階層ベイズモデルによる化学反応曲線の評価

階層ベイズモデルによる化学反応曲線の評価

高橋　啓（情報工学部システムマネジメント学科）

Evaluation of LCST Curves with Hierarchical Bayesian Model

TAKAHASHI Kei　（Department of System Management, Faculty of Information Engineering）

Abstract

In this paper, we propose a method to evaluate substances by modeling the reaction process of a polymer compound with fluctuations using a hierarchical Bayesian model and taking the variance ratio of the model without hierarchy. As a result, it is shown that about half of the variance ratios are systematic errors for the substance in question.

Keywords：Hierarchical Bayesian model, MCMC, LCST curve

1.　はじめに

温度応答性高分子という温度によって親水性や疎水性に状態が変化するような高分子が存在する．これらの高分子は，水中で下限臨界溶液温度（Lower Critical Solution Temperature）を持ち，ある温度を堺に温度が高ければ凝集して白濁する，逆に低ければ水中で溶解し無色透明になるといった可逆的な挙動を示す．この温度によって特殊な状態変化をする性質を利用し，温度応答性高分子はソフトアクチュエータやドラッグデリバリーシステムの材料として注目されており，研究や開発が行われている．近年では，材料工学における新素材の研究に機械学習が活用されている．機械学習を用いることで実験にかかるコストや時間を削減しつつ，有用な材料の特定や特性の予測を行うことが期待できる．

本研究では，LCST を持つ高分子に対して，階層ベイズモデルによる評価を行うことを目的とする．水溶液の温度変化による透過率の変化を計測したデータから，透過率を予測する階層ベイズモデルを作成することで，LCST 曲線に対する評価を行う．階層ベイズモデルはプログラミング言語の R と Stan を用いて，マルコフ連鎖モンテカルロ法の 1 つであるハミルトニアンモンテカルロ法で推定を行う．予測モデルに対しては，広く使える情報量規準（WAIC）を推定結果から求めることで，モデルの予測能力を測る．

2.　LCST 曲線の概要

本研究で使用するデータは，群馬大学の覚知先生に提供していただいた．図のような構造をした LCST を持つ高分子の水溶液に対して，25℃ から 60℃ の間の温度変化による透過率の変化を，0.1℃ごとに 4 回計測したものである．合計 1392 サンプルが得られる．

具体的な高分子の構造は，図 1 のとおりである．この高分子は温度により光の透過率が変化し，温度が上昇すると透過率は急激に減少し，再び下降すると急激に上昇する．この温度と透過率との関係が LCST 曲線である．この変化する温度：下限臨界溶液温度（LCST）である．このような高分子はソフトアクチュエータとして期待されている．

3.　統計モデルとその推定方法

本研究では，トライアルごとに異なる曲線のカーブを推定するために階層ベイズ・モデルを用いる．ここでベイズモデルとは，事前確率分布と尤度関数から事後確率分布をベイズ推定するモデルを指す．階層ベイズモデルとは，事前確率分布についてさらに事前確率分布を考える階層構造を持つモデルである．データを $\boldsymbol{y}$，パラメータを $\boldsymbol{\vartheta}$，ハイパー・パラメータを $\boldsymbol{\alpha}$ とすると，事後分布は次のように表される

$$P(\boldsymbol{\vartheta},\boldsymbol{\alpha}|\boldsymbol{y}) \propto \int P(\boldsymbol{y}|\boldsymbol{\vartheta})P(\boldsymbol{\vartheta}|\boldsymbol{\alpha})P(\boldsymbol{\alpha})$$

