パターン認識と機械学習 第2章：確率分布（後半）

Christopher M. Bishop (2006):
Pattern Recognition and Machine Learning, Springer, pp.95-125

0.もくじ

1.1. ガウス分布に対するベイズ推論

以下では平均 μ と共分散行列 Σ という2つのパラメータの事前分布を導入して、ガウス分布のベイズ主義的な扱い方を導く。

まずは...

- 1変数のガウス分布
- 分散 σ^2 (1変数のため行列ではない)は既知
- 平均 μ の分布をベイズ的に推定

...という場合について。

μ が与えられた際に観測データ列 $\mathbf{x}=\{x_1 \ldots x_N\}$ が生じる確率である尤度関数は

$$p(\mathbf{x} \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

で表される。

*ガウス分布の一般形をかけただけ！

$$p(\mathbf{x}\,|\,\mu)=\prod_{n=1}^{N}p(x_n\,|\,\mu)=\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

この尤度関数を μ の関数と考えると、これは μ についての二次形式の指数の形をとっている。

この形に対応するような共役事前分布はガウス分布！
(* 事後分布 ∝ 尤度関数 × 事前分布であるから、上記の尤度関数をかけても事前分布・事後分布ともに μ についてはおなじ形(共役)であるとよい、ということ)

したがって、事前分布を以下のようにとればよい。

$$p(\mu)=\mathcal{N}(\mu\,|\,\mu_0,\sigma_0^2)$$

これを用ると事後分布は次式で表される。

$$p(\mu\,|\,\mathbf{x})\propto p(\mathbf{x}\,|\,\mu)p(\mu)$$

$$=\left(\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}\right)\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)\right)$$

これをがりがり計算していき、まとめると、事後分布は以下のような平均 μ_N と分散σ_N^2 を持つガウス分布となる。

$$p(\mu \mid \mathbf{x}) = \mathcal{N}(\mu \mid \mu_N, \sigma_N^2)$$

ここで平均 μ_N と分散σ_N^2 は

$$\mu_N = \frac{N\sigma_0^2 \mu_{ML} + \sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

ただし μ_{ML} はサンプル平均、すなわち

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

である。

$$\mu_N = \frac{\sigma_0^2 N \mu_{ML} + \sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

この事後分布の平均 μ_N について、以下のことが言える。

- 事前分布の平均 μ_0 と最尤推定解 μ_{ML} の間をとった形になっている
- 観測データが0(すなわち N=0)なら、事前分布の平均 μ_0 と等しくなる。
- 逆に、N→∞ では最尤推定解解 μ_{ML} となる。

また事後分布の精度(分散の逆数)についても以下のことが言える。

- 観測データがo(すなわち N=o)なら、事前分布の分散 σ_o^2 と等しくなる。
- 事後分布の精度 $1/\sigma_N^2$ は、事前分布の精度に各観測データ点からのデータ精度への影響分を加えたものになり(加算的)、したがって観測データ点が増えるにつれて精度も単調に増加する(分散がoに近づく)。N→∞ では分散はoになる。

(ほかにもあるけれど、上記が特に重要と考え、省略しました)

以上の議論は平均が未知の多次元ガウス分布にもそのまま一般化できる。

さらにこのベイズ推を逐次的に捉えるために x_{N-1} までの式と x_N とに分けてみると...

$$p(\mu \mid \mathbf{x}) \propto \left[p(\mu) \prod_{n=1}^{N-1} p(x_n \mid \mu) \right] p(x_N \mid \mu)$$

カギ括弧内の項は、結局(正規化係数を除いて)N-1個のデータ点を観測したあとの事後分布とちょうど一致する。

すなわち、このカギ括弧内の項(≒N−1個のデータ点を観測したあとの事後分布)を事前分布にとり、新しいデータ点 x_N についての尤度関数をベイズの定理によって結合したもの(この式全体)は、N個のデータ点を観測した後の事後分布とみなすことができるのである。

すごいね！

つぎに…

- 1変数のガウス分布
- 平均 μ は既知
- 分散　σ^2　の分布をベイズ的に推定

という場合を考える。

ただし以下では精度 $\lambda \equiv 1/\sigma^2$ をもって操作することとする(そのほうが楽らしい)。

このとき尤度関数は(先ほどと全く同様に)以下の式で与えられる。

$$p(\mathbf{x} \mid \lambda) = \prod_{n=1}^{N} p(x_n \mid \mu, \lambda^{-1}) = \lambda^{N/2} \exp\left\{ -\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}$$

*分散を精度で置き換えただけ！

$$p(\mathbf{x} \mid \lambda) = \prod_{n=1}^{N} p(x_n \mid \mu, \lambda^{-1}) = \lambda^{N/2} \exp\left\{ -\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}$$

したがって共役な事前分布としては「λのべき乗」と「λの線形関数の指数」の積に比例するものを選びたい…

このような条件を満たし、かつ便利な性質をもつのが以下に示すガンマ分布である。

$$\mathrm{Gam}(\lambda \mid a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

