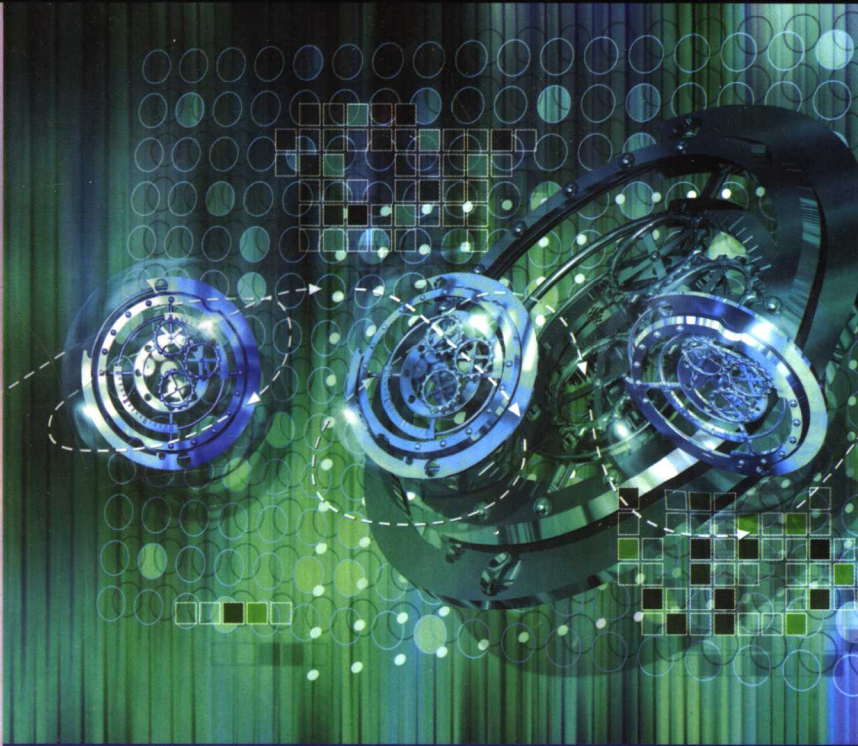


夏新涛 陈晓阳
张永振 杨茹萍 著

滚动轴承乏信息试验分析与评估



科学出版社
www.sciencep.com

滚动轴承乏信息试验分析与评估

夏新涛 陈晓阳 张永振 杨茹萍 著

本书由河南科技大学学术著作出版基金资助出版

科学出版社

北京

内 容 简 介

本书是论述滚动轴承乏信息试验分析与评估方法及其应用技术的学术著作。主要涉及的内容有乏信息系统理论的进展、特征及理论基础,滚动轴承试验中的乏信息数据融合技术、系统属性参数的推断原理、乏信息假设检验原理等;并与目前流行的信息处理方法进行了效果对比分析。本书还给出了滚动轴承乏信息试验的计算方法和大量的计算机仿真试验、工程试验与综合应用实例,并将所提出的乏信息评估方法在军事工程系统、多传感器时间序列融合以及传感器测量特征识别等研究领域进行了应用。

本书可供高等院校教师、研究生和从事滚动轴承研究或试验数据处理与系统信息分析等工作的研究人员使用。

图书在版编目(CIP)数据

滚动轴承乏信息试验分析与评估/夏新涛等著. —北京:科学出版社,2007
ISBN 978-7-03-019200-4

I. 滚… II. 夏… III. 滚动轴承-信息处理 IV. TH133.33

中国版本图书馆 CIP 数据核字(2007)第 091782 号

责任编辑:耿建业 杨 然 / 责任校对:陈玉凤

责任印制:刘士平 / 封面设计:陈 敬

科学出版社出版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

双青印刷厂印刷

科学出版社发行 各地新华书店经销

*

2007年7月第一版 开本:B5(720×1000)

2007年7月第一次印刷 印张:22

印数:1—3 000 字数:421 000

定价:50.00元

(如有印装质量问题,我社负责调换〈双青〉)

