统计信号处理基础

——估计理论

杨文　　电子信息学院
测绘校区教学实验大楼十楼1008室
E-mail: yw@eis.whu.edu.cn

贝叶斯

概率论理论创立人Thomas Bayes。如果你使用过 Google, 你就已经从贝叶斯的理论中受益了。

- 贝叶斯1 7 0 1年出生于英国，比费马晚了整整1 0 0年，比帕斯卡晚了7 8年，他的一生并不辉煌。作为一名皇家协会的会员，他生前在数学领域并未发表任何文章。在他死后，他的论文“如何解决随机原理中某一问题的论述”发表了。当时，人们没有对此引起重视。然而，据彼得·伯恩斯坦说，贝叶斯的论述“是一篇极具创新思想的作品，它使贝叶斯在统计学家、经济学家和社会学家中占有不朽的地位。”
- 贝叶斯定理教给我们一种逻辑分析方法，即为什么在众多可能性中只有某一种结果会发生。从概念上讲这是一种简单的步骤。我们首先基于所掌握的证据为每一种结果分配一个概率。当更多的证据出现时，我们对原有的概率进行调整以反映新的信息。为了成功地应用概率原理，关键的一步是要将历史数据与最近可得的数据相结合，这就是行动中的贝叶斯分析法。

“争论”

- 即便在他的时代，**Bayes**发现他自己置身于主流之外。他于**1702**年出生于伦敦，后来他成为了一名**Presbyterian minister**。虽然他看到了自己的两篇论文被发表了，他的理论很有效，但是《**Essay Toward Solving a Problem in the Doctrine of Chances**》却一直到他死后的第三年，也就是**1764**年才被发表。
- 神学家**Richard Price**和法国的数学家**Pierre Simon LaPlace**成为了早期的支持者。该理论和后来**George Boole**，布尔数学之父的理论背道而驰。**George Boole**的理论是基于代数逻辑的，并最终导致了二进制系统的诞生。也是皇室成员之一的**Boole**死于**1864**年。
- 批评者周期性地声称**Bayes**模型依赖于主观的数据，而让人类去判断答案是否正确。而概率论模型没有完全解决在人类思维过程中存在的细微差别的问题。

贝叶斯定理

贝叶斯定理

$$P(H|E,c)=\frac{P(H|c)\,P(E|H,c)}{P(E|c)}$$

尽管这些符号看起来深奥，但这个理论总体来讲还是相当简单的：通过对事情曾经发生的频率的考察，事情发生的可能性似乎能够被真实地估计到。研究者正在把对这种思想的应用从基因研究推广到fillering email 的研究。

在美国明尼苏达州立大学的网址上能够看到详尽的有关数学上的非议，在游戏理论的网站（gametheory .net）上，a bayes RULE Applet 程序会让你回答诸于＂如果被确信有某种疾病你将如何担忧＂此类的问题。

- 贝叶斯定理为我们提供了不断更新我们原有假设数学程序(这源于贝叶斯所称的先验信息分布)产生一个后序信息分布图。换句话说，先验概率与新的信息相结合就产生了后序概率，从而改变了我们相对的概率机遇。
- 这一切都是如何操作的呢？假设你和你的朋友在某个下午正在玩你们最喜欢的掷骰子跳棋游戏，你们一边玩一边聊着，棋局已接近尾声。这时你朋友说的什么话触动了你想打赌的愿望，但只是友好性地赌注。在掷骰子跳棋游戏中，掷一次骰子直接获得6这一面的机会是1／6，即16%的概率。但这时假设你朋友投了骰子，但很快用手将骰子盖住并偷偷看了一眼，她说：“我可以告诉你，这是一个双数。”有了这条信息，你赌赢的机会就变成了1／3，即33%的概率。正当你在考虑是否改变赌注的时候，你的朋友又开玩笑地说：“这个数不是4。”有了这条信息你赌赢的机会再次改变，变成了1／2，即50%的概率。在这种简单的关系中，你已经实施了贝叶斯的分析方法。每一条新信息都会影响你原来的概率假设，这就是贝叶斯推理。

贝叶斯统计

- 英国学者T.贝叶斯1763年在《论有关机遇问题的求解》中提出一种归纳推理的理论，后被一些统计学者发展为一种系统的统计推断方法，称为贝叶斯方法。
- 采用贝叶斯方法作统计推断所得的全部结果，构成贝叶斯统计的内容。认为贝叶斯方法是唯一合理的统计推断方法的统计学者，组成数理统计学中的贝叶斯学派，其形成可追溯到 20 世纪 30 年代。到50～60年代，已发展为一个有影响的学派。时至今日，其影响日益扩大。

贝叶斯学派

- **Bayes**统计模型的特点是将参数 θ 视为随机变量，并具有先验分布 $p(\theta)$;
- **Bayes**统计学派与经典学派的分歧主要是在关于参数的认识上的分歧，经典学派视 θ 为未知常数； **Bayes**学派视 θ 为随机变量且具有先验分布;
- 两个学派分歧的根源在于对于概率的理解。经典学派视概率为事件大量重复实验频率的稳定值；而**Bayes**学派赞成主观概率，将事件的概率理解为认识主体对事件发生的相信程度，当然，对于可以独立重复实验的事件，概率仍可视为频率稳定值。显然，将 θ 视为随机变量且具有先验分布具有实际意义，能拓广统计学应用的范围。

贝叶斯学派发展和完善了古典统计学

- 在本世纪20和30年代，**Neyman, Person,Cramer**等人奠定了古典统计学基础的同时，**Jeffreys, keynes**等人却对贝叶斯学派的传播进行了大量实质性的工作。
- 从此，古典统计学派和贝叶斯统计学派展开了长期的激烈争论。两种统计学派不仅在解释概率方面存在着哲学上的差异，而且在具体的统计推断理论和方法上也各有不同。
- 战后60年代以来，贝叶斯学派异军突起，其统计理论和方法发展迅速，影响越来越大，大有超过古典统计学的趋势。贝叶斯统计学之所以如此，是由于贝叶斯学派发展和完善了古典统计学。具体地表现在以下几个方面。

贝叶斯学派发展和完善了古典统计学

- 贝叶斯学派吸取了古典统计学派的精华
  - 贝叶斯学派和古典统计学派虽然在统计决策方面有许多不同的看法，但这两个流派的研究目的都是一样的，即通过统计分析，帮助决策者在两个或多个行动中选择一种最佳的行动。
  - 和古典统计推断相比，贝叶斯统计推断虽然在处理方法上有所不同，但贝叶斯估计量往往不是对古典统计估计量的全盘否定，而是保留其精华，弥补其不足。
  - 因此，贝叶斯统计学和古典统计学不仅表现在统计目标上一致，而且其结果也有异曲同工之妙。

贝叶斯学派发展和完善了古典统计学

- 贝叶斯学派发展和完善了古典统计学
  - 贝叶斯学派无论是在信息的利用上，还是在统计理论和方法上都发展和完善了古典统计学。从古典统计学派看来，人们在得到样本之前，除了知道参数落在既定的参数空间以外，其它一无所知。因此，古典统计学派统计推断的根据是样本，而不考虑先验知识。然而在许多统计间题中，由于人们从理论上的分析、实践经验的积累以及主观判断，在抽取样本之前，就可能对参数有一些认识。显然，如果能利用这种先验知识，无疑有助于对参数的推断。贝叶斯学派正是在这个问题上独具匠心，发展了古典统计学。

贝叶斯学派发展和完善了古典统计学

- 贝叶斯学派发展和完善了古典统计学
  - 贝叶斯学派认为，我们应该承认这种先验知识，并加以利用。具体的办法可以把先验信息总结为先验分布然后抽取样本，计算参数的条件分布也称后验分布，最后以此进行统计推断。从以上分析可知，古典统计学派只利用了样本信息，而贝叶斯学派不仅利用了样本信息，而且也利用了试验前的先验信息。从统计推断的理论来看，如果可以利用更多的信息，那么，统计推断必定更可靠、更精确。从这点来讲，贝叶斯学派在信息的利用上丰富了古典统计学。
  - 更为困难的是，在实际工作中，有些信息根本无法取得。例如，要对两家企业合并的影响进行假设检验，这种样本信息就无法搜集到，因为人们不可能为了解决两家企业合并的影响，而将两家企业多次合并。在这种情况下，古典统计学派就束手无策，而贝叶斯学派却可以依据贝叶斯理论进行统计分析。

