetilde{g}f - \widetilde{f}g = 0$.

Nach $(*)$ ist das führende Monom h von $\widetilde{h}$ der größte gemeinsame Teiler von $\mathrm{FM}_\tau(f)$ und $\mathrm{FM}_\tau(g)$. Somit ist $\mathrm{FM}_\tau\left(\widetilde{f}\right) = \mathrm{FM}_\tau(f)/h$ und $\mathrm{FM}_\tau\left(\widetilde{g}\right) = \mathrm{FM}_\tau(g)/h$. Schreiben wir

$$\widetilde{f} = \frac{\mathrm{FM}_\tau(f)}{h} - q \quad \text{und} \quad \widetilde{g} = \frac{\mathrm{FM}_\tau(g)}{h} - p$$

mit geeigneten Polynomen p und q, so zeigt die gleiche Rechnung wie oben, daß

$$\widetilde{g}f - \widetilde{f}g = S(f, g) - (pf + qg)$$

ist. Da die linke Seite verschwindet, folgt $S(f, g) = pf + qg$, wobei die führenden Monome von pf und gq nicht größer als das von $S(f, g)$ sind. Also reduziert das S-Polynom auf Null, und G ist nach dem Kriterium von BUCHBERGER eine GRÖBNER-Basis.

Damit ist der Induktionsanfang $m = 2$ gezeigt. Sei nun $m > 2$ und $G = \{g_1, \ldots, g_m\}$. Falls G seine Super-G-Basis ist, muß für je zwei Elemente g_i und g_j auch $\{g_i, g_j\}$ eine GRÖBNER-Basis von (g_i, g_j) sein, d.h. nach dem gerade bewiesenen Fall $m = 2$ gilt obige Gleichung.

Gilt umgekehrt obige Gleichung für alle (i, j) und ist G' eine Teilmenge von G, so ist zunächst nach dem Fall $m = 2$ für jedes Paar (i, j) mit $g_i, g_j \in G'$ die Menge $\{g_i, g_j\}$ eine GRÖBNER-Basis des Ideals (g_i, g_j). Somit reduziert $S(g_i, g_j)$ nach dem Kriterium von BUCHBERGER modulo $\{g_i, g_j\}$ und damit erst recht modulo G' auf Null. Da dies für alle Paare (i, j) gilt, zeigt eine nochmalige Anwendung des Kriteriums von BUCHBERGER, daß G' eine GRÖBNER-Basis des von G' erzeugten Ideals ist. Damit ist G eine Super-G-Basis.

■

Polynome, für die dieses Kriterium sehr einfach nachzuweisen ist, sind *Distraktionen* (Zerstreuungen) von Monomen:

Definition: $M = X_1^{e_1} \cdots X_n^{e_n}$ sei ein Monom, und für jedes $i = 1, \ldots, n$ sei ein Vektor $a^{(i)} \in k^{\ell_i}$ gegeben, wobei $\ell_i \geq e_i$. Die *Distraktion* von M bezüglich dieser Vektoren ist

$$D(M) = \prod_{i=1}^{n} \prod_{j=1}^{e_i} \left(X_i - a_j^{(i)}\right) .$$

Falls alle Vektoren $a^{(i)}$ Nullvektoren sind, ist $D(M) = M$. Allgemein können wir nur sagen, daß M das führende Monomn von $D(M)$ ist.

Für ein volles faktorielles Design $D = S_1 \times \cdots \times S_n$ mit $\ell_i = \#S_i$ wählen wir Vektoren $a^{(i)}$, deren Komponenten (in irgendeiner Reihenfolge) die verschiedenen Elemente von S_i sind, Dann ist

$$D(X_i^{\ell_i}) = \prod_{j=1}^{\ell_i} \left(X_i - a_j^{(i)}\right) = \prod_{x \in S_i} \left(X_i - x\right)$$

und

$$I(D) = \left(D(X_1^{\ell_1}), \ldots, D(X_n^{\ell_n})\right) .$$

Die uns bereits bekannte Tatsache, daß dieses Erzeugendensystem eine universelle GRÖBNER-Basis ist, folgt nun auch aus dem folgenden

Satz: $M_1, \ldots, M_r$ seien Monome aus $k[X_1, \ldots, X_n]$, und $a^{(1)}, \ldots, a^{(n)}$ seien Vektoren über k mit hinreichend vielen Komponenten. Dann bilden die Polynome $D(M_1), \ldots, D(M_r)$ eine universelle GRÖBNER-Basis des von ihnen erzeugten Ideals. Falls keines der Monome M_i ein anderes teilt, ist diese GRÖBNER-Basis reduziert.

Beweis: Aus der Definition von $D(M)$ folgt sofort, daß alle Monome in $D(M)$ Teiler von M sind; daher ist M bezüglich jeder Monomordnung das führende Monom von $D(M)$. Außerdem ist klar, daß zwei Distraktionspolynome $D(M_i)$ und $D(M_j)$ (bezüglich der gleichen Vektoren) genau dann Teiler voneinander sind, wenn dasselbe für M_i und M_j gilt. Damit ist insbesondere der ggT zweier Polynome $D(M_i)$ und $D(M_j)$ gleich $D\big(\text{ggT}(M_i, M_j)\big)$. Da das führende Monom einer Distraktion $D(M)$ gleich M ist, ist also das führende Monom von $\text{ggT}\big(D(M_i), D(M_j)\big)$ gleich dem ggT der führenden Monome von

$D(M_i)$ und $D(M_j)$, so daß das Kriterium aus obigem Lemma erfüllt ist. Somit bilden die $D(M_i)$ bezüglich jeder Monomordnung eine Super-G-Basis des von ihnen erzeugten Ideals, insbesondere also eine GRÖBNER-Basis.

Diese GRÖBNER-Basis ist minimal genau dann, wenn kein führendes Monom eines $D(M_i)$ das eines $D(M_j)$ mit $j \neq i$ teilt, wenn also kein M_i ein anderes teilt. (Der führende Koeffizient der Distraktion eines Monoms ist natürlich immer gleich eins.) In diesem Fall ist die Basis auch reduziert, denn würde M_i irgendein Monom von $D(M_j)$ teilen, so auch M_j selbst, da alle Monome in $D(M_j)$ Teiler von M_j sind. ∎

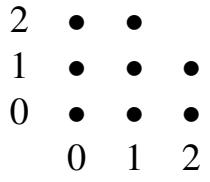
Außer auf den Fall der vollständigen faktoriellen Designs können wir diesen Satz auch auf sogenannte *Stufendesigns* anwenden:

Definition: Ein Design $D \subset \mathbb{N}_0^n$ heißt *Stufendesign,* wenn für jeden Punkt $(x_1, \ldots, x_n) \in D$ auch die sämtlichen Punkte $(u_1, \ldots, u_n) \in \mathbb{N}_0^n$ mit $u_i \leq x_i$ für alle i in D liegen.

Als Beispiel eines Stufendesigns, das kein volles faktorielles Design ist, können wir etwa die Menge

$$D = \{0,1,2\} \times \{0,1\} \cup \{0,1\} \times \{0,1,2\} = \{0,1,2\} \times \{0,1,2\} \setminus \{(2,2\}$$

betrachten. In der graphischen Darstellung sieht man die Stufe:



Lemma: Jedes Stufendesign ist die Nullstellenmenge eines Ideals, das von Distraktionen von Monomen bezüglich hinreichend langer Vektoren $a^{(i)} = (0, 1, \ldots, \ell_i)$ erzeugt wird.

Beweis durch Induktion nach n: Für $n = 1$ gibt es ein $\ell_1 \in \mathbb{N}_0$, so daß

$$D = \{0, 1, \ldots, \ell_1\} = V\big(X_1(X_1 - 1)\ldots(X_1 - \ell_1)\big) = V\big(D(X_1^{\ell_1+1})\big)$$

ist.

Ist $n > 1$ und kann x_n Werte aus $\{0, \ldots, \ell_n\}$ annehmen, so beachten wir zunächst, daß für jeden dieser Werte a auch

$$\left\{(x_1, \ldots, x_{n-1}) \in \mathbb{N}^{n-1} \;\middle|\; (x_1, \ldots, x_{n-1}, a) \in D\right\}$$

ein Stufendesign ist, also Nullstellenmenge einer Menge $\mathcal{M}_a$ von Distraktionen von Monomen in $X_1, \ldots, X_{n-1}$. Offensichtlich ist dann D die Nullstellenmenge der Polynome $D(MX_n^{a+1})$ mit $M \in \mathcal{M}_a$ für $a = 0, \ldots, \ell_n$ und von $D(X_n^{\ell_n+1})$.

∎

Man beachte, daß dieser Beweis konstruktiv ist und daß die Distraktionen nach dem vorigen Satz eine GRÖBNER-Basis des von ihnen erzeugten Ideals bilden. Wir können damit für Stufendesigns ohne den Umweg über den BUCHBERGER-Algorithmus eine GRÖBNER-Basis des zugehörigen Ideals finden. Da diese sogar universell ist, folgt

Korollar: Jedes Stufendesign hat minimalen Fächer.

∎

§4: Fraktionen eines vollen faktoriellen Designs

Nun sei $\mathcal{F}$ eine Teilmenge eines vollen faktoriellen Designs D, und G sei die reduzierte universelle GRÖBNER-Basis von $I(D)$. Dann ist $I(D)$ eine Teilmenge von $I(\mathcal{F})$; es gibt also ein Erzeugendensystem von $I(\mathcal{F})$, das G enthält. Wir können versuchen, das zur Berechnung von Modellen zu $\mathcal{F}$ zu verwenden. Ein erstes technisches Hilfsmittel dazu ist das folgende

Lemma: $\mathcal{O} \subset \mathbb{T}$ sei ein Ordnungsideal. Dann gibt es eine eindeutig bestimmte minimale Teilmenge Min($\mathcal{O}$) von $\mathbb{T}$ derart, daß $\mathbb{T} \setminus \mathcal{O}$ genau aus den Vielfachen der Monome aus Min($\mathcal{O}$) besteht.

Beweis: Das von $\mathbb{T} \setminus \mathcal{O}$ erzeugte monomiale Ideal $I \lhd k[X_1, \ldots, X_n]$ enthält kein Monom aus $\mathcal{O}$, denn wie wir wissen, enthält ein monomiales Ideal genau die Monome, die durch eines der erzeugenden Monome teilbar sind. Da ein Ordnungsideal mit jedem Monom auch dessen sämtliche Teiler enthält, müßte im Falle eines Monoms aus $\mathcal{O}$ in I das Erzeugendensystem $\mathbb{T} \setminus \mathcal{O}$ ein Element von $\mathcal{O}$ enthalten, was natürlich absurd ist.

Nach dem Lemma von DICKSON hat I ein Erzeugendensystem aus endlich vielen Monomen. Ein solches Erzeugendensystem ist minimal, wenn keines der Monome dort durch ein anderes teilbar ist. Ein gegebenes endliches Erzeugendensystem läßt sich problemlos auf ein minimales reduzieren; also gibt es minimale Erzeugendensysteme.

Angenommen, $\{M_1, \ldots, M_p\}$ und $\{N_1, \ldots, N_s\}$ sind zwei solche minimale Erzeugendensysteme. Da jedes N_i im von den M_j erzeugten monomialen Ideal I liegt, muß N_i durch (mindestens) ein M_j teilbar sein. Da I auch von $N_1, \ldots, N_s$ erzeugt wird, muß umgekehrt M_j durch ein N_ℓ teilbar sein, d.h. $N_\ell | M_j | N_i$. Da wir nur minimale Erzeugendensysteme betrachten, folgt $N_\ell = N_i$, also insbesondere $N_i = M_j$. Jedes N_i liegt also im Erzeugendensystem $\{M_1, \ldots, M_p\}$, und da dies ein minimales Erzeugendensystem ist, folgt $\{M_1, \ldots, M_p\} = \{N_1, \ldots, N_q\}$, d.h. es gibt genau ein minimales Erzeugendensystem von I. Da die Monome in I genau die aus $\mathbb{T} \setminus \mathcal{O}$ sind, besteht diese Menge somit genau aus den Vielfachen der Monome aus $\{M_1, \ldots, M_p\}$, so daß wir $\mathrm{Min}(\mathcal{O}) = \{M_1, \ldots, M_p\}$ setzen können.

