2021年6月18日　統計数理研究所　オープンハウス

統計的機械学習の応用研究

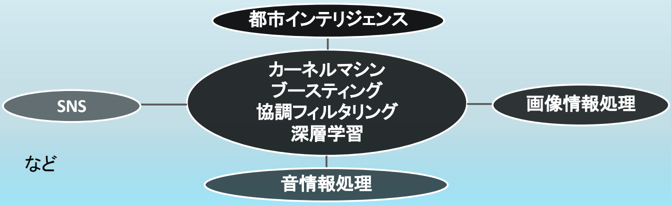
松井　知子　　モデリング研究系　教授

【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

【 A ROBUST HIGH-DIMENSIONAL BAYESIAN FILTER 】

Background

- Task: Estimate a signal from a sequence of noisy observations.
- A general model:

$$p(\boldsymbol{x}_{0:t}, \boldsymbol{y}_{1:t}) = p(\boldsymbol{x}_0)\prod_{k=1}^{t} f_k(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})g_k(\boldsymbol{y}_k|\boldsymbol{x}_k)$$

- Bayesian framework: obtain all relevant info. about x_t from the filtering distribution $p(x_t|y_{1:t})$

✕**Difficult to evaluate the filtering distribution analytically especially for non-linear models.**

Approaches for non-linear models in 1990s:

- Sequential Monte Carlo (**SMC**) / particle filters (**PF**) [Doucet et al., 2000]
  - Numerical integration techniques to approximate $p(x_t|y_{1:t})$

✕**Computational complexity**

✕**Importance sampling design is not efficient for high-dim problems.**

- Ensemble Kalman filter (**EnKF**) [Rebeschini et al., 1994]
  - A main numerical technique for solving him-dim forecasting and data assimilation problems
  - A recursive Monte Carlo filter to approximate $p(x_t|y_{1:t})$ empirically by an ensemble of random samples.

✕**Not robust to observation errors which are heavy-tailed and/or skewed distributed.**

Recent robust approach:

- **GEnKF** [Katzfuss et al., 2019]
  - A hierarchical state space model with *t*-distributed observation errors
  - A combination of an EnKF and a Gibbs sampler

✕**Not robust to observation errors which are skewed distributed.**

Objective

- Generalize **GEnKF** by utilizing a more general and flexible distribution for the observation errors, generalized hyperbolic (**GH**) distribution†.

  †[Barndorff-Nielsen, et al., 1982]
- Enable to model more practical cases with possibly skewed data.

→ Propose **GH-GEnKF**

Stochastic GH-GEnKF: Generalized hyperbolic (GH) distribution

- A normal variance-mean mixture:

$$y = \mu + \theta\gamma + \sigma\sqrt{\theta}z$$

- **GH** distribution: a prominent example
  - θ: Generalized inverse Gaussian (GIG) with 3 parameters $(\lambda \in \mathbb{R}, \chi \geq 0, \psi \geq 0)$
  - Extremely flexible to model heavy tailed and skewed data
  - Include Student *t*, Laplace, hyperbolic, and normal inverse Gaussian as special cases.

- Use the GH distribution as likelihood to obtain a general filtering technique robust to outliers.

$$\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}_t, \gamma_t \sim \mathcal{N}(\boldsymbol{H}_t\boldsymbol{x}_t + \gamma_t\boldsymbol{C}_t \circ \boldsymbol{\theta}_t, \boldsymbol{R}_t(\boldsymbol{\theta}_t))$$
$$\boldsymbol{x}_t|\boldsymbol{x}_{t-1} \sim f_t(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$$ **Independent of parameters**
$$\theta_{t,n} \sim GIG(\lambda_t, \chi_t, \psi_t) \text{ for } n = 1, \ldots, n_{\boldsymbol{y}}$$
$$\gamma_t \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma^2)$$

γ_t : Skewness

$\boldsymbol{R}_t(\boldsymbol{\theta}_t) = \sigma_t^2 \mathrm{diag}(\theta_{t,1}, \ldots, \theta_{t,n_{\boldsymbol{y}}})$: Kurtosis

$(\lambda_t,\ \chi_t,\ \psi_t,\ \sigma_t^2)$: assumed to be known (static)