贝叶斯学派发展和完善了古典统计学

- 贝叶斯学派发展和完善了古典统计学
  - 古典统计推断是由样本推断总体参数,着眼于使统计推断尽可能正确。它仅仅是力求使其结论符合客观实际，而不考虑由于决策失误，采取了错误行动而带来的损失。而贝叶斯学派则是在通过样本弄清参数的情况下，从一开始就考虑损失，并且提出许多在不同情况之下的损失函数。
    - 例如，在古典统计学中，人们常常用“无偏估计的方差愈小愈好”的准则，作为优良估计量的判定标准。
    - 然而，贝叶斯学派却可以采用多种不同的损失函数得到各种不同的推断方法。事实上，由于损失函数的多样性和灵活性，进而引出了大量优良性准则，从而丰富和完善了古典统计学的理论和方法。

贝叶斯学派发展和完善了古典统计学

- 贝叶斯统计推断比古典统计推断更精确
  - 由于贝叶斯统计推断既考察了先验知识，又利用了样本信息，它往往比古典统计推断更为精确。此外，从计算难易来看，贝叶斯理论也有简单易行的优点。贝叶斯方法只要算出了后验分布，就可以求解，即使有什么困难，也只是计算性质的，而在古典统计中，问题的解决往往取决于能否推导出复杂的抽样分布。因此，贝叶斯方法更易于初学者接受、推广和普及。

贝叶斯学派发展和完善了古典统计学

- 贝叶斯统计分析方法更符合客观实际
  - 实践是检验一种学术流派或者学术理论方法的最高标准。如果某一学术流派或者学术理论、方法更符合客观实际，那么它就会有强大的生命力，反之亦然。
  - 通过贝叶斯学派和古典学派的比较分析，我们可以知道，无论是点估计、区间估计、还是假设检验，贝叶斯分析方法都比古典统计分析方法更符合客观实际。许多统计问题是一次性的，要求在严格相同甚至大致相同的条件下重复，事实上是不可能的。因此，在许多情况下，古典统计的概念和方法的频率解释完全没有现实意义。

贝叶斯学派发展和完善了古典统计学

- 贝叶斯统计分析方法更符合客观实际
  - 古典统计学派脱离客观实际还表现在，古典统计的可靠度一般是由事先决定的。具体表现在古典统计的可靠程度在抽样之前就决定了，无论具体的样本如何都不能对此产生影响。不仅如此，古典统计的可靠度没有多少理论依据。

贝叶斯学派发展和完善了古典统计学

- 综上所述，贝叶斯学派补充和完善了古典统计学。不容置疑，我们并不能说，贝叶斯统计发展至今就已经完美无缺。贝叶斯学派在如何确定先验分布以及损失函数等方面还需要进一步探讨。
- 在许多统计间题上，古典学派和贝叶斯学派之间不仅存在着哲学和方法论上的争议，而且两个学派内部也存在有差异。正是由于这些不同的理论、方法和观点，促使统计科学生气勃勃，不断发展。

古典统计学派vs贝叶斯学派

- 经典统计的出发点是根据样本，在一定的统计模型下做出统计推断。
- 贝叶斯统计假设统计模型为参数统计模型，在取得样本观测值前，往往对参数统计模型中的参数有某些先验知识（在数学上，关于先验知识的数学描述就是先验分布）。贝叶斯统计的主要特点就是使用先验分布，而在得到样本观测值后，由观测数据与先验分布提供的信息得到后验分布，该后验分布综合了样本与先验信息，组成较完整的后验信息。后验分布是贝叶斯统计推断的基础。

古典统计学派vs贝叶斯学派

- 经典统计只以样本提供的信息在一定统计模型下作统计推断，对样本量较大的样本，有较好的统计推断效果。
- 贝叶斯统计推断由于利用了先验知识，因而对小样本一般也有较好的统计推断效果。

古典统计学派vs贝叶斯学派

- 经典统计的出发点是根据样本，在一定的统计模型下做出统计推断。
- 贝叶斯统计假设统计模型为参数统计模型，在取得样本观测值前，往往对参数统计模型中的参数有某些先验知识（在数学上，关于先验知识的数学描述就是先验分布）。贝叶斯统计的主要特点就是使用先验分布，而在得到样本观测值后，由观测数据与先验分布提供的信息得到后验分布，该后验分布综合了样本与先验信息，组成较完整的后验信息。后验分布是贝叶斯统计推断的基础。

古典统计学派vs贝叶斯学派

- 经典统计只以样本提供的信息在一定统计模型下作统计推断，对样本量较大的样本，有较好的统计推断效果。
- 贝叶斯统计推断由于利用了先验知识，因而对小样本一般也有较好的统计推断效果。

关于贝叶斯学派的争论

- 贝叶斯学派与频率学派争论的焦点在于先验分布的问题。
  - 所谓频率学派是指坚持概率的频率解释的统计学家形成的学派。
  - 贝叶斯学派认为先验分布可以是主观的，它没有也不需要有频率解释。
  - 而频率学派则认为，只有在先验分布有一种不依赖主观的意义，且能根据适当的理论或以往的经验决定时，才允许在统计推断中使用先验分布，否则就会丧失客观性。

关于贝叶斯学派的争论

- 另一个批评是：贝叶斯方法对任何统计问题都给以一种程式化的解法，这导致人们对问题不去作深入分析，而只是机械地套用公式。
  - 贝叶斯学派则认为：从理论上说，可以在一定条件下证明，任何合理的优良性准则必然是相应于一定先验分布的贝叶斯准则，因此每个统计学家自觉或不自觉地都是“贝叶斯主义者”。他们认为，频率学派表面上不使用先验分布，但所得到的解也还是某种先验分布下的贝叶斯解，而这一潜在的先验分布，可能比经过慎重选定的主观先验分布更不合理。

关于贝叶斯学派的争论

- 其次，贝叶斯学派还认为，贝叶斯方法对统计推断和决策问题给出程式化的解是优点而非缺点，因为它免除了寻求抽样分布，（统计量）这个困难的数学问题。而且这种程式化的解法并不是机械地套公式，它要求人们对先验分布、损失函数等的选择作大量的工作。
- 还有，贝叶斯学派认为，用贝叶斯方法求出的解不需要频率解释，因而即使在一次使用下也有意义。反之，根据概率的频率解释而提供的解,则只有在大量次数使用之下才有意义,而这常常不符合应用的实际。这两个学派的争论是战后数理统计学发展中的一个特色。这个争论目前还远没有解决，它对今后数理统计学的发展还将产生影响。

先验概率忽略现象的发现与争论

- Kahneman 和Tversky 开辟了概率推理这一重要的研究领域。他们在20 世纪70 年代初期的研究首先发现，人们的直觉概率推理并不遵循贝叶斯原理，表现在判断中往往忽略问题中的先验概率信息，而主要根据似然概率信息作出判断。
- 他们一个经典性的研究是：告知被试100 人中有70 人是律师，30 人是工程师，从中随机选出一人，当把该人的个性特征描述得象工程师时，被试判断该人为工程师的概率接近0.90。显然被试忽略了工程师的基础概率只有30％。后来他们还采用多种问题验证先验概率忽略现象，如让被试解决如下出租车问题：一个城市85％的出租车属于绿车公司，15％属于蓝车公司，现有一出租车卷入肇事逃逸事件，根据一目击者确认，肇事车属于蓝车公司，目击者的可靠性为80％。问肇事车是蓝车的概率是多少。结果大多数被试判断为80％，但如果考虑先验概率则应是41％。
- 贝叶斯推理在过去近30 年中得到了较为广泛的研究，特别自Kahneman 和Tversky 发现人们直觉的概率判断忽略先验概率现象以来，出现了许多理论和研究方法的更新，这些都深化了对这一问题的研究。

贝叶斯统计的两个基本概念

- 贝叶斯统计中的两个基本概念是先验分布和后验分布

①先验分布。总体分布参数 θ 的一个概率分布。贝叶斯学派的根本观点，是认为在关于总体分布参数 θ 的任何统计推断问题中，除了使用样本所提供的信息外，还必须规定一个先验分布，它是在进行统计推断时不可缺少的一个要素。他们认为先验分布不必有客观的依据，可以部分地或完全地基于主观信念。

