

ガウス過程状態空間モデルによる非線形な準周期的現象の予測

玉森 聡 モデリング研究系 特任研究員

2015年6月19日 統計数理研究所 オープンハウス

研究背景

身の回りの様々な周期的現象

- 捕食-非捕食関係にある生物の個体数
 - 太陽活動(ex.黒点数)や大気海洋相互作用(ex.エルニーニョ現象)
 - 生体リズム(睡眠, 心拍, 脳波, 基礎体温など)
- ⇒非線形な変動かつ周期に揺らぎ(準周期性)

非線形な準周期的現象の予測により得られるメリット

- 生体リズムを考慮した投薬計画や医薬品開発
- 経済指標の予測精度向上に伴う最適な設備投資額の決定
- 基礎体温や月経周期の予測による女性のQOL向上 など

周期変動データのための時系列解析の手法

- SARIMAモデル, 季節調整モデルなどの状態空間モデル
- 問題点: 自然現象のモデリングには不適(周期固定など) & 周期自体の定量的予測はあまり顧みられてこなかった

ガウス過程状態空間モデルによる準周期的現象のモデル化

- 位相を潜在変数として導入 ⇒ 位相のゆらぎを表現
- 未知の非線形関数をノンパラメトリックに自動選択
- 逐次ベイズフィルタに基づいた位相の逐次予測手法を導出 ⇒ 周期の定量的予測手法を実現

準周期的現象を高精度に予測可能な手法の実現

状態空間モデル

$$\begin{aligned} \mathbf{x}_t &\sim p(\mathbf{x}_t | \mathbf{x}_{t-1}) && (\text{システムモデル}) && \mathbf{x}_t : \text{状態変数} \\ \mathbf{y}_t &\sim p(\mathbf{y}_t | \mathbf{x}_t) && (\text{観測モデル}) && \mathbf{y}_t : \text{観測変数} \end{aligned}$$

- マルコフ性を仮定⇒条件付き分布を簡略化
- 漸化式による効率的な状態推定(一期先予測, フィルタ, 平滑化)
- 観測変数が二項分布やポアソン分布に従う場合もモデル化可能

ガウス過程

確率変数の有限個の集まりが結合ガウス分布に従う性質

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

平均関数 $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$

共分散関数 $k(\mathbf{x}, \mathbf{x}') = \text{Cov}(f(\mathbf{x}), f(\mathbf{x}'))$

- ⇒平均関数と共分散関数による特徴付け
- 共分散関数により複雑な相関関係を扱うことが可能(周期性など)
- 具体的な関数形をデータ駆動でノンパラメトリックに決定

非線形な準周期的現象のモデル化

モデリングの対象: 女性の基礎体温と月経周期

- 第 n 日目の位相(潜在変数) θ_n
 - $0 \leq \theta_n \leq 1$
 - 第 n 日目が月経開始日のとき $\theta_n = 1$
 - その後 $\theta_n = 0$ にリセット
- 第 n 日目の体温(観測変数) y_n

状態空間モデルによる定式化

$$\begin{aligned} \theta_n &= \theta_{n-1} + \epsilon_n \pmod{1} && \epsilon_n \sim \text{ga}(\alpha, \beta) : \text{ガンマ分布} \\ y_n &= g(\theta_n) + \sigma_n && \sigma_n \sim \mathcal{N}(0, \sigma^2) : \text{ガウス分布} \end{aligned}$$

$$g(\theta_n) = a_0 + \sum_{m=1}^M [a_m \cos(2m\pi\theta_n) + b_m \sin(2m\pi\theta_n)] \quad (\text{次数 } M \text{ の級数近似})$$

逐次予測手法(右上へ続く)

- 第 n 日目から第 k 日目 ($n \leq k$) までに加えられる位相の増分

$$\Delta(k | n) \stackrel{\text{def}}{=} \epsilon_n + \epsilon_{n+1} + \dots + \epsilon_k \quad \Delta(k | n) \sim \text{ga}(k\alpha, \beta) \quad (\text{再生性})$$

- 第 k 日目が次回月経開始日となる事象 $\Delta(k | n) > 1 - \theta_n$
- 次回月経開始日を迎えるための位相の増分は $1 - \theta_n \pmod{1}$

逐次予測手法(左下からの続き)

- 事象が実現する確率 (=条件付き累積分布関数)

$$\begin{aligned} F(k | \theta_n) &\stackrel{\text{def}}{=} \Pr(\Delta(k | n) > 1 - \theta_n) = \int_{1-\theta_n}^{\infty} \text{ga}(x; k\alpha, \beta) dx \\ &= 1 - \text{Ga}(1 - \theta_n; k\alpha, \beta) \quad \text{ガンマ分布の累積分布関数} \end{aligned}$$

