module 1感觉关键信息不是非常多，就摘抄出两个方便理解，剩下的内容可以照葫芦画瓢

https://www.zhihu.com/people/chenxran0916 这个知乎作者的文章可以作为补充资料看看，讲了先验后验，共轭，后验预测，蒙特卡洛里的好多东西 变分推断 基本上大体和我们课程一致

## 方差已知的单变量正态

首先求似然函数，按照条件这里方差已知，需要估计的只有theta

图示, 示意图

描述已自动生成

然后考虑下面这个等式

文本, 信件

描述已自动生成

这样做的目的就是去除求和号。因为后验的结果始终是先验乘以likelihood。如果似然带着求和号，那么和先验乘起来的时候没有办法把两个式子融合在一起。所以在考虑到需要去除求和号后，很自然的：**‘样本’**，**‘求和’** 这两个词就会引导出**样本均值**这个概念，再考虑到原式中有n个theta 的平方求和，就首先构造出 这一项，随后再去补齐后面的项。

上面这个式子是处在exp内部的，其中后半部分是常数且由于方差已知也是常数，exp(常数)还是一个常数，所以可以直接舍去。所以就有

图示

中度可信度描述已自动生成

这里也就意味着这是一个均值和样本均值，方差为的正态

随后

文本

描述已自动生成

同上，这里直接整理出和theta相关的部分即可，因为其他都是常数，化简到最后都会变成分布前面的系数。这里先把**前面的负二分之一给忽略了**

图示, 示意图

描述已自动生成

这里的C就是没有用的常数部分（作者估计懒得写了）

文本

低可信度描述已自动生成整理完以后这个就是后验，也就是。将式子和常规的正态做对比

表格

描述已自动生成

通过这样的对比可以发现后验分布的方差和均值就出来了。注意对于前面那个式子，因为我们估计的参数是theta，所以变量就是theta（对应的就是后面的x）。

所以，得出结论：

图片包含 文本

描述已自动生成

有时候，会把方差写成 ，后面那个符号称为precision，利用precision可以让公式看起来更加简洁：

文本

描述已自动生成 这幅图来自老师ppt

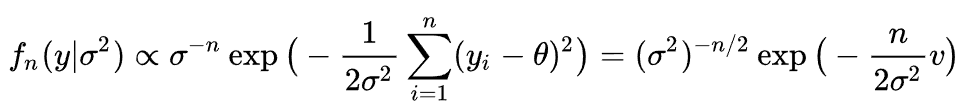
**《知乎百度B站翻一翻共轭先验这个概念》 L2 ppt25页**

[**https://zhuanlan.zhihu.com/p/401258319**](https://zhuanlan.zhihu.com/p/401258319)

**这个知乎更新了非常系统的全面的贝叶斯统计内容值得一看**

## 方差未知的单变量正态

首先求似然函数，按照条件这里均值已知，需要估计的只有sigma。作者这里突然把变量弄成了y不是x。



这里的v被称作 **充分统计量（这概念没什么用而且绕口），**这个v= ( ) 虽然看起来和sigma无关，但是因为前面有一个sigma, 所以还是有关系的，**这一点和前面有所不同需要注意**。最终我们得到了下面这个式子。其实这一块相对来说还简单一点，因为我们关心的参数sigma不在求和号里面，光秃秃的暴露在外面让数学处理变得简单。

文本, 信件

描述已自动生成

根据似然函数的形式‘找到了’ sigma 的先验是逆伽马分布。

图片包含 游戏机, 动物

描述已自动生成

通过 ‘’后验 正比于 先验 \* 似然 ‘’ 这一 永恒不变的规律，可以得到

文本, 信件

描述已自动生成

别忘了 v 就是上面的那个充分统计量

课本上 alpga = 0，beta=0

## Module2

L4 ppt 第三页里写着的

文本, 信件

描述已自动生成

这里指的应该是这两个概率是同分布族的，这也就意味着先验分布是共轭先验。理解这个地方首先先看第二行，第二行是一个全概率公式。等式左边的意思就是由已知的数据y得出新的数据y-tilde。

放到传统机器学习背景下，左边相当于我们用数据训练出来的模型。每一个不同模型对相同的输入会产生不同的输出。这就相当于贝叶斯环境下产生的不同的后验预测分布。等式右边的积分号内部，首先代表的是参数的后验分布，代表的是似然函数。所以后验预测函数实际上是似然函数和后验分布的乘积。通过基础统计学的原理去理解上面那个if，可以知道这个条件相当于新数据和原数据的独立的。当然这也符合传统机器学习的假设：测试集和训练集的数据独立同分布。

