社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－HPC－116 (19)
2008／8／6

頑健で効率的なオンライン自動チューニングのための統計モデル

須田 礼仁[†,††]

自動チューニングでは，プログラムに組み込まれた論理に従って実験・データ解析とチューニングが行われる．このためプログラマが暗黙のうちに想定している状況に一致しているかどうか確認する手段がない．従って，自動チューニングのための実験計画・データ解析・最適化では，プログラマの仮定が間違っていても最適解に到達しうる頑健さを，プログラマの仮定が正しかった場合には効率的に最適に到達する効率性とともに備えるべきである．本稿ではオンライン自動チューニングのためのBayes 逐次実験計画に用いられる正規分布モデル・線形モデルに関して，これらの視点で検討を行う．

Statistical Models for Robust and Efficient Online Automatic Tuning

Reiji SUDA[†,††]

In automatic tuning, experiments, data analysis and tuning are done according to the logic embedded in the program. Thus there is no way for the programmer to check whether his/her implicit assumptions are valid or not. Therefore, design of experiments, data analysis and optimization for automatic tuning must provide such robustness that the optimal solution can be obtained even when the programmer's assumptions are not valid, as well as such efficiency that the optimal solution can be efficiently obtained when the programmer's assumptions are valid. In this paper, normal models and linear models that are used in Bayesian sequential experimental design for online automatic tuning are investigated in those viewpoints.

1. はじめに

自動チューニングは，多様な利用環境にソフトウェアが自ら適応することにより，いつでも良好な性能を発揮できることを目指す技術パラダイムである．ソフトウェアやアルゴリズムを問題や環境に適応させることは以前から個別に研究・利用されているが，それらはそのような必要があって個別に行われてきたものであり，大多数の場合には実際の利用環境に合わせてソフトウェアをチューニングしていたと思われる．しかし，自動チューニングという概念に照らしてみると，既存手法に含まれる様々なパラメタが実はいい加減に設定されていることに気づく．従来適当に設定されていたパラメタが，環境に応じてうまくチューニングされてゆけば，情報システム全般が効率化できるという期待もわく．

自動チューニングを実現するためには，様々な技術的課題を解決しなければならない．その課題は実に多岐にわたるが，大きく分けると次の4種類ある．

1. プログラミング: ATLAS[1] や FFTW[2] のような自動チューニングソフトウェアの構築は極めて容易でない．ABCLibScript[3] のような自動チューニングソフトウェアの構築を助けるプログラミング言語上の工夫が必要である．

2. システム: 自動チューニングは環境に適応することを目標としているが，OS などの仮想化による性能の非現実化や，動的スケジューリングやタスク移送などによる暗黙の環境の変化は，最適化のプロセスに重篤な影響を与える危険性があり，協調動作が必要である．

3. 数理: 自動チューニングは一種の最適化である．パラメタ最適化のための手法，実験計画，データ解析などの手法により，いかに効率的にチューニングできるかが影響を受ける．

4. 個別のチューニング技術: 上記のような一般的な技術のうえに，個別の問題に対するチューニングの技術が必要である．数値計算ライブラリ，集団通信ライブラリ，データベース，実アプリケーションなどのそれぞれの場面でのそれぞれのチューニング技術があってはじめて，その自動化が意味を持つ．

これらの4種類の課題のうち，著者は近年特に3番目に挙げた数理を中心に研究を進めている[4)~6)]．本稿では特にオンライン自動チューニングと呼ばれる種類の自動チューニング手法について望まれる性質を考察し，それを満たすための数理モデルについて検討する．

2. 自動チューニングに求める「望ましい性質」

2.1 コストに関する最適性

しかし実際に自動チューニングを実装しようとすると，様々な問題にぶつかる．そもそも，いい加減に決められているパラメタは，どのように変えればよくなるかがよくわからないためにそうなっている．パラメタを振って実験を行うとしても，パラメタの範囲が広かったり，パ

† 東京大学 情報理工学系研究科 コンピュータ科学専攻
Dpt. of Computer Science, Grad. Schl. of Information Science and Technology, the University of Tokyo
†† JST, CREST