以上より、事前分布 $\mathrm{Gam}(\lambda|a_0,b_0)$ に先ほどの尤度関数をかけあわせることで、以下の事後分布が得られる。（正規化係数であるガンマ関数の部分は省いてある）

$$p(\lambda \mid \mathbf{x}) \propto \lambda^{a_0 - 1} \lambda^{N/2} \exp\left\{ -b_0 \lambda - \frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}$$

この式は以下のように変形することができ…

$$\lambda^{(a_0+N/2)-1}\exp\left\{-\left(b_0+\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right)\lambda\right\}$$

従ってこれは、パラメータを次のように設定したときのガンマ分布 $\mathrm{Gam}(\lambda \mid a_N, b_N)$ であることがわかる。

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{N}{2}\sigma_{ML}^2$$

ここで σ_{ML}^2 は分散の最尤推定量である。

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{N}{2}\sigma_{ML}^2$$

この事後分布のパラメータより以下のことがいえる。

- N 個のデータ点を観測すると、係数 a は$N/2$ 増える。
  - したがって、事前分布のパラメータ a_0 は $2a_0$ の「有効な」観測点が事前にあることを示す、と解釈できる。
- N 個のデータ点は $N\sigma_{ML}^2/2$ だけパラメータ b に影響を与える。

こうしたガンマ分布や（前回出てきた）ディリクレ分布などの指数型分布族では、一般的に共役事前分布を有効な仮想データ点と解釈できる。

同様に
•1変数のガウス分布
•平均と精度の両方が未知
•この両方をベイズ的に推定
という場合は共役事前分布として以下のようなガウス-ガンマ分布を用いる。

$$p(\mu,\lambda) \propto \mathcal{N}(\mu \mid \mu_0, (\beta\lambda)^{-1}) \mathrm{Gam}(\lambda \mid a, b)$$

また、D 次元変数の多変量ガウス分布 $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ で精度が既知の場合、
事前分布は以下のウィシャート分布を用いる。

$$\mathcal{W}(\boldsymbol{\Lambda} \mid \mathbf{W}, \nu) = B \mid \Lambda \mid^{(\nu-D-1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right)$$

さらに、平均と精度の両方が未知の場合、事前分布として以下の
ガウス-ウィシャート分布を用いる。

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \boldsymbol{\mu}_0, \beta, \mathbf{W}, \nu) = \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, (\beta\boldsymbol{\Lambda})^{-1}) \mathcal{W}(\boldsymbol{\Lambda} \mid \mathbf{W}, \nu)$$

詳細についてはpp.98-100を参照のこと。基本的な考え方はこれまでと同様である。

1.2. スチューデントのt分布

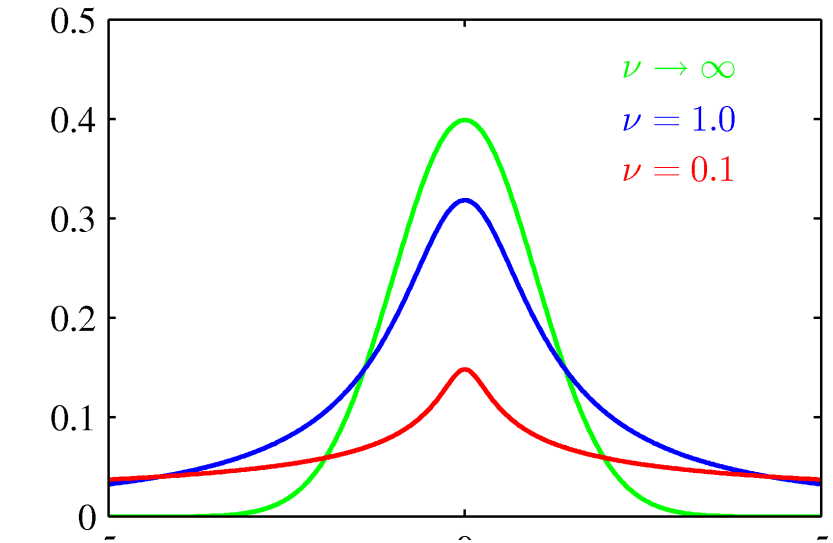
$$\mathrm{St}(x \mid \mu, a, b) = \int_0^\infty N(x \mid \mu, (\eta\lambda)^{-1}) \mathrm{Gam}(\eta \mid \tfrac{\nu}{2}, \tfrac{\nu}{2}) d\eta$$

$$= \frac{\Gamma(\frac{\nu}{2} + \frac{1}{2})}{\Gamma \frac{\nu}{2}} \left(\frac{\lambda}{\pi\nu} \right)^{1/2} \left[1 + \frac{\lambda(x-\mu)^2}{\nu} \right]^{-\frac{\nu}{2} - \frac{1}{2}}$$

