変分推論を頑張るドキュメント ver.0.3
(目的関数の定義までは一応書けてるが、実際の計算例まで追記したい)

このドキュメントは、ちょっと統計学に興味がある程度の心理学者が変分推論を理解するために作った。変分推論、あるいは変分ベイズ法は、ベイズ推測のための1手法であり、事後分布を近似的に求めるために用いられる。心理学者はあくまでユーザーサイドであると思うが、自身が使うツールについて、多少なりとも理解を深めておくのは有用だし、良心のある態度であるだろう。

しかし、それだけではなく、Fristonの自由エネルギー原理なんかでは、「(情報論的) 変分自由エネルギーを下げることが、生命が環境内で持続する条件である」とまで述べる。脳は外界についてベイズ推論しているかのように振舞っていると主張するベイズ脳仮説や、感覚入力とその予測の間の誤差を最小化する計算を神経系は行なっていると定式化する予測符号化理論は、自由エネルギー原理から得られる帰結として捉えることもできる。従って、変分推論について知ることは、単に統計学のテクニカルな側面の勉強以上のポテンシャルを秘めていると思われる[1]。

このドキュメントの目標を設定しよう。これを読めば、変分推論がどのような仮定の下どんな根拠で何をやっているのかがわかるようになることを目指す。あくまで、ユーザー側として「これくらいは知っておいたらよいだろう」という水準の理解を得るのが目的だ。なので、自分で数値計算を実装して頑張りたい人にとっては明らかに物足りない出来ではあると思う。そこは留意してほしい。

事前に必要な知識として、ベイズの定理とカルバック・ライブラー情報量については知っていることを前提とする。それ以外の知識は基本的にこのドキュメントを読めばわかるように説明する。カルバック・ライブラー情報量がわからない人は、別のドキュメントで解説を試みているので、参照してほしい[2]。なお、このドキュメントは Fox and Roberts (2012) と Galdo et al. (2019) を大いに参考にしている。

1. 問題設定：事後分布は通常解析的に求まらない

---

[1] むしろそうでないと、私はこのドキュメントは作っていないだろう。このドキュメントの作者の普段の統計学との付き合い方は、使い方と基本的な仕組みがわかっていればいいというもので、計算のアルゴリズムの詳細や、内側の理論的な側面まではあまり踏み込むことはほとんどない (なので、例えばハミルトニアンモンテカルロ法がなんなのかを説明せよ、とか言われても正直わからない。使っているくせに)。

[2] 「エントロピーを頑張るドキュメント」
https://github.com/HeathRossie/memo/blob/main/%E3%82%A8%E3%83%B3%E3%83%88%E3%83%AD%E3%83%92%E3%82%9A%E3%83%BC%E3%82%92%E9%A0%91%E5%BC%B5%E3%82%8B%E3%83%88%E3%82%99%E3%82%AD%E3%83%A5%E3%83%A1%E3%83%B3%E3%83%88.pdf

データ x を得て、そのデータがどんな分布から生み出されているのか推論するという状況を考えよう。統計モデル (尤度関数) を $p(x|\theta)$ と設定したとする。θ はモデルのパラメータとして、事前分布を $p(\theta)$ とする[3]。ようは θ も x も確率変数だと考えるわけだ。

すると、ベイズの定理より事後分布は

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \cdots (1)$$

である。これは特に深淵な話ではなく、ベイズの定理をそのまま適用したらそうなるというだけの話だ。$p(x) = \int p(x|\theta)p(\theta)d\theta$ は事後分布の積分が1になるようにするための規格化定数であるが、周辺尤度、モデルエビデンス、分配関数といくつか呼び方がある。分配関数は統計力学に由来する名前なので、これ以降出てこない。

今、何をしたいかというと、θ (の分布) を求めたいわけだ。しかし、(1) を直接求めらえるケースは現実世界にはほとんどない。なぜなら、θ が多次元であると、周辺尤度の部分が多重積分になってしまい、解析的に解けなくなるためだ。例えば、θ が2次元なら、$p(x)$ の計算は二重積分になる。それくらいならまだしも、複雑なモデルだと大変な計算になってしまう。従って、なんらかの近似計算が必要になる[4]。変分推論は、そのための近似計算法である。もっと具体的に言えば、変分推論は上記の事後分布を計算するという課題を、ある最適化問題に帰着させる。順を追って見ていこう。