- 条件付き確率質量関数の計算

$$\begin{aligned} p(k | \theta_n) &\stackrel{\text{def}}{=} F(k | \theta_n) - F(k - 1 | \theta_n) \\ &= \text{Ga}(1 - \theta_n; (k - 1)\alpha, \beta) - \text{Ga}(1 - \theta_n; k\alpha, \beta) \end{aligned}$$

- 予測分布の計算 $p(k | y_{1:n}) \stackrel{\text{def}}{=} \int p(k | \theta_n) p(\theta_n | y_{1:n}) d\theta_n$
時刻 n におけるフィルタ分布

- 予測値の計算 $k_{max} = \text{argmax}_k p(k | y_{1:n})$

ガウス過程状態空間モデルによる定式化

$$\begin{aligned} f(\theta) &\sim \mathcal{GP}_f(m_f(\theta; \Phi_f), k_f(\theta, \theta'; \Theta_f)) \\ \theta_n &= f(\theta_{n-1}) + \epsilon_n \pmod{1}, \quad \epsilon_n \sim \text{ga}(\alpha, \beta) \end{aligned}$$

$$\begin{aligned} g(\theta) &\sim \mathcal{GP}_g(m_g(\theta; \Phi_g), k_g(\theta, \theta'; \Theta_g)) \\ y_n &= g(\theta_n) + \sigma_n, \quad \sigma_n \sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

$\Phi_f, \Theta_f, \Phi_g, \Theta_g$: ハイパパラメタ

⇒状態/観測関数, ハイパパラメタをParticle MCMC法により推定

逐次予測手法

- 予測分布の計算⇒フィルタ分布のモンテカルロ積分による近似
- 予測値の計算⇒状態空間モデルによる定式化と同様

実験

女性の基礎体温データから次回月経開始日を予測

実験条件

- 学習データ: 30代女性の口中計測 日次体温データ130日分
- テストデータ: 同女性の口中計測 日次体温データ311日分
- Particle MCMCの サンプル回数10000, 粒子数5000
- GPのカーネル関数: RBFカーネル

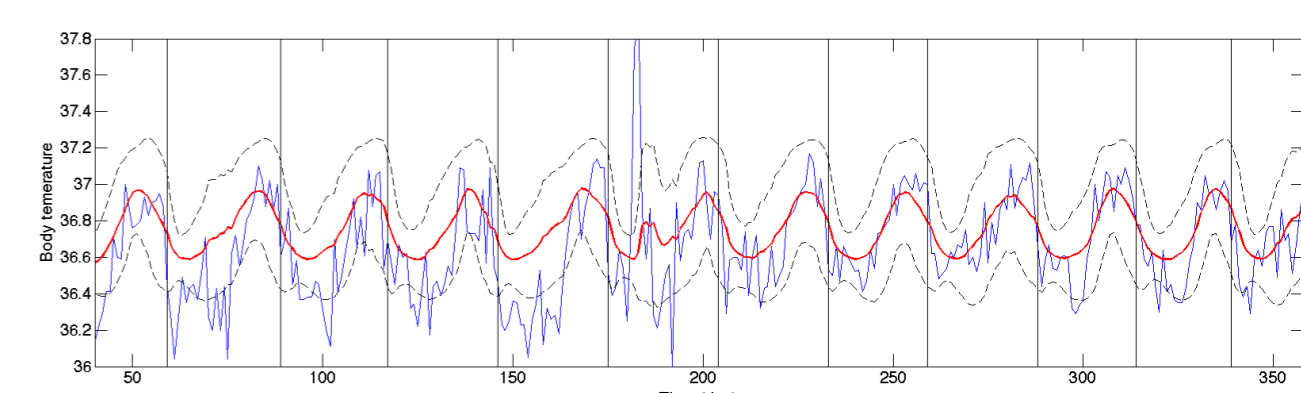
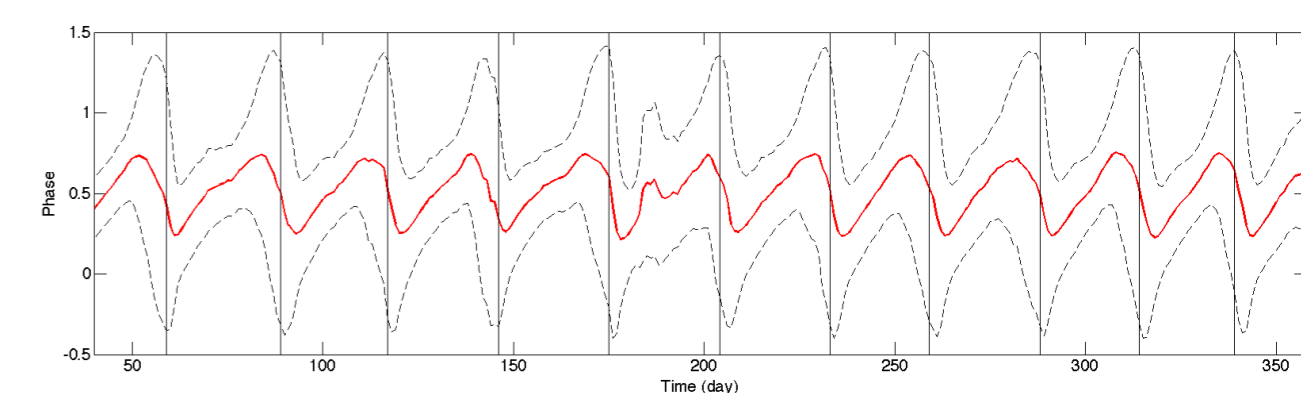
比較手法: 前月の月経開始日から 28日&31日後を予測日とする手法

