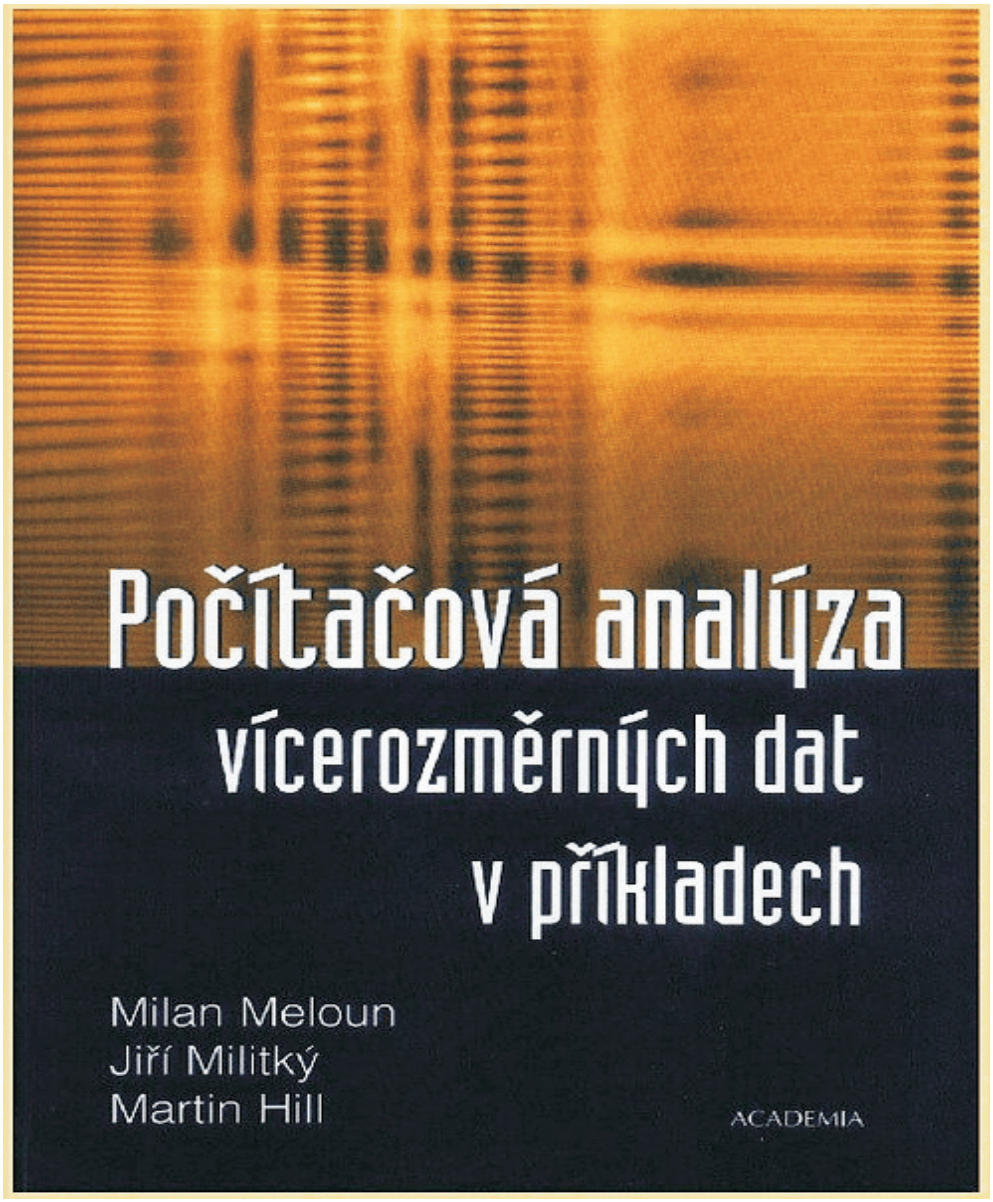
Počítačová analýza
vícerozměrných dat
v příkladech
Milan Meloun
Jiří Militký
Martin Hill
ACADEMIA