2. 変分推論を頑張る：近似分布を事後分布に近づける

さて、前節では駆け足ではあるものの、事後分布 (1) を直接求めるのは困難であるためなんらかの近似が必要であるということを述べた。変分推論では、事後分布 $p(\theta|x)$ の代わりに近似分布 $q(\theta|\lambda)$ を用意して、それを $p(\theta|x)$ に近づけていく、という方策をとる。ここで λ は q のパラメータである。わざわざ厳密な事後分布 $p(\theta|x)$ を直接求めるのをやめて近似分布 q とやらを使うのだから、その分 q という分布は数学的に扱いやすくなければ意味がないだろう。

---

[3] この事前分布を「分析者の信念を反映するものだ」とする文献 (代表的なのは Gelman et al., 2013) もあれば、そのような発想は不要であるという前提で解説する文献 (例えば Watanabe, 2018) もある。しかし、このドキュメントではその点には踏み込まず、とにかくなんらかの $p(\theta)$ を仮定したとする。

[4] マルコフ連鎖モンテカルロ法という名前くらいは聞いたことがある人も多いと思うが、それもそのような近似計算のためのアルゴリズムだ。余談だが、MCMC自体が推論の方法であると思っている人がたまにいるが、正確にはそうではない。ベイズ法という統計的推測の方法があって、それを具体的に実行するための手続きとしてMCMCや変分推論が使える、という関係だ。

2.1. 近似分布に正規分布を使う

そこで、数学的な扱いやすさから、多くの場合正規分布が用いられる。このドキュメントでも近似分布 $q(\theta|\lambda)$ として を正規分布を使うことを考えてみよう。正規分布なら、パラメータは平均 μ と分散 σ^2 の2つだけで、この時点で話が簡単になる雰囲気は感じられるだろう。

1点注意をしておくと、今やろうとしていることは事後分布 $p(\theta|x)$ を $q(\theta|\lambda)$ で近似しようということである。この θ は統計モデル $p(x|\theta)$ のパラメータで「パラメータの従う分布」について近似計算をしようとしているということだ。$p(\theta|x)$ はパラメータ θ の従う分布で、データ x は $p(x|\theta)$ という統計モデルから発生していると考える[5]。例えばみんなよく知ってる線形回帰なら、データ x はなんらかの説明変数 x_{exp} の一次式で表されるので、

$$x = a x_{exp} + b + z$$

$$z \sim Normal(0, \sigma^2)$$

が統計モデルだ。パラメータは傾き、切片、分散で $\theta = \{a, b, \sigma^2\}$ である。一行でまとめると次のようにも書ける。

$$p(x|a, b, \sigma^2) = Normal(a x_{exp} + b, \sigma^2)$$

これで、$p(x|\theta)$ のイメージは明瞭になったんじゃないかと思う。では、この事後分布 $p(\theta|x)$ を求めるとはどういうことか？θ は3つのパラメータ a、b、σ^2 なわけなので、これらの同時分布を求めるということだ。慣れている人には当たり前なんだが、このドキュメントの作者は当初、統計モデル $p(x|\theta)$ と事後分布 $p(\theta|x)$ のところで目が滑りがちで、ごっちゃになってしまうことがよくあった。

2.2. 近似分布を事後分布どうやって近づけるか？

事後分布 $p(\theta|x)$ というのは、どんな形をしているのかは通常わからないし、解析的に導けることは稀である。そこで、$q(\theta|\lambda)$ という近似分布に正規分布を仮定して、それを $p(\theta|x)$ に近づけていくという方針を立てた。これが前節までで行ったことだ。「近づける」というのは、具体的には $q(\theta|\lambda)$ は正規分布なので、平均と分散をいろいろいじくりまわして、$p(\theta|x)$ に形が近くなるようにしていきたいということだ[6]。

$q(\theta|\lambda)$ を $p(\theta|x)$ に近づけるには、この2つの分布がどれくらい似ているかを表す指標が必要だろう。その類似性の指標として、変分推論ではカルバック・ライブラー情報量を用いる。

