Chapter 2. Probability Distributions

> [p67] One role for the distributions discussed in this chapter is to model the probability distribution $p(x)$ of a random variable x

この感覚は大事な気がする．なんというか「正規分布やポアソン分布からデータが発生している」とはあんまり思っていなくて，完全に把握することは不可能な未知の(複雑な)分布 $p(x)$ からデータが発生していて，それをモデル化(あくまで近似)するための分布として，正規分布やポアソン分布を使う，と思っている感じ．

> [p68] In a frequentist treatment, we choose specific values for the parameters by optimizing some criterion, such as the likelihood function. By contrast, in a Bayesian treatment we introduce prior distributions over the parameters and then use Bayes' theorem to compute the corresponding posterior distribution given the observed data.

この頻度主義とベイズ主義の対比，分かりやすい．ベイズ主義では事前情報(主観,知見)を事前分布を通して使うが，頻度主義では一切使わずデータだけを頼る．

> [p68] We shall see that an important role is played by conjugate priors, that lead to posterior distributions having the same functional form as the prior, and that therefore lead to a greatly simplified Bayesian analysis.

共役事前分布はあまりやってこなかったから，重視して読む．

2.1. Binary Variables

ベルヌーイ分布と二項分布については，メモ省略．

2.1.1. The beta distribution

話の流れは...

1. データ集合が小さいと，最尤法では非常に過学習してしまう (偶然 3 枚のコインが表だったら今

後全て表が出ると未来永劫表が出ると予測されてしまう) ことがある.
2. これを避けるため，事前情報(常識)を使ってベイズ主義的に推定したい.
3. そのためにはまず事前情報をどういう事前分布で表せば良いか決めないといけない.
4. 事前分布を決める指針(望ましい性質)として，共役性つまり「事後分布と事前分布の関数型が同じになる」ってのがある．(なぜ嬉しいかは例えば後述の逐次学習を参照)
5. ここで，ベイズの定理より「パラメータ事後分布 $\propto$ 尤度 $\times$ パラメータ事前分布」である.
6. 尤度関数(今回の例だと $\prod \mu_n^x (1-\mu)^{1-x_n}$) は既知なので，こういう事前分布なら事後分布も同じ形になるな，ってやつを探そう.
7. 今回の例(ベルヌーイ分布の μ の推定)では，ベータ分布がまさにそれ.

という感じ.

事後分布 (2.17) の導出を丁寧めに書くと,

$$
\begin{aligned}
p(\mu|m,l,a,b) &= \frac{p(\mu,m,l|a,b)}{p(m,l)} \\
&= \frac{p(m,l|\mu)\, p(\mu|a,b)}{p(m,l)} \\
&\propto p(m,l|\mu)\, p(\mu|a,b) \\
&= {}_N C_m\ \mu^m\ (1-\mu)^l\ \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \\
&\propto \mu^{m+a-1}(1-\mu)^{l+b-1}
\end{aligned}
$$

となり,

> [p72] We see that (2.17) has the same functional dependence on μ as the prior distribution, reflecting the conjugacy properties of the prior with respect to the likelihood function. Indeed, **it is simply another beta distribution**, and its normalization coefficient can therefore be obtained by comparison with (2.13) to give

$$
p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a-1}(1-\mu)^{l+b-1} \qquad (2.18)
$$

という感じで簡単に事後分布が求まる．さらに嬉しいのは,

> [p72] We see that the effect of observing a data set of m observations of $x = 1$ and l observations of $x = 0$ has been to increase the value of a by m, and the value of b by l, in going from the prior distribution to the posterior distribution.

ということ．つまり，新たに得たデータを使って事後分布に更新するには，クラス1の個数とクラス0の個数の分だけベータ分布のパラメータを変えれば良いだけ．(なので最初のベータ分布のパラメータは，事前に何回ずつ表裏が出ているか，みたいな意味合いになる．未知なら$a = b = 1$が無難．)

この利点を活かし「データを観測して事後分布更新 (ベータ分布パラメータ更新)」を繰り返すことを **逐次学習** (図2.2を見るとイメージつきやすい) と呼び,

> [p73] They can be used, for example, in real-time learning scenarios where a steady stream of data is arriving, and predictions must be made before all of the data is seen. Because they do not require the whole data set to be stored or loaded into memory, sequential methods are also useful for large data sets.

などの現実的な大きいメリットがある.

次に，クラスの事後予測密度 (2.19) を丁寧めに導出してみると，

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x, \mu|\mathcal{D})\, d\mu \\
&= \int p(x|\mu, \mathcal{D})\, p(\mu|\mathcal{D})\, d\mu \\
&= \int p(x|\mu)\, p(\mu|\mathcal{D})\, d\mu \\
&= \int \mu^x\, (1-\mu)^{1-x}\, p(\mu|\mathcal{D})\, d\mu \\
&= \begin{cases} \int \mu\, p(\mu|D)\, d\mu & [x=1] \\ \int (1-\mu)\, p(\mu|D)\, d\mu & [x=0] \end{cases} \\
&= \begin{cases} E[\mu \sim p(\mu|D)] & [x=1] \\ 1 - E[\mu \sim p(\mu|D)] & [x=0] \end{cases} \\
&= \begin{cases} \frac{m+a}{m+a+l+b} & [x=1] \\ 1 - \frac{m+a}{m+a+l+b} & [x=0] \end{cases}
\end{aligned}
$$

という感じになる．最後の等号では，事後分布(2.18)がベータ分布であることとベータ分布の期待値(2.15)を使っている．さらにこの結果について

> [p73] Note that in the limit of an infinitely large data set $m, l \to \infty$ the result (2.20) reduces to the maximum likelihood result (2.8).

