



普通高等教育“十一五”国家级规划教材

*TongJi XinHao FenXi
Yu ChuLi*



统计信号分析与处理

侯 强 吴国平 黄 鹰 编著

华中科技大学出版社
<http://www.hustp.com>

普通高等教育“十一五”国家级规划教材

统计信号分析与处理

侯 强 吴国平 黄 鹰 编著

华中科技大学出版社

中国·武汉

图书在版编目(CIP)数据

统计信号分析与处理/侯强 吴国平 黄鹰 编著. —武汉:华中科技大学出版社, 2009年10月
ISBN 978-7-5609-5616-9

I. 统… II. ①侯… ②吴… ③黄… III. ①贝叶斯统计量-信号分析-高等学校-教材 ②贝叶斯统计量-信号处理-高等学校-教材 IV. TN911

中国版本图书馆 CIP 数据核字(2009)第 149697 号

统计信号分析与处理

侯强 吴国平 黄鹰 编著

策划编辑:徐正达
责任编辑:田 密
责任校对:李 琴

封面设计:潘 群
责任监印:周治超

出版发行:华中科技大学出版社(中国·武汉)
武昌喻家山 邮编:430074 电话:(027)87557437

录 排:武汉正风图文照排中心
印 刷:通山金地印务有限公司

开本:710 mm×1000mm 1/16 印张:20.5 字数:390 000
版次:2009年10月第1版 印次:2009年10月第1次印刷 定价:29.80元
ISBN 978-7-5609-5616-9/TN·153

(本书若有印装质量问题,请向出版社发行部调换)

内 容 简 介

本书除了第1章绪论外,包括三大部分。第一部分为基础理论,介绍了全书所关注的理论基础,由第2~4章组成,分别为:统计推断与贝叶斯预测、优化理论与搜索计算以及参数估计与信号检测。这部分主要讨论在贝叶斯统计框架下,搜索与观测数据最佳匹配的模型,并利用各种评价规则来估计模型的参数。第二部分为主题应用,包括第5~8章,包含了四个方面应用:数据建模与系统辨识、自适应信号处理、模式识别的统计方法和基于统计的数据挖掘技术。这部分是全书的应用部分,学生可以根据自己专业的特点有选择地学习。第三部分是本书的提高部分,包括第9章和第10章,分别讨论了人工神经网络和机器学习。

第2~4章是学习本书其余各章节所必不可少的基础,必须仔细体会和琢磨。而有关应用的章节(第5~8章),读者可以按照自己的兴趣或选择阅读或暂时跳过,不必考虑章节次序。最后两章是为学有余力或希望提高自己能力的同学准备的,其他同学目前不研究也没有影响。每章末尾的习题有两个作用:一是加深理解正文的内容;二是介绍一些正文中未能包括的新成果和新应用。每章都介绍一些参考文献。

本书的对象是通信工程、电子信息工程和机电工程专业的高年级本科生和低年级研究生,参考学时32~48。作者希望学习本课程的学生已经学过系统理论课程和概率论与随机过程课程。系统理论课程的内容应包括连续时间系统和离散时间系统的状态变量法和各种变换技术等。

本书备有作者制作的电子教学课件以及相关的仿真程序源代码,免费赠送,需要的任课教师可与责任编辑联系(Email: xuzhengda@163.com)

前 言

信号处理通常被理解为从物理观测结果(数据)中恢复信息(模型)的过程和操作。如果观测结果中明确地包含了信息,并且对出现的任何干扰都能精确地加以描述,那么要求的处理是比较简单的。但观测信息要使用观测装置,传达或观测信息还要通过媒质,它们的物理特性和实际限制,往往使问题复杂化(其中,噪声、畸变和数据丢失,统称为干扰,是使问题复杂化的主要原因)。事实上,干扰通常是随机性的,只能用它的平均特性或统计特性来描述。为恢复信息而对这种随机性观测结果进行的处理称为统计信号处理,本书将致力于研究这个课题和它的一些应用。

在理论研究方面,本书所讨论的统计信号处理,其核心内容是反演、推断和优化,我们将始终围绕这三个关键词展开讨论。反演是一个过程,就是从已知的结果(观测)出发去寻找其未知原因(模型)的过程。统计信号处理就是试图反演这种因果关系。例如,根据观测数据来估计产生数据的可能的密度函数,这个统计学中的主要问题就是反演。而现实中的一些很重要的反演问题又是不适定的,即使这种因果关系形成了一对一的映射,它的反演问题仍然可能是不适定的。贝叶斯统计推断为信号处理和数据建模反演的这种不适定性问题提供了统一的解决框架。在这个框架中,所有的目的都是在寻找与所给数据最佳匹配的模型,并且利用这些模型对数据给出最佳预测。推断是解决不适定性问题的框架,而搜索和优化计算是进行推断和反演具体操作的重要手段。