∎

Falls $\mathcal{O}$ Teilmenge von $\mathrm{Est}_\tau(D)$ für ein volles faktorielles Design D ist, können wir das lineare Modell zu $\mathcal{O}$ natürlich auf Grund von D bestimmen; falls $\mathcal{O}$ allerdings deutlich kleiner als $\mathrm{Est}_\tau(D)$ ist, sollte das auch mit deutlich geringerem Aufwand möglich sein, nämlich mit jeder Fraktion $\mathcal{F}$ von D mit $\mathrm{Est}_\tau(\mathcal{F}) \supseteq \mathcal{O}$. Es genügt, wenn wir die minimalen unter diesen Fraktionen bestimmen, denn jede andere enthält mindestens eine von diesen.

Via GRÖBNER-Basen und Monomordnungen lassen sich (wenn auch mit beträchtlichem Aufwand) alle diese Fraktionen bestimmen; wir wollen uns hier aber mit einem weniger ambitionierten Ziel begnügen und nur die Fraktionen bestimmen, die Nullstellenmengen von Idealen sind, die von Distraktionen von Polynomen erzeugt werden. Die Theorie dazu findet man bei

> LORENZO ROBBIANO, MARIA PIERA ROGANTIN: Full factorial designs and distracted fractions *in* BRUNO BUCHBERGER, FRANZ WINKLER [HRSG]: Gröbner bases and applications Linz, *Cambridge University Press,* 1998, Seite 473–482

wo auch Aussagen über die Anzahl der so zu findenden und der insgesamt existierenden Fraktionen $\mathcal{F}$ zu finden sind. Ansätze zur allgemeinen Lösung des Problems findet man in

> MASSIMO CARBOARA, LORENZO ROBBIANO: Families of Ideals in Statistics *in* KÜCHLIN [HRSG.]: Proceedings of the 1997 ISSAC, *ACM Press,* 1997, Seite 404–409

Wir gehen also aus von einem vollständigen faktoriellen Design D und wollen dort Teilmengen $\mathcal{F}$ bestimmen, die Nullstellenmengen von Distraktionen sind. Da wir D kennen, brauchen wir dazu nicht das volle Ideal $I(\mathcal{F})$, sondern es reicht, wenn wir Funktionen finden, die auf D genau in den Punkten von $\mathcal{F}$ verschwinden. Etwaige gemeinsame Nullstellen außerhalb von D interessieren uns nicht, und natürlich sind auch Polynome, die auf ganz D verschwinden, für uns uninteressant.

Zur Konstruktion verwenden wir Distraktionen der Monome aus $\mathrm{Min}(\mathcal{O})$, und die verschwinden genau dann sogar auf ganz D, wenn das zugehörige Monom bereit im Ideal $\mathrm{FM}_\tau\big(I(D)\big)$ der führenden Monome von $I(D)$ liegen. Da uns diese Monome nicht interessieren, entfernen wir sie aus $\mathrm{Min}(\mathcal{O})$:

Definition: $\mathrm{CutOut}(\mathcal{O}) = \mathrm{Min}(\mathcal{O}) \setminus \mathrm{FM}_\tau\big(I(D)\big)$

Damit gilt

Satz: $D = S_1 \times \cdots \times S_n$ sei ein volles faktorielles Design, G sei die universelle reduzierte GRÖBNER-Basis von $I(D)$, und $\mathcal{O}$ sei ein Ordnungsideal, das in $\mathrm{Est}_\tau(D)$ liegt mit $\mathrm{CutOut}(\mathcal{O}) = \{M_1, \ldots, M_r\}$. Die Vektoren $a^{(i)}$ seien so definiert, daß ihre Komponenten gleich den verschiedenen Elementen von S_i in irgendeiner Reihenfolge sind. Dann bildet G zusammen mit den Polynomen $D(M_1), \ldots, D(M_r)$ bezüglich jeder Monomordnung τ eine GRÖBNER-Basis eines Designideals einer Fraktion $\mathcal{F}$ von D mit $\mathrm{Est}_\tau(\mathcal{F}) = \mathcal{O}$.

Beweis: Wie wir bereits wissen, sind die Elemente der GRÖBNER-Basis G von $I(D)$ die Distraktionen der Monome $X_i^{\#S_i}$ bezüglich der $a^{(i)}$. Daher besteht die Vereinigung von G mit der Menge der M_j nur aus Distraktionen von Monomen bezüglich fester Vektoren und ist damit

nach obigem Satz bezüglich jeder Monomordnung eine GRÖBNER-Basis des davon erzeugten Ideals I. Für $\mathcal{F} = V(I)$ wird $\mathrm{FM}_\tau(I)$ erzeugt von den führenden Monomen der Elemente von G und den M_i, also von CutOut($\mathcal{O}$) und $\mathrm{FM}_\tau\big(I(D)\big)$ und damit von $\mathrm{Min}_\tau(\mathcal{O})$. Somit ist $\mathrm{Est}_\tau(\mathcal{F}) = \mathcal{O}$.

∎

Betrachten wir als Beispiel $D = \{0,1,2\} \times \{0,1\}^{10}$, das volle faktorielle Design für elf Faktoren, deren erster drei Stufen hat; alle weiteren haben nur zwei. D enthält somit $3 \times 2^{10} = 3\,072$ Punkte aus $\mathbb{R}^{11}$.

Wir erwarten nicht, daß jede der 3 072 Faktorkombinationen relevant ist und beschränken und auf das Modell

$$\mathcal{O} = \{1, X_1, \ldots, X_{11}, X_1X_2, X_1X_3, X_1X_4, X_1X_2X_3, X_2X_3, X_1^2\}\,.$$

Es gibt also nur eine Dreierinteraktion, und nur ein Monom kommt als Quadrat vor. Die bezüglich der Teilbarkeitsrelation maximalen Elemente von $\mathcal{O}$ sind $X_5, \ldots, X_{11}, X_1X_2X_3, X_1X_4$ und X_1^2. Die Menge Min($\mathcal{O}$) enthält somit alle Monome der Form MX_i, wobei M irgendeines dieser Monome ist.

$\mathrm{FM}_\tau\big(I(D)\big)$ wird erzeugt von X_1^3 und $X_2^2, \ldots, X_{10}^2$; CutOut($\mathcal{O}$) besteht also aus den davon verschiedenen Monomen aus Min($\mathcal{O}$).

Um kurze Polynome zu bekommen, wählen wir $a^{(1)} = (0,1,2)$ und $a^{(i)} = (0,1)$ für $i \geq 2$, d.h. wir stellen die Null an die erste Stelle. Die Distraktionen der Monome sind dann für alle Monome, die kein Quadrat enthalten, einfach die Monome selbst, und für die $X_1^2X_i$ sind es die Polynome $X_1(X_1-1)X_i$.

Der Nullpunkt und alle Punkte, bei denen genau eine Koordinate von Null verschieden ist, sind Nullstellen aller dieser Polynome; da x_1 auch den Wert zwei annehmen kann, sind dies schon einmal dreizehn Lösungen.

Falls ein x_i mit $i \geq 5$ von Null verschieden ist, müssen alle anderen x_j verschwinden, da X_iX_j eines der Erzeugenden des Ideals ist. Wegen der Polynome $D(X_1^2X_j) = X_1(X_1-1)X_j$ müssen auch im Falle $x_1 = 2$ alle x_j mit $j > 1$ verschwinden.

Ist $x_4 = 1$, so kann wegen der Polynome $X_1X_4X_j$ auch $x_1 = 1$ sein, falls alle anderen x_j verschwinden; es gibt also außer dem Einheitspunkt noch den Punkt $(1, 0, 0, 1, 0, \ldots, 0)$.

Wenn alle x_i mit $i \geq 4$ verschwinden, ist für x_1 bis x_3 jede Kombination aus Nullen und Einsen möglich; dies ergibt acht Lösungen, von denen vier neu sind. Insgesamt haben wir also 18 Punkte, was erwartungsgemäß gleich der Elementanzahl von $\mathcal{O}$ ist.

§5: Berechnung der Est-Mengen

Nachdem wir uns im letzten Abschnitt damit beschäftigt haben, zu einem Ordnungsideal (und damit zu einem statistischen Modell) ein Design zu finden, mit dem sich die Modellparameter schätzen lassen, wollen wir nun zurückkehren zum Hauptthema dieser Vorlesung, also der Frage, welche Modelle wir auf der Grundlage eines vorgegebenen Design identifizieren können. Wir gehen also wieder aus von einem Design $D = \{x^{(1)}, \ldots, x^{(r)}\} \subset k^n$ mit Ideal $I(D) \lhd k[X_1, \ldots, X_n]$.

Beginnen wir mit eine festen Monomordnung τ auf $k[X_1, \ldots, X_n]$. Mit den Methoden, die uns bislang zur Verfügung stehen, müssen wir zur Berechnung von $\mathrm{Est}_\tau(D)$ zunächst eine GRÖBNER-Basis von $I(D)$ bezüglich τ bestimmen und dann die Standardmonome dazu. Wir wollen uns überlegen, wie wir alternativ diese Standardmonome auch berechnen können, indem wir das Design sukzessive aufbauen. Dazu sei

$$D_m = \{x^{(1)}, \ldots, x^{(m)}\} \quad \text{für} \quad m = 1, \ldots, r\,,$$

und wir wollen nacheinander die Mengen $\mathrm{Est}_\tau(D_m)$ bestimmen.

Wie wir wissen, ist jede Est-Menge ein Ordnungsideal, enthält also die Eins, und außerdem hat $\mathrm{Est}_\tau(D_m)$ genauso viele Elemente wie D_m, also m Stück. Somit ist auf jeden Fall $D_1 = \{1\}$.

Lemma: Ist D ein Design und $D' \subset D$, so ist $\mathrm{Est}_\tau(D') \subset \mathrm{Est}_\tau(D)$.

Beweis: Für ein Design D besteht $\mathrm{Est}_\tau(D)$ aus allen Monomen, die durch kein führendes Monom eines Elements einer GRÖBNER-Basis teilbar sind. Da diese führenden Monome das Ideal der führenden Monome von $I(D)$ erzeugen, ist dies äquivalent dazu, daß die Monome aus $\mathrm{Est}_\tau(D)$ durch kein Monom aus $\mathrm{FM}_\tau\big(I(D)\big)$ teilbar sein dürfen.

Für eine Teilmenge $D' \subset D$ ist $I(D)$ eine Teilmenge von $I(D')$, also ist auch $\mathrm{FM}_\tau\big(I(D)\big)$ eine Teilmenge von $\mathrm{FM}_\tau\big(I(D')\big)$. Falls ein Monom durch keines der Monome aus $\mathrm{FM}_\tau\big(I(D')\big)$ teilbar ist, kann es also erst recht durch keines aus $\mathrm{FM}_\tau\big(I(D)\big)$ teilbar sein, was die Behauptung beweist. ∎

Damit ist also $\mathrm{Est}_\tau(D_{m-1}) \subset \mathrm{Est}_\tau(D_m)$, und $\mathrm{Est}_\tau(D_m)$ hat genau ein Element mehr. Es gibt also ein Monom $M \notin \mathrm{Est}_\tau(D_{m-1})$, so daß $\mathrm{Est}_\tau(D_m) = \mathrm{Est}_\tau(D_{m-1}) \cup \{M\}$ ist. In §1 hatten wir zu einem Design D und einer Menge $\mathbb{F}$ von Funktionen f_i die Z-Matrix definiert als die Matrix mit Einträgen $z_{ij} = f_i\big(x^{(j)}\big)$. Da das lineare Modell mit den Monomen aus $\mathrm{Est}_\tau(D_m)$ als Basisfunktionen anhand von D_m eindeutig identifizierbar ist, muß die Z-Matrix zu D_m und $\mathrm{Est}_\tau(D_m)$ invertierbar sein. Außerdem ist $\mathrm{Est}_\tau(D_m)$ ein Ordnungsideal; daher muß M Produkt eines Monoms aus D_{m-1} mit einer der Variablen X_i sein.

Durch diese Bedingungen muß M noch nicht eindeutig bestimmt sein. Angenommen, es gibt zwei Monome M und M' derart, daß sowohl $\mathrm{Est}_\tau(D_{m-1}) \cup \{M\}$ als auch $\mathrm{Est}_\tau(D_{m-1}) \cup \{M'\}$ Ordnungsideale sind und für D_m auf invertierbare Z-Matrizen führen. Dann können wir die Monome M und M' bezüglich der Monomordnung τ miteinander vergleichen; M sei das kleinere der beiden Monome.