ラメタの数が多かったりすると，莫大な実験パターンが必要となる．実験にかけるコストと，最適化の結果とのバランスはどのように取ればよいか，というのは難しい問題である．とりわけ，実行時にならないと情報が得られないオンライン自動チューニングにおいてはこの問題が顕著である．この問題を考慮すると，どんな環境でも高性能であるということを保証するのはなお難しいことになる．未知の環境に対して，最少の実験で最適解を導くためには，きわめて厳しいトレードオフを解決しなければならない．多くの情報を得ることなく，大域的な最適解に到達するというのには，根本的な困難がある．このため，オンラインの自動チューニングでは「最適」という概念を捉え直す必要がある．オンライン自動チューニングにおける「最適」とは，実験コストと実行効率のバランスが最適であることをいう．

この「最適」という言葉について，その背景は正確に理解されるべきである．そもそも「最適」という概念そのものが，何らかの仮定がなければ定義すらできない．そして「最適な自動チューニング」というのは，それぞれの仮定のもとでの最適性に過ぎない．もしその仮定が現実にうまくあっていれば，非常に効率的に最適化することができる．しかし，仮定が現実とかけ離れていると，最適とはかけ離れた結果を生む危険性がある．人手でのチューニングの場合には，仮定が現実に一致しているかどうかは，人間が確認をしていた（より正確には，人間にその確認をする責任がある）．しかし，自動チューニングの場合には，仮定が適切かどうか判断するまでもなく，ソフトウェアは一度組み込まれた仕組みに従って突っ走ってしまう．オンライン自動チューニングでは実験数が最小限に抑えられるから，仮定が適切かどうかを確認するというのが原理的に難しい．そして，仮定は人間が作るものである以上，不適切に設定される危険性は常にゼロでない．従って，仮定が正しくない状況下であっても，効率は低くゆっくりであっても最適解に近づいてくれれば大変ありがたい．

2.2 頑健さ：漸近最適性

このような背景のもと，我々はオンライン自動チューニングに対して「漸近最適性」という概念を定義したい．期待しているのは，想定された仮定と現実との違いが大きくても，致命的な結果にはならないということである．形式的には，**「仮定によらずに漸近的には最適解を見出す」**ということである．「漸近的」とは，使用回数が無限に大きくなることを指している．以下ではパラメタ値の組み合わせが有限の場合を考える（連続パラメタの場合は未考察である）．使用回数が無限であれば，あらゆるパラメタ値の組み合わせを調べ尽くして最適解を見出せばよい．そこで我々はオンライン自動チューニングの手法における「漸近最適性」を，次の 2 つの性質を満たすことと定義する．

漸近最適性条件 1: 任意の仮定 A，任意のパラメタ値の組み合わせ p, 任意の整数 k に対して，ある整数 N があって，使用回数が N 以上であれば，パラメタ値 p での実験回数が k 以上になること

漸近最適性条件 2: 実行時のパラメタ選択において，実験回数が十分に大きくなると，仮定の影響が十分に小さく，実験結果の影響が十分に大きくなること

オンライン自動チューニングの手法が漸近最適性を有していれば，使用回数が十分に大きいときに，すべてのパラメタ値の組み合わせでの実験が十分な回数行われ，仮定の影響は十分に小さくなる．このため，仮定によらず漸近的には最適な解が得られることになる．

自動チューニングの手法が漸近最適であっても，仮定と現実の乖離が大きければ，自動チューニングによる最適化の効率は低下する．このことと，「仮定が現実を適切に表現していれば，自動チューニングによって効率的に最適化できる」ということとは表裏一体であり，後者を期待するなら前者は避けられない．

2.3 効率性：初期実験の有限性

仮定が現実によく合っている場合には，効率的にチューニングがなされてほしい．例えば性能モデルが現実の性能に十分近い場合には，性能が悪いと予想されるパラメタはできるだけ実験してほしくない．これはとりわけ実行回数が少ない時に重要である．実行する回数が少ないときには，事前情報からよさそうだと思われるパラメタを優先的に実行し，性能が悪そうなパラメタは選択しないようにしたい．

また，パラメタ値の組み合わせが無限個ある場合には，最初にすべてのパラメタ値を実験してしまうという手法は実用的な意味を持たなくなってしまう．よって，事前情報と部分的実験から適切な選択肢を選ぶことが必須となる．これを「初期実験の有限性」と呼ぶことにする．