---

[5] ベイズ統計学一般の文献として、初学者でもわかりやすいものとしては、浜田・石田・清水 (2019) が良いだろう。

[6] この説明には誤魔化しがある。θ が多変量のとき、近似分布 $q(\theta|\lambda)$ は多変量正規分布として共分散を考えることもできるはずだ。だが、その点は後ほど指摘するので、ここでは一旦脇において読み進めてほしい。

$$KL(q(\theta|\lambda)\|p(\theta|x)) = \int q(\theta|\lambda)\log\frac{q(\theta|\lambda)}{p(\theta|x)}d\theta \cdots (2)$$

しかし、せっかく定義したはいいものの、このカルバック・ライブラー情報量は計算することができない。なぜなら、式の中に $p(\theta|x)$ が入っているためだ。事後分布 $p(\theta|x)$ を計算できるなら、わざわざ近似分布なんて用意しなくても最初からそっちを求めればいいのだから、このままじゃあまり意味がない。

ただ幸い、多少の式変形で、評価ができない (2) 式をデータと $q(\theta|\lambda)$ だけから計算できる指標に変えることができる。その妙技を見る前に、もう一つ、計算を簡単にするための仮定を置こう。

2.3. 平均場近似にさらに話をシンプルにする

ここまでのところ、パラメータ θ の次元についてはあまり気にせず議論を進めてきた。実際に変分推論を適用するときは、θ は多次元である。例えば、2.1.節で出てきた線形回帰であれば、傾き、切片、分散の3つのパラメータがあるので θ は3次元である。3次元の θ の事後分布 $p(\theta|x)$ を求めるというのは、つまり3つのパラメータの同時分布を求めるということだ。この場合、近似分布 $q(\theta|\lambda)$ に多変量正規分布を考えると、パラメータ間の相関を決める共分散を推定しなくてはいけなくなる。

平均場近似 (mean-field approximation) とは、この点をシンプルにする近似方法だ。具体的には、各パラメータの相関性は考えず、同時分布が単にパラメータの周辺分布の積で表せるとする。つまり、各パラメータが独立であると考えて、同時分布を作る。パラメータの次元が d 次元あるとして、

$$q(\theta|\lambda) = \prod_{i=1}^{d} q_i(\theta_i|\lambda_i) \cdots (3)$$

と、近似分布を構成する。この考え方をスキーマティックに表したのが図1だ。事後分布 $p(\theta|x)$ では、パラメータ間で相関があるかもしれない (図1a)。しかし平均場近似では、各パラメータが独立であると考える。そうすれば $p(\theta_1,\theta_2) = p(\theta_1)p(\theta_2)$ と分解できる。そんなことやってもいいんかいな、と思うかもしれないが、このように仮定することで計算が劇的に簡単になる。しかし、あくまで近似であることは意識しておきたい[7]。

---

[7] パラメータ間に相関を入れて推定する変分推論もある。方法としては、多変量正規分布の分散共分散行列を推定するという手法と、事後的にパラメータ間に相関を入れる手法がある (Galdo et al., 2019) 。聞きかじった話だが、このあたりはモリモリ新手法が開発されているらしい。将来的に、MCMCが不要になってなんでも変分ベイズで解決できるようになったら、このドキュメントの作者はとても嬉しい (変分ベイズの方が速いため)。