Wäre $\mathrm{Est}_\tau(D_m) = \mathrm{Est}_\tau(D_{m-1}) \cup \{M'\}$, so wäre M kein Standardmonom, wäre also teilbar durch das führende Monom eines Elements g der reduzierten GRÖBNER-Basis von $I(D_m)$ bezüglich τ. Da $\mathrm{Est}_\tau(D_{m-1}) \cup \{M\}$ ein Ordnungsideal ist, liegt jeder echte Teiler von M in $\mathrm{Est}_\tau(D_{m-1})$ und damit erst recht in $\mathrm{Est}_\tau(D_m)$. Somit muß M das führende Monom von g sein. Alle anderen Monome in g sind Standardmonome, liegen also in $\mathrm{Est}_\tau(D_{m-1}) \cup \{M'\}$. Das Monom M' kann allerdings nicht vorkommen, da das bezüglich τ kleinere Monom M führendes Monom von g ist. Somit ist g eine Linearkombination von Monomen aus dem Ordnungsideal $\mathrm{Est}_\tau(D_{m-1}) \cup \{M\}$, und als Element von $I(D_m)$ verschwindet g natürlich auf ganz D_m. Wegen der Invertierbarkeit der Z-Matrix muß daher $g = 0$ sein, was für ein Element einer reduzierten GRÖBNER-Basis unmöglich ist. Also kann $\mathrm{Est}_\tau(D_m)$ nicht gleich $\mathrm{Est}_\tau(D_{m-1}) \cup \{M'\}$ sein.

Somit ist $\mathrm{Est}_\tau(D_m) = \mathrm{Est}_\tau(D_{m-1}) \cup \{M\}$ für das bezüglich τ kleinste Monom M, für das $\mathrm{Est}_\tau(D_{m-1}) \cup \{M\}$ ein Ordnungsideal ist und eine invertierbare Z-Matrix über D_m hat.

Dies führt auf folgenden Algorithmus zur Berechnung von $\mathrm{Est}_\tau(D)$:

1. Schritt: Setze $\mathrm{Est}_\tau(D_1) = \{1\}$

m-ter Schritt, $m \geq 2$: Setze $\mathrm{Est}_\tau(D_m) = \mathrm{Est}_\tau(D_{m-1}) \cup \{M\}$ für das bezüglich τ kleinste Monom M, für das $\mathrm{Est}_\tau(D_{m-1}) \cup \{M\}$ ein Ordnungsideal ist und eine invertierbare Z-Matrix über D_m hat.

Betrachten wir etwa $D = \{(0,0), (0,2), (1,1)\}$ für die lexikographische Ordnung mit $X > Y$ und die angegebene Reihenfolge der Punkte. Für $D_1 = \{(0,0)\}$ ist natürlich $\mathrm{Est}_\tau(D_1) = \{1\}$. Dies läßt sich auf zwei Arten zu einem zweielementigen Ordnungsideal erweitern: Entweder zu $\{1, X\}$ oder zu $\{1, Y\}$. Da beide Punkte von $D_2 = \{(0,0), (0,2)\}$ die gleiche x-Koordinate haben, führt $\{1, X\}$ zu einer singulären Z-Matrix; wir müssen also $\{1, Y\}$ nehmen, wofür wir die invertierbare Z-Matrix $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 2 \end{smallmatrix}\right)$ bekommen.

Dies wiederum läßt sich auch auf zwei Arten erweitern: Entweder zu $\{1, Y, Y^2\}$ oder zu $\{1, Y, X\}$. Wir erhalten bezüglich D die beiden nichtsingulären Z-Matrizen

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 4 \\ 1 & 0 & 1 \end{pmatrix}.$$

Da wir mit der lexikographischen Ordnung mit $X > Y$ arbeiten, ist $X > Y^2$, wir müssen also $\{1, Y, Y^2\}$ nehmen. Hätten wir stattdessen mit einer graduiert lexikographischen Ordnung gearbeitet, egal ob für $X > Y$ oder $Y > X$, wäre $Y^2 > X$, so daß wir uns dann für $\{1, Y, X\}$ entscheiden müßten. Dies illustriert, daß die lexikographische Ordnung Modelle bevorzugt, in denen eine (oder mehrere) „kleine“ Variablen dominieren, während die graduiert lexikographische Ordnung zu ausgewogeneren Modellen führt,wobei freilich auch hier die Reihenfolge der Variablen zusätzliche Prioritäten setzt. Andere Monomordnungen können zu Modellen mit wieder anderen Eigenschaften führen. Welche in einem konkreten Zusammenhang nützliche sind, hängt natürlich von

der Anwendung ab: Manchmal dominiert eine der Eingabevariablen, in anderen Zusammenhängen sind alle ziemlich gleichberechtigt, und oft kommt es auch auf das Zusammenspiel verschiedener Variablen, d.h. auf gemischte Terme an. Die Tatsache, daß man anhand desselben Designs D ganz verschiedene Modelle identifizieren kann, bezeichnet man in der Statistik als *confounding* (Verwechseln). Welches der verschiedenen Modelle man in einem konkreten Zusammenhang nehmen sollte, ist keine mathematische Frage, sondern hängt ab vom eventuell vorhandenen Vorwissen über den zu untersuchenden Sachverhalt. Wenn wir zwei Modelle anhand von D und einer Funktion $D \to k$ identifiziert haben, muß die Differenz der beiden entstehenden Polynome natürlich im Designideal $I(D)$ liegen.

Genau wie man in der Numerik bei der Approximation einer transzendenten Funktion praktisch nie ein Interpolationspolynom betrachtet, das an vielen vorgegebenen Stellen exakt den korrekten Wert liefert, sucht man auch in der Statistik praktisch nie nach Modellen, die eine gemessene oder etwa durch Umfragen erhaltene Funktion $D \to k$ exakt darstellt: Stattdessen betrachtet man Modelle, die nur auf einer oft deutlich kleineren Teilmenge von $\mathrm{Est}_\tau(D)$ beruhen. Welche das sind, ist wieder teils abhängig von außermathematischem Vorwissen, teils hilft auch nur Experimentieren mit verschiedenen Ansätzen, für die man dann beispielsweise nach der Methode der kleinsten Quadrate entscheiden kann, welche davon gute Ergebnisse liefern.

Im Falle eines Designs mit minimalem Fächer ist das relativ einfach, da wir mit Teilmengen eines festen Ordnungsideals arbeiten können. Leider führt das aber auch nur auf wenige identifizierbare Modelle. Wenn wir kein Vorwissen über die Art des zu erwartenden Modells haben, ist es daher nützlicher, mit Designs zu arbeiten, deren Fächer möglichst groß ist.

Definition: *a)* Für zwei natürliche Zahlen r, n sei $\mathcal{G}(r, n)$ die Menge aller Ordnungsideale mit r Elementen in n Variablen.
b) Ein Design $D \subset k^n$ aus r Punkten heißt *Design mit maximalem Fächer,* wenn für jedes Ordnungsideal aus $\mathcal{G}(r, n)$ die Z-Matrix über D invertierbar ist.

$\mathcal{G}(r,n)$ ist natürlich stets eine endliche Menge, denn da ein Ordnungsideal mit jedem Element auch dessen sämtliche Teiler enthält, besteht jedes $\mathcal{O} \in \mathcal{G}(r,n)$ aus Monomen vom Grad kleiner r, und da die Menge aller Monome in n Variablen vom Grad kleiner r endlich ist, hat sie auch nur endlich viele Teilmengen.

Die Menge alle Designs aus r Punkten aus k^n ist $\left(k^n\right)^r$, was wir mit k^{nr} identifizieren können. Für jedes $\mathcal{O} \in \mathcal{G}(r,n)$ und jedes Design $D \in k^{rn}$ sind die Einträge der Z-Matrix Polynomfunktionen in den rn Koordinaten, und damit ist auch die Determinante der Z-Matrix eine Polynomfunktion auf k^{rn}. Die Designs mit einer singulären Z-Matrix bezüglich $\mathcal{O}$ liegen somit auf einer Hyperfläche (d.h. der Nullstellenmenge eines Polynoms) in k^{nr}, und die Menge aller Designs, die für *irgendein* $\mathcal{O} \in \mathcal{G}(r,n)$ zu eine singulären Z-Matrix führen, liegt in der Vereinigung endlich vieler solcher Hyperflächen.

Speziell im Fall $k = \mathbb{R}$ wissen wir, daß die Nullstellenmenge eines von Null verschiedenen Polynoms bezüglich des LEBESGUE-Maßes eine Nullmenge ist; die Designs aus $\mathbb{R}^{rn}$, die *keinen* maximalen Fächer haben, liegen also in einer Nullmenge. Wenn wir also zufällig ein Design aus $\mathbb{R}^{rn}$ oder (realistischer) einen beschränkten rn-dimensionalen Quader wählen, haben wir fast sicher ein Design mit maximalem Fächer. Leider hängt dieses Argument ganz entscheidend damit zusammen, daß wir über den rechnerisch eher unzugänglichen reellen Zahlen arbeiten; für uns interessante Teilmengen wie $\mathbb{Z}^{rn}$ und $\mathbb{Q}^{rn}$ sind allesamt Nullmengen, so daß wir dafür überhaupt keine Aussage bekommen. Soweit mir bekannt, gibt es nicht einmal eine Aussage darüber, ob es für jedes Paar (r,n) überhaupt ein Design $D \subset \mathbb{Z}^{rn}$ oder $D \subset \mathbb{Q}^{rn}$ mit maximalem Fächer gibt. Obwohl jedes „zufällig“ aus $\mathbb{R}^{rn}$ gewählte Design maximalen Fächer hat, gibt es meines Wissens auch keinen Algorithmus, der zu gegebenen Werten von r und n auch nur ein solches Design konstruiert.

Wie zu Beginn der Vorlesung bereits erwähnt, gibt es allerdings mathematische Methoden, oftmals algebraischer Art, zur Konstruktion von Designs, anhand derer man möglichst viel Information gewinnen kann. Diese zählen nicht zur algebraischen Statistik und sind vielfach auch deutlich älter als diese. Stattdessen bilden sie ein eigenes Teilgebiet der

Mathematik, das der optimalen Versuchsplanung. Solche Algorithmen können Designs mit nichtminimalem Fächer liefern, und wir möchten auch dazu die eindeutig identifizierbaren Modelle finden.

Glücklicherweise läßt sich der obige Algorithmus zur Bestimmung von $\mathrm{Est}_\tau(D)$ leicht umformulieren zu einem Algorithmus zur Bestimmung *aller* Ordnungsideale, die anhand von D identifiziert werden können.

Auch hier gehen wir wieder punktweise vor, und natürlich ist wieder $\mathrm{Est}_\tau(D_1) = \{1\}$. In den nächsten Schritten betrachten wir nun aber nicht nur das um ein Element größere Ordnungsideal mit dem bezüglich τ kleinstmöglichen neuen Monom, sondern wir betrachten *alle* Ordnungsideale, die durch Hinzufügen eines neuen Monoms entstehen und zu einer invertierbaren Z-Matrix führen. Für jedes der so entstehenden Ordnungsideale werden dann auch alle Folgeschritte durchgeführt. Dies kann zu einem exponentiell ansteigenden Aufwand führen, allerdings werden gelegentlich auch Ordnungsideale entstehen, die bereits vorher bekannt waren, da bei ihrer Konstruktion die gleichen Monome in verschiedener Reihenfolge auftraten. Falls wir diesen Algorithmus vollständig durchühren, erhalten wir am Ende eine Liste aller Ordnungsideale mit $\#D$ Elementen, die anhand von D identifiziert werden können.

Insbesondere können wir so auch entscheiden, ob ein Design maximalen Fächer hat oder nicht. Da sich $\mathcal{G}(r,n)$ problemlos bestimmen läßt, können wir leicht feststellen, ob alle seine Elemente in der durch den Algorithmus berechneten Liste vorkommen

§6: Nichtpolynomiale Modelle

Nicht alle Phänomene lassen sich polynomial modellieren: Bei periodischen Prozessen oder bei exponentiellem Wachstum etwa kann ein polynomialer Ansatz bestenfalls einen beschränkten Zeitraum approximativ beschreiben, und auch das nur mit großem Aufwand. Trotzdem läßt sich die hier behandelte Theorie manchmal auch in solchen Fällen anwenden.

