3. ベイズ推定と機械学習

植野真臣
電気通信大学　大学院
情報理工学研究科

16. ベイズ原理

定義15 (事後分布)
$X=(X_1, \cdots, X_m)$が独立同一分布$f(x|\theta)$に従うn 個の確率変数とする. n 個の確率変数に対応したデータ$\boldsymbol{x} = (x_1, \cdots, x_n)$ が得られたとき,

$$p(\theta|x) = \frac{p(\theta)\prod_{i=1}^{n} f(x_i|\theta)}{\int_{\Theta} p(\theta)\prod_{i=1}^{n} f(x_i|\theta)\, d\theta}$$

を**事後分布**（posterior distribution）と呼び, $p(\theta)$ を**事前分布**（prior distribution）と呼ぶ.

注意：ベイズでのΘの扱い

尤度では、Θは確率変数ではない

ベイズでは事前・事後分布が確率法則に従うのであれば、Θは確率変数となる

$$p(\theta|x) = \frac{p(\theta)\prod_{i=1}^{n} f(x_i|\theta)}{\int_{\Theta} p(\theta)\prod_{i=1}^{n} f(x_i|\theta)\, d\theta}$$

ベイズの推定での利点

ベイズでは、厳密な確率推論がパラメータ推定にも適用できる。

事後分布最大化推定量

定義16 (MAP推定値)
データx を所与として，以下の事後分布最大となるパラメータを求めるとき,

$$\hat{\theta} = \arg max\{p(\theta|x): \theta \in C\}$$

$\hat{\theta}$をベイズ推定値（Bayesian estimator）または，**事後分布最大化推定値**（maximum a posterior estimator, **MAP** 推定値）と呼ぶ.

EAP 推定値

定義17 (EAP 推定値)
データ$\boldsymbol{x}$ を所与として，以下の事後分布によるパラメータの期待値を求めるとき，$\hat{\theta} = E(\theta|x)$ を期待事後推定値（expected a posterior estimator , **EAP** 推定値）と呼ぶ.

ベイズ推定値も強一致性をもつ.

ベイズ推定の一致性

定理11 (ベイズ推定の一致性)

ベイズ推定において推定値$\hat{\theta}$が真のパラメータθ^*の強一致推定値となるような事前分布が設定できる.

定理12 (ベイズ推定の推定値の分散)

事後確率密度関数$p(\theta|x)$が以下で直接求められる.

$$Var(\theta|x)$$

17. 無情報事前分布

$$p(\theta|x) = \frac{p(\theta)\prod_{i=1}^{n} f(x_i|\theta)}{\int_{\Theta} p(\theta)\prod_{i=1}^{n} f(x_i|\theta)\, d\theta}$$

を求めるための事前分布$p(\theta)$ の設定について，どのように設定するかが問題となる. 通常，データを採取するまで，われわれはデータについての情報をもたない. そのために，$p(\theta)$ は無知を表す分布でなくてはならない. このような無知を示す事前分布を**無情報事前分布**(non–informative prior distribution)と呼ぶ.

無情報事前分布(Jeffreys1961)

母数θについて，$\theta \in (-\infty, \infty)$ のみの情報があるとき，事前分布は一様分布となる.

$$p(\theta) \propto const$$

$\int_{-\infty}^{\infty} p(\theta) \neq 1$となり，事前分布$p(\theta)$ は確率の公理を満たさない. このような事前分布を**improper prior distribution**と呼ぶ.

無情報事前分布(Jeffreys1961)

母数θについて，$\theta \in (0, \infty)$のみの情報があるとき，θの対数が一様であるような事前分布を考える. すなわち，$p(\log\theta) \propto const$であるから，変数変換すれば，

$$p(\theta) \propto \frac{1}{\theta}$$

$\int_{-\infty}^{\infty} p(\theta) \neq 1$となり，**improper prior distribution** .

注) 変数変換$\theta \to \phi$

$$p(\theta) \to p\big(f(\theta)\big)$$

$\phi = f(\theta)$とすると

$$p(\phi) = p(\theta)\frac{\partial\theta}{\partial\phi} = p(f^{-1}(\phi))\frac{\partial\theta}{\partial\phi}$$

Proper prior: principle of stable estimation(Edwards et al.1963)

例えば，$\theta \in [a, b]$ であれば，$p(\theta) = \frac{1}{b-a}$となり，$\int_{-\infty}^{\infty} p(\theta) = 1$ と確率の公理を満たす.

$\theta \in [a, b]$ では，$p(\theta) = const$ であるが，$\kappa = \theta^{10}$ としても，ジェフリーズのルールに従えば，$p(\kappa) = const$ となってほしい. しかし，変数変換すれば，そのようにならないことがわかる.

Jefferys prior (Box and Tiao 1973)

パラメータ変換を許容するパラメータ空間でエントロピーを最大にする事前分布は

$$p(\theta) \propto \sqrt{I(\theta)}$$

$I(\theta)$ はフィッシャー情報量を示す.

これが，ジェフリーズが提唱した母数の変換の不変性から導いた分布に一致するので，ジェフリーズの事前分布と呼ばれる.

自然共役事前分布（最も一般的！！）

これまでの事前分布では，データを得る前の事前分布と事後分布は，分布の形状が変化する. しかし，データの有無にかかわらず，分布の形状は同一のほうが自然. そこで，事前分布と事後分布が同一の分布族に属するとき，その事前分布を自然共役事前分布（natural conjugate prior distribution）と呼ぶ.

自然共役事前分布によるベイズ推定例

例7 (二項分布)

$$f(x|\theta,n) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

コインを投げてn 回中x 回表が出たときの

確率θをベイズ推定しよう.

尤度関数は，$\binom{n}{x}\theta^x(1-\theta)^{n-x}$であり，

二項分布の自然共役事前分布は，以下のベータ分布（$Beta(\alpha, \beta)$）である.

