1.Basic Idea

Bayesian school: **the parameters themselves are uncertain.**

We first a prior on the distribution of parameters θ. After receiving new data X, we use bayesian formula to update our belief of the parameter, which is posterior distribution of θ

$$p(\theta|\, x) = \frac{p(x|\theta)p(\theta)}{p(X)}$$

The core idea of Bayesian Estimation is to use the above formula repetitively, combining prior distribution with data distribution to get posterior distribution. (Therefore, what you get is actually a distribution of parameter)

1.1 Prior Distribution Choice

A general rules is to use distributions that do not contain too much information.If the prior distribution contains too much information while the data does not have strong information, the posterior distribution may look almost the same as the prior.

A native idea is to use the uniform distribution. But it has some drawbacks. Another two often used prior distributions are:

1. *normal distribution with large variance (vague of diffuse prior)*
2. *(often used in practice) jeffreys prior*

In reality, we may need to try different prior specifications to check robustness. When the data sample is very large, however, in theory the prior distribution can be neglected.

1.2 The Posterior Distribution

After all, the output of the bayesian estimation is to get a posterior distribution of parameters based on prior distribution and data. What information can this posterior distribution reveal?

1. You can get the marginal distribution for a single parameter
2. You can of course get the moments of the distribution.
3. You can get mean of median of the distribution, can treat it as a ‘point estimation’
4. You can also get the confidence interval. But the interpretation is very different from the conventional confidence interval definition.

When we get the posterior distribution, we can treat it as a revised prior and use it as prior for the future.

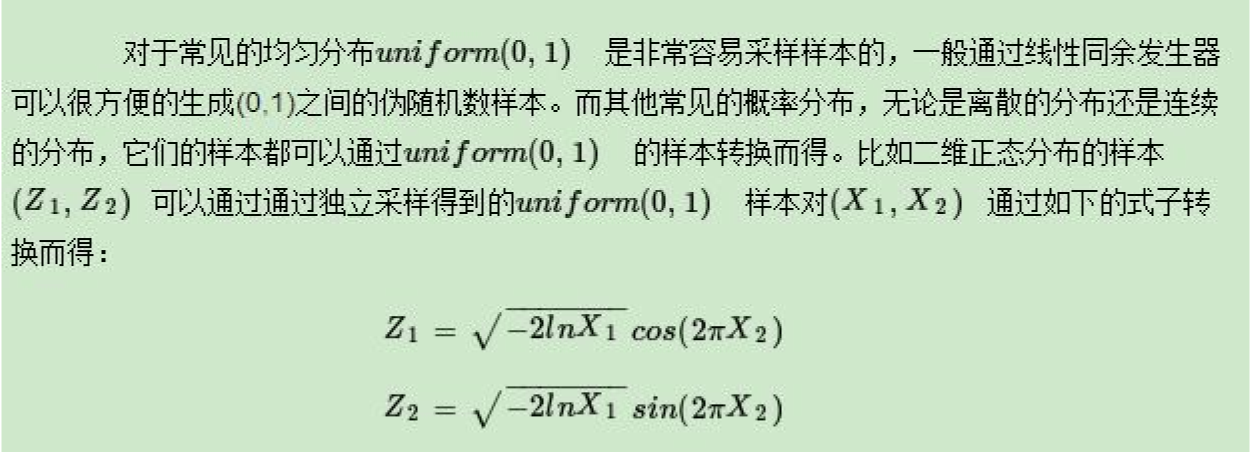
2.key characteristics of bayesian estimation:

1. No need to do optimization. All we have to do is to use bayesian formula.
2. Can the whole posterior distribution directly.
3. Rely on prior, but the dependence on prior declines with the sample size

4. When the data sample is large, typically the posterior distribution of parameter has less bias and variance.

3.Sampling from posterior distribution

Recall that we need to get our conclusion from $p(\theta|x)$. We often need to draw samples from this distribution, and use these samples to plot the distribution or calculate the expectation (using monto carlo), and gain the necessary information. In reality, however, this posterior distribution may be complicated (although it has an analytical density function), making it very difficult to draw samples from this distribution. How to address this problem?
https://zhuanlan.zhihu.com/p/30003899