とあるが，これは $m, l \to \infty$ だと a, b が相対的ににゴミみたいに小さくなって，

$$
p(x|\mathcal{D}) \approx \begin{cases} \frac{m}{m+l} & [x=1] \\ 1 - \frac{m}{m+l} & [x=0] \end{cases}
$$

となり頻度主義で最尤推定した時のクラス予測分布と一致する，てことを言ってる．
別の捉え方としては「パラメータ事後分布(ベータ分布)の分散が $\to 0$ となって最尤推定量(ベータ分布期待値)のとこをピークとした超鋭い分布になる．これってパラメータを最尤推定してるのとほぼ同じだよね」という感じ．これについて，

> [p73] it is a very general property that the Bayesian and maximum likelihood results will agree in the limit of an infinitely large data set

と言ってる．直感的に，観測されたデータが増えればパラメータの不確実性(分散)は減る感じがするし，まあ納得できる．

2.2. Multinomial Variables

ノーテーションが若干違ってくるだけで，２クラスのときと同じことが成り立つ．

ざっくりした対応表としては，

クラス数	単一の試行の結果	複数回の試行の集計	共役事前分布
2クラス	ベルヌーイ分布	二項分布	ベータ分布
多クラス	カテゴリカル分布	多項分布	ディリクレ分布

という感じ.

話の流れをまとめると...

1. 「ある 2 パターンの事象のどちらか一方だけが起こる」というケースは 1 次元のダミー確率変数で表現でき，モデル化する際はベルヌーイ分布が便利.
2. 「$K(\geq 3)$ パターンの事象のいずれか 1 つが起こる」というケースはどういう確率変数(事象の写像)で表現すれば良いだろうか．(2値のケースを包含して拡張させたい)
3. K 次元確率変数ベクトルの one-hot encoding (2.25式) で表現するのが分かりやすい.
4. モデル化する際はカテゴリカル分布 (2.26式) が便利.
5. このカテゴリカル分布のパラメータ(各カテゴリの発生確率)の最尤推定量は，標本比率である (2.33式)．また，十分統計量は各カテゴリの頻度.
6. ベルヌーイ分布の和として二項分布を導入したのと同様に，カテゴリカル分布に i.i.d. な N 個の確率変数ベクトルの和が従う分布として，多項分布っていうのを導入しよう (2.34式)．K パターンの結果が想定される試行を独立に N 回繰り返した時の発生カテゴリ度数が従うのが，多項分布.
7. 応用上，多項分布のパラメータを 1 個の観測値ベクトルから推定(⇔カテゴリカル分布のパラメータを N 個の観測値ベクトルから推定)したいケースは多い.
8. 頻度論的な最尤推定量は先ほど述べたように標本比率 (2.33式) であるが，コイン投げの例で触れたような過適合を防ぎたいので，事前情報を使ってベイズ的に推定したい.
9. そのためにはパラメータベクトル(各カテゴリの真の発生確率)の事前分布を考える必要があり，とりあえず共役なものを探したい.
10. 尤度関数 (2.34式) を見ると明らかなように，定数部分を除いて (2.37) のような密度関数を持つものなら共役になれる.
11. これを全範囲積分してそれをもとに正規化すると (2.38) のようになり，これはディリクレ分布と呼ばれる.
12. ベイズの更新式 (2.40) より，この共役事前分布を使えば事後分布(観測データが与えられたもとでのパラメータ分布)も (2.41) のディリクレ分布となる.
13. この結果はつまり「事前分布としてディリクレ分布を使う場合，データが観測されたら各カテゴリの発生度数をディリクレ分布パラメータに足し合わせるだけで，事後分布を得ることができる」ということを示している．データが系列的に何度も観測される場合でも，この操作を繰り返すだけで，逐次的な学習(パラメータ推定)が可能.
14. このことから分かるように，ディリクレ分布の(ハイパー)パラメータ $\boldsymbol{\alpha}$ は「過去に各カテゴリが何回ずつ発生した(と想定する)か」という事前情報として捉えられる.

という感じ.

> [p77] Note that two-state quantities can either be represented as binary variables and modelled using the binomial distribution (2.9) or as 1-of-2 variables and modelled using the multinomial distribution (2.34) with $K = 2$.

これは少し重要．数式を見て考えれば分かるが，事象をどういう確率変数(写像)で表現するかが違うだけで，カテゴリカル分布はベルヌーイ分布を, 多項分布は二項分布を, ディリクレ分布はベータ分布を完全に包含していて，多変数にそのまま拡張した形になっている．なので当然，分布の性質や推定方法 (頻度主義的な最尤推定量，ベイズ主義的な事後分布や逐次学習) も，全く同じものになっている．

ニューラルネットは多クラス分類(手書き数字認識とか)によく使われてきたので，2クラスから多クラスへ拡張って感じじゃなくて，最初から多クラスの one-hot encoding の形で説明されることが多いのだと思う．ゼロから DL 本とか．

> [p77] Plots of the Dirichlet distribution over the simplex, for various settings of the parameters α_k, are shown in Figure 2.5.