Am einfachsten geht das im Falle von Modellen, deren Basisfunktionen geeignete Exponentialfunktionen sind: Wir betrachten für n feste,

über $\mathbb{Q}$ linear unabhängige reelle Zahlen $\lambda_1, \ldots, \lambda_n$ Funktionen, die Linearkombinationen von Termen der Form

$$e^{a_1\lambda_1 x_1 + \cdots + a_n\lambda_n x_n}$$

mit ganzen Zahlen a_i sind. Mit der Abkürzung Y_i für die Funktion

$$(x_1, \ldots, x_n) \mapsto e^{\lambda_i x_i}$$

können wir so eine Funktion auch als Linearkombination von „Monomen“ $Y_1^{a_1} \cdots Y_n^{a_n}$ schreiben, wobei allerdings im Gegensatz zu den „echten“ Monomen auch negative Exponenten a_i zugelassen sind. Damit befinden wir uns fast im Polynomring $k[Y_1, \ldots, Y_n]$, genauer gesagt in dessen Erweiterung

$$k[Y_1, \ldots, Y_n, Y_1^{-1}, \ldots, Y_n^{-1}]\,,$$

die wir vor allem deshalb brauchen, weil eine Exponentialfunktion und damit auch die Funktion Y_i den Wert Null nicht annehmen kann. Ansonsten können wir in diesem Ring weitgehend so rechnen, wie wir es von Polynomringen gewohnt sind und damit die in diesem Kapitel betrachtete Theorie auch auf solche Modelle erweitern.

Wenn wir den Körper k durch die imaginäre Einheit i erweitern zu $k(i) = k \oplus ki$, können wir auf diese Weise über die EULERschen Formeln

$$\cos x = \frac{e^{ix} + e^{-ix}}{2}, \quad \sin x = \frac{e^{-ix} - e^{-ix}}{2i}, \quad e^{ix} = \cos x + i \sin x$$

auch Modelle für periodische Zusammenhänge behandeln.

Kapitel 4
Markov-Basen für Kontingenztests

Zu den Grundaufgaben der Statistik gehört auch die Frage nach dem Zusammenhang zwischen zwei oder mehreren Zufallsvariablen. Um beispielsweise zu testen, ob ein Medikament besser ist als ein anderes, teilt man die Probanden zufallsgesteuert ein in zwei Kontrollgruppen, die mit je einem der beiden Medikamente behandelt werden, und mißt den Erfolg. Falls die beiden Zufallsvariablen *Medikament* und *Erfolg* voneinander unabhängig sind, haben beide Medikamente gleich viel (oder wenig) Erfolg, andernfalls läßt sich mit einer gewissen Wahrscheinlichkeit eines der beiden als das bessere identifizieren.

§1: Kontingenztafeln

Für zwei unabhängige Zufallsvariablen X und Y mit Werten in $\{1,\dots,r\}$ *bzw.* $\{1,\dots,c\}$ ist $p(X=x,Y=y)=p(X=x)p(Y=y)$ für alle $(x,y)\in\{1,\dots,r\}\times\{1,\dots,c\}$. Wegen der Beziehungen

$$p(X=x)=\sum_{y=1}^{c}p(X=x,Y=y)$$

und

$$p(Y=y)=\sum_{x=1}^{r}p(X=x,Y=y)$$

können wir das auch so ausdrücken, daß $p(X=x,Y=y)$ nur abhängt von den beiden Summen

$$\sum_{y=1}^{c}p(X=x,Y=y)\quad\text{und}\quad\sum_{x=1}^{r}p(X=x,Y=y)\,.$$

Um die Unabhängigkeit der beiden Variablen experimentell zu prüfen, wählen wir eine natürliche Zahl n und betrachten n Werte (x_k, y_k) der Zufallsvariablen $X \times Y$. Für die zugehörige Kontingenztafel definieren wir eine neue Zufallsvariable $U = (U_{ij})_{\substack{i=1\dots r\\ j=1\dots c}}$, wobei U_{ij} die Anzahl von Paaren (x_k, y_k) mit $x_k = i$ und $y_k = j$ bezeichnet. Außerdem betrachten wir noch die Zufallsvariablen

$$U_{i+} = \sum_{j=1}^{c} U_{ij} \quad \text{und} \quad U_{+j} = \sum_{i=1}^{r} U_{ij}$$

für $i = 1, \dots, r$ und $j = 1, \dots, c$. Die Erwartungswerte der Zufallsvariablen U_{ij} sind dann

$$\mathbb{E}(U_{ij}) = np(X = i, Y = j) = n \cdot \sum_{j=1}^{c} p(X = i, Y = j) \cdot \sum_{i=1}^{r} p(X = i, Y = j)$$

Entsprechend sind die Erwartungswerte von U_{i+} und U_{+j} gleich

$$\mathbb{E}(U_{i+}) = n \cdot \sum_{j=1}^{c} p(X = i, Y = j) \quad \text{und} \quad \mathbb{E}(U_{+j}) = n \cdot \sum_{i=1}^{r} p(X = i, Y = j)\,.$$

Durch Vergleich der Erwartungswerte folgen die Beziehungen

$$\mathbb{E}(U_{ij}) = \frac{\mathbb{E}(U_{i+})\mathbb{E}(U_{+j})}{n}\,,$$

$$\mathbb{E}(U_{i+}) = \sum_{j=1}^{c} \mathbb{E}(U_{ij}) \quad \text{und} \quad \mathbb{E}(U_{+j}) = \sum_{i=1}^{r} \mathbb{E}(U_{ij})\,.$$

Natürlich können wir schon wegen der Ganzzahligkeit der u_{ij} nicht erwarten, daß für einen beobachteten Wert u von U auch $u_{ij} = u_{i+}u_{+j}/n$ ist, aber die Abweichung zwischen den beiden Seiten sollte mit großer Wahrscheinlichkeit klein sein. Als Maß der Abweichung wählen wir die Zahl

$$\chi^2(u) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(u_{ij} - \hat{u}_{ij})^2}{\hat{u}_{ij}} \quad \text{mit} \quad \hat{u}_{ij} = \frac{u_{i+}u_{+j}}{n}\,.$$

In der Statistik zeigt man, daß die Verteilung der Zufallsvariablen $\chi^2(U)$ asymptotisch gegen eine χ^2-Verteilung mit $(r-1)(c-1)$ Freiheitsgraden konvergiert, falls alle u_{ij} gegen unendlich gehen.

Leider können aus praktischen Gründen oft nur Experimente realisiert werden, bei denen zumindest einige der u_{ij} recht klein sind. Eine Faustregel besagt, daß man definitiv nicht mit der χ^2-Verteilung arbeiten sollte, wenn nicht alle u_{ij} mindestens gleich fünf sind.

§2: Fishers exakter Test

Das erste Verfahren, die Wahrscheinlichkeit für die Zufälligkeit der Abweichung der Werte von U_{ij} und $U_{i+}U_{+j}/n$ auch im Falle kleiner Werte einiger u_{ij} zu schätzen, war FISHERs exakter Test. Er betrachtet zu einer gegebenen Kontingenztafel (u_{ij}) alle Tafeln (v_{ij}) mit $v_{ij} \in \mathbb{N}_0$ und $v_{i+} = u_{i+}$ für alle i sowie $v_{+j} = u_{+j}$ für alle j. Für jede dieser Tafeln berechnet er das zugehörige χ^2 und schätzt die Wahrscheinlichkeit für die Zufälligkeit der Abweichung als den Anteil aller Tafeln, die zu keinem größeren χ^2 führen als die gegebene Tabelle (u_{ij}). Im Falle von Vierfeldertests ist das auch noch bei moderat großen Zeilen- und Spaltensummen praktikabel, denn offensichtlich sind durch diese Summen alle v_{ij} eindeutig festgelegt, sobald man einen dieser vier Werte kennt. Mit wachsender Zahl der Freiheitsgrade steigt allerdings auch der Aufwand für FISHERs exakten Test dramatisch an, so daß die Betrachtung aller Tabellen mit denselben Randverteilungen wie die gegebene Tabelle nicht mehr mit realistischen Aufwand möglich ist.

§3: Log-lineare Modelle

Bevor wir uns überlegen, wie wir in solchen Fällen vorgehen können, wollen wir zunächst die betrachtete Situation etwas verallgemeinern. Bei der Untersuchung auf Unabhängigkeit sollten die Erwartungswerte der U_{ij} nur abhängen von denen der U_{i+} und der U_{+j}. Auch im Falle abhängiger Größen kann es sein, daß es eine begrenzte Zahl von Linearkombinationen der U_{ij} gibt mit der Eigenschaft, daß die Erwartungswerte aller U_{ij} aus denen dieser Linearkombinationen berechnet werden können. Solche Situationen formalisiert der Begriff eines log-linearen Modells:

Definition: $X_1, \ldots, X_m$ seien Zufallsvariablen, und X_ℓ nehme Werte

aus der Menge $\{1, \ldots, r_\ell\}$ an. Wir setzen

$$\mathcal{R} \underset{\text{def}}{=} \{1, \ldots, r_1\} \times \cdots \times \{1, \ldots, r_m\}$$

und identifizieren für jede Menge $M \subseteq \mathbb{R}$ die Menge $M^{\mathcal{R}}$ der durch $\mathcal{R}$ indizierten Tabellen mit Einträgen aus M über irgendeine feste Anordnung von $\mathcal{R}$ mit $M^{\#\mathcal{R}}$. Für $i = (i_1, \ldots i_m) \in \mathcal{R}$ sei $p_i = p(X_1 = i_1, \ldots, X_m = i_m)$ die Tabelle der Wahrscheinlichkeiten dafür, daß die Variablen X_j die Werte i_j annehmen, identifiziert mit einem Vektor aus $(0, 1)^{\#\mathcal{R}}$. Weiter sei $A \in \mathbb{Z}^{d \times \#\mathcal{R}}$ eine Matrix, deren Spalten alle die gleiche Summe haben. Das log-lineare Modell $\mathcal{M}_A$ zur Matrix A ist die Menge aller Wahrscheinlichkeitstabellen $(p_i)_{i \in \mathcal{R}}$ mit der Eigenschaft, daß der Zeilenvektor $(\log p_i)_{i \in \mathcal{R}}$ im von den Zeilenvektoren von A aufgespannten Untervektorraum von $\mathbb{R}^{\#\mathcal{R}}$ liegt.

Daß wir hier die Logarithmen der Wahrscheinlichkeiten betrachten und nicht diese selbst, liegt natürlich daran, daß bei den hier interessierenden Modellen viele Produkte von Wahrscheinlichkeiten eine Rolle spielen. Durch Übergang zu Logarithmen können wir daraus Summen machen und damit insbesondere auch Methoden aus der Linearen Algebra anwenden.

Betrachten wir als Beispiel das obige Unabhängigkeitsmodell. Hier ist $\mathcal{R} = \{1, \ldots, r\} \times \{1, \ldots, c\}$ und

$$p_{(i,j)} = p(X = i, Y = j) = p(X = i) \cdot p(Y = j)\,.$$

Somit ist $\log p_{(i,j)} = \log p(X = i) + \log p(Y = j)$.