$\boldsymbol{C}_t$: a known vector composed of explanatory variables

- Forecast step:
  - Draw $x_{t-1|t-1}^1, \cdots, x_{t-1|t-1}^N$ from

$$p(\boldsymbol{x}_t|\boldsymbol{\theta}_t, \gamma_t, \boldsymbol{y}_{1:t-1}) = p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})p(\boldsymbol{\theta}_t, \gamma_t|\boldsymbol{y}_{1:t-1})$$

  - Propagate each ensemble member using the transition pdf.

$$\boldsymbol{x}_{t|t-1}^j \sim f_t(\boldsymbol{x}_t|\boldsymbol{x}_{t-1|t-1}^j)$$

- Joint distribution of all variables at time *t*:

$$p(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{\theta}_t, \gamma_t|\boldsymbol{y}_{1:t-1}) = p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}_t, \gamma_t)p(\boldsymbol{\theta}_t)p(\gamma_t) \times p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})$$

Conclusion

- A robust ensemble Kalman filter is proposed which is able to deal with high-dimensional state process in the presence of heavy-tailed and skewed observation errors.
- Numerical simulations empirically show that this proposed **GH-GEnKF** outperforms existing ensemble Kalman filtering techniques.

Numerical simulations

Evaluation points:

- Performance for high-dim problems
- Robustness for observations with heavy-tailed (and possibly skewed) noise

Simulated system:

- The latent process x_t consists in a time-varying spatial physical phenomenon.

  $n_{\boldsymbol{x}} = 100$ on a one-dimensional spatial domain $[1, 100]$

  $n_{\boldsymbol{y}} = 75$ randomly chosen observation locations

Experimental conditions:

- Prior pdf of the state:

$$f_t(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \mathcal{M}(\boldsymbol{x}_{t-1}), \boldsymbol{Q})$$

$\mathcal{M}$: state evolution operator
  - Linear dynamics case: $\mathcal{M}(\boldsymbol{x}_{t-1}) = 0.9\boldsymbol{x}_{t-1}$
  - Nonlinear dynamics case:

$$[\mathcal{M}(\boldsymbol{x}_{t-1})]_l = \frac{x_{t-1,l}}{2} + \frac{25x_{t-1,l}}{1 + x_{t-1,l}^2} \quad \text{for } l = 1, \ldots, n_{\boldsymbol{x}}$$

$\boldsymbol{Q}$: powered exponential covariance function (power 1.8 and scale parameter 10)

- Parameter setting in GH-GEnKF

$$\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}_t, \gamma_t \sim \mathcal{N}(\boldsymbol{H}_t\boldsymbol{x}_t + \gamma_t\boldsymbol{C}_t \circ \boldsymbol{\theta}_t, \boldsymbol{R}_t(\boldsymbol{\theta}_t))$$
$$\boldsymbol{x}_t|\boldsymbol{x}_{t-1} \sim f_t(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$$
$$\theta_{t,n} \sim GIG(\lambda_t, \chi_t, \psi_t) \text{ for } n = 1, \ldots, n_{\boldsymbol{y}}$$
$$\gamma_t \sim \mathcal{N}(0, 1)$$
$$\boldsymbol{R}_t(\boldsymbol{\theta}_t) = 1\ \mathrm{diag}(\theta_{t,1}, \ldots, \theta_{t,n_{\boldsymbol{y}}})$$

$\boldsymbol{C}_t$: binary variables which could reflect the quality of the sensors　(50% of the values: 1 at random observation locations)

Compared methods using 100 Monte Carlo runs:

- EnKF (with Gaussian noise, N=1,500)
- GEnKF (with *t*-distributed noise, N=30)
- GH-GEnKF (with GH distributed noise, N=30)

Simulation result

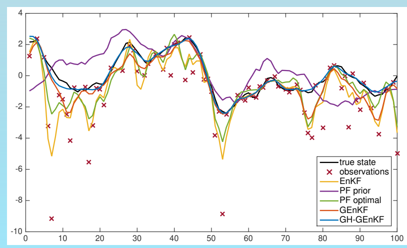


Figure 1. Simulated state and associated observations at time t = 3, together with posterior means obtained by the various methods with model parameters: γ = −2, λ = −1.5, χ = −2λ and ψ = 0.