Kernel Mean Particle Filter with Intractable Likelihoods

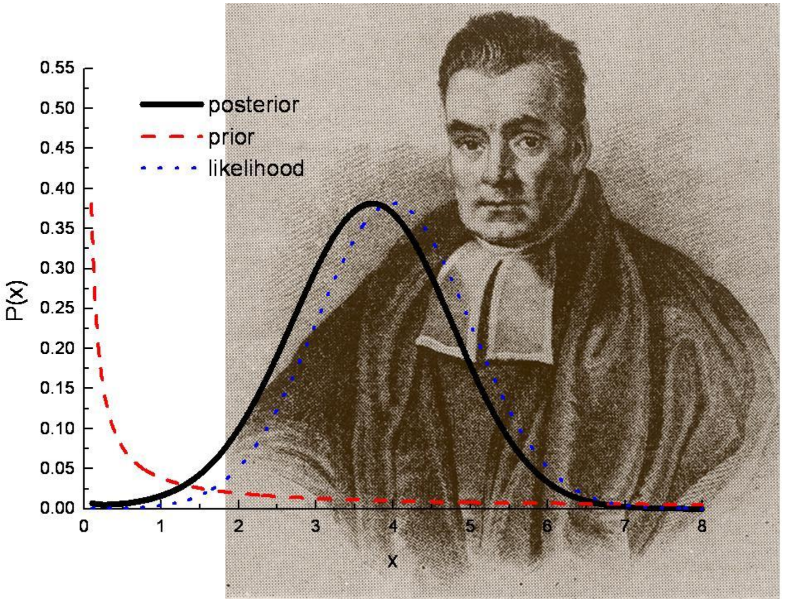
Kenji Fukumizu

The Institute of Statistical Mathematics

Joint work with
Motonobu Kanagawa (SOKENDAI/ISM) and Yoshimasa Uematsu (ISM)

July 17, 2015, STM2015&CSM2015

Intractable likelihood



- Bayesian theorem

$$p(X|Y_{obs}) = \frac{p(Y_{obs}|X)\pi(X)}{p(Y_{obs})}$$

- Bayesian inference needs the value of likelihood $p(Y_{obs}|X)$.

- What should we do if the likelihood is intractable?

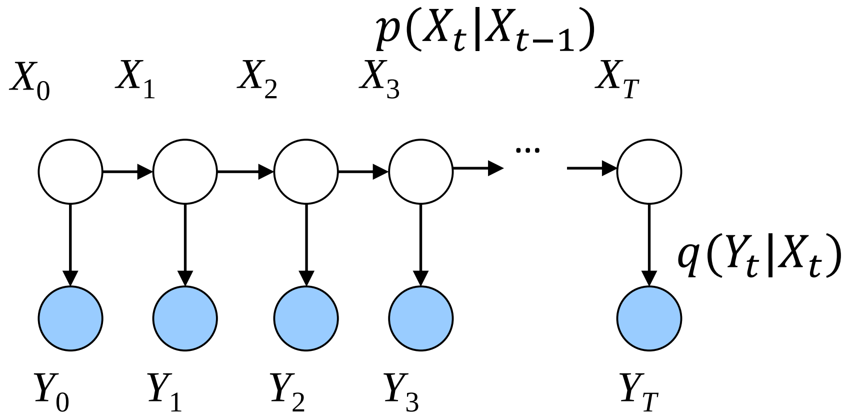
- When this happens?
  - Y may be given only by simulation
    - Population genetics:
      $Y \sim p(y|X)$ is given by a branching process.
    - Epidemiology:
      $Y \sim p(y|X)$ is given by solving (simulating) a stochastic differential equation.
  - Density function $p(y|x)$ is given only by a non-density form.
    - α-Stable distribution: Fourier transform is known
- Recent technology: ABC (Approximate Bayesian Computation)

Problem to solve

- Filtering with intractable likelihood
  - State space model

    $p(X_t|X_{t-1})$: state transition

    $q(Y_t|X_t)$: observation model



  - Assumption:

    Density $q(Y_t|X_t)$ is INTRACTABLE, but sampling is possible.

  - Note:

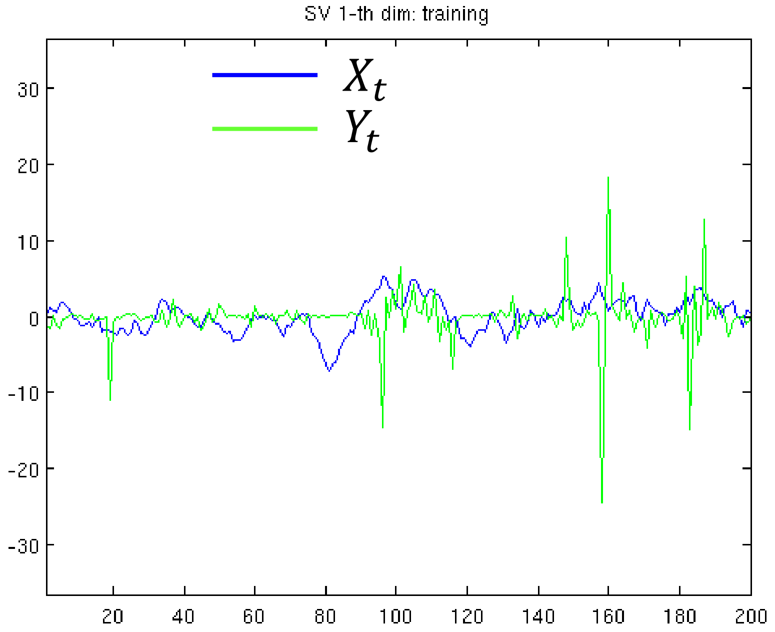
    Standard Sequential MC / Particle Filters are not applicable.

    They need the value $q(Y_t|X_t)$ for importance weighting.