第四和第五页讲述的是同一件事，如何采样出合理分布的过程。通过将正态拆分成两个部分，参数方差为什么要除以n可以参考L2的ppt也可以参考上面关于方差已知正态的内容。然后y\_tilde的分布和原来的分布是相同的，因为之前做过这个假设：测试集和训练集的分布是相同的。这里的参数是参照L2第三页的数据来的。

后面的采样过程的理解是交替采样。采样一个theta，再采样一个y\_tilde。也就是说后验预测分布是在不断更新的。

第7页ppt里提到了 Law of iterated expectations指的是迭代期望法则。现在的情况是，如果我不能知道这个模型的具体信息，该如何获得新数据的期望呢（比如在上一个环节，我们知道error的类型）。这个时候可以用这个期望迭代法则。我们要求的期望y\_tilde实际上是conditional on 参数theta和观测数据y

图标

描述已自动生成首先先conditional图片包含 图形用户界面

描述已自动生成 on theta，和y，这个y就先一直带着，也就是说E(Y|theta,y)看成是一个theta 的函数，这样就可以得到

这时候再conditional on 观测y就可以得到期望。

底下的全方差法则是同样的道理，具体的证明过程一搜一大堆 大概也不是很需要知道。

接下来时间序列预测。给的这个公式没什么营养。对每一个时刻的y减去一个mu再乘上一个系数加起来，最后加一个小小的正态error。可以先回顾一下L2当中提到过一次的时间序列的贝叶斯分析。

文本

描述已自动生成

这上面也没啥意思，倒数第二个意思是假设我们只考虑最近的一个时刻（也就是AR1模型），那么我们可能会得到这样这样的mu。

第14页N-1的意思是N分之一，搞不懂为什么写指数-1，我还以为是什么我看不懂的矩阵的逆。

L5

线性模型预测，ppt第三页讲述了常规的多变量线性模型的矩阵形式。顺便还写了一下似然函数。稍微留意一下，协方差矩阵是一个对角矩阵，也就是说每一个y是独立的，这也符合机器学习的常规，每一个样本都是独立同分布的。

这里提供的是non-informative的prior的求解情况，结果看看就好，这里的方差指的是error服从的均值为0方差为sigma（这个sigma又服从一个逆卡方分布）的情况。

可以参考一下百度百科的‘贝叶斯线性回归’，里面提供了不同先验的等价形式，如正态先验或者使用最大后验方法求解的时候等价于L2的岭回归，使用拉普拉斯先验等价于L1的LASSO回归等。在现在的语境下，共轭先验也就是正态先验。

第五张ppt里面，用了共轭先验，对比先验和后验，可以发现他们属于同一分布族，只是参数不同而已。（提一下 只是为了复习一下共轭先验的概念）。后验这些参数的计算还是通过 似然\*先验 出来的

多项式回归的模型在第6张ppt也写出来了。下面的spline 回归就是样条回归。意思就是在不同的区间内用不同的多项式（老师ppt上只是一个单独的p次方）去拟合，使得每一个knot在交点处的曲线是平滑的（导数是相同的）。

为了防止过拟合 引入了一个penalty lambda。根据不同的beta的先验 可以得到不同的结果，正态先验相当于岭回归，拉普拉斯先验相当于LASSO回归。

接下来的问题就是如何选择lambda。也就是第12页ppt的标题，shrinkage的估计。lambda设置他就服从逆卡方分布。

然后下面那个hierarchical 就是总结了从前面开始的一系列操作：

首先得出样条回归的基础模型 参数beta和数据X， 方差是正态error的方差

然后 beta又是conditional on 方差和lambda的（这步就是上面提到的shinkage）这里添加了一个penalty lambda，接下来就是error方差和lambda的采样。

注意一下第一行和第二行n和m是不一样的。n代表总共有n个样本，因为这个计算结果是一个向量嘛。然后m代表样条回归中总共有m个区间。

对比第五页中共轭先验下的beta，这里我们得到mu0=0, Ω0=lambda\*Im

这里相比于共轭先验的多了一个lambda需要采样

13页做的事情和第五页是一样的通过 似然\*先验的手段计算出每一个参数。

14页：

之前讨论的所有东西都是建立在我们的knots是平均选取的情况下，而实际上呢，我们并不知道knots应该选在哪里（我也不是很想知道）。

所以未知变量中还可以再加入每一个knots的位置作为未知变量。然后给k族找一个先验再去建立hierarchical setup 再 似然\*先验去计算。

看看就好

L6

贝叶斯推断中的分类问题。太好了 第三页才开始第11页就结束了。

这里的argmax的表意是最大后验概率。最大后验概率和极大似然有着非常密切的联系，只是增加了对先验的考量。这里提到了一个词 朴素贝叶斯，朴素贝叶斯在计算后验概率的时候是假设每一个变量之间都是独立的，不然麻烦死了。所以某些贝叶斯模型会考虑变量之间的协方差。比如在某一个任务中 明显有关的两个变量 年收入和平均年旅游开支 ，是可以加入协方差变量去考虑的，这个时候就不是朴素贝叶斯了。可以参考西瓜书或者李航统计学习，里面数学过程讲的比较详细。