まず，3変数のディリクレ分布 (3クラスの発生確率の事前分布) の観測値 $\left(\mu_1, \mu_2, \mu_3\right)$ は，必ず図 2.4 の赤い三角形の面の上にある (そういう性質の分布だから)．当然だが3つの角と軸の交点の値は 1 である．この三角形を x-y 平面にパタっと寝かせて z 軸方向に密度関数値を表したのが，図 2.5 になってる．この形は直感的で「3クラスが1回ずつ観測されただけだと真の発生確率は全然見当もつかないが，3クラスがちょうど10回ずつ観測されたら真の発生確率は大体 1/3 ずつでしょ」という感じになってる．

3変数じゃなくて2変数のディリクレ分布 (2クラスの発生確率の事前分布) を考えると，三角形じゃなくて線分(１辺)の上に (μ_1, μ_2) が乗ることになり，この線分上で密度関数値を考えると，図2.2の形(ベータ分布)になるはず．

2.3. The Gaussian Distribution

ガウス分布の定義

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

ここで，D は確率変数ベクトルの次元で，Σ は正定値対称行列．

(2.53)~(2.64)で示されているように，パラメータ $\boldsymbol{\mu}$ と Σ はこの分布の期待値ベクトルと分散共分散行列である．

また，書かれていないけど「それぞれ 1 変量ガウス分布に従っているような独立な確率変数群の同時分布は，多変量ガウス分布となる」てのもけっこう重要な気がする．

ガウス分布の幾何的な形状

定義から明らかなように，ガウス分布の密度は $\boldsymbol{x}$ と $\boldsymbol{\mu}$ とのマハラノビス距離 ($\Sigma = I$ のときはユークリッド距離)

$$\Delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \qquad (2.44)$$

の単調減少関数である．$\boldsymbol{x}$ はこれを通してのみ密度に影響するので，このマハラノビス距離が等しい $\boldsymbol{x}$ では密度も等しくなり，等高線みたいな感じになる．

ここで，Σ は実対称行列なので，D 個の実数固有値(重解含む) $\lambda_1, \ldots, \lambda_D$ に対応する固有ベクトルとして，正規直交 (2.46式) な $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_D$ を選ぶことができる．(対称行列 $\subset$ 正規行列では異なる固有値に対応する固有ベクトルは直交するので，重解がなければ自動的に固有ベクトルが全て直交する．n 重解な固有値があってもそれに対応し直交する n 本の固有ベクトルが存在するので，それを選べば全て直交する固有ベクトルを取得できる．)

直交($\Rightarrow$ 線形独立)な D 本の固有ベクトルが得られているので，それを使って $U^{-1}\Sigma U = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$ と対角化でき，これを固有値分解の形 $\Sigma = U\mathrm{diag}(\lambda_1, \ldots, \lambda_D)U^{-1}$ で見てブロックで考えると，

$$\Sigma = \sum_{i=1}^{D} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T \qquad (2.48)$$

のように表せる．また，同様に

$$\begin{aligned}
\Sigma^{-1} &= (U^{-1})^{-1}\mathrm{diag}(\lambda_1, \ldots, \lambda_D)^{-1}U^{-1} \\
&= U\mathrm{diag}(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_D})U^{-1} \\
&= \sum_{i=1}^{D} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^T \qquad (2.49)
\end{aligned}$$

と表せる．この (2.49) を (2.44) に代入すると

$$
\begin{aligned}
\Delta^2 &= (\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\
&= (\boldsymbol{x}-\boldsymbol{\mu})^T\left(\sum_{i=1}^{D}\frac{1}{\lambda_i}\boldsymbol{u}_i\boldsymbol{u}_i^T\right)(\boldsymbol{x}-\boldsymbol{\mu}) \\
&= \sum_{i=1}^{D}\left[\frac{1}{\lambda_i}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{u}_i\boldsymbol{u}_i^T(\boldsymbol{x}-\boldsymbol{\mu})\right] \\
&= \sum_{i=1}^{D}\left[\frac{1}{\lambda_i}\left\{\boldsymbol{u}_i^T(\boldsymbol{x}-\boldsymbol{\mu})\right\}^T\boldsymbol{u}_i^T(\boldsymbol{x}-\boldsymbol{\mu})\right] \\
&= \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i} \qquad (2.50)
\end{aligned}
$$

ここで，$y_i = \boldsymbol{u}_i^T(\boldsymbol{x}-\boldsymbol{\mu})$ と置いていて，すると，前述のように正規直交基底 $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_D$ からなる U は直交行列なので,

$$
\begin{aligned}
&(y_1,\ldots,y_D)^T = (\boldsymbol{u}_1,\ldots,\boldsymbol{u}_D)^T(\boldsymbol{x}-\boldsymbol{\mu}) \\
&\Leftrightarrow\ \boldsymbol{y} = U^T(\boldsymbol{x}-\boldsymbol{\mu}) \\
&\Leftrightarrow\ \boldsymbol{y} = U^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \qquad (2.52)
\end{aligned}
$$

と表すことができる．すなわち，$\boldsymbol{x}$ を「原点(座標軸の交点)を $\boldsymbol{\mu}$ に平行移動し，座標軸(基底ベクトル)を固有ベクトル $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_D$ に変更(回転)した新しい座標系」で表したのが $\boldsymbol{y}$ である.

