

地盤データを対象とした自己相関関数のモデル選択の基礎検討

東京都市大学 ○学生会員 鈴木 悠也 東京都市大学 学生会員 津田 悠人
東京都市大学 正会員 吉田 郁政

1. はじめに

機械学習において、目的に応じて適切なモデルを正しく選択することが重要である。また、V&V(Verification and Validation)においてもモデルと観測データの適合性を表す指標は有用である。V&Vとは、計算結果の信頼性を具体的に立証するためのプロセスであり、分野によって様々な基準がある。適切なモデル選定により過学習を防ぐことができる。過学習を防ぐ能力である汎化性能を評価する手法として普及している CV(Cross Validation)は、一般に大量のデータを必要とするため、本研究では Evidence を用いた評価の性能について検討を行う。

一般的な回帰理論で近年注目されているガウス過程回帰 (Gaussian process regression, GPR) はクリギングと基本的に同じであり、地盤物性値の空間物性推定に対して有効な方法である。Yoshida et al.¹⁾ は確率場を重ね合わせた GPR による地盤物性値の3次元空間分布推定手法を提案している。本報告では GPR の自己相関関数(Autocorrelation function, ACF)についてのモデル選択について述べる。

2. 確率場の作成に用いた自己相関関数

確率場を作成する際に共分散行列が必要である。共分散行列 $\mathbf{M}$ の各要素は、式(1)より得られる。

$$\text{cov}(s, s') = \sigma_k^2 \times \rho_k(d|\delta_k) \quad (1)$$

ここで、 s : 位置座標、 σ_k : 場の標準偏差、 $\rho_k(d|\delta_k)$: 自己相関関数 (ACF)、 δ : Scale of fluctuation (SOF)²⁾、 d : 2点間距離である。ACFとしては、Gaussian型と Whittle-Matérn型³⁾を想定する。式(2)にGaussian、式(3)にWhittle-Matérnを示す。

$$\rho(d|\delta) = \exp\left(-\pi\left(\frac{d}{\delta}\right)^2\right) \quad (2)$$

$$\rho(d|\delta) = \frac{2}{\Gamma(v)} \left(\frac{\sqrt{\pi}\Gamma(v+0.5)d}{\Gamma(v)\delta} \right)^v K_v \left(\frac{2\sqrt{\pi}\Gamma(v+0.5)d}{\Gamma(v)\delta} \right) \quad (3)$$

ここで、式(3)について Γ はガンマ関数、 K_v は変形された次数 v の第2種のベッセル関数、 v は smoothness parameter(SP)である。式(4)で示される負の対数尤度を最小化することでパラメータを求める。最小化には、BFGS法を用いる。

$$-\ln L = \frac{1}{2} \mathbf{z}^T \left(\sum_{k=1}^n \mathbf{M}_{k,11} \right)^{-1} \mathbf{z} + \frac{1}{2} \ln \left| \sum_{k=1}^n \mathbf{M}_{k,11} \right| + \frac{m}{2} \ln(2\pi) \quad (4)$$

ここで、 m はデータ数である。

3. モデル選択のための指標

与えられた観測データに基づいて、各モデルのデータとの一致度とモデルの複雑さのトレードオフの観点から、最も適したモデルを選択する必要がある。Evidence (marginal likelihood) $p(\mathbf{z})$ はモデル選択のための有効的な指標であるが、Evidence を計算するための積分は一般に解析が存在しない。

$$p(\mathbf{z}) = \int p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \quad (5)$$

簡易にモデル選択できる方法として AIC や BIC などの情報量規準が提案されている。式(6)に AIC と式(7)に BIC を示す。

$$\text{AIC} = -2\ln p(\mathbf{z}|\hat{\boldsymbol{\theta}}) + 2n_h \quad (6)$$

$$\text{BIC} = -2\ln p(\mathbf{z}|\hat{\boldsymbol{\theta}}) + n_h \ln(m) \quad (7)$$

モデル選択では、AIC と BIC の値は小さいほうがよく、Evidence は大きいほうがよい。BIC は AIC よりもモデルの複雑さに大きなペナルティを与える。

4. 模擬データを用いたモデル選択

モデル選択の例として、2つの確率場を足し合わせた模擬データを用いた検討を行った。真のモデルとして第1成分(トレンド): ACF=Gaussian model, SD=10, SOF=5 とし、第2成分(ランダム): ACF=Whittle-Matérn model, SD=5, SOF=0.5, SP=0.7 とした。観測点数に関しては 51, 101, 201 点

キーワード モデル選択, 自己相関関数, Evidence, AIC, BIC

連絡先 〒158-8557 東京都世田谷区玉堤 1-28-1 東京都市大学 都市工学科 TEL 03-5707-0104

の3ケースを考えた。51点のデータ間隔は、201点のデータの4倍である。作成した模擬データの一例を図-1に示す。作成した模擬データに関して、以下のモデルA,B,C,Dの4つのACFモデルからモデル選択を行う。Cが真のモデルである。