前 言

本书的研究内容属于乏信息系统理论的范畴。尽管早在 1763 年贝叶斯公式被提出,1848 年香农创立了信息论,1965 年查德创立了模糊集合理论,1977 年 B. Efron 发展了自助原理,1982 年 Z. Pawlak 提出了粗集理论;但实际上,乏信息研究主要起源于华中科技大学邓聚龙教授 1979 年前后创立的灰色系统理论。随后,各种数学思想开始大量融入灰色系统理论,并聚焦于乏信息系统,由此形成乏信息系统理论。

乏信息又叫贫信息,是指信息缺乏或严重缺乏。例如,在机械系统评估中,系统总体的概率分布未知或概率分布很复杂,同时(或)仅有小子样数据可供参考,就属于乏信息问题;无系统总体的任何概率分布信息,而仅有极少数据的评估属于严重乏信息问题。乏信息也包括趋势项的先验资料问题,趋势项的任何规律性先验信息未知的评估也属于乏信息范畴。

系统的评估主要包括三个问题:估计、预报与假设检验。这些问题在机械系统属性与性能的研究中很重要,传统方法以经典统计学为基本依据,要求概率分布已知或要求大量的数据。随着科学技术的持续发展,经典统计学本身的缺陷逐渐暴露出来。针对滚动轴承的研究来说,军品轴承的小批量多品种研制、高成本的综合性能破坏性试验、新型轴承单元制造误差分析与性能预报等,由于先验性的知识积累少,难以用经典统计学有效地解决。随着社会经济与科学技术的不断发展与完善,诸如此类的问题越来越引起人们的重视,成为近年来信息科学领域中的热点与前沿课题之一。

目前,对小子样问题的研究文献很多,但对概率分布未知同时小子样问题的研究文献较少。现有文献的相关研究基本可以分为统计方法与非统计方法两大类。具体来说,有经典统计方法、熵方法、灰方法、模糊方法、贝叶斯方法、自助法和其他方法等。这些方法将理论上数据个数为无穷减少到数据个数为有限多,或对系统总体的概率分布从正态分布扩展到特定的已知概率分布,从而丰富了统计方法。但对本书所研究的乏信息系统评估问题,以上各种方法都有其局限性。

本书是作者近年来的主要研究成果的总结。针对滚动轴承加工、测量和性能试验等乏信息问题,提出乏信息评估方法。其主要特点是:可以解决试验数据很少,并且系统总体概率分布很复杂或未知时的估计、预报和假设检验等问题,特别是在非平稳随机过程中,针对数据序列中的趋势项先验信息尚未知的条件下,提出解决非平稳随机过程动态推断与检验的原理与方法。其中某些研究问题用目前流

行方法难以解决。

本书将目前流行的数学方法进行综合考虑与融合,取各种数学思想之长而避其短,在不同的问题研究中,都尽可能有机地吸取两种或两种以上数学思想,可以使研究结果明显优于其他方法,解决其他方法难以解决的问题。

本书研究特别注重理论与实践的结合,注重与其他研究方法的效果对比分析,并希望在三个方面取得一些成果:一是提出新的假设检验方法,完善现有假设检验方法,提出在信息缺乏条件下的假设检验方法;二是提出新的静态评估方法,突破统计评估特定概率分布的限制,在数据个数很少的条件下,建立给定置信水平的区间估计模型及系统性能预报模型;三是提出新的动态评估方法,突破灰预报对原始数据有特殊要求的禁区,完善区间估计的自助方法,在外界干扰规律未知的条件下,取得比灰预报和自助预报更准确的预报效果,提高预报的可靠度。这些研究成果形成了比较完整的滚动轴承乏信息试验分析与评估的基本原理和方法。