- Example:
  α-stable Stochastic Volatility model

$$X_t = \phi X_{t-1} + \eta_t, \qquad \eta_t \sim N(0, \sigma_s^2)$$
$$Y_t = e^{X_t/2} w_t, \qquad w_t \sim S(\alpha, 0, \sigma_o)$$

X_t: log volatility, Y_t: return
Popular in mathematical finance



- Existing methods
  - Convolution particle filter (KDE-based) (Campillo & Rossi 2009)
  - ABC filter (Jasra et al 2012; Calvet & Czellar 2014)

Our approach

Kernel method for particle representation of a distribution

- Kernel mean embedding
  - Positive definite kernel / RKHS is used for nonparametric estimation
  - Good for (moderately) high-dimensional data

- A new way of Bayesian inference
  - Kernel mean can be regarded as “particle” representation.
  - Negative weights may appear (signed measure)
  - Bayesian inference is done by matrix computation

Representing distributions with kernel means

Positive definite kernel

<u>Definition</u>

Ω: set. $k: \Omega \times \Omega \to \mathbf{R}$ is a positive definite kernel if

(1) $k(x, y) = k(y, x)$

(2) For any $x_1, \dots, x_n$ in Ω, the Gram matrix $k(x_i, x_j)$ is positive semidefinite, i.e.,

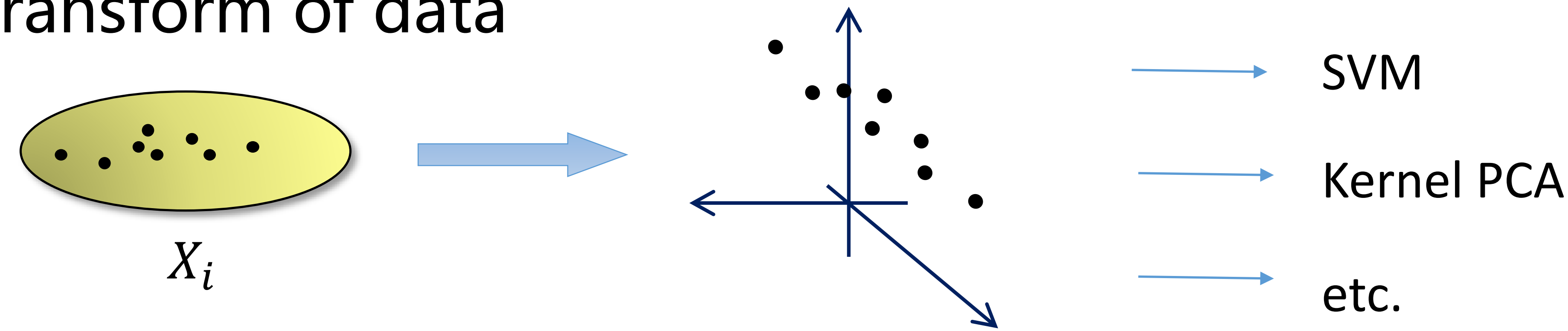
$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$$

for any $c_1, \dots, c_n \in \mathbf{R}$.

It is known that k uniquely defines a reproducing kernel Hilbert space (RKHS), which is a function space and used for a feature space.

Kernel method at a glance

- Nonlinear transform of data



Data space

$\Phi(X_i) = k(\cdot, X_i)$

Feature map

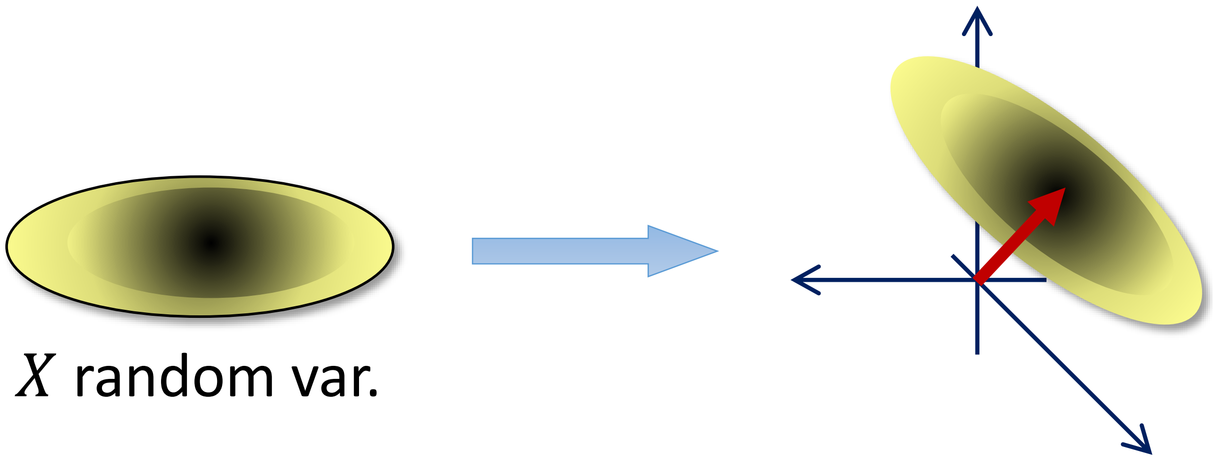
Feature space
（function space, RKHS）

- Kernel trick: special, efficient computation of inner product

$$\langle \Phi(x), \Phi(y) \rangle_{H_k} = k(x, y)$$

e.g. Gaussian RBF kernel $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$

Kernel mean



X random var.

