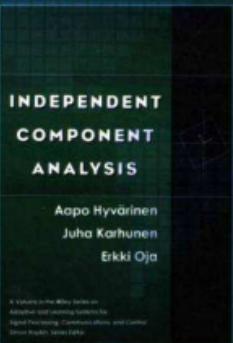




独立成分分析

Independent Component Analysis



Aapo Hyvärinen

[芬兰] Juha Karhunen 著
Erkki Oja

周宗潭 董国华 徐昕 胡德文 等译



電子工業出版社

Publishing House of Electronics Industry

<http://www.phei.com.cn>

独立成分分析

Independent Component Analysis

本书为学生和实际工作者提供了一个关于ICA技术的全面综合的介绍。

独立成分分析(ICA)是神经网络、高级统计学和信号处理等研究领域中最令人振奋的主题之一。本书是首次对ICA这门新技术进行全面综合介绍的专著，其中还包括了为理解和使用该技术相应的数学基础背景材料。本书不仅介绍了ICA的基本知识与总体概况，给出了重要的求解过程及算法，而且还涵盖了图像处理、无线通信、音频信号处理以及更多其他的应用。

独立成分分析的内容共分为四个部分，分别包括：

- 掌握ICA所需的一般性数学概念
- 基本ICA模型及其求解方法
- 基本ICA模型的各种扩展
- ICA模型的实际应用

本书的作者是Aapo Hyvärinen, Juha Karhunen和Erkki Oja，他们对发展ICA所做的贡献在业界是广为人知的，而这里还包含了更多的相关理论、新算法以及在各个领域的应用。不同学科领域的科研探索人员、学生和实际工作者将会从这本容易理解的书中获得帮助和教益。

作者介绍

Aapo Hyvärinen: 博士，芬兰科学院资深院士，目前在芬兰赫尔辛基技术大学神经网络研究中心工作。

Juha Karhunen 和 Erkki Oja 均为芬兰赫尔辛基技术大学神经网络研究中心教授。



ISBN 978-7-121-04293-5



9 787121 042935 >



责任编辑：李秦华
责任美编：毛惠庚

本书贴有激光防伪标志，凡没有防伪标志者，属盗版图书。

定价：49.00 元

国外电子与通信教材系列

独立成分分析

Independent Component Analysis

Aapo Hyvärinen

[芬兰] Juha Karhunen 著

Erkki Oja

周宗潭 董国华 徐 昕 胡德文 等译

電子工業出版社

Publishing House of Electronics Industry

北京 · BEIJING

内 容 简 介

独立成分分析(ICA)已经成为近年来神经网络、高级统计学和信号处理等研究领域中最令人振奋的主题之一。ICA 源自对客观物理世界的抽象，它能够有效地解决许多实际问题，具有强大的生命力和广阔的工程应用前景。本书(英文原版)是国际上第一本对 ICA 这门新技术进行全面介绍的综合性专著，其中还包括了为理解和使用该技术的相应数学基础背景材料。本书不仅介绍了 ICA 的基本知识与总体概况、给出了重要的求解过程及算法，而且还涵盖了图像处理、无线通信、音频信号处理以及更多其他应用。

全书分为四个部分，共 24 章。第一部分（第 2 章至第 6 章）介绍了本书所用到的主要数学知识，第二部分（第 7 章至第 14 章）是本书的重点，详细讲述了基本 ICA 模型及其求解过程，第三部分（第 15 章至第 20 章）讨论了基本 ICA 模型的多种扩展形式，第四部分（第 21 章至第 24 章）对 ICA 方法在不同领域的应用做了生动的阐述。

本书可作为不同工程应用领域的大学教师、研究生和科技工作者的 ICA 入门教材；而对于探索 ICA 技术的专业研究人员来说，本书也是一本极有价值的参考书。

0-471-40540-X, Independent Component Analysis by Aapo Hyvärinen, Juha Karhunen and Erkki Oja.

Original English language edition copyright © 2001, John Wiley & Sons, Inc. All Rights Reserved. This translation published under license.

本书中文简体字翻译版由 John Wiley & Sons Inc. 授予电子工业出版社。未经出版者预先书面许可，不得以任何方式复制或抄袭本书的任何部分。

版权贸易合同登记号 图字：01-2006-0894

图书在版编目 (CIP) 数据

独立成分分析 / (芬) 海韦里恩 (Hyvärinen, A.) 等著；周宗潭等译。

北京：电子工业出版社，2007.6

(国外电子与通信教材系列)

书名原文：Independent Component Analysis

ISBN 978-7-121-04293-5

I. 独... II. ①海... ②周... III. 信号处理 - 教材 IV. TN911.7

中国版本图书馆 CIP 数据核字 (2007) 第 060682 号

责任编辑：李秦华

印 刷：北京季蜂印刷有限公司

装 订：三河市万和装订厂

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱 邮编：100036

开 本：787 × 1092 1/16 印张：23.25 字数：595 千字

印 次：2007 年 6 月第 1 次印刷

定 价：49.00 元

凡所购买电子工业出版社的图书有缺损问题，请向购买书店调换；若书店售缺，请与本社发行部联系。联系及邮购电话：(010) 88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：(010) 88258888。

读者调查表

感谢对我们的支持！非常欢迎留下您的宝贵意见，帮助我们改进出版和服务工作。我们将从信息意见完备的读者中抽取一部分赠阅一本我们的样书（赠书定价限 50 以内，品种我们会与获赠读者沟通）。

姓名：_____ 单位：_____ 职务 / 职称：_____

邮寄地址：_____ 邮编：_____

电话：_____ 手机：_____ E-mail：_____ 专业方向：_____

您购买的出版物名称					
先进性和实用性	<input type="checkbox"/> 很好	<input type="checkbox"/> 好	<input type="checkbox"/> 一般	<input type="checkbox"/> 不太好	<input type="checkbox"/> 差
图书文字可读性	<input type="checkbox"/> 很好	<input type="checkbox"/> 好	<input type="checkbox"/> 一般	<input type="checkbox"/> 不太好	<input type="checkbox"/> 差
(光盘使用方便性)	<input type="checkbox"/> 很好	<input type="checkbox"/> 好	<input type="checkbox"/> 一般	<input type="checkbox"/> 不太好	<input type="checkbox"/> 差
图书篇幅适宜度	<input type="checkbox"/> 很合适	<input type="checkbox"/> 合适	<input type="checkbox"/> 一般	<input type="checkbox"/> 不合适	<input type="checkbox"/> 差
出版物中差错	<input type="checkbox"/> 极少	<input type="checkbox"/> 较少	<input type="checkbox"/> 一般	<input type="checkbox"/> 较多	<input type="checkbox"/> 太多
封面(盘面及包装)设计水平	<input type="checkbox"/> 很好	<input type="checkbox"/> 好	<input type="checkbox"/> 一般	<input type="checkbox"/> 不太好	<input type="checkbox"/> 差
图书(包括光盘)印装质量	<input type="checkbox"/> 很好	<input type="checkbox"/> 好	<input type="checkbox"/> 一般	<input type="checkbox"/> 不太好	<input type="checkbox"/> 差
纸张质量(光盘材质)	<input type="checkbox"/> 很好	<input type="checkbox"/> 好	<input type="checkbox"/> 一般	<input type="checkbox"/> 不太好	<input type="checkbox"/> 差
定价	<input type="checkbox"/> 很便宜	<input type="checkbox"/> 便宜	<input type="checkbox"/> 合理	<input type="checkbox"/> 贵	<input type="checkbox"/> 太贵
您从何处获取出版物信息	<input type="checkbox"/> 书目	<input type="checkbox"/> 电子社宣传材料	<input type="checkbox"/> 书店	<input type="checkbox"/> 他人转告	<input type="checkbox"/> 网站
您的具体意见或建议					

您或周围人士有何著述计划 _____

您希望我处增添何种类型的图书 _____

电子工业出版社高等教育分社

联系人：冯小贝 E-mail: fengxiaobei@phei.com.cn, te_service@phei.com.cn

地址：北京市万寿路 173 信箱 1102 室 邮编：100036 电话：010-88254555

传真：010-88254560

反侵权盗版声明

电子工业出版社依法对本作品享有专有出版权。任何未经权利人书面许可，复制、销售或通过信息网络传播本作品的行为；歪曲、篡改、剽窃本作品的行为，均违反《中华人民共和国著作权法》，其行为人应承担相应的民事责任和行政责任，构成犯罪的，将被依法追究刑事责任。

为了维护市场秩序，保护权利人的合法权益，我社将依法查处和打击侵权盗版的单位和个人。欢迎社会各界人士积极举报侵权盗版行为，本社将奖励举报有功人员，并保证举报人的信息不被泄露。

举报电话：(010) 88254396; (010) 88258888

传 真：(010) 88254397

E-mail : dbqq@phe.com.cn

通信地址：北京市万寿路173信箱

电子工业出版社总编办公室

邮 编：100036

序

2001年7月间，电子工业出版社的领导同志邀请各高校十几位通信领域方面的老师，商量引进国外教材问题。与会同志对出版社提出的计划十分赞同，大家认为，这对我国通信事业、特别是对高等院校通信学科的教学工作会很有好处。

教材建设是高校教学建设的主要内容之一。编写、出版一本好的教材，意味着开设了一门好的课程，甚至可能预示着一个崭新学科的诞生。20世纪40年代MIT林肯实验室出版的一套28本雷达丛书，对近代电子学科、特别是对雷达技术的推动作用，就是一个很好的例子。

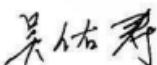
我国领导部门对教材建设一直非常重视。20世纪80年代，在原教委教材编审委员会的领导下，汇集了高等院校几百位富有教学经验的专家，编写、出版了一大批教材；很多院校还根据学校的特点和需要，陆续编写了大量的讲义和参考书。这些教材对高校的教学工作发挥了极好的作用。近年来，随着教学改革不断深入和科学技术的飞速进步，有的教材内容已比较陈旧、落后，难以适应教学的要求，特别是在电子学和通信技术发展神速、可以讲是日新月异的今天，如何适应这种情况，更是一个必须认真考虑的问题。解决这个问题，除了依靠高校的老师和专家撰写新的符合要求的教科书外，引进和出版一些国外优秀电子与通信教材，尤其是有选择地引进一批英文原版教材，是会有好处的。

一年多来，电子工业出版社为此做了很多工作。他们成立了一个“国外电子与通信教材系列”项目组，选派了富有经验的业务骨干负责有关工作，收集了230余种通信教材和参考书的详细资料，调来了100余种原版教材样书，依靠由20余位专家组成的出版委员会，从中精选了40多种，内容丰富，覆盖了电路理论与应用、信号与系统、数字信号处理、微电子、通信系统、电磁场与微波等方面，既可作为通信专业本科生和研究生的教学用书，也可作为有关专业人员的参考材料。此外，这批教材，有的翻译为中文，还有部分教材直接影印出版，以供教师用英语直接授课。希望这些教材的引进和出版对高校通信教学和教材改革能起一定作用。

在这里，我还要感谢参加工作的各位教授、专家、老师与参加翻译、编辑和出版的同志们。各位专家认真负责、严谨细致、不辞辛劳、不怕琐碎和精益求精的态度，充分体现了中国教育工作者和出版工作者的良好美德。

随着我国经济建设的发展和科学技术的不断进步，对高校教学工作会不断提出新的要求和希望。我想，无论如何，要做好引进国外教材的工作，一定要联系我国的实际。教材和学术专著不同，既要注意科学性、学术性，也要重视可读性，要深入浅出，便于读者自学；引进的教材要适应高校教学改革的需要，针对目前一些教材内容较为陈旧的问题，有目的地引进一些先进的和正在发展中的交叉学科的参考书；要与国内出版的教材相配套，安排好出版英文原版教材和翻译教材的比例。我们努力使这套教材能尽量满足上述要求，希望它们能放在学生们的课桌上，发挥一定的作用。

最后，预祝“国外电子与通信教材系列”项目取得成功，为我国电子与通信教学和通信产业的发展培土施肥。也恳切希望读者能对这些书籍的不足之处、特别是翻译中存在的问题，提出意见和建议，以便再版时更正。



中国工程院院士、清华大学教授
“国外电子与通信教材系列”出版委员会主任

出版说明

进入21世纪以来，我国信息产业在生产和科研方面都大大加快了发展速度，并已成为国民经济发展的支柱产业之一。但是，与世界上其他信息产业发达的国家相比，我国在技术开发、教育培训等方面都还存在着较大的差距。特别是在加入WTO后的今天，我国信息产业面临着国外竞争对手的严峻挑战。

作为我国信息产业的专业科技出版社，我们始终关注着全球电子信息技术的发展方向，始终把引进国外优秀电子与通信信息技术教材和专业书籍放在我们工作的重要位置上。在2000年至2001年间，我社先后从世界著名出版公司引进出版了40余种教材，形成了一套“国外计算机科学教材系列”，在全国高校以及科研部门中受到了欢迎和好评，得到了计算机领域的广大教师与科研工作者的充分肯定。

引进和出版一些国外优秀电子与通信教材，尤其是有选择地引进一批英文原版教材，将有助于我国信息产业培养具有国际竞争能力的技术人才，也将有助于我国国内在电子与通信教学工作中掌握和跟踪国际发展水平。根据国内信息产业的现状、教育部《关于“十五”期间普通高等教育教材建设与改革的意见》的指示精神以及高等院校老师们反映的各种意见，我们决定引进“国外电子与通信教材系列”，并随后开展了大量准备工作。此次引进的国外电子与通信教材均来自国际著名出版商，其中影印教材约占一半。教材内容涉及的学科方向包括电路理论与应用、信号与系统、数字信号处理、微电子、通信系统、电磁场与微波等，其中既有本科专业课程教材，也有研究生课程教材，以适应不同院系、不同专业、不同层次的师生对教材的需求，广大师生可自由选择和自由组合使用。我们还将与国外出版商一起，陆续推出一些教材的教学支持资料，为授课教师提供帮助。

此外，“国外电子与通信教材系列”的引进和出版工作得到了教育部高等教育司的大力支持和帮助，其中的部分引进教材已通过“教育部高等学校电子信息科学与工程类专业教学指导委员会”的审核，并得到教育部高等教育司的批准，纳入了“教育部高等教育司推荐——国外优秀信息科学与技术系列教学用书”。

为做好该系列教材的翻译工作，我们聘请了清华大学、北京大学、北京邮电大学、南京邮电大学、东南大学、西安交通大学、天津大学、西安电子科技大学、电子科技大学、中山大学、哈尔滨工业大学、西南交通大学等著名高校的教授和骨干教师参与教材的翻译和审校工作。许多教授在国内电子与通信专业领域享有较高的声望，具有丰富的教学经验，他们的渊博学识从根本上保证了教材的翻译质量和专业学术方面的严格与准确。我们在此对他们的辛勤工作与贡献表示衷心的感谢。此外，对于编辑的选择，我们达到了专业对口；对于从英文原书中发现的错误，我们通过与作者联络、从网上下载勘误表等方式，逐一进行了修订；同时，我们对审校、排版、印制质量进行了严格把关。

今后，我们将进一步加强同各高校教师的密切关系，努力引进更多的国外优秀教材和教学参考书，为我国电子与通信教材达到世界先进水平而努力。由于我们对国内外电子与通信教育的发展仍存在一些认识上的不足，在选题、翻译、出版等方面的工作中还有许多需要改进的地方，恳请广大师生和读者提出批评及建议。