本书共有 10 章。第 1 章主要论述信息处理方法的进展及其特征、乏信息系统理论的起源及其特征、乏信息系统理论的应用和滚动轴承制造与性能试验中的问题;第 2 章介绍模糊集合理论,灰色系统理论,最大熵原理和自助原理及统计理论等乏信息系统理论的基础知识;第 3 章提出乏信息系统融合的概念描述并论述直接解法、定性融合、定量融合和本征融合方法在轴承试验分析中的应用技术等内容;第 4 章重点论述所提出的各种乏信息数据融合数学模型,如逐步均值累加融合、模糊融合、自助融合和灰自助融合等;第 5 章重点论述所提出的各种乏信息系统属性参数推断的数学模型,如模糊估计与预报、自助估计与预报、基于自助的最大熵估计与预报、基于最大熵的自助估计与预报等;第 6 章为计算机仿真试验与验证;第 7 章主要论述乏信息评估方法在滚动轴承制造误差试验研究、滚动轴承性能试验研究中的应用,以及该方法与其他方法的对比分析等内容;第 8 章主要论述所提出的乏信息假设检验数学模型,如模糊假设检验、基于两个数据序列的灰关系与灰假设检验,并与其他方法进行对比分析;第 9 章为基于乏信息系统理论的圆锥滚子轴承振动的试验研究;第 10 章主要涉及乏信息评估方法在军事工程系统评估中的应用、在多传感器时间序列融合中的应用和在传感器测量特征识别中的应用等问题,并与其他方法进行对比分析。

鉴于乏信息系统理论起源于灰色系统理论,作者特意将一篇未发表的关于学习灰色系统理论感想的文章“灰色系统理论中的科学观初探”作为本书的结束语,以体现“科学技术在行动中”的思想,与读者共勉。

本书获河南科技大学学术著作出版基金资助。书中的研究工作还得到了国家自然科学基金(50375011,50675011)、国防科工委项目子项(MKPT-2004-51ZD/7)和河南科技大学重大科技前期预研专项(2003ZD03)等多项基金的资助。其中的部分内容已经在 *Measurement Science and Technology*、*The Journal of Grey*

System (UK)、*Journal of Grey System* (China, Taiwan)、机械工 程学报、中国机械 工程、计量学报、航空动力学报、岩石力学与工程学报、哈尔滨工业大学学报、吉林 大学学报(工学版)、四川大学学报(工程科学版)、轴承和机床与液压等国内外学术 期刊及 SPIE 等国际学术会议上公开发表。

本书由夏新涛(河南科技大学教授,上海大学在读博士生)、陈晓阳(上海大学 教授,博士,博士生导师)、张永振(河南科技大学教授,博士,博士生导师)以及杨茹 萍(河南科技大学副教授)撰写,一些内容来自河南科技大学硕士研究生的研究成 果和书中各章的参考文献。

为便于读者阅读和研究,书中给出了几乎所有原始数据,并在附录中列出了部 分试验数据和所使用的试验仪器等资料。

洛阳轴研科技股份有限公司李建华、朱孔敏与汤洁,北京航空航天大学王中 宇,洛阳 LYC 轴承有限公司田振山、张立红与刘宏业,以及人本集团轴承有限公司 柏有才等人对作者的研究工作提供了很多帮助,河南科技大学和上海大学的领导 与主管部门,科学出版社的编辑与工作人员为本书的顺利出版付出了大量的心血, 在此谨向他们表示衷心的感谢!

感谢作者的家人们在作者的学术研究和成书过程中所给予的理解与支持!

由于作者水平所限,书中难免存在错误与不妥之处,敬请读者批评指正,以便 再版时修正。

目前,乏信息系统理论仍处于发展的初期阶段,希望通过作者和读者的共同努 力,使之不断完善与进步!

作 者

2007 年 5 月 5 日

目 录

前言

第 1 章 绪论	1
1.1 概述	1
1.1.1 乏信息的含义与表现	1
1.1.2 乏信息系统评估的内容	2
1.2 信息处理方法的进展及其特征	3
1.2.1 信息处理方法的进展	3
1.2.2 信息处理方法的特征	12
1.3 乏信息系统理论的起源及其特征	13
1.3.1 乏信息系统理论的起源	13
1.3.2 乏信息系统理论的特征	15
1.4 乏信息系统理论的应用	16
1.4.1 应用领域	16
1.4.2 应用实例	17
1.5 滚动轴承制造与性能试验中的问题	20
1.5.1 滚动轴承零件制造过程中的问题	21
1.5.2 滚动轴承测量过程中的问题	21
1.5.3 滚动轴承产品性能试验中问题	21
1.6 本章小结	22
参考文献	22
第 2 章 乏信息系统的理论基础	32
2.1 概述	32
2.2 灰色系统理论	32
2.2.1 灰数与白化权函数	32
2.2.2 灰关联性分析	34
2.2.3 灰色预测模型 GM(1,1)	38
2.2.4 灰色预测模型 GM(1,N)	40
2.3 模糊集合理论	41
2.3.1 隶属函数	41
2.3.2 距离	42