$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

事後分布は，$p(\theta|n,x,\alpha,\beta) = \frac{\Gamma(\mathrm{n}+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(\mathrm{n}-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$

とやはりベータ分布となる.

対数をとり，以下の対数事後分布を最大化すればよい.

$$\log p(\theta|n,x,\alpha,\beta)$$
$$= \log\frac{\Gamma(\mathrm{n}+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(\mathrm{n}-x+\beta)} + (x+\alpha-1)\log\theta + (n-x+\beta-1)\log(1-\theta)$$

$\frac{\partial \log p(\theta|n,x,\alpha,\beta)}{\partial\theta} = 0$のとき，対数事後分布は最大となるので，

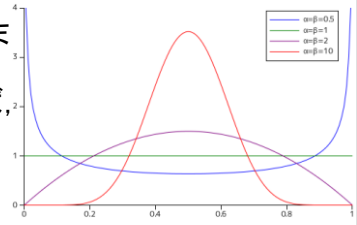
$$\frac{\partial \log p(\theta|n,x,\alpha,\beta)}{\partial\theta} = \frac{(x+\alpha-1)}{\theta} - \frac{(n-x+\beta-1)}{1-\theta}$$
$$= \frac{x+\alpha-1-x\theta-\alpha\theta+\theta-n\theta+x\theta-\beta\theta+}{\theta(1-\theta)}$$
$$= \frac{x+\alpha-1-(n+\alpha+\beta-2)\theta}{\theta(1-\theta)} = 0$$

$\theta(1-\theta) \neq 0$とすると

$$\hat{\theta} = \frac{x+\alpha-1}{n+\alpha+\beta-2}$$

がベイズ推定値となる．さて，α, βは事前分布のパラメータであるが，これをハイパーパラメータ（hyper parameter）と呼ぶ．

ハイパーパラメータによって，事前分布はさまざまな形状をとる（図）．例えば，事前分布が一様となる場合（$Beta(1, 1)$）の推定値は，$\hat{\theta} = \frac{x}{n}$となり，最尤解に一致する．



EAP推定量

$$\hat{\theta} = \frac{x+\alpha}{n+\alpha+\beta}$$

となり、例えば，事前分布が一様となる場合（$Beta(1, 1)$）の推定値は

$$\hat{\theta} = \frac{x+1}{n+2}$$

データがない場合は、$\hat{\theta} = \frac{1}{2}$となり、データが増えるごとに真値に近づく。

EAP推定量でジェフリーズ事前分布

$$\hat{\theta} = \frac{x+\alpha}{n+\alpha+\beta}$$

となり、例えば，事前分布が一様となる場合（$Beta(1, 1)$）の推定値は

$$\hat{\theta} = \frac{x+1/2}{n+1}$$

データがない場合は、一様分布同様に$\hat{\theta} = \frac{1}{2}$となるが、一様分布よりもデータに速く影響を受ける。

例8 (正規分布)

$$P(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}\}$$

$(x_1, \cdots, x_n)$ を得たときのμ, σ^2 を求めよう．

尤度は，$L = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$

$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$

このとき，自然共役事前分布は$\sigma_0^2 = \frac{\sigma^2}{n_0}$（注：$n_0$事前分布への信念の強さ）

$$p(\mu) = N(\mu_0, \sigma_0^2)$$
$$= \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$
$$\propto \left(\frac{\sigma^2}{n_0}\right)^{-\frac{1}{2}}\exp\left\{-\frac{n_0(\mu-\mu_0)^2}{2\sigma^2}\right\}$$

$$p(\sigma^2) = Ig(\nu_0, \lambda_0)$$
$$= \frac{(\lambda_0/2)^{\frac{1}{2}\nu_0}}{\Gamma(\frac{1}{2}\nu_0)} (\sigma^2)^{-\frac{1}{2}\nu_0 - 1} \exp\left(-\frac{\lambda_0}{2\sigma^2}\right)$$

(逆ガンマ分布)

$$\propto (\sigma^2)^{-\frac{1}{2}\nu_0 - 1} \exp\left(-\frac{\lambda_0}{2\sigma^2}\right)$$

事前分布はこれらの積の形で以下のように表される．自由度$\nu_0 = n_0 - 1$ とすると

$$p(\mu, \sigma^2) = p(\mu|\mu_0, \sigma_0^2) p(\sigma^2|\nu_0, \lambda_0)$$
$$\propto \left(\frac{\sigma^2}{n_0}\right)^{-\frac{1}{2}} \exp\left\{-\frac{n_0(\mu - \mu_0)^2}{2\sigma^2}\right\}$$
$$(\sigma^2)^{-\frac{1}{2}\nu_0 - 1} \exp\left(-\frac{\lambda_0}{2\sigma^2}\right)$$
$$\propto (\sigma^2)^{-\frac{1}{2}(\nu_0+1)-1} \exp\left\{-\frac{\lambda_0 + n_0(\mu - \mu_0)^2}{2\sigma^2}\right\}$$

ここで$n_0 = \nu_0 + 1$

事前分布を尤度に掛け合わせて事後分布を導くのであるが，計算の簡便さのために，以下のように尤度を変形させる．

$$L = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}\right\}$$

ここで指数部分$\exp\left\{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}\right\}$を三平方の定理により，推定平均$\bar{x}$を介して，以下のように分解する．

$$\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{2\sigma^2} + \frac{(\bar{x} - \mu)^2}{2\sigma^2}$$

尤度Lは，$L = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}\right\}$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{2\sigma^2}\right\}$$
$$\exp\left\{-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right\}$$
$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{S^2 + n(\mu - \bar{x})^2}{2\sigma^2}\right\}$$

ただし，ここで，　$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, S^2 = \sum_{i=1}^n (x_i - \bar{x})^2$

$$p(\mu, \sigma^2|x) \propto L \times p(\mu, \sigma^2)$$
$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{S^2 + n(\mu - \bar{x})^2}{2\sigma^2}\right\}$$

