Sequential Monte Carlo Samplers for Applications in High Dimensions

Alexandros Beskos

National University of Singapore

KAUST, 26th February 2014

Joint work with:

Dan Crisan, Ajay Jasra, Nik Kantas, Alex Thiery

Imperial College, NUS, Imperial College, NUS

Outline

Outline

Main References

Presentation based on papers:

α) *SMC Methods for High-Dimensional Inverse Problems: A case study for the Navier-Stokes equations*, (under revision, SIAM Journal of Uncertainty Quantification).

β) *On the Stability of SMC Methods in High Dimensions*, (forthcoming at The Annals of Applied Probability).

General Picture

- This talk is part of a broader collaborative research effort that aims at developing efficient principled Monte-Carlo methods for filtering problems in high dimensions.
- An important area of application is Data Assimilation, where the state of the art in terms of practical applications is probably the Ensemble Kalman Filter (Evensen, 09).
- A concern about Kalman-Filter-type methods is that they employ rather ad-hoc linearisations, thus their properties when applied to non-linear systems are yet to be fully understood.

Background

- Perceived idea in
  Data Assimilation (DA) - Sequential Monte-Carlo (SMC)
  communities that solving the full Bayesian problem for practical DA applications using particle filtering is infeasible.
- Due to weight degeneracy happening very fast.
- So, standard practice is to apply Kalman-Filter-type methods using Gaussian approximations.
- Yet, there have been new attempts trying to confront weight degeneracy for SMC from DA community (e.g. van Leeuwen (10), Chorin et al. (10)).
- Talk will show some efforts towards this direction from group from (mainly) SMC community.

Outline

SMC Method

- **Objective:** Obtain samples from sequence of target distributions of increasing dimension:

$$\Pi_1(x_{1:1})\,,\ \Pi_2(x_{1:2}),\ \ldots\ ,\Pi_n(x_{1:n})\,,\ldots$$

  Index n can represent time, or be fictitious.

- The construction involves also some kernel which increases the dimension, $M_n(x_{1:(n-1)}, dx_n)$.

- **Method:** Exploit sequential structure via:

  i) Importance Sampling

  ii) Resampling

  to generate sequence of weighted particles:

$$\{x_{1:n}^{(i)}, W_n^{(i)}\}_{i=1}^N \quad s.t. \quad \Pi_n(dx_{1:n}) \approx \sum_{i=1}^N W_n^i\, \delta_{x_{1:n}^i}(dx_{1:n})$$

Important Example: Particle Filtering

- **Model:** Consider State Space Model:

$$x_n|x_{n-1} \sim p(x_n|x_{n-1})\ , \quad y_n|x_n \sim p(y_n|x_n)\ .$$

- Of interest here is the posterior of the signal:

$$\Pi_n(x_{1:n}) \equiv p(x_{1:n}|y_{1:n})$$

- Here, we have that:

$$M_n(x_{1:(n-1)}, dx_n) = p(x_n|x_{n-1})dx_n$$

General SMC Algorithm

- Del Moral et al. (06).
- The Algorithm:

0. Initialise $x_{1:1}^{(i)} \sim M_1(x_{1:1})$ with $W_1^{(i)} = \frac{\Pi_1}{M_1}(x_{1:1}^{(i)})$. Set $n = 1$.
1. Given $(x_{1:n}^{(i)}, W_n^{(i)})$, get $x_{n+1}^{(i)} \sim M_{n+1}(x_{1:n}^{(i)}, dx_{n+1})$ and assign:

$$W_{n+1}^{(i)} = W_n^{(i)} \cdot \frac{\Pi_{n+1}(x_{1:(n+1)}^{(i)})}{\Pi_n(x_{1:n}^{(i)})\, M_{n+1}(x_{1:n}^{(i)}, x_{n+1}^{(i)})}$$

2. Calculate Effective Sample Size:

$$ESS_{n+1} = \frac{(\sum_{i=1}^{N} W_{n+1}^{(i)})^2}{\sum_{i=1}^{N} (W_{n+1}^{(i)})^2}$$

If $\frac{ESS_{n+1}}{N} < \alpha \in (0, 1)$ then resample and set $W_{n+1}^{(i)} = 1$.