Hilbert space-valued random variable

Data space

$\Phi(X) = k(\cdot, X)$

Feature map

Feature space, RKHS
（function space）

Kernel mean: $m_P = E_X[\Phi(X)]$

Kernel mean = Representation of distribution

- No information loss (with suitable choice of kernel, e.g. Gaussian)
  → Feature space is infinite dimensional (infinite components)

- Integral transform

$$m_P = E_X[\Phi(X)] = \int k(\cdot, x)dP(x)$$

Function again.

c.f. Characteristic function $\phi_P(\omega) = \int e^{\sqrt{-1}\omega^T x} dP(x)$

Kernel mean as a particle representation

- Estimator of kernel mean

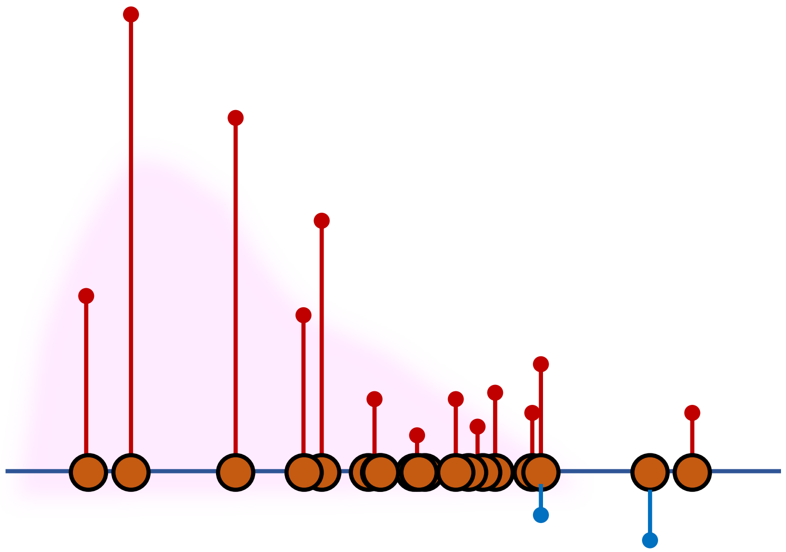
$$\hat{m}_P = \frac{1}{N}\sum_{i=1}^{N} k(\cdot, X_i)$$

$X_1, \dots, X_N \sim P,$ i.i.d.

- More generally

$$\hat{m}_P = \sum_{i=1}^{N} w_i k(\cdot, X_i)$$

Weighted sample expression (X_i, w_i)



Kernel version of importance weight

- Prior π: kernel mean $\hat{m}_\pi = \frac{1}{N}\sum_i k(\cdot, X_i)$
- Likelihood $p(y|x)$: intractable,
  but sampling possible

$$Y_i \sim p(y|x = X_i) \quad (i = 1, \dots, N)$$

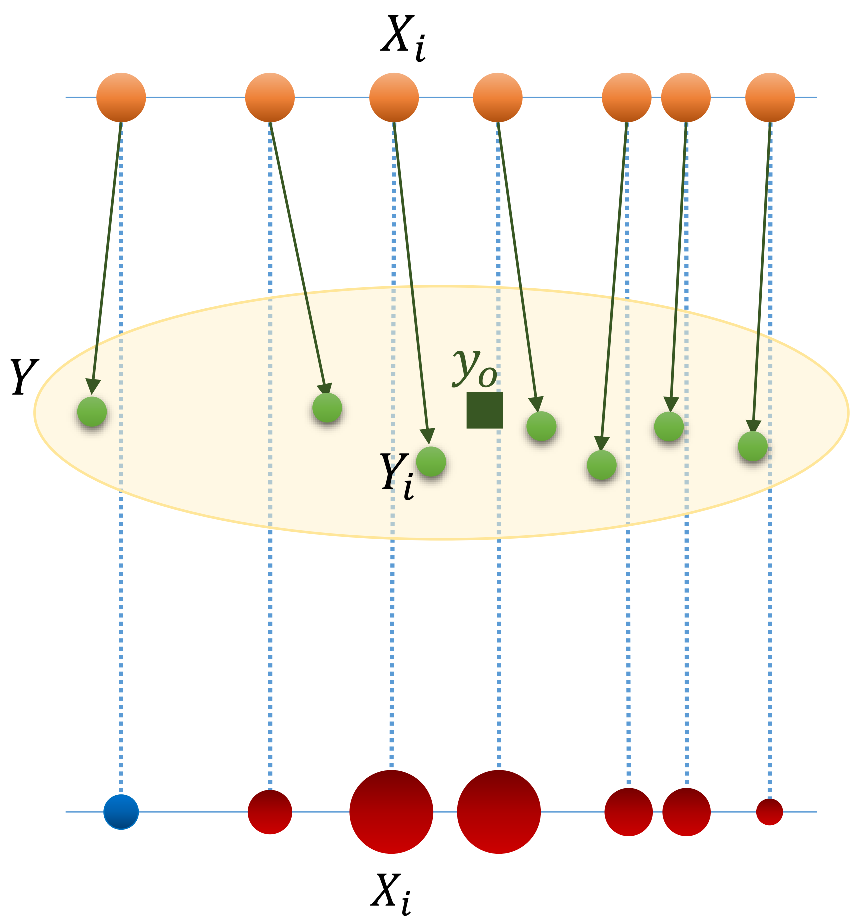
- Kernel mean of posterior given y_o : (X_i, w_i)

$$\hat{m}_{post} = \sum_i w_i k(\cdot, X_i)$$

$$w = (G_Y + \lambda I_N)^{-1} \mathbf{k}_Y(y_o)$$

ridge regression

$$G_Y = \left(k(Y_i, Y_j)\right), \quad \mathbf{k}_Y(y_o) = \left(k(Y_1, y_o), \dots, k(Y_N, y_o)\right)^T$$



Theory: convergence

Theorem

- $\widehat{m}_\pi = \frac{1}{N}\sum_{i=1}^{N} k(\cdot, X_i)$ is a consistent estimator of m_π with convergence rate $\|\widehat{m}_\pi - m_\pi\|_H = O_p\left(N^{-b}\right)$ $(0 < b \leq 1/2)$.
- $E[k(Y, Y')|X = x, X' = x']$ is a function in $H_X \otimes H_X$ as a function of (x, x'), where $Y \sim p(y|x), Y' \sim p(y'|x')$ independently.

