

## Lernen durch Matrixvervollständigung

F. Kiraly  
(University College London)

f.kiraly@ucl.ac.uk



---

### Ein Sudoku-Exkurs

---

Vermutlich werden Sie, geneigter Leser, und die meisten Menschen in Ihrem Bekanntenkreis, von Sudoku gehört haben - den ursprünglich aus den USA stammenden Zahlenrätseln, bei denen man in eine mit Zahlen unvollständig ausgefüllte Tabelle weitere Zahlen einfüllen muss, nach einem fest vorgegebenen Muster. Zum Beispiel dieses hier:

○	2	3			1	7		
		8	4	6			1	
9				5			4	8
5		4	3				2	○
	9		8	7		1		
1			○		4	9		5
	7				6	8		2
8		1	7		2			
	6			3	○			1

**Tabelle 1.** Das erste veröffentlichte Sudoku (bzw. eines der zwei ersten), im Jahre 1979 erschienen unter dem Namen "Number Place" in einer amerikanischen Rätselzeitschrift [3, 8]. Ziel ist es, die ganzen Zahlen zwischen (einschließlich) 1 und 9 so einzufüllen, dass in jedem 3x3-Block, jeder Zeile, und jeder Spalte jede dieser Zahlen genau einmal auftritt. Zusätzliche Hilfestellung: in der korrekten Lösung enthalten die Kreise die Zahlen 4,5,7,8.

Sudokus (oder Sudokui? Sudokuta? Sudöker?), wie zum Beispiel das in Tabelle 1, entsprechen somit dankbarerweise dem gemeinhin verbreiteten und akzeptierten Klischee der Mathematik: überall Zahlen, kompliziert, schwer verständlich, und vollständig unnützlich. Dies

harmoniert passgenau mit dem Bild des Klischeemathematikers, der den ganzen Tag über geistesabwesend im Kopf oder an der Kreidetafel Sudokus löst, während er sich ab und zu den von Essensresten angegilbten, knielangen Bart krault, dabei manchmal Dinge raunt wie "ja, das ist es" oder "elementar, Watson" (in Abwesenheit eines Watson). Eines Tages springt er aus seinem seltenen Bade, läuft "Heureka" schreiend, wild gestikulierend, nur vom Barte bedeckt, durch die Gassen, und wird von einer Pferdekutsche überfahren. Am Rande nur sei angemerkt, dass alle Klischeemathematiker selbstverständlich männlich sind, auch die weiblichen (siehe: Bart). Klischees und Stereotype zu hinterfragen ist natürlich immer unangebracht.

Wie dem auch sei, in der real existierenden Welt verhalten sich die meisten Mathematiker doch relativ unauffällig, genauso die glücklicherweise in immer größerer Anzahl real existierenden Mathematikerinnen. Die überwältigende Mehrheit beider ist im Allgemeinen mehr als nur einen Sicherheitsabstand entfernt von der stationären Einweisung.

Und anstelle von abgehobener Gehirnakrobatik geht es in der Mathematik vor allem um folgerichtiges Schlussfolgern, das heißt, möglichst irrtumsfrei Aussagen zu treffen. Man kann das nun, wie Kunst, zum Selbstzweck betreiben, und daran ist auch nichts grundlegend falsch. Zweifelsohne von weitaus größerer gesellschaftlicher Relevanz sind aber die mathematischen und statistischen Modelle, die auf reale Vorgänge anwendbar sind, und somit folgerichtige Schlussfolgerungen über die Realität erlauben. Beispielsweise Maschinen, Autos oder Computer zu bauen; das Wetter vorherzusagen; oder die beste Therapie für eine Krankheit zu finden.

Am Ende dieses Artikels werden wir ein Modell sehen, Sudokus nicht ganz unähnlich, welches Sportergebnisse vorhersagen kann, mit relativ elementarem Wissen über Algebra und Statistik. Was aber viel wichtiger ist: wir werden auch sehen, wie man folgerichtig überprüfen kann, ob ein solches Modell und dessen Vorhersagen grober Unfug sind oder nicht. Schließ-







lich ist etwas ja nicht allein dadurch richtig, dass es irgendwo steht, oder in komplizierten Formeln aufgeschrieben wurde - sondern es muss irrtumsfrei begründet werden. Das ist Mathematik.

## Über das Vervollständigen der Matrix

Das am Ende vorgestellte Modell (welches wie gesagt unter anderem Sportergebnisse vorhersagen kann) beruht auf einem Prinzip, das dem des Sudoku nicht unähnlich ist: in eine Tabelle nach einem bestimmten Prinzip Zahlen einzutragen. Das Prinzip ist natürlich ein bisschen anders als beim Sudoku, im Folgenden wird die nötige Theorie dazu eingeführt.

### Was ist eine Matrix?

Die bessere Frage zu Beginn ist vielleicht: wieso kann man überhaupt etwas vorhersagen, indem man Zahlen in eine Tabelle einfüllt. Die einfache Antwort lautet: dadurch, dass die eingefüllte Zahl selbst die Vorhersage ist. Und dadurch, dass die Tabelle eine Tabelle von Daten bzw. Messwerten ist. Ein Beispiel:

NETFLIX		Users		
				
Movies		☆☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆
		☆☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆
		☆☆☆☆☆	☆☆☆☆☆	?

**Tabelle 2.** Schematisch vereinfachte Darstellung des "Netflix Prize" Problems, für dessen Lösung die Firma Netflix im Jahr 2006 eine Million Dollar ausschrieb [11]. Zeilen entsprechen Filmen, Spalten entsprechen Benutzern, Einträge entsprechen Bewertungen. Ziel ist es, eine gute Vorhersage für fehlende Bewertungen einzufüllen.

Tabelle 2 enthält Benutzerbewertungen von Filmen, wie sie zum Beispiel bei Netflix vorgenommen werden. Die verschiedenen Zeilen entsprechen verschiedenen Benutzern, die Spalten entsprechen verschiedenen Filmen. Ein Eintrag ist die Bewertung, die der Benutzer in derselben Zeile dem Film in derselben Spalte gegeben hat. Nicht jeder Benutzer bewertet jeden Film, und nicht jeder Film wird von jedem Benutzer bewertet, daher ist die Tabelle nicht komplett ausgefüllt.