电子工业出版社

教材出版委员会

主任	吴佑寿	中国工程院院士、清华大学教授
副主任	林金桐	北京邮电大学校长、教授、博士生导师
	杨千里	总参通信部副部长、中国电子学会会士、副理事长 中国通信学会常务理事、博士生导师
委员	林孝康	清华大学教授、博士生导师、电子工程系副主任、通信与微波研究所所长 教育部电子信息科学与工程类专业教学指导分委员会委员
	徐安士	北京大学教授、博士生导师、电子学系主任
	樊昌信	西安电子科技大学教授、博士生导师 中国通信学会理事、IEEE 会士
	程时昕	东南大学教授、博士生导师
	郁道银	天津大学副校长、教授、博士生导师 教育部电子信息科学与工程类专业教学指导分委员会委员
	阮秋琦	北京交通大学教授、博士生导师 计算机与信息技术学院院长、信息科学研究所所长 国务院学位委员会学科评议组成员
	张晓林	北京航空航天大学教授、博士生导师、电子信息工程学院院长 教育部电子信息科学与电气信息类基础课程教学指导分委员会副主任委员 中国电子学会常务理事
	郑宝玉	南京邮电大学副校长、教授、博士生导师 教育部电子信息与电气学科教学指导委员会委员
	朱世华	西安交通大学副校长、教授、博士生导师 教育部电子信息科学与工程类专业教学指导分委员会副主任委员
	彭启琮	电子科技大学教授、博士生导师、通信与信息工程学院院长 教育部电子信息科学与电气信息类基础课程教学指导分委员会委员
	毛军发	上海交通大学教授、博士生导师、电子信息与电气工程学院副院长 教育部电子信息与电气学科教学指导委员会委员
	赵尔沅	北京邮电大学教授、《中国邮电高校学报（英文版）》编委会主任
	钟允若	原邮电科学研究院副院长、总工程师
	刘 彩	中国通信学会副理事长兼秘书长，教授级高工 信息产业部通信科技委副主任
	杜振民	电子工业出版社原副社长
	王志功	东南大学教授、博士生导师、射频与光电集成电路研究所所长 教育部高等学校电子电气基础课程教学指导分委员会主任委员
	张中兆	哈尔滨工业大学教授、博士生导师、电子与信息技术研究院院长
	范平志	西南交通大学教授、博士生导师、信息科学与技术学院院长

目 录

第 1 章 引论	1
1.1 多元数据的线性表示	1
1.2 盲源分离	2
1.3 独立成分分析	4
1.4 ICA 的历史	7
第一部分 数学预备知识	
第 2 章 随机向量和独立性	10
2.1 概率分布和概率密度	10
2.2 期望和矩	13
2.3 不相关性和独立性	17
2.4 条件密度和贝叶斯法则	20
2.5 多元高斯密度	22
2.6 变换的密度	25
2.7 高阶统计量	25
2.8 随机过程*	31
2.9 小结与文献引述	36
习题	37
计算机练习	40
第 3 章 梯度和最优化方法	42
3.1 向量和矩阵梯度	42
3.2 无约束优化和学习规则	46
3.3 约束优化的学习规则	54
3.4 小结与文献引述	56
习题	56
计算机练习	57
第 4 章 估计理论	58
4.1 基本概念	58
4.2 估计器的性质	59
4.3 矩方法	62
4.4 最小二乘估计	64
4.5 极大似然法	67
4.6 贝叶斯估计*	69

4.7 小结与文献引述	73
习题	74
计算机练习	77
第5章 信息论	78
5.1 熵	78
5.2 互信息	81
5.3 极大熵	82
5.4 负熵	83
5.5 通过累积量逼近熵	83
5.6 用非多项式函数近似熵	85
5.7 小结与文献引述	89
习题	89
计算机练习	90
本章附录:有关证明	90
第6章 主成分分析和白化	93
6.1 主成分	93
6.2 在线学习的 PCA	97
6.3 因子分析	101
6.4 白化	102
6.5 正交化	103
6.6 小结与文献引述	105
习题	105

第二部分 独立成分分析基本模型

第7章 什么是独立成分分析	108
7.1 动机	108
7.2 独立成分分析的定义	111
7.3 ICA 的实例	114
7.4 ICA 比白化更加强大	115
7.5 高斯变量为何不能适用	117
7.6 小结与文献引述	118
习题	119
计算机练习	119
第8章 极大化非高斯性的 ICA 估计方法	120
8.1 非高斯就是独立的	120
8.2 用峭度来度量非高斯性	123
8.3 用负熵度量非高斯性	129
8.4 估计多个独立成分	137
8.5 ICA 与投影寻踪	139

8.6 小结与文献引述	141
习题	141
计算机练习	142
本章附录:有关证明	143
第 9 章 ICA 的极大似然估计方法	145
9.1 ICA 模型中的似然度	145
9.2 极大似然估计算法	147
9.3 信息极大原理	151
9.4 例子	151
9.5 小结与文献引述	153
习题	154
计算机练习	154
本章附录:有关证明	155
第 10 章 极小化互信息的 ICA 估计方法	156
10.1 用互信息定义 ICA	156
10.2 互信息和非高斯性	157
10.3 互信息和似然估计	157
10.4 极小化互信息的算法	158
10.5 例子	158
10.6 小结与文献引述	159
习题	159
计算机练习	159
第 11 章 基于张量的 ICA 估计方法	160
11.1 累积张量的定义	160
11.2 由张量特征值得到独立成分	160
11.3 用幂法计算张量分解	162
11.4 特征矩阵的联合近似对角化	163
11.5 加权相关矩阵方法	164
11.6 小结与文献引述	165
习题	165
计算机练习	165
第 12 章 基于非线性去相关和非线性 PCA 的 ICA 估计方法	166
12.1 非线性相关和独立性	166
12.2 Hérault-Jutten 算法	168
12.3 Cichocki-Umbenauen 算法	169
12.4 估计函数方法*	170
12.5 通过独立性的等变自适应分离(EASI)	171
12.6 非线性主成分	172

12.7 非线性 PCA 指标和 ICA	175
12.8 非线性 PCA 指标的学规则	176
12.9 小结与文献引述	182
习题	182
第 13 章 实际的考虑	184
13.1 时间滤波作为预处理	184
13.2 用 PCA 进行预处理	186
13.3 应该估计多少个成分	188
13.4 算法选择	189
13.5 小结与文献引述	189
习题	189
计算机练习	190
第 14 章 基本 ICA 方法的综述和比较	191
14.1 目标函数和算法	191
14.2 ICA 估计原理的联系	191
14.3 统计最优非线性函数	193
14.4 ICA 算法的实验比较	195
14.5 参考文献	199
14.6 基本 ICA 方法小结	200
本章附录: 有关证明	201

第三部分 ICA 的扩展及其相关方法

第 15 章 有噪声的 ICA 模型	204
15.1 定义	204
15.2 传感器噪声和信号源噪声	204
15.3 噪声成分数目较少的情况	205
15.4 混合矩阵的估计	205
15.5 估计无噪声的独立成分	208
15.6 通过稀疏编码收缩而去噪	211
15.7 小结	211
第 16 章 具有超完备基的 ICA 模型	212
16.1 独立成分的估计	212
16.2 估计混合矩阵	213
16.3 小结	217
第 17 章 非线性 ICA	218
17.1 非线性 ICA 与 BSS	218
17.2 后非线性混合的分离	221
17.3 采用自组织映射的非线性 BSS	222

17.4	非线性 BSS 的一种生成拓扑映射方法*	223
17.5	非线性 BSS 的一种集成学习方法	227
17.6	其他方法	234
17.7	小结	235
第 18 章	使用时间结构的方法	237
18.1	通过自协方差实现分离	237
18.2	利用方差的非平稳性实现分离	241
18.3	统一的分离原理	245
18.4	小结	246
第 19 章	卷积性混合和盲去卷积	247
19.1	盲去卷积	247
19.2	卷积性混合的盲分离	251
19.3	小结	256
本章附录:离散时间滤波器和 z 变换		257
第 20 章	ICA 的其他扩展	259
20.1	混合矩阵的先验信息	259
20.2	放宽独立性假设	264
20.3	复值数据的处理	268
20.4	小结	271

第四部分 ICA 的应用

第 21 章	基于 ICA 的特征提取	274
21.1	线性表示	274
21.2	ICA 和稀疏编码	277
21.3	从图像中估计 ICA 的基向量	278
21.4	压缩稀疏编码用于图像去噪	279
21.5	独立子空间和拓扑 ICA	281
21.6	与神经生理学的联系	283
21.7	小结	283
第 22 章	ICA 在脑成像中的应用	285
22.1	脑电图和脑磁图	285
22.2	EEG 和 MEG 中的伪迹鉴别	286
22.3	诱发磁场分析	288
22.4	ICA 使用于其他的测量技术中	290
22.5	小结	291
第 23 章	无线通信	292
23.1	多用户检测和 CDMA 通信	292
23.2	CDMA 信号模型和 ICA	295

23.3	衰落信道的估计	297
23.4	卷积 CDMA 信号的盲分离*	301
23.5	采用复值 ICA 改进多用户检测*	305
23.6	小结与文献引述	309
第 24 章	ICA 的其他应用	310
24.1	金融方面的应用	310
24.2	音频分离	313
24.3	更多的应用领域	314
参考文献		316
中英文术语对照		341

第1章 引 论

独立成分分析(ICA)是从多元(多维)统计数据中寻找其内在因子或成分的一种方法。ICA有别于其他方法的地方是,它寻找的是既统计独立又非高斯的成分。这里我们简要地介绍ICA的基本概念、应用及其估计原理。

1.1 多元数据的线性表示

1.1.1 一般统计框架

寻找多元数据的一种好的表示法,一直是统计学及相关领域中长期存在的问题。在这里,表示这个词指的是我们以某种方式对数据进行变换,使得其本质结构更显著或更容易理解。

在神经计算中,这个基本问题属于无监督学习的范畴,因为该表示必须从数据自身学习得出,而不需要从一个指导“教师”那里获取任何外部输入。获得一个合理的表示也是数据挖掘和探索性数据分析的许多技术的核心目标。在信号处理中,同样的问题出现在特征提取问题以及下面即将考虑的盲源分离问题中。

假设数据由已经得到其观测的一组变量构成。记变量数目为 m , 观测数目为 T 。这样可将数据记为 $x_i(t)$, 其中,下标取值 $i = 1, \dots, m$ 而 $t = 1, \dots, T$ 。维数 m 和 T 可能非常大。

该问题的一种非常通用的表述如下:从 m 维空间到 n 维空间的什么函数可使得变换后的变量能够凸显原本隐藏在大量数据集中的信息。也就是说,变换后的变量应是内在因子或成分,它们描述了数据的本质结构。我们当然希望这些成分对应于数据生成过程中的某些物理原因。

大多数情况下我们只考虑线性函数,因为这样可使表示的解释与计算更加简单。这样,每个成分,比方说 y_i ,可表示成观测变量的一个线性组合:

$$y_i(t) = \sum_j w_{ij} x_j(t), i = 1, \dots, n, j = 1, \dots, m \quad (1.1)$$

式中, w_{ij} 是定义上述表示的某个系数。于是,原来的问题可以重新表述成如何确定系数 w_{ij} 的另外一个问题。利用线性代数,我们可以将公式(1.1)中的线性变换表示成矩阵乘法。将系数 w_{ij} 纳入矩阵 \mathbf{W} ,则该方程变成:

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{bmatrix} = \mathbf{W} \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_m(t) \end{bmatrix} \quad (1.2)$$

一种基本的统计处理方法是将 $x_i(t)$ 看成 m 个随机变量的 T 个实现。这样,每组 $x_i(t), t = 1, \dots, T$ 是某个随机变量的一组样本;我们将随机变量记为 x_i 。在此框架下,可以根据变换后成分 y_i 的统计特性来确定矩阵 \mathbf{W} 。在以下各节中,将讨论一些后面用到的统计性质;其中之一将导致独立成分分析。

1.1.2 降维方法

选择矩阵 W 的一条统计方面的原则是, 将成分 y_i 的数目限制为非常小(可能仅为 1 或 2)并使得 y_i 包含数据中尽可能多的信息。这导致了一类称为主成分分析或因子分析的技术。

在一篇经典文献中, Spearman [409] 考虑了在不同科目学习学生的成绩数据, 另外还补充了一些实验室测量数据。Spearman 寻找单个线性组合并使之解释数据中最大的差异, 通过这种方法确定了 W 。他声称找到了智力的一个普遍的因子, 于是创立了因子分析方法, 同时, 也开始了心理学中一场旷日持久的论战。

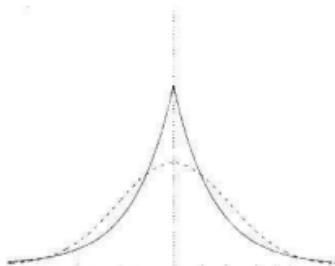
1.1.3 独立性作为一个指导原则

确定 W 的另一个指导原则是独立性: 成分 y_i 之间应统计独立。这意味着, 任何一个成分的取值不能给出其他成分取值的任何信息。

事实上, 在因子分析中, 经常声称各因子是独立的, 但这仅仅部分正确, 因为因子分析假设数据服从高斯分布。若数据是高斯的, 寻找独立的成分非常简单, 因为对于高斯数据, 不相关成分总是独立的。然而, 现实中数据通常并不服从高斯分布, 这时情况就不像那些方法假设的那样简单。例如, 许多现实数据集具有超高斯分布。超高斯分布是指该随机变量取值出现在零附近和较大数值处更为频繁。换句话说, 相比具有相同方差的高斯密度而言, 数据的概率密度函数在零处具有尖峰, 且有较长的拖尾(在远离零处取值较大)。图 1.1 显示了此种概率密度的一个例子。

上述内容是 ICA 研究的出发点。我们希望, 在数据是非高斯的一般情况下, 把统计独立的成分找出来。

图 1.1 拉普拉斯分布的密度函数, 该分布是一个典型的超高斯分布。虚线是作为对比的相应高斯密度函数。拉普拉斯密度函数在零处有一个尖峰, 其边缘的拖尾较长。两种分布都被归一化为零均值和单位方差



1.2 盲源分离

下面我们从与寻找好的表示法不同的另外一种角度来探讨同一个问题。该问题属于信号处理的范畴, 它也展现了 ICA 出现的历史背景。

1.2.1 未知信号的观测混合

让我们来考虑这样的一般情况: 有这么一组信号, 是由几个物理对象或物理源发出的, 物理源可以是发出电信号的不同脑区、可以是在同一房间讲话的人, 也可能是发射无线电波的移动电话。进一步假设存在多个传感器或接收机, 而这些传感器安置在不同位置, 从而, 每个传感器可以分别以略为不同的权重记录各物理源信号的某种混合。

为使问题的陈述更为简单, 我们假定有三个源信号, 同时有三个观测信号。把观测信号记为 $x_1(t), x_2(t)$ 和 $x_3(t)$, 它们是所记录的信号在 t 时间点处的幅值; 原始信号记为 $s_1(t), s_2(t)$ 和 $s_3(t)$ 。这样, $x_i(t)$ 是 $s_i(t)$ 的加权和, 而加权系数依赖于源和传感器之间的距离:

$$\begin{aligned}x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t) \\x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t) \\x_3(t) &= a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t)\end{aligned}\quad (1.3)$$

式中, a_{ij} 是常值系数, 表示混合的权重。 a_{ij} 是未知的, 因为我们不可能了解物理混合系统的全部特性(这通常是极其困难的), 所以也无法知道 a_{ij} 的值。源信号 s_i 也同样是未知的, 而这正是要解决的问题: 因为我们不能对它们进行直接记录。

作为一个直观示例, 可以考虑图 1.2 中的波形。这些波形分别是一些源信号的三个线性混合量 x_i , 它们看起来像纯粹的噪声, 但事实上这些观测信号里, 隐藏着一些具有相当结构化特性的源信号。