在实际应用方面,统计信号处理有极其广阔的应用领域。例如,在因特网领域,目前的现状是大量获得的信息中大约只有1%可以被使用。消耗了大量资源的信息不仅未能被有效地使用,而且由于有用的信息正在更深地被掩埋在无用信息之中,变得更难以利用。花费了大量人力、物力而获得的信息,却无法有效使用,这与未获得信息毫无区别。统计信号处理可以帮助我们有效利用这些被掩埋的有用信息。与此有关的研究在不同的计算机分支有着不同的说法。例如,在图像处理中,它称为基于内容的检索,在文本分析中称为文本摘抄、文本检索,在数据分析中称为数据挖掘等,但是,这些任务的本质都是试图从不同媒体表示的信息(观测数据)中发掘出对用户有意义的知识(规律和模型)。无论从何种角度理解和解释这个反演或推断的任务,建立模型和理解数据是两个必须解决的问题。特别是图像与文本这类非结构化或半结构化数据,直接使用这类数据建立模型存在着大量的困难。因此,通常情况下,我们首先利用用户的领域知识或先验知识(贝叶斯先验分布)进行特征提取,以将这类非结构化或半结构化数据变换为结构化数据,并根据这个结构化数据建立模型。由

此, 统计信号处理就成为解决这类反演、推断问题的有力工具。

编者在书中阐述的某些学术观点, 可能仅为一家之言, 同时由于编者水平有限, 错漏之处恳请广大读者批评指正。

本书的出版要感谢中国地质大学(武汉)“十一五”教材建设项目的资助, 同时感谢湖北省统计科研基金项目“3S 集成中误差传播模型的蒙特卡罗模拟研究与应用(批准号: HB082-08)”的资助以及国家自然科学基金“油气圈闭异常场源参数全域灰色自适应解法”(批准号: 40674069-D040901)的资助。特别要对以下老师表示感谢: 中国地质大学(武汉)机电学院张友纯教授、许鸿文副教授和地球物理与空间信息学院的潘和平教授, 没有他们的大力帮助本书不可能顺利出版。

编 者

2008 年 12 月 24 日

目 录

第 1 章 绪论	(1)
1.1 引言与导学	(1)
1.2 随机信号的概念和系统的表征	(3)
1.3 统计信号处理的贝叶斯框架	(6)
1.4 病态条件下的逆问题(反演)及其求解思路	(8)
1.5 搜索及优化计算.....	(10)
1.6 如何有效地利用本书.....	(12)
1.7 总体思路与写作布局.....	(13)
第 2 章 统计推断与贝叶斯预测	(17)
2.1 引言与导学.....	(17)
2.2 贝叶斯估计基础.....	(18)
2.3 贝叶斯估计.....	(26)
2.4 期望-最大算法	(37)
2.5 高斯混合模型的设计.....	(39)
2.6 贝叶斯分类.....	(42)
2.7 随机过程空间的建模.....	(48)
参考文献	(50)
第 3 章 优化理论与搜索计算	(52)
3.1 引言与导学.....	(52)
3.2 最优化问题的下降迭代搜索.....	(53)
3.3 一维搜索(线性搜索).....	(57)
3.4 无约束最优化方法.....	(67)
3.5 约束最优化方法.....	(79)
习题	(91)
参考文献	(92)
第 4 章 参数估计与信号检测	(94)
4.1 引言与导学.....	(94)
4.2 参数估计初步.....	(95)
4.3 最大似然估计	(102)
4.4 线性最小均方估计	(105)
4.5 最小二乘估计	(109)
4.6 信号检测基础	(111)
4.7 判决准则	(115)