Then for any $f(x)$ with $\int f(x)^2 \pi(x) dx < \infty$ and $\int f(x) p(x|y =\cdot) dx \in R(C_{YY})$ (range of covariance operator C_{YY}),

$$\sum_{i=1}^{N} w_i f(X_i) - \int f(x) p(x|Y = y_{obs}) dx = O_p\left(N^{-b/2}\right) \qquad (N \to \infty).$$

Recap: "Standard" particle methods

- Importance weight

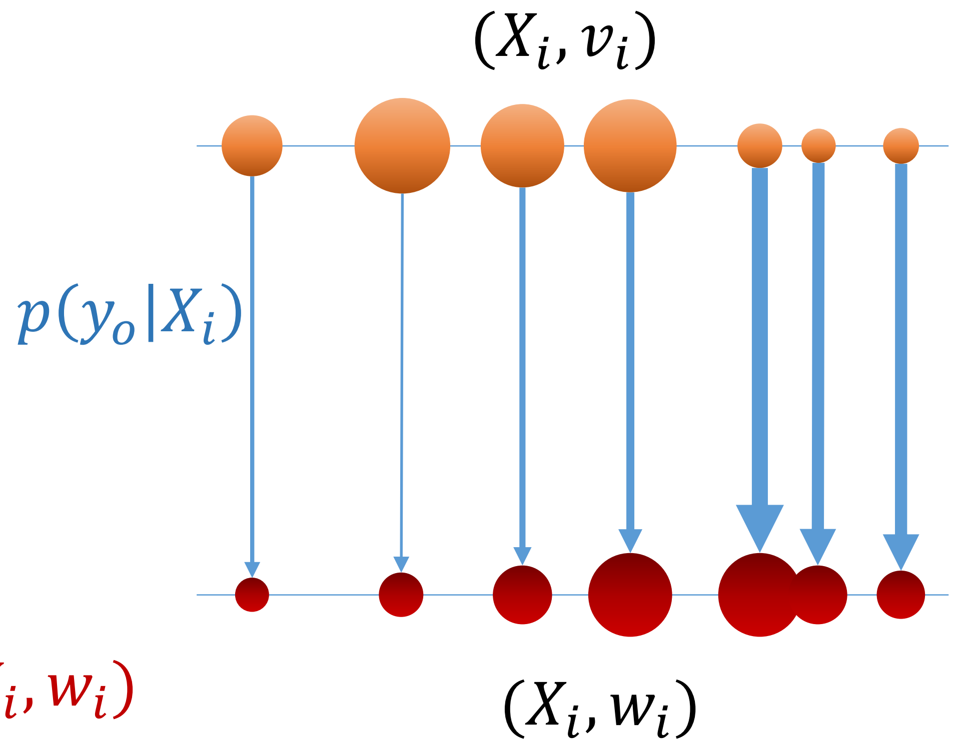
$$p(x|y_o) = \frac{p(y_o|x)\pi(x)}{\int p(y_o|x)\pi(x)dx}$$

- Prior π : (X_i, v_i) particle

$$\hat{\pi} = \sum_i v_i \delta_{X_i}$$

- Likelihood $p(y|x)$: known
- Given observation y_o,
  posterior $p(x|y_o)$ is represented by (X_i, w_i)

$$w_i \propto v_i \underbrace{p(y_o|X_i)}_{\text{Importance weight}}$$



Comparison: kernel vs standard particles

Kernel mean

$$\widehat{m}_P = \sum_{i=1}^{N} w_i k(\cdot, X_i)$$

- Estimator of kernel mean m_P
- Allows negative weights
- Bayesian inference with linear algebra

Standard

$$\sum_{i=1}^{N} w_i \delta_{X_i}$$

- Estimation by atomic probability
- (w_i) is a probability on N points.
- Bayesian inference with importance sampling

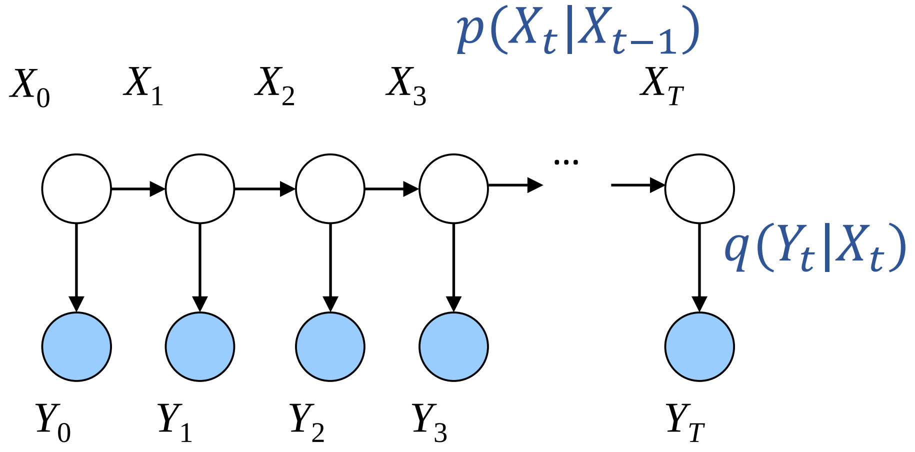
Kernel Mean Particle Filter

Re: Filtering with intractable likelihood

- Filtering with intractable likelihood
  - State space model

    $p(X_t|X_{t-1})$: state transition

    $q(Y_t|X_t)$: observation model



  - Assumption:

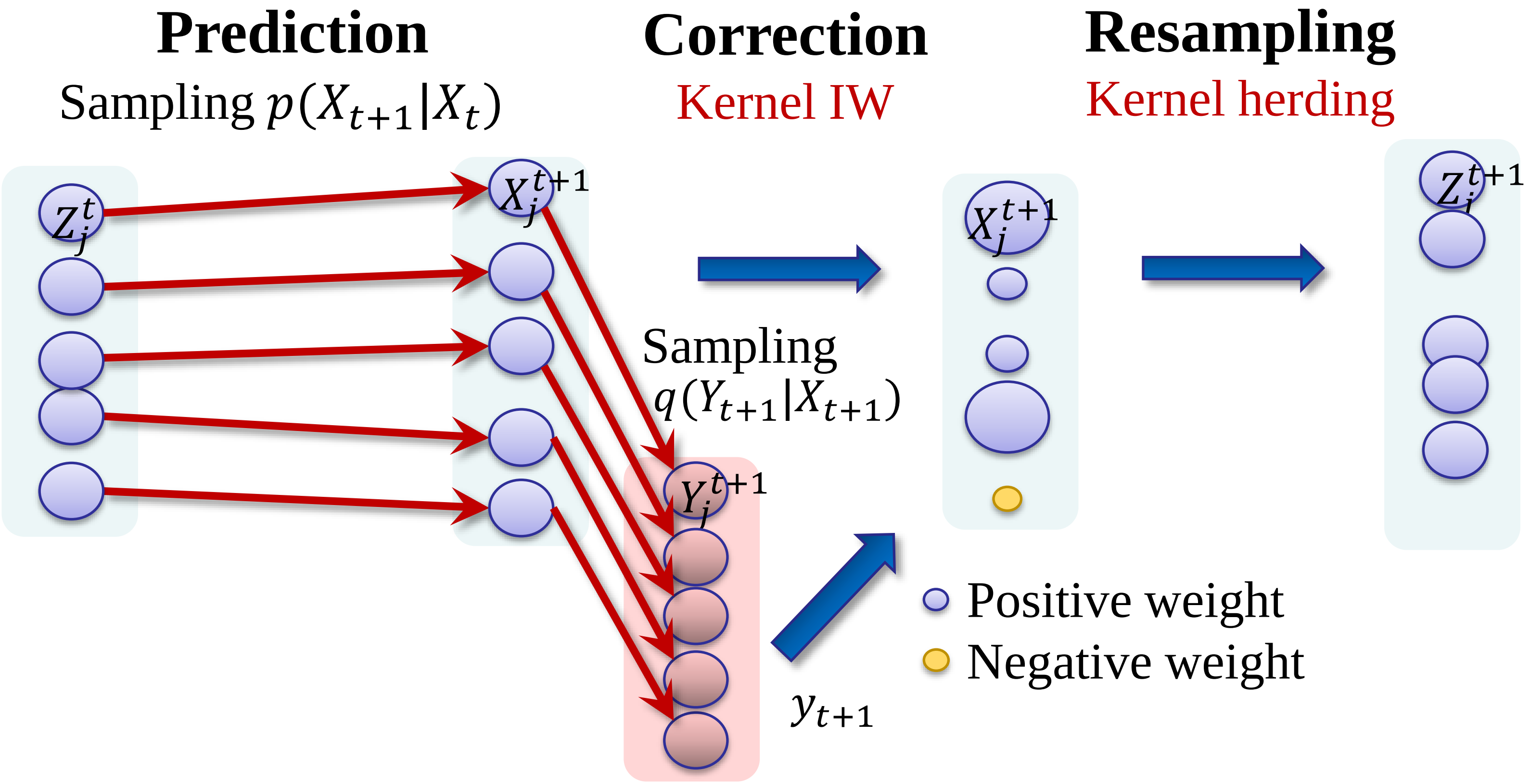
    $q(Y_t|X_t)$ is INTRACTABLE, but sampling is possible.

  - Apply the kernel IW for the intractable likelihood!

Kernel Mean Particle Filter (Fukumizu et al 2015+)

$p(X_t|y_1,\dots,y_t)$ $\quad$ $p(X_{t+1}|y_1,\dots,y_t)$ $\quad$ $p(X_{t+1}|y_1,\dots,y_{t+1})$ $\quad$ $p(X_{t+1}|y_1,\dots,y_{t+1})$

$\frac{1}{N}\sum_{j=1}^N k(\cdot,Z_j^t)$ $\quad$ $\frac{1}{N}\sum_{j=1}^N k(\cdot,X_j^{t+1})$ $\quad$ $\sum_{j=1}^N w_j^{t+1}k(\cdot,X_j^{t+1})$ $\quad$ $\frac{1}{N}\sum_{j=1}^N k(\cdot,Z_j^{t+1})$



Resampling by kernel herding

- Kernel herding (Chen et al 2010)
  - Find points $Z_1, \dots, Z_N$ so that the kernel mean $m_P = \int k(\cdot, x) dP(x)$ is approximated:

$$\min_{Z_1,\dots,Z_N} \left\| m_P - \frac{1}{N} \sum_{i=1}^{N} k(\cdot, Z_i) \right\|_H$$

  - Kernel herding solves $Z_1, Z_2, \dots$ sequentially.

$$Z_{\ell+1} = \arg\max_{Z} \; m_P(Z) - \frac{1}{\ell+1} \sum_{i=1}^{\ell} k(Z, Z_i)$$

  - KH shows good approximation accuracy in theory and practice.
    $O(1/N)$ in norm (NOT squared) for finite dimensional RKHS.

c.f. ABC filter

- Approximate Bayesian Computation (ABC)
  - Likelihood $p(y|x)$: intractable, but sampling possible
  - Simplest rejection method: Repeat 1-3.
    1. $X_i \sim \pi$
    2. $Y_i \sim p(y|x = X_i)$
    3. If $d(Y_i, y_{obs}) < \varepsilon$, Accept X_i; otherwise Reject.
  - If $\varepsilon \to 0$, the accepted sample approaches to a sample from $p(x|y_{obs})$, but acceptance rate becomes low.
  - For high-dimensional y, acceptance rate is lower. Low dimensional (sufficient) statistics are preferably used.
- ABC filter: apply ABC to the correction step in the particle filter.