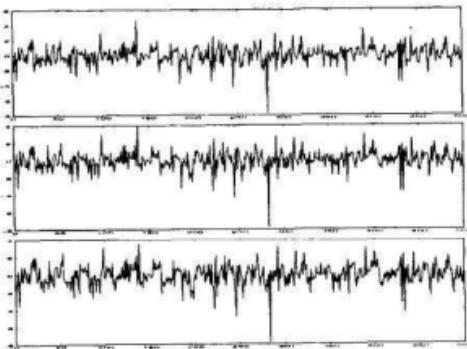


图 1.2 一组观测信号的波形, 假定它们是一些内在源信号的混合

我们想要做的就是利用 $x_1(t)$, $x_2(t)$ 和 $x_3(t)$ 这些混合量找出原始信号。这就是盲源分离(BSS)问题。盲意味着我们对于原始信号所知甚少。

不妨假设混合系数 a_{ij} 具有足够的差异, 使得它们构成的矩阵可逆。因此存在一个以元素 w_{ij} 为系数的矩阵 W , 使得我们可以用它分离出源信号 s_i :

$$\begin{aligned}s_1(t) &= w_{11}x_1(t) + w_{12}x_2(t) + w_{13}x_3(t) \\s_2(t) &= w_{21}x_1(t) + w_{22}x_2(t) + w_{23}x_3(t) \\s_3(t) &= w_{31}x_1(t) + w_{32}x_2(t) + w_{33}x_3(t)\end{aligned}\quad (1.4)$$

如果我们已经知道公式(1.3)中的那些系数 a_{ij} , 将它们形成的矩阵进行求逆, 即可得到矩阵 W 。

现在我们可以看到, 上述问题实际上和前面希望为 $x_i(t)$ 中的随机数据寻找一个好的表示法[参见式(1.2)]的问题在数学上非常类似。事实上, 我们可以将每个信号 $x_i(t)$, $t = 1, \dots, T$ 看做随机变量 x_i 的一组样本, 使得随机变量的值可由该信号在所记录时间点处的幅值给出。

1.2.2 基于独立性的源分离

现在的问题是: 如何估计公式(1.4)中的系数 w_{ij} ? 我们希望获得一种普适性的方法, 使它能适用于许多不同的场合, 并给最开始提出的问题——即为多元数据寻找一个好的表示法, 提供

一种答案。但是我们只能使用非常一般的统计性质,因为 x_1, x_2 和 x_3 是我们的全部观测。我们还希望找到一个矩阵 \mathbf{W} ,使得这个好的表示法可以用源信号 s_1, s_2 和 s_3 给出。

仅仅通过考虑信号的统计独立性,就可以找到上述问题的一个令人惊奇的简单求解方式。事实上,如果信号是非高斯的,那么只需确定系数 w_{ij} ,使得信号:

$$\begin{aligned}y_1(t) &= w_{11}x_1(t) + w_{12}x_2(t) + w_{13}x_3(t) \\y_2(t) &= w_{21}x_1(t) + w_{22}x_2(t) + w_{23}x_3(t) \\y_3(t) &= w_{31}x_1(t) + w_{32}x_2(t) + w_{33}x_3(t)\end{aligned}\quad (1.5)$$

是统计独立的即可。如果信号 y_1, y_2 和 y_3 是统计独立的,那么它们就等同于原始信号 s_1, s_2 和 s_3 (它们之间可能是某种标量常数乘积的关系,也就是说一个信号可能是另一个信号乘以一个标量的比例常数,但这并不重要)。事实上,仅仅利用统计独立性的信息,我们就可以估计出图 1.2 中信号所对应的系数矩阵 \mathbf{W} ,从而得到如图 1.3 所示的源信号(这些信号是用我们将会在本书后面多个章节里遇到的 FastICA 算法估计出的)。可以看到,从一个貌似噪声的数据集中,利用一个只用到统计独立性信息的算法,就能将源信号估计出来。而估计得到的信号确实等于我们用于产生图 1.2 中混合信号的源信号(原始信号没有给出,不过它们确实与算法找出的信号在本质上是等价的)。而在源分离的问题中,原始信号就是数据集的“独立成分”。

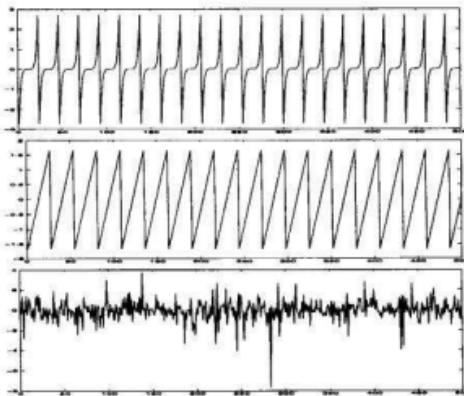


图 1.3 源信号波形的估计,它们是仅用图 1.2 中的观测混合信号估出的,估计结果相当精确

1.3 独立成分分析

1.3.1 定义

前面我们已经看到,盲源分离问题可以归结为寻找一个线性表示,使得该表示对应的成分统计独立。在实际情形下,我们一般不可能找到一个其成分真正独立的表示,但是至少能够找到一个其成分尽可能独立的表示。

这使我们能够对 ICA 进行下述的定义(更为详细的内容将在第 7 章中给出):给定随机变量的一组观测 $(x_1(t), x_2(t), \dots, x_n(t))$,其中 t 是时间或者样本标号,假设它们由独立成分线性混合而产生:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} = \mathbf{A} \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_n(t) \end{bmatrix} \quad (1.6)$$

式中, \mathbf{A} 是某个未知矩阵。在我们仅能观测到 $x_i(t)$ 的情况下, 独立成分分析就要同时估计出矩阵 \mathbf{A} 和 $s_i(t)$ 。注意此处我们还假定独立成分 $s_i(t)$ 的数目与观测变量的数目相同; 其实这只是一个简化假设, 而不是必需的。我们可以将 ICA 定义为另一种形式: 寻找一个类似于公式(1.2)中矩阵 \mathbf{W} 确定的线性变换, 使得随机变量 $y_i, i=1, \dots, n$ 尽可能地独立。这种表述和前一个表述并没有很大差异, 因为如果矩阵 \mathbf{W} 能估计出, 对其求逆就能得到矩阵 \mathbf{A} 。

可以表明(参见 7.5 节), 该问题是适定的, 也就是说公式(1.6)中的模型可估, 当且仅当各成分 s_i 是非高斯的。这是一个基本要求, 它也反映了 ICA 与因子分析之间的主要差别, 后者并没有考虑数据的非高斯性。实际上, ICA 可以认为是一种非高斯的因子分析, 因为在因子分析中, 我们也是将数据建模成某些内在因子的线性混合。

1.3.2 应用

由于 ICA 模型具有一般性, 它在许多领域中都可以应用, 其中一些在第四部分中讨论。下面是一些应用的例子:

- 在脑成像中。大脑内部不同的信号源发出的信号在头部以外的传感器中混合起来, 该过程符合基本盲源分离模型(参见第 22 章)。
- 在计量经济学中。我们可以获得并行的时间序列, ICA 可以将它们分解成独立成分, 从这些成分可以分析和洞察数据集的内在结构(参见 24.1 节)。
- 图像特征提取中有一个稍微不同的应用。这里我们希望找到尽可能独立的特征(参见第 21 章)。

1.3.3 如何寻找独立成分

在只有独立性外没有其他任何假设的情况下, 仍能从线性混合中估计出独立成分, 这个结论可能让人觉得非常吃惊。那么下面我们试图简要地解答上述情况为什么是可能的, 以及如何实现两个基本疑问; 当然, 这正是本书的主题(尤其是第三部分的主题)。

仅仅不相关是不够的:首先要注意的是, 独立性是比不相关性强得多的性质。对于盲源分离问题, 我们实际上可以找到信号的许多不同的不相关表示法, 但这些表示未必独立, 也未必能将源信号分离出来。不相关性就其本身而言是不足以分离这些成分的。这也是主成分分析(PCA)或因子分析不能分离信号的原因: 它们给出的成分除了不相关外就没有更多信息了。

我们可用用一个简单例子来说明该问题。两个独立成分具有均匀分布, 也就是说, 这些成分能够以相同概率取到某个特定区间内的所有值。图 1.4 中画出了来自这两个成分的数据。由于成分间独立, 这些数据均匀分布在一个正方形内。

图 1.5 显示了这两个独立成分的两个不相关混合。虽然两个混合是不相关的, 我们可清楚地看出, 分布并不一样。独立成分是静态混合起来的, 使用的是一个正交混合矩阵, 它相当于平面的一个旋转。也可以看出, 图 1.5 中的成分并不独立: 如果水平轴上的成分在正方形最右边的角附近取得某值, 这显然限制了垂直轴上成分的可能取值。

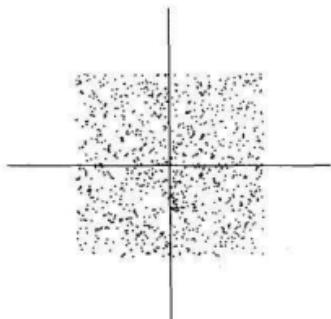


图 1.4 具有均匀分布的独立成分 s_1 和 s_2 的一组样本散布图(水平轴: s_1 , 垂直轴: s_2)

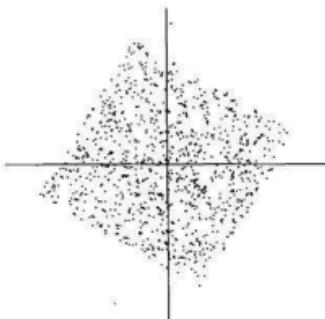


图 1.5 不相关的混合量 x_1 和 x_2
(水平轴: x_1 , 垂直轴: x_2)

事实上,利用我们熟知的去相关方法,可以将独立成分的任何线性混合变成不相关成分,其中的混合变换是正交的(这点将在 7.4.2 节中证明)。这样,ICA 的要点就是估计去相关之后留下的未知正交矩阵。这是经典方法所不能估计的,因为它们和去相关方法一样,是基于协方差信息的。

非线性去相关是基本 ICA 方法:表达独立性如何强于不相关性的一种说法是,独立性本身就蕴含了非线性不相关性:若 s_1 和 s_2 独立,那么任何非线性变换 $g(s_1)$ 和 $h(s_2)$ 都是不相关的(在它们之间协方差为零的意义下)。与此形成鲜明对比的是,对于两个仅仅不相关的随机变量,这样两个非线性变换一般不再具有零协方差。这样,我们可以通过一种更强形式的去相关运算来实现 ICA:即,寻找一个表示,使得 y_i 即使经过非线性变换仍然不相关。这给出了估计矩阵 W 的一个简单原理:

ICA 估计原理 1:非线性去相关。寻找矩阵 W ,使得对任何 $i \neq j$,成分 y_i 和 y_j 不相关,而且变换后的成分 $g(y_i)$ 和 $h(y_j)$ 也不相关,其中, g 和 h 是某些适当的非线性函数。

这是估计 ICA 的一个有效方法:如果非线性函数选得适当,此方法的确能找到独立成分。事实上,如果我们计算图 1.5 中两个混合量的非线性相关矩阵,就能立即发现它们并不是相互独立的。

虽然这个原理非常直观,但它却遗留了一个重要的问题:非线性函数 g 和 h 该如何选择?该问题的答案可以从估计理论和信息论中找到。估计理论提供了可以估计任何一个统计模型的最为经典的方法:极大似然估计法(参见第 9 章)。信息论可以给出独立性的一些准确度量,如互信息(参见第 10 章)。利用其中任何一种理论,我们都能确定出满意的非线性函数 g 和 h 。

独立成分是极大非高斯性成分:ICA 估计另一个非常直观和重要的原则是极大非高斯性(参见第 8 章)。其思路是,根据中心极限定理,非高斯随机变量之和比原变量更接近高斯变量。因此,如果我们取观测混合变量的一个线性组合 $y = \sum_i b_i x_i$ (因为混合模型是线性的,因此该线性组合同时也是独立成分的一种线性组合),如果它等于独立成分之一,那么它的非高斯性达到极大。这是因为,如果它确实是两个或更多成分的混合,按中心极限定理,该混合将更接近于高斯分布。

这样,相关的原理可以表述如下:

ICA 估计原理 2: 极大非高斯性。在 y 的方差为常数的约束下,求线性组合 $y = \sum_i b_i x_i$ 非高斯性的局部极大值。每个局部极大给出一个独立成分。

为了在实际应用中度量非高斯性,我们可以使用一些量,比方说峭度。峭度是一个高阶累积量,它是方差的某种推广(利用高阶多项式)。累积量具有一些有趣的代数和统计性质,这也是它们在 ICA 理论中起着重要作用的原因。

举例来说,比较图 1.4 和图 1.5 中各坐标轴方向对应成分的非高斯性,我们可以看到图 1.5 中的非高斯性更小,因而图 1.5 的坐标轴方向给出的成分不可能是独立成分(参见第 8 章)。

有意思的一点是,极大非高斯性原理表明了 ICA 和独立发展技术,称为投影寻踪的这项技术之间的紧密关系。在投影寻踪方法里,我们实际上也是寻找具有极大非高斯性的线性组合,并用于可视化或其他目的。这样,独立成分可以解释成投影寻踪的方向。

当 ICA 用于提取特征时,极大非高斯性原理也表明了它与特征提取的神经科学理论中使用过的稀疏编码之间的紧密关系(参见第 21 章)。稀疏编码的思路是将数据用成分表示,使得只有很少数量的成分是同时“激活”的。在某些情形下,这等价于寻找极大非高斯性成分。

ICA 与投影寻踪和稀疏编码的这些联系,都和一个更为深入的结果有关,该结果表明,ICA 给出了一个尽可能结构化的线性表示。此论断可以用信息论概念给出其严格的意义(参见第 10 章),并且相关结果还表明,独立成分在许多方面比原始随机变量更容易处理。特别地,独立成分比原始变量更容易编码(压缩)。

ICA 估计所需信息要比协方差更多: 还有很多其他方法也可以估计 ICA 模型,其中很多都将在本书中讨论。这些方法的共同点是,它们考虑了没有包含在协方差矩阵中的某些统计量(协方差矩阵包含的是所有 x_i 对之间的协方差)。

利用协方差矩阵,我们可以在通常线性意义下去除各成分间的相关性,但仅此而已。故所有 ICA 方法都用到了某种形式的高阶统计量,高阶特别地意味着这些信息并未包含在协方差矩阵中。到此为止我们已经遇到了两类高阶信息:非线性相关性和峭度。也可以使用其他很多类型的高阶统计量。

数值方法是重要的: 除了估计原理外,我们还必须找到一个具体算法,以实现所需的计算。由于估计原理使用的是非二次的函数,所需的计算通常不能用简单的线性代数来表达,因此算法方面的要求可能是很高的。这样,数值算法就成为 ICA 估计方法一个不可缺少的组成部分。

数值方法通常是基于对某种目标函数的优化。基本的优化方法是梯度法,而特别有意思的是一个称为 FastICA 的不动点算法,它似乎是特别为 ICA 问题量身定制的,可以充分地挖掘问题的特殊结构。我们可以采用两种方法的任意一个找到用峭度绝对值所度量的非高斯性极大解。

1.4 ICA 的历史

J. Hérault, C. Jutten 和 B. Ans [178, 179, 16] 在 20 世纪 80 年代早期就已经引入了 ICA 技术,虽然那时还没有采用 ICA 这个名称。就像 Jutten[227]最近综述的那样,问题是 1982 年首先在一个神经生理学的背景下提出的。在肌肉收缩运动编码的一个简化模型中,输出 $x_1(t)$ 和 $x_2(t)$ 是度量肌肉收缩的两类敏感信号,而 $s_1(t)$ 和 $s_2(t)$ 是运动点的角位置和角速度。那么,这