4.8 检测性能及其蒙特卡罗仿真	(125)
习题	(129)
参考文献	(131)
第5章 数据建模与系统辨识	(133)
5.1 引言与导学	(133)
5.2 数据建模与系统辨识基础	(135)
5.3 AR(1)模型	(140)
5.4 ARMA(n, m)模型	(145)
5.5 AR模型参数的直接估计法	(149)
5.6 AR模型在语音分析与合成中的应用	(157)
习题	(160)
参考文献	(161)
第6章 自适应信号处理	(163)
6.1 引言与导学	(163)
6.2 性能测量方法	(165)
6.3 基本自适应算法	(176)
习题	(190)
参考文献	(190)
第7章 模式识别的统计方法	(192)
7.1 引言与导学	(192)
7.2 模式的特征与基于模板匹配的认识	(195)
7.3 基于统计决策理论的识别	(200)
7.4 语音信号的产生机理、模型与搜索算法	(201)
7.5 语音信号处理中的统计模式识别	(205)
习题	(213)
参考文献	(214)
第8章 基于统计的数据挖掘	(216)
8.1 引言与导学	(216)
8.2 非参数模型	(218)
8.3 标准线性模型	(221)
8.4 广义线性模型	(224)
8.5 图模型	(231)
8.6 基于统计检验标准的数据挖掘方法评价	(237)
8.7 基于计分函数的标准	(241)
8.8 贝叶斯标准	(242)

8.9 计算标准	(243)
参考文献	(246)
第 9 章 人工神经网络及其应用	(248)
9.1 引言与导学	(248)
9.2 误差信号的反向传播算法	(250)
9.3 径向基函数网络	(259)
9.4 自组织化映射	(262)
9.5 学习向量量化	(264)
9.6 Hopfield 神经网络	(265)
9.7 双向联想存储及其改进	(270)
9.8 玻尔兹曼机和平均场理论学习	(274)
9.9 神经网络在人脸识别中的应用	(279)
习题	(283)
参考文献	(287)
第 10 章 机器学习及其应用	(289)
10.1 引言与导学	(289)
10.2 机器学习的基本问题和方法	(292)
10.3 统计学习理论的核心内容	(296)
10.4 支持向量机	(305)
10.5 支持向量机的应用与研究	(312)
参考文献	(318)

第1章 绪 论

1.1 引言与导学

统计信号处理是现代智能技术中的重要方面,对科学研究的整个过程正起着越来越大的支持作用,已经在生物信息学、计算金融学、遥感信息处理、机器人、工业过程控制、信息安全等方面有卓越贡献。其基本的研究思路是考察从观测数据(样本)出发寻找目前尚不能通过原理分析得到的规律,然后利用这些规律对未来数据或无法观测的数据进行预测。其目的是根据给定的训练样本(观测数据)估计(求解)某系统输入、输出之间的依赖关系,使它能够对未知输出作出尽可能准确的预测。因此统计信号处理的核心内容就是反演、推断(有时也称为解释)和优化。本书将始终围绕这三个关键词展开讨论,在讨论之前,我们先来看看什么是信号处理吧。

信号处理通常被理解为从物理观测结果中恢复信息的过程与操作。如果观测结果中明确地包含了信息,并且对出现的任何干扰都能精确地加以描述,那么要求的处理是比较简单的。但观测信息要使用观测装置,传达或观测信息还要通过媒质,它们的物理特性和实际限制往往使问题复杂化(其中,噪声、畸变和数据丢失统称为干扰,是使问题复杂化的主要原因)。事实上,干扰通常是随机性的,只能用它的平均特性或统计特性来描述。为恢复信息而对这种随机性观测结果进行的处理称为统计信号处理。本书将致力于研究这个课题和它的一些应用。

概率统计是描述复杂世界,尤其是带有随机性质的物理世界的通用语言(我们用最简洁、有力的语言称之为随机信号)。传统的统计学主要包括官方统计、工业统计和现场统计,而现代统计学则注重挖掘数据包含的信息,用已知来预测未知,用过去预测未来。物质运动看似杂乱无章,对个体来说有很多不确定性,但总体上都是服从一些统计规律的。无论医学、经济学、社会科学、工业生产或是科学实验,它们所得到的都是数据,统计学就是对这些数据进行加工和提炼,找出规律,预测未知的学科。统计信号处理在人类活动的各个领域已经得到了广泛的应用,它的应用范围遍及地震信号处理、计算机辅助诊断和治疗、空间目标跟踪、交通管理等许多方面。因此,本书也可以称为随机信号的统计处理,或者直接称为统计信号处理。

为了阐明统计信号处理问题的特点,我们来分析一个空间飞行目标的无线电定位问题。为此可以向目标方向发射一束电磁能,观测反射的电磁波。这时有两个问题要研究。首先是要判断究竟有没有目标存在(检测问题)。如果我们判定目标存在,可能还要求知道与目标有关的某些参数,如它的距离或速度(估计问题)。如果没