(2.50) より，この新しい座標系において，密度関数の等高線が(平行移動や回転ナシの)楕円形になっていることが分かる．また，固有値ベクトル(座標軸) $\boldsymbol{u}_i$ 方向における楕円の幅は対応する固有値を使って $2(\Delta^2\lambda_i)^{1/2}$ となっている．この様子を可視化のが図 2.7 である．(Σ は正定値行列としているので $\lambda_i > 0$ であり問題ない．)

この結論(図 2.7)は，主成分分析とも若干関係がありそう.
第 1 主成分ベクトル(最も標本分散を最大化する軸,基底ベクトル,線形結合の係数ベクトル)は標本分散共分散行列の最大固有値に対応する固有ベクトルだった.
ガウス分布の場合も，最大固有値に対応する固有ベクトルの方向において，楕円等高線の幅(分散)が一番広くなっている.
ついでに思ったが，分散 ≈ エントロピーなので，主成分分析は「観測値から得られる平均的な情報量がなるべく大きい軸を探す(変数を作る)」と捉えられる.

補足と復習：座標変換について
例えば，3本の標準基底ベクトル (普通の x,y,z 軸) で表現されている3次元ベクトル $\boldsymbol{x}$ を，自分で用意した3本の基底ベクトル(軸) $\boldsymbol{a}_1,\boldsymbol{a}_2,\boldsymbol{a}_3$で表現したいとする．この3軸で表現した時の座標(成分) は,

$$
\begin{aligned}
\boldsymbol{x} &= c_1\boldsymbol{a}_1 + c_2\boldsymbol{a}_2 + c_3\boldsymbol{a}_3 \\
&= (\boldsymbol{a}_1,\boldsymbol{a}_2,\boldsymbol{a}_3)(c_1,c_2,c_3)^T \\
\Leftrightarrow (c_1,c_2,c_3)^T &= (\boldsymbol{a}_1,\boldsymbol{a}_2,\boldsymbol{a}_3)^{-1}\boldsymbol{x}
\end{aligned}
$$

という感じで，基底ベクトル(軸)を列に持った行列の逆行列を掛けることで取得できる.
ここで，$\boldsymbol{a}_1,\boldsymbol{a}_2,\boldsymbol{a}_3$ が正規直交基底 (長さ1で直交している軸)の時は $(\boldsymbol{a}_1,\boldsymbol{a}_2,\boldsymbol{a}_3)$ が直交行列となりその逆行列は転置行列なので,

$$
(c_1,c_2,c_3)^T = (\boldsymbol{a}_1,\boldsymbol{a}_2,\boldsymbol{a}_3)^T\boldsymbol{x}
$$

という感じで求められる.

ガウス分布の制限

1つめの制限として,

> [p83] A general symmetric covariance matrix Σ will have $D(D+1)/2$ independent parameters, ... For large D, the total number of parameters
> therefore grows quadratically with D, and the computational task of manipulating and inverting large matrices can become prohibitive.

が挙げられている. 対策として, モデルの柔軟性を犠牲にして $\Sigma = \mathrm{diag}(\sigma_i^2)$ や $\Sigma = \sigma^2 I$ という制約を課すことが挙げられている.
例えば $[\mathrm{diag}(d_1, \ldots, d_p)]^{-1} = \mathrm{diag}(\frac{1}{d_1}, \ldots, \frac{1}{d_p})$ という感じで, 大きい行列でも即座に逆行列を求められるし, 積とかの計算も楽.

こういう「計算コストとモデルの柔軟性のトレードオフ」みたいなものは, ガウス分布だけでなく色々ありそう.

2つめの制限として,

> [p84] A further limitation of the Gaussian distribution is that it is intrinsically unimodal (i.e., has a single maximum) and so is unable to provide a good approximation to multimodal distributions.

が挙げられている.

上記２つの制限について

> [p84] Thus the Gaussian distribution can be both too flexible, in the sense of having too many parameters, while also being too limited in the range of distributions that it can adequately represent.

とまとめていて, 2.3.9 節の潜在変数を用いた混合ガウス分布で, 両方の解決を目指すらしい.

2.3.1. Conditional Gaussian distributions

2.3.2. Marginal Gaussian distributions

結論として,

> [p85] An important property of the multivariate Gaussian distribution is that if two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian. Similarly, the marginal distribution of either set is also Gaussian.

という事実が重要.
図 2.9 が分かりやすく，確かにガウスになってそうと思える．(条件付き分布と周辺分布の概念の違いを理解する上でも良い図で，いつも頭の中でイメージしてたやつと同じグラフ．)

具体的なパラメータ値は (2.94)~(2.98) を参照 (導出は瀬尾先生の授業で少しやった).
この式について,

- $\boldsymbol{x}_a$ の条件付き期待値は与えられた確率変数ベクトル $\boldsymbol{x}_b$ の線形関数(アフィン変換=線形変換+平行移動)になっている.
- $\boldsymbol{x}_a$ の条件付き分散は与えられた確率変数ベクトル $\boldsymbol{x}_a$ に依存せず独立である.
- 条件付き分散は精度行列のブロックを使うと表現しやすいが，一方，周辺分散は共分散行列のブロックを使った方が表現しやすい.

などとコメントされている.
前半２つについては，図 2.9 からも納得できる．x_b の値を動かして(赤線を動かしてどこの断面を見るかを変えて)条件付き平均をプロットすると直線になりそう．そして，x_b の値がなんであろうと(赤線をどこに引いてもどこの断面を見ても)，条件付き分散は変わらなそう.

2.3.3. Bayes’ theorem for Gaussian variables

前との繋がりや話の流れをまとめると...