Ist A die $(r + c) \times rc$-Matrix, deren Spalte mit Index (i, j) in den Komponenten i und $r + j$ eine Eins stehen hat und sonst lauter Nullen, so ist

$$(\log p_{(1,1)}, \ldots, \log p_{(r,c)}) = \sum_{i=1}^{r} \log p(X = i) a^{(1)} + \sum_{j=1}^{c} \log p(Y = j) a^{(r+j)}\,,$$

wobei $a^{(\ell)}$ für den ℓ-ten Zeilenvektor von A steht. Für $r = 3$ und $c = 2$

etwa ist

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

und $(\log p_{(1,1)} \quad \log p_{(1,2)} \quad \log p_{(2,1)} \quad \log p_{(2,2)} \quad \log p_{(3,1)} \quad \log p_{(3,2)})$ ist gleich

$$\begin{array}{rlcccccc} & \log p(X=1) & \times & (1 & 1 & 0 & 0 & 0 & 0) \\ + & \log p(X=2) & \times & (0 & 0 & 1 & 1 & 0 & 0) \\ + & \log p(X=3) & \times & (0 & 0 & 0 & 0 & 1 & 1) \\ + & \log p(Y=1) & \times & (1 & 0 & 1 & 0 & 1 & 0) \\ + & \log p(Y=2) & \times & (0 & 1 & 0 & 1 & 0 & 1)\,. \end{array}$$

Wenn wir eine Kontingenztafel wie gerade eben als einen Vektor u auffassen, gibt uns das Matrixprodukt Au die sämtlichen Zeilen- und Spaltensummen; beim Unabhängigkeitsmodell hängen die Erwartungswerte der Variablen U_{ij} nur von diesen ab. Entsprechend erwarten wir bei einem beliebigen log-linearen Modell, daß auch dort die Erwartungswerte der Zufallsvariablen U_i, $i \in \mathcal{R}$, nur vom Zufallvsvariablenvektor AU abhängen sollten. Sobald wir wissen, wie sich die Erwartungswerte der U_i aus denen von AU berechnen lassen, können wir FISHERs exakten Test auch auf diese Situation verallgemeinern, indem wir zu einer gegebenen Tafel u aller Tafeln $v = (v_i)_{i\in\mathcal{R}}$ betrachten, für die $Av = Au$ ist.

Definition: Die *Faser* $\mathcal{F}(u)$ zu einer gegebenen Tafel $(u_i)_{i\in\mathcal{R}}$ bezüglich des Modells $\mathcal{M}_A$ ist die Menge aller $v \in \mathbb{N}_0^{\#\mathcal{R}}$ mit $Av = Au$.

Auch hier wird die Faser oft zu groß sein als daß wir χ^2 für jedes einzelne Element ausrechnen können, so daß wir uns mit einer Stichprobe begnügen müssen, die wir uns über eine Irrfahrt verschaffen wollen. Die einzelnen Schritte der Irrfahrt sind durch Elemente einer sogenannten MARKOV-Basis gegeben:

Definition: Eine MARKOV-Basis des log-linearen Modells $\mathcal{M}_A$ ist eine endliche Menge $\mathcal{B}$ von Tafeln $b \in \mathbb{Z}^{\#\mathcal{R}}$ mit $Ab = 0$, für die gilt: Für jede

Tafel $u \in \mathbb{N}_0^{\#\mathcal{R}}$ und je zwei Elemente $v, v' \in \mathcal{F}(u)$ gibt es eine Folge von Elementen $b_1, \ldots, b_L \in \mathcal{B}$, so daß gilt:
1.) Für $\ell = 1, \ldots, L$ ist $v + b_1 + \cdots + b_\ell \in \mathbb{N}_0^{\#\mathcal{R}}$
2.) $v + b_1 + \cdots + b_L = v'$

Der russische Mathematiker ANDREĬ ANDREEVIČ MARKOV (Андрей Андреевич Марков, 1856–1922) studierte in Sankt Petersburg, wo er später auch Professor wurde. Er beschäftigte sich zunächst hauptsächlich mit Zahlentheorie und Analysis; erst später folgen die wahrscheinlichkeitstheoretischen Arbeiten, für die er heute vor allem bekannt ist. Der Name Марков wird in lateinischen Buchstaben verschieden transkribiert; MARKOVs französische Arbeiten erschienen mit der Schreibweise MARKOFF; nach den klassischen deutschen Transkriptionsregeln müßte man MARKOW schreiben. Die Schreibweise MARKOV entspricht den englischen Regeln und scheint sich mittlerweile in der Mathematik ziemlich durchgesetzt zu haben.

Im Falle des Unabhängigkeitsmodells für zwei Zufallsvariablen können wir leicht eine MARKOV-Basis finden: Wir nehmen einfach alle Tafeln $b^{(ijk\ell)}$, $i, k = 1, \ldots, r$ und $j, \ell = i, \ldots, c$, deren Einträge allesamt verschwinden mit Ausnahme von

$$b_{ij}^{(ijk\ell)} = b_{k\ell}^{(ijk\ell)} = 1 \quad \text{und} \quad b_{i\ell}^{(ijk\ell)} = b_{kj}^{(ijk\ell)} = -1\,.$$

Für die klassischen Vierfeldertests mit $r = c = 2$ sind das gerade die beiden Tafeln

$$\begin{matrix} 1 & -1 \\ -1 & 1 \end{matrix} \quad \text{und} \quad \begin{matrix} -1 & 1 \\ 1 & -1 \end{matrix}\,,$$

für $r = 2$ und $c = 3$ haben wir die drei Tafeln

$$\begin{matrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{matrix} \quad \begin{matrix} 1 & 0 & -1 \\ -1 & 0 & 1 \end{matrix} \quad \text{und} \quad \begin{matrix} 0 & 1 & -1 \\ 0 & -1 & 1 \end{matrix}$$

sowie deren Negative.

Später in diesem Kapitel werden wir sehen, daß es stets mindestens eine MARKOV-Basis gibt und daß wir diese auf dem Umweg über einen Polynomring finden können. Zunächst aber wollen wir sehen, wozu MARKOV-Basen nützlich sind. Die Grundidee besteht darin, daß wir eine MARKOV-Basis für eine Irrfahrt durch eine Faser einer gegebenen Tafel

benutzen und an jedem Punkt dieser Irrfahrt den dortigen Wert von χ^2 vergleichen mit dem unserer Tafel. Nach einer hinreichend langen Fahrt bekommen wir so einen Schätzwert für die Wahrscheinlichkeit, daß das χ^2 einer zufällig aus der Faser gewählten Tafel größer oder gleich dem unserer Tafel ist.

§4: Markov-Ketten

Ein häufiger zitiertes Vorbild einer Irrfahrt ist der Heimweg eines Betrunkenen über einen Platz mit vielen Laternen. Jedesmal, wenn er sich den Kopf an einer Laterne anschlägt, geht er (mit gleicher Wahrscheinlichkeit) nach rechts oder nach links. Seine Entscheidungen an den verschiedenen Laternen sind unabhängig voneinander, hängen also insbesondere nicht davon ab, *wie* er an die aktuelle Laterne gekommen ist. MARKOV-Ketten sind stochastische Prozesse mit entsprechenden Eigenschaften:

Definition: *a)* Ein stochastischer Prozeß ist eine Folge $X^{(1)}, X^{(2)}, \ldots$ von Zufallsvariablen mit Werten in einer festen Menge A. Wir betrachten hier nur den Fall einer endlichen Menge $A = \{a_1, \ldots, a_m\}$.
b) Ein solcher Prozeß heißt MARKOV-Prozeß oder MARKOV-Kette, wenn für alle $n \in \mathbb{N}$ gilt

$$\begin{aligned} &p(X^{(n+1)} = x^{(n+1)} \mid X^{(1)} = x^{(1)}, \ldots, X^{(n)} = x^{(n)}) \\ &\quad = p(X^{(n+1)} = x^{(n+1)} \mid X^{(n)} = x^{(n)}) \,. \end{aligned}$$

c) Eine MARKOV-Kette heißt *zeitinvariant,* wenn die bedingte Wahrscheinlichkeit $p(X^{(n+1)} = y \mid X^{(n)} = x)$ nicht von n abhängt. In diesem Fall setzen wir setzen wir $p_{ij} = p(X^{(n+1)} = a_j \mid X_n = a_i)$ und bezeichnen die $m \times m$-Matrix P mit Einträgen p_{ij} als die *Übergangsmatrix* des Prozesses.

Wir betrachten im folgenden nur zeitinvariante MARKOV-Ketten. Zeitinvarianz muß selbstverständlich nicht bedeuten, daß alle Zufallsvariablen $X^{(n)}$ dieselbe Verteilung haben, sie sind allerdings auch nicht unabhängig voneinander: Bezeichnet $p^{(n)} \in \Delta_{m-1}$ die Verteilung

von $X^{(n)}$, so ist

$$p_j^{(n+1)} = p(X^{(n+1)} = a_j) = \sum_{i=1}^{m} p(X^{(n+1)} = a_j \mid X^{(n)} = a_i) = \sum_{i=1}^{m} p_i^{(n)} p_{ij} \,;$$

wenn wir die $p^{(n)}$ als Zeilenvektoren schreiben, ist also $p^{(n+1)} = p^{(n)} P$.

Definition: *a)* Eine MARKOV-Kette heißt *stationär,* wenn alle $X^{(n)}$ dieselbe Wahrscheinlichkeitsverteilung haben.
b) Eine MARKOV-Kette heißt *reversibel,* wenn für alle i, j, n gilt: $p_i^{(n)} p_{ij} = p_j^{(n)} p_{ji}$.
c) Eine MARKOV-Kette heißt *irreduzibel,* wenn es für je zwei mögliche Werte a_i, a_j stets ein $r \in \mathbb{N}$ gibt, so daß $p(X^{(r+1)} = a_j \mid X^{(1)} = a_i) > 0$ ist.
d) Der Wert a_i heißt *aperiodisch,* wenn der ggT aller natürlicher Zahlen r mit $p(X^{(r+1)} = a_i \mid X^{(1)} = a_i) > 0$ gleich eins ist.

Satz: Ist eine MARKOV-Kette reversibel, irreduzibel und aperiodisch, so ist sie stationär.

Zum *Beweis* sei auf Lehrbücher zur Theorie der MARKOV-Ketten verwiesen, zum Beispiel

ACHIM KLENKE: Wahrscheinlichkeitstheorie, *Springer,*^{2}2008, Satz 18.18

Wir werden im folgenden, soweit nicht explizit etwas anderes gesagt ist, stets annehmen, daß unsere MARKOV-Ketten zeitinvariant sind. In diesem Fall können wir die Wahrscheinlichkeitsverteilungen aller Zufallsvariablen aus der von X_1 und der Übergangsmatrix berechnen: Ist allgemein $p_i^{(n)}$ die Wahrscheinlichkeit, mit der X_n dem Wert a_i annimmt, so ist

$$\begin{aligned} &p(X_n = a_{i_0}, X_{n+1} = a_{i_1}, \ldots, X_{n+r} = a_{i_r}) \\ &= p(X_n = a_{i_0}) \prod_{\ell=1}^{r} p(X_{n+\ell} = a_{i_\ell} \mid X_{n+\ell-1} = a_{\ell-1}) \,. \end{aligned}$$

Für $r = 1$ wird das zu

$$p(X_n = a_i, X_{b+1} = a_j) = p(X_n = a_i) p(X_{n+1} = a_j \mid X_n = a_i) = p_i^{(n)} p_{ij} \,,$$

was wir auch einfacher mit Matrizen und Vektoren formulieren können: Ist $\mathfrak{p}^{(n)} = (p_1^{(n)}, \ldots, p_m^{(n)})^T$ der Spaltenvektor der Wahrscheinlichkeitsverteilung zu X_n, so ist $\mathfrak{p}^{(n+1)} = A^T \mathfrak{p}^{(n)}$ und damit $\mathfrak{p}^{(n)} = \left(A^T\right)^{n-1} \mathfrak{p}^{(1)}$.

Somit bestimmen $\mathfrak{p}^{(1)}$ und die Übergangsmatrix A die Wahrscheinlichkeitsverteilungen aller X_n und erlauben damit auch die Berechnung der Wahrscheinlichkeiten aller Teiltupel, die von der MARKOV-Kette produziert werden.

§5: Der Algorithmus von Metropolis und Hastings

Der Algorithmus von METROPOLIS und HASTINGS ist eine Montecarlo-Methode auf der Basis von MARKOV-Ketten.

Wir gehen aus einem log-linearen Modell $\mathcal{M}_A$ und einer MARKOV-Basis $\mathcal{B}$ von $\mathrm{Kern}_{\mathbb{Z}}(A)$. Zu einer beobachteten Tabelle u betrachten wir eine Zufallsvariable U mit Werten in $\mathcal{F}(u)$ und interessieren uns für die Wahrscheinlichkeit

$$p\big(\chi^2(U) \geq \chi^2(u)\big) .$$

Wir gehen auf eine Irrfahrt durch die Faser von u, wobei wir in jedem Schritt ein zufällig gewähltes Element von $\mathcal{B}$ addieren. Um eine unverzerrte Schätzung der gesuchten Wahrscheinlichkeit zu erhalten, müssen wir allerdings gelegentlich auch auf der Stelle treten. Konkret sieht der Algorithmus folgendermaßen aus:

Initialisierung: Wähle ein beliebiges Element $u^{(1)}$ aus $\mathcal{F}(u)$.

t-ter Iterationsschritt: Wähle zufällig ein Element $b_t \in \mathcal{B}$, wobei jedes Element mit gleicher Wahrscheinlichkeit auftreten kann. Dann ist

$$u^{(t+1)} = \begin{cases} u^{(t)} + b_t & \text{mit Wahrscheinlichkeit } q \\ u^{(t)} & \text{mit Wahrscheinlichkeit } 1 - q \end{cases}$$

mit $q = \min\left(1, \dfrac{p\big(U = u^{(t)} + b_t \;\big|\; U \in \mathcal{F}(u)\big)}{p\big(U = u^{(t)} \;\big|\; U \in \mathcal{F}(u)\big)}\right)$.