对于常见的均匀分布$uniform(0, 1)$ 是非常容易采样样本的，一般通过线性同余发生器可以很方便的生成(0,1)之间的伪随机数样本。而其他常见的概率分布，无论是离散的分布还是连续的分布，它们的样本都可以通过$uniform(0, 1)$ 的样本转换而得。比如二维正态分布的样本(Z_1, Z_2) 可以通过通过独立采样得到的$uniform(0, 1)$ 样本对(X_1, X_2) 通过如下的式子转换而得：

$$Z_1 = \sqrt{-2lnX_1}\,cos(2\pi X_2)$$

$$Z_2 = \sqrt{-2lnX_1}\,sin(2\pi X_2)$$

3.1. Method: Rejection Sampling

https://en.wikipedia.org/wiki/Rejection_sampling

The basic idea is very simple: for example, if you want to sample from p(x) but it is very complicated, then you can actually consider a simpler distribution q(x), and a constant k, such that k*q(x)> p(x) always holds. Each time,

1. you simply draw a sample from the q(x). Denote this sample as z.
2. Given this z, you know that the probability of getting it is q(z).
3. Now consider uniform distribution U(0, k*q(z)). Get a sample s from this distribution.
4. If s>p(z), then reject this sample z. Otherwise, put this sample z into your sample set.

One limitation of this method is that it is hard to deal with high dimensional distribution p(x): it will be super hard to find a proper q(x) and k.

3.2. Method: MCMC

Recall that a markov chain. (If it satisfies some property), will finally reach a stable limit distribution, which is irrelevant to initial distribution. **In other words, given our target distribution, if we treat it as a stable limit distribution and recover the corresponding**

markov chain, we can, starting from any initial distribution, draw out the sample, and given the markov chain, draw the next sample….repeat this many times, and cut the initial many samples and keep the remaining. We get a sample that mimics our targe distribution. This is the basic idea of MCMC.

To be more specific, consider p(x), which is very hard to sample directly. Treating it as a limit distribution, suppose that we can work out a markov chain, which specifies q(x’ | x).Suppose this q is easy to sample.

Now, get a sample x1 from any initial distribution. Using the q(), get x2.

Given x2, using the q(), get x3……

Repeat this many times N. also take a large number K<N, assuming that after K times, the markov chain reaches the limit distribution p(x), we can regard x(K+1),...x(N) as a sample set drawn from the distribution p(x). One big challenge, however, is how to get the q(.|.) from p(x)? To address this issues, we first see one proposition

$if\ \pi(i)q(i,j) = \pi(j)\ q(j,i)\ holds\ for\ all\ i,j\ \ \pi()\ is\ the\ stationary\ distribution\ of\ q(|)$

This proposition is easy to verify. Theoretically, using this proposition we can get the q(|) using $\pi()$ .But this is still not easy to find q. Often times,

$$\pi(i)q(i,j) \neq \pi(j)\ q(j,i)$$

But we can do some small tricks .Given $\pi()$, suppose that we just randomly get $q()$. of course,

$$\pi(i)q(i,j) \neq \pi(j)\ q(j,i)$$

But now, we add something at both sides of the equation to make things equal.