$\nu_0 = n_0 - 1$より

$$\times (\sigma^2)^{-\frac{1}{2}(\nu_0+1)-1} \exp\left\{-\frac{\lambda_0 + n_0(\mu - \mu_0)^2}{2\sigma^2}\right\}$$
$$\propto (\sigma^2)^{-\frac{1}{2}(n+n_0)-1}$$
$$\exp\left\{-\frac{\lambda_0 + S^2 + n_0(\mu - \mu_0)^2 + n(\mu - \bar{x})^2}{2\sigma^2}\right\}$$

さらに，指数部分のうち，$\lambda_0 + S^2$以外の部分に平方完成を行うと，　$p(\mu, \sigma^2|x) \propto (\sigma^2)^{-\frac{1}{2}(n+n_0)-1} \exp\left\{-\frac{\lambda_* + (n_0+n)(\mu - \mu_*)^2}{2\sigma^2}\right\}$

ただし，　$\lambda_* = \lambda_0 + S^2 + \frac{n_0 n(\bar{x} - \mu_0)^2}{n_0 + n}, \mu_* = \frac{n_0\mu_0 + n\bar{x}}{n_0 + n}$

この事後分布もまた，正規分布と逆ガンマ分布の積となり，

$$N \times IG(n_0 + n, \mu_*, \nu_0 + n, \lambda_*)$$

事後分布は，μとσ^2の同時事後確率分布

μの周辺事後分布

このように，複数のパラメータを同時に最大化させる場合，つぎのような周辺化（marginalization）を行い，個々のパラメータの分布を導く．このような分布を周辺事後分布（marginal posterior distribution）と呼ぶ．$p(\mu|x) = \int_0^\infty p(\mu, \sigma^2|x)p(\sigma^2)d\sigma^2$

$$\propto \frac{\Gamma\left[\frac{(\nu_*+1)}{2}\right]}{\sqrt{\frac{\nu_*\pi\lambda_*}{n_*}}\Gamma\left(\frac{\nu_*}{2}\right)}\left\{1+\frac{(\mu-\mu_*)^2}{\mu_*}\right\}^{-\frac{1}{2}(\nu_*+1)}$$

$$\equiv t(\nu_*, \mu_*, \lambda_*/n_*)$$

μ の周辺事後分布はt 分布 $t(\nu_*, \mu_*, \lambda_*/n_*)$ に従う

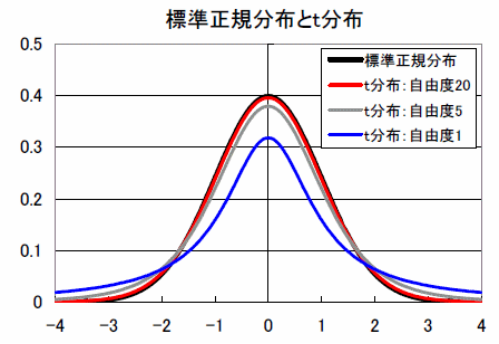
MAP推定値

事後確率最大化によるベイズ推定値は，t 分布のモードがμ_*であることより，

μ のMAP推定値は，

$$\hat{\mu} = \frac{n_0\mu_0 + n\bar{x}}{n_0 + n}$$

正規分布とt分布



σ^2の周辺事後分布

σ^2についての周辺事後分布は

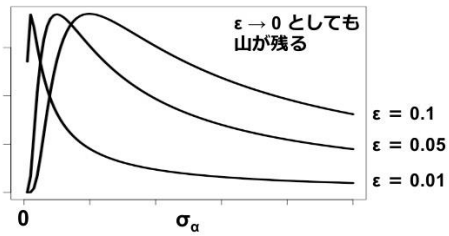
$$p(\sigma^2|x) = \int_0^\infty p(\mu, \sigma^2|x)p(\mu)d\mu$$

$$\propto \frac{\lambda_*^{\frac{\nu_*}{2}}}{2^{\frac{\nu_*}{2}}\Gamma\left(\frac{\nu_*}{2}\right)}(\sigma^2)^{-\frac{\nu_*}{2}-1}\exp\left(-\frac{\lambda_*}{2\sigma^2}\right)$$

となり，σ^2の周辺事後分布は，逆ガンマ分布$IG(\nu_*/2, \lambda_*/2)$ に従うことがわかる．

逆ガンマ分布

- $\sigma_\alpha \sim$ InvGamma(ε, ε)



MAP推定値

σ^2のベイズ推定値は，逆ガンマ分布のモードが$\frac{\lambda_{*/2}}{\nu_{*/2}+1} = \frac{\lambda_*}{\nu_*+2}$であることより，$\sigma^2$のMAP推定値は，

$$\widehat{\sigma^2} = \frac{\left\{\lambda_0 + S^2 + \frac{n_0 n(\bar{x}-\mu_0)^2}{n_0+n}\right\}}{\nu_* + 2}$$

EAP推定値

μ のEAP推定値は，平均値とモードが同一なので

$$\hat{\mu} = \frac{n_0\mu_0 + n\bar{x}}{n_0 + n}$$

σ^2 のMAP推定値は，逆ガンマ分布のモードが$\frac{\lambda_{*/2}}{\nu_{*/2}-1} = \frac{\lambda_*}{\nu_*-2}$であることより，

$$\widehat{\sigma^2} = \frac{\left\{\lambda_0 + S^2 + \frac{n_0 n(\bar{x}-\mu_0)^2}{n_0+n}\right\}}{\nu_* - 2}$$