これらの要請は漸近最適性と対照的である．実行数が少なく，仮定が現実に近いときには，仮定を優先し，一部のパラメタ値を除いて実験は行わないということである．

2.4 望ましい性質の保証

漸近最適性と初期実験の有限性は，極端に性質の悪い手法を排除するものであるから，チューニングの手法と事前情報を注意深く選択・設定すれば比較的容易に達成できる．しかし，性能モデルなどのパラメタは定量的な値まで設計時に与えることは難しく，実験・実測から推定しなければならないことが多い．この推定方法によっては，運が悪いと漸近最適性や初期実験の有限性が達成できない推定値が得られてしまう危険性がある．このため，性能モデルのパラメタ推定などにおいて，望ましい性質が満たされない特殊な値がほとんど確実に生じない（生じる確率が 0 となる）こと，また，性能モデルなどが適切に推定されている時には，望ましい性質がほとんど確実に（確率 1 で）満たされること，が必要である．これは注意深く構築された数理モデルを利用することで実現できる．

3. 正規分布モデル

オンライン自動チューニングの問題には Bayes 統計[7]に基づく逐次実験計画が適している．Bayes 統計は極めて自由度の高い枠組みで，確率分布と確率パラメタの事前分布はほとんど自由に仮定することができ，「積分」さえ実行できれば，ほとんどあらゆる問題を明解に解決することができる（ように見える）．

しかし実際にやってみると意外とうまくゆかない．仮定している確率分布と事前分布に相性があるようである．いい加減に事前分布を工夫してみると，時折とんでもない事後分布が得られてしまう．また，高次元の積分は精度よく短時間で行うことは難しいため，できるだけ多くの次元の積分を解析的に実行し，数値積分を最小限に抑える必要がある．解析的に積分できるためには，確率分布と事前分布をうまく選択しなければならない．

これらの性質から，オンライン自動チューニングで実用になる確率分布と事前分布は限られていると思われる．とりわけ正規分布のような連続分布については，前節で述べた望ましい性質を満たし，実用的な時間で積分が実行できる，適切な確率分布と事前分布の組み合わせを調べ上げておき，ユーザはそれを参照する形にするのが望ましい．

以下では一般的な正規分布モデルとその特徴をオンライン自動チューニングの視点で検討する．

以下では第 i 回目の実験の測定値 y_i は，確率的に独立に，平均 μ，分散 σ^2 の正規分布に従うと仮定する．これを

$$y_i \sim N(\mu, \sigma^2)$$

と表現する．実行時間などは正の値しかとらないが，実行時間に比べて分散が極端に大きくなければ正規分布でかなりよく近似できる．

3.1 分散既知の正規分布

分散既知の正規分布では，平均値 μ の事前分布を

$$\mu \sim N(\mu_0, \tau^2) \tag{1}$$

と仮定する．ここで μ_0 は性能モデルなどの事前情報から推定した値，τ^2 は μ と μ_0 がどの程度離れているかを分散で表現したもので，最小二乗法では平均二乗誤差に相当するパラメタである．この μ_0 や τ^2 は別途推定し，固定しておく．また，観測の分散 σ^2 も別途推定して固定値としておく．

n 個の測定値 $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ が得られたとき，平均 μ の事後分布は

$$p(\mu|\boldsymbol{y}) \propto \exp\left(-\frac{(\mu-\mu_n)^2}{2\tau_n^2}\right),$$
$$\tau_n^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1},$$
$$\mu_n = \tau_n^2\left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau^2}\right)$$

となる．ただし $\bar{y}$ は標本平均

$$\bar{y} = \frac{\sum y_i}{n}$$

であり，μ_n は標本平均と事前平均との重み付き平均になっている．

次の測定値 y_{n+1} の値の分布は，取りうる μ の範囲についてその事後確率を重みとする平均値を取った

$$p(y_{n+1}|\boldsymbol{y}) = \int_{-\infty}^{\infty} p(y_{n+1}|\mu)p(\mu|\boldsymbol{y})d\mu$$

で推定されるが，これは解析的に積分できて

$$y_{n+1}|\boldsymbol{y} \sim N(\mu_n, \sigma^2 + \tau_n^2)$$

となる．

この手法は σ^2 と τ^2 が正であれば，漸近最適性と初期実験の有限性の両方が満たされる．

ただし分散 σ^2 が事前に十分な精度で推定されているという仮定はあまりもっともらしくない．分散 σ^2 の推定精度が悪いと，以下に述べる分散未知の正規分布に比べて性能が低下する．前報告に見るように，分散の不確定性が取り込めていないため，悪いものは悪いと決めつけてしまう傾向にあり，実験が不足しがちである．