有任何干扰,反射波通过传输媒质也没发生畸变,则很容易求得答案。只要监测反射信号,根据信号峰值出现的时间来观测发射波和反射波间的延时 τ 即可。如果没有目标,也就没有尖峰信号;如果有目标,则我们可以估计它的距离 $r = \tau c / 2$, 其中 c 是电磁波的传播速度。

如果存在干扰(噪声),则解答就不那么简单了。干扰可能起因于经过传输媒质时产生的畸变或测量设备的热噪声。干扰掩盖了我们要监测的回波信号尖峰。没有目标时,我们可能得到一个虚假的回波尖峰,而在有目标时,又可能辨别不出目标回波尖峰。无论哪种情况,由于有噪声存在,都有可能做出错误的判决。我们的任务是监测某一段时间的信号,做出关于目标是否存在的判决。这就是检测问题(详细的讨论见第 4 章),它属于一般的统计判决问题。

如果我们已判定目标存在,并试图根据观测到的延时来确定距离,在这过程中我们还会碰到困难。因为干扰会使回波尖峰出现的时间、位置发生误差。这时我们面临一个根据含有噪声的观测结果来恢复信息(目标距离)问题,这就是前面提出过的估计问题(详细的讨论见第 4 章)。

检测和估计这两个问题是所有统计信号处理技术的基础,不仅出现在无线电定位问题中,也出现在其他领域中。系统辨识(第 5 章说明)、信号处理(第 6 章讨论)、模式识别(第 7 章介绍)等领域都会碰到与此类似的问题。在模拟通信系统中,发送的消息经常在传输过程中遭到畸变。在接收端往往用观测结果中的噪声来刻画这种畸变作用。许多情况下,消息(如声音或图像信号)可以用随机信号来描述。所以,接收机恢复消息的问题可以表述成在有随机噪声存在的情况下,估计随机信号的问题。数字通信系统中,把消息编码形成二进制数字序列(一般来说,是编码成若干符号构成的序列)。典型的例子是用 1 或 0 表示这些数字,借助于发送适当选择的脉冲来传输它们,传输的过程中脉冲还要受到畸变。畸变效应使接收机不再能确定发送了哪种波形。我们可以用接收机中的随机噪声来描述传输过程的畸变。这个问题同样是一个判决(检测)问题,要求根据含有噪声的观测结果判断发送了对应 1 或 0 的哪个波形。

高效语言传输方面的最新研究成果,是用语音谱的某些参数来表征语音波形(语音分析)。把这些参数发送给接收机,接收机再根据这些参数综合出语言波形(语音合成)。提取这些参数是一个辨识问题,实质上就是在适当选定的语音波形模型中估计参数问题。这些参数在传输过程中也要遭到畸变,因此接收机的任务是从含有噪声的观测结果中恢复(估计)这些参数。

模式识别系统也包括统计信号处理系统。例如,我们要设计一个自动机来区别两个手写字母(如字母 a 和 b)。这些字母的特征与书写者有关。每一个字样都可以看成是总体(它的统计特性是已知的)的一个取样。我们要设计这样一台机器,当给它一个取样时,它应区分出两种字迹(字母 a 或 b)。另外,总体的统计特性也往往是未知的,还要根据每种字迹的取样来确定它。

上述一切都说明了检测问题和估计问题在各种各样的应用场合中都会出现,或是单独出现,或是同时出现。虽然表面上看,这两个问题好像是统计信号处理的两个分支,实际上这两个问题的结构基本上是相似的,这种相似性有助于解决许多信号处理问题。在第4章参数估计与信号检测中将阐述检测和估计理论的基本原理,把它作为统计信号处理的基础,并且说明这些概念和方法在各领域中的应用。关于应用领域我们选择了数据建模与系统辨识(第5章)、自适应信号处理(第6章)、统计模式识别(第7章)和基于统计的数据挖掘(第8章)等内容。我们举出这些领域绝不是说已经包罗一切了,而是能针对相当广泛的读者就已经足够了。