Möchte man nun vorhersagen, wie dem Benutzer X der Film Y gefallen würde, entspricht das der Aufgabe, einen plausiblen Wert für den Eintrag in der X-ten Spalte und der Y-ten Zeile zu finden. Macht man das für alle Filme, die X noch nicht gesehen hat, kann man daraus eine Filmempfehlung ableiten, indem man Filme

mit hoher vorhergesagter Bewertung empfiehlt. Ähnlich erhält man eine Produktempfehlung, wenn man "Film" durch "Produkt des Onlinegroßhändlers O" ersetzt. Und so weiter.

Und nun zum Sport:

Athlet	200m	400m	800m	1500m
Bolt	19.19	45.28		
Farah			1:48.69	3:28.81
Johnson	19.32	43.18		
Kipketer		46.85	1:41.11	
Rudisha		45.45	1:40.91	
Yego			1:42.67	3:33.68

**Tabelle 3.** Persönliche Bestzeiten einiger Weltspitzenläufer. Ziel ist es, diese in einem Modell zu verstehen. Interessante Fragen in dieser Tabelle wären z. B. die nach Rudishas Potenzial auf 1500m, oder Johnsons 800m-Zeit.

Tabelle 3 enthält Bestleistungen von Weltspitzenläufern. Zeilen entsprechen den Athleten, Spalten verschieden langen Laufstrecken, und die Einträge den Bestzeiten. Vorhersagen sind hier interessant in Vorbereitung auf eines der längeren Rennen, für die oft eine mehrmonatige Trainingszeit erforderlich ist, wie zum Beispiel dem Marathon; oder, um eine Leistung außerhalb der "typischen" Lauflängen eines Athleten zu schätzen. Zusätzlich und insbesondere ist hier auch wichtig, diese Vorhersagen im Rahmen eines wissenschaftlich spärlichen Modelles zu sehen, welches im Optimalfall Rückschlüsse auf die zugrundeliegende Physiologie erlaubt.

Dem aufmerksamen Leser wird aufgefallen sein, dass die obigen Ausführungen zwar erklären, wieso man durch gutes Einfüllen der Einträge eine gute Vorhersage treffen kann, aber nichts darüber gesagt ist, wie man denn die Einträge nun praktisch einfüllt. Schließlich ist eine Vorhersage nicht allein deswegen gut, weil man irgendetwas eingefüllt hat. Der nächste Abschnitt wird das mathematische Prinzip hinter einer guten Einfüllstrategie beschreiben.

Dazu sei noch kurz die anfängliche Frage beantwortet. Eine Matrix (im mathematischen Sinne) ist die formelle Abstraktion einer Tabelle von Zahlen (ohne Zeilen- und Spaltenbeschriftung). Präziser:

**Definition.** Eine (reelle) Matrix von Größe  $(m \times n)$ , ist ein tabellarisch geordnetes Tupel  $A$  von reellen Zahlen  $A_{ij} \in \mathbb{R}$ , mit Indizes  $i \in \{1, 2, \dots, m\}$  und  $j \in \{1, 2, \dots, n\}$ . Man schreibt kurz  $A \in \mathbb{R}^{m \times n}$  und interpretiert  $A_{ij}$  als den Eintrag in der  $i$ -ten Zeile und  $j$ -ten Spalte. Die  $i$ -te Zeile ist eine  $(1 \times n)$ -Matrix und wird mit  $A_{i*}$  notiert; die  $j$ -te Spalte ist eine  $(m \times 1)$ -Matrix und wird mit  $A_{*j}$  notiert.

Dem Leser sind Matrizen vermutlich aus den Abiturklassen oder aus dem Studium vertraut, inklusive der üblichen Konventionen im Bezug auf Addition und Multiplikation, die als bekannt vorausgesetzt werden. Auch einem Leser mit abgeschlossenen Mathematikstudium sei die Lektüre der folgenden Abschnitte nahegelegt, da die Interpretation einer Matrix als Modell,

und nicht als abstraktes Konzept, möglicherweise unge-  
 wohnt vorkommen kann.

### Low-Rank Matrix Completion

Ein Einfüllprinzip, mit dem man gute Vorhersagen für  
 Bewertungen und die sportlichen Leistungen bekommt,  
 beruht auf der Annahme, dass die vollständige Matrix  
 einen niedrigen Rang besitzt. Eine mögliche (wenn auch  
 weniger gebräuchliche) Definition für den Rang einer  
 Matrix lautet wie folgt:

**Definition.** Sei  $A \in \mathbb{R}^{m \times n}$  eine Matrix. Der Rang  
 von  $A$  ist die kleinste Anzahl  $r$  an Zeilenprototypen  
 $u_1, \dots, u_r$ , sodass jede Zeile von  $A$  als Linearkombi-  
 nation der  $u_j$  dargestellt werden kann, d.h. für alle Zei-  
 lennummern  $i$  existieren  $\lambda_{ij} \in \mathbb{R}$ , sodass

$$A_{i*} = \lambda_{i1}u_1 + \lambda_{i2}u_2 + \dots + \lambda_{ir}u_r.$$

Die Zeilenprototypen haben eine unmittelbare Inter-  
 pretation in den Beispielszenarien. Wenn man zum Bei-  
 spiel annimmt, dass die Matrix mit den Filmbewertungen  
 Rang  $r$  hat, entspricht das der Annahme, dass es  $r$  pro-  
 totypische Filme oder "Grundgenres" wie Action, Dra-  
 ma, Komödie, etc gibt, aus denen jeder Film (zumindest  
 im Hinblick auf die Bewertungen) zusammengesetzt ist.  
 Die  $\lambda_{ij}$  beschreiben den Anteil des  $j$ -ten Genres an  $i$ -  
 ten Film. Die Bewertungen der einzelnen Filme werden  
 erklärt als lineare Mischung der Bewertung, die die Pro-  
 totypen erhalten würden. Ähnlich verhält es sich mit den  
 Sportlern - hier hat man prototypische Athleten, so wie  
 den Sprinter oder den Langstreckenläufer.