1. 前節では，まず２つの確率変数ベクトル $\boldsymbol{x}_a, \boldsymbol{x}_b$ の同時分布 $p(\boldsymbol{x}_a, \boldsymbol{x}_b)$ として多変量ガウス分布を用意した．その仮定のもとで色々と調べていき，以下のような結果を得られた.
    - 周辺分布 $p(\boldsymbol{x}_b)$ もガウスで，期待値と分散はもとの同時分布の一部を切り取ってきたもの.
    - 条件付き分布 $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ もガウスで，条件付き期待値は与えられた確率変数ベクトル $\boldsymbol{x}_b$ の線形関数(アフィン変換)，条件付き分散は $\boldsymbol{x}_b$ とは独立.
2. この節では逆に，１での結果(を一般化したもの)を事前の仮定としたとき，どのような結果が得られるかを考える．(ベイズ線形回帰とかで便利な結果が得られるから考えてるってのもある．)
3. 具体的には，２つの確率変数ベクトル $\boldsymbol{x}, \boldsymbol{y}$ について，次の仮定をおく.
    - $p(\boldsymbol{x})$ はガウス分布であるとする．(2.99, 2.113式)
    - 条件付き分布 $p(\boldsymbol{y}|\boldsymbol{x})$ はガウス分布であるとし，条件付き期待値は $\boldsymbol{x}$ の任意の線形関数(アフィン変換)で条件付き分散は $\boldsymbol{x}$ とは独立であるとする．(2.100, 2.114式)
4. 同時分布を求めて，周辺化して，条件付き分布を求めて，という手順を踏むと，次のような結果が得られる.
    - 同時分布 $p(\boldsymbol{x}, \boldsymbol{y})$ はガウス分布で，平均と分散は (2.108),(2.105) 式で与えられる.
    - 与えられていない方の $\boldsymbol{y}$ の周辺分布 $p(\boldsymbol{y})$ は (2.115) のガウス分布となる.
    - 逆の(結果が与えられた時の原因の)条件付き分布 $p(\boldsymbol{x}|\boldsymbol{y})$ は (2.116) のガウス分布となる.

5. ベイズっぽく見ると，(2.114)の尤度 $p(\boldsymbol{y}|\boldsymbol{x})$ と，(2.113) のガウス事前分布 $p(\boldsymbol{x})$ から，事後分布 (2.115) を導いたということ．それがガウスだったので，(2.114) の尤度に対して (2.115) のガウス分布が共役事前分布だということが分かったことになる．
6. 実際 1.2.5 でやったベイズ的な多項式回帰は，$\boldsymbol{y}$ が 1次元のときの例になってる．尤度(1.61) が (2.114) に対応し，事前分布 (1.65) が (2.113) に対応する．あのときはパラメータ事後分布まで考えず最大値だけ見て「MAP推定量がリッジ推定量と等価」という話で終わらせたが，ガウス分布になってたのか．

2.3.4. Maximum likelihood for the Gaussian

多変量ガウス分布において，平均ベクトルの MLE は標本平均ベクトル (2.121)，分散共分散行列の MLE は標本分散共分散行列 (2.122) である．これの導出は瀬尾先生の授業でやった気がする．

平均ベクトルでは MLE が不偏推定量にもなってるが，分散共分散行列では MLE は不偏推定量になってない．この辺は 1 変数と同じ．

2.3.5. Sequential estimation

Robbins-Monro アルゴリズムは本書ではもう使われないし，一旦流れやモチベだけ抑える．

1. 前の 2.1.1 で，逐次的なベイズ推定について，このようなメリットが挙げられていた．過去のデカいデータを DB に取っておいたり，それを推定値算出のためにメモリに載せる必要がないよ，ということ．

   > [p73] They can be used, for example, in real-time learning scenarios where a steady stream of data is arriving, and predictions must be made before all of the data is seen. Because they do not require the whole data set to be stored or loaded into memory, sequential methods are also useful for large data sets.

2. 最尤推定を逐次的にやりたいケースももちろんある．同じメリットあるし．
3. 例えばガウス分布の母平均の最尤推定 (つまり標本平均ベクトルの算出) は，明らかなように，(2.126) のように逐次的に行うことができる．
4. 他のケースの最尤推定ではどうなんだろうか．ガウスほど簡単にはいかなそう．
5. 一般の最尤推定に使える汎用的な逐次学習のアルゴリズムとして，Robbins-Monro アルゴリズムがある．

2.3.6. Bayesian inference for the Gaussian

平均のベイズ推定(分散は既知)

この節は 2.1.1 の「二項分布(ベルヌーイ分布)のパラメータ(母比率)のベイズ推定」と対応させて理解すべき．流れは同じで,

0. (まず分散は既知として平均の推定を考える)
1. 最尤推定では，観測されたデータのみ(尤度だけ)をもとに推定を行う．(特に小標本の時に極端な結論が導かれる可能性がある．)
2. 事前情報(主観,知見)も取り入れてベイズ的に推定したい時もある.
3. そのために，事前情報を表現するパラメータ事前分布を決めないといけない．扱いやすいので,とりあえず共役事前分布を探そう.
4. 事後密度 $\propto$ 尤度関数 $\times$ 事前密度 と既に決まっている尤度関数(2.137)から考えて，ガウス分布なら μ の共役事前分布になりそう．(いまは分散既知として平均μ だけ推定しようとしているので，μ だけを変数と見て考えれば良い)

という感じ.