如何从浩如烟海的数据中提炼信息是每一个从事科研工作的人孜孜以求的事情。信号处理就是从数据中提炼信息,从信息中汇总知识的一门学科。提炼的方法大体可以分为两种:一是确定性的方法,二是统计性的方法。两者各有千秋。对待物理现象我们研究其规律、预测其未来。确定论者认为只要有充分的信息,任何现象都可以为之建立模型并准确地预测其未来。确定论者认为不能准确预测未来的原因是先验信息不足。而概率论的世界观首先就否定了准确预测未来的可行性:一是不可能,二是不必要。在数理统计中,对系统未来的预测或过去的建模统称为统计推断。本书就是在贝叶斯统计推断的总体框架下来讨论信号处理的算法和应用的。

在此我们先对全书所关注的知识要点和教学方法来一个高度的概括和预览。1.2节我们讨论随机信号的概念和系统的表征(representation)问题,有时也称为正演问题。我们研究的物理世界完全可以用随机信号的概念来表征,信号的处理和世界的改造前提假设是系统的表征问题。在目前,随机信号处理都是以贝叶斯推断为总体的统一框架下进行的,在这个框架下信号得到妥善的处理,这是1.3节的内容。在1.4节,我们讨论病态条件下的逆问题及其求解思路。因为绝大多数的统计信号分析与处理的应用问题都是反问题(反演问题),而在工程的条件下,反问题一定是病态的。而病态问题的求解通常可以转化为1.5节所讨论的搜索及优化计算方法。1.6节说明如何最好地利用本书。1.7节介绍总体思路与写作布局。

我们用一句话来结束这一部分:对于物理世界中的各种实际应用问题,我们用随机信号和系统来表征和建模,并且在贝叶斯统计推断框架下进行处理,同时借助优化和搜索的手段来解决信号处理反演中的病态问题。

1.2 随机信号的概念和系统的表征

离散时间随机信号的分析方法和许多结果与连续时间随机信号的有不少相似之处。因此,本章不准备将所有随机过程的问题从头再仔细讨论一遍,而主要是按时间离散的观点和方法讨论各种在数字信号处理中将要用到的基本结果。这些结果是后面各章内容必不可少的基础。本节讨论的重点是离散的相关函数(相关序列)、协方

差函数(协方差序列)和功率谱密度函数(简称功率谱),以及线性非时变离散系统的输入、输出关系。

时间上离散的信号(又称为时间序列、观测数据)和时间上连续的信号一样可以分为确定的和随机的。所谓确定信号,就是其每个时间点上的值可以用某个数学表达式或图表唯一确定的信号。对于这种离散时间的确定信号及其 Z 变换和傅里叶变换的表征,我们假定读者已很熟悉。本节将讨论离散时间的随机信号(即随机序列或时间序列)的表征及有关问题。

随机序列有以下性质。

(1) 随机序列中的任何一个点上的取值都是不能先验确定的随机变量。

以最简单的抛掷硬币这种随机试验为例,每次抛掷结果有两种可能,一是硬币的正面朝上,二是反面朝上。每次抛掷时并不能先验地确定是正面朝上还是反面朝上。如果把正面朝上用 $x=+1$ 表示,反面朝上用 $x=-1$ 表示,连续地抛掷,可以得到一个由 $+1$ 和 -1 组成的序列 $X_i[n]$,如图 1.1 所示。这个序列在任何 n 值点上的取值都是不能先验确知的。因此,称抛掷硬币过程所产生的是一个随机过程。或者说,抛掷硬币产生的序列是一个随机序列。概括这个概念,即一个随机信号(或序列)是一个随机过程,在它的每个时间点上的取值都是随机的,可用一个随机变量表示,或者说,一个随机过程是一个随机试验所产生的随机变量依时序组合得到的序列。今后用 $X_i[n]$ 表示一个随机过程,而用 x_n 表示时间为 n 的点上的一个随机变量。显然任何一个具体试验所得到的序列(见图 1.1 的序列 $X_i[n]$)都只能是随机序列的一个样本序列(或一次实现)。

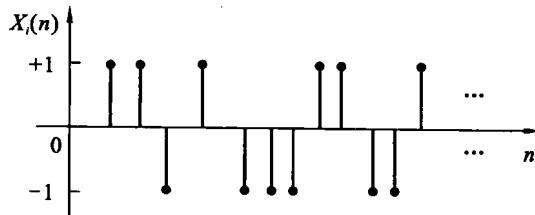


图 1.1 抛掷硬币得到的随机样本序列

(2) 随机序列可以用它的统计平均特性来表征。

虽然上述抛掷硬币试验所得到的随机序列在任何 n 值点上的取值都是不能先验确定的,但是,这种先验不确定的过程中却含有确定的统计规律。不难理解,如果试验条件稳定不变,那么抛掷硬币结果出现 $+1$ 或 -1 的概率都是确定的。又如果硬币的结构是正、反面完全对称的,则出现 $+1$ 或 -1 的概率将都是确定的 $1/2$ 。由其他随机试验产生的随机序列,同样地,在各时间点上的随机变量的取值是服从某种确定的概率分布的。因此,一个随机序列中的每一个随机变量,都可以用确定的概率分布