②后验分布。根据样本分布和未知参数的先验分布，用概率论中求条件概率分布的方法，求出的在样本已知下，未知参数的条件分布。因为这个分布是在抽样以后才得到的，故称为后验分布。贝叶斯推断方法的关键是任何推断都必须且只须根据后验分布，而不能再涉及样本分布。

先验分布与后验分布

- 对参数的传统认识
  - 设$x \sim p(x;\theta)$，传统上，我们把**x**当作随机变量，而把θ 当作确定的未知常量。
  - $p(x;\theta)$ 提供的知识与信息是关于**x**的，它反映了**x**取值的规律性，它没有反映θ 的变化规律。$p(x;\theta)$对确定**(**估计**)** 会有帮助。 **x**提供的知识包含了θ 的信息。

先验分布与后验分布

- **Bayes** 统计对参数的认识
  - **Bayes**学派认为：θ 不是常量参数，而是随机变量，它可能取各种不同的值，取各种不同值的概率分布$p(\theta)$也是确定的。
  - 例**3**：某厂每天的废品率p。
  - p的算法：在当天的成品中，进行产品全检，计算其废品率p ；或者抽取部分成品，估计其废品率p 。
  - 从当天看， p是一个单纯的未知常数，但从较长的时间看，每天都有一个p值，其值因随机因素的作用，会产生波动，当天的p值可合理地视为随机变量p的一个可能值。如果我们有相当长一个时期的检验记录，则可以相当精确地定出p的概率分布。

先验分布

- 形式上，把参数θ看成一个随机变量，并给出的概率分布$P(\theta)$，或概率密度$p(\theta)$，这个分布在抽样前就给出了，把它称为θ的先验分布。
  - 注**1**：有时，把参数θ看成随机变量有其合理性，但把所有的未知参数都视为随机变量则牵强。例，要估计某铁矿的含铁量***p***，把***p***看成随机变量，就要设想这个铁矿是无穷多“类似”铁矿的一个样本，这是不自然的，不如干净利落地把***p***看作一个孤立的未知常数。
  - 注**2**：有时，把参数θ看成随机变量有其合理性，但人们的先验知识没有确切到能用概率分布把表达出来。于是，引出了一系列先验分布的确定方法。

先验分布定义

(1)参数θ的参数空间Θ上的一个概率分布称为θ的先验分布，即Bayes参数统计模型中的参数θ是参数空间Θ上的随机变量，它的概率分布叫参数θ的先验分布，记为$\{p(\theta):\theta\in\Theta\}$

(2)样本$\mathbf{X}=(X_1,X_2,\cdots,X_n)^T$的条件密度函数族$\{f(x|\theta):\theta\in\Theta\}$称为样本分布族

(3)先验分布$\{p(\theta):\theta\in\Theta\}$与样本分布族$\{f(x|\theta):\theta\in\Theta\}$构成Bayes参数统计模型

后验分布

定义：在$X=\mathrm{x}$的条件下，θ的条件分布称为θ的后验分布，后验分布由后验概率密度函数$p\{(\theta|\mathrm{x}):\theta\in\Theta\}$描述

理解：后验分布的意义在于综合了关于 θ 的先验信息（反映在先验分布 $p(\theta)$ 中）和样本观测值**x**关于 θ 的信息（反映在样本分布$p(\mathrm{x}|\theta)$中）。先验分布概括了在试验前对 θ 的认识，而得到样本观测值**x**之后，对 θ 的认识有了深化，这集中反映在后验分布中。**Bayes**公式反映了先验分布到后验分布的转化，即**Bayes**自己所说的“归纳推理”的统计方法。

Bayes统计推断的原则

- **Bayes统计推断的原则**
  - 样本x的作用:对Bayes统计而言，样本x的唯一作用在于把对 θ 的认识由先验分布转化成后验分布。
  - Bayes统计推断的原则:对参数 θ 所作的任何推断(估计、检验等)必须基于且只能基于 θ 的后验分布。
  - 对原则的理解:一经由样本x算出了 θ 的后验分布，就设想我们除了这一后验分布外，其余的东西(样本值、样本分布、先验分布)全忘记了。这时，对 θ 的推断的唯一凭借就是这一后验分布。传统的许多统计推断原则许多都不能用了。如，无偏性原则，$\hat{\theta} = T(\mathrm{x}), E(\hat{\theta}) = \theta$ ，完全利用样本(样本的函数)在进行推断没有利用后验分布，不符合的Bayes统计推断的原则。

先验PDF的选择

围绕贝叶斯估计量的使用上有许多争议，这源于在实践中不能证明先验PDF

- 贝叶斯估计的结果是在“平均”的意义上或者是假定 θ 的先验PDF的情况下的最佳估计
- 先验知识很难以先验PDF的形式表达
- 如果先验统计不准确，那么贝叶斯估计量将是高度有偏的，这类似于在经典估计量问题中使用不正确的数据模型设计估计量。
- 除非先验概率是建立在物理约束的基础上，否则还是使用经典估计比较合适。

选取先验分布的方法

- **Bayes**统计中，选取先验分布是一个相当重要的问题。
- 若对参数 θ 选用均匀分布，但其函数**g(** θ **)**往往不服从均匀分布，即往往不再服从**Bayes**假设；
- 又由**Bayes**统计推断原则，后验分布是统计推断的基础，而只有正确选择的先验分布，才有正确的后验分布。因此在**Bayes**统计中，必须深入探讨选取先验分布的方法。

先验分布的确定方法

- 客观法
  - 以前的资料积累较多，对 θ 的先验分布能作出较准确的统计或估计。在这种情况下，分布的确定没有渗杂多少人的主观因素，故称之为客观法。
  - 如果能用客观法确定 θ 的先验分布 $p(\theta)$，对贝叶斯学派持否定态度的统计学者也不反对用贝叶斯方法去作数据处理。
  - 在不少情况下，以往积累的资料并不是直接给出了参数在当时的取值，而只是一种估计。例如，某厂产品的废品率，不可能是全检(可能是破坏性检验)。有些资料不是直接关于 θ 取值分布的记录，但我们可以利用这些资料对 θ 的先验分布作出经验性的推断。

先验分布的确定方法

- 主观概率法
  - 按**Bayes**学派的说法，这是一种通过“自我反省”去确定先验分布的方法。就是说，对参数 θ 取某某值的可能性多大，通过思考，觉得该如何，而定下一个值。
  - 主观先验分布反映了个人以往对 θ 的了解，包括经验知识和理论知识，其中有部分可能是通过他人获取的，也可能是他人对 θ 的了解。
  - 对过去的经验和知识，必须经过组织和整理。
  - 这样提出的先验分布，在主观上是正确的，但不能保证合乎某种客观标准。

先验分布的确定方法

- 同等无知原则
  - 这一原则称为**Bayes**假定。以产品的废品率p为例，当我们对p一无所知时，我们只好先验地认为，以同等机会取[0,1]内各种值，因而以[0,1]内均匀分布$U[0,1]$作为p的先验分布。
  - 注：这一原则会出现矛盾：如果我们对p无知，对p^3也同样无知。按同等无知原则，可以取$U[0,1]$作为p^3的分布，但这时p的分布就不是$U[0,1]$了。

Bayes假设是在对参数“无信息”的条件下，认为参数在其取值范围内，取各个值的可能性都相同，无所偏爱。也称“无信息先验分布”

先验分布的确定方法

- **共轭分布方法**

H.Raiffa, R.Schlaifer提出先验分布应取共轭分布

设样本X的分布族为$p\{(x|\theta):\theta\in\Theta\}$, 若先验分布$p(\theta)$与后验分布$p(\theta|x)$属于同一分布类型，则先验分布$p(\theta)$称为$p(\theta|x)$的共轭分布。

*共轭分布要求先验分布和后验分布具有同一形式，注意到

$$p(\theta|x)\propto p(\theta)L(\theta|x)$$

可见共轭分布要求先验分布$p(\theta)$提供的信息与样本分布$L(\theta|x)$提供的信息综合以后，不改变θ的总的分布规律。

先验分布的确定方法

- 共轭分布方法
  - 共轭分布方法实质上认为由样本提供的信息是主要的。它要求先验分布与后验分布属于同一类型，就是要求过去的经验知识通过样本信息转化为同一类型的经验知识。在不断的取得新的样本观测值前，现时的后验分布可看成进一步试验或观测的先验分布。这样，人们对 θ 的认识就能不断深化。因此，共轭分布方法确实应作为选取先验分布的重要方法。