事前分布の意味を考える例題

以下のどちらのかけを選ぶと得か？

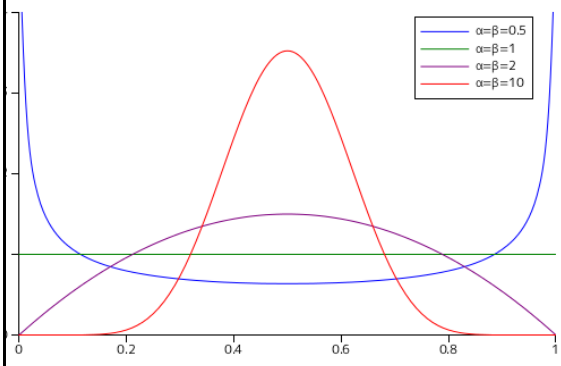
1. 50個の赤玉と50個の白玉が入った壺から一つ玉を取り出し，それが赤玉であったら1万円もらえる。白玉であったら1万円支払う。
2. 赤玉と白玉が合わせて100個入った壺から一つ玉を取り出し，それが赤玉であったら1万円もらえる。白玉であったら1万円支払う。

追加例題

以下のどちらのかけを選ぶと得か？

1. 50個の赤玉と50個の白玉が入った壺から一つ玉を取り出し，それが赤玉であったら1万円もらえる。白玉であったら1万円支払う。これを10回繰り返す。
2. 赤玉と白玉が合わせて100個入った壺から一つ玉を取り出し，それが赤玉であったら1万円もらえる。白玉であったら1万円支払う。

   これを10回繰り返す。



赤玉の確率$P(A) = \psi$

事前分布の例題2

いま、外見がまったく同じ2つの封筒の中に、現金が入っているものとする．それぞれの封筒の中の金額は知らされていないが、片方にはもう一方の2倍が入っていることが分かっている。今、AとBの二人に封筒がランダムに分けられ、自分の中身だけ見て交換してもよいルールとなった。Aの封筒には10ドル入っていた。交換したほうがよいか？

18. データから統計モデルを選択

統計モデルのパラメータ(母数)をデータから推定するには，尤度最大化により漸近的一致性が得られた．

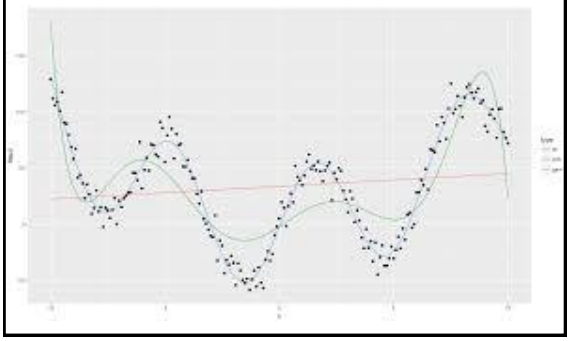
ひとつのデータに対して，複数のモデルからどのモデルが一番よいかを決定するときに，尤度最大化は使えるのであろうか？

→

モデル選択基準

例；多項式のデータへのあてはめ

$$y = a_k x^k + a_{k-1} x^{k-1} + \cdots * a_1 x + a_0$$



パラメータ数が増えると予測が劣化

$$y = a_k x^k + a_{k-1} x^{k-1} + \cdots * a_1 x + a_0$$

パラメータ数=*k+1*

パラメータ数が増える（モデルが複雑になる）とデータとの誤差が単調減少し、尤度は単調増加する。

データ数=パラメータ数のとき既知のデータへのあてはまり誤差は0になるが、未知のデータへの予測は非常に悪くなる。この現象を過学習(over fitting)という。

尤度最大化はモデル選択に使えない

複雑なモデルほど尤度が高くなってしまうので
尤度最大化では、モデルの選択はできない

予測を最大にするモデルを選択手法は何か？

AIC(Akaike Information Criterion 1973)

$$AIC = -2\mathrm{E}[\ln L] \approx -2\ln L + 2k$$

ここで，$\ln L$は対数最大尤度、kはモデルのパラメータ数

Akaike, H., "Information theory and an extension of the maximum likelihood principle", *Proceedings of the 2nd International Symposium on Information Theory*, Petrov, B. N., and Caski, F. (eds.), Akadimiai Kiado, Budapest: 267-281 (1973).