実はそもそも分散というのは精度よく推定するのが難しい．よく知られているように，$\sum(y_i-\bar{y})^2/\sigma^2$ は自由度 $n-1$ の χ^2 分布に従う．この分布は平均が $n-1$，分散が $2(n-1)$ である．すなわちよく用いられる不偏推定 $\sum(y_i-\bar{y})^2/(n-1)$ は標準偏差が $\sigma^2\sqrt{2/(n-1)}$ もある．標準偏差を $0.1\sigma^2$，つまり相対誤差を 0.1 程度におさえるためには，$n = 201$ が必要である．オンライン自動チューニングのように，実験回数を最小限にしようと思えば，分散を高精度で推定するというのは，極めて非現実的である．

3.2 分散未知の正規分布

分散の不確定性を Bayes 統計で表現するには，分散 σ^2 に関する事前分布を仮定すればよい．標準的な手法では

$$p(\sigma^2) = \frac{(\nu_0\sigma_0^2/2)^{\nu_0/2}}{\Gamma(\nu_0/2)}\frac{1}{(\sigma^2)^{1+\nu_0/2}}\exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) \tag{2}$$

を仮定する．これは逆カイ二乗分布と呼ばれている．ただしこのままでは解析的な積分ができないため，平均値の事前分布の τ^2 に対して

$$\tau^2 = \sigma^2/\kappa_0$$

という仮定をおく．ここで κ_0 は既知の定数であるとする．

すると平均と分散の事後分布が

$$p(\mu, \sigma^2|\boldsymbol{y}) \propto \frac{1}{(\sigma^2)^{1+(\nu_0+n+1)/2}}$$
$$\exp\left(-\frac{\kappa_0(\mu-\mu_0)^2 + n(\mu-\bar{y}) + n\hat{\sigma}_n^2 + \nu_0\sigma_0^2}{2\sigma^2}\right)$$

ただし $\hat{\sigma}_n^2$ は標本分散

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum(y_i-\bar{y})^2$$

である（なお，本稿では標本分散の分母に $n-1$（不偏推定）ではなく n（最尤推定）を用いる．このほうがわずかに式が短くなる）．これは σ^2 に関して解析的に積分

できて

$$p(\mu|\boldsymbol{y}) \propto \left(1+\frac{\kappa_n(\mu-\mu_n)^2}{\nu_n\sigma_n^2}\right)^{-(\nu_n+1)/2},$$
$$\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y},$$
$$\kappa_n = \kappa_0+n,$$
$$\nu_n = \nu_0+n,$$
$$\sigma_n^2 = \nu_n^{-1}\left(\frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2 + n\hat{\sigma}_n^2 + \nu_0\sigma_0^2\right)$$

となる．これは自由度 ν_n の t 分布である．次の測定 y_{n+1} の分布を計算すると

$$p(y_{n+1}|\boldsymbol{y}) \propto \left(1+\frac{\kappa_n(y_{n+1}-\mu_n)^2}{(\kappa_n+1)\nu_n\sigma_n^2}\right)^{-(\nu_n+1)/2}$$

となりやはり自由度 ν_n の t 分布となる．

しかしこの方法にはいくつかの問題がある．まず，t 分布は自由度が 1 以下では平均値を持たない（積分が発散する）．このため，$\nu_0 \leq 1$ の場合，実験を最低でも 1 回は行わないと，性能を推定することができなくなってしまう．すなわち，初期実験の有限性が保証できない．ν_0 は分散の不確定性を表すパラメタで，分かりやすく言うと，分散に関する事前情報が実験何回分に相当するかを表している．計算の所要時間などについて実験してみると，分散のばらつきは非常に大きく，数桁程度の不確定性がある（このような分散の大きなばらつきは，「分散既知の正規分布」の適切性に疑問を投げかけるものでもある）．ある実験では $\nu_0 = 0.4$ 程度の値が得られており，これでは初期実験の有限性が達成できない．無理やり $\nu_0 > 1$ の値に設定してしまうということも考えられるが，1 ではだめということもあり，一体いくつにすればよいかはっきりしない．1 に近ければやはり初期実験の量が無限に多くなるので，かなり大きめに設定せざるを得ない．しかし分散を既知に固定するよりはましと思われる．

