統計的機械学習の応用研究

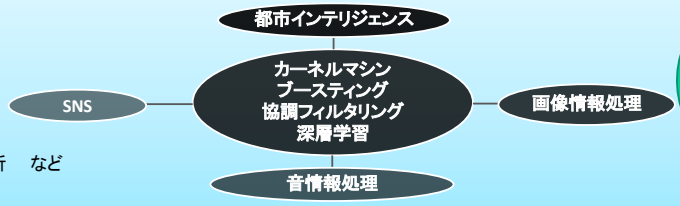
松井　知子　　モデリング研究系　教授

【概要】

本研究室では統計的学習機械を用いて、音声/音楽/画像/SNSなどを処理する方法について研究しています。具体的にはカーネルマシン、ブースティング、協調フィルタリング、深層学習の手法を用いて、

1. 音声・話者認識
2. 音楽情報処理
3. 画像識別
4. SNS解析
5. 都市インテリジェンス
6. 新型コロナウイルス感染症関連データの解析　など

の研究課題に取り組んでいます。



本研究室では統計的機械学習とその応用研究に興味のある学生さんを募集しています！

【統計的機械学習】

- 統計科学を用いて、
  - データから、内在する数学的な構造を発見する。
  - その数学的な構造に基づいて、予測や判別などの情報処理を行う。
- 帰納的アプローチ

  v.s.
- 自然科学でよく見られる演繹的アプローチ
  - 仮説をたて、推論し、実験的または理論的に検証する。

- カーネルマシン
  - 自動的な特徴(/モデル)選択機構を含む。
  - 非線形の扱いに優れている。
  - サポートベクターマシン(SVM)、罰金付ロジスティック回帰マシン
- いろいろな確率モデルによる方法
  - 混合ガウス分布モデル
  - 隠れマルコフモデル
- ガウス過程状態空間モデル
- 深層学習　　など

【Tukey g-and-h ランダムフィールドモデルを用いた都市部の熱波の時空間解析】

背景

■都市部のヒートアイランド（UHI）問題の深刻化
- ◆ヨーロッパの大熱波（2003年、死者推定 52,000 人以上）
- ◆ロシアの熱波(2010年、死者推定約 15,000 人)
- ◆インド、中東、オーストラリア等の熱波による高い死亡率

■都市環境における空間的に細かい粒度で温度過程を定量化し、モデル化する必要性
- ◆気候変動への取り組みやスマートシティの設計
- ◆CO2排出削減基準の政策の企画・運用
- ◆高齢化の進む我が国において熱波対策は特に重要

目的

■時間的・空間的な温度モデルの解像度の向上

■異なるデータソースを組合せ、局所的な都市環境における気温と地表面温度（地温）をモデル化する方法の開発

■首都圏：人口3,500万のうち高齢者人口は世界最大級

方針

■空間的メッシュの気温・地温エミュレータの作成
- ◆異なるデータソースの利用
  - 衛星リモートセンシング画像
  - 地上の様々な モニタリングサイトデータ

■Tukey g-and-h random fieldモデルによる解析
- ◆温度過程の空間的・時間的側面を捉える
- ◆Tukey g-and-h random field～ガウス過程（GP）の拡張
  - 歪度と尖度を明示的にパラメータ化できる変換やワーピング特性の組み込み可能

従来：都市環境における時間的・空間的温度モデリング

■従来アプローチ①：GP（クリギング）の利用
- ◆標本点数n、観測回数Tを大きく取ることにより時間・空間分解能を向上
- ◆推定と予測にかかる計算コスト の 問題 $O((nT)^3)$
  - 時間分解能：分単位 → Tは非常に大きくなる

■従来アプローチ②：max-stable過程の利用
- ◆多変量極値分布をGPの枠組みに拡張
  - 尤度は閉形式を持たない → nやTが大きい場合に問題となる
- ◆時系列の特徴づけに関する研究は少ない
- ◆極値分布が空間と時間で均質であることを仮定

気温・地温データ

■地温（地表面温度）
- ◆都市の道路や建物などからの熱の吸収と放射を反映
- ◆熱波やUHIの強さを決定

■気温（大気温度）データソース
- ◆気象庁データ：1時間ごとの78観測地点の気温
- ◆NTTドコモデータ：1分ごとの206観測地点の気温

■地温データソース
- ◆MODISのリモートセンシング画像データ
  - 毎日1時30分、10時30分、13時30分、22時30分
  - 空間分解能1kmの全世界の地温分布図を生成