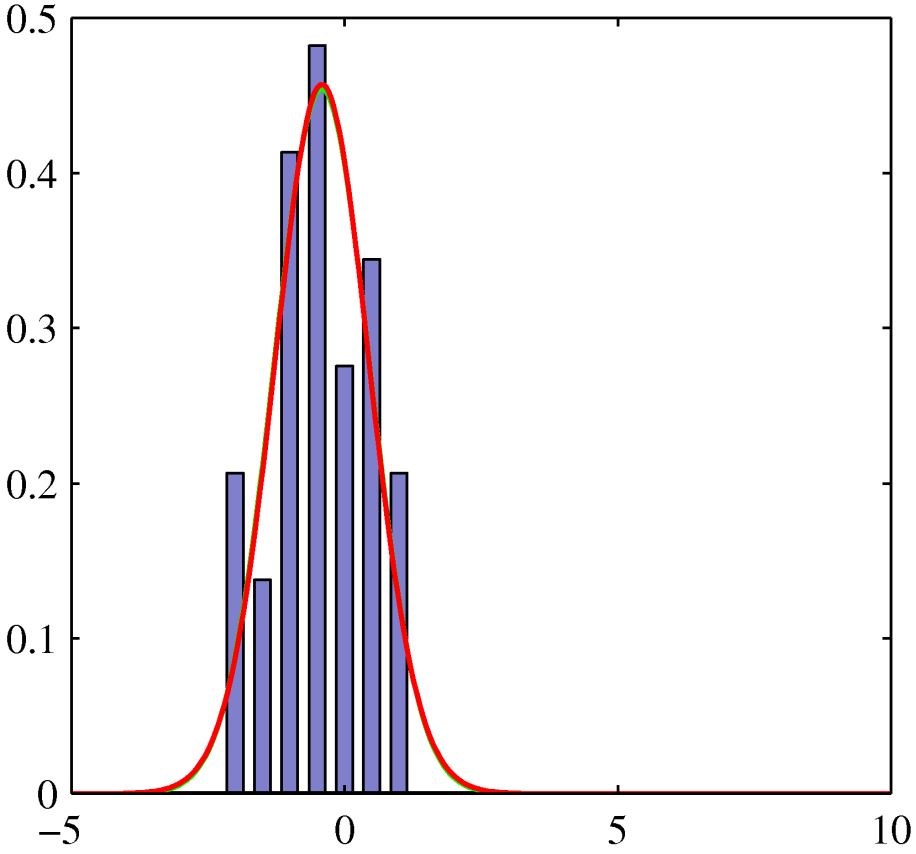
上式で表される分布をスチューデントのt分布と呼ぶ。これはガンマ分布に尤度関数をかけたものから精度を積分消去した結果として得られ、λをt分布の精度、νを自由度と呼ぶ。特にν=1のとき、これをコーシー分布と呼ぶ。また、$\nu \to \infty$ の極限ではガウス分布と一致する。

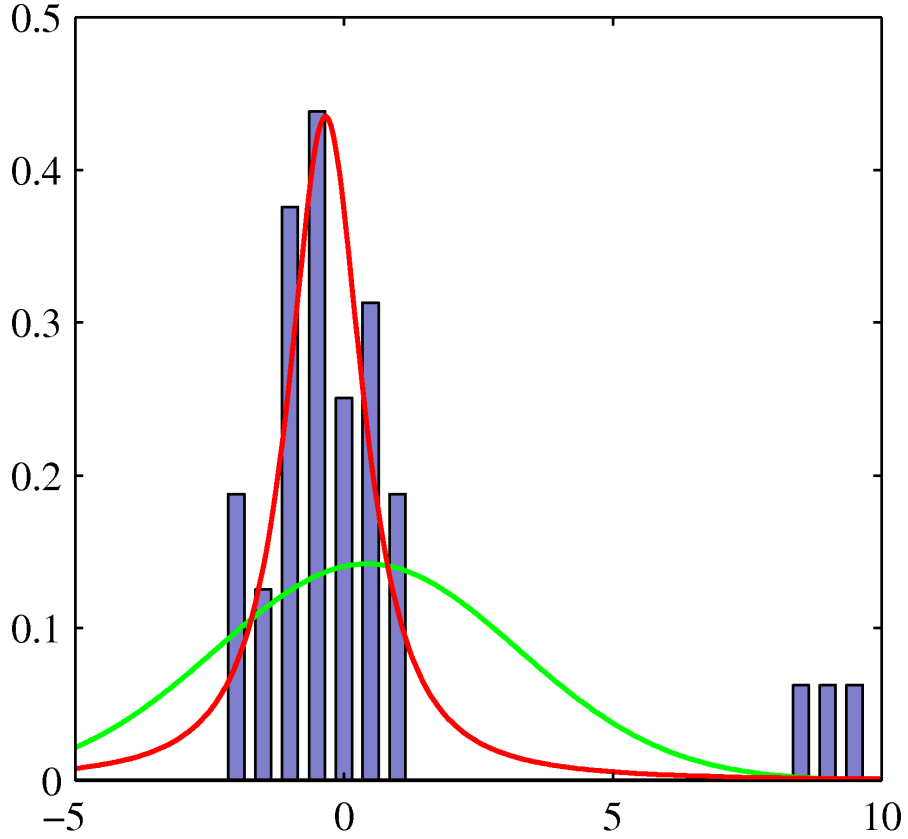
積分消去の過程から分かるように、スチューデントのt分布は、平均は同じだが精度が異なるようなガウス分布を無限解足し合わせた無限混合分布である。