2.3.3 贴近度	43
2.3.4 模糊关系	43
2.4 最大熵原理	44
2.4.1 信息熵的概念	44
2.4.2 最大熵原理	45
2.5 自助原理	47
2.5.1 经验分布函数	47
2.5.2 自助样本及其特征参数	47
2.5.3 自助分布	48
2.6 范数理论	48
2.6.1 范数的基本概念	48
2.6.2 数学建模的范数原理	49
2.7 统计理论	50
2.7.1 回归分析与统计建模	50
2.7.2 统计假设检验	56
2.8 本章小结	59
参考文献	59
第3章 乏信息系统融合的概念描述及其在轴承试验分析中的应用方法	61
3.1 乏信息系统融合的概念描述	61
3.1.1 直接解法	61
3.1.2 乏信息的定性融合	61
3.1.3 乏信息的定量融合	62
3.1.4 乏信息的本征融合	63
3.2 乏信息融合技术的应用方法	64
3.2.1 定性融合的应用方法	64
3.2.2 定量融合的应用方法	67
3.2.3 本征融合的应用方法	70
3.3 本章小结	72
参考文献	72
第4章 滚动轴承试验数据的乏信息融合原理与方法	73
4.1 概述	73
4.1.1 静态问题及其乏信息表现	73
4.1.2 动态问题及其乏信息表现	75
4.1.3 乏信息融合方法	76
4.2 单个数据序列的静态融合	76

4.2.1 加权逐步均值累加融合	76
4.2.2 模糊融合	78
4.2.3 自助融合	82
4.2.4 灰自助融合	83
4.3 单个数据序列的动态融合(SDSDFM 模型)	84
4.3.1 滚动自助融合	85
4.3.2 滚动灰自助融合	86
4.3.3 滚动融合的统计量	87
4.4 多个数据序列的融合	90
4.4.1 基本思路	90
4.4.2 融合方法	91
4.5 本章小结	101
参考文献	101
第5章 滚动轴承乏信息试验的参数估计与预报	104
5.1 概述	104
5.1.1 经典统计理论的假设及其局限性	104
5.1.2 滚动轴承试验的乏信息参数估计的基本思想	105
5.2 点估计与数据融合	107
5.2.1 真值估计中的问题	107
5.2.2 参数估计与融合方法	108
5.3 区间数融合	111
5.3.1 随机信息和尺度信息的提取	112
5.3.2 随机信息区间和尺度信息区间的融合	112
5.4 基于模糊范数融合的参数估计与预报	113
5.4.1 隶属函数的建立	113
5.4.2 区间估计与预报	115
5.4.3 经验概率密度函数与置信水平	115
5.4.4 模糊范数融合的计算步骤	116
5.5 基于滚动灰自助融合的动态参数估计与预报 GBM(1,1)模型	118
5.5.1 滚动灰自助分布	118
5.5.2 灰自助动态评估的参数指标	120
5.5.3 灰自助动态评估模型 GBM(1,1)的特点	127
5.6 基于自助原理的最大熵融合的参数估计	129
5.6.1 自助样本	129
5.6.2 最大熵概率密度函数	129