もうひとつの問題は定数である κ_0 の値である．定義より $\kappa_0 = \sigma^2/\tau^2$ となるが，上述のように σ^2 は不確定性が大きく，モデルの精度を表す τ^2 も必ずしも精度よく推定できるわけでもない．これらの比を表す κ_0 には大きな不確定性が含まれているのに，これを定数として扱ってしまっている．この κ_0 は実験計画にかなりの影響を及ぼす定数なので，それが精度よく推定できないことは問題である．κ_0 に事前分布を設定することも考えられるが，数値積分が避けられない．

3.3 準共役正規分布

分散未知の正規分布では，測定値の分散 σ^2 とモデルの精度 τ^2 を κ_0 というパラメタで結びつけている．しかし，正規分布においては標本平均と標本分散は確率的に独立であるから，標本分散から推定される σ^2 と，標本平均から推定するべき τ^2 とは，別々に推定することができてもおかしくない．

そこでこの結びつきを取り払ったものが，準共役な事前分布である．すなわち平均 μ について式 (1) を，分散 σ^2 について式 (2) を仮定する．これで平均 σ^2 について積分すると

$$p(\mu|\boldsymbol{y}) \propto \exp\left(-\frac{(\mu-\mu_0)^2}{2\tau^2}\right)\left(1+\frac{n(\mu-\bar{y})^2}{n\hat{\sigma}_n^2+\nu_0\sigma_0^2}\right)^{-\nu_n/2} \quad (3)$$

となる．これは μ に関して解析的には積分できず，例えば μ の期待値や最尤推定などは数値計算で求めなければならない．次の測定 y_{n+1} の予測分布も同様の式になってしまうが，y_{n+1} については解析的に積分することができるため，数値積分の必要性は μ に関する 1 次元積分だけとなる．

この方法の利点は，$0 < \nu_0 \leq 1$ で $n = 0$ でも積分が収束することである．このため，実測を行わなくてもある程度の性能の推定ができ，初期実験の有限性が実現できる．

ところが，こんどは漸近最適性に問題が発生する．式 (3) を見ると，平均 μ と事前情報 μ_0 との乖離は指数関数的におさえられているのに対して，平均 μ と標本平均 $\bar{y}$ との乖離は多項式的にしかおさえられていない．このため，標本平均 $\bar{y}$ が事前情報 μ_0 と極端に離れているときには，事前情報が限りなく優先され，標本平均が無視されてしまう．その結果，次の実験結果がいくら小さくても事後平均値は有限の範囲に抑え込まれてしまう．このため，事前情報における平均値がある程度以上大きい選択肢は，永久に測定候補に含まれなくなってしまうのである．

この問題は μ に対する事前分布を無理やり自由度 ν_n の t 分布にすれば解決するが，測定回数に依存する事前情報というのは不自然だし，$\nu_0 \leq 1$ で積分が収束するという性質も同時に失われてしまう．

3.4 モデル修正項つき準共役正規分布

実際には，事前情報 μ_0 は実測値と性能モデルを用いて推定されている．このため，標本平均 $\bar{y}$ が変わると事前情報 μ_0 も違う値になる．そこで，次の測定 y_{n+1} によって標本平均が変化し，それに伴い事前情報 μ_0 もずれる効果を入れてみる．具体的には，式 (3) の μ_0 のところに $\mu_0 + \delta(y_{n+1} - \bar{y})$ を代入する．ここで δ は性能モデルとそれまでの測定回数によって決まる定数とする．これを計算すると

$$p(y_{n+1},\mu,\sigma^2|\boldsymbol{y}) \propto \frac{1}{(\sigma^2)^{1+(\nu_0+n+1)/2}} \exp\left(-\frac{(\mu-y_{n+1})^2+n(\mu-\bar{y})^2+n\hat{\sigma}_n^2+\nu_0\sigma_0^2}{2\sigma^2}\right) \exp\left(-\frac{(\mu-\mu_0-\delta(y_{n+1}-\bar{y}))^2}{2\tau^2}\right) \quad (4)$$