t分布は頑健性、すなわち外れ値に影響されにくいという重要な性質をもつ。



緑線がガウス分布（= t分布でν→∞の極限）赤/青線がt分布である。
下図はデータに対する最尤フィッティングの結果。頑健性が示されている。





1.3. 周期変数

ガウス分布の周期変数への応用、例えば
•風向の分布
•24時間や1年といった時間的周期を持つ量のモデル化
…といった量は、角座標 $0 \leqq \theta \leqq 2\pi$ を用いると便利に表現できる。

しかし、単純にある方向を原点に選んだ周期関数を使ってガウス分布を適用するだけではうまくいかない。

(例) $\theta_1=1^\circ$ ，$\theta_2=359^\circ$ の2つの観測値があるとき…
•原点を0° に選ぶと… 平均が180°，標準偏差が179°
•原点を180° に選ぶと… 平均が0°，標準偏差が1°

原点のとりかたによって結果に大きな違いが出てしまう。

…つまり、周期関数を扱うためには、特殊な方法が必要ということ！

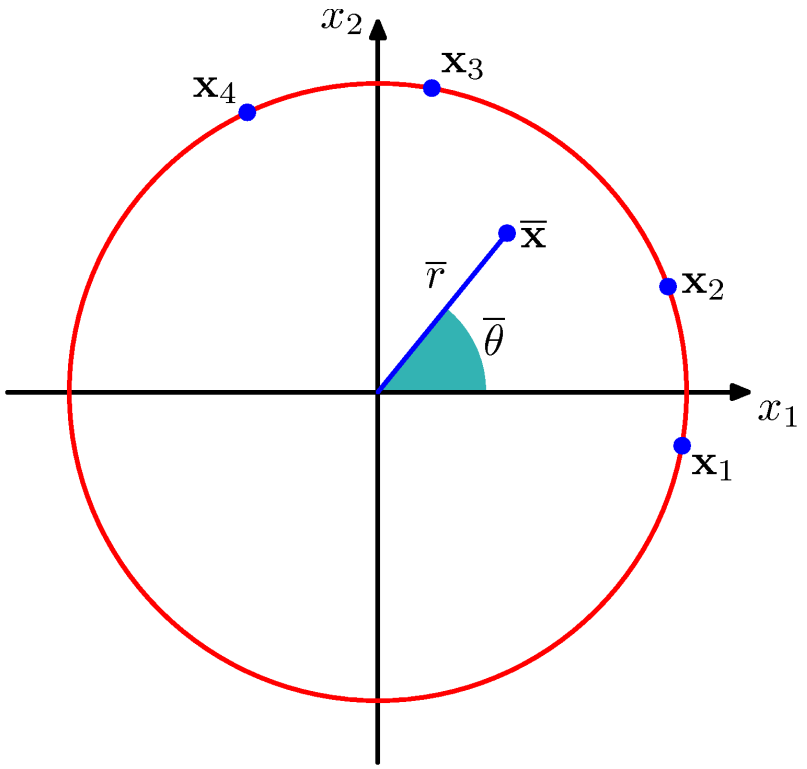
そこで…

周期変数の観測値の集合 $\mathcal{D}=\{\theta_1 \ldots \theta_n\}$の平均を求める際に、これを下図のような2次元単位ベクトル $x_1 \ldots x_n$ で表す。

角度の平均の代わりにこれらのベクトル $\{x_n\}$ の平均、すなわち

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

を求め、これに対応する角度を求める。



すなわち $x_n = r\cos\theta$, $y_n = r\sin\theta$ より、

$$\bar{\theta} = \tan^{-1}\left\{\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n}\right\}$$

と表すことができる。これは明らかに原点のとり方によらない。

では、周期変数上のガウス分布はどのようになるのか？
…ここで出てくるのが以下で導出するフォン・ミーゼス分布。

まず、求めたい分布は以下の条件（非負、積分して1、周期が2π）を満たさなければならない。

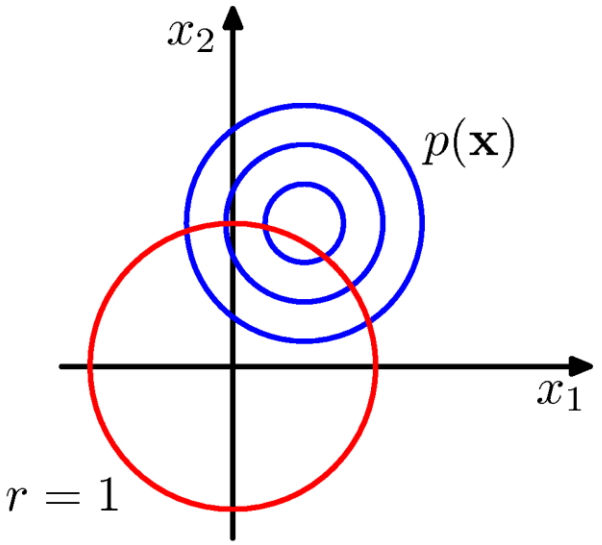
$$p(\theta) \geq 0$$

$$\int_0^{2\pi} p(\theta) d\theta = 1$$

$$p(\theta + 2\pi) = p(\theta)$$

ここで、θ が x_1, x_2 という2つのパラメータで表現されていたことを利用して、2次元ガウス分布（ただし、2変数が独立で、分散が等しいとする）を考えると、下図のような等高線を持つ平面上の分布となる。