Zu dieser neuen Tabelle wird χ^2 berechnet, und anhand einer hinreichend langen Folge dieser Werte wird die obige Wahrscheinlichkeit geschätzt.

Um die Schätzun möglichst unabhängig von $u^{(1)}$ zu machen, werden dabei meist die ersten Glieder der Folge ignoriert.

Um zu sehen, daß wir so tatsächlich eine unverzerrte Schätzung der Wahrscheinlichkeit bekommen, müssen wir zeigen, daß die erzeugte Folge von χ^2-Werten als erwartete Häufigkeit den gesuchten Anteil aller Elemente der Faser mit einem Wert von χ^2, der den für die gegebene Tabelle von u nicht übersteigt. Siehe dazu das oben zitierte Buch von KLENKE, insbesondere die Paragraphen 18.2 und 18.3.

§6: Tafeln und Ideale

Wir gehen weiterhin aus von einem log-linearen Modell $\mathcal{M}_A$ und wollen diesem algebraische Objekte zuordnen. Kontingenztafeln haben nichtnegative ganze Zahlen als Einträge, und die Elemente einer MARKOV-Basis sind Vektoren mit ganzzahligen Einträgen. Der Zusammenhang zu Polynomen wird nun dadurch hergestellt, daß solche Vektoren aufgefaßt werden als Exponentenvektoren von Monomen. Falls der Vektor auch negative Komponenten enthält, liegen diese Monome in keinem Polynomring, sondern im sogenannten Ring der LAURENT-Polynome. Die Elemente der MARKOV-Basis liegen in $\mathbb{Z}^d$; wir betrachten daher den Ring der LAURENT-Polynome in d Variablen $T_1, \ldots, T_d$. Er ist definiert als

$$k[T, T^{-1}] \underset{\text{def}}{=} k[T_1, \ldots, T_d, T_1^{-1}, \ldots, T_d^{-1}],$$

das heißt seine Elemente sind endliche Linearkombinationen von Monomen der Form $T^e = T_1^{e_1} \cdots T_d^{e_d}$ mit $e = (e_1, \ldots, e_d) \in \mathbb{Z}^d$.

Die Darstellung in diesem Paragraphen folgt weitgehenjd dem 4. Kapitel von

BERND STURMFELS: Gröbner Bases and Convex Polytopes, *University Lecture Series vol. 8, American Mathematical Society,* 1996

Zur Matrix A betrachten wir die Menge $\mathcal{A} = \{a^{(1)}, \ldots, a^{(n)}\}$ ihrer Zeilenvektoren ($n = \#\mathcal{R}$), und für jedes $a^{(\ell)}$ haben wir ein Monom $T^{a^{(\ell)}} \in k[T, T^{-1}]$. Außerdem betrachten wir den Polynomring

$k[X] \underset{\text{def}}{=} k[X_1, \ldots X_n]$ und die Abbildungen

$$\pi: \begin{cases} \mathbb{Z}^n \to \mathbb{Z}^d \\ u \mapsto \sum u_\ell a^{(\ell)} \end{cases} \quad \text{und} \quad \hat{\pi}: \begin{cases} k[X] \to k[T, T^{-1}] \\ X_\ell \mapsto T^{a^{(\ell)}} \end{cases} .$$

Den Kern von $\hat{\pi}$ bezeichnen wir als das *Gitterideal* $I_{\mathcal{A}}$.

Lemma: $I_{\mathcal{A}}$ ist ein Primideal und wird als k-Vektorraum erzeugt von den Binomen $X^u - X^v$ mit $u, v \in \mathbb{N}_0^n$ so, daß $\pi(u) = \pi(v)$ ist.

Beweis: Zunächst ist $I_{\mathcal{A}}$ ein Primideal, denn liegt ein Produkt fg in $I_{\mathcal{A}}$, so verschwindet $\hat{\pi}(fg) = \hat{\pi}(f)\hat{\pi}(g)$ in $k[T, T^{-1}]$. Da dieser Ring nullteilerfrei ist, muß mindestens einer der beiden Faktoren $\hat{\pi}(f)$ und $\hat{\pi}(g)$ verschwinden, d.h. mindestens eines der Polynome f und g liegt in $I_{\mathcal{A}}$.

Ein Monom X^u wird von $\hat{\pi}$ abgebildet auf $T^{\pi(u)}$, ein Binom $X^u - X^v$ liegt also genau dann im Kern von $\hat{\pi}$, wenn $\pi(u) = \pi(v)$ ist. Wir müssen zeigen, daß jedes $f \in I_{\mathcal{A}}$ als k-Linearkombination solcher Binome geschrieben werden kann.

Angenommen, es gibt Polyome $f \in I_{\mathcal{A}}$, für die das nicht der Fall ist. Dann betrachten wir ein solches Gegenbeispiel f, das bezüglich irgendeiner Monomordnung auf $k[X]$ das kleinste führende Monom hat. Da f im Kern von $\hat{\pi}$ liegt, ist $f(T^{a^{(1)}}, \ldots, T^{a^{(\ell)}}) = 0$. Schreibt man dies aus, erhält man eine Linearkombination, in der auch $\hat{\pi}(X^u) = T^{\pi(u)}$ auftritt, wobei u das führende Monom von f bezeichnet. Da die gesamte Linearkombination verschwindet, muß in f noch mindestens ein Monom X^v vorkommen mit $v \neq u$ und $\pi(v) = \pi(u)$. Durch Subtraktion eines geeigneten skalaren Vielfachen von $X^u - X^v$ erhält man ein Polynom $g \in I_{\mathcal{A}}$, dessen führendes Monom kleiner ist als das von f. Wegen der Minimalität von f läßt sich g als Linearkombination von Binomen $X^p - X^q$ mit $\pi(p) = \pi(q)$ schreiben, und da f die Summe von g und einem skalaren Vielfachen von $X^u - X^v$ ist, gilt dasselbe für f. Das ist ein Widerspruch, und damit ist das Lemma bewiesen. ∎

Für $u \in \mathbb{Z}^n$ definieren wir $u^+, u^- \in \mathbb{N}_0^n$ durch

$$u_i^+ = \max(u_i, 0) \quad \text{und} \quad u_0^- = \max(-u_i, 0) .$$

Dann ist $u = u^+ - u^-$, und dies ist die einzige Darstellung von u als Differenz zweier Elemente $v, w \in \mathbb{N}_0^n$, bei der für kein $i \in \mathcal{R}$ sowohl u_i als auch v_i von Null verschieden sind. Offensichtlich liegt u genau dann im Kern von π, wenn $\pi(u^+) = \pi(u^-)$ ist. Das gerade bewiesene Lemma läßt sich daher auch so formulieren, daß $I_{\mathcal{A}}$ als k-Vektorraum erzeugt wird von den Binomen $X^{u^+} - X^{u^-}$ mit $u \in \operatorname{Kern} \pi$.

Korollar: Für jede Monomordnung $<$ auf $k[X]$ gibt es eine endliche Teilmenge $\mathcal{G}_< \subset \operatorname{Kern} \pi$, so daß

$$\left\{ X^{u^+} - X^{u^-} \;\middle|\; u \in \mathcal{G}_< \right\}$$

die reduzierte GRÖBNER-Basis von $I_{\mathcal{A}}$ bezüglich dieser Monomordnung ist.

Beweis: Zunächst hat $I_{\mathcal{A}}$ ein endliches Erzeugendensystem aus Binomen $X^{u^+} - X^{u^-}$ mit $u \in \operatorname{Kern} \hat{\pi}$, denn die Gesamtheit dieser Monome erzeugt das Ideal, und wenn keine Teilmenge dazu ausreichen sollte, hätten wie eine unendliche Folge von Binomen mit der Eigenschaft, daß das von den ersten r Folgegliedern erzeugte Teilideal I_r stets echt kleiner wäre als I_{r+1}. Eine solche unendliche aufsteigende Folge von Idealen kann es aber nicht geben, denn nach dem HILBERTschen Basissatz ist die Vereinigung aller I_r endlich erzeugt, und wenn wir uns ein endliches Erzeugendensystem hernehmen, liegt jedes von dessen Elementen in einem I_r und damit auch in allen folgenden. Für den größten unter diesen Indizes r liegen also alles Erzeugenden in I_r, so daß I_{r+1} nicht größer sein kann.

Es gibt daher ein Erzeugendensystem aus endlich vielen Binomen, und darauf können wir den BUCHBERGER-Algorithmus anwenden. $f = X^{u^+} - X^{u^-}$ und $g = X^{v^+} - X^{v^-}$ seien zwei Binome die in irgendeinem Stadium des Algorithmus im gerade betrachteten Erzeugendensystem liegen. Dabei können wir o.B.d.A. annehmen, daß bezüglich der betrachteten Monomordnung $X^{u^+} > X^{u^-}$ und $X^{v^+} > X^{v^-}$ ist: Andernfall können wir einfach u durch $-u$ *bzw.* v durch $-v$ ersetzen, was nichts an der Erzeugung ändert. Mit dem Monom $M = \operatorname{kgV}(X^{u^+}, X^{v^+})$

ist dann

$$\begin{aligned} S(f,g) &= \frac{M}{X^{u^+}}\left(X^{u^+} - X^{u^-}\right) - \frac{M}{X^{v^+}}\left(X^{v^+} - X^{v^-}\right) \\ &= \frac{M}{X^{v^+}}X^{v^-} - \frac{M}{X^{u^+}}X^{u^-} \end{aligned}$$

wieder ein Binom aus Kern $\hat{\pi}$, und auch wenn wir darauf den Divisionsalgorithmus anwenden mit Binomen als Divisoren ändert sich daran nichts, denn bei jedem Schritt wird einfach ein Monom durch ein anderes ersetzt. Auch beim Übergang zur reduzierten GRÖBNER-Basis wird nur gestrichen oder ein Binom durch ein anderes ersetzt, womit die Behauptung folgt.

■

Der obige Beweis beruhte auf dem HILBERTschen Basissatz, ist also nicht konstruktiv. Die gleiche Idee, mit der wir am Ende von §1 des zweiten Kapitels zu einer in Parameterdarstellung gegebenen Menge die darauf verschwindenden Polynome bestimmt hatten, führt aber zu einem Algorithmus, mit der wir eine explizite GRÖBNER-Basis von $I_{\mathcal{A}}$ konstruieren können: Der Homomorphismus $\hat{\pi}$ bildet X_ℓ ab auf $T^{a^{(\ell)}} = T^{a^{(\ell)}+}/T^{a^{(\ell)}-}$. Im Polynomring $k[T_0, \ldots, T_d, X_1, \ldots, X_n]$ verschwinden daher modulo $I_{\mathcal{A}}$ die Polynome $X_\ell T^{a^{(\ell)}-} - T^{a^{(\ell)}+}$.

Wir betrachten das von diesen Polynomen und $T_0 T_1 \cdots T_d - 1$ erzeugte Ideal J. Wie in Kapitel 2 folgt, daß $I_{\mathcal{A}}$ der Durchschnitt von J mit $k[X_1, \ldots, X_n]$ ist. Berechnen wir also bezüglich einer Monomordnung, in der alle T_j größer sind als jedes X_ℓ eine reduzierte GRÖBNER-Basis von J, so ist deren Durchschnitt mit $k[X_1, \ldots, X_n]$ eine reduzierte GRÖHBNER-Basis von $I_{\mathcal{A}}$.

Diese GRÖBNER-Basis hängt natürlich ab von der gewählten Monomordnung auf $k[X_1, \ldots, X_n]$, und für $n \geq 2$ gibt es unendlich viele solche Monomordnungen. Trotzdem gilt

Satz: Jedes Ideal $I \lhd k[X_1, \ldots, X_n]$ hat nur endlich viele reduzierte GRÖBNER-Basen.