Die "Prototypen" müssen hierbei nicht als reale Fil-  
 me oder Athleten existieren, damit das Modell die Ma-  
 trizen gut beschreibt. Es sei hier erwähnt, dass obwohl  
 das Modell vielleicht plausibel erscheint, es im Prinzip  
 weder zutreffend noch hilfreich sein muss. Ob die Mo-  
 dellannahme sinnvoll war, muss bei einer praktischen  
 Anwendung immer erst überprüft werden.

Ein grundlegendes Resultat der Matrixalgebra be-  
 sagt, dass man in der Definition oben für den Rang  
 statt Zeilenprototypen auch Spaltenprototypen nehmen  
 könnte, die Anzahl ist die gleiche. Am einfachsten ist  
 das vielleicht zu sehen, indem man die Definition in  
 Matrix-Multiplikations-Notation umschreibt:

**Definition<sup>1</sup>.** Sei  $M \in \mathbb{R}^{m \times n}$  eine Matrix. Der Rang  
 von  $M$  ist die kleinste Zahl  $r$ , sodass Matrizen  $U \in$   
 $\mathbb{R}^{m \times r}$  und  $V \in \mathbb{R}^{r \times n}$  existieren mit  $M = UV$ .

Die Zeilen von  $U$  sind  $u_i$ , und die Einträge von  $V$   
 sind die  $\lambda_{ij}$  in der ursprünglichen Definition. Die Spal-  
 ten von  $V$  sind Spaltenprototypen, und durch Transpo-  
 nieren von  $A$  sieht man, dass die kleinstmögliche An-  
 zahl Zeilenprototypen gleich der kleinsten Anzahl Spal-  
 tenprototypen ist.

In den Beispielen bedeutet das, dass im Low-Rank-  
 Modell zu den Filmgenres auch Benutzertypen gehören:  
 Action-Fan, Drama-Fan, Komödien-Fan, usw. Und zu  
 den Musterathleten auch prototypische Aufgaben: z. B.  
 Sprinten, und Langstreckenlauf.

<sup>1</sup>Diese Definition des Ranges wird auch der "Zerlegungsrang" genannt. Wer andere Definitionen kennt wie den Zeilen- oder Spaltenrang, die im Abiturunterricht und in Universitätsvorlesungen wesentlich häufiger zu finden sind, sollte die Äquivalenz relativ schnell beweisen können.

Low-Rank Matrix Completion, also das Einfüllen  
 von Einträgen unter der Annahme niedrigen Ranges  
 (wir wollen den gebräuchlicheren, englischen Ausdruck  
 benutzen), bekommt so die Form eines Rang-udokus.  
 Zum Beispiel diese unvollständige  $(10 \times 10)$ -Matrix in  
 Tabelle 4.

Die (von praktischem Nutzen komplett befreite,  
 aber illustrative) Aufgabe ist es, sämtliche Einträge ex-  
 akt oder mit einer numerischen Präzision von  $1e-3$  der-  
 art einzufüllen, dass die vollständige Matrix einen Rang  
 von 2 besitzt. Es ist durchaus erlaubt, dass eine Zahl  
 pro Zeile oder Spalte mehrmals auftritt, und es dürfen  
 auch irrationale oder komplexe Zahlen eingefüllt wer-  
 den, wenn es denn hilft.

-4			2	5			3		
4	4	-2							-4
4		-4			-3			6	
			-1	5	0			3	
	-2	2					-4	6	
			0		2	6			-2
					0	4	1		-4
6			-3	-9			-5		
	6					-6		9	-6
	6	-3		-3		-6			

**Tabelle 4.** Rang-Rätsel. Ziel ist es, komplexe Zahlen der-  
 art einzufüllen, dass die vollständige Matrix Rang 2 hat. Die  
 Lösung ist mit drei Nachkommastellen Präzision gesucht.

1	2	3	4	5	6	7	8	9	10
2	1	4	3	6	5	8	7	10	9
1	1								
2	1								
1	2								
1	4								
3	1								
4	2								
3	5								
0	1								

**Tabelle 5.** Rang-Rätsel. Ziel ist es, komplexe Zahlen der-  
 art einzufüllen, dass die vollständige Matrix Rang 2 hat. Die  
 exakte Lösung gesucht.

## Wie man einfüllt

Falls Sie sich am Rätsel versucht haben - falls nicht, würde ich an dieser Stelle darum bitten, dass Sie sich kurz Gedanken machen, wie Sie anfangen würden - wird Ihnen vermutlich aufgefallen sein, dass die Definition des Ranges (Zerlegungsrang, auch der Ihnen möglicherweise bekannte Zeilen-/Spaltenrang) nicht sehr hilfreich ist, um die Einträge einzufüllen. Naives Probieren hilft im Allgemeinen auch nicht, das im Gegensatz zum Sudoku im Prinzip unendlich viele Zahlen in Frage kommen.

Wie so oft besteht auch hier ein gewisser Unterschied zwischen der theoretischen Charakterisierung der Lösung und deren praktischer Konstruktion - den meisten angewandten Wissenschaftlern, sowie den Lesern des Rundbriefs dürfte dieser Unterschied wohl vertraut sein; dieser besteht in der Angabe einer konstruktiven Rechenvorschrift, eines sogenannten Algorithmus. Diesem auf die Spur kommt man vielleicht am besten in einer einfacheren Version des Rätsels, in Tabelle 5.

Die Aufgabe ist wieder die gleiche: alle fehlenden Einträge einzufüllen, sodass die Matrix Rang 2 besitzt, d.h., jede Zeile ist die Summe zweier "Prototypenzeilen". Im Beispiel kann man sich relativ schnell überlegen, dass man ohne Beschränkung der Allgemeinheit die ersten zwei (vollständigen) Zeilen als die Prototypen wählen kann. Da zwei Einträge in jeder anderen Zeile vorhanden sind, kann man die Lösung auf eine Übungsaufgabe in Gauss-Elimination zurückführen.