Milan Meloun
Jiří Militký
Martin Hill

Počítačová analýza vícerozměrných dat v příkladech

v oborech přírodních, technických i společenských věd

ACADEMIA 2005

O autorech

Prof. RNDr. **MILAN MELOUN**, DrSc., (*1943) vystudoval přírodovědeckou fakultu Univerzity J. E. Purkyně (dnešní Masarykova) v Brně 1965. Je profesorem analytické chemie a chemometrie na katedře analytické chemie Chemickotechnologické fakulty Univerzity v Pardubicích. Vyučoval statistické metody a analytickou chemii na Bagdádské univerzitě v Iráku a na Královské technice The Royal Institute of Technology ve Stockholmu. Je autorem a spoluautorem více než 150 originálních sdělení, 15 monografií a 11 vysokoškolských učebnic, 8 patentů a zlepšovacích návrhů a na konferencích přednesl více než 200 přednášek. Byl víc jak 10 let členem redakčních rad zahraničních odborných časopisů Talanta a Analytica Chimica Acta a je předsedou sekce chemometrie při České společnosti chemické.

Většina jeho publikací se týká počítačově orientované analýzy instrumentálních dat při studiu rovnováh v roztocích a určování reakčního produktu. Knižně se uvedl dvoudílnou monografií *Computation of Solution Equilibria*, Folia UJEP Brno 1984 (spoluautor Josef Havel), která pak doplněná o extrakční rovnováhy (spoluautor Erik Högfeldt) vyšla v roce 1988 v anglickém nakladatelství Ellis Horwood, Chichester. K této problematice se spolu s Josefem Havlem vrátil ve dvou obsáhlých kapitolách monografie *Computational Methods for the Determination of Formation Constants*, Plenum Press, New York 1985, jejímž editorem byl David Leggett.

Ve vědecké práci prof. Melouna představuje počítač spojovací článek mezi statistikou a analytickou chemií. Výsledkem je analytická chemometrie, předmět, který přednáší na Univerzitě v Pardubicích od roku 1978. Dokladem je řada učebních textů, např. *Chemometrie*, VŠCHT Pardubice 1985. V zahraničí přednášel dle textů svých učebnic, a to při dvouletém pobytu na univerzitě v Bagdádu *Data Analysis by Statistical and Computing Technique*, University Baghdad Press, Baghdad 1980. Při hostování na Královské univerzitě ve Stockholmu používal svou sbírku příkladů *Introduction to Chemometrics*, která je postavena na interaktivní analýze pomocí statistického systému STATGRAPHICS. Vlastní přístup k analýze experimentálních dat se promítá i do kapitoly *Chemometrics in the Instrumental Laboratory* v monografii, vydané editorem Jaroslavem Churáčkem: *Advanced Instrumental Methods of Chemical Analysis*, Academia, Praha 1993, nebo kapitoly *Hodnocení analytických výsledků* ve Vláčilově sbírce *Příklady z chemické a instrumentální analýzy*, SNTL, Praha 1983. Obsáhlá rešerše a zkušenosti spoluautora Jiřího Militkého přinesly řadu novějších postupů ze statistické analýzy dat, průzkumové analýzy a především interaktivní přístup k analýze dat na osobním počítači. Společně tak vzniklo první vydání učebnice *Chemometrie – Zpracování experimentálních dat na IBM PC.* Text byl přeložen do angličtiny a po doplnění o kapitoly vícerozměrné statistiky Michelem Forinou vyšel postupně jako dvoudílná učebnice u nakladatelství Ellis Horwood, Chichester 1991, s názvem *Chemometrics for Analytical Chemistry – Volume I. PC-Aided Statistical Data Analysis, Volume II. Regression Model Building and Testing.*

Na Univerzitě v Pardubicích přednáší v řádném studiu chemometrii, organizuje postgraduální a dvouleté licenční studium chemometrie a i krátkodobé intenzivní kurzy chemometrie pro aplikaci v průmyslu. V těchto formách studia chemometrie jsou užívány především 2 učebnice Milan Meloun a Jiří Militký: *Statistické zpracování experimentálních dat,* Finish, Pardubice 1992, PLUS, Praha 1994 a East Publishing 1998, Academia, Praha 2004 a Milan Meloun, Jiří Militký: *Kompendium statistického zpracování dat,* Academia Praha 2002.