5.6.3	最大熵概率分布	130
5.6.4	参数估计	130
5.7	基于最大熵的自助估计与预报	134
5.7.1	最大熵分布	134
5.7.2	样本的参数估计	135
5.7.3	自助推断	135
5.8	基于最大熵的灰自助估计与预报	139
5.9	最大熵分布的数值解法	139
5.9.1	牛顿数值解法	139
5.9.2	积分区间的映射	140
5.9.3	初始值的选取	141
5.10	本章小结	141
	参考文献	142
第6章	计算机仿真试验与验证	145
6.1	概述	145
6.1.1	仿真与计算机仿真	145
6.1.2	滚动轴承乏信息试验计算机仿真的重要性	146
6.2	随机数与随机数的产生	147
6.2.1	随机数与伪随机数	147
6.2.2	随机数的产生	147
6.2.3	随机数排序数据图	149
6.3	基于模糊范数融合的仿真试验与验证	150
6.3.1	小样本数据的静态评估	151
6.3.2	大样本数据的静态评估	154
6.4	基于区间数据融合与模糊范数融合的仿真试验与验证	156
6.4.1	典型分布的静态预报	156
6.4.2	混合分布的静态预报	160
6.5	基于滚动灰自助融合的GBM(1,1)模型仿真试验与验证	161
6.5.1	动态预报的参数选择	162
6.5.2	典型分布的动态评估	163
6.5.3	混合分布的动态评估	164
6.6	基于自助融合的真值评估	166
6.6.1	单个数据序列的静态评估	166
6.6.2	多个数据序列的动态评估	167
6.7	各种方法的对比分析	169

6.7.1 典型分布的静态值评估	169
6.7.2 典型分布的动态值预报	170
6.7.3 典型分布的动态区间预报	174
6.8 本章小结	176
参考文献	177
第7章 滚动轴承乏信息试验分析与评估方法的工程试验与应用	179
7.1 概述	179
7.2 滚动轴承零件加工质量的试验研究	180
7.2.1 滚动轴承零件尺寸误差的试验研究	180
7.2.2 滚动轴承零件形状误差的试验研究	181
7.2.3 滚动轴承零件相互位置误差的试验研究	184
7.2.4 滚动轴承零件表面波纹度的试验研究	184
7.2.5 滚动轴承零件表面粗糙度的试验研究	185
7.3 滚动轴承产品性能的试验研究	186
7.3.1 滚动轴承摩擦力矩的试验研究	186
7.3.2 滚动轴承振动速度与加速度的试验研究	197
7.3.3 诊断滚动轴承振动速度影响因素的试验研究	203
7.3.4 滚动轴承噪声的试验研究	205
7.3.5 滚动轴承寿命的无失效数据可靠性试验研究	208
7.4 本章小结	210
参考文献	211
第8章 滚动轴承试验中的乏信息假设检验原理	214
8.1 概述	214
8.1.1 假设检验的发展	214
8.1.2 统计假设检验的局限性	215
8.2 模糊假设检验及其应用	215
8.2.1 参数的模糊估计	215
8.2.2 包含点分析法	216
8.2.3 非包含点分析法	217
8.2.4 综合检验	217
8.2.5 与统计检验的对比(正态分布)	217
8.2.6 与统计检验的对比(非正态分布)	219
8.2.7 讨论	220
8.3 基于两个数据序列的灰关系	220
8.3.1 概述	220

8.3.2	非排序灰关系的定义与定理	221
8.3.3	排序灰关系的定义与定理	225
8.3.4	灰关系的应用问题	226
8.4	非排序灰假设检验及其应用	227
8.4.1	非排序灰假设检验原理	227
8.4.2	非排序灰假设检验的仿真试验	227
8.4.3	非排序灰假设检验的工程试验	229
8.5	排序灰假设检验及其应用	230
8.5.1	排序灰假设检验原理	230
8.5.2	排序灰假设检验的仿真试验	231
8.5.3	排序灰假设检验的工程试验	234
8.5.4	讨论	236
8.6	多个时间序列融合的相容性假设检验仿真试验	237
8.7	乏信息假设检验在机械工程系统分析中的应用问题	239
8.8	本章小结	240
	参考文献	240
第 9 章	基于乏信息系统理论的圆锥滚子轴承振动的试验研究	242
9.1	概述	242
9.1.1	滚动轴承振动与噪声研究的发展	242
9.1.2	圆锥滚子轴承振动的研究状况	243
9.1.3	圆锥滚子轴承振动的研究内容与方法	243
9.2	32210 型圆锥滚子轴承振动的试验研究	244
9.2.1	32210 型圆锥滚子轴承振动的试验	244
9.2.2	影响轴承振动的主要因素	246
9.2.3	建立轴承振动的试验模型	250
9.2.4	轴承振动试验模型的验证	253
9.3	30204 型圆锥滚子轴承减振降噪研究	256
9.3.1	轴承参数对振动的影响	256
9.3.2	建立轴承振动的实验模型	260
9.4	不同型号轴承振动的影响因素分析	265
9.4.1	轴承参数对振动的影响	265
9.4.2	不同型号轴承的参数对振动的影响	265
9.5	本章小结	268
	参考文献	269
第 10 章	乏信息系统理论在其他领域中的推广应用与对比分析	271