AICの意味

-½ AIC = 尤度（モデルのてはまり）
- パラメータ数（モデルの複雑さ）

モデルのあてはまりと モデルの複雑さのトレードオフが存在する。

AICは一致性を持たない

しかし、AICはデータ数を増やしても真のモデルを選択する確率が1.0に収束しないという問題がある。

準備：分布P(θ)と分布Q(θ)の距離

カルバックライブラー距離

$$\int_\theta P(\theta)\log\frac{P(\theta)}{Q(\theta)}d\theta$$

AICの導出の考え方

真の分布$P^*(\theta)$と分布の推定値$P(\theta)$のカルバックライブラー距離

$$\int_\theta P^*(\theta)\log\frac{P^*(\theta)}{P(\theta)}d\theta$$

$$=\int_\theta P^*(\theta)\log P^*(\theta)\,d\theta$$

$$-\int_\theta P^*(\theta)\log P(\theta)\,d\theta$$

AICの導出の考え方

真の分布$P^*(\theta)$と分布の推定値$P(\theta)$のカルバックライブラー距離

$$\int_\theta P^*(\theta)\log\frac{P^*(\theta)}{P(\theta)}d\theta$$

$$=\int_\theta P^*(\theta)\log P^*(\theta)\,d\theta \quad \text{Const}$$

$$-\int_\theta P^*(\theta)\log P(\theta)\,d\theta$$

ここだけ考えればよい

クロス　エントロピー

AICの導出の考え方

$$-\int_\theta P^*(\theta)\log P(\theta)\,d\theta$$

$$\approx -\int_\theta P(\theta)\log P(\theta)\,d\theta$$

$$\approx -\mathrm{E}[\ln L]$$

LnLを二回　テーラー近似し、

$$-\mathrm{E}[\ln L]$$

$$\approx -\ln L + k$$

これを最小化すればよい。

問題

$P^*(\theta)$を$P(\theta)$に置き換えてしまうとクロスエントロピーは真の分布との距離を反映しない。

結局、期待対数尤度を最大化してしまうので過学習が起こり、複雑なモデルを好んでしまう。

AICは一致性を持たない

尤度はモデルを複雑にするといくらでも大きくなってしまう。そこでその平均を考えるとモデルの複雑さ（パラメータ数）をペナルティとして考えないといけないことがわかる。

しかし、AICはデータ数を増やしても真のモデルを選択する確率が1.0に収束しない。

ベイズではモデルの確率を考える

m:モデル，M:モデル候補集合，x:データ

$$p(m|x) = \frac{p(x|m)p(m)}{\sum_{i=1}^{M} p(x|m_i)p(m_i)}$$

今、すべての$p(m)$が同一だと考えると

$p(x|m)$ が最大となるモデルを選択すればよい。

ここで

$$p(x|m) = \int_{\Theta} p(x|\theta, m)p(\theta|m)d\theta$$

を周辺尤度と呼ぶ。

19 周辺尤度

ベイズ統計では，一般的に，モデル選択のために以下の周辺尤度を最大にするモデルを選択する.

定義19

データxを所与としたモデルmの尤度を周辺化して周辺尤度（marginal likelihood），**ML** と呼ぶ.

$$p(x|m) = \int_{\Theta} p(x|\theta, m)p(\theta|m)d\theta$$

BIC(Bayesian Information Criterion)

周辺尤度は、モデルごとにパラメータ空間を積分消去しなければならない。より、簡単に用いるために 周辺尤度の漸近近似としてBICが求められた。これは漸近一致性を持つ。

$$\mathrm{BIC} = \ln(L) - \frac{1}{2}k\ln(n)$$

ここで，$\ln L$は対数最大尤度、kはモデルのパラメータ数，nはデータ数.

Schwarz, Gideon E. (1978), "Estimating the dimension of a model", *Annals of Statistics*, **6** (2): 461–464

MDL(minimum description length)

Jorma Rissanen により導入された。MDLでは、データをモデルを用いて圧縮・送信する際の符号長の最小化を考える。これはノイズを含むデータから意味のある規則性を抽出することにあたる。最初はBICと等価な基準が提案されたが、その後NML（ Normalized Maximum Likelihood ）も提案されている。基本、符号問題、離散データの圧縮問題に用いる理論的仮定がある。

Rissanen, J. (1978). "Modeling by shortest data description". *Automatica*. **14** (5): 465–658

MDL(minimum description length)

NMLのアイデアは尤度を確率になるように標準化する。そのためにはデータのとりえるパターンの尤度をすべて列挙して計算しなければならないので計算量の問題、またデータがスパースなパターンがあるのでデータスパース問題がありえる。

BICの数式は そもそも周辺尤度の近似であるがNMLの近似としても導ける。

20. 予測分布

データやモデルを用いて推論を行う重要な目的の一つに，未知の事象の予測が挙げられる. この予測問題のためには，最もよく用いられるのは，

$$p(y|\hat{\theta})$$

で示されるplug–in distribution と呼ばれる分布である. しかし，$\hat{\theta}$ は推定値であるためにそのサンプルのとり方によってこの分布は大きく変化する. ベイズ的アプローチでは，この$\hat{\theta}$ のばらつき（$\hat{\theta}$ の事後分布）を考慮し，以下のように予測分布を定義する.

20. 予測分布

定義18

モデルm から発生されるデータ$\boldsymbol{x}$ により，未知の変数y の分布を予測するとき，以下の分布を予測分布（predictive distribution）と呼ぶ．

$$p(y|\boldsymbol{x},m)=\int_{\Theta}p(y|\theta,m)p(\theta|\boldsymbol{x},m)d\theta$$

例9 (二項分布) ベータ分布を事前分布とした二項分布の予測分布は，以下のようになる．

$$p(y|x)=\int_{\Theta}p(y|\theta)p(\theta|x)d\theta=$$
$$\int_{\Theta}\binom{n}{y}\theta^{y}(1-\theta)^{n-y}\frac{\Gamma(\mathrm{n}+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}\times$$
$$\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}d\theta$$
$$\propto\binom{n}{y}\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$
$$\frac{\Gamma(x+\alpha)(n-x+\beta)}{\Gamma(n+\alpha+\beta)}$$

$$=\frac{n!}{y!\,(n-y)!}\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$
$$\frac{\Gamma(x+\alpha)(n-x+\beta)}{\Gamma(n+\alpha+\beta)}$$

特に，α, β が整数のとき

$$p(y|x)\propto\frac{n!}{y!\,(n-y)!}\frac{y!\,(n-y)!}{(n+1)!}$$
$$\frac{(x+\alpha-1)!\,(n-x+\beta-1)!}{(n+\alpha+\beta-1)!}$$

例10 (正規分布) 事前分布を$N(\mu,\sigma^2)$ 分布

$$p(\mu,\sigma^2)=p(\mu|\sigma^2)p(\sigma^2)$$
$$\propto\left(\frac{\sigma^2}{n_0}\right)^{-\frac{1}{2}}\exp\left\{-\frac{n_0(\mu-\mu_0)^2}{2\sigma^2}\right\}(\sigma^2)^{-\frac{1}{2}\nu_0-1}$$
$$\exp\left(-\frac{\lambda_0}{2\sigma^2}\right)$$
$$=(\sigma^2)^{-\frac{1}{2}(\nu_0+1)-1}\exp\left\{-\frac{\lambda_0+n_0(\mu-\mu_0)^2}{2\sigma^2}\right\}$$

事後分布は

$$p(\mu,\sigma^2|x)\propto(\sigma^2)^{-\frac{1}{2}(n+n_0)-1}$$
$$\exp\left\{-\frac{\lambda_*+(n_0+n)(\mu-\mu_*)^2}{2\sigma^2}\right\}$$