第4页逻辑斯蒂回归，模型和普通机器学习背景下的模型是一致的。然后底下的Probit模型是正态模型。那个大写的 意思是标准正态，然后输入的参数就是x\*beta。

第五页大样本近似后验使用泰勒展开去近似。

根据后验众数(posterior mode)的定义，就是后验概率最大的点，后验概率最大的点对应的x，也就是一阶导数为0的点。所以这里不能只考虑到一阶导数要考虑到至少二阶导数。这也一来，泰勒展开项当中只剩下了0阶导和2阶导的项。二阶导项反应了观测在后验众数的信息。看一下exp里面的项，将它和常规正态对比，可以发现他能写成正态分布的形式。其中

J(theta\_tilde) 作为方差，theta\_tilde 作为均值。

第6页的例子poisson model 首先给出的theta的后验是gamma先验乘以泊松似然得出的结果，后验也是gamma形式（牢记共轭先验）。这个结果在第一学期统计课的贝叶斯推断习题里经常能看见。

先求一下泰勒展开的一阶导，从这个例子可以发现在求导的时候也是不需要考虑系数的只需要考虑函数的形状就可以了。一阶导的结果就是后验众数。接下来将后验众数theta=theta\_tilde 带入二阶导的式子得到J(theta\_tilde)。接下来根据之前的推到，我们就可以写出在这个模型中参数theta服从的分布。

这里有一个很有意思的点，因为alpha和beta都是常数，所以最终theta服从的分布是和观测信息有关的（这是废话），而且随着观测信息的增多，后验也更倾向于‘信任’来自观测的信息而非先验。考虑n趋于无穷大的时候，此时alpha和beta是可以不用考虑的，这时候后验分布就完全取决于观测的信息。而且由于在这里使用的是后验众数去做估计，也就是后验最大的点做估计，最大后验在某种程度上就是 极大似然+先验。如果观测信息非常多导致先验的影响降到很低，此时这个 贝叶斯大样本估计 感觉和极大似然就没什么区别了。

第8页将模型从低维升级到了高维，在高维的条件下，我们寻找后验众数的求导过程就变得比较操蛋，因为确实不太好求。此时就需要借助计算机去做，老师ppt里说常规 的做法是使用optim.r去做，相信都已经非常熟悉这个了。optim输入的参数如果是一个向量，那么会返回最优的向量同时也会返回hessian矩阵，也就是我们要的theta和J。那个re-parametrization先暂时不管他好了，下一张会碰到。这里重尾分布的概念感觉不是非常重要。这里需要选取t分布的自由度。

第9页就是加了一个log将正态变成了对数正态，感觉没啥好说的，就是一个技巧。对比第10页和第7页的图，可以看到放了log以后theta的取值不在可能就0以下了。

第11页举了一个例子，想说的是虽然theta的分布是正态的，但是G(theta)不一定是正态的。前面讲了大半天，讲的是模型的参数theta的东西，模型怎么样还没讲呢。

看看Gini的例子。一直到第4行应该都没什么问题，这个模型一个两个参数mu和sigma,将mu和sigma打包成一个二维的theta，利用上面讲的各种方法去估计，最后可以得到theta 的分布，接着采样theta，采样了theta就可以去计算sigma然后再去计算Gini系数完成最终的采样。

## Module 3

L7 ppt有28页 慢慢来

第三页讲的是大数定理，利用样本均值来估计总体均值。

第四页 inverse CDF 计算统计学过了。这里提到一个概念 Direct sampling直接采样 概念，

暂时先记一下这个概念，到后面碰到不同的概念对比起来讲才有意义。

第五第6页 吉布斯采样 如果想不起来可以看看

<https://cosx.org/2013/01/lda-math-mcmc-and-gibbs-sampling> 这个网址我也贴在了计算统计的东西里。

第9页 提到了gibbs采样是一个dependent draws，他和直接采样是有区别的，因为这里每一个theta是基于上一个theta来的，因为采样的时候需要全条件概率，这个全条件概率是在不断更新的。而像inverse cdf采样过程每一个iteration之间是独立的。