評価基準: 最小二乗誤差 (RMSE) および最小絶対誤差 (MAE)

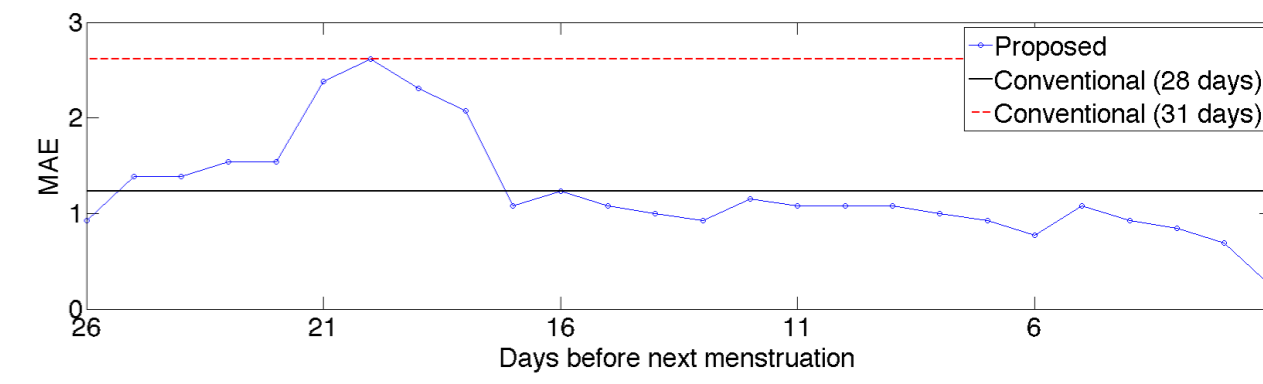
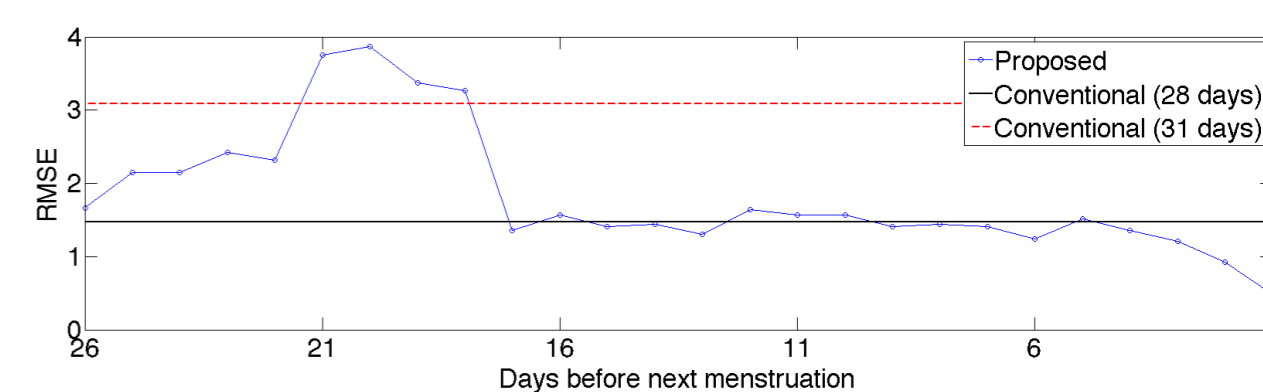
実験結果: 31日後に予測する手法からの精度改善を確認