些信号之间 ICA 模型成立的假设并非毫无道理。神经系统以某种方式通过测量响应 $x_1(t)$, $x_2(t)$ 而推断出位置和速度信号 $s_1(t), s_2(t)$ 。一种可能的方法是采用简单的神经网络,并利用非线性去相关原理,学习其逆模型。Hérault 和 Jutten 提出了一个特别的反馈电路,用于解决该问题。该方法包含在第 12 章的内容中。

在整个 20 世纪 80 年代中,对 ICA 最熟悉的主要还是法国的研究者,但其国际影响有限。在 20 世纪 80 年代中期,国际神经网络会议有少量的 ICA 介绍,但被那时到处泛滥传播的、对 BP 网络、Hopfield 网络,以及 Kohonen 的自组织映射(SOM)方面的兴趣所掩盖。另一个相关的领域是高阶谱分析,它的第一届国际研讨会于 1989 年召开。在这次研讨会上,出现了 J.-F. Cardoso [60] 和 P. Comon [88] 在 ICA 方面的早期文章。J.-F. Cardoso 使用代数方法,特别是高阶累积量,最终形成了 JADE 算法[72]。四阶累积量在更早以前已由 J.-L. Lacoume [254] 提出。在信号处理领域,参考文献[228, 93, 408, 89] 是法国学派的经典文章。参考文献[227] 对历史发展有很好的解说,其中还附有一份更完备的参考文献目录。

信号处理领域中,在与 ICA 相关的盲信号去卷积问题[114, 398] 上很早以前就有了一些进展。特别需要提及的是,多通道盲去卷积中使用的结果与 ICA 技术非常接近。

20 世纪 80 年代科学家的工作被一些研究者进一步扩展,特别是 A. Cichocki 和 R. Unbehauen,他们首先提出了目前最流行的 ICA 算法中的一个[82, 85, 84]。20 世纪 90 年代早期,ICA 和信号分离方面的其他文章有参考文献[57, 314]。“非线性 PCA”方法由本书作者引入[332, 232]。然而,直到 20 世纪 90 年代中期,ICA 研究仍然只是一个小且范围狭窄的研究领域。这个时期提出了几个可用的算法,不过通常只能适于有限的一些问题。不过这些方法与统计优化指标的严格联系,直到后来才逐渐被揭示出来。

20 世纪 90 年代中期,A.J. Bell 和 T.J. Sejnowski 发表了他们基于信息极大原理的方法[35, 36]后,ICA 吸引了更为广泛的关注,人们对它的兴趣也不断增长。该算法被 S.-I. Amari 及其合作者们用自然梯度法[12]进一步实现了细化,同时建立了它和极大似然估计以及 Cichocki-Unbehauen 算法之间的基本联系。几年后,本书作者提出了不动点或 FastICA 算法[210, 192, 197],由于其计算效率,对 ICA 在大规模问题上的应用做出了贡献。

自 20 世纪 90 年代中期以来,已经出现了大量致力于 ICA 的文章、研讨会和专题分组会议。ICA 的第一届国际研讨会 1999 年 1 月在法国的 Aussois 召开,第二届研讨会 2000 年在芬兰赫尔辛基(Helsinki)召开。两次会议都吸引了上百位 ICA 和盲源分离方面的研究者,他们的贡献使得 ICA 已经成为一个被人们承认并业已成熟的研究领域。

第一部分

数学预备知识

- 第 2 章 随机向量和独立性
- 第 3 章 梯度和最优化方法
- 第 4 章 估计理论
- 第 5 章 信息论
- 第 6 章 主成分分析和白化

第2章 随机向量和独立性

在本章中,我们回顾概率论、统计学和随机过程中的核心概念,而重点放在多元统计学和随机向量上。本书后面章节将要用到的概念,包括统计独立性和高阶统计量等,会讨论得更详细一些。假定读者已具备了单变量概率论的基本知识,从而已经熟悉了诸如概率、基本事件和随机变量等定义。那些已经掌握了多元统计学的读者可以跳过本章的大部分内容。对于那些想对高级内容有更广泛的了解,或者希望得到更多信息的读者,有许多优秀的教科书可以选读,程度从初级到高级不等。参考文献[353]就是一本广为使用的教材,内容覆盖了概率论、随机变量和随机过程。

2.1 概率分布和概率密度

2.1.1 随机变量的定义

本书中如无特别声明,我们假设随机变量是取连续值的。随机变量 x 在点 $x = x_0$ 处的累积分布函数(cdf) F_x 定义为 $x \leq x_0$ 的概率为:

$$F_x(x_0) = P(x \leq x_0) \quad (2.1)$$

令 x_0 从 $-\infty$ 变到 $+\infty$,就定义了全部的 cdf。

很明显,对连续随机变量,其 cdf 是非负且非减(通常还单调增加)的连续函数,且取值落在区间 $0 \leq F_x(x_0) \leq 1$ 中。由定义立即可知 $F_x(-\infty) = 0$ 且 $F_x(+\infty) = 1$ 。

我们通常用密度函数而不是它的 cdf 来刻画一个概率分布。连续随机变量 x 的概率密度函数(pdf) $p_x(x)$ 可以从形式上由它的累积分布函数求导得到:

$$p_x(x_0) = \frac{dF_x(x)}{dx} \Big|_{x=x_0} \quad (2.2)$$

实际上,根据导数和积分的互逆关系,从已知的 pdf 通过计算也可得到其 cdf:

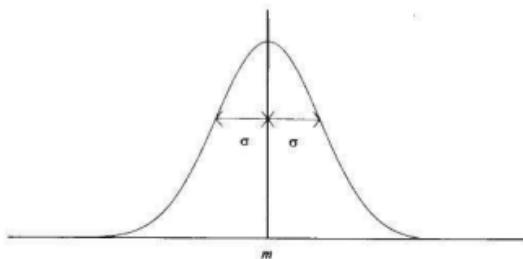
$$F_x(x_0) = \int_{-\infty}^{x_0} p_x(\xi) d\xi \quad (2.3)$$

为简单起见,常常把 $F_x(x)$ 记为 $F(x)$,相应地,把 $p_x(x)$ 记为 $p(x)$ 。当可能产生混乱时,用下标指代对应的随机变量。

例 2.1 高斯(或正态)概率分布在大量模型和应用中使用,比如用于描述加性噪声。其密度函数为:

$$p_x(x) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \quad (2.4)$$

式中,参数 m (均值)决定了其对称密度函数的峰值点,而 σ (标准差)是它的有效宽度(峰的平坦或尖锐程度)。参见图 2.1 中的例子。

图 2.1 一个高斯概率密度函数, 其均值为 m , 标准差为 σ

一般来说, 不能由公式(2.3)算得高斯密度一个封闭形式的 cdf。密度公式(2.4)中前面的项 $1/\sqrt{2\pi\sigma^2}$ 是一个归一化因子, 它保证当 $x_0 \rightarrow \infty$ 时该 cdf 值为 1。不过, 该 cdf 值可以用数值方法计算, 如用下述误差函数的查表值得到:

$$\text{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \exp\left(-\frac{\xi^2}{2}\right) d\xi \quad (2.5)$$

该误差函数与归一化的高斯密度的 cdf 密切相关: 对于归一化的高斯密度, 其均值 $m = 0$, 方差 $\sigma^2 = 1$ 。具体细节请参见参考文献[353]。

2.1.2 随机向量的分布

假设 \mathbf{x} 是一个 n 维随机向量:

$$\mathbf{x} = [x_1, x_2, \dots, x_n]^T \quad (2.6)$$

式中, T 表示转置(此处取转置, 是因为本书中的向量都是列向量)。列向量 \mathbf{x} 的分量 x_1, x_2, \dots, x_n 是连续随机变量。概率分布的概念很容易推广到这种随机向量上。特别地, \mathbf{x} 的累积分布函数定义为:

$$F_{\mathbf{x}}(\mathbf{x}_0) = P(\mathbf{x} \leq \mathbf{x}_0) \quad (2.7)$$

式中, $P(\cdot)$ 仍指括号中事件的概率, 而 \mathbf{x}_0 是随机向量 \mathbf{x} 的某个常值向量。记号 $\mathbf{x} \leq \mathbf{x}_0$ 意思是, 向量 \mathbf{x} 的每个分量小于或等于向量 \mathbf{x}_0 的相应分量。公式(2.7)中的多元 cdf 具有和单个随机变量的 cdf 类似的性质, 即关于每个分量它是非减函数, 取值于区间 $0 \leq F_{\mathbf{x}}(\mathbf{x}_0) \leq 1$ 。当 \mathbf{x}_0 的所有分量趋于无穷时, $F_{\mathbf{x}}(\mathbf{x}_0)$ 取得其上限 1; 当 \mathbf{x}_0 的任一分量 $x_{0,j} \rightarrow -\infty$ 时, $F_{\mathbf{x}}(\mathbf{x}_0) \rightarrow 0$ 。

\mathbf{x} 的多元概率密度函数 $p_{\mathbf{x}}(\mathbf{x}_0)$ 定义为累积分布函数 $F_{\mathbf{x}}(\mathbf{x}_0)$ 关于自变量 \mathbf{x}_0 的所有分量的混合偏导数^①:

$$p_{\mathbf{x}}(\mathbf{x}_0) = \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \cdots \frac{\partial}{\partial x_n} F_{\mathbf{x}}(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{x}_0} \quad (2.8)$$

① 严格地说, 下面使用的 $F_{\mathbf{x}}(\mathbf{x})$ 和 $p_{\mathbf{x}}(\mathbf{x})$ 等记号中, 下标和自变量使用同一个符号 \mathbf{x} 是不合法的, 原因有两点: 首先, 下标 \mathbf{x} 是随机向量, 而自变量则是确定性的向量; 其次, 无论对 F 的微分运算, 还是对 p 的积分运算, 都只是关于括号中的自变量进行的, 与下标变量无关。但这种不严谨做法的好处是能节省一半的符号开销, 而且比较直观。只要读者心中有数, 就不致产生混淆, 所以译者在符号上基本未做变动。原著中有些地方也注意到了这一点, 比如, 第 5 章就使用了较规范的记号——译者注。

因而：

$$F_x(x_0) = \int_{-\infty}^{x_0} p_x(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{x_{0,1}} \int_{-\infty}^{x_{0,2}} \cdots \int_{-\infty}^{x_{0,n}} p_x(\mathbf{x}) dx_n \dots dx_2 dx_1 \quad (2.9)$$

式中， $x_{0,i}$ 是 \mathbf{x}_0 的第 i 个分量。显然：

$$\int_{-\infty}^{+\infty} p_x(\mathbf{x}) d\mathbf{x} = 1 \quad (2.10)$$

该式给出了一个真正的多元概率密度函数 $p_x(\mathbf{x}_0)$ 必须满足的归一化条件。

在许多情况下，随机变量的概率密度函数仅在某些特定区间上具有非零值。下面是此种情形的一个例子。

例 2.2 假设二维随机向量 $\mathbf{z} = (x, y)^T$ 的概率密度函数为：

$$p_z(\mathbf{z}) = p_{x,y}(x, y) = \begin{cases} \frac{3}{7}(2-x)(x+y), & x \in [0, 2], y \in [0, 1] \\ 0, & \text{其他} \end{cases}$$

我们来计算 \mathbf{z} 的累积分布函数。通过对 x 和 y 的积分可以算出，积分时要考虑密度非零区域的界限。当 $x \leq 0$ 或 $y \leq 0$ 时，密度 $p_z(z_0)$ 以及 cdf 都是零。在 $0 < x \leq 2$ 且 $0 < y \leq 1$ 的区域中，其 cdf 由下式给出：

$$\begin{aligned} F_z(\mathbf{z}) = F_{x,y}(x, y) &= \int_0^y \int_0^x \frac{3}{7}(2-\xi)(\xi+\eta) d\xi d\eta \\ &= \frac{3}{7}xy \left(x + y - \frac{1}{3}x^2 - \frac{1}{4}xy \right) \end{aligned}$$

在 $0 < x \leq 2$ 且 $y > 2$ 的区域中，对 y 的积分上界等于 1，相应的 cdf 可在前一表达式中令 $y = 1$ 得到。类似地，在 $x > 2$ 且 $0 < y \leq 1$ 的区域中，相应的 cdf 可在前一表达式中令 $x = 2$ 得到。最后，若 $x > 2$ 且 $y > 1$ ，cdf 变成 1，表明该概率密度函数密度 $p_z(z)$ 已经正确地归一化。将这些结果集中起来，可以得到：

$$F_z(\mathbf{z}) = \begin{cases} 0, & x \leq 0 \text{ 或 } y \leq 0 \\ \frac{3}{7}xy \left(x + y - \frac{1}{3}x^2 - \frac{1}{4}xy \right), & 0 < x \leq 2, 0 < y \leq 1 \\ \frac{3}{7}x \left(1 + \frac{3}{4}x - \frac{1}{3}x^2 \right), & 0 < x \leq 2, y > 1 \\ \frac{6}{7}y \left(\frac{2}{3} + \frac{1}{2}y \right), & x > 2, 0 < y \leq 1 \\ 1, & x > 2 \text{ 和 } y > 1 \end{cases}$$

2.1.3 联合与边缘分布

两个随机向量的联合分布可以用类似的方式处理。特别地，令 \mathbf{y} 是另一个随机向量，其维数 m 一般不同于 \mathbf{x} 的维数 n 。向量 \mathbf{x} 和 \mathbf{y} 可以连起来形成一个“超向量” $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$ ，这样前面的公式就可以直接使用。由此产生的 cdf 称为 \mathbf{x} 和 \mathbf{y} 的联合分布函数，由下式给出：

$$F_{x,y}(\mathbf{x}_0, \mathbf{y}_0) = P(\mathbf{x} \leq \mathbf{x}_0, \mathbf{y} \leq \mathbf{y}_0) \quad (2.11)$$

式中， \mathbf{x}_0 和 \mathbf{y}_0 是分别与 \mathbf{x} 和 \mathbf{y} 具有同样维数的常值向量。公式(2.11)定义了事件 $\mathbf{x} \leq \mathbf{x}_0$ 和 $\mathbf{y} \leq \mathbf{y}_0$ 的联合概率。

\mathbf{x} 和 \mathbf{y} 的联合密度函数 $p_{x,y}(\mathbf{x}_0, \mathbf{y}_0)$ 在形式上仍然可以定义为联合分布函数 $F_{x,y}(\mathbf{x}_0, \mathbf{y}_0)$ 关于所有分量的混合偏导数。因而，下面的关系式成立：

$$F_{x,y}(x_0, y_0) = \int_{-\infty}^{x_0} \int_{-\infty}^{y_0} p_{x,y}(\xi, \eta) d\eta d\xi \quad (2.12)$$

而且当 $x_0 \rightarrow \infty$ 且 $y_0 \rightarrow \infty$ 时, 该积分值趋近于 1。

x 和 y 的边缘密度 $p_x(x_0)$ 与 $p_y(y_0)$ 可以通过在联合密度 $p_{x,y}(x_0, y_0)$ 中对其中另一个向量进行积分得到:

$$p_x(x) = \int_{-\infty}^{\infty} p_{x,y}(x, \eta) d\eta \quad (2.13)$$

$$p_y(y) = \int_{-\infty}^{\infty} p_{x,y}(\xi, y) d\xi \quad (2.14)$$

例 2.3 考虑例 2.2 中给出的联合密度。随机变量 x 和 y 的边缘密度是:

$$\begin{aligned} p_x(x) &= \int_0^1 \frac{3}{7}(2-x)(x+y) dy, \quad x \in [0, 2] \\ &= \begin{cases} \frac{3}{7}(1 + \frac{3}{2}x - x^2) & x \in [0, 2] \\ 0 & \text{其他} \end{cases} \end{aligned}$$

$$\begin{aligned} p_y(y) &= \int_0^2 \frac{3}{7}(2-x)(x+y) dx, \quad y \in [0, 1] \\ &= \begin{cases} \frac{3}{7}(2+3y), & y \in [0, 1] \\ 0, & \text{其他} \end{cases} \end{aligned}$$