对于iid的samople，均值的方差就是原方差除以N，挺好理解的，但是对于自适应的样本，为什么要乘上一个inefficiency系数暂时还没想明白，在网上我也找不太到，就记者就好了。这里的sigmak 举个例子说明 sigma1 就表示所有的采样中 t和t-1的相关系数，但是很奇怪的是，难道所有的t和t-1的相关系数是一致的吗？

第10页给了吉布斯的一个例子，首先给出了theta1和theta2的联合分布，接下来分别求条件概率，采样的时候先给一个初始值(t1,t2)

将t2代入theta1的条件概率，采样出新的t1

将新的t1带入t2的条件概率采出下一代t2。。。。

这也循环往复 不断带入新的（同计算统计内容）

第16页和L2第四页联动

在L2第四页，这个theta是未知的，方差sigma是已知的，我们有theta的先验就是那个mo0，tao0^2的正态，可以算出theta 的分布。

在L7 16页 这里我们并不知道sigma（最底下那行字说的就是如果sigma知道那么和L2就是一样的）。所以只能用吉布斯的方法去采样，用最新的sigma去更新mu，用最新的mu去更新sigma。sigma里面的vn是自由度，这个可以看看L3第4页，选取不同的sigma先验会有不同的后验，这里并没有给出先验到底是啥，只说了是共轭先验（共轭先验能保证后验是同一分布族，但是每一个参数的数值是多少还是得看先验的具体数值，这里我估计是没有算出来直接就写在这了）。

17页 这里将吉布斯应用到AR的过程中。这里涉及的由三个参数，mu和sigma是1维的，而φ是高维的。采样就是和普通吉布斯一样，先采mu，再一个一个采φ，最后采sigma。

18页讲了一个数据增强方法。

类似于高斯混合模型，样本是来自两个不同高斯的混合，我们并不知道他是哪一个高斯，所以添加一个隐变量(latent variable)来指示。采样的时候就以概率Π来从第一个高斯采样，以1-Π从第二个高斯采样，这也非常合理。

19页开始讲吉布斯在这上面的采样。首先再怎么吉布斯也逃不开求后验，求后验就逃不开先验\*似然这么一个过程，而这个概率它是和的形式写出来的，所以最后似然的形式长得就像：

(...+...)\*(...+...)\*(...+...).......\*(...+...)。

这种形式基本上是没办法操作的，所以必须要设置这个隐变量来指示我们的样本来自哪一个模型。通过引入未知变量Is（indicators，假设我们知道每一个样本是来自哪一个高斯），我们相当于增加了参数的维度，减少了问题的复杂程度。然后现在的问题是我们并不知道每一个样本的标签是啥，所以设置另一个参数Π，通过采样Π，我们可以设置当前从哪个高斯采样。

如同21页中讲述的采样过程，先由一个n1和n2，采样Π,采样Π的时候其实只需要用到indicator就可以了。然后再根据是l1还l2去采样sigma1和sigma2，mu1和mu2，这个采样的过程就是和前面未知方差是一样的。需要注意的是，采样sigma1和sigma2，mu1和mu2的时候肯定是需要观测的，这个观测就是l1和l2来告诉我们的，比如总共5个数据，

indicator向量是(1,1,1,0,0)，说明 前三个数据是indicator为1的观测，为0的是indicatior为2的观测。通过这样将原来的混合数据分开采样。

这看起来已经结束了呀，我们有了两个高斯的全部采样，但实际上并不是这样，因为一开始设置的n1和n2并不一定是真正的高斯混合的比例，这时候我们要通过底下的公式去采样新的l，然后更新一下n1和n2，当然这个过程中I本身的分布也是会变的（比如从上面的1,1,1,0,0 -> 1,0,0,1,1），所以我们可以看到最后采样的I是有下标的，因为他是n个indicator一个一个独立采样采出来的。

第23页probit regression的数据增强。前面讲过probit ，那个φ就是标准正态。后面那些吉布斯都大同小异，总体过程都是一样的，理解了上面的吉布斯过程，对于后面的理解也不是非常困难。

L8

L8 这里讲的是马尔可夫链的一些东西，建议直接看下面的链接 看完了再来过一遍PPT

[LDA-math-MCMC 和 Gibbs Sampling | 统计之都](https://cosx.org/2013/01/lda-math-mcmc-and-gibbs-sampling)

直接从第7页开始 第七页讲的这个rejection sampling 和计算统计里的一摸一样，target density和proposal density，然后还有一个常数C。