先验分布的确定原则—Jeffreys原则

- Jeffreys原则

Jeffreys提出的选取先验分布的原则是一种不变原理，较好地解决了**Bayes**假设中的一个矛盾，即若对参数 θ 选用均匀分布，则其函数 $g(\theta)$往往不是均匀分布。

Jeffreys原则：设按照原则决定 θ 的先验分布为$p(\theta)$,若以 $g(\theta)$ 作为参数，按同一原则决定的$\eta = g(\theta)$的先验分布是$p_g(\eta)$,则应用关系式：

$$p(\theta) = p_g[g(\theta)]\,|\,g'(\theta)\,|$$

若选取的 $p(\theta)$符合上式，则用 θ 或 θ 的函数 $g(\theta)$ 导出的先验分布总是一致的。

困难之处在于如何找到满足上式的 $p(\theta)$，**Jeffreys**利用**Fisher**信息量的不变性，找到了符合要求的$p(\theta)$

先验分布的确定原则—Jeffreys原则

引理：设$\eta=g(\theta)$与θ具有相同维数p，则有

$$|I(\theta)|^{1/2}=\left|\frac{\partial g(\theta)}{\partial \theta}\right||I(\eta)|^{1/2}$$

$|I(\theta)|^{1/2},|I(\eta)|^{1/2}$分别表示$I(\theta),I(\eta)$的行列式的平方根，而$\left(\frac{\partial g(\theta)}{\partial \theta}\right)_{p\times p},\left|\frac{\partial g(\theta)}{\partial \theta}\right|$是其行列式绝对值。

由此引理与*Jeffreys*原则，可取

$$p(\theta)\propto|I(\theta)|^{1/2}$$

该式对标量参数和矢量参数都适用。

先验分布的确定原则—Jeffreys原则

例1：设X是来自正态总体N(μ,1)的IID样本，求μ的先验分布。

可求得$I(\mu)=n$,故$p(\mu)\propto 1$

例2：设X是来自正态总体N(0, σ^2)的IID样本，求σ与$\delta=\sigma^2$的先验分布。

可求得$I(\sigma)=\frac{2n}{\sigma^2}, I(\delta)=\frac{2n}{\delta^2}$,故

$$p(\sigma)\propto\frac{1}{\sigma}, p(\delta)\propto\frac{1}{\delta}$$

先验分布的确定原则—最大熵原则

- 最大熵原则
  - 熵是信息论的一个基本概念，是随机变量不确定性的度量。不确定性越大，则熵越大，在“无信息”的情况下，应取熵最大的分布为先验分布。
  - 最大熵原则：无信息先验分布应取参数 θ 的变化范围内使熵最大的分布。
  - 回顾信息论中离散和连续情况下的最大熵定理

贝叶斯公式和贝叶斯假设

- 贝叶斯学派的起点是贝叶斯的两项基本工作：贝叶斯定理和贝叶斯假设。贝叶斯定理（公式）在通常的概率论的教科书中都有叙述，而贝叶斯假设几乎都不提及

贝叶斯公式

- 贝叶斯公式的事件形式是：

  假定$A_1,\ldots,A_k$是互不相融的事件，他们之和 $\bigcup_{i=1}^{k} A_i$ 包含事件B，即B$\subset$ $\bigcup_{i=1}^{k} A_i$ ，则有：

$$P(A_i \mid B)=\frac{P(A_i)P(B \mid A_i)}{\sum_{j=1}^{n} P(A_i)P(B \mid A_j)} \qquad (i=1,2,\ldots,n)$$

贝叶斯公式

- 贝叶斯公式的随机变量形式：
- 设两个随机变量x，y，p(x|y)是x对y的条件密度，q(y)是y的边缘密度，于是y对x的条件密度是：

$$h(y|x)=\frac{q(y)p(x|y)}{\int q(y)p(x|y)\mathrm{d}y}$$

- 贝叶斯公式的离散随机变量形式：
- 只要将字母中的积分号改为求和号即可

贝叶斯方法梗概

1. 将未知参数看成随机变量（或随机向量），记为 θ，当 θ 已知时，样本 $x_1, \ldots, x_n$ 的联合分布密度 $p(x_1, \ldots, x_n; \theta)$ 就看成是 $x_1, \ldots, x_n$ 对 θ 的条件密度，记为 $p(x1, \ldots, xn|\theta)$，简写为 $p(x|\theta)$
2. 设法确定先验分布 π（θ），这是根据以往对参数 θ 的知识来确定的，引起争议

贝叶斯方法梗概

3. 利用条件分布密度$p(x_1, \ldots, x_n \mid \theta)$和先验分布$\pi(\theta)$，可以求出$x_1, \ldots, x_n$与$\theta$的联合分布和样本$x_1, \ldots, x_n$的分布，于是就可以用它们求得$\theta$对$x_1, \ldots, x_n$的条件分布密度，也就是用贝叶斯公式求得后验分布密度$h(\theta \mid x_1, \ldots, x_n)$
4. 利用后验分布密度$h(\theta \mid x_1, \ldots, x_n)$做出对$\theta$的推断（估计$\theta$或对$\theta$做检验）

贝叶斯假设

- 通过上面的讨论可以看出，应用贝叶斯公式时，需要知道参数θ的分布—先验分布，然后才能导出**h(**θ|a**)**分布。
- 在2中，如果没有任何以往的知识来帮助我们确定先验分布π（θ），贝叶斯提出可以采用均匀分布作为π（θ），即参数在他变化的范围内，取到各个值的机会是相同的。这种确定先验分布的原则，就称为贝叶斯假设

对贝叶斯方法的批评

- 对贝叶斯方法的批评集中在以下几点：
- 一是把 θ 看成随机变量是否妥当
- 二是贝叶斯假设认为 θ 没有任何先验信息，既然无先验信息，又哪来先验分布呢？（就是先验分布是否存在）
- 三是均匀分布的存在性：均匀分布在有限区域内是存在的，但在无限区域内还存在吗？

贝叶斯方法— θ 是否随机变量

- 比如在打靶问题中，对某人的打靶技术事先一无所知，只能凭n次打靶的结果来估计，此时把每次命中的概率 θ 看成是随机变量似乎有些勉强
- 但是进一步考虑，正因为对每次的命中率没有任何知识，它在0与1之间取哪个值的可能性完全是相同的，也就是 θ 取各个不同的值都有相同的机会，也就可以把 θ 可以看作是随机变量

贝叶斯方法—先验分布是否存在

- 在讨论一些理论问题时，假定先验分布密度是已知的，这是完全可以的；然而在实际工作中，有时候对参数 θ 是没有任何过去的值是可以借鉴，而希望通过实验结果来获得，这时的先验分布称为无信息先验分布**(Non-informative Priors)**。如何获得无信息先验，是贝叶斯方法的一个重大的理论问题

贝叶斯方法—先验分布是否存在

- 这里无信息的含义是确切的，它指的是没有任何信息可以帮助我们去选用一个特定的分布作为先验分布，并不是说对参数θ的其他情况一无所知。
- 至少知道两点：参数θ与样本x1,...,xn联合分布密度的关系，经典方法认为已知x1,...,xn的联合密度是p(x1,...,xn;θ)，从贝叶斯的观点来看，就是已知x1,...,xn对θ的条件密度；其次我们也知道θ的取值范围。

贝叶斯方法—均匀分布的存在性

- 这是贝叶斯假设面临的一个困难，在打靶问题中，每次的命中概率 θ 在[0，1]内均匀分布可以接受，但是正态分布的两个参数 μ 和 σ^2，他们的变化范围都是无限的区间，而在无限区间上，均匀分布是不存在的
- 因为均匀分布相应的密度是一个常数，即密度函数f(x)=c，c是一个常数，而

$$\int f(x)dx=c\int dx=c*\infty=\infty$$

贝叶斯方法—均匀分布的存在性

- 另一方面，未知参数的选择具有任意性，例如正态分布$N(\mu, \sigma^2)$的参数σ^2既可以取σ也可以取σ^2为参数。如果σ在$(0, \infty)$上均匀分布（假定存在），则σ^2就不可能均匀；反之亦然
- 但是根据贝叶斯假设，无论对σ还是σ^2，都应选均匀分布，这就是贝叶斯假设面临的另一个问题