特性来统计地表征它。或者,可以通过统计平均特性来统计地表征它。因为统计平均特性反映着随机变量的概率分布特性。一个随机变量的各种统计平均特性,是这个随机变量的各种函数按概率加权求平均的运算结果。例如,随机变量 x_n (一次幂) 的统计平均值 $E(x_n)$ 为

$$E(x_n) = \sum_i x_i p(x_i) \quad (1.1)$$

其中, $p(x_i)$ 表示 x_n 取值为 x_i 的概率,它表示 x_n 的概率特性。

总之,一个随机变量的取值尽管是不能先验确知的,但是可以用它取各种可能值的概率特性 $p(x_i)$ 或它的全部完整的统计平均特性 $E(f(x_n))$, 包括 $E(x_n)$, $E(x_n^2)$, $E(x_n^3)$, \dots 来表征它。它们分别从不同侧面描述了 x_n 的取值特性。

一个随机序列在各个 n 值上的取值都是随机变量,但这个随机序列并不是把各种随机试验产生的随机变量任意地放在一起随意编序排列而构成的。我们所遇到的随机序列,常常是在一个作随机运动的系统中某一端口上所观测到的采样数据,依照时序排列构成的时间序列,它的各时间点上的取值之间,往往有前后波及影响。这种波及影响是该系统的各种惯性所决定的。波及影响的统计特性可由描述此序列各时间点取值的多维概率特性来表征。因此,对于一个随机序列,不仅需要知道它在各个 n 值点上的取值特性,还需要知道它的各个不同点间取值的相互关联性(波及性)。这就不仅需要用一阶,而且需要用它的高阶统计平均特性来表征它。

(3) 平稳随机序列是无限持续期、无限能量的时间序列。

我们在实际中遇到的随机序列,不少可认为是平稳序列。平稳的概念是,这种随机序列在各点上的概率特性不随时间平移而变化。平稳序列是无始无终的。

为了信号处理的目的,常常需要将信号转入变换域来表征和认识,即常常需要研究信号的各种变换。因为,一个在时域里的(离散)信号常常是变换到频域(或 Z 域)中作数学处理更为方便。在先行课程中,已详细讨论了确知序列的傅里叶变换和 Z 变换。但是,一个序列的傅里叶变换只有当它的能量是有限的時候才可能存在。对于一个无始无终的平稳随机信号,它的能量是无限的,傅里叶变换是不存在的(它的 Z 变换也是不存在的),因而不能求得这种随机信号的频谱。但是,一个序列如果能量无限并不一定意味着它的功率是无限的。所以,一个随机信号 $X_i[n]$ 的频谱在数学上虽不存在,但它的功率谱是可以存在的。因此,人们可以用功率谱统计地在频域内表征一个随机过程的谱特性。我们知道,一个信号的功率谱是这个信号的自相关函数的傅里叶变换,功率谱和自相关函数是一对傅里叶变换对,它们相互唯一地确定,它们都是信号的一种(二阶)统计平均表征,它们从不同域的侧面表征着一个随机过程的最本质的性质。因此,对于一个观测到的随机信号,重要的是确定它的功率谱密度函数和自相关函数。

从以上所列三点性质及随机信号的概念和系统的表征问题,我们可以概括为以

下两点:一是一个随机信号在各时间点上的取值以及不同点上取值之间的相互关联性只能用概率特性或统计平均特性来表征,它的确定值是无法先验表达的;二是一个无限持续期、无限能量,但有限功率的平稳随机序列,它的频谱是无法求得的,对它来讲,作为统计表达手段,重要的是功率谱密度函数和自相关函数。