第8页 random walk。首先有一个初始theta，接下来基于初始theta采样，采样从高斯里采，均值是上一个theta，方差是c\*协方差矩阵，这个c和协方差都是自己选的，简单点可以直接选一个sigma\*单位矩阵。sigma也自己选。

然后这个接受率就是上一次采样的theta对应的概率密度取值和这一次采出来的theta对应的概率密度取值的比值和1中取最小值。

意思就是 如果这次采样 采出来的概率比上一次大 就直接接受；如果这一次采样的概率比上一次小，那么就以概率alpha接受，如果被拒绝了就保持当前theta。从主观上理解，这样的做法能让概率密度大的点更多的出现在采样链中，让概率密度小的点出现的次数变少，河里！

第9页讲的东西是基于第8页的。第8页中我们需要计算后验概率密度。但是有些时候如果先验不是共轭先验那么后验会长的乱七八糟甚至可能不能算，这个时候可以用贝叶斯公式将后验的比值转换为 似然\*先验的比值。因为先验和似然我们肯定是有的，这样相当于直接做数值计算。

第10页果然讲了通常的协方差的取值，要么选择单位矩阵 要么选择之前在L7里面接触到的泰勒展开的二阶导的那一项。然后设置C的时候 一般最好是能让接受率维持在25-30%之间。

如下图，我们的后验分布就是那个蓝蓝的圈子p(theta|y),然后红圈圈的中心就是我们每次采样时上一次的theta，方差小的时候 他确实每一次都能走在里面，但是想完整的走遍整个C显然需要很长的时候而且很容易走不到边缘区域因为步子太小了导致最后的分布和原来的分布不太一样。如果C太大，也不行，经常走出去也就经常被拒绝。

文本, 白板

描述已自动生成

第11页开始正式讲MH算法

MH算法很不一样的一点在于他的proposal distribution不是symmetric的，对称的意思就是q(x|y) = q(y|x),前面讲的东西是基于MC的，而MC有一个要求叫细致平稳条件（detailed balance），这一点可以回顾我之前放的一个讲MCMC的网站。细致平稳条件要求两个状态转移的概率是完全相等的，然而如果我们现在不想要（可能不对称的能更符合后验ppt11页也说了 这是一个更general的情况，适用于所有非对称的proposal）这样那该怎么做呢（其实最好还是要对称，这样能少算很多东西）。这里这个q就代表我们选的那个proposal，因为是任意的一个非对称，所以参数是啥也不好写出来，就用 ‘·’来表示。

所以MH算法和MCMH算法还是不太一样的，基于MC的都是对称的，因为细致平稳条件.所以对比之前的接受概率里，后面多乘了一个东西，可以看到如果proposal是对称的，那么这一项就可以删掉，那就变成前面一摸一样的东西了。

第12页就更加general了，现在用independence sampler。也就是说 我们的采样和上一次的结果已经没关系了。用老师的话说，不在locally走了要globally走，放在下图就是不考虑红圈圈的束缚了，每次都在q的分布下乱走。

徽标, 公司名称

描述已自动生成

这里举了一个例子，多变量t分布。我们有了后验以后，通过optim.r计算出后验众数以及hessian矩阵（和L6第5页里大样本估计是一样一样的），然后带入多变量t分布里面去采样，就是一个例子。这个效率会非常高，因为每一个iteration完全是独立的。

这里有一点关于尾部的概念，我们要保证选取的propoosal的尾部要比后验的要厚。这句话的含义可以由重尾分布这个词来，重尾分布的定义就是尾部比指数分布要厚的分布。我们知道指数分布是exp(-lambda\*x)，而下面的式子将概率取了倒数，然后给了这样一个数学定义，所以我们只要照葫芦画瓢就可以判断我们的proposal是不是要比后验厚。

图示

描述已自动生成

所以proposal比后验 厚的时候，

theta0 -> ∞时 (p(theta0|y)^-1)\*Proposal[theta>theta0] = ∞,从几何层面想想这个数学定义是符合直观感受的。

接下来讲了Gibbs过程中可以怎么用MH，Gibbs需要全概率分布，有时候全概率分布中某一步可能不是非常好求，这个时候对于那一步我们可以用MH。

接下来那个efficiency和之前讲的差不多，之前是在时序里提到的，现在这个每一轮相当于一个时序，所以问题不大 差不多。

15页收敛性验证，记得在计算统计里用过一个包，这里讲的差不多是类似的东西

L9

L9主要考虑的是高纬度的参数采样，这时候如果用MH算法不太好去做。可能会考虑用吉布斯采样，但是吉布斯采样有一个条件就是全概率后验必须要比较好表示出来，否则没有条件后验，哪来的吉布斯。