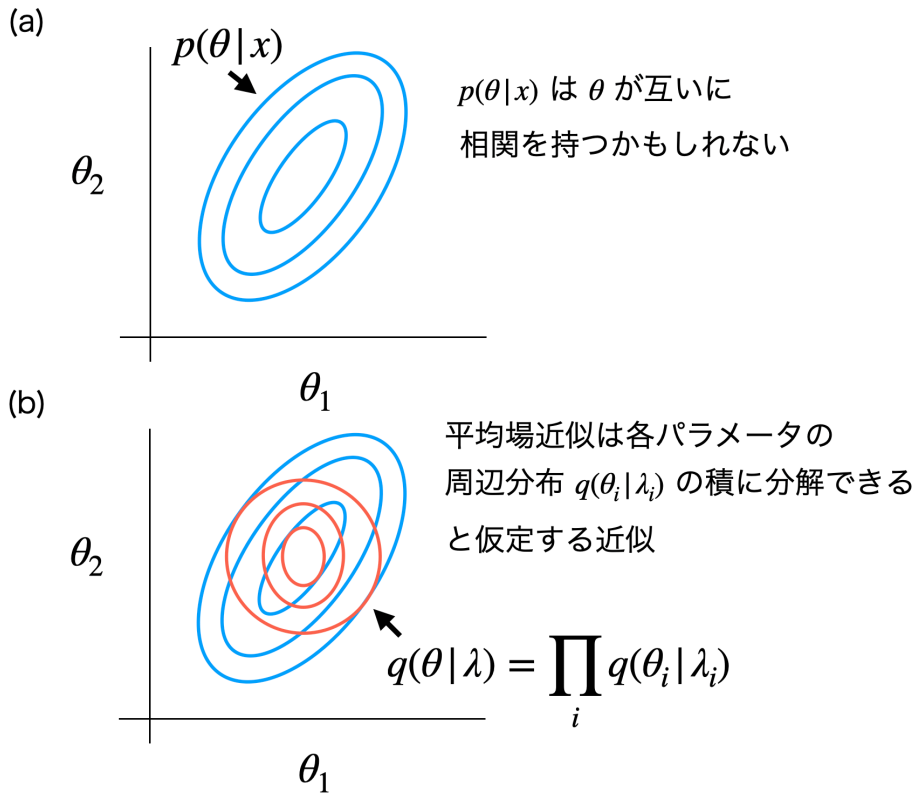

図1：平均場近似を直観的に掴む

2.1-2.3で何をやったか、一度おさらいしよう。まず、事後分布 $p(\theta|x)$ はベイズの定理から決まるが、これは直接求めるのが難しい (2.1節)。そこで、より簡単な形をしている近似分布 $q(x|\lambda)$ を代わりに用意して、$p(\theta|x)$ との間のカルバック・ライブラー情報量式 (2) 式を最小化しようという方針を立てた (2.2節)。しかし、$q(x|\lambda)$ に正規分布を仮定しても、パラメータ θ が多次元のとき、それでもやや面倒くさい。そこで、平均場近似という近似法 (3) 式を導入した (2.3節)。しかし、2.2節で述べたように、このままではまだ近似分布 $q(\theta|\lambda)$ を最適化することができない！なぜなら、(2) 式で定義されたカルバック・ライブラー情報量には事後分布 $p(\theta|x)$ が入っているため、計算できないからだ。次節ではこの点を解決し、近似分布 $q(x|\lambda)$ を事後分布 $p(\theta|x)$ に近づけられるようにする。

2.4. 変分下界とは何か

この節が最も重要だが、やることはシンプルだ。このままでは評価できないカルバック・ライブラー情報量 (2) 式をゴリゴリ変形させた結果、評価可能な量が出てくる。そいつを最適化すれば $q(\theta|\lambda)$ を $p(\theta|x)$ の近づけていることに相当する、ということをこれから示す。早速始めよう。まず、カルバック・ライブラー情報量の式をもう一度示そう。

$$KL(q(\theta|\lambda)\|p(\theta|x)) = \int q(\theta|\lambda)\log\frac{q(\theta|\lambda)}{p(\theta|x)}d\theta$$

ここでベイズの定理 (1) より $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ である。これを上式に代入する次のように変形できる。

$$KL(q(\theta|\lambda)\|p(\theta|x)) = \int q(\theta|\lambda)\log\frac{q(\theta|\lambda)p(x)}{p(x|\theta)p(\theta)}d\theta$$

$$= \int q(\theta|\lambda)\Big[\log p(x) + \log q(\theta|\lambda) - \log p(x|\theta)p(\theta)\Big]d\theta$$

$$= \int q(\theta|\lambda)\log p(x)d\theta + \int q(\theta|\lambda)\log q(\theta|\lambda)d\theta - \int q(\theta|\lambda)\log p(x|\theta)p(\theta)d\theta$$

二段目では $\log ab = \log a + \log b$、並びに$\log\frac{x}{y} = \log x - \log y$ となる対数の性質を使っている。第1項 $\int q(\theta|\lambda)\log p(x)d\theta$ の中の $p(x)$ は、θ には依存しない。よって、確率分布の定義より

$\int q(\theta|\lambda)d\theta = 1$ となるはずなので

$$\int q(\theta|\lambda)\log p(x)d\theta = \log p(x)\int q(\theta|\lambda)d\theta = \log p(x)$$