$$\pi(i)q(i,j)\alpha(j,i) = \pi(j)\ q(j,i)\alpha(i,j)$$

On one hand, this implies

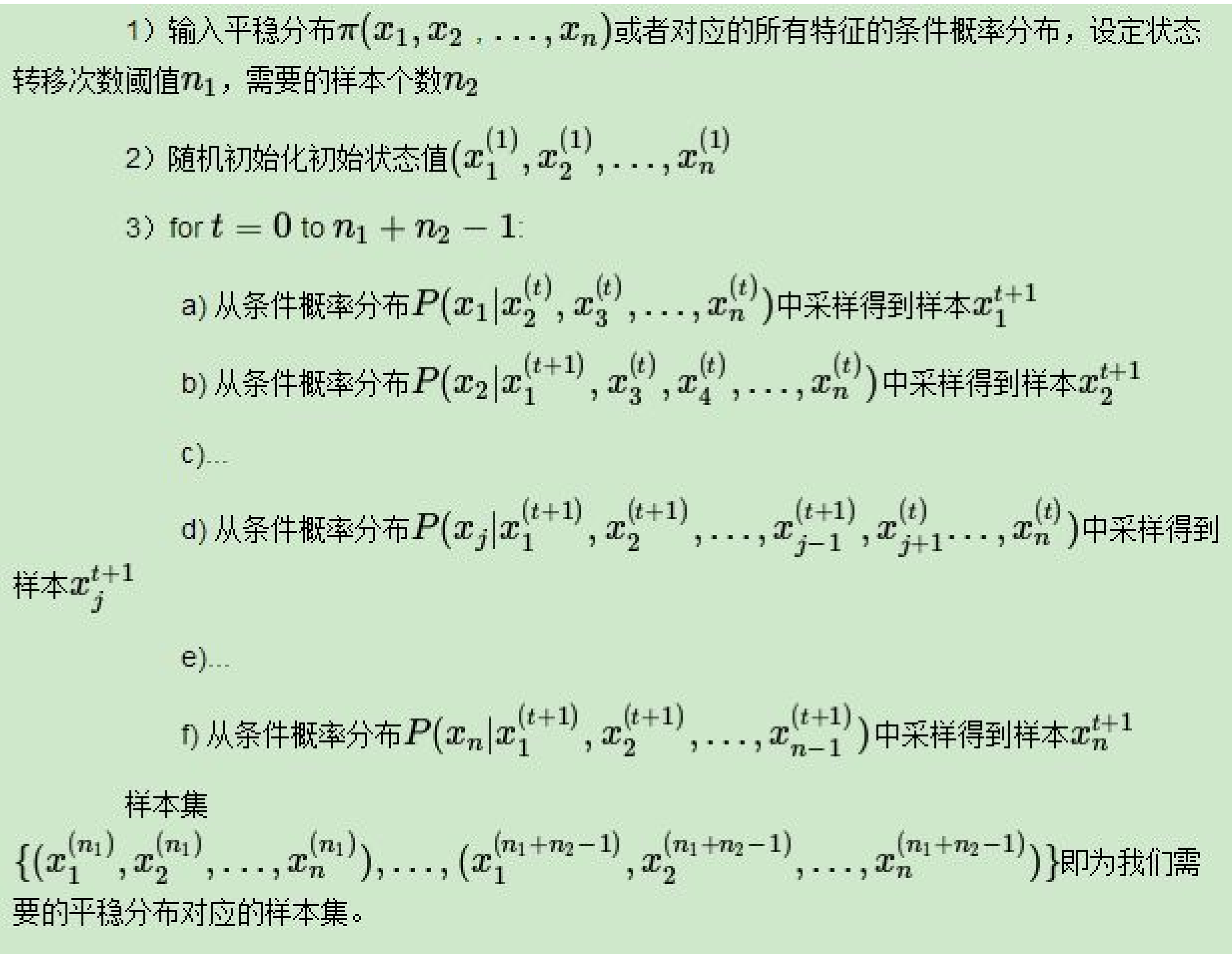
$$\alpha(j,i) = \pi(j)\ q(j,i)$$
$$\alpha(i,j) = \pi(i)\ q(i,j)$$

This implies that, given $\pi()$, we can specify a arbitrary markov chain $q()$, and construct

$$Q(i,j) \ = \ q(i,j)\alpha(j,i)$$
$$Q(j,i) \ = \ q(j,i)\alpha(i,j)$$

Therefore, theoretically, using Q, we can do the repeat sampling.But this method is still not good enough… The currently two most used methods are

1. M-H sampling
2. Gibbis (rolling over each marginal distribution) (mostly used)



```
1）输入平稳分布$\pi(x_1, x_2, \dots, x_n)$或者对应的所有特征的条件概率分布，设定状态转移次数阈值$n_1$，需要的样本个数$n_2$

2）随机初始化初始状态值$(x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)}$

3）for $t = 0$ to $n_1 + n_2 - 1$:

    a) 从条件概率分布$P(x_1|x_2^{(t)}, x_3^{(t)}, \dots, x_n^{(t)})$中采样得到样本$x_1^{t+1}$

    b) 从条件概率分布$P(x_2|x_1^{(t+1)}, x_3^{(t)}, x_4^{(t)}, \dots, x_n^{(t)})$中采样得到样本$x_2^{t+1}$

    c)...

    d) 从条件概率分布$P(x_j|x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_{j-1}^{(t+1)}, x_{j+1}^{(t)} \dots, x_n^{(t)})$中采样得到样本$x_j^{t+1}$

    e)...

    f) 从条件概率分布$P(x_n|x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_{n-1}^{(t+1)})$中采样得到样本$x_n^{t+1}$

样本集$\{(x_1^{(n_1)}, x_2^{(n_1)}, \dots, x_n^{(n_1)}), \dots, (x_1^{(n_1+n_2-1)}, x_2^{(n_1+n_2-1)}, \dots, x_n^{(n_1+n_2-1)})\}$即为我们需要的平稳分布对应的样本集。
```

Remark

1. How to judge that resampling iteration has converged?
   http://web.sfc.keio.ac.jp/~maunz/BS14/BS14-11.pdf
2. If the independent variables are highly correlated , then posterior coefficients may be highly correlated, and this may lead to a much longer time to converge. Try to regress them and get residual to get orthogonal components of these variables !