（今さら思ったが，最尤推定では尤度関数の最大化を考えているのに対し，ベイズ MAP 推定ではそこに事前密度を掛けた 事後密度 $\propto$ 尤度関数 $\times$ 事前密度 の最大化を考えている．客観的に尤度だけを見るか，そこに主観的な事前密度を取り入れたものを見るか，の違い．）

実際に共役であることを確かめていくと,

$$
\begin{aligned}
p(\mu|\mathbf{x}) &\propto p(\mathbf{x}|\mu)\, p(\mu) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}\exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}x_i^2-2\mu\sum_{i=1}^{N}x_i+N\mu^2\right)-\frac{1}{2\sigma_0^2}\left(\mu^2-2\mu\mu_0+\mu_0^2\right)\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2\sigma_0^2}\left[(N\sigma_0^2+\sigma^2)\mu^2-2\left(\sigma_0^2\sum_{i}^{N}x_i+\sigma^2\mu_0\right)\mu+\left(\sigma_0^2\sum_{i=1}^{N}x_i^2+\sigma^2\mu_0^2\right)\right]\right\} \\
&= \exp\left\{-\frac{N\sigma_0^2+\sigma^2}{2\sigma^2\sigma_0^2}\left[\mu^2-2\left(\frac{\sigma_0^2\sum_{i}^{N}x_i+\sigma^2\mu_0}{N\sigma_0^2+\sigma^2}\right)\mu+\left(\frac{\sigma_0^2\sum_{i=1}^{N}x_i^2+\sigma^2\mu_0^2}{N\sigma_0^2+\sigma^2}\right)\right]\right\} \\
&= \exp\left\{-\frac{N\sigma_0^2+\sigma^2}{2\sigma^2\sigma_0^2}\left[\mu-\left(\frac{\sigma_0^2\sum_{i}^{N}x_i+\sigma^2\mu_0}{N\sigma_0^2+\sigma^2}\right)\right]^2+\mathrm{Const}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)\left[\mu-\left(\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0+\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}\right)\right]^2\right\} \\
\Rightarrow\ p(\mu|\mathbf{x}) &= \mathcal{N}\left[\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0+\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}},\ \left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)^{-1}\right]
\end{aligned}
$$

という感じで，(2.140) ~ (2.143) を導出できた.

この事後分布からベイズ主義と頻度主義の関係を考察していて,

> [p98] First of all, we note that the mean of the posterior distribution given by (2.141) is a compromise between the prior mean μ_0 and the maximum likelihood solution μ_{ML}. If the number of observed data points $N = 0$, then (2.141) reduces to the prior mean as expected. For $N \to \infty$, the posterior mean is given by the maximum likelihood solution.

と述べられている．これは 2.1.1 (p71~74) のベータ分布の時に，同じような話をしている．
二項分布の比率パラメータも正規分布の平均パラメータも,

- 事後分布の期待値は，事前分布の期待値(主観的な推定値)と最尤推定量(客観的な推定値)の間の値となっている．サンプルサイズ N が主観と客観のどっちをどれくらい重視するかに関わってくる.
- 事後分布の分散は，サンプルサイズ N が大きくなるにつれて小さくなっていく．これはデータが多いほどパラメータ推定の精度(信頼度)が上がることを表している.
- $N \to \infty$ のとき，事後分布は最尤推定量のところに集中した尖った分布となる．つまりベイズ推定と最尤推定の結果が一致する.

という感じの性質を持つ．ベータ分布による比率推定では図 2.2 が，正規分布による平均推定では図 2.12 が分かりやすい.

また，ベータ分布による比率推定でもそうであったように「追加で N 個目の観測値 x_N が得られた時，最初の事前分布 $p(\mu)$ を $x_1, x_2, \ldots, x_N$ で更新して求められる事後分布と，$N-1$ 個目の観測値を使って得られていた事後分布 $p(\mu|x_1, \ldots, x_{N-1})$ を事前分布と思ってそれを x_N で更新した事後分布は，一致する」というのが即座に示せる (2.18, 2.144)．これは(共役事前分布を使った)逐次的ベイズ推定の話で,

> [p99] We have already seen how the maximum likelihood expression for the mean of a Gaussian can be re-cast as a sequential update formula ... In fact, the Bayesian paradigm leads very naturally to a sequential view of the inference problem. ... This sequential view of Bayesian inference is very general and applies to any problem in which the observed data are assumed to be independent and identically distributed.

という感じで推されている．ベイズ推定では最尤推定よりも直感的に逐次学習を導ける．(2.126)とか Robbins-Monro アルゴリズムを考えるより「データが観測される度にその尤度でパラメータ分布を更新する」と考えて逐次学習を構築したほうが自然で分かりやすいでしょ，という主張かな.

分散のベイズ推定(平均は既知)

流れは大体同じで,

1. 今度は平均を既知として分散をベイズ推定することを考える．計算の簡便さのため精度パラメータ$\lambda = \frac{1}{\sigma^2}$を推定することとする.
2. 尤度関数を精度 λ の関数と意識して眺めると(2.145式)，λ の共役事前分布としてガンマ分布 (2.146式) が使えそうだとわかる.
3. 実際に計算すると(2.149式)，確かに事後分布もガンマ分布になっていて，パラメータを (2.150),

(2.151)式のように更新すれば良いと分かる.

という感じ.