Die Lösung lässt sich explizit angeben (Kenntnis der Matrixinversen und Determinante vorausgesetzt): ist  $D$  die fehlende  $(8 \times 8)$ -Matrix, so lassen sich die fehlenden Einträge bestimmen als  $D = BA^{-1}C$ , wo  $A$  die  $(2 \times 2)$ -Matrix der ersten zwei Spalten und Zeilen,  $B$  die  $(8 \times 2)$ -Matrix der ersten zwei Zeilen und letzten acht Spalten, und  $C$  die  $(2 \times 8)$ -Matrix der letzten acht Zeilen und ersten zwei Spalten. Das heißt, für jeden fehlenden Eintrag  $D_{ij}$  gilt die Gleichung  $D_{ij} = B_i A^{-1} C_j$ , wo  $B_i$  die  $i$ -te Zeile von  $B$ , und  $C_j$  die  $j$ -te Spalte von  $C$  ist. Letzteres ist äquivalent (nach Umstellung und Multiplikation mit  $\det(A)$ ) zur Formel  $\det M^{(i,j)} = 0$ , wo  $M^{(i,j)}$  die Untermatrix mit den drei Zeilen 1, 2,  $i$  und drei Spalten 1, 2,  $j$  ist.

Mit anderen Worten, sobald man drei Zeilen und drei Spalten gefunden hat, in deren Schnittmenge insgesamt 8 ( $= 3 \times 3 - 1$ ) Einträge bekannt sind, kann man diese benutzen, um den neunten durch Auflösung nach der  $(3 \times 3)$ -Determinante einzufüllen<sup>2</sup>. Tatsächlich beruht hierauf bereits die Hauptidee des praktischen Algorithmus für Low-Rank Matrix Completion, der sich in [6] findet - mittels weniger beobachteter Einträge die unbekanntesten einzeln einzufüllen.

Leider ist die Determinanten-Strategie in der obigen Form weder theoretisch noch praktisch zureichend. Zum einen erlaubt sie nicht, alle Einträge einzufüllen, die theoretisch einfüllbar wären. In der Praxis ist sie so nicht verwendbar, da eine Datenmatrix normalerweise mit Messungenauigkeiten behaftet ist und einer Matrix

von niedrigerem Rang nur nahe kommt.

Die Praxis wird im übernächsten Abschnitt behandelt (schließlich wollen wir ja Sportergebnisse vorhersagen). Der nächste Abschnitt ("Circuits") gibt einen kurzer Ausblick auf interessante theoretische Phänomene und Lösungshinweise zum Rätsel in Tabelle 4, kann jedoch vom Leser auch übersprungen werden; für eine genauere Abhandlung der Theorie sei auf die entsprechende Publikation [6] verwiesen.

## Circuits

Ein natürlicher Ausgangspunkt der theoretischen Betrachtungen ist eine weitere Alternativdefinition des Ranges, der sogenannte Determinanten-Rang:

**Definition.** Sei  $M \in \mathbb{R}^{m \times n}$  eine Matrix. Der Rang von  $M$  ist die kleinste Zahl  $r$ , sodass die Determinanten aller  $(r+1 \times r+1)$ -Untermatrizen von  $M$  verschwinden.

Die Äquivalenz zur den vorigen Definitionen des Ranges ist nicht trivial, aber klassisch; man kann diese Äquivalenz verstehen als Begründung einer allgemeineren (für Algebraiker relativ offensichtlichen) Determinanten-Strategie. Leider zeigt zum Beispiel das erste Rang-udoku, dass man im Allgemeinen mit der Determinanten-Strategie alleine nicht weit kommt (das Rätsel ist mit genau dieser böartigen Absicht konstruiert). Es stellt sich heraus, dass man zu einer Verallgemeinerung der Determinante übergehen muss, sogenannten "Circuits":

**Definition.** Sei  $C \subseteq [m] \times [n]$  eine Menge von Indizes.  $C$  heißt ein Circuit (von Rang  $r$ ), falls:

- (i) Für beliebige  $e \in C$  lassen sich an den Indizes  $C \setminus \{e\}$  beobachtete Einträge zu einer (nicht notwendigerweise eindeutigen) Matrix von Rang  $r$  vervollständigen.
- (ii) Im Allgemeinen lassen sich an den Indizes  $C$  beobachtete Einträge nicht zu einer Matrix von Rang  $r$  vervollständigen.

Zum Beispiel ist jede  $(r+1 \times r+1)$ -Untermatrix ein Circuit; denn fehlt ein Eintrag, lässt sich dieser mit der Determinanten-Strategie zu einer Rang- $r$ -Matrix vervollständigen (ebenso potenziell fehlende Einträge in anderen Zeilen und Spalten), somit ist Bedingung (i) erfüllt. Kennt man jedoch bereits alle Einträge, hat man im Allgemeinen keine Vervollständigung zu einer Matrix von Rang  $r$  - nämlich genau dann wenn die Matrix bereits Rang  $r+1$  hat, da eine solche Matrix auch nicht Untermatrix einer Rang- $r$ -Matrix sein kann.

Allerdings ist nicht jeder Circuit von der Form einer Untermatrix. Beispielsweise kann man zeigen: das Indexmuster

<sup>2</sup>da mit  $\det(A)$  multipliziert wurde, ist es notwendig, dass die bekannte  $(2 \times 3)$ -Untermatrix vollen Rang besitzt; für eine "allgemeine" oder zufällige Matrix ist das der Fall, im Allgemeinen aber nicht.


ist ein (für das Rätsel in Tabelle 4 übrigens hilfreicher) Circuit von Rang 2. Interessanterweise hat dieser Circuit für allgemeine Matrizen genau zwei Vollständigungen zu einer  $(4 \times 4)$ -Matrix.

Ähnlich wie im Fall von Untermatrizen, kann man von einem Circuit eine Rekonstruktionsvorschrift ableiten:

**Theorem.** Sei  $C \subseteq [m] \times [n]$  ein Circuit (von Rang  $r$ ). Dann existiert ein (bis auf Skalarmultiplikation) eindeutig bestimmtes Polynom  $P_C$  in den Einträgen einer Matrix  $M_c$ ,  $c \in C$ , sodass folgendes gilt: Die Einträge  $M_c$ ,  $c \in C$  können zu einer Rang- $r$ -Matrix vervollständigt werden genau dann, wenn gilt, dass  $P_C$  ausgewertet auf den  $M_c$ ,  $c \in C$  verschwindet, i.e.,  $P_C(M_c, c \in C) = 0$ .