10.1 概述	271
10.2 在军事工程系统评估中的应用与对比分析	271
10.2.1 捷联惯组的误差系数检验	271
10.2.2 武器系统效能检验	274
10.2.3 武器故障的可靠性融合分析	275
10.2.4 导弹无失效数据的评估	277
10.3 在滑坡体多传感器时间序列融合中的应用与对比分析	278
10.3.1 基本思路	278
10.3.2 滑坡体的试验研究	279
10.3.3 对试验分析结果的讨论	287
10.4 在智能温度仪表数据序列评估中的应用与对比分析	288
10.4.1 温度数据序列的真值估计	288
10.4.2 温度数据序列的静态预报	290
10.4.3 温度数据序列的动态预报	291
10.5 在传感器特征识别中的应用与对比分析	293
10.6 本章小结	295
参考文献	296
结束语 灰色系统理论中的科学观初探	298
附录	304
附录 A 滚动轴承零件表面和结构参数测量仪主要技术参数	304
附录 B 滚动轴承产品性能测量仪主要技术参数	305
附录 C 32210 与 30204 圆锥滚子轴承的有关试验数据	307
附录 D 滚动轴承摩擦力矩试验数据	319
附录 E 乏信息试验分析与评估方法分类表	333

第 1 章 绪 论

乏信息系统理论研究是近年来信息科学和系统科学领域的热点问题与前沿课题之一。本章主要论述信息处理方法的进展及其特征、乏信息系统理论的起源及其特征、乏信息系统理论的应用和滚动轴承制造与性能试验中的某些关键问题。本章的概论性评述将为后续章节理论与实践方面的深入研究打下基础。

1.1 概 述

试验是科学技术研究的必要环节,试验数据(test data)分析是获取试验对象(系统)特征信息最重要的一个过程。在科学技术迅速发展的现代社会里,人们经常会面临如何进行数据和信息分析的问题。近年来,数据和信息分析理论已经取得重大进展,其发展势态已经形成了两个明显的研究方向:一个是数据挖掘(data mining)或信息挖掘,从海量信息中提炼能描述研究对象特征极少量有用信息,由此形成信息挖掘与信息融合(information fusion)理论;另一个是乏信息(poor information)分析,是从研究对象表现出的极其有限的少量信息中,尽可能多地挖掘对象的总体特征,由此形成乏信息系统理论(information poor system theory)。这两个方向的目的的一致,且有很多联系,也建立在几乎相同的理论基础上。本书主要涉及乏信息系统理论在滚动轴承(rolling bearing)试验分析与评估(assessment; evaluation)方面的应用问题。

1.1.1 乏信息的含义与表现

乏信息常常被称为贫信息,是指信息缺乏或严重缺乏。在许多信息科学与系统科学研究的理论中,乏信息系统被描述为信息不完备的不确定性(uncertainty)系统,有时还有数据残缺等。例如,灰色系统理论(grey system theory)、模糊集合理论(fuzzy set theory)、粗集理论(rough set theory)和非统计理论(non-statistical theory)等都可以归属于乏信息系统理论。

在机械系统评估中,系统总体(population)的概率分布(probability distribution)未知或概率分布很复杂,同时(或)仅有小子样(small sample)数据可供参考,就属于乏信息问题;无系统总体的任何概率分布信息,而仅有极少个数据的评估属于严重乏信息问题。乏信息也包括趋势项(trend)的先验(prior)资料问题,无趋势

项的任何规律性先验信息的评估也属于乏信息范畴。

在机械系统的研究中主要有两类乏信息现象。

第一类是已有产品的改进。例如,新型军用装备轴承(bearing)单元,在已经具有类似产品的先验知识(如技术资料)背景下,仅对很少量产品进行试验分析,以获取、验证或预测改进产品的总体性能参数,并将研究结果与现有技术资料作对比分析,作为新的先验知识进行补充。这种现象体现了先验知识是在不断沉淀和积累过程中得到丰富与完善的。