である。上式のカルバックライブラー情報量を書き換えると、

$$KL(q(\theta|\lambda)\|p(\theta|x)) = \log p(x) + \int q(\theta|\lambda)\log q(\theta|\lambda)d\theta - \int q(\theta|\lambda)\log p(x|\theta)p(\theta)d\theta \cdots (4)$$

というふうになることがわかる。ここで、第１項目はベイズの定理 (1) の分母の対数なので、周辺対数尤度であることがわかる。私たちは近似分布 $q(\theta|\lambda)$ を事後分布に近づけるに当たって、制御できるのは $q(\theta|\lambda)$ のパラメータ λ である。$p(x)$ は q に依存していないため、今は無視して構わない。つまり、(4) 式の第2項と第3項を最小化すれば、$q(\theta|\lambda)$ と $p(\theta|x)$ のカルバック・ライブラー情報量が最小になることがわかる。

文献によって表されかたが違うので、ここでもいくつか書いてみて、慣れ親しもう。

$$L'(\lambda) = \int q(\theta|\lambda)\log q(\theta|\lambda)d\theta - \int q(\theta|\lambda)\log p(x|\theta)p(\theta)d\theta$$

$$= \int q(\theta|\lambda)\log q(\theta|\lambda)d\theta - \int q(\theta|\lambda)\log p(x,\theta)d\theta$$

$$= \int q(\theta|\lambda)\log\frac{q(\theta|\lambda)}{p(x,\theta)}d\theta$$

$$= KL(q(\theta|\lambda)\|p(x,\theta)) \cdots (4)'$$

ここで定義した $L'(\lambda)$ が小さいほど $KL(q(\theta|\lambda)\|p(\theta|x))$ は小さくなるため、$L'(\lambda)$ を最小化すれば近似分布 $q(\theta|\lambda)$ を最大限事後分布 $p(\theta|x)$ に近づけられたと言える。

さらに、変分下界 (evidence lower bound; ELBO) という概念をここで導入する。カルバック・ライブラー情報量は常に0以上であるため、

$$KL(q(\theta|\lambda)\|p(\theta|x)) \geq 0$$

である。(4) 式をここに代入すれば

$$\log p(x) + \int q(\theta|\lambda)\log q(\theta|\lambda)d\theta - \int q(\theta|\lambda)\log p(x|\theta)p(\theta)d\theta \geq 0$$

$$\log p(x) \geq \int q(\theta|\lambda)\log p(x|\theta)p(\theta)d\theta - \int q(\theta|\lambda)\log q(\theta|\lambda)d\theta$$

(4)’ で定義した $L'(\lambda)$ の符号反転を

$$L(\lambda) = \int q(\theta|\lambda)\log p(x|\theta)p(\theta)d\theta - \int q(\theta|\lambda)\log q(\theta|\lambda)d\theta \quad ・・・(5)$$

とおけば、上式は次のように簡潔に表せる。

$$\log p(x) \geq L(\lambda) ・・・(6)$$

$L(\lambda)$ は $L'(\lambda)$ の正負を変えただけだ。よって、$L(\lambda)$ を最大化することが(4) 式のカルバック・ライブラー情報量を最小化することに相当するわけだが、それは周辺対数尤度 $p(x)$ の下限を押し上げていることに相当することがわかると思う。この $L(\lambda)$ を変分下界 (ELBO) という。なお、「周辺対数尤度の下限を押し上げる」と言われても、その嬉しさがよくわからないという人向けに、周辺対数尤度についての補足を4.2につけた。

また、(4)’ より

$$L(\lambda) = \int q(\theta|\lambda)\log p(x,\theta)d\theta - \int q(\theta|\lambda)\log q(\theta|\lambda)d\theta ・・・(7)$$

が成り立つのも大丈夫だろう。ついでに述べておくと、知っている人にとっては当たり前かもしれないが、x を何かの確率変数、$p(x)$ をその確率分布、$f(x)$ を x の関数とすると、$\int p(x)f(x)dx$ は $p(x)$ で $f(x)$ の期待値を取っていることになる。それをよく $\mathbb{E}_{p(x)}[f(x)] = \int p(x)f(x)dx$ と略記することがある。なので、 (7) 式は次のように表記されることもある。

$$L(\lambda) = \mathbb{E}_{q(\theta)}[\log p(x,\theta)] - \mathbb{E}_{q(\theta)}[\log q(\theta|\lambda)] ・・・(7)’$$