Nach der Definition von Circuits muss jede Variable  $M_c$ ,  $c \in C$  im Circuit nichttrivial auftreten, somit kann man einen einzelnen fehlenden Eintrag  $e$  in  $C$  durch Auflösen von  $P_C$  nach der Variablen  $M_e$  bis auf endlich viele Alternativen genau bestimmen. Das Polynom für den  $(r+1 \times r+1)$ -Untermatrix-Circuit ist die Determinante (als Polynom in den Einträgen betrachtet).

Man sieht leicht, dass Zeilen- und Spaltennummerierung an der Tatsache, dass man einen Circuit vor sich hat, nichts wesentlich ändert; man kann zeigen, dass es selbst mit dieser Äquivalenz unendliche viele Circuits von festem Rang  $r$  gibt. Alle Circuits von Rang 1 sind bekannt, für höheren Rang ist das offen.

Es wird vermutet, dass eine vollständige Charakterisierung schwierig ist, da Verbindungen zu anderen, bereits längere Zeit offenen Problemen in der Kombinatorik gibt; aber bitte lassen Sie sich davon auf keinen Fall entmutigen. Beweise und weitere Details finden sich in [6].

## Wie man nachprüft, dass die obigen Ideen praktisch funktionieren und die Vorhersagen gut sind

Im vorhergehenden Abschnitt haben wir gesehen, wie sich mittels Determinanten (oder allgemeiner: Circuits) Einträge in eine unvollständige Matrix einfüllen lassen. Selbst schöne theoretische Resultate sind aber keine Garantie, dass die Strategie in der Praxis auch funktioniert, z. B. für die Filmbewertungen oder die Sportergebnisse. Zum einen müssen etwaige Annahmen nicht in den Daten erfüllt sein; und selbst wenn, in der modernen Wissenschaft ist nur systematische Beobachtung von (rea-

len, tatsächlichen) guten Vorhersagen ein ausreichender Nachweis für die Fähigkeit der Methode, gute Vorhersagen zu treffen.

### Empirische Validierung

Wie ein solcher Nachweis nach dem Stand der statistischen Wissenschaft erfolgen kann, wird im Folgenden erklärt. Daran lässt sich dann ablesen, ob die naive Strategie “finde eine  $(r+1 \times r+1)$ -Untermatrix mit einem fehlenden Eintrag, fülle diesen mit der Determinante ein” gut funktioniert oder nicht.

Für eine Vorhersagemethode ist, nach dem statistischen Induktionsschluss, eine gute Vorhersagefähigkeit empirisch bestätigt, wenn unter ähnlichen Bedingungen systematisch gute Vorhersagen beobachtet wurden. Um eine solche Vorhersage zu simulieren, geht man beispielsweise folgendermaßen vor: ein Eintrag aus der unvollständigen Datenmatrix wird entfernt, und mittels der Methode aus den anderen Einträgen vorhergesagt. Das wird mehrmals wiederholt, und man erhält mehrere vorhergesagte Einträge. Die Vorhersagen werden dann mit den jeweils echten Einträgen verglichen, eine gute Vorhersagemethode besitzt im Durchschnitt einen niedrigen Fehler. Niedrig ist hier immer im Vergleich zu naiven Methoden zu sehen, wie den Eintrag uninformiert zu raten, oder den Mittelwert der Zeile/Spalte einzufüllen.

Wir beschreiben etwas formaler eine Möglichkeit für ein solches sogenanntes Kreuzvalidierungsschema<sup>3</sup>: Seien  $x_1, \dots, x_N \in [m] \times [n]$  Indizes, an denen Einträge beobachtet wurden. Seien  $y_1, \dots, y_N \in \mathbb{R}$  die dazugehörigen Einträge. Fixiere eine (abstrakte) Vorhersagemethode  $V$ .

**Schritt 1:** Ziehe zufällig  $k$  Indizes  $x_{i_1}, \dots, x_{i_k}$  (ohne Zurücklegen).

**Schritt 2:** Benutze die Methode  $V$ , um Vorhersagen  $f_{i_1}, \dots, f_{i_k}$  für die Indizes zu machen. Zur Vorhersage von  $f_{i_j}$  werden alle Index/Eintrags-Paare benutzt außer  $(x_{i_j}, y_{i_j})$ , damit es sich auch tatsächlich um ein Vorhersage handelt.

**Schritt 3:** Berechne eine Vorhersagefehlerstatistik, zum Beispiel denn mittleren Absolutfehler (mean absolute error, MAE), der definiert ist als

$$\text{MAE von } V = \frac{1}{k} \sum_{j=1}^k |f_{i_j} - y_{i_j}|.$$

Standardfehler oder Konfidenzintervalle erhält man aus der Stichprobe der Absolutfehler  $|f_{i_j} - y_{i_j}|$ . Zwei Vorhersagemethoden  $V_1, V_2$  (z. B. Determinanten-Strategie, Mittelwert einfüllen) lassen sich über gepaarte Stichprobentests auf den Absolutfehlern vergleichen.

Falls der MAE der Determinanten-Strategie signifikant unter dem MAE einer naiven Strategie (z. B. Mittelwert einfüllen) liegt, kann dann der Schluss getroffen werden, dass die Determinanten-Strategie eine Vorhersage erlaubt, auf dem betrachteten Datensatz.

<sup>3</sup>Das vorgestellte Validierungsschema ist eine Variante der “leave-one-out cross-validation”. Der Subsampling-Schritt Nummer 1 ist eine Möglichkeit, um bei großen  $N$  Zeit zu sparen; wann immer möglich, wählt man bei der leave-one-out cross-validation  $k = N$ . Eine Übersicht über andere ebenfalls sinnvolle Validierungsschemata und quantitative Schätzgarantien findet sich in [4], Kapitel 7.

Normalerweise werden Methoden auf mindestens zwei Arten von Datensätzen verglichen: einem synthetischen Datensatz, der maschinell mit korrekten Modelleigenschaften (z. B. Rang 2) erzeugt wurde, und einem realen Datensatz.