Beweis: Ist G eine reduzierte GRÖBNER-Basis von I bezüglich irgendeiner Monomordnungen, so bilden die führenden Monome der Basisele-

mente ein minimales Erzeugendensystem des von den führenden Monomen der Elemente aus I erzeugten Ideals FM(I). Ein solches minimales Erzeugendensystem ist eindeutig bestimmt, denn haben wir zwei solche Systeme, so muß jedes Monom aus dem einen teilbar sein durch eines aus dem anderen und umgekehrt, aber kein Monom eines minimalen Erzeugendensystems kann ein anderes daraus teilen. Also bestehen beide Systeme aus denselben Monomen. Haben wir daher zwei Monomordnungen, bezüglich derer die Ideale der führenden Monome übereinstimmen, stimmen auch die führenden Monome der Elemente der GRÖBNER-Basen überein. Wir betrachten ein Element g der ersten Basis und ein Element g' der zweiten, die beide dasselbe führende Monom haben. Dann liegt auch $g - g'$ in I, und im Falle $g \neq g'$ hat dieses Polynom bezüglich der ersten Ordnung ein führendes Monom, das kleiner ist als das von g. Da wir eine GRÖBNER-Basis haben, muß es durch das führende Monom eines anderen Elements der Basis teilbar sein. Damit kann es sich nicht um ein Monom aus g handeln, denn in einer reduzierten GRÖBNER-Basis ist kein Monom durch das führende Monom eines anderen Basiselements teilbar. Also handelt es sich um ein Monom aus g', ist also durch das führende Monom eines Elements der zweiten Basis teilbar. Beide Bassen haben aber dieselben führenden Monome, so daß dies nicht möglich ist. Somit ist $g = g'$, d.h. die beiden reduzierten GRÖBNER-Basen sind gleich.

Falls es daher unendlich viele reduzierte GRÖBNER-Basen gibt, ist auch die Menge Σ_0 der Ideale der führenden Terme zu I unendlich. Wir wählen ein Element $f_1 \in I$. Da es nur endlich viele Monome enthält, ist für mindestens eines dieser Monome die Menge Σ_1 aller Ideale aus Σ_0, die dieses Monom m_1 enthalten, unendlich. Insbesondere gibt es daher Monomordnungen, für die das Hauptideal (m_1) eine echte Teilmenge von FM(I) ist. Die Monome, die nicht in FM(I) liegen, sind die Standardmonome, und deren Restklassen bilden eine Vektorraumbasis von $k[X]/I$. Die Monome, die nicht in (m_1) liegen, sind daher linear abhängig modulo I. Daher gibt es ein Polynom $f_2 \in I$, von dessen Monomen keines in (m_1) liegt. Wie oben folgt, daß f_2 ein Monom m_2 enthalten muß, das in unendlich vielen Idealen aus Σ_1 liegen muß. Die Menge dieser Ideale sei Σ_2.

Wieder muß es ein FM(I) $\in \Sigma_2$ geben, das echt größer ist als (m_1, m_2), und die Monome, die nicht in (m_1, m_2) liegen, sind linear abhängig modolo I, so daß es ein Polynom $f_3 \in I$ gibt, von dessen Monomen keines in (m_1, m_2) liegt, und so weiter.

Auf diese Weise erhalten wir eine unendliche echt aufsteigende Folge

$$(m_1) \subset (m_1, m_2) \subset (m_1, m_2, m_3) \subset \cdots ,$$

und wie wir schon mehrfach in dieser Vorlesung gesehen haben, ist das nach dem HILBERTschen Basissatz unmöglich. ∎

Definition: Wir bezeichnen die Vereinigung aller reduzierter GRÖBNER-Basen von $I_{\mathcal{A}}$ als *die* universelle GRÖBNER-Basis $\mathcal{U}_{\mathcal{A}}$ von $I_{\mathcal{A}}$.

(Üblicherweise bezeichnet man in der Computeralgebra jede Menge von Polynomen, die bezüglich jeder Monomordnung eine GRÖBNER-Basis ist, als eine universelle GRÖBNER-Basis. Die Definition hier ist sehr viel spezieller.)

Da jede reduzierte GRÖBNER-Basis nur Binome enthält, besteht $\mathcal{U}_{\mathcal{A}}$ somit aus endlich vielen Binomen.

Definition: Ein Binom $X^{u^+} - X^{u^-}$ aus $I_{\mathcal{A}}$ heißt *primitiv,* wenn es kein Binom $X^{v^+} - X^{v^-}$ aus $I_{\mathcal{A}}$ mit $v \neq u$ gibt, für das X^{v^+} ein Teiler von X^{u^+} ist und X^{v^-} ein Teiler von X^{u^-}.

Lemma: Jedes Binom aus $\mathcal{U}_{\mathcal{A}}$ ist primitiv.

Beweis: $X^{u^+} - X^{u^-}$ sei ein Element einer reduzierten GRÖBNER-Basis, und wir nehmen wieder an, daß X^{u^+} das führende Monom sei. Dann ist X^{u^+} Teil eines minimalen Erzeugendensystems von FM($I_{\mathcal{A}}$), so daß das kleinere Monom X^{u^-} nicht in FM($I_{\mathcal{A}}$) liegt. Angenommen, es gäbe ein von u verschiedenes $v \in \operatorname{Kern} \pi$ mit $X^{v^+} | X^{u^+}$ und $X^{v^-} | X^{u^-}$. Falls X^{v^+} das führende Monom von $X^{v^+} - X^{v^-}$ wäre, wäre X^{u^+} kein minimales Erzeugendes, und wäre X^{v^+} das führende Monom, so läge X^{u^-} in FM($I_{\mathcal{A}}$). Beides ist nicht möglich, also ist $X^{u^+} - X^{u^-}$ primitiv. ∎

Unser nächstes Ziel ist eine Schranke für die Grade der primitiven Polynome aus $I_{\mathcal{A}}$). Über mehrere Lemmata wollen wir den folgenden Satz beweisen:

Satz: Die Matrix $A \in \mathbb{Z}^{d\times n}$ habe Rang d, und $D(\mathcal{A})$ sei der maximale Betrag einer $d \times d$-Unterdeterminante von A. Dann ist der Grad eines jeden primitiven Polynoms aus $I_{\mathcal{A}}$ kleiner als $(d+1)(n-d)D(\mathcal{A})$.

Wir beschränken uns zunächst spezielle Elemente:

Definition: *a)* Der *Träger* eines Vektors $u \in \operatorname{Kern}\pi$ ist die Menge aller Indizes i, für die u_i nicht verschwindet.
b) Ein primitives Element heißt ein *Kreis,* wenn sein Träger minimal ist und $\operatorname{ggT}(u_1, \ldots, u_n) = 1$ ist.

Lemma: Der Träger eines Kreises besteht aus höchstens d+1 Elementen.

Beweis: $u \in \operatorname{Kern}\pi$ sei ein Vektor mit $r \geq d+2$ von Null verschiedenen Komponenten, und B sei die $d \times r$-Untermatrix von A aus den Spalten, in denen u eine nichtverschwindende Koordinate hat. Dann ist der Kern der durch Multiplikation mit B definierten linearen Abbildung mindestens zweidimensional. Durch eine geeignete Linearkombination zweier linear unabhängiger Vektoren erhalten wir einen Vektor $v' \neq 0$ mit mindestens einer verschwindenden Koordinate. Diesen erweitern wir zu einem Vektor $v \in \mathbb{Z}^n$, indem wir für alle Indizes, für die u_i verschwindet, Null einsetzen. Dann liegt v in $\operatorname{Kern}\pi$, und sein Träger ist kleiner als der von u. Somit kann u kein Kreis sein. ∎

Lemma: Ist $u = (u_1, \ldots, u_n) \in \operatorname{Kern}\pi$ ein Kreis, so ist $|u_i| \leq D(\mathcal{A})$ für alle i.

Beweis: Der Träger von u sei $\{i_1, \ldots, i_r\}$, und B sei die $d \times r$-Matrix mit Zeilen $a_{i_1}, \ldots, a_{i_r}$. Da u im Kern von π liegt, hat sie höchstens Rang $r-1$. Tatsächlich muß sie Rang $r-1$ haben, denn wäre der Rang kleiner, so könnten wir wie im Beweis des vorigen Lemmas ein Element von $\operatorname{Kern}\pi$ konstruieren, dessen Träger echt in dem von u enthalten wäre. Daher können wir Indizes $i_{r+1}, \ldots, i_{d+1}$ finden derart,

daß die $d \times (d+1)$-Matrix B mit den Zeilen $a_{i_1}, \ldots, a_{i_{d+1}}$ Rang d hat. Wir wollen das homogene lineare Gleichungssystem mit Matrix B nach der CRAMERschen Regel lösen. Dazu müssen wir es ergänzen zu einem Gleichungssystem mit einer nichtsingulären quadratischen Matrix. Dies erreichen wir, indem wir eine Gleichung hin zufügen, die postuliert, daß zum Beispiel für einen der Indizes aus dem Träger von u die Komponente einen bestimmten Wert haben soll. Dann können wir die Regel anwenden und, da es bei der Lösung des homogenen Gleichungssystems auf skalare Vielfache nicht ankommt, die Nenner ignorieren. Dies führt auf eine ganzzahlige Lösung, deren k-te Komponente gleich

$$(-1)^k \det(a_{i_1}, \ldots, a_{i_{k-1}}, a_{i_{k+1}}, \ldots, a_{i_{d+1}})$$

ist. Der Vektor mit den entsprechenden Komponenten von u ist natürlich ebenfalls eine ganzzahlige Lösung, und da u ein Kreis ist, haben deren Komponenten den ggT eins. Die nach CRAMER berechnete Lösung muß deshalb ein ganzzahliges Vielfaches dieser Lösung sein, und das zeigt die Behauptung.

■

Definition: Ein Vektor $u \in \mathbb{Z}^n$ ist *konform* zu $v \in \mathbb{Z}^n$, wenn der Träger von u^+ in dem von v^+ enthalten ist und der von u^- in dem von v^-.

Lemma: Jeder Vektor v aus dem Kern von π kann dargestellt werden als rationale Linearkombination mit nichtnegativen Koeffizienten von $n - d$ Kreisen, die allesamt konform zu v sind.

Beweis durch Induktion nach n: Für $n \leq d+1$ ist die Behauptung klar; sei also $n \geq d+2$, und v ein Vektor aus dem Kern, der nicht schon selbst ein Kreis ist. Wir können annehmen, daß der Träger von v gleich $\{1, 2, \ldots, n\}$ ist, denn andernfalls können wir Spalten, deren Nummer nicht im Träger vorkommt, aus der Matrix A streichen und erhalten eine Matrix mit geringerer Spaltenzahl. Nach Induktionsannahme läßt sich daher v als rationale Linearkombination von weniger als $n - d$ Kreisen schreiben.

Nun sei u irgendein Kreis mit $u_1 v_1 > 0$, und λ sei der kleinste Wert unter den positiven der Brüche v_i/u_i. Dann ist der Vektor $u - \lambda v$ konform zu v, und seine i-te Koordinate verschwindet, d.h. sein Träger ist

kleiner als der von v. Daher können wir wieder die Induktionsannahme anwenden und $u - \lambda v$ als rationale Linearkombination mit nichtnegativen Koeffizienten von $n - 1 - d$ Kreisen schreiben. Addieren wir λu dazu, haben wir die gewünschte Darstellung von v.

■

Mit diesen Lemmata können wir nun den obigen Satz beweisen:

v sei ein primitiver Vektor aus dem Kern von π. Falls v ein Kreis ist, folgt die Behauptung aus dem vorletzten Lemma. Andernfalls ist nach dem gerade bewiesenen Lemma

$$v = \lambda_1 u_1 + \cdots + \lambda_{n-d} u_{n-d}$$

mit nichtnegativen rationalen Zahlen λ_i und Kreisen u_i, die allesamt konform zu v sind. Damit ist auch

$$v^+ = \lambda_1 u_1^+ + \cdots + \lambda_{n-d} u_{n-d}^+ \quad \text{und} \quad v^- = \lambda_1 u_1^- + \cdots + \lambda_{n-d} u_{n-d}^- .$$

Wegen der Primitivität von v folgt daraus, daß alle λ_1 kleiner als eins sein müssen.