$$p(x_1, x_2) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2}{2\sigma^2}\right\}$$



$$x_1 = r\cos\theta \qquad \mu_1 = r_0\cos\theta_0$$
$$x_2 = r\sin\theta \qquad \mu_2 = r_0\sin\theta_0$$

r=1 (単位円！)であることに注意して極座標に変換すると、指数部分は...

$$-\frac{(x_1-\mu_1)^2+(x_2-\mu_2)^2}{2\sigma^2}$$
$$=-\frac{(r\cos\theta-r_0\cos\theta_0)^2+(r\sin\theta-r_0\sin\theta_0)^2}{2\sigma^2}$$
$$=\frac{2r_0(\cos\theta\cos\theta_0+\sin\theta\sin\theta_0)-(1+r_0^2)}{2\sigma^2}$$
$$=\frac{r_0}{\sigma^2}\cos(\theta-\theta_0)+const.$$

と変形される。ここで $const.=-(1+r_0^2)/2\sigma^2$ で、$const.$は θ とは独立な項である。
そのためこれを指数部分の係数の一部として分離してよい。ただしこの係数は、θ についての積分を1にするために適切な正規化係数として適切に設定される必要があることに注意。

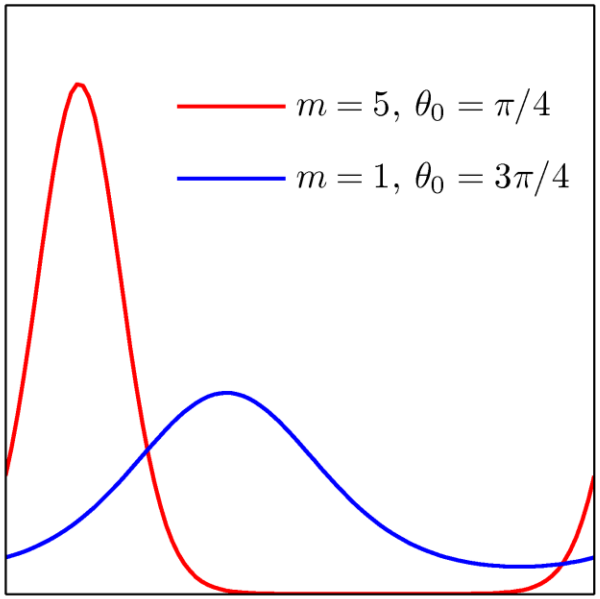
ここで $m=r_o/\sigma^2$ とおくと、結局 p は次のようなフォン・ミーゼス分布で表される。

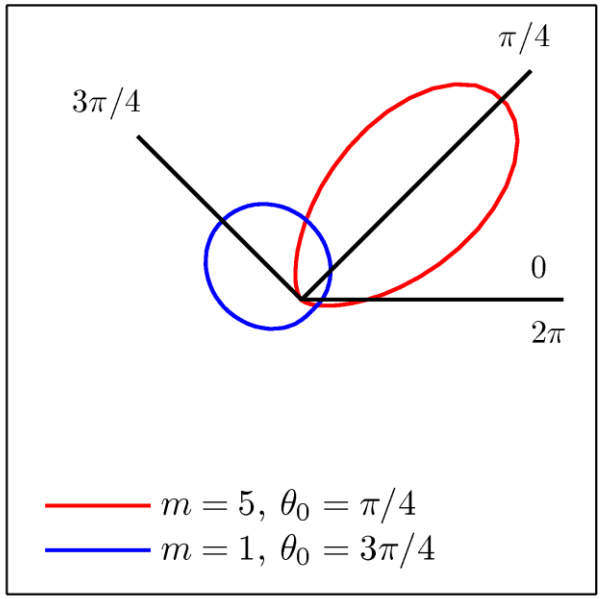
$$p(\theta \mid \theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\{m\cos(\theta - \theta_0)\}$$

ここで正規化係数 $I_o(m)$ は以下のようなo次の第1種変形ベッセル関数(なんだそれ)で、

$$I_0(m) = \frac{1}{2\pi}\int_0^{2\pi} \exp\{m\cos\theta\} d\theta$$

で定義される。
ここで θ_0 は平均に相当し、m は集中度パラメータ(≒精度＝逆分散)と呼ばれる。
この p は充分に大きな m に対しては近似的にガウス分布となる(らしい)。
p を縦軸に、θ を横軸にとったものが左図、極座標 (p, θ) で図示したものが右図である。