统计概率

统计概率:若在大量重复试验中，事件A发生的频率稳定地接近于一个固定的常数p，它表明事件A出现的可能性大小，则称此常数p为事件A发生的概率，记为P(A)，即

$$p=P(A)$$

可见概率就是频率的稳定中心。任何事件A的概率为不大于1的非负实数，即

$$0<P(A)<1$$

条件概率

条件概率:我们把事件B已经出现的条件下，事件A发生的概率记做为P(A|B)。并称之为在B出现的条件下A出现的条件概率，而称P(A)为无条件概率。

若事件A与B中的任一个出现，并不影响另一事件出现的概率，即当P(A)＝P(A·B)或P(B)＝P(B·A)时，则称A与B是相互独立的事件。

加法定理

两个不相容(互斥)事件之和的概率，等于两个事件概率之和，即

$$P(A+B) = P(A) + P(B)$$

若A、B为两任意事件，则：

$$P(A+B) = P(A) + P(B) - P(AB)$$

乘法定理

设A、B为两个任意的非零事件，则其乘积的概率等于A(或B)的概率与在A(或B)出现的条件下B(或A)出现的条件概率的乘积。

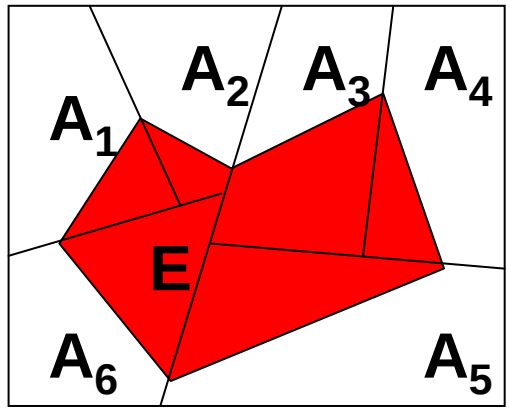
$$P(A \cdot B) = P(A) \cdot P(B|A)$$

或 $$P(A \cdot B) = P(B) \cdot P(A|B)$$

贝叶斯规则

$$p(A_i \mid B) = \frac{p(B \mid A_i)p(A_i)}{p(B)} = \frac{p(B \mid A_i)p(A_i)}{\sum_i p(B \mid A_i)p(A_i)}$$

$i=1\ldots n$



- 设事件A_1，A_2，…，A_n构成互不相容的事件组，贝叶斯公式如上给出，先验信息以$\{P(A_i), i=1\ldots n\}$给出，即先验分布。由于事件B的发生，可以对A_1，A_2，…，A_n发生的概率重新估计。
- 贝叶斯公式综合了先验信息与试验提供的新信息，获得了后验信息，以后验概率$\{P(A_i \mid B), i=1\ldots n\}$体现出来，贝叶斯公式反映了先验分布向后验分布的转化。

巴菲特vs查理·蒙格

- “人类并没有被赋予随时随地感知一切、了解一切的天赋。但是人类如果努力去了解，去感知—通过筛选众多的机会—就一定能找到一个错位的赌注。而且，”查理说：“聪明的人会在世界提供给他这一机遇时下大赌注。当成功概率很高时他们下了大赌注，而其余的时间他们按兵不动，事情就是这么简单。”
- “对这种初级数学，你必须学以致用而且日积月累地应用于生活中。如果你不能将这种初级数学中的初级概率应用于生活中的方方面面(尽管应用的有些不自然)，那么你的一生就像一个瘸腿的人参加赛跑，永远处于不利的地位。如果拥有这种数学能力，你就会比别人拥有巨大优势。”

贝叶斯参数估计

- **Bayes**学派认为：后验分布族 $p(\theta \mid \mathrm{x}, \theta \in \Theta)$ 是统计推断的出发点。这里样本观测值是确定的，而 θ 是随机的。而经典统计中，其出发点是样本分布族，其中 θ 是未知参数，而样本观测值**x**只是无限次可能试验结果的一个具体实现，总体来说，样本**x**是随机的。
- 无偏性与样本分布族有关，因而“无偏性”不符合**Bayes**统计推断原则。
- 总之， **Bayes**统计推断的任务是根据已知的样本观测值**x**对未知的随机变量根据后验分布作出推断，这里**x**是具体的值，没有必要将其放在“无限多可能值之一”中考察。

贝叶斯估计

- 问题：传统估[illegible]情况下有其理论上的[illegible]重复大量的试[illegible]得到少量的[illegible]是否优良，是[illegible]小样本情况下[illegible]

设x(x[i], i=0, 1,[illegible]较大时，用传统方法估计θ，估计[illegible]。[illegible]，特别是当N=1或N=2时，传统估计不是很可[illegible]

过去的看法、记忆或经验，常常支配着我们对事物的判断(估计、评判)。

- 出路：用过去的经验，用人们过去对的了解(或部分了解)，给出较可靠、较切合实际的估计。

先验知识和估计

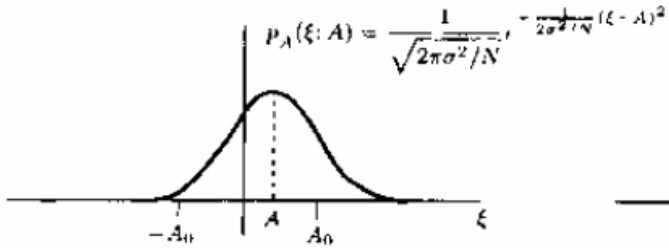
- 在贝叶斯估计中，先验**PDF**的选择是很关键的，错误的选择将导致差的估计量，这类似于在经典估计量问题中使用不正确的数据模型设计估计量。
- 围绕贝叶斯估计量的使用上有许多争议，这源于在实践中不能证明先验**PDF**。完全可以认为，除非先验概率是建立在物理约束的基础上，否则还是使用经典估计比较合适。

先验知识和估计

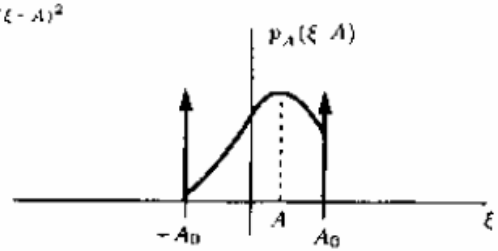
比较 两种估计量的 MSE，我们会注意到，对于区间 $-A_0 \leqslant A \leqslant A_0$ 上的任何 A，

$$
\begin{aligned}
\operatorname{mse}(\hat{A}) &= \int_{\infty}^{\infty} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi \\
&= \int_{-\infty}^{-A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi \\
&\quad + \int_{A_0}^{\infty} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi \\
&> \int_{-\infty}^{-A_0} (-A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi \\
&\quad + \int_{A_0}^{\infty} (A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi \\
&= \operatorname{mse}(\check{A})
\end{aligned}
$$

截断样本均值估计量$\check{A}$从MSE上来看要优于样本均值估计量。尽管$\hat{A}$仍然是MVU估计量，我们允许估计量是有偏的来减小均方误差。（在经典情况下使用最佳MSE准则常常导出不可实现的估计量，在贝叶斯方法中，这个问题就不存在了）。