となる（δ を含む項に $\bar{y}$ が出てきており $n = 0$ の時に問題が生じるが，これは μ_0 で置き換えることで解決できる）．これは変数変換が必要なものの σ^2 と y_{n+1} について積分することができ，やはり 1 次元積分に帰着させることができる．

これにより，どんなに推定値 μ_0 が悪くても，極端にすばらしくよい結果 y_{n+1} を出せば，最適解よりもよくなる可能性がわずかながらに残る．このため，漸近最適性は確保される．

ところが今度は $\nu_0 \leq 1$ ではまったく実験していない $n=0$ で積分が収束しなくなってしまい，初期実験の有限性が失われる．これは最初の実験値 y_1 の不確定性が非常に高いため，それに影響される μ_0 もあわせて不確定になってしまうからである．この問題は y_{n+1} が μ_0 におよぼす影響を非線形にすれば解決できる可能性があるが，そうすると解析的に積分ができず，2 次元の数値積分が必要となってしまう．無限領域の 2 次元積分を高精度に計算したいとなると，計算量がかなり問題となりそうである．

3.5 本節のまとめ

このように，「初期実験の有限性」と「漸近最適性」という視点で Bayes 統計モデルを見てみると，様々な問題点が明らかとなる．本節で論じた手法の中で，この両方を満足しているのは，最も簡単な分散既知の正規分布だけである．ただし分散既知の正規分布には，実験不足になりやすいという欠点がある．

やや無理やりの感があるが，実験をしていない $n=0$ では 3.3 節の式 (3) を用い，実験が 1 回でもあれば 3.4 節の式 (4) を用いることにより，初期実験の有限性と漸近最適性の両方を満足させることができる．

別の方法として，分散未知の正規分布において，積分の計算のときのみ ν_0 を十分大きく再設定するという方法が考えられる．分散を固定している分散既知の正規分布よりはましという程度で，数理的根拠はない．

さらに別の方法として，分散既知の正規分布の実験不足の解消のため，分散を大きめに推定しておくという方法も考えられる．逆カイ二乗分布では，平均値が $\nu\sigma_0^2/(\nu-2)$ になるのが参考になるが，$\nu > 2$ でないとそもそも平均値が収束しない．

4. 線形モデル

平均値の推定 μ_0 は，モデル（フィッティング）を用いて推定するのがよい．推定の安定性ということでは線形モデルが優れている．問題によっては非線形のモデルを用いたくなることも多いが，非線形のフィッティングは予想外のデータが与えられた時にとんでもない推定を出してしまうことがあるため，どうしても避けられない場合に限り，十分な注意を払い，技術的工夫を施して使わなければならない．

以下，i 番目の選択肢の j 回目の測定値 y_{ij} は正規分布 $N(\mu_i, \sigma_i^2)$ に従うとする☆．実際に用いるのは標本平均 $\bar{y}_i$ のみである．

標準的な線形モデルでは，平均 μ_i とモデル g_i との乖離は $N(0, \sigma_i^2/\kappa_i)$ のように設定する．これは分散未知の正規分布にマッチしているが，前述のとおり分散未知の正規分布では初期実験の有限性が得られない．このため本稿では，分散とは（少なくとも形式的には）独立して $N(0, \tau^2/w_i)$ のように仮定する．ここで w_i は重みであり，平均とモデルの乖離の期待値や，標本平均 $\bar{y}_i$ の分散の推定値などを考慮して決めるべきものである．重み w_i は十分慎重に選ぶ必要がある．そうでないと，やはりとんでもない推定を出してしまうことがある．

線形モデルは

$$g_i = \boldsymbol{x}_i^T \boldsymbol{b}$$

のように書くことができる．ここで $\boldsymbol{x}_i$ は既知のベクトルでモデルを表し，$\boldsymbol{b}$ は未知のベクトルでモデルの係数を表している．すべてのデータに関してまとめると

$$\boldsymbol{g} = X\boldsymbol{b}$$

と書くことができる．行列 X は計画行列と呼ばれる．先の仮定を

$$\boldsymbol{y} - \boldsymbol{\mu} \sim N(\boldsymbol{0}, \Sigma)$$

（Σ は σ_i^2/n_i を要素とする対角行列）および

$$\boldsymbol{\mu} - \boldsymbol{g} \sim N(\boldsymbol{0}, \tau^2 W)$$

（W は $1/w_i$ を要素とする対角行列）と書くと，

$$V = \tau^2 W + \Sigma$$

として $\boldsymbol{g}$ の最尤推定は

$$\boldsymbol{\gamma} = X(X^T V^{-1} X)^{-1} X V^{-1} \boldsymbol{y}$$

のように得られる．これは重み付き線形最小二乗の標準的な結果である．

しかし事前には τ^2 も Σ も未知であるので，V はこれらの推定値を考慮して適当に決められることになる．さいわい，極端におかしな V でなければ，使い物になる結果が得られる．以下記法を簡単にするため