次に最尤推定量を求める。

ここで対数尤度関数 $\ln p$ は以下のように表される。

$$\ln p(\mathcal{D}\,|\,\theta_0, m) = -N\ln(2\pi) - N\ln I_0(m) + m\sum_{n=1}^{N}\cos(\theta_n - \theta_0)$$

ここで θ_o についての導関数をoとおくと次式を得る。

$$\sum_{n=1}^{N}\sin(\theta_n - \theta_0) = 0$$

$$\sum_{n=1}^{N}(\sin\theta_n\cos\theta_0 - \cos\theta_n\sin\theta_0) = 0$$

$$\cos\theta_0\sum_{n=1}^{N}\sin\theta_n = \sin\theta_0\sum_{n=1}^{N}\cos\theta_n$$

これを θ_o について解くと、以下の最尤解を得る。これは先ほどの平均と同じ形である。

$$\theta^{ML} = \tan^{-1}\left\{\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n}\right\}$$

また、m についての最尤解も求めたいのだが...

$$\ln p(\mathcal{D} \mid \theta_0, m) = -N \ln(2\pi) - N \ln I_0(m) + m \sum_{n=1}^{N} \cos(\theta_n - \theta_0)$$

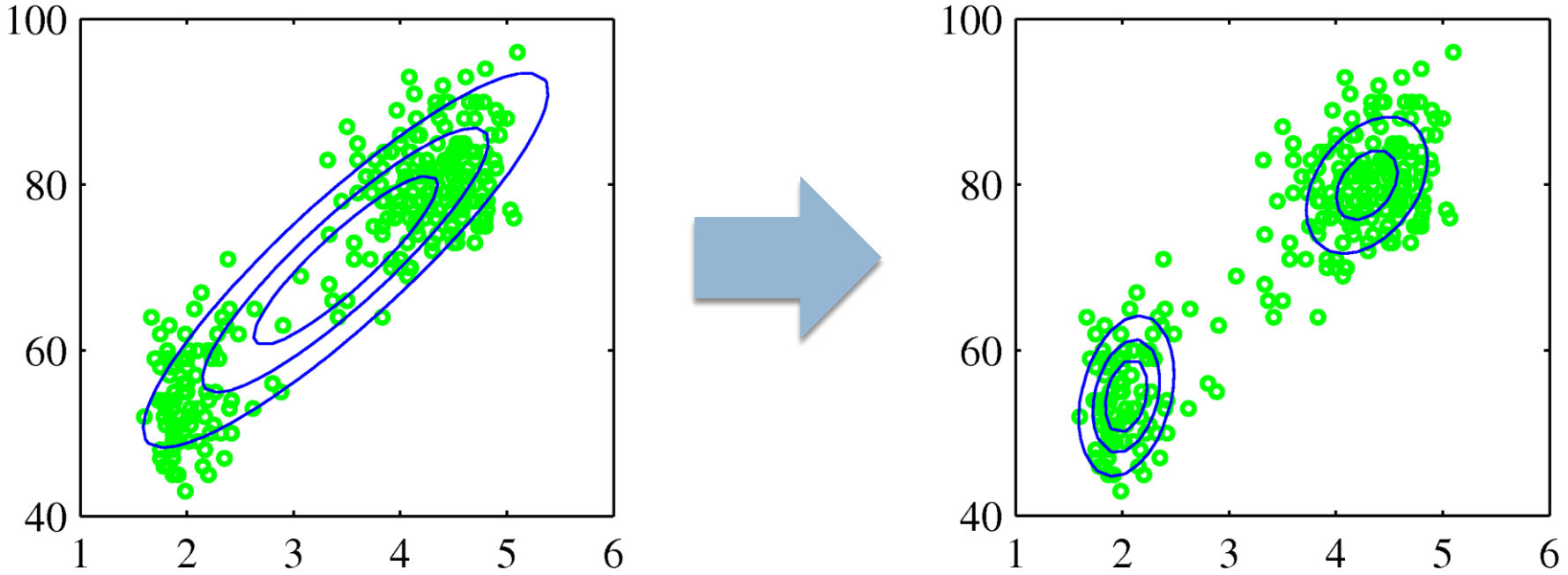
どうもこの導出は難しいらしく(Abramowitz and Stegun 1965)、結果だけ示されていました。すなわち

$$\frac{I_0'(m_{ML})}{I_0(m_{ML})} = \frac{1}{N} \sum_{n=1}^{N} \cos(\theta_n - \theta_{ML})$$