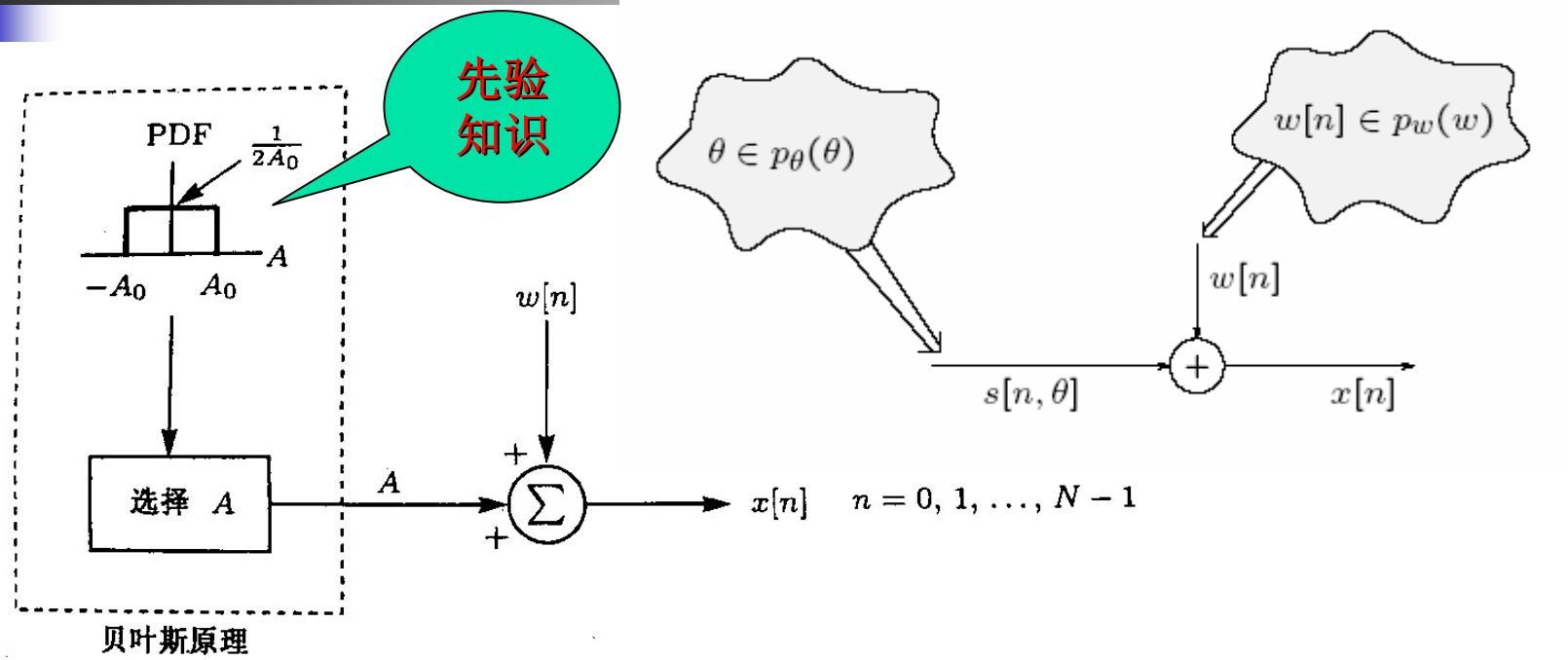


(a) 样本均值的 PDF



(b) 截断样本均值的 PDF

贝叶斯估计的数据建模



根据给定的PDF选择A的行为是Bayes方法和经典方法的不同之处。**以往的问题是估计A的值或者随机变量的实现，而现在我们可以把A是如何选择的知识结合进来**。

先验知识和估计

在贝叶斯方法中，我们可以寻找一种估计量$\hat{A}$,使如下定义的贝叶斯MSE最小，

$$Bmse(\hat{A})=E[(A-\hat{A})^2]$$

与经典估计误差定义$\hat{A}-A$相反，我们定义误差$A-\hat{A}$。在上式的计算过程中，要注意期望运算是对联合PDF$p(x,A)$求取的，这是MSE与经典估计本质上的不同。

经典MSE：　$mse(\hat{A})=\int(\hat{A}-A)^2 p(x;A)dx$

贝叶斯MSE：　$Bmse(\hat{A})=\iint(A-\hat{A})^2 p(x,A)dxdA$

可见，在贝叶斯MSE中，可以通过积分消除对参数的依赖性。

先验知识和估计

下面推导使贝叶斯*MSE*最小的估计量，首先，应用贝叶斯原理，得到

$$p(\mathrm{x}, A) = p(A/\mathrm{x})\,p(\mathrm{x})$$

因此，$Bmse(\hat{A}) = \int\left[\int (A-\hat{A})^2\, p(A/\mathrm{x})dA\right]p(\mathrm{x})d\mathrm{x}$

由于对于所有的x而言有$p(\mathrm{x}) \geq 0$，如果括号内的积分对每一个x能够最小那么贝叶斯MSE将达到最小。因此，固定x使$\hat{A}$是一个标量变量，则有

$$\frac{\partial}{\partial \hat{A}}\int (A-\hat{A})^2\, p(A/\mathrm{x})dA = \int \frac{\partial}{\partial \hat{A}}(A-\hat{A})^2\, p(A/\mathrm{x})dA$$

$$= \int -2(A-\hat{A})\,p(A/\mathrm{x})dA$$

$$= -2\int Ap(A/\mathrm{x})dA + 2\hat{A}\int p(A/\mathrm{x})dA$$

令偏导数为零，得：$\hat{A} = \int Ap(A/\mathrm{x})dA$,即$\hat{A} = E(A/\mathrm{x})$

可见，使贝叶斯MSE最小得最佳估计量是后验PDFp(A/x)的均值

先验知识和估计

后验*PDF*是指得到观测数据后*A*的*PDF*。与此相对，$p(A)$或者

$$p(A)=\int p(\mathrm{x},\mathrm{A})d\mathrm{x}$$

可看作*A*的先验*PDF*,它表示在数据被观测到之前的*PDF*。今后我们称使贝叶斯*MSE*最小的估计量为最小均方误差(MMSE)估计量。在确定MMSE估计量时，首先需要知道后验PDF,可以利用贝叶斯规则确定后验PDF,

$$p(A/\mathrm{x})=\frac{p(\mathrm{x}/\mathrm{A})p(A)}{p(\mathrm{x})}=\frac{p(\mathrm{x}/\mathrm{A})p(A)}{\int p(\mathrm{x}/\mathrm{A})p(A)d\mathrm{A}}$$

注意到分母正好是一个与A无关的归一化因子。

先验知识和估计

参数估计的贝叶斯方法假设要估计的参数是随机变量θ的一个现实。我们对它指定一个先验$PDF p(\theta)$,在观测到数据后，后验$PDF p(\theta / \mathrm{x})$概况了我们对这个参数的了解情况。对$\theta$的所有现实和x，使平均MSE最小的估计量定义为最佳估计量，即所谓的贝叶斯MSE。该估计量是后验PDF的均值，即$\hat{\theta} = E(\theta / \mathrm{x})$,可表示为

$$\hat{\theta} = E(\theta / \mathrm{x}) = \int \theta p(\theta / \mathrm{x}) d\theta$$

一般而言，MMSE估计量依赖于先验知识和数据，如果先验知识相对于数据较弱，那么估计量将忽略先验知识。否则，估计量将偏向于先验均值。如期望的那样，利用先验知识通常能改善估计精度。

MMSE估计的性质

- MMSE估计被确定为是E(θ |x)或后验PDF的均值，因此它通常也称为条件均值估计量。
  - 对于线性变换，它可以交换。即满足线性特性；
  - 对于独立的数据集，MMSE估计量具有叠加性质。
  - 在联合高斯情况下，MMSE是数据的线性函数。

高斯PDF的特性

- 高斯先验PDF在实际中非常有用是因为它的再生性：如果$p(x, A)$是高斯的，那么边缘PDF 也是高斯的，且后验PDF $p(A|x)$ 也是高斯的。
- 高斯先验PDF在很多实践问题中都会遇到。

多维条件高斯PDF

• 高斯情况：

如果 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 是联合高斯，$\boldsymbol{x}$ 是 $k\times 1$，$\boldsymbol{y}$ 是 $\ell\times 1$ 矢量，均值矢量为 $[E(\boldsymbol{x}),E(\boldsymbol{y})]^T$，分块协方差矩阵

$$C=\begin{pmatrix} C_{xx}, & C_{xy} \\ C_{yx}, & C_{yy} \end{pmatrix}$$

$$p(x,y)=\frac{1}{(2\pi)^{\frac{k+l}{2}}[\det(C)]^{\frac{1}{2}}}\exp\left\{-\left[\frac{1}{2}\begin{bmatrix}\boldsymbol{x}-E(\boldsymbol{x})\\ \boldsymbol{y}-E(\boldsymbol{y})\end{bmatrix}^T\cdot C^{-1}\begin{bmatrix}\boldsymbol{x}-E(\boldsymbol{x})\\ \boldsymbol{y}-E(\boldsymbol{y})\end{bmatrix}\right]\right\}$$