ただし，$\lambda_*=\lambda_0+S^2+\frac{n_0n(\bar{x}-\mu_0)^2}{n_0+n}$,

$$\mu_*=\frac{n_0\mu_0+n\bar{x}}{n_0+n}$$

予測分布は

$$p(x_{n+1}|\boldsymbol{x})$$
$$=\int\int p(x_{n+1}|\mu,\sigma^2)p(\mu,\sigma^2|x_1,\cdots,x_n)d\mu d\sigma^2$$

ここで，$p(x_{n+1}|\mu,\sigma^2)\propto$
$(\sigma^2)^{-1}\exp\left\{-\frac{(x_{n+1}-\mu)^2}{2\sigma^2}\right\}$

$$p(x_{n+1}|\boldsymbol{x}) = \int\int p(x_{n+1}|\mu,\sigma^2)p(\mu,\sigma^2|x_1,\cdots,x_n)d\mu d\sigma^2$$

$$\propto \int\int (\sigma^2)^{-\frac{\nu+1}{2}-2}\exp\left[-\frac{(x_{n+1}-\mu)^2+S^2+n(\mu-\bar{x})^2}{2\sigma^2}\right]d\mu d\sigma^2$$

$$= \int\int (\sigma^2)^{-\frac{\nu+1}{2}-2}\exp\left[-\frac{1}{2\sigma^2}\left\{(n+1)(\mu-\bar{\mu})^2+S^2 + \frac{n}{n+1}(x_{n+1}-\bar{x})^2\right\}\right]d\mu d\sigma^2$$

$$\propto \int (\sigma^2)^{-\frac{\nu+1}{2}-2}\exp\left[-\frac{1}{2\sigma^2}\left\{S^2+\frac{n}{n+1}(x_{n+1}-\bar{x})^2\right\}\right]d\sigma^2$$

$$\propto \left\{S^2+\frac{n}{n+1}(x_{n+1}-\bar{x})^2\right\}^{-\frac{\nu+1}{2}}$$

$$\propto \left[1+\left\{\frac{x_{n+1}-\bar{x}}{\sqrt{\frac{n+1}{n\nu}S^2}}\right\}^2/\nu\right]^{-\frac{\nu+1}{2}}$$

ただし,ここで

$$\bar{\mu} = \frac{n\bar{x}+x_{n+1}}{n+1}$$

ここで, $t = \frac{x_{n+1}-\bar{x}}{\sqrt{\frac{n+1}{n\nu}S^2}}$

とおくとき,tは自由度νのt分布に従う.

21.マルコフ連鎖モンテカルロ法 (MCMC法)

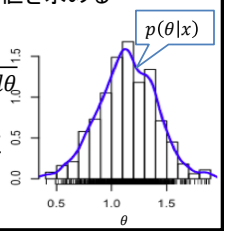
確率分布をサンプリング近似する手法

ベイズ推定では，パラメータの事後分布を推定し，得られた分布形に基づいて推定値を求める

$$p(\theta|\boldsymbol{x}) = \frac{p(\theta)\prod_{i=1}^{n} f(x_i|\theta)}{\int_\Theta p(\theta)\prod_{i=1}^{n} f(x_i|\theta)\,d\theta}$$

$\mathrm{argmax}_\theta\, p(\theta|x) \rightarrow$ MAP推定値

$\mathrm{E}_\theta[\,p(\theta|x)\,] \rightarrow$ EAP推定値



代表的なMCMCアルゴリズム

1. ギブスサンプリング
2. メトロポリスヘイスティングス

他のMCMCアルゴリズム:

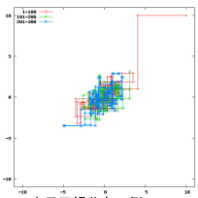
スライスサンプリング

ハミルトニアンモンテカルロ

以降では，多次元パラメータ$\boldsymbol{\theta} = \{\theta_1,\cdots,\theta_K\}$の事後分布をMCMCで推定することを想定

1. ギブスサンプリング

事後分布$p(\boldsymbol{\theta}|\boldsymbol{x})$から直接にはサンプリングできないが，パラメータごとの条件付き分布 $p(\theta_i|\boldsymbol{x},\boldsymbol{\theta}^{\setminus i})$からはサンプリングができる場合に利用できる手法（ここで，$\boldsymbol{\theta}^{\setminus i} = \boldsymbol{\theta} \setminus \{\theta_i\}$ ）パラメータごとの条件付き分布から順にサンプリングを繰り返す



2次元正規分布の例
http://d.hatena.ne.jp/jetbead/20120119/1326987540 より

アルゴリズム

以下を十分な回数繰り返す

$$\theta_1 \sim p(\theta_1|x,\boldsymbol{\theta}^{\setminus 1})$$
$$\theta_2 \sim p(\theta_2|x,\boldsymbol{\theta}^{\setminus 2})$$
$$\vdots$$
$$\theta_K \sim p(\theta_K|x,\boldsymbol{\theta}^{\setminus K})$$

サンプリングしたパラメータ値 $\boldsymbol{\theta}$を保存

例：正規分布のパラメータ推定

x_i~N(μ, σ^2)とするn個のサンプル$\boldsymbol{x} = \{x_1, \cdots, x_n\}$を所与としてパラメータ$\mu, \sigma^2$を推定

パラメータの同時事後分布はサンプリング可能な既知の分布とならないため，この分布から直接サンプリングすることはできない

$$p(\mu, \sigma^2|\boldsymbol{x}) = \frac{p(\mu)p(\sigma^2)\prod_{i=1}^{n} f(x_i|\mu, \sigma)}{\int p(\mu)p(\sigma^2)\prod_{i=1}^{n} f(x_i|\mu, \sigma)\, d\mu, d}$$