Wenn die Methode auf dem synthetischen Datensatz Vorhersagen treffen kann, beweist das, dass sie unter den eingebauten Modellannahmen funktioniert. Wenn die Methode auf dem realen Datensatz vorhersagen kann, beweist dies - lediglich genau das. Es zeigt nämlich strenggenommen nicht, dass der reale Datensatz ebenfalls den Modellannahmen folgt, obwohl es das plausibel macht, wenn andere Gründe unplausiblen scheinen. Umgekehrt, wenn die Methode auf dem synthetischen Datensatz vorhersagen kann, aber nicht auf dem realen, liefert das einen starker Hinweis dafür, dass die Modellannahmen für den realen Datensatz nicht zutreffen.

Tabelle 6 zeigt MAE der Determinanten-Strategie "finde eine  $(3 \times 3)$ -Untermatrix mit einem fehlenden Eintrag, fülle diesen mit der Determinante ein" und der Strategien "Median der Spalte", und "Mittelwert der Spalte", geschätzt wie oben mit  $k = 100$ , auf zwei Datensätzen. Bei Datensatz 1 handelt es sich um die komplette  $(10 \times 10)$ -Matrix aus dem zweiten Rätsel, mit Rang 2. Bei Datensatz 2 handelt es sich um die Datenmatrix der offiziellen Sportergebnisse von 101.775 männlichen britischen Läufern seit 1954, wie sie in [1] verwendet wurde.

Die Aufgabe in Datensatz 1 ist die Vorhersage eines gleichverteilt zufälligen Eintrages, die Aufgabe in Datensatz 2 ist die Vorhersage einer gleichverteilt zufällig gewählten Marathonzeit.

Methode	Tabelle 5	Athleten
Spalten-Median	$1.1 \pm 0.2$	$32 \pm 3\text{min}$
Spalten-MW	$1.5 \pm 0.2$	$34 \pm 3\text{min}$
Determinanten	$0 \pm 0$	$27 \pm 4\text{min}$

**Tabelle 6.** Empirische Vorhersagegüten verschiedener Vorhersagestrategien (Zeilen) auf zwei Datensätzen (Spalten). Tabelle 5 ist eine  $(10 \times 10)$ -Matrix von Rang exakt 2; der Athleten-Datensatz ist in [1] beschrieben. Die mittleren Absolutfehler sind nach dem beschriebenen Kreuzvalidierungsverfahren, mit  $k = 100$  geschätzt. Der berichtete Standardfehler ist die empirische Standardabweichung des Mittelwertes der einzelnen Vorhersagefehler (Absolutresiduale).

Die Ergebnisse zeigen empirisch, dass die Determinanten-Strategie auf der (zweiten) Rätselmatrix perfekt funktioniert, wie erwartet. Leider ist sie auf den Athleten nicht wesentlich besser (im Rahmen des statistischen Unsicherheit) als die naiven Strategien, die überhaupt keine Information über die einzelnen Athleten benutzen - und selbst wenn sie ein kleines bisschen besser wäre, ist die Vorhersage einer Marathonzeit mit einem mittleren (nicht maximalen) Fehler in der Größenordnung von 30 Minuten Abweichung in der Praxis kaum zu gebrauchen. Wie oben ausgeführt, ist sind das alles starke Hinweise darauf, dass ein Rang-2-Modell für die Matrix der Athleten nicht zutrifft, oder

präziser, dass mindestens eine der impliziten Annahmen der Determinanten-Vorhersagestrategie nicht, oder nicht gut genug erfüllt ist.

### Es rauscht

Nun fragen Sie sich sicher, geneigter Leser, was hier denn falsch gelaufen ist. Immerhin wurde gerade der Nachweis erbracht, dass das oben beschriebene Einfüllen mit Determinanten für die Sportdaten ganz und gar nicht gut funktioniert. Beziehungsweise, genau genommen wurde lediglich beschrieben, wie man die Güte der vorgeschlagenen Methode überprüft. Den nachprüfbaren Nachweis in der Form eines der Tabelle 1 vergleichbaren Ergebnisses erhalten Sie, sobald Sie den entsprechenden Code auf den erwähnten Daten ausgeführt haben ([5]). Es sei angemerkt, dass das durch Glauben der Tabelle 6 nicht ersetzt werden kann, daher sind Sie eingeladen, das auch tatsächlich praktisch in der Realität mit einem Computer zu tun ([5]), und erst daraus korrekt zu schließen, dass die Methode schlecht ist.

Was schief gelaufen ist, ist allerdings nicht nur die Methode. Tatsächlich wurde hier die Erbsünde der modernen Wissenschaft begangen: es wurde von einem Modell ausgegangen, bevor die Daten überhaupt betrachtet wurden; es wurde eine Erklärung für die Realität angeboten, bevor diese überhaupt beobachtet und in einem Experiment geprüft wurde. Das low-rank-Modell wurde Ihnen präsentiert mit Verweis auf die Sport-Anwendung, ohne dass genau erklärt wurde, warum gerade das ein gutes Modell sein soll, oder besser als andere Modelle. Dieser Versuchung zu erliegen ist umso einfacher, je mathematisch schöner und intuitiv plausibler das Modell oder die Erklärung ist, aber Schönheit und Plausibilität einer Annahme machen sie leider nicht wahr.

Natürlich ist das auch ein bisschen der Erzählreihenfolge in diesem Artikel geschuldet, der existierende Resultate wiedergibt; wenn aber erstmals behauptet wird "Modell X ist gut für die Daten Y", dann sind zuerst die Daten Y gut zu untersuchen. Wenn man das im für den Fall der Sportergebnisse macht (siehe [5]), dann sieht man (qualitativ) folgendes ein:

Größere Blöcke in der Matrix haben fast Rang 1, wenn man sie logarithmiert, und

$(3 \times 3)$ -Untermatrizen (nach Logarithmieren) haben fast Rang 2, aber eben nicht ganz.