平均の推定に使った事前ガウス分布では，その(ハイパー)パラメータの意味は直感的に分かりやすかった(ガウスだから)．ただ，分散の推定に使おうとしているガンマ事前分布の(ハイパー)パラメータ a_0, b_0 の意味はパッと分からない．これについて，

> [p101] From (2.150), we see that the effect of observing N data points is to increase the value of the coefficient a by $N/2$. Thus we can interpret the param2eter a_0 in the prior in terms of $2a_0$ 'effective' prior observations. Similarly, from (2.151) we see that the N data points contribute $N\sigma^2_{ML}/2$ to the parameter b, where σ^2_{ML} is the variance, and so we can interpret the parameter b_0 in the prior as arising from the $2a_0$ 'effective' prior observations having variance $2b_0/(2a_0) = b_0/a_0$.

と説明されているが，これを少し解説.

1. 二項(多項)分布のパラメータ推定に使ったベータ(ディリクレ)事前分布のハイパーパラメータは，その更新式に着目することで解釈できた．各クラスが事前に何回ずつ観測された(と想定する)かという意味.
2. ガンマ事前分布でも同様に，パラメータの更新式 (2.150), (2.151) に着目しよう.
3. まず，パラメータ a はデータが観測される度に $+N/2$ される．なので，最初の事前分布に設定するパラメータ値 a_0 は「事前に観測された(と想定する)サンプルサイズ $/2$」として解釈できる．別の言い方をすると $2a_0$ 'effective' prior observations となる.
4. 次に，パラメータ b はデータが観測される度に $+N\sigma^2_{ML}/2$ される．ここでσ_{ML} は標本分散.なので，最初に設定する b_0 は「事前に観測された(と想定する)データのサンプルサイズ$\times$その標本分散$/2$」として解釈できる.
5. 3 と 4 を合わせて(3を4に代入して式変形して)解釈すると， a_0, b_0 によって「事前に標本分散が b_0/a_0 な $2a_0$ 個のデータが観測されていた」という事前情報が表されていることになる．これを $2a_0$ 'effective' prior observations having variance $2b_0/(2a_0) = b_0/a_0$ と言ってる.

この解釈を図 2.13 のガンマ分布と見比べると納得できる．$a = 4, b = 6$ は「事前に 8 個のデータが観測されて標本分散が $6/4$ だった」という状況で，ちゃんと精度(分散の逆数)が $4/6$ あたりのところに分布の山がある．$a = 1, b = 1$ だとサンプルサイズ 2 ってことでまだ信頼度あまりないので $1/1$ のとこに山はできていない.

さらに一般化して

> [p101] we shall see that the interpretation of a conjugate prior in terms of effective fictitious data points is a general one for the exponential family of distributions.

となるらしい．ベータ分布($\subset$ディリクレ分布)でもガンマ分布でも，共役事前分布として使った場合にパラメータが「事前に観測された(と想定する)データの数」みたいな意味を持ってた．ディリクレならカテゴリ度数，ガンマならサンプルサイズ．こういう特徴は，ここに限った話じゃなく指数型分布族を共役事前分布として使った場合に一般的に成り立つらしい.

平均と分散の両方をベイズ推定(両方とも未知)

この問題設定でも流れは同様で,

1. 平均 μ と精度 λ の両方が未知でベイズ推定したい場合,
   まずはパラメータベクトル (μ, λ) の(2変量)共役事前分布を探すべき.
2. 尤度関数を μ, λ の関数だと意識して眺めると(2.153式), ガウス-ガンマ分布(2.154式, 図2.14) が使えそうだとわかる.

という感じ. この説では, 事後分布(への更新式)の導出まではやってない.

多変量ガウス分布のパラメータのベイズ推定

ここまでは 1 変数ガウス分布の話だったが, ここでは多変量ガウス分布のパラメータのベイズ推定について説明されている.

共役事前分布についてまとめると,

尤度(データ発生分布)	推定したいパラメータ	共役事前分布
1 変数ガウス分布	平均 μ	1 変数ガウス分布
1 変数ガウス分布	精度 λ	ガンマ分布
1 変数ガウス分布	μ と λ を同時に	ガウス-ガンマ分布
多変量ガウス分布	平均ベクトル $\boldsymbol{\mu}$	多変量ガウス分布
多変量ガウス分布	分散共分散行列 Σ	ウィシャート分布
多変量ガウス分布	$\boldsymbol{\mu}$ と Σ を同時に	ガウス-ウィシャート分布
ベルヌーイ(二項)分布	発生確率 μ	ベータ分布
カテゴリカル(多項)分布	クラス発生確率ベクトル $\boldsymbol{\mu}$	ディリクレ分布

となる.

2.3.7. Student's t-distribution

ここまではパラメータ事前分布 $p(\theta)$ とパラメータ事後分布 $p(\theta|\mathcal{D})$ だけを議論してきた. パラメータ推定にはこれで十分だが「将来にどういうデータが発生するか」という予測をするには, データ予測事後分布 $p(x|\mathcal{D})$ を知って導出する必要がある.

予測分布 $p(x|\mathcal{D})$ は, 同時分布を $p(x, \theta|\mathcal{D}) = p(x|\theta)\, p(\theta|\mathcal{D})$ で求めてこれを θ で積分して周辺化することで得られる.

まず，ガウス分布の平均 μ を共役事前ガウス分布でベイズ推定した場合を考える(分散は既知).

μ の事後分布は (2.140)~(2.142)のガウス分布になるので,
$p(x, \mu|\mathbf{x}) = p(x|\mu)\, p(\mu|\mathbf{x}) = \mathcal{N}(x|\mu, \sigma^2)\, \mathcal{N}(\mu|\mu_N, \sigma_N^2)$ を μ で全範囲積分すれば予測分布 $p(x|\mathbf{x})$ が偉られる．この節では示されていないが，この予測分布はガウス分布となる.