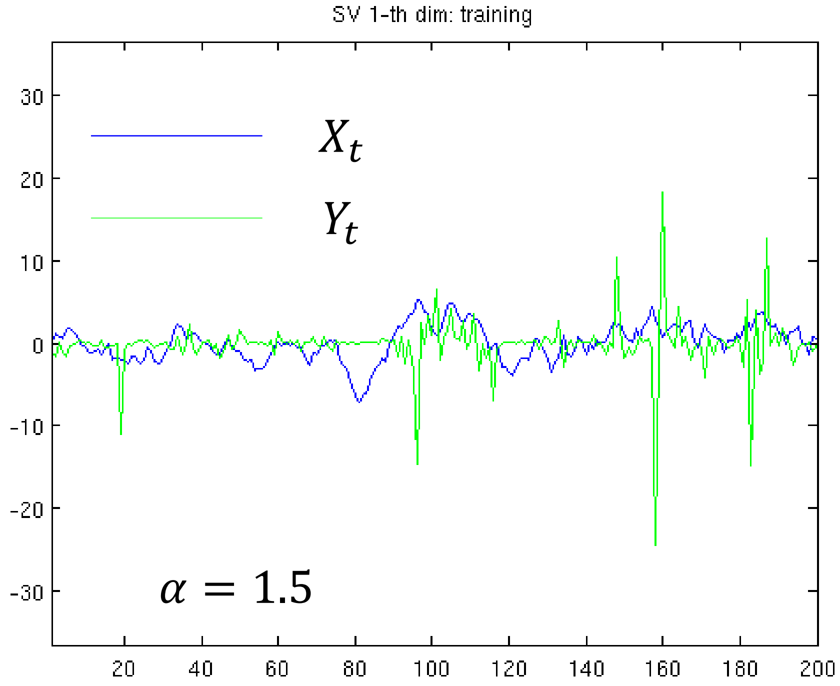
Application: Stochastic Volatility model

- Multivariate α-stable Stochastic Volatility model

$$X_t = \Phi X_{t-1} + v_t, \qquad \Phi = \mathrm{Diag}(\phi_1, \dots, \phi_d), \; v_t \sim N\left(0, \sigma_p^2 I_d\right)$$

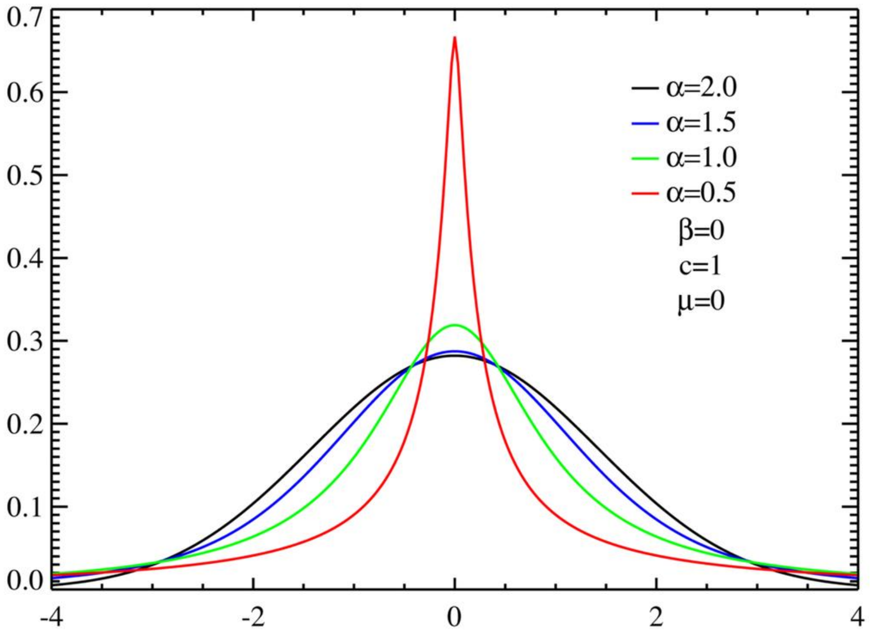
$$Y_t = V_t^{1/2} w_t, \qquad V_t = \mathrm{Diag}(e^{X_{t,1}}, \dots, e^{X_{t,d}}), \quad w_{t,i} \sim S(\alpha, 0, \sigma_o), \quad v_t \perp w_t$$

$S(\alpha, 0, \sigma_o)$: α-Stable distribution.
$\alpha = 2$: normal; $\alpha = 1$: Cauchy.
For general α: no analytic form for density, but sampling is possible.



- A model for volatility (degree of variations) of securities.
- Used popularly in mathematical finance.