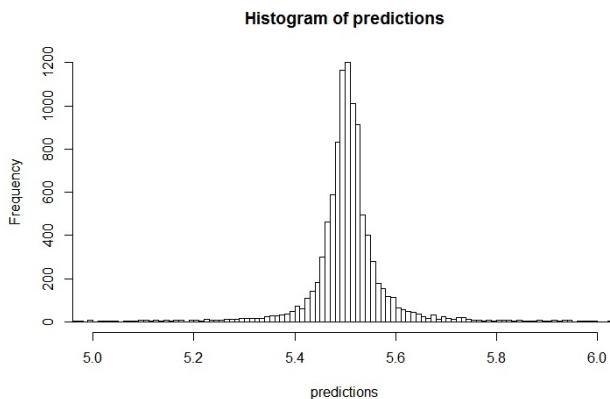
Beides lässt sich z. B. mit geeigneter Anwendung der Singulärwertzerlegung feststellen.

Aus den obigen Beobachtungen kann man die Hypothese ableiten, dass der *Logarithmus* der Datenmatrix eine *mit Messungenauigkeiten behaftete* Rang-2-Matrix ist. Der Logarithmus macht im Rahmen bekannter Potenzgesetze für sportliche Leistungen Sinn, und Messungenauigkeiten sind ein Standardphänomen auf realen Daten.

Nur den Logarithmus zu nehmen hilft etwas bei der Vorhersage, aber nicht allzu viel. Der genauen Struktur



der Messungengenauigkeiten kommt man auf die Spur, indem man einen Eintrag fixiert, und viele Vorhersagen durch die Determinanten-Strategie (an Hand verschiedener positionierter Untermatrizen) betrachtet.



**Schaubild. Typisches Histogramm von Vorhersagen durch zufällige Determinanten auf der logarithmierten Matrix der Athleten. Die x-Achse zeigt den vorhergesagten Wert, die y-Achse die Absoluthäufigkeit im jeweiligen Intervall. Zu beachten ist eine Streuung um den tatsächlichen Wert (hier 5.5), mit einigen Outliern. Typisch ist auch das vereinzelte Auftreten absurd großer Vorhersagen (hier nicht gezeigt).**

Das Schaubild zeigt ein typisches Histogramm der Vorhersagen. Man sieht, dass die Vorhersagen um den tatsächlich beobachteten Wert streuen, an den Enden mitunter extrem, aber um den tatsächlichen Wert konzentrieren. Das legt eine leicht verbesserte Strategie<sup>4</sup> nahe: zuerst Logarithmus ziehen; viele Vorhersagen mit der Determinanten-Strategie machen, dann den Median nehmen (robust gegenüber extremen Werten); schließlich exponentieren.

Und siehe da, das funktioniert auch: der MAE einer solchen Strategie auf dem Datensatz beträgt  $3.4(\pm 0.5)$  Prozent, das entspricht der gegenwärtig besten bekannten Vorhersagegüte, und einem mittleren Fehler von 3 bis 4 Minuten beim Marathon. Für eine technisch präzisere Beschreibung eines Algorithmus, empirische Ergebnisse, sowie eine ausführliche Behandlung des Themas siehe [1].

Wie oben wird zum Überprüfen dieser Behauptung nachdrücklich eingeladen ([5]). Das Wichtigste ist dazu die empirische Bestimmung des Vorhersagefehlers per Kreuzvalidierung; als zweiten Schritt würde man dieselbe Strategie auf einem oder mehreren synthetischen Datensätzen validieren, der dem realen nachempfunden

<sup>4</sup>Im Fachsprech wird diese Strategie, d.h., viele Vorhersagen auf verschiedenen Teildatensätzen machen und dann den Median nehmen, "bragging" genannt; ein Kofferwortsynonym für "robust bootstrap aggregation". Bragging ist verwandt mit der bekannteren und älteren, jedoch mathematisch ein bisschen komplizierteren "bagging"-Strategie nach Breiman, welche für die Determinanten-Vorhersagen in [1] verwendet wird. Auf dem Athleten-Datensatz sind bagging und bragging von Determinanten ungefähr gleich gut. Eine Übersicht über Bragging und andere Varianten des Bagging findet man in [2].

<sup>5</sup>wenn auch nicht der einzige. Es stimmt zwar, dass ich (a) an der Kreidetafel und im Kopf schon öfters Rang-Rätsel wie in den Tabellen 5 un 6 gelöst habe, (b) einen Bart trage, (c) beinahe schon überfahren wurde, und (d) mich als Mann fühle. Aber: (a) wie im Artikel erklärt wurde, sind diese tatsächlich praktisch nützlich, (b) es ist ein Schnurrbart, der ohnehin wieder in Mode kommt, (c) das geht den meisten Londonern genauso, und (d) ungefähr die Hälfte unserer Studenten sowie meiner Doktoranden sind weiblich. Mit diesen vier Dingen habe ich übrigens nicht nur kein Problem, sondern bin auch ganz glücklich darüber. Mit Ausnahme von (c) vielleicht.

ist, und prüfen, ob die Modellannahmen ausschlaggebend waren für die gute Vorhersage.

Des Weiteren würde man immer, auch hier, gerne exakt nachprüfen, warum denn funktioniert, was funktioniert - liegt es am Rang (hier: 2); daran, dass ein Low-Rank-Modell angenommen wurde; und/oder an der speziellen Methode, mit der man die Einträge einfüllt? Vielleicht werden ja nicht alle Komponenten benötigt. Tatsächlich stellt sich heraus, dass andere Methoden, inklusive bekannterer Strategien, die auf der Low-Rank-Annahme beruhen, auch solche mit Rang 2, nicht vergleichbar gut sind; und dass ein Rang von 3 sogar (unter Umständen) noch besser funktioniert. Das genau herauszuarbeiten bedarf aber entsprechend präziser empirischer Vergleiche, daher sei dafür abschließend ein letztes Mal auf das Athletik-Paper [1] verwiesen, und zum eigenen Experimentieren eingeladen.

## Ausklang

Und so kann man tatsächlich schlussfolgern, dass Sportergebnisse mit einer Sudoku-ähnlichen Strategie vorhergesagt werden können. Wobei eine solche Schlussfolgerung notwendigerweise auf einer irrtumsfrei folgerichtigen (= mathematisch korrekten) Argumentation und einer präzisen empirisch-experimentellen Validierung beruhen muss - was insbesondere bedeutet, dass irgendwo tatsächlich Sportergebnisse vorhergesagt werden müssen, und dass in Zahlen präzisiert werden muss, wie gut das gelungen ist.