これは比較的容易に、数値的に求めることが可能であるらしい。です。

1.4. 混合ガウス分布

左図のようなデータ分布は、単一のガウス分布ではうまく捉えることができない。



しかし、右図のような2つのガウス分布の線形結合を用いることで、このデータ分布の特徴をよく表すことができる！！

…ということで、次はいくつかの分布を線形結合してつくる混合分布についてです。

このように、十分な数のガウス分布を用い、線形結合する重みの係数と平均、共分散を調節すれば、ほぼ任意の連続な密度関数を任意の精度で近似することができる。

このような混合ガウス分布の一般形は

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$$

で表される。

この分布を構成する各々のガウス分布は混合要素と呼ばれる（もちろん、一般の混合分布の混合要素はガウス分布に限られない）。また、重み付けのためのパラメータ π_k を混合係数と呼び、正規化のため以下の条件を満たさなければならない。

$$\sum_{k=1}^{K} \pi_k = 1$$

ただしこのとき、各々の混合分布が正規化されており、また、すべての k について $\pi_k \geqq 0$ を満たしている必要がある。すなわち $0 \leqq \pi_k \leqq 1$ である。

以上のような混合係数の条件(0以上1以下で総和が1)から、混合係数もまた確率の条件を満たしていることがわかる。

従って $\pi_k = p(k)$をk番目の混合要素を選択する事前確率とし、$\mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) = p(\mathbf{x}|k)$ をk が与えられたときの $\mathbf{x}$ の条件付き密度と考えれば、$p(\mathbf{x})$ は $\mathbf{x}$ の周辺密度として与えられ、

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)\, p(\mathbf{x}\,|\,k)$$

と表される。当然これは当初の $p(\mathbf{x})$ についての式に等しい。

ここで事後確率 $p(k|\mathbf{x})$ は負担率としても知られ、重要な役割を果たす。らしい。
この負担率を求めるには単純にベイズの定理を用いればよく、

$$p(k\,|\,\mathbf{x}) = \frac{p(k)\, p(\mathbf{x}\,|\,k)}{\sum_l p(l)\, p(\mathbf{x}\,|\,l)}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}\,|\,\mu_k, \mathbf{\Sigma}_k)}{\sum_l \pi_l \mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu}_l, \mathbf{\Sigma}_l)}$$

となる。

ただしこのような混合分布についての最尤解は、もはや closed form の解析解では得られない（対数尤度関数の内部に混合要素についての和がある！）。

$$\ln p(\mathbf{X} \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \mu_k, \Sigma_k) \right\}$$

そのため、パラメータ推定には繰り返し的な数値最適化法を用いるか、EMアルゴリズムを用いることになる。

混合分布やEMアルゴリズムの詳細については9章で。

2.1. 指数型分布族

これまで出てきた確率分布は（混合ガウス分布を除いて）指数型分布族と呼ばれる分布の大きな族の例となっている。

「指数型分布族」とは次式で定義される分布の集合である。

$$p(\mathbf{x}|\eta) = h(\mathbf{x})g(\eta)\exp\{\eta^T\mathbf{u}(\mathbf{x})\}$$

ここで $\mathbf{x}$ はスカラーでもベクトルでも、また離散でも連続でもよい。
また、η は分布の自然パラメータと呼ばれ、$\mathrm{u}(\mathbf{x})$ は $\mathbf{x}$ の任意の関数。関数 $g(\eta)$ は分布を正規化するための係数である。

正規分布はもちろんのこと、ベルヌーイ分布や多項分布、ディリクレ分布、ベータ分布、t分布等はいずれもこの指数型分布族に属している。

以下ではこれらの分布（だけでなく、上式の形をしたあらゆる分布）をひっくるめて、一般的に扱えることについて見ていきますよ。

というわけで、最尤推定によって指数型分布族の一般形のパラメータベクトル η を推定する問題を考える。まず一般の指数型分布 $p(\mathbf{x}|\eta)$ について次式が成り立つことは明らかである。

$$g(\eta)\int h(\mathbf{x})\exp\{\eta^{\mathrm{T}}\mathbf{u}(\mathbf{x})\}d\mathbf{x}=1$$

η について両辺の勾配をとると

$$\nabla g(\eta)\int h(\mathbf{x})\exp\{\eta^{\mathrm{T}}\mathbf{u}(\mathbf{x})\}d\mathbf{x}+g(\eta)\int h(\mathbf{x})\exp\{\eta^{\mathrm{T}}\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})d\mathbf{x}=0$$

従って

$$-\frac{1}{g(\eta)}\nabla g(\eta)=g(\eta)\int h(\mathbf{x})\exp\{\eta^{\mathrm{T}}\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})d\mathbf{x}$$

となる。右辺は$\mathbf{u}(\mathbf{x})$ の期待値と見なせるからこれを $\mathbf{E}[\mathbf{u}(\mathbf{x})]$ とおき、左辺が対数の勾配の形をとっていることに注意すると、次式を得る。

$$-\nabla \ln g(\eta)=\mathbf{E}[\mathbf{u}(\mathbf{x})]$$

したがって指数型分布族では、$\ln g(\eta)$ の負の勾配が $\mathbf{u}(\mathbf{x})$ の期待値となる。

（ちなみに、もういちど勾配をとることでこれが $\mathbf{u}(\mathbf{x})$ の共分散となり、同様により高次のモーメントを求めることもできる。すなわち $\ln g(\eta)$ は$\mathbf{u}(\mathbf{x})$ のモーメント母関数となっている）