本研究：都市環境における時間的・空間的温度モデリング

■本アプローチ：Tukey g-and-h random field（TGH-RF）の利用
- ◆歪度と尖度を明示的にパラメータ化
- ◆非ガウス型自己回帰過程を捉えることができる
- ◆計算コストの問題への対応
  - 低ランクTGH-RFモデリング
  - スパースTGH-RFモデリング
    - ➢Local approximate GP (laGP) [Gramacy et al., 2015]の考えに基づく拡張

首都圏の高解像度エミュレート地温データセットの作成

TGH-RFモデルを適用して熱波解析

地温の空間的解析

■予備検討：地温の分布には歪みと尖りがあり、その強さは場所によって大きく異なる。

■TGH-RF (Tukey g-and-h random field) モデルを用いて、歪度や尖度に基づいて地温データの背後にあるプロセスを分析

■問題：TGH-RFモデルの計算コスト
- →低ランクモデルへの拡張
  - 空間的に一定と仮定して歪度・尖 度を推定
- →スパースモデルへの拡張
  - 局所的に歪度・尖度を推定

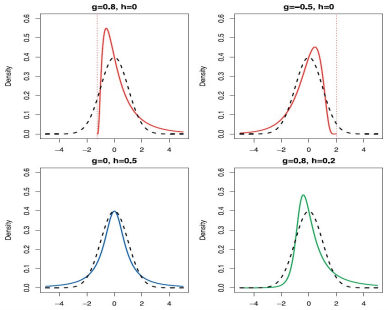
Tukey g-and-h (TGH)分布

$$\tilde{Y} = a + b\tau_{g,h}[Z], \quad (6)$$

$$\tau_{g,h}[Z] = \frac{1}{g}exp\left(gZ-1\right)exp\left(h\frac{Z^2}{2}\right), \quad (7)$$

- $Z \sim N(0,1)$
- a：場所
- b：スケール
- g：歪度
- h：尖度
- もし$g,h \to 0$ならば、a,b は $\tilde{Y}$ の平均と標準偏差へ

TGH分布
歪度と尖度の分析に有効



TGH random field (TGH-RF) モデル

■分布の歪度と尖度＋空間依存性

■ガウス過程 (GP) モデルに歪度と尖度のパラメータを組み込む

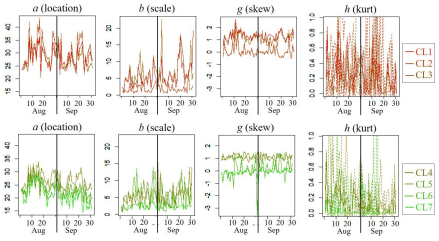
$$\tilde{Y}_t(s_i) = a + b\tau_{g,h}[Z_t(s_i)] \quad (8)$$

$$\tau_{g,h}[Z_t(s_i)] = \frac{1}{g}exp\left(gZ_t(s_i)-1\right)exp\left(h\frac{Z_t^2(s_i)}{2}\right), \quad (9)$$

- $Z_t(s_i)$：ガウス確率変数
- $E[Z_t(s_i)] = 0$
- $Var[Z_t(s_i)] = 1$
- $Cov[Z_t(s_i), Z_t(s_j)] = c[d(s_i, s_j)]$

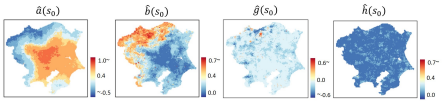
TGH-RFモデル
g,hを最尤推定
問題：計算コスト

TGH-RFモデルによるパラメータ推定結果

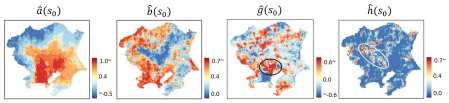


TGH-RFモデルによる推定結果

[7月19日を除く11日間の平均値で推定されたパラメータ]



[最高気温の平均が最も高くなった7月19日の推定パラメータ]
黒い円は都心部、白い円は内陸部で他の地域よりも気温が高い



まとめ

■地上観測とリモートセンシング観測を組合せてエミュレートした地温の空間的・時間的な歪度と尖度を解析
- ◆空間的解析：低ランクでスパースなTGH-RFモデル
- ◆時間的解析：l-moment matching法

■熱波モデリングにおいて歪度と尖度を考慮することの重要性を示した。
- ◆相当な暑さに見舞われることが知られている都心部では強い歪度値を推定
- ◆低ランク・スパースTGH-RFモデルは大規模な空間データのテール構造を明らかにするのに有効