今後の課題

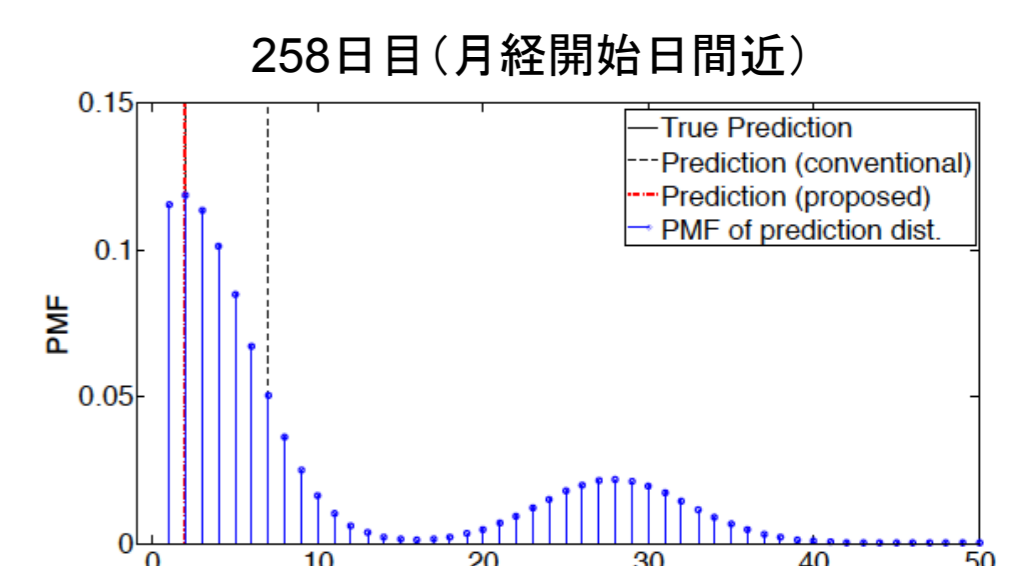
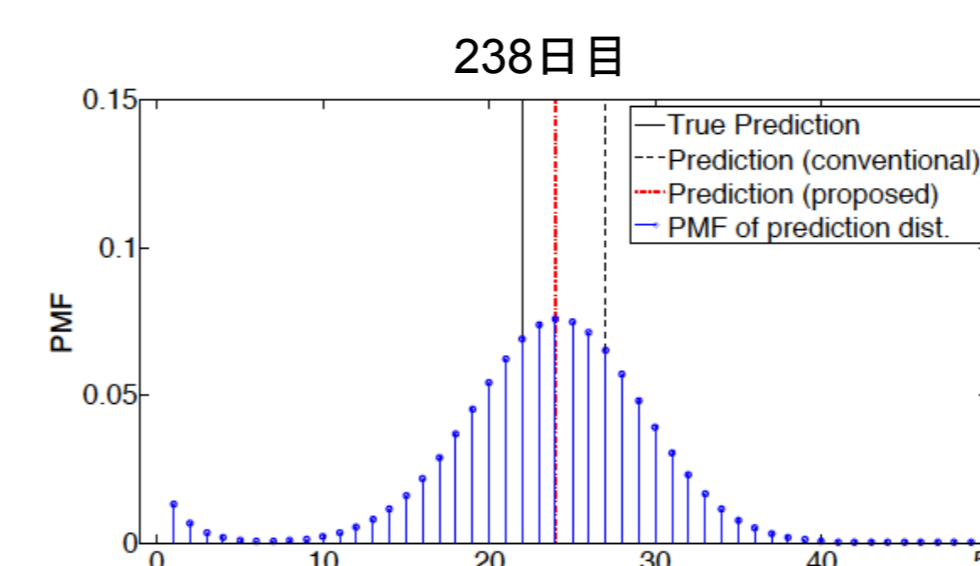
- 観測の異常値に対しても頑健なモデリング手法の検討
- SARIMAモデルや季節調整モデルとの比較実験
- 周期カーネルを適用した比較実験
- 被験者数を増やした実験



上段: テストデータに対する位相の推定平均(赤線)
下段: テストデータ(青線)と体温の推定平均(赤線)
ただし点線は(平均)+(標準偏差)×2, 縦線は月経開始日(既知), 横軸は時間(日にち)を表す



次回月経開始日の予測誤差(上段: RMSE, 下段: MSE): 横軸は次回開始日までの日数を表す



予測の例: 青のプロットが予測分布, 黒線が次回月経開始日までの日数(真の予測値), 赤点線が提案法の予測値, 黒点線が比較手法の予測値. 横軸は予測分布のインデックス k , 縦軸は予測分布の値を表す