Model A: WhittleMatérn,

Model B: WhittleMatérn+White Noise(WN),

Model C(TURE): Whittle-Matérn+Whittle-Matérn,

Model D: Whittle-Matérn+Whittle-Matérn+WN,

上記の4つのACFモデルについて、乱数の種を替えて100個の模擬データを作成して、Evidence, AIC または BIC による ACF モデルの正解率を各観測点数について求めた。Evidence はモンテカルロ法で算出し、その際通常の乱数と超一様分布列である Sobol 列での比較を行った。それぞれ 10000 個発生させ、Evidence を算出した結果を図-2に示す。縦軸は算出された Evidence を規準化したものである。通常の乱数よりも Sobol 列の方が早く収束していることがわかる。Evidence を算出では Sobol 列 4000 個を用いた。図-3 は、観測点数が 101 点のときの Evidence, AIC, BIC によって選択された ACF モデルの割合を比較したものである。Evidence, AIC, BIC の順に観測点数が多いほど正解率は高く、Evidence では観測点数 101 点で 8 割程度の正解率である。

5. おわりに

本報告では、模擬データを用いて、ACF のモデル選択に関する基礎的な検討を行った。紙面の都合で図は示していないが、当然のことながら観測点数が多いほど正解率は高くなった。観測点数が少ない場合(51点)は、AIC, BIC ではより単純なモデルが選択され Evidence ではどのモデルも大きな差はないとの結果となった。今後の検討では、CV を用いたモデル選択との比較を行う予定である。

参考文献

- 1) Yoshida et al.: Estimation of trend and random components of conditional random field using Gaussian process regression, Computers and Geotechnics, Vol.136, 2021.
- 2) Vanmarcke, E. H.: *Random fields: Analysis and synthesis*. Revised and expanded new edition, World Scientific Publishing, 2010.

- 3) Ching et al.: Identification of sample path smoothness in soil spatial variability, Structural Safety, Vol.81, 2019.

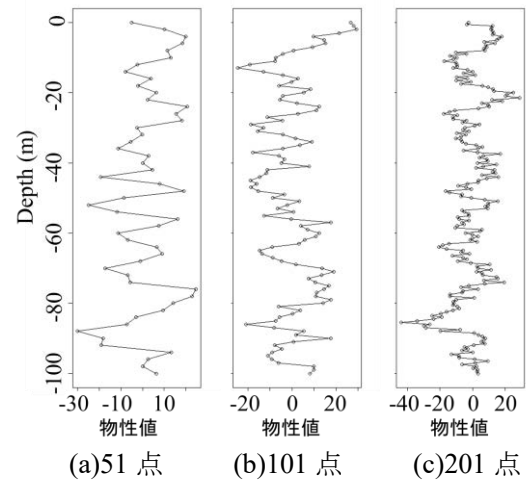


図-1 観測点数による模擬データの一例

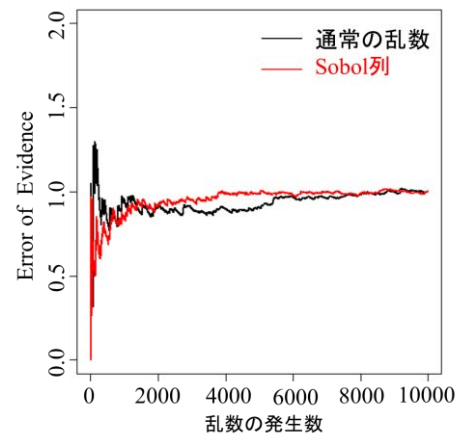


図-2 Evidence を算出する際に用いる乱数の数による比較

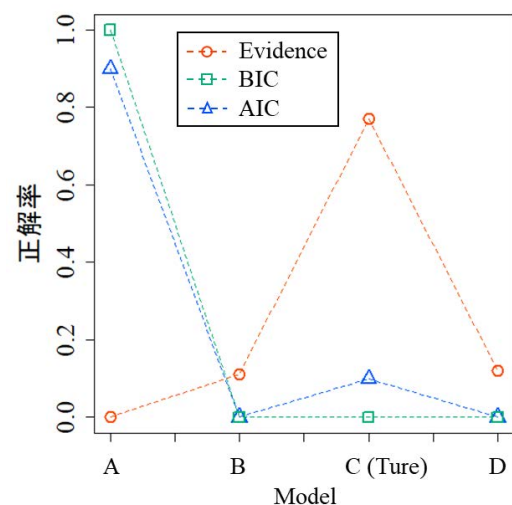


図-3 観測点数 101 点、各モデルの Evidence, AIC, BIC の正解率の比較