1.3 统计信号处理的贝叶斯框架

在 1.1 节中谈到的推断也称为统计推断(statistical inference)或统计推理,是根据带随机性的观测数据(样本)以及问题的条件和假定(模型),而对未知事物做出的,以概率形式表述的推测。它是数理统计学的主要任务,其理论和方法构成数理统计学的主要内容。统计推断的一个基本特点是,其所依据的条件中包含带随机性的观测数据。以随机现象为研究对象的概率论,是统计推断的理论基础。在数理统计学中,统计推断问题常表述为:所研究的问题有一个确定的总体,其总体分布未知或部分未知,通过从该总体中抽取的样本(观测数据)做出与未知分布有关的某种结论。例如,某一群人的身高构成一个总体,通常认为身高是服从正态分布的,但不知道这个总体的均值,随机抽取部分人,测得身高的值,用这些数据来估计这群人的平均身高,这就是一种统计推断形式,即总体的感兴趣的问题(参数)估计。若感兴趣的问题是“平均身高是否超过 1.7 m”,就需要通过样本检验此命题是否成立,这也是一种推断形式,即假设检验。因此我们看到,统计推断在实际应用中有两种基本形式,即参数估计和假设检验。如果把要检验的假设参数化,则这两者几乎是等价的。同学们在后续的学习中一定要仔细体会这层意思。由于统计推断是由部分(样本)推断整体(总体),因此根据样本对总体所作的推断不可能是完全精确和可靠的,其结论要以概率的形式表达。统计推断的目的,是利用问题的基本假定及包含在观测数据中的信息,做出尽量精确和可靠的结论。

贝叶斯方法是统计推断中重要的理论和分支,它有鲜明的特点和独到的处理方法。对于信号处理来说,贝叶斯推断给出了随机信号统计处理的一般框架。全面系统地介绍贝叶斯学派的观点、理论和方法超出了本书的范围,有兴趣的同学可以参考书后的参考文献。而对于着眼于应用统计信号处理理论来解决实际问题的人来说,更重要的是贝叶斯具有特色的一些处理方法以及相应的理论(这是我们重点关注的焦点)。因此,凡是一般教材中有的概念、定理,我们都不加证明地予以引用,但这一部分内容是不多的。所以,本书的写法有一个特点,经常把贝叶斯方法所得的结果与经典方法,即通常教科书中的方法所得的结果加以比较,以便更容易了解、领会贝叶斯方法的特点。如果对经典方法的结果并不了解,则遇到这样的比较就无法理解了,所以要求读者先要具备一定的概率统计的知识。本书的第 2 章着重对贝叶斯方法在

统计信号处理中的应用给予介绍,让读者了解其观点的不同和处理具体问题的方法。

需要强调的是,所谓推断,就是依据事实推导出结论的过程。在统计估计问题中,常假定信号或参数的数值是由观测得到的。在从相关信号对随机过程进行预测或估计中,贝叶斯体系的基本思路是,依据过程概率分布的先验知识,将包含在信号中的事实进行组合。因此,本书在第2章和第4章分别介绍了基于贝叶斯推断的系统预测和参数估计的内容,本节只是简单概述系统预测的一般框架——贝叶斯推断。

系统预测是统计推断的一种表现形式,在时间轴上表现为利用过去或现在的知识和信息来猜测系统未来的状况。贝叶斯推断或预测就是运用贝叶斯统计方法进行的一种预测。那么,什么是贝叶斯统计呢?粗略地讲,在统计推断中使用先验分布的方法进行的统计基本上都是贝叶斯统计。把这个说法略作一点延伸,就可以说,是否使用先验分布是区分贝叶斯统计和非贝叶斯统计的标志。这个说法大体上抓住了问题的要点所在。非贝叶斯统计在做统计推断时只依据两类信息:一是模型信息,二是数据信息。而贝叶斯统计在做统计推断时,除了以上两类信息外,尚需利用另外一种信息,即有关总体分布的未知参数的先验信息(或者称为猜测信息)。由于这类信息是在进行试验以前就有的,故一般称为先验信息(即先于试验的信息)。贝叶斯统计要求这类信息如果能以未知参数 θ 的一个概率分布来表示,这个概率分布就称为先验分布。所以,贝叶斯统计在做统计推断时,既考虑客观信息,又考虑主观信息。一般模式是这样的:由先验分布加上样本信息以及模型的假设信息推出后验分布,然后基于后验分布的最大化对系统进行预测或对信号进行处理。