Prof. Ing. **JIŘÍ MILITKÝ**, Csc., EURING, (*1949) ukončil fakultu textilní, specializaci textilní chemie, na VŠST v Liberci roku 1973 s vyznamenáním. V letech 1974 až 1976 pracoval ve Státním výzkumném ústavu textilním v Liberci, v oddělení matematického modelování textilních struktur. V letech 1976 až 1989 pracoval ve Výzkumném ústavu zušlechťovacím ve Dvoře Králové nad Labem, kde se věnoval převážně zpracování experimentálních dat s využitím výpočetní techniky. Od roku 1990 je vedoucím katedry textilních materiálů na Technické univerzitě v Liberci. V roce 1982 obhájil kandidátskou disertační práci z oblasti fyziky textilních vláken. V roce 1989 byl jmenován docentem a v roce 1992 se habilitoval. V prosinci 1993 byl jmenován řádným profesorem. Od 1.2.1994 do r. 1999 zastával funkci děkana fakulty textilní TUL. Od roku 1999–2003 zastával funkci prorektora pro vědu, výzkum a zahraniční spolupráci a od r. 2003 je opět děkanem fakulty textilní. V roce 1995 byl jmenován akademikem Ukrajinské akademie inženýrských věd a v r. 1996 obdržel titul EURING. Je členem několika vědeckých společností (The Textile Institute, JČMF) a pracuje ve výboru sekce chemometrie při České společnosti chemické.

Jeho publikační činnost zahrnuje oblasti textilního inženýrství, modelování kinetických procesů v pevné fázi a zpracování experimentálních dat. Je autorem nebo spoluautorem 606 vědeckých příspěvků (publikací, monografií, referátů a článků). Jeho první kniha *Modifikovaná PES vlákna* (spoluautoři Jiří Kryštůfek, Jiří Vaníček a Oldřich Hartych) vyšla v SNTL v roce 1984. Zcela přepracované a rozšíření vydání bylo publikováno nakladatelstvím Elsevier v roce 1991. S Jiřím Kryštůfkem zpracoval knihu *Barvení akrylových vláken a směsí*, která vyšla v Praze v nakl. SNTL v roce 1987. Ve spolupráci s Milanem Melounem publikoval učebnice a monografie z oblasti využití interaktivních statistických metod v chemometrii. Jiří Militký publikoval celkem 10 knih, z nichž tři jsou zaměřeny do oblasti zpracování experimentálních dat s využitím výpočetní techniky. Moderní metody interaktivní statistické analýzy dat zpracoval do rozsáhlého seriálu příruček *Statistické metody v textilní praxi I – IV,* vydaného v letech 1982 až 1985 v Domě techniky Pardubice. Přehled metod regrese a matematického modelování publikoval v seriálu skript *Tvorba matematických modelů I – IV*, vydaných v letech 1983 až 1989 v Domě techniky Ostrava. Vytvořil systém programů pro zpracování experimentálních dat v jazyce HPL. Tyto programy jsou charakteristické tím, že, kromě stránky statistické, vycházejí vždy nejdříve z ověřování předpokladů o modelech, datech a použité metodě a využívají také progresivních numerických postupů (zejména v oblasti lineární a nelineární regrese). Tyto algoritmy se později staly jádrem originálního programového systému ADSTAT.

Přednášel na odborných akcích v USA, Kanadě, Japonsku, Tchaj-wanu, Austrálii, Hongkongu, Vietnamu, Egyptě, Maroku a v řadě zemí Evropy. Je aktivně zapojen do činnosti v řadě odborných společností. Je členem výboru „International Textile Academy“, české pobočky „The Textile Institute“ a předsedou českého monitorovacího výboru FEANI. Je členem výboru České statistické společnosti a České chemické společnosti.