ベルヌーイ分布や多項分布が指数型分布族に属することの確認は pp. 111-113 を見てもらうとして…とりあえずここでは1変数ガウス分布の例のみを紹介する。

$$p(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\}$$

このガウス分布を、次式で表される指数型分布族の一般形と比較すると…

$$p(\mathbf{x} \mid \eta) = h(\mathbf{x}) g(\eta) \exp\{\eta^T \mathbf{u}(\mathbf{x})\}$$

以下のように対応付けができる。（$g(\eta)$はなんでこうなるのかよくわかんない）

$$\eta = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} \qquad h(x) = (2\pi)^{-1/2}$$

$$\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \qquad g(\eta) = (-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right)$$

独立で同分布に従うデータの集合 $\mathbf{X}=\{x_1 \ldots x_N\}$ がある場合、これに対する尤度関数は...

$$p(\mathbf{X}\,|\,\eta)=\left(\prod_{n=1}^{N}h(\mathbf{x}_n)\right)g(\eta)^N\exp\left\{\eta^T\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)\right\}$$

$$\ln p(\mathbf{X}\,|\,\eta)=\sum_{n=1}^{N}\ln h(\mathbf{x}_n)+N\ln g(\eta)+\eta^{\mathrm{T}}\sum_{N=1}^{N}\mathbf{u}(\mathbf{x}_n)$$

という形になる。
これを最大化する η を求めるために η についての勾配をoとおくと、最尤解 η_{ML} が満たすべきは以下のように求まる。

$$-\nabla\ln g(\eta)=\frac{1}{N}\sum_{N=1}^{N}\mathbf{u}(\mathbf{x}_n)$$

したがって、最尤推定の解は、データに $\Sigma_n\mathbf{u}(\mathbf{x}_n)$ を通じてのみ依存することがわかる。
このように $\Sigma_n\mathbf{u}(\mathbf{x}_n)$ を十分統計量と呼び、最尤推定の解を求めるためにはデータ集合全体を保持する必要はなく、この値だけを保持しておけばよい、というものである。

例えば、ガウス分布では $\mathbf{u}(x)=(x, x^2)$ であるが、$\{x_n\}$ の和と $\{x_n^2\}$ の和の両方を保持する必要がある、ということである。

また、一般の指数型分布族に対する共役事前分布* は次式で表される。

$$p(\eta \mid \mathcal{X}, \nu) = f(\mathcal{X}, \nu) g(\eta)^{\nu} \exp\{\nu \eta^{\mathrm{T}} \mathcal{X}\}$$

ただし、$f(\mathcal{X},\nu)$ は正規化係数である。

これに先ほどの尤度関数をかければ以下のような事後分布が得られる。（ただし、正規化係数は除いてある）

$$p(\eta \mid \mathbf{X}, \mathcal{X}, \nu) \propto g(\eta)^{\nu+N} \exp\left\{\eta^{\mathrm{T}}\left(\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}) + \nu \mathcal{X}\right)\right\}$$

これは確かに事前分布と同じ形になっている。

* 例えば、ベルヌーイ分布に対してはベータ分布。ガウス分布の平均についてはガウス分布、精度についてはウィシャート分布であった。

2.2. 無情報事前分布

ベイズ推論では、事前にある知識を事前分布として便宜的に表現することでこれを利用できる。しかし一方で、分布がよくわからん場合には事後分布への影響がなるべく少なくなるような事前分布を選びたい。

このような場合に用いられるのが無情報事前分布である。

単純に考えれば「一様分布を使いたい* 」と思うかもしれないが…

- パラメータが有限個の値しか取りえないような離散型確率変数であれば、特に問題はない。
- しかし、連続型のパラメータだと…

* パラメータ λ で定められる分布$p(x|\lambda)$ に対して事前分布 $p(\lambda)=\text{const.}$ とすること

連続型のパラメータ（λとする）に対する事前分布として一様分布を選ぶと、以下のような問題がおこる。

1. （λの定義域が有界でないなら）λ上での積分が1にならず発散してしまうため、正規化できない！確率分布は積分が1でなければならないはず！
2. 非線形な変数変換をしたときの確率密度の変化に起因する問題がある。

1. については…

こうした正則化できない事前分布は変則事前分布と呼ばれているが、そこから得られる事後分布が正則化できるならば、使ってもよい。

(例)

$$p(\mu) = \mathcal{N}(\mu \,|\, \mu_0, \sigma_0^2)$$

ガウス分布の平均についての事前分布である上式のガウス分布について、$\sigma_0^2 \to \infty$ とする場合を考えれば、これは均一な分布となる。（そしてこれは、変則である）

…すると事後分布の平均と分散はそれぞれ以下のようになり、事後分布は正則化できるのだ！

$$\mu_N = \frac{\sigma_0^2 N \mu_{ML} + \sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2} \to \mu_{ML}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \to \frac{N}{\sigma^2}$$