$$L = X(X^T V^{-1} X)^{-1} X V^{-1}$$

とおく．$L^2 = L$ に注意する．

4.1 線形モデルの精度 τ^2 の推定

次に係数 τ^2 を推定したい．このために適当な重み u_i を定数で与えて

$$e = \sum u_i (\bar{y}_i - \gamma_i)^2$$

を計算してみる．u_i を並べた対角行列を U としてこの e の期待値を計算すると

$$\begin{aligned}
E(e) &= E\left((\boldsymbol{y}-\boldsymbol{\gamma})^T U (\boldsymbol{y}-\boldsymbol{\gamma})\right) \\
&= E\left(\boldsymbol{y}^T (I-L)^T U (I-L)\boldsymbol{y}\right) \\
&= E\left(\boldsymbol{\mu}^T (I-L)^T U (I-L)\boldsymbol{\mu}\right) \\
&\quad + E\left((\boldsymbol{y}-\boldsymbol{\mu})^T (I-L)^T U (I-L)(\boldsymbol{y}-\boldsymbol{\mu})\right) \\
&= E\left((\boldsymbol{\mu}-\boldsymbol{g})^T (I-L)^T U (I-L)(\boldsymbol{\mu}-\boldsymbol{g})\right) \\
&\quad + E\left((\boldsymbol{y}-\boldsymbol{\mu})^T (I-L)^T U (I-L)(\boldsymbol{y}-\boldsymbol{\mu})\right) \\
&= E\left(tr\left((I-L)^T U (I-L)(\boldsymbol{\mu}-\boldsymbol{g})(\boldsymbol{\mu}-\boldsymbol{g})^T\right)\right) \\
&\quad + E\left(tr\left((I-L)^T U (I-L)(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})^T\right)\right) \\
&= tr((I-L)^T U (I-L) W \tau^2) \\
&\quad + tr((I-L)^T U (I-L)\Sigma)
\end{aligned}$$

となる．ここで，$\boldsymbol{y}-\boldsymbol{\mu}$ と $\boldsymbol{\mu}-\boldsymbol{g}$ が独立と仮定している．

☆ 前節とは記号に若干の食い違いがある

$$\phi = tr((I-L)^T U(I-L)W)$$
$$\psi = tr((I-L)^T U(I-L)\Sigma)$$

とおけば

$$E(e) = \phi\tau^2 + \psi$$

であるから，

$$\hat{\tau}^2 = (e-\psi)/\phi$$

で推定できそうである．

実際にやってみると，測定数 n_i が十分に大きく，標本平均 $\bar{y}_i$ も分散も σ_i^2 も十分な精度で推定できている場合には，確かに本来の τ^2 に近い値を出す．ところがオンライン自動チューニングのように，最適と思われていないものは1回ぐらいしか測定されておらず，平均も分散もよい精度で推定できていない場合には，かなりの割合で $\hat{\tau}^2$ が負の値になってしまう．さすがに負の値を出すのは意味がないので，

$$\hat{\tau}^2 = \max\{0, (e-\psi)/\phi\}$$

で推定する．Bayes 統計を用いて最尤推定などをしても似たようなことになり，測定値の不確定性が高い場合には，τ^2 の推定値が 0 となってしまうことがある．これは本質的な問題であるようで，いろいろ試みてみたが，点推定をする限りは避けられそうにない．さいわい，次節で述べるような事情により，$\hat{\tau}^2 = 0$ と推定しても漸近最適性などに影響する危機的な状況にはならない．

逆に τ^2 に比べて $\hat{\tau}^2$ が極端に大くなることもある．これは重み行列が影響しているらしい．例えば分散が大きいサンプルに対して無理に w_i を大きくしてモデルにフィットさせようとすると，推定される誤差 τ^2 は大きくなってしまう．このあたり，σ^2 と τ^2 を完全に分離して考えることには無理があることがうかがえる．