α-stable distribution



$S(\alpha, \beta, c, \mu)$

α: index, β: skewness, c: scale, μ: shift

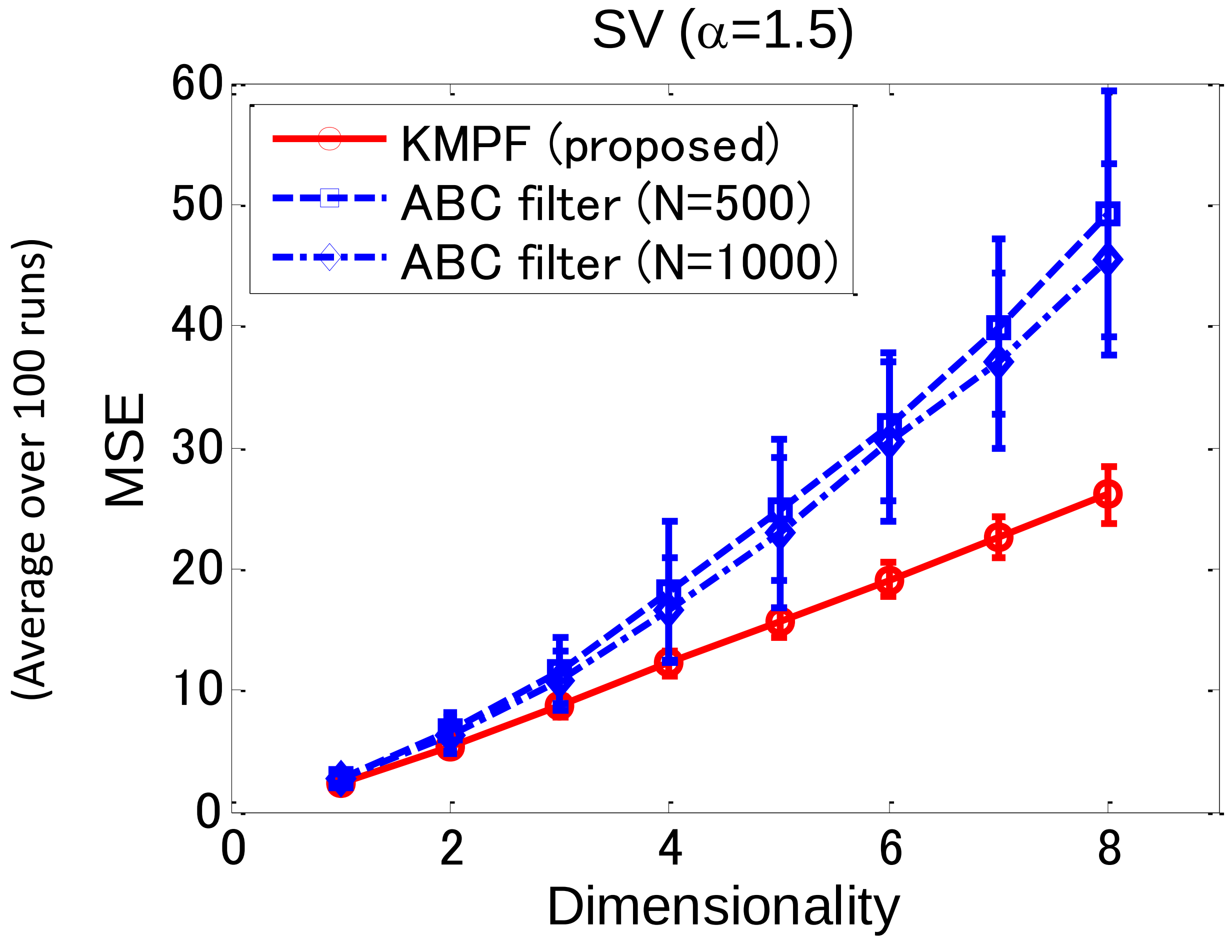
- Characteristic function

$$\phi(\omega; \alpha, \beta, c, \mu) = \int e^{\sqrt{-1}\omega x} dP(x; \alpha, \beta, c, \mu)$$
$$= \exp(\sqrt{-1}\mu\omega - |c\omega|^{\alpha}(1 - \sqrt{-1}\beta\omega\, sgn\,(\omega)\Phi)\,)$$

where $\Phi = \tan(\pi\alpha/2)$ for $\alpha \neq 1; \Phi = \frac{2}{\pi}\log|\omega|$ for $\alpha = 1$.

- $S(\alpha, \mu, \sigma) \coloneqq S(\alpha, 0, \sigma, \mu)$ (No skewness)