第二类是新产品研制与开发。新产品研制与开发的批量一般很小,特别是一些新型航天轴承(space bearing)组件,品种很多但每个品种每次只有十几套或几套的极小批量,能用于试验研究的更少,几乎没有关于概率分布等可靠的背景资料,属于信息严重缺乏的系统。对这种情况,只能通过极少次的试验研究,来评估产品的总体性能参数,并将研究结果作为后续生产的先验知识进行储备与逐渐积累。

1.1.2 乏信息系统评估的内容

机械系统的评估,从概率论与统计学而言,实际上就是统计推断(statistical inference)的主要内容,可以分为两大类,一类是估计(estimate)问题,另一类是假设检验(hypothesis testing)问题。对系统总体而言,估计又有:点估计(point estimation)与区间估计(interval estimation)、参数估计与非参数估计(non-parametric estimation)、静态估计与动态估计等分类。最常见的估计参数有系统的真值(truth; true value)以及伴随真值的标准差(standard deviation)与置信区间等。系统的参数预报(forecasting; prediction)作为估计的另一种非常特殊的形式,通常是单独进行分类的。在系统总体的信息不全时,有时要推断系统总体的某些未知特性,就要提出对系统总体的某些假设,并根据数据样本对提出的假设做出接受或拒绝的决策,就是假设检验。

因此,机械系统的评估主要包括三个问题:估计、预报与假设检验。这些问题在机械系统属性(attribute)与性能的研究中是很重要的,传统的解决方法是以经典统计学为基本依据的,要求系统概率分布已知或要求具有大量的采样数据。随着科学技术的持续发展,经典统计学本身的缺陷逐渐暴露出来^[1,2]。特别是在现代科学技术的研究中,经常遇到概率分布未知和小子样的乏信息分析问题,如航天试验、高成本的破坏性试验和新产品试制等,难以用经典统计学有效地解决,越来越引起人们的关注与重视,成为近年来系统科学领域的热点问题与前沿课题之一^[3~10]。

从经济角度考虑,该问题的研究可以在试验次数减到最少的条件下获取有用的信息,给出比较可靠的预报,降低研究成本,有比较高的应用价值和经济价值。特别是,该问题的研究是对经典统计学的有益补充。

这些机械系统参数评估问题,很难用经典的统计学方法解决。本书将针对滚动轴承加工、测量和性能试验等乏信息问题,提出一些乏信息分析与评估方法,主要目的是解决试验数据很少且系统总体概率分布很复杂或未知时的参数推断(inference)问题,并针对趋势先验信息缺乏的非平稳随机过程(nonstationary random process)动态评估等问题提出解决方案和基本思路。

1.2 信息处理方法的进展及其特征

1.2.1 信息处理方法的进展

目前,对小子样问题的研究文献很多,但对概率分布未知同时小子样问题的研究文献比较少。从现有文献来看,相关问题研究的方法基本可以分为两大类:一类是统计方法,另一类是非统计方法^[11~17]。具体来说有经典统计学方法、熵(entropy)方法^[18]、灰(grey)方法^[19]、模糊(fuzzy)方法^[20]、贝叶斯方法(Bayes approach, Bayes method)^[21]、自助法(bootstrap)^[2]和其他方法等。这些方法在解决关于概率分布未知和小子样的乏信息问题方面,各自有许多不同的特点和用途。