介绍了MHC算法，在后验中添加了一个动量来控制采样当中的步长。

<https://bocaiwen.github.io/Hamiltonian-Monte-Carlo.html>

网站讲的比较清晰，从对这个动力系统的解读到将该动力系统与概率采样的结合.看完这个再回来看ppt第三页底下MHC加入了一个动量参数，这个参数和数据y是独立的，所以采样可以用后验乘以动量的分布来做。

第四页具体讲了这个MHC算法当中的一些细节，前两点可以在网站中直接找到，网站中甚至讲的更加的具体。

第四页底下首先设置 了一个U（theta），这个theta字啊这里就是代表着位置，theta是我们要采样的参数，log里面的东西就是theta的先验乘以似然，所以这里也就是对后验取了一个minus-log。取minus-log的原因在那个网站里面有提到一些，原因是所有的分布都可以写成A\*exp(log(…))的形式。这样方便数学处理。

mass matrix是质量矩阵，在这里这个质量矩阵是要我们自己选的，我们会在stan里面使用默认的设置。如同网站里提到的那样，这个K函数代表着速度的函数也要写成minus-log的形式。看到这里如果不了解为什么要这样写 就需要把网站重新看一遍。后面那个等号并不重要，他是分析力学里的一个公式（<https://zh.wikipedia.org/wiki/%E8%B4%A8%E9%87%8F%E7%9F%A9%E9%98%B5>）

第五页theta对t的导数可以直接从这个公式得到，*φ*对t的导数等于U对theta的导数，也就是后验对theta 的求导。得到这个导数的目的是为了使用接下来的leapfrog算法。

从这个算法可以看到他是用离散的手段去迭代。epsilon是我们要自己选的参数，步长，stan里也用默认的就好了。就像randomwalk 里面选的C一样。然后算位置theta的东西，走一整步。最后再走半步*φ*。如果放到一个循环里去考虑这几步的话，其实就是第一次先走半步，之后就是theta更新一整步，*φ*再走一整步。

第6页的采样里面，这个接受率和MH算法是差不多的，因为离散化之后会有误差，毕竟原值+导数\*位移只是新值的一个近似，所以有可能会走出hv等值线样本不一定是能被接受的。这里就以概率去接受拒绝。在MCMH中我们只需要考虑后验，在普适MH（theta的转移不一定是对称的）中我们需要考虑两个状态转移的概率的比值。在这里前面的网站中曾提到这个方法是要求细致平稳条件的，也就是说在这里theta的proposal的选取必须是对称的，所以只需要考虑后验的比值就可以了。前面在第三页的时候提到了这个算法中的后验是写成theta的后验乘以*φ*的分布的，所以这里当然也要考虑*φ。*

那个NO U-TURN是stan里的一个设置，可以直接参考stan这个函数的说明文档就可。

## Model 4

L10

P3 明确以下nested model 和non-nested model。中文翻译为嵌套模型和非嵌套模型。嵌套与非嵌套是针对一组模型来说的。比如我们有一个回归模型A

Points = β0 + β1(minutes) + β2(height) + β3(position) + β4(shots) + ε

又有一个回归模型B

Points = β0 + β1(minutes) + β2(height) + ε

其中 beta0，1，2是相同的，也就是说 **模型B的系数向量是模型A中系数向量的一部分，那么这个时候就可以说模型B是模型A的嵌套模型（B is nested in model A）.**

P4 讲了一种基于先验的比较方法，通过对先验和似然的乘积求积分

文本

中度可信度描述已自动生成

可以得到一个叫边缘似然的东西，这是针对第K个模型Mk和对应的参数thetaK。

这样我们可以计算贝叶斯因子B。

P5 给了一个例子，这个例子有点特殊。假设检验可以看作是对两个模型的选择，只是在这里我们的M1和M2是同一个模型，只不过参数不同。通过计算上面的积分，然后可以分别计算出来两个边缘似然，最后求得贝叶斯因子。伯努利模型的似然在统计课上也求过几百次了。

P6 是一个离散情况的例子。在上面的例子中，Pk(thetak)是先验，是一个连续的函数。那么在这两幅图中，先验变成了等概率的离散情况，即该模型的先验是离散的均匀分布。

P7 给了一个例子，1：几何模型基于beta先验 2：泊松模型基于gamma先验。

首先根据公式计算两个边缘似然分布。P8的图就是给大家看看两个分布长什么样。

P9中当进行两个比较的时候，这里提到了一点就是两个模型的期望值得是相同的。如何保证两个模型的期望是相同的呢？那当然得从先验的参数入手。这里用到了之前提到的 期望迭代法则。

首先对于模型M1，他是基于beta先验的几何分布.