Ing. **Martin Hill**, DrSc. (1962). V roce 1986 absolvoval VŠCHT v Praze, obor chemické a energetické zpracování paliv. V letech 1987 až 1992 byl vědeckým aspirantem v Ústavu geotechniky v Praze. V roce 1992 obhájil hodnost kandidát technických věd za práci v oblasti kinetiky zplynování tuhých paliv. Je autorem nebo spoluautorem přes 150 odborných prací převážně v mezinárodních impaktovaných časopisech. Většina z těchto prací je postavena na využití pokročilých metod statistického zpracování vícerozměrných lékařských a biochemických dat.V roce 1996 absolvoval stáž v Conservatoire National des Arts et Métiers v Paříži, kde se zabýval antiglukokortikoidními účinky derivátů DHEA a pregnenolonu. Od roku byl nebo je hlavním řešitelem sedmi grantových projektů z nichž většina byla zařazena do soutěže o cenu ministra zdravotnictví a spoluřešitelem řady dalších z nichž jeden se umístil na prvním místě v uvedené soutěži. Za sérii metodických prací v oblasti analýzy méně běžných steroidů získal v roce 1998 Cenu Endokrinologické společnosti. V roce 1999 absolvoval dvouleté licenční studium chemometrie na Univerzitě Pardubice, které uzavřel prací na téma vícerozměrné statistické metody v analýze dat s negaussovským rozdělením. V roce 2001 obhájil na základě práce „Steroid analysis and data treatment for physiological and diagnostic conclusions“ vědeckou hodnost „Doktor chemických věd“ v oboru Analytická chemie. Má zkušenosti v oblasti kapalinové a plynové chromatografie, hmotové spektrometie, imunoanalýzy, chemometrie a statistiky, které uplatňuje jako vědecký pracovník v Oddělení steroidních hormonů Endokrinologického ústavu, kde je zaměstnán od r. 1992. V současnosti se zabývá vývojem nových metod analýzy steroidů ve speciální steroidní diagnostice a aplikací moderních metod statistické analýzy při zpracování biochemických dat.

Obsah

Předmluva

Zpracování vícerozměrných dat v technické praxi využívá poznatků *přírodních věd*, *matematické statistiky a informatiky* v kombinaci se speciálními počítačově orientovanými postupy. Současné výkonné osobní počítače umožňují interaktivnost při zpracování vícerozměrných dat a interpretaci získaných výsledků. To klade stále větší nároky na znalosti pracovníků, kteří data zpracovávají a analyzují. Nabídka a možnosti počítačově orientovaného statistického zpracování dat nutí ke komplexnější analýze problémů, což vede většinou i k radikální změně pohledu na metodiku jejich zkoumání.

Při zpracování reálných vícerozměrných dat se běžně naráží na řadu problémů a omezení:

(a) rozsahy zpracovávaných dat nejsou vzhledem k rozměrnosti problémů obyčejně dostatečně velké,

(b) v datech se vyskytují výrazné nelinearity, neaditivity a vzájemné vazby, které je třeba identifikovat a popsat,

(c) rozdělení dat jen zřídka odpovídá normálnímu běžně předpokládanému ve standardní statistické analýze,

(d) v datech se vyskytují podezřelá a odlehlá měření a různé heterogenity,

(e) statistické modely se často tvoří na základě předběžných informací z dat (datově orientované přístupy),

(f) existuje jistá neurčitost při výběru modelu, popisujícího chování dat.