何によって期待値を取っているか誤解がないときは、$\mathbb{E}_{q(\theta)}[]$ を $\mathbb{E}[]$ と略記している文献もある[8]。ただ、このドキュメントでは、どんな計算をしているのかを明示したいので、見やすさは犠牲になるが $\mathbb{E}_{q(\theta)}[]$ という表記はしない。なので、これ以降この記号は出てこない。

近似分布 q は λ によって形が決まる (正規分布で近似するなら、平均と分散) のだから、ELBOを最大にするというのは、$L(\lambda)$ を最大化する λ を探すということだ。そのような λ を $\lambda*$ と書こう。この $\lambda*$ を求めることができたら、変分推論が完了したということだ。このこともあからさまに書いておこう。

$$\lambda* = \arg\max_{\lambda} L(\lambda) \cdots (8)$$

ここまでで導いてきたものを整理しよう。出発点は何であったか。事後分布 $p(\theta|x)$ は事前分布 $p(\theta)$ と統計モデル $p(x|\theta)$ を定めれば (1) として定義されるのであった。しかし、事後分布は一般的に解析的に求めることができない。故に何らかの形で近似が必要ということで、もっと簡単な形をしている近似分布 $q(\theta|\lambda)$ をできるだけ事後分布 $p(\theta|x)$ に近づけるという方策をとることにした。具体的にはカルバック・ライブラー情報量 $KL(q(\theta|\lambda)||p(\theta|x))$ を最小化すればいいのだが、この値は事後分布 $p(\theta|x)$ が項に入っているので評価できない。しかし、式を整理すると変分下界 $L(\lambda)$ を最大化することが、結局カルバック・ライブラー情報量を最小化するのに等しいことを示してきた。変分下界 $L(\lambda)$ は、統計モデル $p(x|\theta)$、事前分布 $p(\theta)$、近似分布 $q(\theta|\lambda)$ から算出できる。これらはいずれも計算できるものなので、$L(\lambda)$ は評価可能だ。これで、事後分布の導出を変分下界の最適化問題へと置き換えることに成功した。やったぜ。これが変分推論である。

実際にどのように最適化を実行するか？λ について勾配をとって、$L(\lambda)$ が最大になる点を探せばいいわけだが、それについては色々なアルゴリズムが提案されている。自分で最新手法を追いかけて数値計算を実装する、みたいなことをやる人は心理学者にはあまりいないと思う。従って、そこまで踏み込む必要性は薄い。「とにかく世界のどこかの頭のいい人が効率よく (8) 式を実現するアルゴリズムを開発したんだなあ」でも支障をきたすことは稀だろう。

とはいえ計算例があった方がいいと思うので、追記予定

補足

full-rankの変分推論

脚注7で触れたが、平均場近似せず、パラメータ間の相関を考慮する変分推論もある。例えば近似分布 $q(\theta|\lambda)$ に多変量正規分布を仮定して、平均場近似しないで変分推論する場合、パラメータは

---

[8] 個人的には原則としてちゃんと書いてほしい・・・。

$\lambda = \{\mu, \Sigma\}$ で、Σ が分散共分散行列になる。Stanならvb(algorithm = “fullrank”) でそのような近似法になる。平均場近似というのは、Σ の対角成分以外が0の場合に相当する。

故に当たり前だが、平均場近似より推定すべきパラメータは多くなるため、計算量は増える。パラメータ間の相関を考慮できると言っても、図1に示した問題が常に解決されるわけではない点にも注意しよう。パラメータ同士の相関が線型とは限らないからだ。

Reference

Fox, C. W., & Roberts, S. J. (2012). A tutorial on variational Bayesian inference. Artificial intelligence review, 38(2), 85-95.

Galdo, M., Bahg, G., & Turner, B. M. (2019). Variational Bayesian methods for cognitive science. Psychological Methods.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis. CRC press.

浜田・石田・清水 (2019). 社会科学のための ベイズ統計モデリング. 朝倉書店.
Watanabe, S. (2018). Mathematical theory of Bayesian statistics. CRC Press.