実際にフィッティングするモデルは，線形混合モデルの一種であるシグモイド・カーブに変量効果を加えたモデルである．変量効果の有無で 4 種類のモデルを作成する．具体

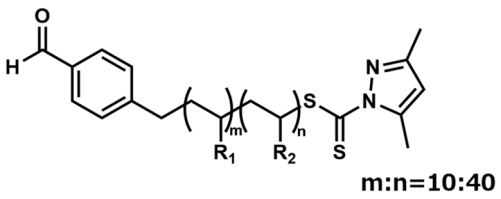

図 1　対象高分子の構造

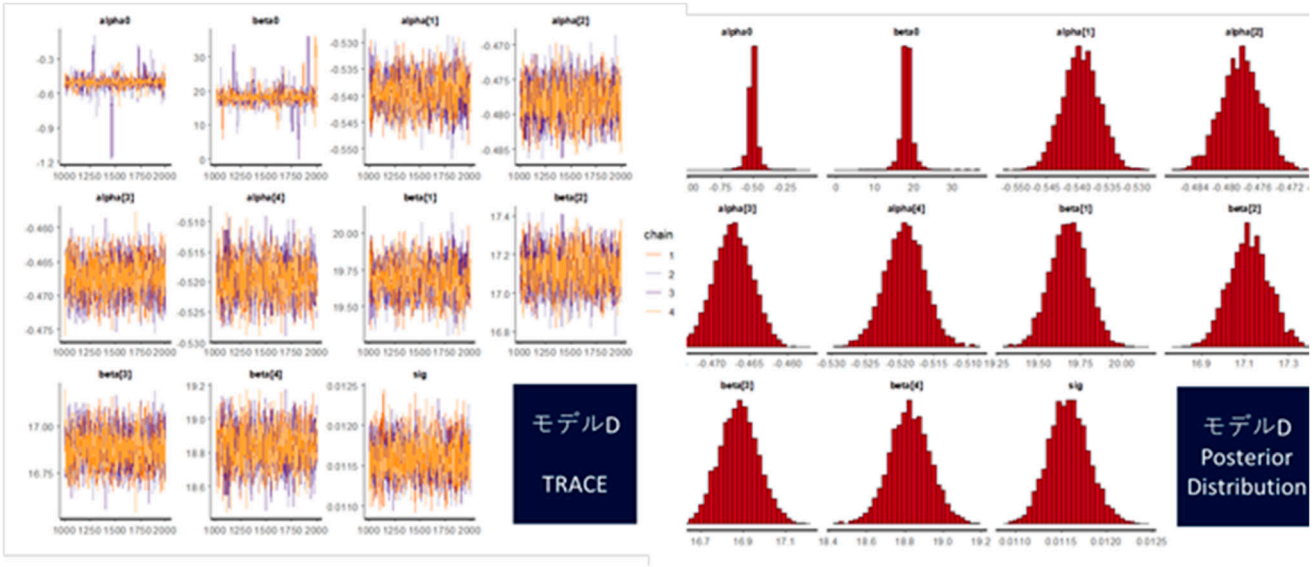

図 3 モデル D の軌跡（左図）および事後分布（右図）

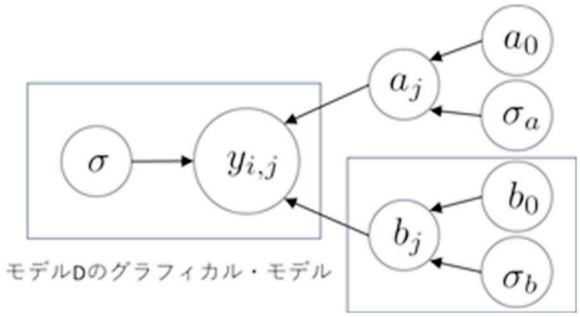

図 2　モデル D のグラフィカルモデル

的には，モデル A：変量効果なし，モデル B：回帰係数に変量効果あり，モデル C：切片に変量効果あり，モデル D：両方の 4 つである．次にモデル D の式を示す．

$$\mathrm{Y_{i,j}} \sim \mathrm{N}\left(\frac{1}{1+\exp\left(-\left(\mathrm{a_j x_i + b_j + \epsilon_{i,j}}\right)\right)}, \sigma^2\right)$$

$$\mathrm{a_j} \sim \mathrm{N}(\mathrm{a_0}, \sigma_a^2), \quad \mathrm{b_j} \sim \mathrm{N}(\mathrm{b_0}, \sigma_b^2)$$

ここで，$\mathrm{Y_{i,j}}$ はトライアル j における i 番目のデータ（透過率），$\mathrm{x_i}$ は i 番目の温度である．このモデルをグラフィカルモデルで表したものが図 2 である．

このモデルのパラメータを推定する場合，階層構造があるため，一般的な最尤推定を用いることができない．その代わりに確率的グラフィカル構造を利用し，Markov Chain Monte-Carlo の一手法である Hamiltonian Monte-Carlo（HMC）法を用いる．HMC は補助変数の導入により，高次元でも容易に推定可能となる手法である．

実際の推定の設定は次に示すとおりである．全部で 2000 回サンプリングし，系列相関を避けるために 4 チェーンで推定を行う．初期値の依存を考え，はじめの 1,000 回は除去し，事後分布には含めない．分布収束については Gelman-Rubin 統計量から収束を判断する．

4.　結果と評価

図 3 にモデル D の各パラメータのマルコフチェーンの軌跡と事後分布を示す．各パラメータとも Gelman-Rubin 統計量の値が 1.001 以下と分布収束していることは確かめられている．

予測性能の良さについては，WAIC で評価を行う．階層ベイズモデルは特異点のあるモデルであり，AIC や BIC といった選択基準を用いることができない．代わりに特異点解消可能かつ事後尤度分布から評価可能である WAIC を用いる．WAIC は，具体的には次のように書き下せる：

$$\mathrm{WAIC} = -2\,\mathrm{lppd} - p_{\mathrm{WAIC}}$$

$$\mathrm{lppd} = \sum_{i=1}^{348}\sum_{j=1}^{4} \ln\left(\Pr\left(y_{i,j}\right)\right), p_{\mathrm{WAIC}} = \sum_{i=1}^{348}\sum_{j=1}^{4} V\left(\ln\left(\Pr\left(y_{i,j}\right)\right)\right).$$

WAIC による評価の結果は，モデル D：-8452，モデル C:-8064，モデル B:-7995，モデル A:-7186 となりモデル D が予測の面からはよいという結果となった．

物質としての評価は，モデル A と D の分散比で行う．この比較により全誤差のうち系統誤差の占める割合がわかる．この物質の場合，0.47 と約半分が系統誤差であることが示された．

謝辞

本研究は本学情報科学研究所の 2021 年度研究スタートアップ支援制度により実施したものである．

文　　献

(1) Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B.: “*Bayesian Data Analysis Third Edition*”, CRC Press (2013)

(2) Mueller, T., Kusue, A. G., and Ramprasad, R.: Machine Learning in Materials Science. *Reviews in Computational Chemistry*, Vol. 29, pp. 186-273 (2016).

(3) Vehtari, A., Gelman, A., and Gabry, J.: Practical Bayesian model evaluation using leave one-out-cross-validation and WAIC. *Statistics and Computing*, Vol. 27, pp. 1413-1432 (2017)