Die besondere Schönheit einer solchen Validierung (wie im vorigen Abschnitt vorgestellt und z. B. in [4], Kapitel 7, ausführlicher erklärt) besteht darin, dass diese unabhängig von der Methode möglich ist, mit der die Vorhersagen erzielt werden; somit können beliebige Methoden auf ihre Vorhersagefähigkeit hin geprüft werden, sei es Raten, Determinanten, oder Krakenorkeln [12].

Um auf die anfängliche Diskussion zurückzukommen, genau darin beruht auch der wichtigste und größte<sup>5</sup> Unterschied zwischen Mathematiker und Klischeemathematiker: in der sorgfältigen Überprüfung von mathematischen und wissenschaftlichen Schlussfolgerungen. Wenn man sich für die reale Welt interessiert, bedeutet das insbesondere eine quantitative und empirisch-experimentelle Bewertung der Hypothesen (z. B. Methode X kann Y gut vorhersagen), die sich dann als falsch oder richtig herausstellen - und dadurch potentiell als praktisch nützlich. Ein bisschen Gehirnakrobatik mag ja dazugehören, aber das ist nicht so wichtig.

Ich würde Sie, geneigter Leser, nach diesen Worten herzlich dazu einladen, Tabellen, Daten, oder einfach Fakten, die Sie interessieren, aus einem solchen empirischen Blickwinkel zu betrachten, und in ihnen möglicherweise bisher ungesehene mathematische Gesetzmäßigkeiten - unter Umständen abseits Ihrer Lieblingsmethoden oder Lieblingsichtweisen - zu entdecken.

Falls Sie dafür gerne Anregungen hätten: wie wäre es denn mit den Filmbewertungen, die am Anfang nur relativ kurz besprochen wurden. Den 1-Million-Dollar-Netflix-Preis können Sie zwar leider nicht mehr gewinnen (dieser wurde im Jahre 2009 verliehen), und der Netflix-Preis-Datensatz ist aus Datenschutzgründen ebenfalls nicht mehr verfügbar [7]. Ein gerne genommener Ersatz ist aber z. B. der frei verfügbare MovieLens-Datensatz [10]; oder der Jester-Datensatz, benannt nach einem sich an den Benutzer anpassenden Witzempfehlungsalgorithmus [9].

Oder, falls Sie ein bisschen enttäuscht waren, dass es sich bei dem “Sport” in “Sportergebnisse” um Leichtathletik handelte: wie wäre es mit einem Rätsel wie in Tabelle 7, das einem Sudoku ebenfalls relativ ähnlich ist, und wo in der Regel<sup>6</sup> ebenfalls ganze Zahlen kleiner als 10 eingefüllt werden. So elementar die Fragestellung auch klingt, mit einer guten (d.h. wie beschrieben wissenschaftlich stringenten und statistisch validierten, und nicht einfach irgendeiner Unfugs-)Lösung wären Sie ganz vorne mit dabei im gerade aufkeimenden Forschungsgebiet der quantitativen Sportwissenschaften.

	2		1	0				3
4		1	4			2		2
	1		3	1			3	
2				1		0		1
0	1				0		3	
	3		0		2	2		2
0	2				3		1	
		1		2	2			3
0		0	2				3	

**Tabelle 7.** Eine Anregung zur Matrixvervollständigung. Ziel ist es, eine Methode zu finden, die die fehlenden Zahlen gut vorhersagt, und im Optimalfall auch modelliert. Hilfestellung: die (fehlende) Beschriftung der ersten Zeile ist “Bayern München”. Außerdem wurde die alberne Bezeichnung “Fußboku” absichtlich vermieden.

Ein Rätselhinweis zum Schluss: Low-Rank Matrix Completion ist möglicherweise nicht die beste Methode für die genannten Probleme, und auch nicht die einfachste Methode, die man probieren könnte. Aber vielleicht fällt Ihnen bei der kritischen Betrachtung der Daten ohnehin etwas Besseres ein, genau das ist ja das Spannende an der Wissenschaft.

## Literatur

- [1] Duncan AJ Blythe, Franz J Király. Quantification and prediction of individual athletic performance. *arXiv*, 1505:01147, 2015.
- [2] Peter Bühlmann. Bagging, subbagging and bragging for improving some prediction algorithms. *In: Recent Advances and Trends in Nonparametric Statistics*, Elsevier, 2003.
- [3] Howard Garns. Number Place. *Dell Pencil Puzzles & Word Games*, 16:6, 1979.
- [4] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Second Edition. *Springer Series in Statistics*, Springer, 2009.
- [5] Franz J Király.  
<http://mloss.org/software/view/524>
- [6] Franz J Király, Louis Theran, Ryota Tomioka. The Algebraic Combinatorial Approach for Low-Rank Matrix Completion. *Journal of Machine Learning Research*, 16(Aug):1391–1436, 2015.
- [7] Arvind Narayanan, Vitaly Shmatikov. Robust De-anonymization of Large Sparse Datasets *IEEE Symposium on Security and Privacy*, 111–125, 2008.
- [8] Ed Pegg Jr. Sudoku Variations. *mathpuzzle.com*, Sep 6, 2005.
- [9] Jester 5.0 - Jokes for your sense of humour  
<http://eigentaste.berkeley.edu/>  
Abgerufen Oktober 16, 2015.
- [10] The MovieLens data  
<http://grouplens.org/datasets/movielens/>  
Abgerufen Oktober 16, 2015.
- [11] The Netflix Prize Rules  
<http://www.netflixprize.com/rules>  
Abgerufen Oktober 16, 2015.
- [12] Wikipedia: Paul (Krake)  
[https://de.wikipedia.org/wiki/Paul\\_\(Krake\)](https://de.wikipedia.org/wiki/Paul_(Krake))  
Abgerufen Oktober 16, 2015.

<sup>6</sup>aber nicht immer, wie z. B. am 16. Spieltag der Saison 71/72