次に，ガウス分布の精度 λ を共役事前ガンマ分布でベイズ推定した場合を考える(平均は既知).

λ の事後分布は (2.149)~(2.145) のガンマ分布になるので,
$p(x, \lambda|\mathbf{x}) = p(x|\lambda)\, p(\lambda|\mathbf{x}) = \mathcal{N}(x|\mu, \lambda^{-1})\, \mathrm{Gam}(\lambda|a_N, b_N)$ を λ で全範囲積分して周辺化すれば予測分布 $p(x|\mathbf{x})$ が得られる．これを計算しているのが (2.158) 式で，結果として予測分布は (2.159) 式の t 分布となることが示されている.

μ, λ, υ の３つのパラメータを持つ (2.159) は，仮説検定で使ってたような自由度パラメータ υ のみの t 分布を一般化したもので 一般化 t 分布 と呼ばれているらしい．(2.164),(2.165)より μ, λ は期待値と分散に関するパラメータであることがわかる．自由度 υ は分布の裾の長さに関するパラメータで, $\upsilon = 1$ ではコーシー分布(裾が長く期待値すら存在しない)となり，$\upsilon \to \infty$ では裾の短い正規分布に漸近する．図 2.15 が分かりやすい.

t 分布とガウス分布の関係として,

> [p103] From (2.158), we see that Student's t-distribution is obtained by adding up an infinite number of Gaussian distributions having the same mean but different preci- sions. This can be interpreted as an infinite mixture of Gaussians.

が挙げられている．t 分布は「$\mathcal{N}(x|\mu, \lambda^{-1})\, \mathrm{Gam}(\lambda|a_N, b_N)$ を λ で全範囲積分して周辺化」という操作で得られると言ったが，この操作は積分を和で捉えれば「平均 μ を固定して精度 λ を少しずつずらした無限個のガウス分布を，ガンマ密度で重み付けしたうえで，足し合わせる」と解釈できる．重みの総和(ガンマ分布の全範囲積分)はちゃんと 1 になるし，t 分布は無限混合ガウス分布みたいに解釈できる.

モデリングにおける t 分布のアドバンテージとして,

> The result is a distribution that in general has longer 'tails' than a Gaussian, as was seen in Figure 2.15. This gives the t-distribution an important property called robustness, which means that it is much less sensitive than the Gaussian to the presence of a few data points which are outliers.

が挙げられている．ガウス分布と違い，t 分布は自由度パラメータで裾の長さを調整できるので，外れ値が含まれるような分布を無理なく表現できる．はなっから外れ値が出てくるかもと想定しているので，外れ値が観測されても推定結果はそこまで変わらない(図2.16)．すなわち外れ値の影響を受けづらくロバストである.

次に，データが多変量ガウス分布から発生する場合の予測分布について考える.

この節では説明されていないが，1変数と同様に,

- 平均ベクトルを多変量ガウス事前分布を使ってベイズ推定した場合，予測分布は多変量ガウス分布
- 分散共分散行列をウィシャート事前分布を使ってベイズ推定した場合，予測分布は多変量 t 分布

となる.

2.3.8. Periodic variables

もし実務で使うことになったら，そのときに勉強.

2.3.9. Mixtures of Gaussians

9.2 節で詳細に扱われるので，そっちも一部参照しながらメモする.

混合ガウス分布は，２通りのアプローチで定式化できる．それを,

> [p430] In Section 2.3.9 we motivated the Gaussian mixture model as a simple linear superposition of Gaussian components, aimed at providing a richer class of density models than the single Gaussian. We now turn to a formulation of Gaussian mixtures in terms of discrete latent variables.

で言っている．具体的には,

- 単体のガウス分布で表現するのが難しい複雑な分布(ex.多峰)を表現するために，複数のガウス分布を重ね合わせた分布．(2.188 式)
- データの背後に潜在的なクラスが複数存在し，それぞれが異なるガウス分布を持つとする．そして各観測値 $\boldsymbol{x}$ は潜在的なクラスのいずれかに属し，そのガウス分布から発生するとする．このときの $\boldsymbol{x}$ の周辺分布．(2.191 式) (9.7 ~ 9.12 式)

という２通りの見方ができる.

後者の見方をすることでメリットが色々あって,

> [p431,432] It might seem that we have not gained much by doing so. However, we are now able to work with the joint distribution $p(x, z)$ instead of the marginal distribution $p(x)$, and this will lead to significant simplifications, most notably through the introduction of the expectation-maximization (EM) algorithm. Another quantity that will play an important role is the conditional probability of z given x. We shall use $\gamma(z_k)$ to denote $p(z_k = 1|x)$ ... As we shall see later, $\gamma(z_k)$ can also be viewed as the responsibility that component k takes for 'explaining' the observation x.

と説明している．簡単にまとめると,

- 潜在変数を導入したことで，広汎な応用を持つ EM アルゴリズムで自然に数値的最尤推定でき

る．(詳しくは 9 章で)
- ベイズの定理を使って「データが与えられた下での各潜在クラスへの所属確率 $p(z_k|x)$」を得ることができる．これは responsibility (負担率) と呼ばれ，ソフトクラスタリングとして捉えれば各クラスタの寄与率．なんかの知識発見とか意思決定に繋がるかもしれないし，新たに得られたデータの分類にも使える．

という感じ．

scikit learn で簡単に試してみる．