2.2 期望和矩

2.2.1 定义和一般性质

实际中的向量或标量值随机变量, 其准确概率密度函数通常是未知的。不过, 我们可以取该随机变量的某个函数的期望, 用它们来实现有用的分析和处理。期望的一个很好的特点是: 虽然它在形式上是通过密度函数定义的, 但可以直接用数据估计得到。

令 $\mathbf{g}(\mathbf{x})$ 是从随机向量 \mathbf{x} 导出的任一个量。 $\mathbf{g}(\mathbf{x})$ 可以是标量、向量, 甚至是一个矩阵。 $\mathbf{g}(\mathbf{x})$ 的期望记为 $E[\mathbf{g}(\mathbf{x})]$, 其定义为:

$$E[\mathbf{g}(\mathbf{x})] = \int_{-\infty}^{\infty} \mathbf{g}(\mathbf{x}) p_x(\mathbf{x}) d\mathbf{x} \quad (2.15)$$

式中, 积分符号要遍及 \mathbf{x} 的所有分量。积分运算对该向量或矩阵的每一个元素分别进行, 产生另一个同样大小的向量或矩阵。若 $\mathbf{g}(\mathbf{x}) = \mathbf{x}$, 则得到 \mathbf{x} 的期望 $E[\mathbf{x}]$; 这将在下一节中更详细地讨论。

期望具有一些重要的基本性质:

1. 线性性质: 设 \mathbf{x}_i ($i = 1, \dots, m$) 是一组不同的随机向量, 而 a_i ($i = 1, \dots, m$) 是某些非随机的标量系数。那么:

$$E\left[\sum_{i=1}^m a_i \mathbf{x}_i\right] = \sum_{i=1}^m a_i E\{\mathbf{x}_i\} \quad (2.16)$$

2. 线性变换性质: 设 x 是一个 m 维的随机向量, 而 A 和 B 分别是 $k \times m$ 和 $m \times l$ 维的非随机矩阵。那么:

$$E\{Ax\} = AE\{x\}, \quad E\{xB\} = E\{x\}B \quad (2.17)$$

3. 变换不变性: 设 $y = g(x)$ 是随机向量 x 的某个向量值函数。那么:

$$\int_{-\infty}^{\infty} y p_y(y) dy = \int_{-\infty}^{\infty} g(x) p_x(x) dx \quad (2.18)$$

于是有 $E\{y\} = E\{g(x)\}$, 虽然等式两边是对不同概率密度函数的积分。

这些性质可以利用期望算子的定义和概率密度函数的性质来证明。它们在实际应用中很重要: 能帮助我们将一个包含期望的表达式进行简化, 而无须实际计算任何积分(可能除了最后一步外)。

2.2.2 均值向量和相关矩阵

矩是用于刻画随机向量 x 的一种典型的期望。当 $g(x)$ 由 x 分量的乘积构成时, 就得到了矩。特别地, 随机向量 x 的一阶矩称为 x 的均值向量 μ_x , 它定义为 x 的期望:

$$\mu_x = E\{x\} = \int_{-\infty}^{\infty} x p_x(x) dx \quad (2.19)$$

μ_x 的各个分量 μ_{x_i} 由下式给出:

$$\mu_{x_i} = E\{x_i\} = \int_{-\infty}^{\infty} x_i p_x(x) dx = \int_{-\infty}^{\infty} x_i p_{x_i}(x_i) dx_i \quad (2.20)$$

式中, $p_{x_i}(x_i)$ 是 x 的第 i 个分量 x_i 的边缘密度。上式成立是因为: 根据边缘密度的定义, 对 x 的其他所有分量积分时就得到 1。

另一组重要的矩阵由 x 的各分量对之间的相关构成。 x 的第 i 分量和第 j 分量之间的相关 r_{ij} 由如下二阶矩给出:

$$r_{ij} = E\{x_i x_j\} = \int_{-\infty}^{\infty} x_i x_j p_x(x) dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_i x_j p_{x_i, x_j}(x_i, x_j) dx_j dx_i \quad (2.21)$$

注意, 相关可正可负。

向量 x 的 $n \times n$ 相关矩阵:

$$R_x = E\{xx^T\} \quad (2.22)$$

以一种方便的形式表达了它所有分量之间的相关, 其中 r_{ij} 是 R_x 第 i 行第 j 列的元素。

相关矩阵具有一些重要性质:

1. 它是一个对称矩阵: $R_x = R_x^T$ 。
2. 它是一个半正定矩阵: 即对所有 n 维向量 a , 有:

$$a^T R_x a \geq 0 \quad (2.23)$$

实际上, R_x 常常是正定的, 即对任何向量 $a \neq 0$, 不等式(2.23)严格成立。

3. R_x 的所有特征值是非负实数(当 R_x 为正定矩阵时, 特征值为正)。进一步, R_x 的所有特征向量都是实值的, 并且总可以选成使它们标准正交。

更高阶的矩也可以用类似的方式定义,但对它们的讨论将推后到2.7节。下面我们首先考虑两个不同随机向量的中心矩和二阶矩。

2.2.3 协方差和联合矩

中心矩的定义方式与其他矩相似,只是在计算期望之前减去了相应随机向量的均值向量。很明显,中心矩只有在高于一阶的情况下才有意义。与相关矩阵 \mathbf{R}_x 相应的量称为 x 的协方差矩阵 \mathbf{C}_x ,由下式给出:

$$\mathbf{C}_x = E\{(x - m_x)(x - m_x)^T\} \quad (2.24)$$

$n \times n$ 维矩阵 \mathbf{C}_x 中的元素:

$$c_{ij} = E\{(x_i - m_i)(x_j - m_j)\} \quad (2.25)$$

称为协方差,它们是与公式(2.21)中定义的相关^① r_{ij} 对应的中心矩。

协方差矩阵 \mathbf{C}_x 具有与相关矩阵 \mathbf{R}_x 相同的一些特性。利用期望算子的性质,容易看出:

$$\mathbf{R}_x = \mathbf{C}_x + m_x m_x^T \quad (2.26)$$

若均值向量 $m_x = 0$,那么相关矩阵和协方差矩阵是一样的。如有必要,可以从数据向量中减去均值向量(的估计),经过这一个预处理步骤后,数据就是零均值的了。这是独立成分分析中常用的做法,在后面各章中,我们将相关/协方差矩阵简记为 \mathbf{C}_x ,为简单起见,甚至经常不带下标 x 。

对于单个随机变量 x ,均值向量归结为其平均值 $m_x = E[x]$;相关矩阵则归结为其二阶矩 $E[x^2]$;而协方差矩阵归结为 x 的方差:

$$\sigma_x^2 = E\{(x - m_x)^2\} \quad (2.27)$$

这时,关系式(2.26)变为如下的简单形式: $E[x^2] = \sigma_x^2 + m_x^2$ 。

通过联合密度,可将期望运算扩展到两个不同随机向量 x 和 y 的函数 $\mathbf{g}(x, y)$ 上去:

$$E\{\mathbf{g}(x, y)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbf{g}(x, y) p_{x,y}(x, y) dy dx \quad (2.28)$$

此公式中的积分计算要遍历 x 和 y 的所有分量。

在联合期望中应用最广的是互相关矩阵:

$$\mathbf{R}_{xy} = E\{xy^T\} \quad (2.29)$$

以及互协方差矩阵:

$$\mathbf{C}_{xy} = E\{(x - m_x)(y - m_y)^T\} \quad (2.30)$$

注意, x 和 y 的维数可以是不一样的,因而互相关矩阵和互协方差矩阵未必一定是方阵,而且它们一般不对称。不过定义容易得到:

$$\mathbf{R}_{xy} = \mathbf{R}_{yx}^T, \quad \mathbf{C}_{xy} = \mathbf{C}_{yx}^T \quad (2.31)$$

^① 经典统计学中用的是相关系数 $p_{ij} = \frac{c_{ij}}{(c_{ii}c_{jj})^{1/2}}$,由它们组成的矩阵称为相关矩阵。在本书中,相关矩阵由公式(2.22)定义,在信号处理、神经网络和工程应用中通常采用这种形式的定义。

若 x 和 y 的均值向量为零,那么它们的互相关矩阵和互协方差矩阵相同。实际应用中经常需要得到具有相同维数的两个随机向量 x 和 y 之和的协方差矩阵 C_{x+y} ,容易看出:

$$C_{x+y} = C_x + C_y + C_{yx} + C_y \quad (2.32)$$

相关和协方差是用二阶统计量来度量随机变量之间的依赖性。这可从如下例子中看出。

例 2.4 考虑零均值随机变量 x 和 y 的两个不同的联合分布 $p_{x,y}(x,y)$,如图 2.2 和图 2.3 所示。图 2.2 中, x 和 y 明显地具有负协方差(或相关), x 取正值通常隐含着 y 取负值,反之亦然。而在图 2.3 中,却不可能通过观测 x 对 y 的值做出什么推断,因而,它们的协方差 $c_{xy} \approx 0$ 。

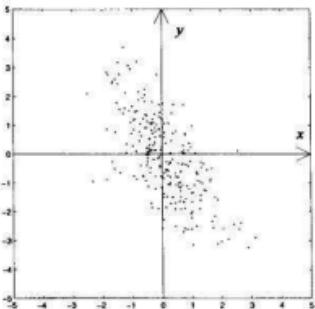


图 2.2 随机变量 x 和 y 具有
负协方差的一个例子

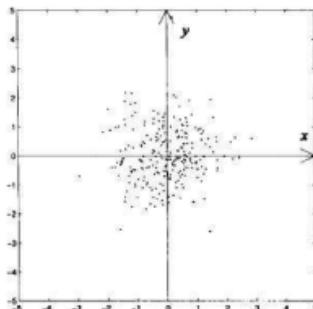


图 2.3 随机变量 x 和 y 具有
零协方差的一个例子

2.2.4 期望的估计

随机向量 x 的概率密度一般是未知的,但是,通常可以从 x 得到一组(K 个)样本 x_1, x_2, \dots, x_K 。利用这些样本,通过对样本取平均,用如下公式将期望(2.15)估计出来[419]:

$$E\{\mathbf{g}(\mathbf{x})\} \approx \frac{1}{K} \sum_{j=1}^K \mathbf{g}(\mathbf{x}_j) \quad (2.33)$$

举例来说,运用公式(2.33),我们可以得到 x 的均值向量 \mathbf{m}_x 的标准估计,即样本平均为:

$$\hat{\mathbf{m}}_x = \frac{1}{K} \sum_{j=1}^K \mathbf{x}_j \quad (2.34)$$

式中, $\hat{\mathbf{m}}$ 上的符号[^],是一个量的估计的标准记号。

类似地,如果不知道随机向量 x 和 y 的联合密度 $p_{x,y}(x,y)$,但是我们有 K 对样本 $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_K, \mathbf{y}_K)$,则可用下面的式子估计期望[参见式(2.28)]:

$$E\{\mathbf{g}(\mathbf{x}, \mathbf{y})\} \approx \frac{1}{K} \sum_{j=1}^K \mathbf{g}(\mathbf{x}_j, \mathbf{y}_j) \quad (2.35)$$

例如,对于互相关矩阵,可以得到其估计的公式:

$$\hat{\mathbf{R}}_{xy} = \frac{1}{K} \sum_{j=1}^K \mathbf{x}_j \mathbf{y}_j^\top \quad (2.36)$$

对于其他类型的相关矩阵 R_{xx} , C_{xx} 或 C_{xy} , 也可以立即可以得出类似的公式。

2.3 不相关性和独立性

2.3.1 不相关性和白化

我们称两个随机向量 x 和 y 不相关, 若它们的互协方差矩阵 C_{xy} 是零矩阵:

$$C_{xy} = E\{(x - m_x)(y - m_y)^\top\} = 0 \quad (2.37)$$

这等价于如下条件:

$$R_{xy} = E\{xy^\top\} = E\{x\}E\{y^\top\} = m_x m_y^\top \quad (2.38)$$

在两个不同的标量型随机变量 x 和 y 的特殊情形下(例如, 作为某随机向量 z 的两个分量), 若它们的协方差 c_{xy} 等于零, 则 x 和 y 是不相关的:

$$c_{xy} = E\{(x - m_x)(y - m_y)\} = 0 \quad (2.39)$$

或等价地:

$$r_{xy} = E\{xy\} = E\{x\}E\{y\} = m_x m_y \quad (2.40)$$

可以再次看到, 在零均值变量情形下, 零协方差等价于零相关。

另一种重要的特例涉及到单个随机向量 x 各分量之间的相关性, 这可由公式(2.24)定义的协方差矩阵 C_x 给出。此种情况下, 不可能存在等价于公式(2.37)的条件, 原因是 x 的每个分量肯定与自身完全相关。我们所能遇到的最极端的特例也只能是 x 的分量之间两两互不相关, 这种情况对应下面的不相关性条件:

$$C_x = E\{(x - m_x)(x - m_x)^\top\} = D \quad (2.41)$$

这里矩阵 D 是 $n \times n$ 的对角矩阵:

$$D = \text{diag}(c_{11}, c_{22}, \dots, c_{nn}) = \text{diag}(\sigma_{x_1}^2, \sigma_{x_2}^2, \dots, \sigma_{x_n}^2) \quad (2.42)$$

而 D 的 n 个对角元素则是 x 的各分量 x_i 的方差。

特别地, 如果一个随机向量具有零均值和单位协方差(从而也具有单位相关)矩阵(或者也可以是单位协方差再乘上一个常值方差 σ^2), 称该随机向量是白化的(white, 或白的)。因此, 白化随机向量满足如下条件:

$$m_x = 0, \quad R_x = C_x = I \quad (2.43)$$

其中 I 是 $n \times n$ 的单位矩阵。

现在我们假设将一个 $n \times n$ 的矩阵 T 所定义的正交变换作用于随机向量 x 。在数学上可将它表示成:

$$y = Tx, \text{ 其中 } T^\top T = TT^\top = I \quad (2.44)$$

正交矩阵 T 在 n 维空间中定义了一个旋转(改变坐标轴), 该旋转保持范数和距离。假设 x 是白化的, 我们可以得到:

$$m_y = E\{Ty\} = TE\{x\} = Tm_x = 0 \quad (2.45)$$

以及

$$\begin{aligned} \mathbf{C}_x &= \mathbf{R}_x = E\{\mathbf{T}\mathbf{x}(\mathbf{T}\mathbf{x})^T\} = \mathbf{T}E\{\mathbf{x}\mathbf{x}^T\}\mathbf{T}^T \\ &= \mathbf{T}\mathbf{R}_s\mathbf{T}^T = \mathbf{T}\mathbf{T}^T = \mathbf{I} \end{aligned} \quad (2.46)$$

这就表明了,旋转后的向量 y 也是白化的。因此我们可以得出结论:正交变换下,白化的性质保持不变。事实上,原始数据的白化可以按无数种方式进行。第6章将会更加详细地讨论白化,因为它非常有用。而且在独立成分分析中,白化也是广泛采用的一个预处理步骤。