老师说这里我们用的是 “失败i-1次，第i次成功”这样的模式，

所以几何分布的期望为

文本

描述已自动生成

接下来就要求 E()。这个不能直接计算E(theta)再求倒数。

手机屏幕截图

描述已自动生成

这是从维基百科上找到的 [被称为 Harmonic mean](https://en.wikipedia.org/wiki/Beta_distribution#Harmonic_mean)。所以是。

对于Model2，如法炮制求出。这里我下标都没有标，但是可以看出来如果要求这两个期望相同是可以得到ppt上的等式的。

后面几页看看就好 没啥特别的，12页给了一个例子，是一个研究人员通过这种方法在100个不同的参数模型中选择出最好的10个（有点类似于网格搜索调参）。

P13 讲了一些不痛不痒的东西，贝叶斯因子的计算具有传递性，很好理解因为他是除法定义的。如果我们定义的模型中有真实符合数据分布的模型，那么当观测充分大时，我们总能找到那个真实的模型（概率为1）。如果我们很不幸并没有将真实模型囊括在我们的备选list中，那么当观测足够多时，我们总会选中KL散度最小的那一个。这个‘选’的意思就是利用贝叶斯因子去判断哪一个更好哪一个更差。

奥卡姆剃刀法则：我们倾向于选择简单的模型。

如果先验比较次，那么是没办法拿来算贝叶斯因子的。

P14讲的东西就是条件概率公式，第二条式子乘积部分会让前面是yi和theta，后面是y1到yi-1，再把theta积分积掉就只剩yi了。后面讲的两句话指的是这个out-of-sample 的预测是首先基于先验来做的，所以一开始的样本会极大程度的受到先验的影响，而后面的样本则不会。

P15来源于L3 第7页，假设参数符合正态的形式，L3P7中的k0等价于这里的。

简单分析一下。对于theta来说，mu0 = 0, 根据L3P7的公式准备采集第i个样本的时候已经有i-1个样本了，所以这时候把m0=0，n=i-1, y拔 用i-1个样本均值带入即可。方差就更加简单了，直接加上i-1就可以了。对于y的采样在L4P4的正态模型里。对比一下可以发现P15里的yi相当于L4P5里的y\_tilde，其他y1~i-1相当于L4P5里的y，所以简单带入一下就可以得到答案了。

底下几张杂技一样的图就是简单得提了一提这个过程，从第17页的放大图中可以看出来确实是这么一回事，一开始样本数量小的时候差别还是有的，别看曲线那么接近那是因为这个值取了一个log实际上差的还是比较多的。

第18页讲了一个LPS的分数。因为前面我们提到这个采样过程的前期情况并不是非常理想，即不够稳定，所以这里我们需要牺牲一部分前期的数据（当作是以一种训练），然后打分只在后面你那部分去打分。

19页 看看就好。

L11

主要是如何计算边缘似然。第一行给的公式是条件似然\*先验，再把参数积分积掉，这个是最基本的公式，有时候不一定可行，毕竟不可积的式子还是比较多的，尤其是指数带平方在概率模型中也是比较常见。

这个共轭形式的边缘似然看着十分之眼熟，利用贝叶斯公式去推导。但是用这个方法算稍稍有一个小条件，那就是先验得是似然的共轭先验，不然后验长得太难看做起来会很复杂。从下面那个伯努利例子也可以看出来，如果分子分母的形式不统一，这个过程是十分痛苦的。

P4第一行那个公式是变量函数的期望公式，假设theta服从p(theta),然后theta的函数p(y|theta)的期望就是这个积分。

老师提到了确实通常积分计算是比较复杂的。可以参考蒙特卡洛方法去估计这个分布。只是这个方法在先验和后验不同的时候（非共轭先验）会不够稳定。

底下还提到了一个方法，可以找另一个分布g，乘以g再除以g，这样一来变量theta还是不变，只不过这次theta的分布变成了g，用这种方法，我们就可以采样采到我们想要的theta的分布，而并非一定要和先验是一样的。

最后的那个东西是一个cut-off的形式。如果tehta超出某一个范围了，那么取值就取0，也就是说，我们采样只能采到一定范围内的theta。

第五页 首先用之前的泰勒展开到二阶导项，然后两边同时取exp再乘以p(theta),最后得出的结果可以配凑配出一个多变量正态的密度函数，theta\_hat是后验均值，J的逆就是协方差矩阵。