- $\alpha = 1.5$: intractable



CPU time / run

KMPF (500): 10.8 sec

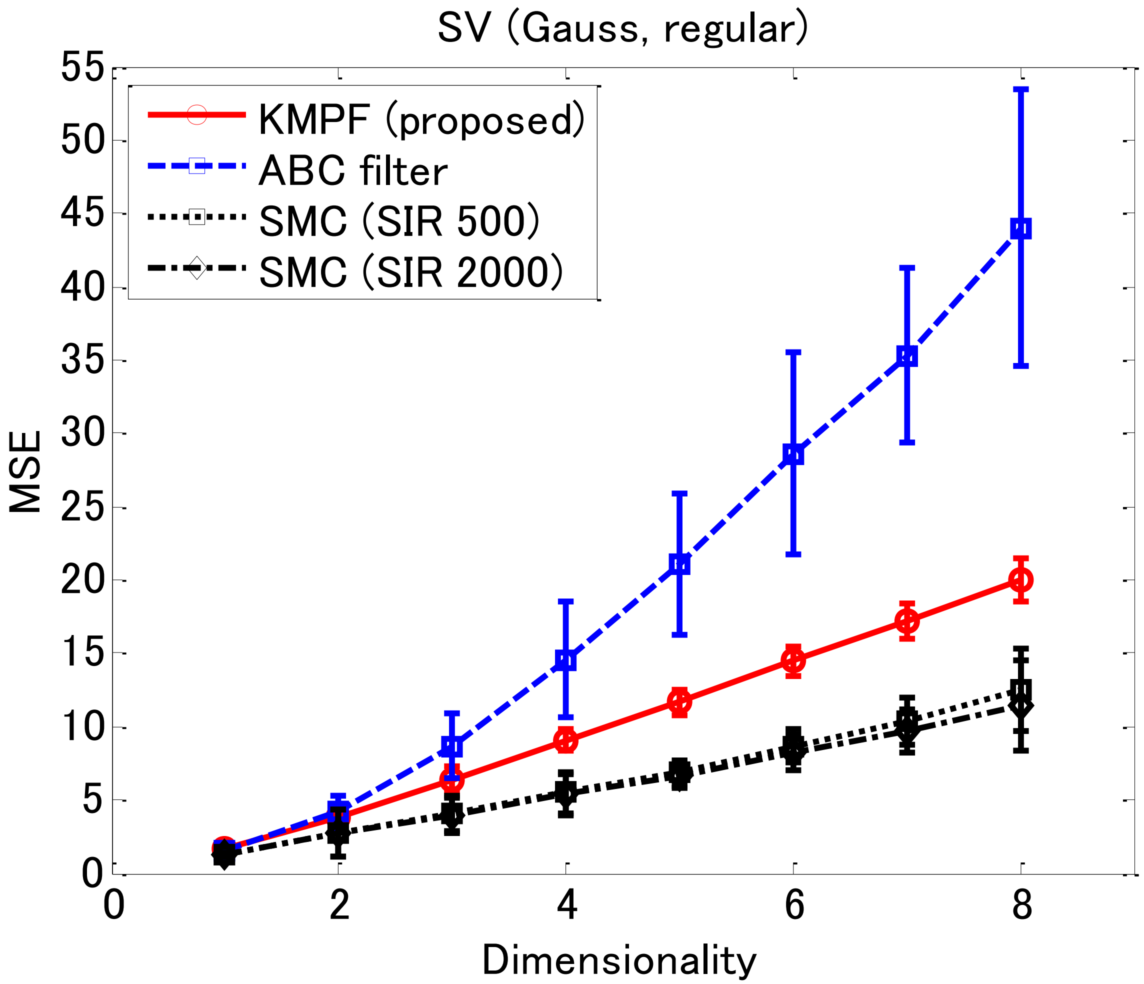
ABC filter (500): 5.6 sec

ABC filter (1000): 19.2 sec

Mean square errors in estimating X_t (point estimates, average over $T = 500$)

- $\alpha = 2$ (Gaussian, $w_t \sim N(0, \Sigma_o)$) Tractable case (standard SMC applicable)

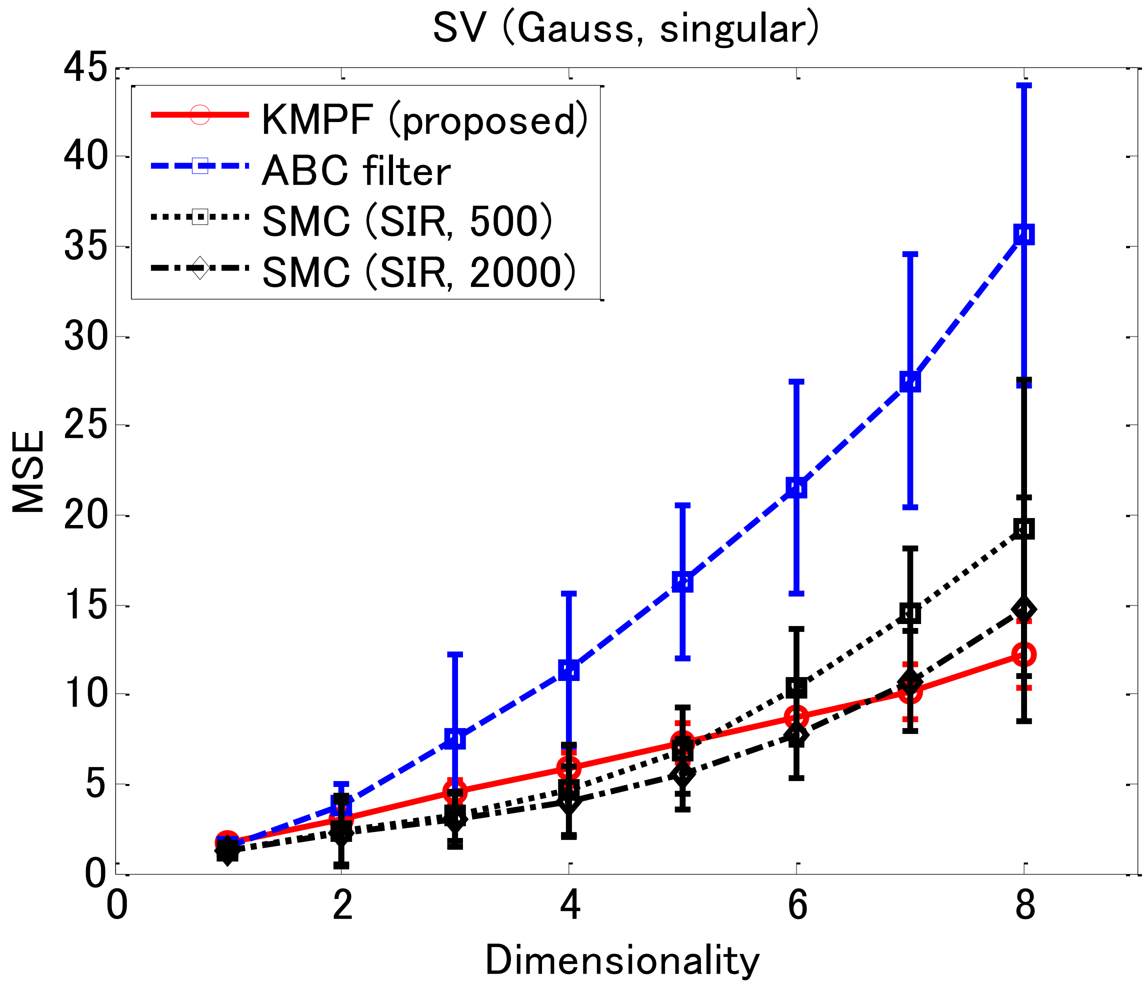
Σ_o: regular



CPU time / run

KMPF (500):	24.9 sec
ABC filter (500):	5.6 sec
SMC (SIR 500):	17.7 sec
SMC (SIR 2000):	142.6 sec

Σ_o: close to singular



Concluding remarks

- Kernel mean particle filter for intractable likelihoods
  - Kernel mean "particle" expression of distributions
    - Allows negative weights.
    - Matrix computation for updating weights.
    - Resampling by kernel herding.
  - Effective for filtering with intractable likelihoods
    - Works better than state-of-the-art ABC filters in high dimensional cases
    - Even better than standard SMC (SIR) in difficult cases.
      (Needs more comparisons.)
- Future directions
  - Estimation of parameters in state transition

Thank you.