很清楚,也同样存在无穷多种方式将原始数据进行去相关,因为白化性质是不相关性质的特例。

例 2.5 考虑线性信号模型:

$$\mathbf{x} = \mathbf{As} + \mathbf{n} \quad (2.47)$$

式中, x 是 n 维随机或数据向量, A 是 $n \times m$ 的常值矩阵, s 是 m 维随机信号向量, n 是 n 维随机向量,通常描述加性噪声。这样, x 的相关矩阵变成:

$$\begin{aligned} \mathbf{R}_x &= E\{\mathbf{x}\mathbf{x}^T\} = E\{(\mathbf{As} + \mathbf{n})(\mathbf{As} + \mathbf{n})^T\} \\ &= E\{\mathbf{A}\mathbf{s}\mathbf{s}^T\mathbf{A}^T\} + E\{\mathbf{A}\mathbf{s}\mathbf{n}^T\} + E\{\mathbf{n}\mathbf{s}^T\mathbf{A}^T\} + E\{\mathbf{n}\mathbf{n}^T\} \\ &= A E\{\mathbf{s}\mathbf{s}^T\} A^T + A E\{\mathbf{s}\mathbf{n}^T\} + E\{\mathbf{n}\mathbf{s}^T\} A^T + E\{\mathbf{n}\mathbf{n}^T\} \\ &= \mathbf{AR}_s\mathbf{A}^T + \mathbf{AR}_{sn} + \mathbf{R}_{ns}\mathbf{A}^T + \mathbf{R}_n \end{aligned} \quad (2.48)$$

通常,噪声向量 n 假设具有零均值,且与信号向量不相关。这样,信号和噪声向量之间的互相关矩阵等于零:

$$\mathbf{R}_{ns} = E\{\mathbf{s}\mathbf{n}^T\} = E\{\mathbf{s}\}E\{\mathbf{n}^T\} = 0 \quad (2.49)$$

类似地, $\mathbf{R}_{nn} = 0$,因此, x 的相关矩阵可简化为:

$$\mathbf{R}_x = \mathbf{AR}_s\mathbf{A}^T + \mathbf{R}_s \quad (2.50)$$

另一个经常做出的假设是:假定噪声是白化的。这意味着噪声向量 n 的各分量互不相关,并且具有相等的方差 σ^2 ,从而在公式(2.50)中有:

$$\mathbf{R}_s = \sigma^2 \mathbf{I} \quad (2.51)$$

有时候,比如说在 ICA 模型的某种含噪声版本(参见第 15 章)中,信号向量 s 的分量也是互不相关的,从而信号的相关矩阵也是对角矩阵:

$$\mathbf{D}_s = \text{diag}(E\{s_1^2\}, E\{s_2^2\}, \dots, E\{s_m^2\}) \quad (2.52)$$

式中, s_1, s_2, \dots, s_m 是信号向量 s 的分量。这时,公式(2.50)可以写成如下形式:

$$\mathbf{R}_x = \mathbf{AD}_s\mathbf{A}^T + \sigma^2 \mathbf{I} = \sum_{i=1}^m E\{s_i^2\} \mathbf{a}_i \mathbf{a}_i^T + \sigma^2 \mathbf{I} \quad (2.53)$$

式中, a_i 是矩阵 A 的第 i 列。

含噪声的线性信号或数据模型式(2.47)在信号处理和其他领域中会经常遇到,而对 s 和 n 所做的假设也依赖于要解决的具体问题。可以很直接地看出,本例中导出的结果,对相应的协方差矩阵也是成立的。

2.3.2 统计独立性

构成独立成分分析基础的一个关键概念是统计独立性。为简单起见,先考虑两个不同标量

值随机变量 x 和 y 的情形。如果知道随机变量 y 的值并不能给出随机变量 x 取值的任何信息,那么我们说 x 独立于 y 。独立的情况包括: x 和 y 是毫无关联的两个事件的输出;或者是两个非常不同的、之间不存在任何联系的物理过程产生的随机信号。这样的独立随机变量的实例有:掷骰子和抛硬币得到的值,或者语音信号和某个通风系统在特定时刻产生的背景噪声。

数学上,统计独立性是通过联合概率密度来定义的。称随机变量 x 和 y 为独立的,当且仅当:

$$p_{x,y}(x,y) = p_x(x)p_y(y) \quad (2.54)$$

换言之, x 和 y 的联合密度 $p_{x,y}(x,y)$ 必须能分解成它们的边缘密度 $p_x(x)$ 和 $p_y(y)$ 之积。可以通过累积分布函数等价地来定义:在定义式(2.54)中将概率密度函数换成相应的累积分布函数,那么联合累积分布函数也必须是可分解的。

独立的随机变量满足如下基本性质:

$$E\{g(x)h(y)\} = E\{g(x)\}E\{h(y)\} \quad (2.55)$$

式中, $g(x)$ 和 $h(y)$ 分别是关于 x 和 y 的任意绝对可积函数。这是因为:

$$\begin{aligned} E\{g(x)h(y)\} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)p_{x,y}(x,y)dydx \\ &= \int_{-\infty}^{\infty} g(x)p_x(x)dx \int_{-\infty}^{\infty} h(y)p_y(y)dy = E\{g(x)\}E\{h(y)\} \end{aligned} \quad (2.56)$$

式(2.55)展示了这样一个事实:统计独立性是比不相关性强得多的性质。式(2.40) 定义的不相关性可以作为独立性[参见式(2.55)]的特例——即把 $g(x)$ 和 $h(y)$ 都取成线性函数而得出,它只考虑了二阶统计量(相关或协方差)。然而,若随机变量具有联合高斯分布,那么独立性和不相关性就是一回事。这是高斯分布非常特殊的性质,将在 2.5 节中更详细地讨论。

独立性定义式(2.54)可以自然地推广到多于两个的随机变量以及多个随机向量的情形。令 $\mathbf{x}, \mathbf{y}, \mathbf{z}, \dots$, 是随机向量(一般可能具有不同维数)。 $\mathbf{x}, \mathbf{y}, \mathbf{z}, \dots$ 的独立性条件是:

$$p_{x,y,z,\dots}(x,y,z,\dots) = p_x(x)p_y(y)p_z(z)\dots \quad (2.57)$$

而基本性质(2.55)则推广为:

$$E\{g_x(x)g_y(y)g_z(z)\dots\} = E\{g_x(x)\}E\{g_y(y)\}E\{g_z(z)\}\dots \quad (2.58)$$

其中, $g_x(x), g_y(y)$ 和 $g_z(z)$ 分别是随机变量 x, y, z 的任意函数,只要这些函数使得公式(2.58)中定义的期望存在。

定义式(2.57)给出了统计独立性标准观念的一种推广。随机向量 \mathbf{x} 的分量本身就是标量值的随机变量, \mathbf{y} 和 \mathbf{z} 也是一样。很显然, \mathbf{x} 的分量之间可以相互依赖,但它们可以和其他随机向量的分量相互独立,并且使得式(2.57)成立。对随机向量 \mathbf{y} 和 \mathbf{z} 也有类似的论断。

例 2.6 首先考虑例 2.2 和例 2.3 中讨论过的随机变量 x 和 y 。为方便起见,这里再次写出 x 和 y 的联合密度:

$$p_{x,y}(x,y) = \begin{cases} \frac{3}{7}(2-x)(x+y), & x \in [0,2], y \in [0,1] \\ 0, & \text{其他} \end{cases}$$

它不等于例 2.3 中算过的边缘密度 $p_x(x)$ 和 $p_y(y)$ 的乘积。因而方程(2.54)不满足,故我们得出结论: x 和 y 不独立。事实上,直接观察也可看出,上面给出的联合密度 $p_{x,y}(x,y)$ 是不可分解的,因为它不能写成仅依赖于 x 和 y 的两个函数的乘积。

其次考虑参考文献[419]中给出的二维随机向量 $x = (x_1, x_2)^T$ 和一维随机向量 $y = y$:

$$p_{x,y}(x, y) = \begin{cases} (x_1 + 3x_2)y, & x_1, x_2 \in [0, 1], y \in [0, 1] \\ 0, & \text{其他} \end{cases}$$

同样运用上述论证,可以看出随机向量 x 和 y 是统计独立的,但是 x 的分量 x_1 和 x_2 并不独立。这些结论验证的细节留做练习。

2.4 条件密度和贝叶斯法则

至此,我们已经遇到了通常的概率密度、联合密度和边缘密度的概念。还有一类概率密度函数,即条件密度。它们在估计理论中特别重要,估计理论将在第4章中研究。条件密度的概念是在回答如下问题时产生的:“如果随机向量 y 取固定值 y_0 ,那么随机向量 x 的概率密度是什么?”这里 y_0 通常是测量向量 y 的某个特定实现。

假设 x 和 y 的联合密度 $p_{x,y}(x, y)$ 以及它们的边缘密度存在,给定 y 之下 x 的条件概率密度可定义为:

$$p_{x|y}(x|y) = \frac{p_{x,y}(x, y)}{p_y(y)} \quad (2.59)$$

对此定义,可解释如下:假设随机向量 y 落在区域 $y_0 < y \leq y_0 + \Delta y$ 中,则 x 落在区域 $x_0 < x \leq x_0 + \Delta x$ 中的概率是 $p_{x,y}(x_0, y_0)\Delta x$ 。此处 x_0 和 y_0 是常值向量, Δx 和 Δy 都是小量。类似地:

$$p_{y|x}(y|x) = \frac{p_{x,y}(x, y)}{p_x(x)} \quad (2.60)$$

在条件密度中,式(2.59)中的 y 和式(2.60)中的 x 这些条件量,可被看成类似于非随机的参数向量,尽管它们本身实际上是随机向量。

例 2.7 考虑图 2.4 中所描绘的二维联合密度 $p_{x,y}(x, y)$ 。对于给定的常值 x_0 ,其条件分布的密度为:

$$p_{y|x}(y|x_0) = \frac{p_{x,y}(x_0, y)}{p_x(x_0)}$$

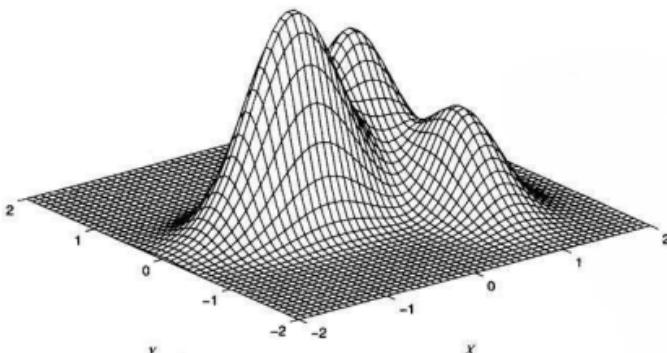


图 2.4 随机变量 x 和 y 的一个二维联合密度

因此它是一个一维密度，可通过在点 $x = x_0$ 处将联合分布做平行于 y 轴的“切片”得到。注意，分母 $p_x(x_0)$ 仅仅是一个尺度常数，它并不影响作为 y 的函数的条件密度 $p_{y|x}(y|x_0)$ 的形状。

类似地，条件分布 $p_{x|y}(x|y_0)$ 在几何上可以从 $y = y_0$ 处平行于 x 轴将联合分布进行“切片”得到。图 2.5 中显示了当 $x_0 = 1.27$ 时的条件分布，图 2.6 中显示的是 $y = -0.37$ 时的条件分布。

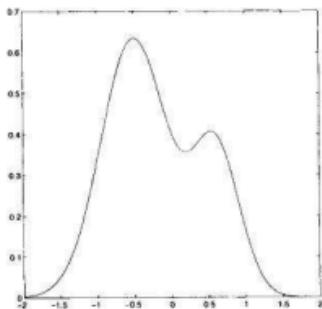


图 2.5 条件概率密度 $p_{y|x}(y|x=1.27)$

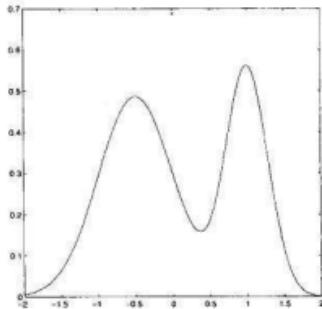


图 2.6 条件概率密度 $p_{x|y}(x|y=-0.37)$

由公式(2.13)和公式(2.14)中给出的 x 和 y 的边缘密度 $p_x(x)$ 和 $p_y(y)$ 的定义，我们可以看到，公式(2.59)和公式(2.60)中的分母可以通过对无条件随机向量的联合密度 $p_{x,y}(x,y)$ 进行积分得到。这也就表明了，条件密度的确是真正的概率密度，且满足：

$$\int_{-\infty}^{\infty} p_{x|y}(\xi|y) d\xi = 1, \quad \int_{-\infty}^{\infty} p_{y|x}(\eta|x) d\eta = 1 \quad (2.61)$$

如果随机向量 x 和 y 统计独立，那么条件密度 $p_{x|y}(x|y)$ 就等于 x 的无条件密度 $p_x(x)$ ，因为 x 不以任何方式依赖于 y ，类似地，有 $p_{y|x}(y|x) = p_y(y)$ 。而且公式(2.59)和公式(2.60)都能写成如下形式：

$$p_{x,y}(x,y) = p_x(x)p_y(y) \quad (2.62)$$

它恰好就是随机向量 x 和 y 的独立性定义。

在一般情形下，我们可以从公式(2.59)和公式(2.60)得到 x 和 y 联合密度的两个不同的表达式：

$$p_{x,y}(x,y) = p_{y|x}(y|x)p_x(x) = p_{x|y}(x|y)p_y(y) \quad (2.63)$$

由此，可以找到以 x 为条件 y 的密度的一个解(反之亦然)：

$$p_{y|x}(y|x) = \frac{p_{x|y}(x|y)p_y(y)}{p_x(x)} \quad (2.64)$$

如有必要，其中的分母可以通过将分子积分而算出：

$$p_x(x) = \int_{-\infty}^{\infty} p_{x|y}(x|\eta)p_y(\eta)d\eta \quad (2.65)$$

公式(2.64)和公式(2.65)合称为贝叶斯法则。该法则在统计估计理论中尤为重要。其中, $p_{x|y}(x|y)$ 是测量向量 x 的条件密度, 而 y 则是未知随机参数向量。假设知道随机参数 y 的验前密度 $p_y(y)$, 贝叶斯法则[参见公式(2.64)]能够在给定特定测量(观测)向量 x 下, 计算出参数 y 的验后密度 $p_{y|x}(y|x)$ 。这些问题将在第4章中更详细地讨论。

条件期望可以用与前面的期望类似的方式定义, 只不过出现在积分中的概率密度函数现在则是适当的条件密度。例如:

$$E\{g(x, y)|y\} = \int_{-\infty}^{\infty} g(\xi, y)p_{x|y}(\xi|y)d\xi \quad (2.66)$$

这仍然是随机向量 y 的函数, 在计算上述期望时则可把 y 看做是非随机的。关于 x 和 y 的完全期望, 可以通过取公式(2.66)关于 y 的期望而得到:

$$E\{g(x, y)\} = E\{E\{g(x, y)|y\}\} \quad (2.67)$$

事实上, 这只不过是计算期望[参见式(2.28)]的另一个两步过程, 可以很容易由贝叶斯法则得出。

2.5 多元高斯密度

多元高斯密度或成为正态密度具有一些特殊的性质, 使得它在众多的密度函数中独一无二。由于其重要性, 本节中我们将对它进行透彻的讨论。

考虑一个 n 维随机向量 x 。我们称 x 为高斯的, 若它的概率密度函数具有如下形式:

$$p_x(x) = \frac{1}{(2\pi)^{n/2} (\det C_x)^{1/2}} \exp\left(-\frac{1}{2}(x - m_x)^T C_x^{-1}(x - m_x)\right) \quad (2.68)$$