上記の手法では，選択の自由度がある重み行列 U がある．また，W と V の選択にもある程度自由度があるものと考えられる．τ^2 が安定してよい精度で推定できるような U, V, W の決め方について今後検討する必要がある．

4.2 線形モデルの不確定性

τ^2 は真のモデル g_i と真の平均値 μ_i との乖離を示すパラメタであるが，実際には g_i は分からないので，γ_i で代用することになる．従って，本当に必要なのは μ_i と γ_i との乖離度である．

$$\begin{aligned}\boldsymbol{\mu}-\boldsymbol{\gamma} &= (I-L)\boldsymbol{\mu} + L(\boldsymbol{\mu}-\boldsymbol{y}) \\ &= (I-L)(\boldsymbol{\mu}-\boldsymbol{g}) + L(\boldsymbol{\mu}-\boldsymbol{y})\end{aligned}$$

であるから，$\mu_i - \gamma_i$ の分散・共分散行列は

$$\begin{aligned}&E\left((\boldsymbol{\mu}-\boldsymbol{\gamma})(\boldsymbol{\mu}-\boldsymbol{\gamma})^T\right) \\ &= (I-L)W(I-L)^T\tau^2 + L\Sigma L^T \qquad (5)\end{aligned}$$

となる．この第1項がモデルと真の平均との誤差，第2項が標本平均と真の平均との誤差に起因する項である．$\hat{\tau}^2$ の推定で負の値を許さなければ，$L\Sigma L^T$ のおかげで分散は正の値になる．

各 i に関して個別に見れば，式 (5) の右辺の第 i 対角要素を採用すればよい．

しかし，式 (5) は $\mu_i - \gamma_i$ と $\mu_j - \gamma_j$ は独立ではないことを示している．これは性能の比較を行う自動チューニングにおいては無視できない影響を及ぼす可能性がある．共分散が正であれば，μ_i と μ_j は同じ方向にずれる傾向にあるので，性能差の不確定性は（個別に不確定性を扱うよりも）小さめであり，逆に共分散が負であれば，μ_i と μ_j は逆方向にずれる傾向にあるので，性能差の不確定性は大きめである．オンライン自動チューニングでは，不確定性の大小が実験計画に影響するため，共分散の存在は結果に影響を及ぼす．しかし，著者が提案している実験計画[4),5)] において共分散を考慮することは極めて自明でない修正が必要となりそうである．

5. おわりに

本稿では，著者が研究を続けているオンライン自動チューニングのための Bayes 統計のための正規分布モデル，線形モデルについて考察した．これらの考察は HOKKE で発表した実験結果[6)] を精査するうちに見つかった課題をきっかけに進めたものだが，まだ十分な知見には到達していない．早急に検討を進め，基本的な数理モデルの確立を急ぎたい．

謝辞

本研究の一部は JST CREST プロジェクト，文部科学省科学研究費による支援を受けています．

参考文献

1) Whaley, R. C. and Dongarra, J. J.: Automatically Tuned Linear Algebra Software, *Proceedings of SC98*, (CD-ROM), 1998.
2) Frigo, M. and Johnson, S.G.: FFTW: an adaptive software architecture for the FFT, *Proceedings of ICASSP '98*, Vol. 3, pp. 1381–1384, 1998.
3) Katagiri, T., Kise, K., Honda, H., and Yuba, T.: ABCLibScript: A directive to support specification of an auto-tuning facility for numerical software, *Parallel Computing*, Vol. 32, No. 1, pp. 92–112, 2006.
4) Suda,R.: A Bayesian Method for Online Code Selection: Toward Efficient and Robust Methods of Automatic Tuning, *Proc. 2nd Int'l Workshop on Automatic Performance Tuning (iWAPT2007)*, pp. 23–32 (2007).
5) 須田礼仁: オンライン自動チューニングのための Bayes 統計に基づ逐次実験計画法, 情報処理学会 HPCS2008, pp. 73–80 (2008).
6) 須田礼仁: オンライン自動チューニングのための Bayes 逐次実験計画の解析モデルによる性能評価, 情報処理学会研究報告, Vol. 2008, No. 19, pp. 211–216 (2008).
7) Carlin, B. P. and Louis, T. A.: *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., Chapman and Hall, 2000.