文本

中度可信度描述已自动生成

这时候再回来考虑上面的这个式子，如果在拉普拉斯近似的上下文去计算，后面那部分直接就是1，因为他是一个密度函数。前部分与theta无关，可以直接提到积分号外面，这时候积分结果为电脑屏幕的照片上有字

低可信度描述已自动生成

两边再取一个log，就得到了我们拉普拉斯近似的结果，这个theta\_hat就是后验众数，J就是hassian矩阵，和前面讲的用optim函数获取的是一样的。

BIC近似是对拉普拉斯近似的一种近似简化。他在样本量比较大的时候忽略了

徽标

描述已自动生成 文本

低可信度描述已自动生成

且将协方差矩阵简单近似成了n\*Ip这样的对角矩阵。

P7开始是另一种方法来判断模型的好坏，AIC越小越好。从传统机器学习的视角看，前部分带负号的部分相当于是损失，后面加上一个正则化。后一页DIC里的S表示posterior draw的个数总共有S个。

第9页WAIC老师说比之前的都好一些，问题就在于他的数据得满足一定条件，必须是可以分成n组的数据。好处是这个方法随便选先验。

首先看lppd，我们对于参数theta，会有很多posterior draw,对于每一个draw，我们对第i个观测计算他的期望（密度函数在y=yi处的值 的均值），取log再加起来。后一项也是对每一个观测而言。然后还是和前面一样，对于每一个draw，我们有密度函数在yi点的值，对于每一个样本有S个draw，取样本方差再基于所有的观测求和。

LOO-CV 留一验证法

假设有n+1个观测，每一次对于1个观测，我们取另外n个观测组成subdata，然后基于这个subdata推断出theta，然后基于那个独立的观测用和上面一样的方法得到后验概率均值，就可以了。这个方法不需要penalty，也是越小越好。留一验证法齐硕就相当于是WAIC计算n次这里n是观测总量。所以计算量是非常之大的。

第11页开始是变量选择（选取哪些特征，如这个线性回归的例子，我们选哪些x来做回归），有点类似于数据降维。

第12页描述的和之前做data augmentation的时候差不多，我们先用贝叶斯推断去推断每一个indices，这里对于这个贝叶斯公式的理解可以把X给忽略掉，因为X他不是一个变量，他是固定的观测，是一个fixed的数据。只是客观上因为这些东西都是基于观测得来的，所以始终把X写在被conditional on的位置。去掉X以后这个式子就是一个常见的2变量贝叶斯公式。下面给了一个常规的关于indices的初始化选择，从伯努利模型中选。当然也可以对每一个I做独特的选择，比如我觉得第三个变量百分之200重要，直接拉成1也可以。 后面的东西看起来都不是很重要，就不写了。

## 习题

### 第一次习题

总结一下第一次习题的套路。第一题就是先验\*似然得到后验，没啥好说的。

第二题和第三题讲的都是正态似然配上不同的先验。核心思想都是把不相关的项去掉把正态exp里面的写成参数theta的多项式形式

图片包含 图示

描述已自动生成

比如这个，就是去掉mu\_n二次项以后的，因为这一项不带theta，对于theta的分布来说属于常数。这是标准正态的形式，将所有的都化成这样的形式可以得出均值和方差。

第三题的(b)相当于是3个正态的结合，

计算过程也是一样的，答案是从(a)的结果来的，分开写也是一样的，最终写成

theta二次项里面的就是新的方差，所以需要解方程，这个tao\_n 解出来就是新的方差

同理去解均值，解均值的时候用的方程 “theta的系数就是新均值除以方差”

### 习题二

第一题

有几个东西要记一记，课上有提到，均值已知方差未知的正态的共轭先验是逆卡方分布。

然后这里用到的先验长这样。

然后这个word开头有讲到关于为什么这里写似然的时候不能和之前一样忽视掉一些看起来像常数的东西，因为这里的参数变成了方差，所以似然最终是,这里的theta是已知的均值。

把这两个乘起来可以发现处在分母的项是可以合并的，处在exp里面的项也是可以合并的，且平方和可以用样本方差s代替（这里的样本方差是除以n不是除以n-1，然后减去的数不是样本均值而是期望均值，也就是我们已知的参数theta）。最终得到的结果

所以，这里可以直接根据分母写出新的自由度 v+n

然后解方差的时候，从式1可以看到方差就是exp里分子部分和自由度相乘的项，所以

可以解出新的方差，这个新的自由度和新的方差就是后验逆卡方分布的参数

第二题，通过这题可以更好的理解后验分布和后验预测分布的关系。后验分布顾名思义指的是后验的分布，而后验预测分布是有了后验，又有了结果，再去算分布