我们应还没有忘记, n 是 x 的维数, m_x 是它的均值, 而 C_x 是它的协方差矩阵。记号 $\det A$ 则表示矩阵 A 的行列式, 在这里表示的是 C_x 的行列式。容易看出, 对于单个随机变量 x ($n=1$), 密度公式(2.68)可以归结成例2.1中简要讨论过的一维高斯 pdf[参见公式(2.4)]。另外需要注意的是, 协方差矩阵 C_x 被假定为严格正定的, 这也意味着它的逆存在。

可以证明, 对于密度公式(2.68), 有:

$$E\{x\} = m_x, \quad E\{(x - m_x)(x - m_x)^T\} = C_x \quad (2.69)$$

因此, 分别称 m_x 和 C_x 为该多元高斯密度的均值向量和协方差矩阵。

2.5.1 高斯密度的性质

以下我们将不加证明地列举多元高斯密度的一些最重要的性质。相应的证明可以在许多书里找到, 如参考文献[353, 419, 407]。

决定高斯密度仅需要一阶和二阶统计量:有关 x 的均值向量 m_x 和协方差矩阵 C_x 的信息就足以完全决定高斯密度[参见公式(2.68)]。这样, 所有高阶矩必然仅依赖于 m_x 和 C_x , 即这些高阶矩中已经没有关于高斯密度任何新的信息了。上述事实(以及高斯 pdf 的形式)的一个重要推论是:对于高斯数据, 基于一阶矩和二阶矩的线性处理方法通常是最优的。举例来说, 对于高斯数据, 独立成分分析并不能比标准的主成分分析带来更多新的信息(后面会讨论到); 类似地, 对于高斯数据的滤波, 经典统计信号处理中使用的线性时不变离散时间滤波器是最优的。

高斯随机向量线性变换后仍然是高斯的：若 x 是高斯随机向量，而 $y = Ax$ 是它的一个线性变换，那么 y 也是高斯的， y 的均值向量是 $\mathbf{m}_y = A\mathbf{m}_x$ ，协方差矩阵是 $C_y = AC_xA^T$ 。此结果的一种特例是：联合分布为高斯的高斯随机变量的任何线性组合仍然是高斯的^①。这个结果在标准的独立成分分析中将有重要的隐含意义：不可能对高斯数据估计其ICA模型，也就是说，在没有关于源信号的额外信息情况下，我们不可能从其混合量中分离出高斯源来^②，正如第7章中将要看到的那样。

高斯分布的边缘和条件密度也是高斯的：现在考虑随机向量 x 和 y ，它们的维数分别是 n 和 m 。将它们合成单个随机向量 $z^T = (x^T, y^T)$ ，其维数为 $n+m$ 。该向量的均值向量 \mathbf{m}_z 和协方差矩阵 C_z 是：

$$\mathbf{m}_z = \begin{bmatrix} \mathbf{m}_x \\ \mathbf{m}_y \end{bmatrix}, \quad C_z = \begin{bmatrix} C_x & C_{xy} \\ C_{yx} & C_y \end{bmatrix} \quad (2.70)$$

该式中次对角线上的两个互协方差是互为转置的： $C_{xy} = C_{yx}^T$ 。

现假设 z 具有联合高斯分布。可以证明，联合高斯密度 $p_z(z)$ 的边缘密度 $p_x(x)$ 和 $p_y(y)$ 都是高斯的。另外，条件密度 $p_{y|x}$ 和 $p_{x|y}$ 分别是 n 维和 m 维的高斯密度。条件密度 $p_{y|x}$ 的均值和协方差矩阵是：

$$\mathbf{m}_{y|x} = \mathbf{m}_y + C_{yx} C_x^{-1} (\mathbf{x} - \mathbf{m}_x) \quad (2.71)$$

$$C_{y|x} = C_y - C_{yx} C_x^{-1} C_{xy} \quad (2.72)$$

对条件密度 $p_{x|y}$ ，同样可以得到均值 $\mathbf{m}_{x|y}$ 和协方差矩阵 $C_{x|y}$ 的类似表达式。

不相关性和几何结构的关系：我们在前面已经提到过，不相关高斯随机变量也是独立的，这是其他分布一般所没有的性质。这个重要结果的推导，将作为练习留给读者。如果多元高斯密度[参见公式(2.68)]的协方差矩阵 C_x 不是对角的， x 的分量将是相关的。因为 C_x 是对称正定矩阵，它总可以表示成如下形式：

$$C_x = E D E^T = \sum_{i=1}^n \lambda_i e_i e_i^T \quad (2.73)$$

其中， E 是一个正交矩阵（也就是说它定义了一个旋转），它的各列 e_1, e_2, \dots, e_n 是 C_x 的 n 个特征向量，而 $D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ 是由 C_x 的各相应特征值 λ_i 组成的对角矩阵。现在容易验证，对 x 施加下面的旋转：

$$\mathbf{u} = E^T (\mathbf{x} - \mathbf{m}_x) \quad (2.74)$$

将使具有高斯分布的 \mathbf{u} 向量的各分量不相关，因而也相互独立。

另外，协方差矩阵 C_x 的特征值 λ_i 和随机向量 e_i 揭示了该多元高斯分布的几何结构。任何

① 在原文“高斯随机变量的任何线性组合仍然是高斯的”这句话前面加入了“联合分布为高斯的”，否则该断言不真：因为存在这样的随机向量，它的每个分量（的边缘分布）都是高斯的，但是它们的联合分布非高斯，从而它们的线性组合未必是高斯的。具体例子可以在复旦大学《概率论》第一册第180页例16等处找到。按作者的原意，也可以将这句话改成：“高斯随机向量的任何线性变换仍然是高斯的”或“高斯向量的分量的任何线性组合仍是高斯的”——译者注。

② 然而，在某些特定条件下，利用二阶时间统计量，有可能将相关（非白化的）高斯源在时间上分离开来。这种技术和标准的独立成分分析是很不相同的。这些技术将在第18章中讨论。

pdf 的等值线定义为该密度的常值曲线,由方程 $p_x(\mathbf{x}) = \text{常数}$ 给出。对于多元高斯密度,这等价于要求其指数等于某常数 c :

$$(\mathbf{x} - \mathbf{m}_x)^T \mathbf{C}_x^{-1} (\mathbf{x} - \mathbf{m}_x) = c \quad (2.75)$$

利用公式(2.73),容易看出[419],多元高斯密度的等值线是中心在均值向量 \mathbf{m}_x 处的超椭球面。该超椭球面的主轴平行于特征向量 \mathbf{e}_i ,而特征值 λ_i 是相应的方差,参见图 2.7。

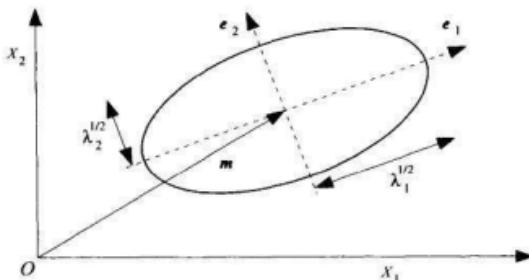


图 2.7 一个多元高斯密度的例子

2.5.2 中心极限定理

高斯分布重要性的另一个原因由中心极限定理给出。令:

$$x_k = \sum_{i=1}^k z_i \quad (2.76)$$

为某独立同分布随机变量的 $\{z_i\}$ 的部分和序列。因为当 $k \rightarrow \infty$ 时, x_k 可能无界地增长,进而考虑标准化的变量:

$$y_k = \frac{x_k - m_{x_k}}{\sigma_{x_k}} \quad (2.77)$$

式中, m_{x_k} 和 σ_{x_k} 是 x_k 的均值和方差。

可以说明,当 $k \rightarrow \infty$ 时, y_k 的分布收敛于具有零均值和单位方差的某个高斯分布。此结果就是我们熟知的中心极限定理。该定理存在另外几种不同形式,这些形式减弱了关于独立性和同分布的假设。中心极限定理是将许多随机现象建模成高斯随机变量的根本原因。比如说,加性噪声常常被认为是产生于大量基本影响效果之和,因此将其建模成高斯随机变量是很自然的。

中心极限定理很容易推广到具有共同均值 \mathbf{m}_z 和协方差矩阵 \mathbf{C}_z 的随机向量 \mathbf{z}_i 的情形。随机向量序列:

$$\mathbf{y}_k = \frac{1}{\sqrt{k}} \sum_{i=1}^k (\mathbf{z}_i - \mathbf{m}_z) \quad (2.78)$$

的极限分布是一个具有零均值和协方差为 \mathbf{C}_z 的多元高斯分布。

中心极限定理在独立成分分析和盲源分离中具有重要的推论。数据向量 \mathbf{x} 的一个混合或者说一个成分,它的一个典型形式是:

$$x_i = \sum_{j=1}^m a_{ij} s_j \quad (2.79)$$

其中, $a_{ij}, j = 1, \dots, m$ 是常值的混合系数, $s_j, j = 1, \dots, m$, 是 m 个未知的源信号。即使源的数目很少(比方说 $m = 10$), 混合 x_k 的分布也通常更接近于高斯分布。实际上, 在不同源密度极不相同且远非高斯的情况下, 上述结论看来仍然成立。此性质的例子可以在第 8 章, 以及参考文献 [149] 里找到。

2.6 变换的密度

现假定 \mathbf{x} 和 \mathbf{y} 都是 n 维随机向量, 通过下述向量映射相联系:

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) \quad (2.80)$$

此映射的逆映射为:

$$\mathbf{x} = \mathbf{g}^{-1}(\mathbf{y}) \quad (2.81)$$

存在且唯一。可以说明, 通过如下公式, 我们能利用 \mathbf{x} 的密度 $p_{\mathbf{x}}(\mathbf{x})$ 得到 \mathbf{y} 的密度 $p_{\mathbf{y}}(\mathbf{y})$:

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{1}{|\det J\mathbf{g}(\mathbf{g}^{-1}(\mathbf{y}))|} p_{\mathbf{x}}(\mathbf{g}^{-1}(\mathbf{y})) \quad (2.82)$$

式中, $J\mathbf{g}$ 是如下雅可比矩阵:

$$J\mathbf{g}(\mathbf{x}) = \begin{bmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial g_n(\mathbf{x})}{\partial x_1} \\ \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_n(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1(\mathbf{x})}{\partial x_n} & \frac{\partial g_2(\mathbf{x})}{\partial x_n} & \cdots & \frac{\partial g_n(\mathbf{x})}{\partial x_n} \end{bmatrix} \quad (2.83)$$

而 $g_j(\mathbf{x})$ 是向量函数的第 j 个分量。

在变换式(2.80)是线性、非奇异的特殊情形下, 有 $\mathbf{y} = \mathbf{Ax}$ 且 $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$, 公式(2.82)可简化为:

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{1}{|\det \mathbf{A}|} p_{\mathbf{x}}(\mathbf{A}^{-1}\mathbf{y}) \quad (2.84)$$

概率论教材中还讨论了其他类型的变换[129, 353]。例如: 和式 $z = x + y$ 在实际中经常出现, 此处 x 和 y 是统计独立的随机变量。但由于这种情况下, 随机变量之间的变换不是一对一的, 故前述结果不能直接使用。但是, 可以说明, z 的 pdf 是 x 和 y 密度的卷积积分[129, 353, 407]。

公式(2.82)是在应用中很重要的一种特殊情形, 即所谓的概率积分变换。若 $F_x(x)$ 是某随机变量 x 的累积分布函数, 那么, 随机变量:

$$z = F_x(x) \quad (2.85)$$

在区间 $[0, 1]$ 上是均匀分布的。利用该结果, 可以从均匀分布的随机变量生成具有所需分布的随机变量: 首先计算所需密度的 cdf, 然后确定式(2.85)的逆变换。由此, 只要能够算出式(2.85)的逆变换, 我们就能得到具有所需密度的随机变量 x 。

2.7 高阶统计量

至此, 我们主要使用二阶统计量来刻画随机向量。在线性离散时间系统中, 统计信号处理中的标准方法正是基于使用这种统计信息。该理论非常成熟, 而且在许多情形下很有用。然而, 它局限于高斯、线性和平稳性等假设。

从 20 世纪 80 年代中期开始, 在信号处理领域中对高阶统计量的兴趣与日俱增。与此同时, 随着几种新的、有效的学习范式的进展, 神经网络开始流行起来。神经网络中的一个基本思想[172, 48]是对输入数据的分布式非线性处理。神经网络是由互相连接的一些称为神经元的简单计算单元构成的, 每个神经元的输出非线性地依赖于它的输入。这些非线性函数, 如双曲正切 $\tanh(u)$, 也隐含地在其处理中引入了高阶统计量。这一点可以通过将非线性函数展开成泰勒级数看出:

$$\tanh(u) = u - \frac{1}{3}u^3 + \frac{2}{15}u^5 - \dots \quad (2.86)$$

在许多神经网络中, 标量 u 是神经元的权向量 w 和输入向量 x 的内积 $u = w^T x$ 。将它代入式(2.86), 可以看出, 计算中将涉及向量 x 分量的高阶统计量。

独立成分分析和盲源分离要求通过非线性直接或间接地使用高阶统计量。因此下面我们将在后面用到的一些相关的基本概念和结果。

2.7.1 峭度与概率密度分类

本节中我们讨论单个标量随机变量的简单高阶统计量。这些统计量虽然简单, 但在很多场合里却非常有用。

考虑一个具有概率密度函数 $p_x(x)$ 的随机变量 x 。 x 的第 j 阶矩 α_j 定义为如下期望:

$$\alpha_j = E\{x^j\} = \int_{-\infty}^{\infty} \xi^j p_x(\xi) d\xi, \quad j = 1, 2, \dots \quad (2.87)$$

相应地定义 x 的第 j 阶中心矩 μ_j 为:

$$\mu_j = E\{(x - \alpha_1)^j\} = \int_{-\infty}^{\infty} (\xi - m_x)^j p_x(\xi) d\xi, \quad j = 1, 2, \dots \quad (2.88)$$

这样, 中心矩是围绕 x 的均值 m_x 计算的, 而均值 m_x 等于一阶矩 α_1 。二阶矩 $\alpha_2 = E|x^2|$ 是 x 的幂的平均。可以看出, 零阶和一阶中心矩 $\mu_0 = 1$ 和 $\mu_1 = 0$ 是无关紧要的, 而二阶中心矩 $\mu_2 = \sigma_x^2$ 就是 x 的方差。

在继续下去之前首先需要指出, 存在某些分布, 它们的所有阶矩都不是有限的。矩的另一个缺点是, 即使知道了全部矩, 也未必能够唯一地确定概率密度函数。幸运的是, 对于经常出现的大多数分布, 它们的所有阶矩都是有限的, 而且知道了有关这些矩的信息, 在实际上就等价于知道了对应的概率密度[315]。

三阶中心矩:

$$\mu_3 = E\{(x - m_x)^3\} \quad (2.89)$$

称为偏度。它是 pdf 非对称性的一个有用的度量。容易看出, 关于均值对称的概率密度其偏度为零。

现在我们仔细考察一下四阶矩。高于四阶的矩和其他统计量在实践中极少使用, 因此我们将不对它们展开讨论。四阶矩 $\alpha_4 = E|x^4|$ 由于其简单性, 在某些 ICA 算法中得到了应用。除四阶中心矩 $\mu_4 = E|(x - m_x)^4|$ 外, 一种称为峭度的四阶统计量因其具有四阶中心矩没有的一些有用性质, 使得它在实际中经常得到应用。峭度的概念将在下一节中介绍累积量的一般理论时正式引出。此处先对它进行一些讨论, 是因为它的简单性和在独立成分分析与盲源分离中的重要性。