しかし，以下の条件付き分布はそれぞれ既知の分布になるため，サンプリングが可能

$$p(\mu|\boldsymbol{x}, \sigma^2),\ p(\sigma^2|\boldsymbol{x}, \mu)$$

μ，σ^2の事前分布に一様分布を仮定すると

$$p(\mu|\boldsymbol{x}, \sigma^2) = N\left(\frac{1}{N}\sum_{i=1}^{n} x_i, \frac{\sigma^2}{N}\right)$$

$$p(\sigma^2|\boldsymbol{x}, \mu) = IG\left(\frac{n}{2} + 1, \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2}\right)$$

正規分布や逆ガンマ分布$IG()$からの乱数生成手法は既知

多くのプログラミング言語にはこれらの乱数生成器が実装されている

2. メトロポリスヘイスティングス

条件付き分布からもサンプリングできないときに利用

Step1:

現在のパラメータ値$\boldsymbol{\theta}$の付近の候補値$\boldsymbol{\theta}^*$を，

提案分布（proposal distribution）$p(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ から生成

一般に $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = MN(\boldsymbol{\theta}^*|\boldsymbol{\theta}, \boldsymbol{I}\sigma)$

MNは多次元正規分布，$\boldsymbol{I}$は単位行列，σは微小な値(0.01等)

2. メトロポリスヘイスティングス

Step2:

以下の採択確率に基づいて候補値$\boldsymbol{\theta}^*$を採択

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\boldsymbol{x})q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}|\boldsymbol{x})q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}\right\}$$

($q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$
$= MN(\boldsymbol{\theta}^*|\boldsymbol{\theta}, \boldsymbol{I}\sigma)$ のとき) $\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta})$

$$= \min\left\{1, \frac{p(\boldsymbol{\theta}^*|x)}{p(\boldsymbol{\theta}|x)}\right\}$$

棄却された場合には$\boldsymbol{\theta}^* = \boldsymbol{\theta}$とする

考え方

$$p(\boldsymbol{\theta} \to \boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta})\ \alpha(\boldsymbol{\theta}^*)$$
$$p(\boldsymbol{\theta}^* \to \boldsymbol{\theta}\ \ ) = q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)\ \alpha(\boldsymbol{\theta})$$

$$\frac{\alpha(\boldsymbol{\theta}^*)}{\alpha(\boldsymbol{\theta})} = \frac{p(\boldsymbol{\theta}^*|\boldsymbol{x})q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}|\boldsymbol{x})q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}$$

採択確率計算時のポイント

事後分布の分母は多重積分を含むため計算困難

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{\theta})\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

しかし，採択確率の計算ではこの項は消去可能

$$\frac{p(\boldsymbol{\theta}^*|\boldsymbol{x})}{p(\boldsymbol{\theta}|\boldsymbol{x})} = \frac{\frac{p(\boldsymbol{\theta}^*)\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}^*)}{\int p(\boldsymbol{\theta}^*)\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}^*)\, d\boldsymbol{\theta}^*}}{\frac{p(\boldsymbol{\theta})\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})\, d\boldsymbol{\theta}}}$$
$$= \frac{p(\boldsymbol{\theta}^*)\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta})\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})}$$

3. メトロポリス with ギブス

メトロポリスヘイスティングスでは，パラメータ数が増加すると，パラメータ値が改悪される方向に進むときに，採択確率$\frac{p(\boldsymbol{\theta}^*|\boldsymbol{x})}{p(\boldsymbol{\theta}|\boldsymbol{x})}$が極端に小さくなり，更新が進まなくなることがある

メトロポリスヘイスティングス with ギブス

パラメータごとにメトロポリスヘイスティングスを実行する手法

アルゴリズム

Init $\boldsymbol{\theta} = \{\theta_1 \cdots \theta_K\}$ # 初期値をランダムに設定
For loop = 1 to Max Loop:
　For $i = 1, \cdots, K$:
　　・現在の値を所与として θ_iの候補値θ_i^*を生成

$$\theta_i^* \sim N(\theta_i, \sigma^2)$$

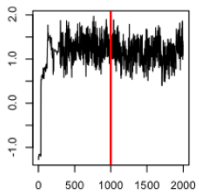
　　・採択確率に基づきθ_i^*を採択（または棄却）

$$\alpha(\theta_i^*, \theta_i) = \min\left\{1, \frac{p(\theta_i^*|x, \boldsymbol{\theta}^{\setminus i})}{p(\theta_i|x, \boldsymbol{\theta}^{\setminus i})}\right\}$$

　end for
　現在のパラメータ値を保存
end for
得られたサンプル集合を事後分布からのサンプルとみなし，所望の統計量を計算

サンプルの選択（バーンイン）

アルゴリズム初期のサンプルは，初期値に依存するため，一定回数サンプリングを繰り返した後のサンプルを利用する
初期値に依存しなくなったとみなすまでの時間をバーンイン期間と呼ぶ



サンプルの選択（インターバル）

メトロポリスヘイスティングスは，サンプル間の自己相関（隣接するサンプル間の依存性）が高いため，一定区間でサンプルを間引いて用いる必要がある

間引く間隔をインターバル期間と呼ぶ

アルゴリズム（修正版）

Init $\boldsymbol{\theta} = \{\theta_1 \cdots \theta_K\}$ # 初期値をランダムに設定
For loop = 1 to Max Loop:
　For $i = 1, \cdots, K$:
　　・現在の値を所与として θ_iの候補値θ_i^*を生成

$$\theta_i^* \sim N(\theta_i, \sigma^2)$$

　　・採択確率に基づきθ_i^*を採択（または棄却）

$$\alpha(\theta_i^*, \theta_i) = \min\left\{1, \frac{p(\theta_i^*|x, \boldsymbol{\theta}^{\setminus i})}{p(\theta_i|x, \boldsymbol{\theta}^{\setminus i})}\right\}$$

　end for
　If loop > バーンイン期間:
　　If loop % interval = 0:
　　　現在のパラメータ値$\boldsymbol{\theta}$を保持
end for
得られたサンプル集合を用いて所望の統計量を計算

22. （従来手法）統計的仮説検定

- ある仮説が正しいかどうかを標本（データ）から判定する手法.
- 統計的仮説（Statistical Hypothesis）：
  - 帰無仮説（null hypothesis）：棄却されることを前提とした仮説を表し H_0 とする.
  - 対立仮説（alternative hypothesis）：帰無仮説が棄却されたときの採用される仮説を表し H_1とする.
- 有意水準α：ユーザが設定する帰無仮説を棄却する基準であり，誤って帰無仮説を棄却してしまう確率を表す.