则条件 PDF：$p(\boldsymbol{y}\mid\boldsymbol{x})$ 也是高斯的，且有：

$$E(\boldsymbol{y}\mid\boldsymbol{x})=E(\boldsymbol{y})+C_{yx}\cdot C_{xx}^{-1}(\boldsymbol{x}-E(\boldsymbol{x})) \qquad (3)$$

$$C_{y|x}=C_{yy}-C_{yx}\cdot C_{xx}^{-1}C_{xy} \qquad (4)$$

这里若 y 是待估计参数 θ，$\boldsymbol{x}$ 是数据矢量，（3）式就是 Baysian 估计，（4）式就是估计方差的表达式。

贝叶斯线性模型

如果观测数据x可以写为：　　x=Hθ+w

其中x是一个N×1的数据矢量，H是一个已知的N×p矩阵，θ是一个p×1的具有先验概率PDFN(μ_θ, C_θ)的随机矢量，w是一个N×1的噪声矢量，具有PDFN(0, C_w)，且与θ无关。它和经典的一般线性模型的区别在于，将θ看作为一个具有高斯先验PDF的随机变量。

如果观测数据x满足上面的模型，那么后验PDFp($\theta|$x)是高斯分布的，它的均值和协方差分别为

$$E(\theta|x)=\mu_\theta+C_\theta H^T(HC_\theta H^T+C_w)^{-1}(x-H\mu_\theta)$$

$$C_{\theta|x}=C_\theta-C_\theta H^T(HC_\theta H^T+C_w)^{-1}HC_\theta$$

为确保$HC_\theta H^T+C_w$的可逆性，那么H不必是满秩的。

贝叶斯线性模型

后验*PDF*的均值和协方差还可以表达为

$$E(\theta \mid x) = \mu_\theta + \left(C_\theta^{-1} + H^T C_w^{-1} H\right)^{-1} H^T C_w^{-1} \left(x - H\mu_\theta\right)$$

且

$$C_{\theta|x} = \left(C_\theta^{-1} + H^T C_w^{-1} H\right)^{-1}, \text{即} C_{\theta|x}^{-1} = C_\theta^{-1} + H^T C_w^{-1} H$$

对于无先验知识的情况，$C_\theta^{-1} \to 0$，因此

$$\hat{\theta} = E(\theta \mid x) \to \left(H^T C_w^{-1} H\right)^{-1} H^T C_w^{-1} x$$

这是一般线性模型的*MVU*估计量

结论：在Bayes线性模型中没有线性知识的时候，MMSE估计量与经典线性模型的MVU估计量有着相同的形式。

贝叶斯线性模型下的MMSE估计量的性能

贝叶斯线性模型下MMSE估计量的性能：

如果观测数据x可以使用贝叶斯线性模型表示，那么*MMSE*估计量为

$$\hat{\theta}=\mu_\theta+C_\theta H^T(HC_\theta H^T+C_w)^{-1}(x-H\mu_\theta)$$

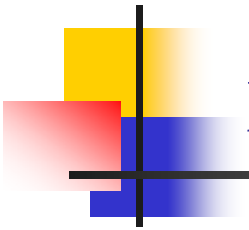
$$=\mu_\theta+\left(C_\theta^{-1}+H^TC_w^{-1}H\right)^{-1}H^TC_w^{-1}\left(x-H\mu_\theta\right)$$

估计量的性能是通过误差$\varepsilon=\theta-\hat{\theta}$来度量的，它的PDF是高斯的，均值为零，协方差矩阵为

$$C_\varepsilon=E_{x,\theta}(\varepsilon\varepsilon^T)=C_\theta-C_\theta H^T(HC_\theta H^T+C_w)^{-1}HC_\theta$$

$$=\left(C_\theta^{-1}+H^TC_w^{-1}H\right)^{-1}$$

误差协方差矩阵也是最小的*MSE*矩阵$M_{\hat{\theta}}$,其对角线上的元素产生最小贝叶斯MSE, 即$\left[\mathrm{M}_{\hat{\theta}}\right]_{ii}=\left[\mathrm{C}_\varepsilon\right]_{ii}=Bmse(\hat{\theta}_i)$



多余参数

- 许多估计问题都是由一个未知参数集来表征，而我们真正感兴趣的只是它们的一个子集。剩余的参数只会使问题变得复杂，我们称其为多余参数。
- 在贝叶斯方法中，我们可以通过对多余参数积分来消除它们。

多余参数

假设要估计的未知参数是θ，出现的附加的多余参数为α。那么如果$p(\theta,\alpha/\mathrm{x})$代表后验$PDF$,我们可以确定$\theta$的$PDF$为

$$p(\theta/\mathrm{x})=\int p(\theta,\alpha/\mathrm{x})d\alpha,$$

或者表示为

$$p(\theta/\mathrm{x})=\frac{p(\mathrm{x}/\theta)p(\theta)}{\int p(\mathrm{x}/\theta)p(\theta)d\theta},\text{其中}p(\mathrm{x}/\theta)=\int p(\mathrm{x}/\theta,\alpha)p(\alpha/\theta)d\alpha,$$

如果进一步假设多余参数和希望的参数是独立的，可有

$$p(\mathrm{x}/\theta)=\int p(\mathrm{x}/\theta,\alpha)p(\alpha)d\alpha$$

从理论角度讲，贝叶斯方法不会遇到像经典估计量的多余参数使一个估计量失效那样的问题。

确定性参数的贝叶斯估计

- 严格来讲，贝叶斯方法只能应用在 θ 是随机变量的情况，然而在实践中，贝叶斯方法叶经常用到确定性参数的估计中。即利用贝叶斯假设获得一个估计量，然后就把 θ 作为非随机变量一样应用。

- 例如，如果**MVU**估计量不存在，我们就不大可能找到一个无偏估计量，这个估计量从方差的角度来说，能够一致地好于其它估计量。然而在贝叶斯框架内，**MMSE**估计量总是存在的，因此它至少提供了一个估计，平均而言（选择不同的 θ 值）它的性能是好的。当然对于一个特定的 θ，估计量的性能并不好，这就是我们把贝叶斯估计量应用于确定性参数的估计所带来的风险。

- 对于确定性参数利用贝叶斯估计量通常是可以的，这是建立在不含信息量的先验**PDF**不向问题增加任何信息的基础之上。

BAYES ex1

The data $x[n]$ for $n = 0, 1, ..., N-1$ are observed, each sample having the conditional PDF

$$p(x[n]|\theta) = \begin{cases} exp[-(x[n]-\theta)] & , x[n] > \theta \\ 0 & , x[n] < \theta, \end{cases}$$

and conditioned on θ the observations are independent. The prior PDF is

$$p(\theta) = \begin{cases} exp(-\theta) & , \theta > 0 \\ 0 & , \theta < 0. \end{cases}$$

Find the MMSE estimator of θ. Note: Two random variables x and y are said to be conditionally independent of each other if the joint PDF factors as

$$p(x, y|z) = p(x|z)p(y|z)$$

where z is the conditioning variable. (Kay: Problem 10.3)

BAYES so1

We should find the MMSE estimator of θ. We first formulate the conditional PDF

$$p(x|\theta) = \begin{cases} e^{(-\sum_{n=0}^{N-1}(x(n)-\theta))} & , x(n) > \theta \\ 0 & , x(n) < \theta \end{cases}$$

The prior PDF for θ is

$$p(\theta) = \begin{cases} e^{(-\theta)} & , \theta > 0 \\ 0 & , \theta < 0 \end{cases}$$

Then we formulate the Bayesian posterior PDF $(0 < \theta \leq x_{min})$

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int_0^{x_{min}} p(x|\theta)p(\theta)d\theta}$$

$$= \frac{e^{(-\sum_{n=0}^{N-1}(x(n)-\theta))}e^{-\theta}}{\int_0^{x_{min}} e^{(-\sum_{n=0}^{N-1}(x(n)-\theta))}e^{-\theta}d\theta}$$

$$= \frac{e^{(N-1)\theta-\sum_{n=0}^{N-1}x(n)}}{\int_0^{x_{min}} e^{(N-1)\theta-\sum_{n=0}^{N-1}x(n)}d\theta}$$

$$= \frac{e^{(N-1)\theta-\sum_{n=0}^{N-1}x(n)}}{e^{-\sum_{n=0}^{N-1}x(n)}\int_0^{x_{min}} e^{(N-1)\theta}d\theta}$$