在零均值的情况下, 峭度通过如下方程定义:

$$kurt(x) = E\{x^4\} - 3[E\{x^2\}]^2 \quad (2.90)$$

也可以用规范化的峭度，其定义为：

$$\bar{\kappa}(x) = \frac{E\{x^4\}}{[E\{x^2\}]^2} - 3 \quad (2.91)$$

对于白化的数据， $E|x^2| = 1$ ，因此峭度的两个定义都归结为：

$$kurt(x) = \bar{\kappa}(x) = E\{x^4\} - 3 \quad (2.92)$$

这意味着对于白化数据，四阶矩 $E|x^4|$ 可以替代峭度来刻画 x 的分布。峭度基本上可以认为是四阶矩的一个规范化的版本。

峭度的一个有用性质是它的可加性。若 x 和 y 是统计独立的随机变量，那么下式成立^①：

$$kurt(x+y) = kurt(x) + kurt(y) \quad (2.93)$$

然而应该注意，可加性质对四阶矩不成立，这也说明了用累积量代替矩带来的一个重要优点。另外，对于标量参数 β ，有：

$$kurt(\beta x) = \beta^4 kurt(x) \quad (2.94)$$

因而，峭度关于其自变量不是线性的。

峭度另一个重要的特征是，它是在统计上能指示一个随机变量非高斯性的最简单的量。可以说，若 x 具有高斯分布，其峭度 $kurt(x)$ 为零。这样，峭度比四阶矩更具“规范化”意义，对于高斯变量，四阶矩并不是零。

具有零峭度的分布在统计文献中也称为是“中间峭度”(mesokurtic)的分布。一般地，具有负峭度的分布称为是次高斯的(或者在统计学中称为 platykurtic, 扁峭度)。若峭度为正，相应分布称为超高斯的(或 leptokurtic)。次高斯概率密度倾向于比高斯密度更平坦或者多峰。典型的超高斯概率密度比高斯概率密度函数具有更尖锐的峰和更长的拖尾。

峭度常用做对随机变量或随机信号非高斯性的一个定量度量，但是有些地方还是应该引起注意：超高斯信号的峭度可以具有很大的正值(原则上最大值可以是无穷的)，但是次高斯信号其峭度的负值有下界，最小的可能值是 -2 (当方差归一化为 1 时)^②。这样，单纯使用峭度的取值来比较超高斯信号和次高斯信号的非高斯程度是不适当的。不过，如果被比较的信号是同一类型的：或者都是超高斯的或者都是次高斯的，那么峭度就可以作为一种简单的度量来衡量它们的非高斯程度。

在用计算机进行仿真时，一个常用的次高斯分布是均匀分布。对于一个零均值、均匀分布的随机变量 x ，它的 pdf 是：

$$p_x(x) = \begin{cases} \frac{1}{2a}, & x \in [-a, a] \\ 0, & \text{其他} \end{cases} \quad (2.95)$$

其中，参数 a 决定了该 pdf 的宽度(和高度)，参见图 2.8。广泛使用的一个次高斯分布是拉普拉斯分布，或称双指数分布。其概率密度(仍假设零均值)是：

^① 式(2.93)中，必须假设随机变量 x 和 y 的均值是 0，否则不成立，读者若直接将该式两边用中心矩表达出来、再进行比较，就可以看出这一点——译者注。

^② 由 Cauchy-schwarz 不等式可得(假定 $E|x^2| = 1$)：

$$Kurt(x) = E|x^4| - 3 \geq (E|x^2|)^2 - 3 = -2$$

——译者注。

$$p_x(x) = \frac{\lambda}{2} \exp(-\lambda|x|) \quad (2.96)$$

唯一的参数 λ 同时决定了该拉普拉斯密度的方差和峰值的高度。容易看出，随着参数 λ 的增加，该拉普拉斯分布的方差减少，而它在 $x=0$ 处的峰值的高度 $\lambda/2$ 变得更高，参见图 2.9。

均匀密度和拉普拉斯密度都可以作为广义高斯或指数幂类 pdf[53, 256] 的特例得到。该密度族的一般表达式为（假设具有零均值）：

$$p_x(x) = C \exp\left(-\frac{|x|^v}{v E\{|x|^v\}}\right) \quad (2.97)$$

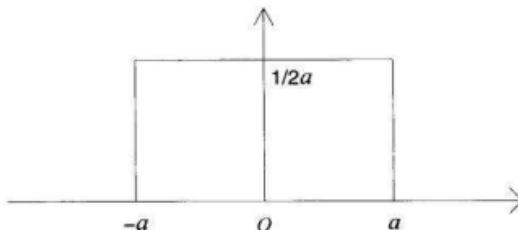


图 2.8 零均值均匀密度的例子

正实值幂 v 决定了分布的类型， C 是尺度常数，它将该分布归一化为具有单位面积（参见参考文献[53]）（分母中的期望也是一个归一化常数）。若参数 $v=2$ ，则得到通常的高斯密度。取 $v=1$ 得到拉普拉斯密度，而 $v \rightarrow \infty$ 给出均匀密度。在式(2.97)中，参数 $v < 2$ 时产生超高斯密度，而 $v > 2$ 时则产生次高斯密度。当 $0 < v < 1$ 时，从式(2.97)中得到的是脉冲型分布。

2.7.2 累积量、矩以及它们的性质

下面我们给出累积量的一般定义。假设 x 是一个实值、零均值的连续标量随机变量，其概率密度函数为 $p_x(x)$ 。那么 x 的第一特征函数 $\varphi(\omega)$ 定义为其 $p_x(x)$ 的连续傅里叶变换：

$$\varphi(\omega) = E\{\exp(j\omega x)\} = \int_{-\infty}^{\infty} \exp(j\omega x) p_x(x) dx \quad (2.98)$$

其中， $j = \sqrt{-1}$ ， ω 是相应于 x 的变换变量。每一个概率分布都可以由它的特征函数唯一决定，反之亦然[353]。将特征函数展成泰勒级数得到[353, 149]：

① (1) 在式(2.97)中，左边的下标 x 和右边括号里分母中的 x 指的是随机变量，另两处 x 则是概率密度函数的自变量；(2) 常数 C 的值等于 $v^{1-\frac{1}{v}} [2(E\{|x|^v\})]^{\frac{1}{v}} \Gamma(\frac{1}{v})^{-1}$ ，其中 $\Gamma(\cdot)$ 是欧拉伽马函数；(3) 计算可知，

$$\tilde{K}(x) = \frac{\Gamma(\frac{1}{v}) \Gamma(\frac{5}{v})}{[\Gamma(\frac{3}{v})]^2} - 3; \quad (4) \text{为了讨论 } v \rightarrow \infty \text{ 时 } p_x(x) \text{ 的极限行为，必须补充 } E\{|x|^v\} \text{ 关于 } v \text{ 的函数关系。例如，}$$

若 $E\{|x|^v\} = \sigma^v$ ， $\sigma > 0$ ，可知，当 $|x| > \sigma$ 时， $p_x(x) \xrightarrow{v \rightarrow \infty} 0$ ；面对 $|x| < \sigma$ ， $p_x(x) \approx C = \frac{\sigma^2}{2\sigma}$ ，在这个意义上， $v \rightarrow \infty$ 时， $p_x(x)$ 趋于平均密度。若不对 $E\{|x|^v\}$ 关于 v 的函数形式做任何假设，则得出任何确定的结论；(5)“脉冲型分布”指的是，当 $0 < v < 1$ 时，函数 $p_x(x)$ 在 $x=0$ 处有一个不光滑的尖峰，形如脉冲。这只是形象的说法，和“脉冲函数”、“脉冲信号”等术语中的“脉冲”不太一样——译者注。

$$\varphi(\omega) = \int_{-\infty}^{\infty} \left(\sum_{k=0}^{\infty} \frac{x^k (\mathrm{j}\omega)^k}{k!} \right) p_x(x) dx = \sum_{k=0}^{\infty} \mathrm{E}\{x^k\} \frac{(\mathrm{j}\omega)^k}{k!} \quad (2.99)$$

展开式中的系数项就是矩 $\mathrm{E}|x^k|$ (假设它们存在)。由于这个原因, 特征函数 $\varphi(\omega)$ 也被称为矩生成函数。

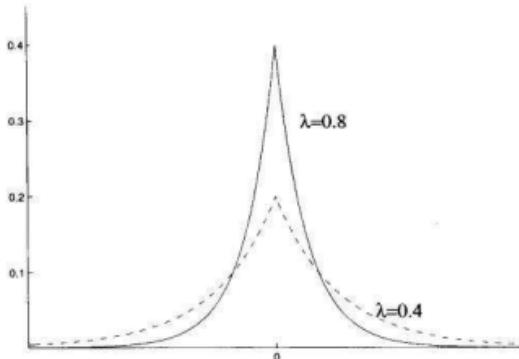


图 2.9 拉普拉斯密度的例子

通常需要使用 x 的第二特征函数 $\phi(\omega)$, 或称累积量生成函数, 这样命名的理由下面就会讨论到。此函数由第一特征函数 [参见式(2.98)] 的自然对数给出:

$$\phi(\omega) = \ln(\varphi(\omega)) = \ln(\mathrm{E}\{\exp(\mathrm{j}\omega x)\}) \quad (2.100)$$

x 的累积量 k_k 类似地定义为与第二特征函数 [参见式(2.100)] 的泰勒级数展开式的系数相对应的矩:

$$\phi(\omega) = \sum_{k=0}^{\infty} \kappa_k \frac{(\mathrm{j}\omega)^k}{k!} \quad (2.101)$$

第 k 个累积量由如下的导数得到:

$$\kappa_k = (-\mathrm{j})^k \left. \frac{d^k \phi(\omega)}{d\omega^k} \right|_{\omega=0} \quad (2.102)$$

对于零均值随机变量 x , 前四个累积量分别是:

$$\kappa_1 = 0, \quad \kappa_2 = \mathrm{E}\{x^2\}, \quad \kappa_3 = \mathrm{E}\{x^3\}$$

以及

$$\kappa_4 = \mathrm{E}\{x^4\} - 3[\mathrm{E}\{x^2\}]^2 \quad (2.103)$$

因此, 前三个累积量等于相应的矩, 而第四个累积量 κ_4 , 我们发现它正是前面在式(2.90)中定义过的峭度。

当 x 的均值 $\mathrm{E}\{x\}$ 非零时, 我们在下面罗列出其各累积量的相应表达式 [319, 386, 149]:

$$\begin{aligned} \kappa_1 &= \mathrm{E}\{x\} \\ \kappa_2 &= \mathrm{E}\{x^2\} - [\mathrm{E}\{x\}]^2 \\ \kappa_3 &= \mathrm{E}\{x^3\} - 3\mathrm{E}\{x^2\}\mathrm{E}\{x\} + 2[\mathrm{E}\{x\}]^3 \\ \kappa_4 &= \mathrm{E}\{x^4\} - 3[\mathrm{E}\{x^2\}]^2 - 4\mathrm{E}\{x^3\}\mathrm{E}\{x\} + 12\mathrm{E}\{x^2\}[\mathrm{E}\{x\}]^2 - 6[\mathrm{E}\{x\}]^4 \end{aligned} \quad (2.104)$$

这些表达式是对第二特征函数 $\phi(\omega)$ 经过繁琐的计算得到的。更高阶累积量的表达式越来越复杂[319, 386]，由于实际中极少使用，因而在此处忽略。

现在简要地考虑一下多元的情形。令 x 是一个随机向量，其概率密度函数为 $p_x(x)$ 。 x 的特征函数仍然是该pdf的傅里叶变换：

$$\varphi(\omega) = E\{\exp(j\omega x)\} = \int_{-\infty}^{\infty} \exp(j\omega x) p_x(x) dx \quad (2.105)$$

这里 ω 是行向量，且与 x 具有相同维数，积分符号是对 x 全部分量进行计算的。 x 的矩和累积量可以用同标量一样的方式得到： x 的矩是第一特征函数 $\varphi(\omega)$ 的泰勒展开式的系数，而累积量是第二特征函数 $\phi(\omega) = \ln(\varphi(\omega))$ 的展开式的系数。在多元的情况下，累积量常常称为互累积量(cross-cumulants)，以便于和互协方差的术语相对应。

可以说明，对零均值的随机向量 x ，它的第二、第三和第四阶累积量分别是：

$$\begin{aligned} \text{cum}(x_i, x_j) &= E\{x_i x_j\} \\ \text{cum}(x_i, x_j, x_k) &= E\{x_i x_j x_k\} \\ \text{cum}(x_i, x_j, x_k, x_l) &= E\{x_i x_j x_k x_l\} - E\{x_i x_j\} E\{x_k x_l\} \\ &\quad - E\{x_i x_k\} E\{x_j x_l\} - E\{x_i x_l\} E\{x_j x_k\} \end{aligned} \quad (2.106)$$

因此，二阶累积量等于二阶矩 $E|x_i x_j|$ ，它同时又是变量 x_i 和 x_j 的相关 r_{ij} 或协方差 c_{ij} 。类似地，三阶累积量 $\text{cum}(x_i, x_j, x_k)$ 等于三阶矩 $E|x_i x_j x_k|$ 。然而，四阶累积量不同于随机变量 x_i, x_j, x_k 和 x_l 的四阶矩 $E|x_i x_j x_k x_l|$ 。

一般来说，高阶矩对应于二阶统计量中的相关，而累积量则是协方差的高阶对应。矩和累积量包含同样的统计信息，因为累积量可以表示成矩的乘积之和。通常我们更愿意使用累积量，原因是它们能更清楚地展示高阶统计量中的附加信息。可以特别说明，累积量具有下述性质，矩则没有这些的性质[319, 386]：

1. 令 x 和 y 是具有相同维数的统计独立随机向量，那么它们的和 $z = x + y$ 的累积量等于 x 和 y 的累积量之和。此性质对于多于两个的独立随机向量之和也同样成立。
2. 若随机向量或随机过程 x 的分布是多元高斯的，那么它的三阶或更高阶的所有累积量都等于零。

这样，高阶累积量度量了一个随机向量与具有相同均值向量和协方差矩阵的高斯随机向量之间的偏差。这个性质非常有用，使得我们能够从一个信号中抽取出其非高斯部分。另外，采用累积量还能使我们忽略混杂在非高斯信号中的加性高斯噪声。

矩、累积量和特征函数还具有这里没有讨论的其他的一些性质。可以参阅参考文献[149, 319, 386]等，并从中获取更多的信息。另外还值得引起注意的是，矩和累积量都具有一些对称性质，在对它们进行估计时，这些对称性质可以用于减轻计算量[319]。

为估计矩和累积量，我们可以采用2.2.4节中引入的例程。但是四阶累积量不能直接估计出：正如在公式(2.106)中显示的那样，必须首先估计那些必须首先计算的矩。实用的估计公式可以在参考文献[319, 315]里找到。

使用高阶统计量也有其相应的缺点：其中之一是，要实现高阶矩和高阶累积量的可靠估计，比二阶统计量的估计需要更多的样本[318]。另一个缺点是，高阶统计量可能对数据中的野值(coutlier)非常敏感(参见8.3.1节)。举例来说，那些具有最大绝对值的少数样本数据可能在很大程度上决定了峭度的值。使用非线性双曲正切函数 $\tanh(u)$ 则能以一种更为鲁棒的方式利用

有关此电子图书的说明

本人由于一些便利条件，可以帮您提供各种中文电子图书资料，且质量均为清晰的 PDF 图片格式，质量要高于网上大量传播的一些超星 PDG 的图书。方便阅读和携带。只要图书不是太新，文学、法律、计算机、人文、经济、医学、工业、学术等方面 的图书，我都可以帮您找到电子版本。所以，当你想要看什么图书时，可以联系我。我的 QQ 是：85013855，大家可以在 QQ 上联系我。

此 PDF 文件为本人亲自制作，请各位爱书之人尊重个人劳动，敬请您不要修改此 PDF 文件。因为这些图书都是有版权的，请各位怜惜电子图书资源，不要随意传播，否则，这些资源更难以得到。