Falls im Binom $X^{v^+} - X^{v^-}$ der Gesamtgrad gleich dem des ersten Terms ist, erhalten wir nach der Dreiecksungleichung und den drei Lemmata die Abschätzung

$$\left\|v^+\right\|_1 \leq \sum_{i=1}^{n-d} \lambda_i \left\|u_i^+\right\|_1 < \max_{i=1}^{n-d} \left\|u_i^+\right\| \leq (n-d)(d+1)D(\mathcal{A}) ,$$

und falls der Gesamtgrad von $X^{v^+} - X^{v^-}$ gleich dem von X^{v^-} ist, können wir analog argumentieren.

Damit ist der Satz bewiesen.

■

Wir bezeichnen die Menge aller Kreise in $I_\mathcal{A}$ mit $\mathcal{C}_\mathcal{A}$ und die Menge aller primitiver Binome als die GRAVER-Basis $Gr_\mathcal{A}$. Wie der Beweis eines der obigen Lemmata zeigt, kann $\mathcal{C}_\mathcal{A}$ zumindest im Prinzip mit Hilfe der CRAMERschen Regel berechnet werden. Die Computeralgebra kennt allerdings effizientere Verfahren sowohl zur Berechnung von $\mathcal{C}_\mathcal{A}$ als auch zu der der universellen GRÖBNER-Basis $\mathcal{U}_\mathcal{A}$.

Lemma: Für jede endliche Teilmenge $\mathcal{A} \subset \mathbb{Z}^d$ ist $\mathcal{C}_\mathcal{A} \subseteq \mathcal{U}_\mathcal{A} \subseteq Gr_\mathcal{A}$.

Beweis: Wir wissen bereits, daß jedes Polynom aus $\mathcal{U}_\mathcal{A}$ ein primitives Binom ist, woraus die zweite Inklusion folgt. Für die erste müssen wir zeigen, daß jeder Kreis Element einer reduzierten GRÖBNER-Basis ist. Dazu betrachten wir einen Vektor u aus dem Kern von π und eine Monomordnung $<$, bezüglich derer die Variablen X_i mit i aus dem Träger von u kleiner sind als die, für die i nicht im Träger liegt. Dann ist $X^{u^+} - X^{u^-}$ Element der reduzierten GRÖBNER-Basis bezüglich dieser Eliminationsordnung: Da u in $I_\mathcal{A}$ liegt, gibt es ein Element $X^{v^+} - X^{v-}$ der reduzierten GRÖBNER-Basis, dessen führender Term X^{v^+} den führenden Term X^{u^+} teilt. Insbesondere liegt dann der Träger von v^+ im Träger von u. Enthielte der Träger von v^- ein i, das nicht im Träger von u liegt, wäre X_i größer als alle X_j mit j aus dem Träger von u, also insbesondere auch als die mit einem j aus dem Träger von v^+. Damit könnte X^{v+} nicht das führende Monom sein.

Somit liegt der gesamte Träger von v im Träger von u, und da u ein Kreis ist, muß v ein ganzzahliges Vielfaches von u sein.Wegen $X^{v^+}|X^u$ geht das nur, wenn $v = u$ ist, d.h.$X^{u^+} - X^{u^-}$ ist ein Element der reduzierten GRÖBNER-Basis. ∎

Man kann Beispiele von Teilmengen $\mathcal{A} \subset \mathbb{Z}^d$ konstruieren, für die beide Inklusionen strikt sind, aber auch solche, in denen nur die erste oder nur die zweite oder beide Gleichheitszeichen sind.

Was noch fehlt bei den obigen Inklusionen sind die Basen, die uns hier am meisten interessieren, die MARKOV-Basen. Sie haben allerdings einen entscheidenden Unterschied zu den anderen bisher betrachteten Basen: Da man mit einer MARKOV-Basis nicht nur von einem Vektor u zu einem Vektor v kommen kann, sondern auch wieder zurück, werden sie in vielen Fällen außer einem Vektor b auch den Vektor $-b$ enthalten, d.h. bei der algebraischen Betrachtung wird außer $X^{b^+} - X^{b^-}$ auch $X^{b^-} - X^{b^+}$ in der Basis sein. Bei reduzierten GRÖBNER-Basen ist so etwas natürlich nie der Fall; daher empfiehlt es sich, die Definition an den algebraischen Fall anzupassen:

Neudefinition: Ab jetzt ist eine MARKOV-Basis des log-linearen Modells $\mathcal{M}_A$ eine endliche Menge $\mathcal{B}$ von Tafeln $b \in \mathbb{Z}^{\#\mathcal{R}}$ mit $Ab = 0$, für die gilt: Für jede Tafel $u \in \mathbb{N}_0^{\#\mathcal{R}}$ und je zwei Elemente $v, v' \in \mathcal{F}(u)$ gibt es eine Folge von Elementen $b_1, \ldots, b_L \in \mathcal{B}$, so daß gilt:
1.) Für $\ell = 1, \ldots, L$ ist $v + \epsilon_1 b_1 + \cdots + \epsilon_L b_\ell \in \mathbb{N}_0^{\#\mathcal{R}}$
2.) $v + \epsilon_1 b_1 + \cdots + \epsilon_L b_L = v'$

(Zur Erinnerung: Wir haben $n = \#\mathcal{R}$ gesetzt und identifizieren Tafeln mit Vektoren aus $\mathbb{N}_0^n$.)

Arbeitet man mit dieser Neudefinition, muß natürlich auch der Algorithmus von METROPOLIS-HASTINGS entsprechend moodifiziert werden: In jedem Schritt wählt man zusätzlich zum zufälligen Element b_t der MARKOV-Basis noch ein $\varepsilon_t \in \{-1, 1\}$, wobei jeweils jedes Element die gleiche Wahrscheinlichkeit hat. Statt mit b_t arbeitet man dann mit dem Produkt $\varepsilon_t b_t$.

Bezüglich der neuen Definition gilt dann der folgende

Satz: Eine endliche Teilmenge $\mathcal{B} \subset \operatorname{Kern} \pi$ ist genau dann eine MARKOV-Basis für das log-lineare Modell $\mathcal{M}_\mathcal{A}$, wenn die Binome $X^{b^+} - X^{b^-}$ das Ideal $I_\mathcal{A}$ erzeugen.

Beweis: Zunächst sei $\mathcal{B}$ eine MARKOV-Basis. Wie wir bereits zu Beginn dieses Paragraphen gesehen haben, wird $I_\mathcal{A}$ als Vektorraum erzeugt von den Binomen $X^v - X^{v'}$ mit $\pi(v) = \pi(v')$. Es genügt daher zu zeigen, daß jedes solche Polynom im Erzeugnis der Binome $X^{b^+} - X^{b-}$ liegt.

Nach Definition einer MARKOV-Basis gibt es zu zwei Vektoren vs und v' mit $\pi(v) = \pi(v')$ eine Folge von Vektoren $b_1, \ldots, b_L$ aus $\mathcal{B}$ mit der Eigenschaft, daß

$$v + \sum_{i=1}^{\ell} b_i \in \mathbb{N}_0^n \quad \forall \ell \leq L \quad \text{und} \quad v + \sum_{i=1}^{L} b_i = v'$$

ist. Somit ist

$$X^{v'} = X^v + \sum_{i=1}^{L} X^{v+\varepsilon_1 b_1 + \cdots + \varepsilon_{i-1} b_{i-1} - \varepsilon_i b_i^-} (X^{b_i^+} - X^{b_i^-}),$$

wobei die Summe im Exponenten für jedes i nur nichtnegative Komponenten hat. Somit ist $X^v - X^{v'}$ als monomiale Linearkombination der $X^{b_i^+} - X^{b_i^-}$ dargestellt, liegt also im von diesen erzeugten Ideal.

Umgekehrt sei $I_{\mathcal{A}}$ erzeugt von den Monomen $X^{b^+} - X^{b^-}$ mit $b \in \mathcal{B}$, und v, v' seien zwei Vektoren mit $\pi(v) = \pi(v')$. Dann ist $X^v - X^{v'}$ eine polynomiale und damit auch eine monomiale Linearkombination von Binomen $X^{b^+} - X^{b^-}$ mit $b \in \mathcal{B}$. Da $X^v - X^{v'}$ nur ± 1 als Koeffizienten hat und wir im gesuchten Pfad von v nach v' nach der Neudefinition außer einem Vektor b auch sein Negatives $-b$ verwenden dürfen, können wir nötigenfalls b durch $-b$ ersetzen und somit annehmen, daß alle Koeffizienten gleich eins sind. Es gibt daher Vektoren $m_i \in \mathbb{N}_0^n$ und $b_i \in \mathcal{B} \cup -\mathcal{B}$, so daß $X^v - X^{v'} = \sum_{i=1}^{L} X^{m_i}(X^{b_i^+} - X^{b_i^-})$ ist für ein geeignetes $L \in \mathbb{N}$. Durch Indexpermutation können wir erreichen, daß $X^v = X^{b_1^+}$ ist und anschließend $X^{m_i} X^{p_i^-} = X^{m_{i+1}} X^{b_{i+1}^+}$. Zum Schluß muß dann $X^{m_L} X^{b_i^-} = X^{v'}$ sein und damit

$$v = v' + \sum_{i=1}^{L} b_i \quad \text{und} \quad v' + \sum_{i=1}^{\ell} b_i \in \mathbb{N}_0^n \quad \forall \ell \leq L\,.$$

Somit ist $\mathcal{B}$ eine MARKOV-Basis. ∎

Als erste Folgerung daraus können wir zeigen

Satz: Zu jedem log-linearen Modell $\mathcal{M}_{\mathcal{A}}$ gibt es eine MARKOV-Basis.

Beweis: Nach dem Korollar zum ersten Lemma dieses Paragraphen gibt es zu jeder Monomordnung $<$ auf $k[X_1, \ldots, X_n]$ eine endliche Teilmenge $\mathcal{G}_< \subset \operatorname{Kern} \pi$, so daß die reduzierte GRÖBNER-Basis des Ideals $I_{\mathcal{A}}$ bezüglich $<$ aus den Binomen $X^{u^+} - X^{u^-}$ mit $u \in \mathcal{G}_<$ besteht. Nach dem gerade bewiesenen Satz ist $\mathcal{G}_<$ somit eine MARKOV-Basis für $\mathcal{M}_{\mathcal{A}}$. Sobald wir also bezüglich irgendeiner Monomordnung eine GRÖBNER-Basis von $I_{\mathcal{A}}$ bestimmt haben, kennen wir eine MARKOV-Basis für das Modell $\mathcal{M}_{\mathcal{A}}$. Insbesondere folgt daraus, daß es zu jedem log-linearen Modell eine MARKOV-Basis gibt. ∎

Verschiedene Monomordnungen können zu verschiedenen reduzierten GRÖBNER-Basen führen. Daher muß diese MARKOV-Basis nicht eindeutig sein. Da außerdem oft bereits eine Teilmenge einer GRÖBNER-Basis ausreicht, um das Ideal zu erzeugen, kann es auch vorkommen, daß bereits eine echte Teilmenge von $\mathcal{G}_<$ eine MARKOV-Basis ist.

Definition: Eine MARKOV-Basis $\mathcal{B}$ für das Modell $\mathcal{M}_\mathcal{A}$ heißt *minimal,* wenn keine echte Teilmenge von $\mathcal{B}$ eine MARKOV-Basis für $\mathcal{M}_\mathcal{A}$ ist.

Man kann sich fragen, ob für jede minimale MARKOV-Basis die zugehörigen Binome Teilmenge einer reduzierten GRÖBNER-Basis sind. In der mir bekannten Literatur ist dazu leider nichts zu finden.

Bekannt ist aber, daß alle minimalen MARKOV-Basen die gleiche Anzahl von Elementen haben, und aus dem Beweis dieser Tatsache kann man auch einen (im Allgemeinen recht aufwendigen) Algorithmus zur Bestimmung einer (oder aller) minimaler MARKOV-Basen herleiten. Siehe dazu Theorem 1.3.2 und den nachfolgenden Algorithmus im Buch

MATHIAS DRTON, BERND STURMFELS, SETH SULLIVANT: Lectures on Algebraic Statistics, *Oberwolfach Seminars* **39**, *Birkhäuser,* 2009