仮説検定の手順

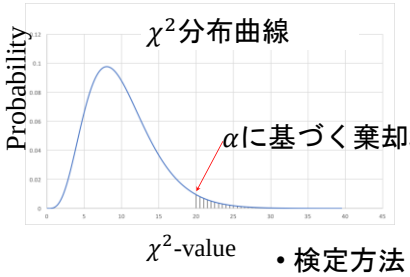
1. 帰無仮説H_0，対立仮説H_1を決める．
2. 得られたデータから統計量を求める．
   - 用いる統計量：T（T分布），F（F分布），
     χ^2（χ^2分布）
3. 用いる統計量が確率分布にどれだけ従っているかを表す確率p値を求める．p値は帰無仮説が正しい確率とも言われ，有意水準αより小さければ，帰無仮説H_0は棄却し対立仮説H_1を採用する．

独立性検定

- 帰無仮説：2変数間が独立
- 対立仮説：2変数間が従属
- 一般的にχ^2 統計量を用いて自由度dfのχ^2分布との適合度により独立性を検定する

$$\text{検定統計量} = \sum\sum \frac{(\text{観測度数} - \text{期待度数})^2}{\text{期待度数}} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(x_{ij} - y_{ij})^2}{y_{ij}}$$

例：自由度10のχ^2検定



- 検定方法：
  p値 $< \alpha$ → 従属
  p値 $> \alpha$ → 独立

仮説検定法の問題点

- 検定の精度：p値と有意水準 α に依存する．

これによって引き起こされる問題

- 真に帰無仮説が正しいが，誤って棄却してしまう．
  →第一種の過誤（Type I error）と呼ばれる．
- 真に対立仮説が正しいが，帰無仮説を棄却しない．
  →第二種の過誤（Type II error）と呼ばれる．

ベイズ的アプローチによる検定

- Bayes factor：
  二つのモデルの周辺尤度の比により検定する．
- 漸近的に真の独立性検定が可能である．
- データセットを$\mathbf{X}$，独立なモデルを$g_1: p(x_1, x_2) = p(x_1)p(x_2)$，従属なモデルを$g_2: p(x_1, x_2) = p(x_1|x_2)p(x_2)$としたときの周辺尤度の比　Bayes factor（BF）：

$$BF = \frac{p(\mathbf{X} \mid g_1)}{p(\mathbf{X} \mid g_2)}$$

$BF > 1$：独立
$BF < 1$：従属　と判定する

シミュレーション実験1
-Type I errorの検証-

- 2ノード間が真に独立である構造を用いて実験を行う．
- χ^2 統計量を用いた検定ではデータ数を増やしたとしても
  Type I errorが発生するが，Bayes factorでは漸近的に収束することを示す．

実験結果 – 確率パラメータ:0.8

表：Type Ⅰ errorの発生率

	10	50	100	500	1000	5000
BF	0.26	0.07	0.02	0.0	0.0	0.0
χ^2	0.16	0.0	0.0	0.03	0.08	0.07
G^2	0.17	0.05	0.02	0.03	0.08	0.06

※BF: Bayes factor

実験結果 – 確率パラメータ:0.7

表：Type Ⅰ errorの発生率

	10	50	100	500	1000	5000
BF	0.23	0.09	0.03	0.02	0.0	0.0
χ^2	0.08	0.08	0.07	0.07	0.05	0.02
G^2	0.14	0.11	0.08	0.07	0.05	0.03

※BF: Bayes factor

実験結果 – 確率パラメータ:0.6

表：Type Ⅰ errorの発生率

	10	50	100	500	1000	5000
BF	0.12	0.04	0.01	0.01	0.0	0.0
χ^2	0.02	0.06	0.04	0.14	0.03	0.07
G^2	0.08	0.06	0.04	0.14	0.03	0.06

※BF: Bayes factor

シミュレーション実験2
-Type Ⅱ errorの検証-

- 2ノード間が真に従属である構造を用いて実験を行い，Type Ⅱ errorの発生率とp値を検証する.

実験結果 – 確率パラメータ:0.8

表：Type Ⅱ errorの発生率

	10	20	30	40	50	100
BF	0.3	0.29	0.19	0.11	0.02	0
χ^2	0.61	0.42	0.19	0.1	0.02	0
G^2	0.49	0.32	0.18	0.11	0.02	0

※BF: Bayes factor

実験結果 – 確率パラメータ:0.7

表：Type Ⅱ errorの発生率

	10	20	30	40	50	100	200
BF	0.62	0.61	0.45	0.44	0.31	0.09	0
χ^2	0.89	0.75	0.43	0.39	0.23	0.02	0
G^2	0.72	0.65	0.43	0.37	0.24	0.02	0

※BF: Bayes factor

実験結果 – 確率パラメータ:0.6

表：Type II errorの発生率

	40	50	100	200	500	1000
BF	0.77	0.7	0.65	0.33	0.05	0
χ^2	0.73	0.62	0.54	0.19	0.01	0
G^2	0.73	0.61	0.54	0.19	0.01	0

※BF: Bayes factor

仮説検定の比較

従来の仮説検定では、結果が不安定で必ずType I誤差が残るのに対して、ベイズ検定では漸近的に正しい仮説を選ぶことができる。