To vše klade zvýšené nároky na techniky umožňující snižování rozměrnosti, hledání vnitřních skrytých vazeb v datech respektive vhodné zobrazení vícerozměrných dat. Pro tyto účely existuje celé spektrum méně či více dokonalých komplexních programů a programových systémů. Některé jsou budovány jako univerzálně použitelné a některé jsou zaměřené na specifické oblasti (chemometrie, biometrie, ekonometrie, medicínská statistika, obchodní statistika, statistika pro sociology, psychology, atd.). Jejich účinné využití není možné bez znalostí alespoň základů příslušných metod, které jsou základem pro interpretaci výsledků. Vlastní interpretace má obyčejně jak statistickou stránku tak i stránku související

s daným oborem.

Kniha je výsledkem snahy autorů překlenout rozdíl mezi pokrokem ve vývoji softwarových balíků obsahujících metody statistického zpracování vícerozměrných dat na jedné straně a v praxi stále nedostatečně využívanými možnostmi jejich užitečné aplikace na straně druhé. Přesto, že manipulace s daty je u moderních přístrojů a systémů snadná, vlastní výpočet je zpravidla otázkou sekund, počítačové výstupy z příslušného programu bývají přehledné a jejich interpretace ve většině případů nevyžaduje detailní matematickou znalost metodiky. Mezi odbornou, nematematicky zaměřenou veřejností stále přetrvávají obavy z využívání vícerozměrné statistické analýzy dat. Problémy činí především vlastní formulace úlohy a dostatečná interpretace výsledků. Specifickým problémem bývá také nepochopení výsledků a interpretace výstupů statistické vícerozměrné analýzy dat u části odborné veřejnosti. Přesto, že výsledky jsou pro člověka byť jen s povrchní znalostí uvedených metodik zcela zřejmé, stává se, že odborníci v daném oboru nejsou z důvodů vlastní neznalosti schopni těchto informací využít. Přitom efektivita využití informací z dat při použití vícerozměrných statistických technik je podstatně větší než u jednorozměrných dat.

Cílem této knihy je zpřístupnit vícerozměrné statistické techniky, které jsou dnes již běžnou součástí statistických softwarových balíků široké nematematické přírodovědné a technické veřejnosti a zejména studentům. Zvláštní důraz je kladen na sestavení úlohy a na interpretaci výsledků. Proto jsou těžištěm knihy návody s podrobnými postupy a komentované příklady s množstvím grafických výstupů umožňujících kromě diagnostiky kvality vstupních dat také snadnou interpretaci výsledků. Kromě sestavení úlohy bývá, zvláště u dat z oblasti biochemie a medicíny, problémem jejich nesymetrie rozdělení, nekonstantní rozptýlení a výskyt nehomogenit. Z těchto důvodů byla pozornost věnována i kvalitě vstupních dat a možnostem jejich transformace aby tak byly splněny základní předpoklady správného provedení vícerozměrné statistické analýzy. Příklady jsou voleny ze širokého spektra přírodních a technických věd, často z oblasti biochemie ale také z klinické praxe. Je na nich demonstrován zejména praktický přínos vícerozměrných statistických technik a zpravidla také jejich nezastupitelnost jednoduššími metodikami.

Kniha vychází a velmi úzce souvisí s nedávno zveřejněnou publikací Meloun M., Militký J.: Statistická analýza experimentálních dat, Academia Praha 2004, která obsahuje výklad jednotlivých vícerozměrných metod do větší hloubky ale s menším zaměřením na počítačově orientovanou analýzu resp. výsledky jednotlivých programů. Předpokládáme, že se kniha stane základem zpřístupnění vícerozměrného statistického zpracování dat širokému okruhu čtenářů, pracovníkům přírodovědných a technických oborů, studentům, lékařům resp. dalším specialistům, kteří zpracovávají vícerozměrná statistická data.

Na závěr je naší milou povinností poděkovat všem spolupracovníkům, studentům a doktorandům, které není možné ani všechny jmenovat a kteří nám pomáhali či přispěli praktickými úlohami, radami či konstruktivní kritikou. Vděk patří také všem pedagogům a studentům řádného i licenčního studia, kteří nám poskytli cenné dotazy, podněty a připomínky k řešeným příkladům.

Milan Meloun, Jiří Militký a Martin Hill

Pardubice, Liberec a Praha, leden 2005