1. 统计方法

传统的统计方法一般是指用经典统计学和概率论为理论基础的研究方法。

统计方法是应用最广泛的信息处理方法之一。早在17世纪,费尔玛(Fermat)等人就提出了数学期望的概念。在18世纪,统计理论已经受到关注。1794年,德国数学家、测量学家和天文学家高斯(C. F. Gauss)首先阐述了最著名的最小二乘法的基本原理^[6,22]。而统计理论的具体技术应用可以追溯到19世纪初,当时,美国人口普查局就用统计方法对人口死亡率进行过报道^[23],加拿大多伦多采矿局的年鉴就包含了安大略湖(Ontario)采矿工业的统计信息^[24]。但最著名的应用是1805年法国数学家勒让德(A. M. Legendre)在决定彗星轨道的新方法和1809年高斯在天体沿圆锥截面围绕太阳运动的理论中对天文理论与观测数据的最小二乘处理。法国天文学家和数学家拉普拉斯(P. S. Laplace)1812年在分析概率论中最早给出了概率定义,并设计了观测误差理论,谈到了最小二乘法,给出了二项分布极限为正态分布这一定理的证明。高斯和勒让德的主要贡献包括误差分析、正态分布和最小二乘法^[1,6,22]等。到19世纪中期,据NTIS(美国国家技术情报局, National Technical Information Service)等有关统计学应用的报道已数以千万计,内容已延伸到社会学、渔业、医学、光学、军事和工业等领域^[25~29]。从19世纪末到20世纪上半叶,皮尔逊(K. Pearson)、费歇尔(R. A. Fisher)和奈曼(J. Neyman)等数学家的杰出工作创立了经典统计学^[1,8]。从此,经典统计学的理论和方法在社会科学和自然科技领域里得到了广泛应用。多年来,经典统计学的持续应

用也持续地推进着统计理论的发展,而经典统计学本身的某些缺陷也逐渐暴露出来^[1,2,6,8]。

经典统计学的最重要的理论基础是大数定律和中心极限定理。

大数定律论述了算术平均值的稳定性和频率的稳定性问题。频率的稳定性是指,随着独立重复试验次数的增加,事件发生的频率逐渐收敛于事件的概率,当独立重复试验次数很大时,可以用频率代替概率;算术平均值的稳定性是指,当相互独立的随机变量的个数为无限增大时,它们的算术平均值几乎变成一个常数。

中心极限定理认为,许多随机变量是由大量的相互独立的随机因素的综合影响所形成的,其中每一个别因素在总的影响中所起的作用很微小,当这种随机变量的个数增加时,其和的分布趋于正态分布。

统计推断就是以此为基本依据的。但大数定律和中心极限定理实际上隐含着很多假设,至少有三个:试验数据很多,理论上是无穷大;随机变量个数很多且对总体的影响是微小的;相互独立性。因此,统计推断是有条件限制的。事实上,很多工程问题的概率分布不是正态分布,如圆度误差的瑞利分布、滚动轴承疲劳寿命的威布尔分布等,甚至有些系统的概率分布是未知的,如滚动轴承的振动与噪声^[30]以及摩擦与润滑问题等,特别是很难获得系统的大批量数据即试验数据很少。

在统计推断方面,为了弥补经典统计学的缺陷,许多新的数学和物理概念被引入到传统的统计方法中。主要的数学和物理概念有熵、模糊集合、贝叶斯统计、范数^[31]、神经网络^[6,32]、自助再抽样和粗集等。这些方法将理论上数据个数为无穷多减少到有限多,或对系统总体的概率分布从正态分布扩展到特定的已知的概率分布,从而丰富与完善了传统的统计方法,使其应用更加广泛。

2. 信息熵方法

1848年香农(C. E. Shannon)^[18]在创立信息论时,提出了一个用于度量信息源不确定性的量,这个量与热力学和统计力学中熵的数学形式和物理意义都相近,就称之为熵,在信息论中又叫信息熵或香农熵。

设系统 P 的离散随机变量 x 取值 x_i 的概率为 $p_i(x_i)$,熵的定义为

$$H(x) = -k \sum_{i=1}^n p_i(x_i) \lg p_i(x_i) \tag{1-1}$$

选用不同的对数基, k 的取值不同。常用 $k=1$,选自然对数 \ln 。对于连续的随机变量 x , $p(x)$ 为概率密度函数,有

$$H(x) = -k \int_{-\infty}^{+\infty} p(x) \lg p(x) dx \tag{1-2}$$

熵越小,不确定性越小;反之,则越大。在实际应用中,主要涉及最小熵^[33]、最大熵^[34]、相对熵^[35]和经验熵^[36]等概念^[37~39]。

相对熵考察两个系统 P 和 Q 的属性差异问题^[35]。其离散表达式为