如上所述,先验分布反映了在做试验前我们关于未知参数的知识,有了样本带来的信息后,这个知识有了改变(改变的部分在数学分析中称为增量——改变量,在统计信号处理中称为新息——新增信息),其结果就反映在后验分布中,或者说,后验分布综合了先验分布的信息和样本的信息。在应用上往往根据增量或新息求出后验分布,然后根据最大化的后验分布做出推断。贝叶斯预测正是建立在利用先验分布求出后验分布的基础上的,只不过中间环节要利用一下新息。

贝叶斯预测的精度取决于贝叶斯参数估计的性能。贝叶斯预测包括了许多传统的预测方法,例如,线性回归、指数平滑和线性时间序列模型(具体内容在第5章中有详细介绍)都是贝叶斯动态模型和预测模型的特殊情况。传统的预测方法和贝叶斯预测方法有一个大的差别,前者总是利用过去的的数据信息,建立统计模型,产生常规预测,并以纯粹的机械形式将输入信息转换为输出信息。另外,传统的预测方法处理不了异常情况的发生。而贝叶斯预测方法不仅利用了过去的的数据信息,还包括人的因素,人对预测负有责任,并提供信息。贝叶斯预测利用客观信息和主观信息相结合的方法进行预测,这是贝叶斯预测的一个重要基本思想。

贝叶斯推断方法包含了诸如最大后验法(MAP)、最大似然法(ML)、最小均方误差法(MMSE)和最小平均绝对值误差法(MAVE)等经典估计方法。这些方法的介

绍和具体的应用将在后续章节中陆续展开。而在模式识别领域得到广泛应用的隐马尔可夫模型则是贝叶斯模型的一个特例(在第7章我们将利用隐马尔可夫模型(HMM)对语音信号进行处理和识别)。贝叶斯推断是基于最小化贝叶斯风险函数的,其中包含了在给定观测和代价函数条件下未知参数的后验模型(在第10章我们将这一理论应用于统计机器学习理论中)。

1.4 病态条件下的逆问题(反演)及其求解思路

在1.1节我们谈到统计信号处理的基本思路是考察从观测数据(样本)出发,寻找目前尚不能通过原理分析得到的规律,然后利用这些规律对未来数据或无法观测的数据进行预测,我们把这个过程称为逆问题(反演)。简单概括地讲,逆问题就是由果(观测数据)求因(内在规律)的问题。其特点通常有两个:一是逆问题通常是病态的或不适宜的;二是它是由“部分信息”求“全部信息”。逆问题的研究必须建立在对正问题透彻研究的基础之上。其中的问题病态原因有三:一是没有任何原因会造成某结果,二是不同的原因造成相同的结果,三是差别很大的原因造成特别相似的结果。从应用的角度看,逆问题可分为两类:一是优化设计问题(在第3章详细讨论),又称为综合问题;二是参数识辨问题(在第4章和第5章详细介绍)。参数辨识方法通常用以下3种方法定义目标函数:

- (1) 最小二乘(LSM)估计,该方法不需要待估参数或测量误差变化的先验知识;
- (2) 最大似然(ML)估计,该方法需要关于测量误差变化的先验知识;
- (3) 最大后验(MAP)估计,该方法需要测量误差变化和待估参数的先验知识。

对于逆问题,信息量获取和评价是非常重要的。一方面,信息的量对逆问题的求解至关重要,信息越多,模型或问题的解就被刻画得越精确,更多的歧义就会被排除,可能的解集(模型空间)就越小,因而搜索的范围也就越小;另一方面,信息的“质”会显著影响信息的“量”。因此,对于逆问题的研究,目前人们主要从以下两方面展开:一方面,从严格数学模型出发,讨论解的存在性和唯一性理论;另一方面,从合理定义正则解开始,将问题转化为某个评价函数(目标函数)的极小值问题,再设定一些先验条件作为极值问题的约束条件。其中很容易想到的办法是穷举法,即逐一计算所有可能的解的目标函数,然后找出那个(那些)目标函数值最小的解。穷举法原理很简单,但是,当列举量无穷大或列举量随解空间维数剧烈增长时,穷举法理论上或实际上是行不通的。为了避免穷举,可以在迭代过程中根据一定的策略,只列举部分可能解进行试验,以得到或逼近最优解。根据不同的策略,可以将优化问题的求解方法分为两类,一是确定性方法,二是随机性方法。确定性方法是指在迭代过程中根据每步迭代所确定的搜索方向和步长一步一步地进行搜索的方法,当然后一步迭代解所对应的目标函数值一定要比前一次迭代解对应的目标函数值要小。不同的确定性方法