BAYES so1 cont

Then we can determine nominator of MMSE estimator for θ using the following formula (denominator of posterior $p(\theta|x)$ does not depend on θ):

$$\int_0^{x_{min}} \theta e^{(N-1)\theta - \sum_{n=0}^{N-1} x(n)} d\theta$$

$$= e^{-\sum_{n=0}^{N-1} x(n)} \int_0^{x_{min}} \theta e^{(N-1)\theta} d\theta$$

$$= e^{-\sum_{n=0}^{N-1} x(n)} \frac{1}{N-1} e^{(N-1)\theta} \theta /_0^{x_{min}} - \int_0^{x_{min}} \frac{1}{N-1} e^{(N-1)\theta} d\theta$$

$$= e^{-\sum_{n=0}^{N-1} x(n)} \frac{x_{min}}{N-1} e^{(N-1)x_{min}} - \frac{1}{(N-1)^2}(e^{(N-1)x_{min}} - 1)$$

Now we can write the entire MMSE estimator for θ as

$$\frac{e^{-\sum_{n=0}^{N-1} x(n)} \frac{x_{min}}{N-1} e^{(N-1)x_{min}} - \frac{1}{(N-1)^2}(e^{(N-1)x_{min}} - 1)}{e^{-\sum_{n=0}^{N-1} x(n)} \frac{1}{N-1}(e^{(N-1)x_{min}} - 1)}$$

which reduces to

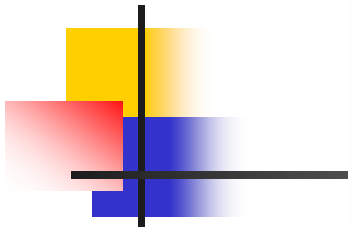
$$MMSE_\theta = \frac{x_{min} e^{(N-1)x_{min}} - \frac{1}{N-1}(e^{(N-1)x_{min}} - 1)}{e^{(N-1)x_{min}} - 1}$$

BAYES ex2

Repeat Problem 1 but with conditional PDF

$$p(x[n]|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \leq x[n] \leq \theta \\ 0, & \text{otherwise,} \end{cases}$$

and the uniform prior PDF $\theta \sim \mathcal{U}[0, \beta]$. What happens if β is very large so that there is little prior knowledge? (Kay: Problem 10.4)



BAYES so2

We want to find the Bayesian MMSE estimator for a parameter θ having

$$f(\theta) = \frac{1}{\beta}, 0 \leq \theta \leq \beta \quad (0 \text{ elsewhere})$$

in the case of sample PDF

$$f(x(n)|\theta) = \frac{1}{\theta}, 0 \leq \ x(n) \leq \theta \quad (0 \text{ elsewhere})$$

i.e.

$$f(\mathbf{x}|\theta) = \frac{1}{\theta^N}, 0 \leq x(0), x(1), ..., x(N-1) \leq \theta \quad (0 \text{ elsewhere})$$

for $\mathbf{x} = (x(0), x(1), ..., x(N-1))^T$. The Bayesian posterior PDF is now (note that $max\,\mathbf{x} \leq \theta \leq \beta$)

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)f(\theta)}{\int_{max\,\mathbf{x}}^{\beta} f(\mathbf{x}|\theta)f(\theta)d\theta}$$

$$= \frac{1/\theta^N}{(\beta^{1-N} - (max\,\mathbf{x})^{1-N})/(1-N)}$$

$$= \frac{(1-N)}{\theta^N(\beta^{1-N} - (max\,\mathbf{x})^{1-N})}$$

BAYES so2 cont

Finally, the MMSE for θ is

$$\hat{\theta} = E(\theta|\mathbf{x}) = \int_{max\,\mathbf{x}}^{\beta} \theta f(\theta|\mathbf{x}) d\theta$$

$$= \frac{1-N}{\beta^{1-N} - (max\,\mathbf{x})^{1-N}} \int_{max\,\mathbf{x}}^{\beta} \frac{1}{\theta^{N-1}} d\theta$$

$$= \frac{1-N}{\beta^{1-N} - (max\,\mathbf{x})^{1-N}} \frac{1}{2-N} \left[\frac{1}{\theta^{N-2}} \right]_{max\,\mathbf{x}}^{\beta}$$

$$= \frac{1-N}{2-N} \frac{\beta^{2-N} - (max\,\mathbf{x})^{2-N}}{\beta^{1-N} - (max\,\mathbf{x})^{1-N}}$$

$$= \frac{1-N}{2-N} \left(max\,\mathbf{x} + \frac{(max\,\mathbf{x})^{N} (max\,\mathbf{x} - \beta)\beta}{-(max\,\mathbf{x})^{N}\beta + max\,\mathbf{x}\beta^{N}} \right).$$

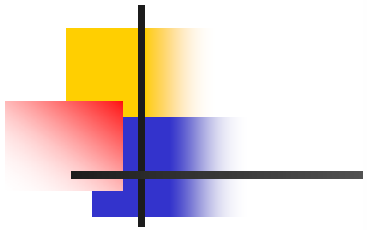
Let now $\beta \longrightarrow \infty$, whence the latter expression in parenthesis tends to 0, and hence $\hat{\theta} \longrightarrow \frac{1-N}{2-N} max\,\mathbf{x}$. Let then β be fixed and let $N \longrightarrow \infty$, whence again the latter expression in the parenthesis tends to 0, since $\beta \geq max\,\mathbf{x}$, but also the coefficient $\frac{1-N}{2-N} \longrightarrow 1$. Thus we get closer and closer the "classical" estimator $max\,\mathbf{x}$, when the sample increases.

BAYES ex3

If $[x\,y]^T \sim \mathcal{N}(0, C)$, where

$$\mathbf{C} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

let $g(y) = p(x_0, y)$ for some $x = x_0$. Prove that $g(y)$ is maximized for $y = \rho x_0$. Also, show that $E(y|x_0) = \rho x_0$. Why are they the same? If $\rho = 0$, what is the MMSE estimator of y based on x? (Kay: Problem 10.12)



BAYES so3

We have the assumption

$$\mathbf{x} = [xy]^T \sim N(0, \mathbf{C}), \quad \text{where}$$

$$\mathbf{C} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Hence

$$f(\mathbf{x}) = \frac{1}{2\pi|\mathbf{C}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-E(\mathbf{x}))^T \mathbf{C}^{-1}(\mathbf{x}-E(\mathbf{x}))}$$

$$= \frac{1}{2\pi|\mathbf{C}|^{1/2}} e^{-\frac{1}{2}\mathbf{x}^T \mathbf{C}^{-1}\mathbf{x}}$$

$$= \frac{1}{2\pi|\mathbf{C}|^{1/2}} e^{-\frac{1}{2}\frac{1}{|\mathbf{C}|}(x^2-2\rho xy+y^2)}.$$

Define now

$$g(y) = f(x_0, y) = \frac{1}{2\pi|\mathbf{C}|^{1/2}} e^{-\frac{1}{2}\frac{1}{|\mathbf{C}|}(x_0^2-2\rho x_0 y+y^2)}. \quad (*)$$

Furthermore

$$g'(y) = \frac{1}{2\pi|\mathbf{C}|^{1/2}}(-\frac{1}{2}\frac{1}{|\mathbf{C}|}(-2\rho x_0 + 2y))e^{-\frac{1}{2}\frac{1}{|\mathbf{C}|}(x_0^2-2\rho x_0 y+y^2)} = 0$$

$$\Leftrightarrow -2\rho x_0 + 2y = 0$$

$$\Leftrightarrow y = \rho x_0.$$

BAYES so3 cont

It is easy to check that this zero point is a maximum (by noticing that the argument of the exponential in (*) is a parabola with a negative second order term). By using Theorem 10.1., we have

$$E(y|x) = \underbrace{E(y)}_{=0} + \underbrace{\frac{Cov(x,y)}{Var(x)}}_{=1}(x - \underbrace{E(x)}_{=0}) = Cov(x,y)x = \rho x.$$

These values are the same, since when x_0 is fixed, we get a new PDF $f(y|x_0) = g(y)$, and thus the expected value of $y \sim g(y)(E(y|x_0))$ is $argmaxg(y)$ (See, for example, Kay, Fig. 10.6 and 10.7). Let then $x = x_0$ be fixed. Hence $E(y|x_0) = \rho x_0$, and since it is the MMSE of y based on x_0, for $\rho = 0, E(y|x_0) = 0, x_0 = 0$.

谢谢大家