4. Set $n = n + 1$. Return to Step 1.

Static Case

- Sequence of interest is on fixed dimension:

$$\Pi_1(x)\,,\,\Pi_2(x),\,\ldots\,,\Pi_n(x)\,,\ldots$$

- This can be cast into the general SMC framework of increasing dimension as long as for $x_{1:n} \sim \Pi_n(x_{1:n})$ we have $x_n \sim \Pi_n(x_n)$ (Del Moral et al. 06).

- A standard way for developing the SMC sampler is by specifying kernels $K_n(x, dx')$ such that:

$$\Pi_n K_n = \Pi_n$$

- MCMC methodology provides several candidates for K_n.

- For instance, Random-Walk Metropolis or Independence Samplers have been used in applications.

SMC Sampler (a version of it)

- Neal (01); Chopin (02); Del Moral et al. (06).
- The Algorithm:

0. Initialise $x_1^{(i)} \sim \Pi_1$ with weights $W_1^{(i)} = 1$. Set $n = 1$.
1. Given $(x_n^{(i)}, W_n^{(i)})$, move $x_{n+1}^{(i)} \sim K_{n+1}(x_n^{(i)}, dx)$.
2. Assign weights $W_{n+1}^{(i)} = W_n^{(i)} \cdot \frac{\Pi_{n+1}}{\Pi_n}(x_n^{(i)})$ to get $(x_{n+1}^{(i)}, W_{n+1}^{(i)}) \sim \Pi_{n+1}$.
3. Calculate Effective Sample Size:

$$ESS_{n+1} = \frac{(\sum_{i=1}^{N} W_{n+1}^{(i)})^2}{\sum_{i=1}^{N} (W_{n+1}^{(i)})^2} .$$

   If $\frac{ESS_{n+1}}{N} < \alpha \in (0, 1)$ then resample and set $W_{n+1}^{(i)} = 1$.
4. Set $n = n + 1$. Return to Step 1.

Adaptive SMC Samplers

- A critical property of SMC samplers is that they can use current particle information to tune kernels k_n 'on the fly'.
- **SMC Adaptation** (an example):

  Assume having $(x_n^{(i)}, W_n^{(i)}) \sim \Pi_n$, we can estimate:

$$\hat{\mu}_n = \frac{\sum_{i=1}^N W_n^{(i)} x_n^{(i)}}{\sum_{i=1}^N W_n^{(i)}} \ , \quad \hat{\Sigma}_n^2 = \frac{\sum_{i=1}^N W_n^{(i)} (x_n^{(i)} - \hat{\mu}_n)^2}{\sum_{i=1}^N W_n^{(i)}}$$

  and correspond K_n to a RWM kernel with proposal:

$$x_{n+1}^{(i),pr} = x_n^{(i)} + \ell \cdot N(0, \hat{\Sigma}_n^2)$$

Adaptation and Consistency

- Adaptive SMC is widely used in practical applications.
- Adaptation affects the consistency properties of MC estimates.
- We have found (Beskos et al. (14)) that, for many cases of practical interest:
  i) The effect of adaptation in the accuracy of MC estimates is small $\mathcal{O}(\frac{1}{N})$ compared to MC error $\mathcal{O}(\frac{1}{\sqrt{N}})$.
  ii) Asymptotic variances at the CLT for MC estimates using the adaptive kernels are the same as using the 'ideal' kernel.
- Estimates of normalising constants are not unbiased any more, thus adaptation cannot be used yet in recent popular 'pseudo-marginal' MCMC methods.

General Guidelines

- Ingredients for a potentially Stable SMC Sampler:
  - Successive Π_n should not be "too different", so that incremental weights $\frac{\Pi_{n+1}}{\Pi_n}(x_n^{(i)})$ are stable.
  - MCMC move steps should be "uniformly effective" over the sequence of targets.
- We have actually quantified these principles in a particular context (Beskos et al. (14)).

Example Static SMC

- We have i.i.d. target distribution:

$$\Pi(x_{1:d}) = \prod_{j=1}^{d} \pi(x_j)$$

and will use particles, N, from:

$$\Pi_1(x_{1:d}) = \{\Pi(x_{1:d})\}^{\phi_1}$$

for some small $\phi_1 > 0$.

- We would require $N = \mathcal{O}(\kappa^d)$, $\kappa > 1$, for direct Importance Sampling:

$$x^{(i)} \sim \Pi_1 \ , \ \ W^{(i)} = \frac{\Pi}{\Pi_1}(x^{(i)}) \ , \quad \{x^{(i)}, W^{(i)}\}_{i=1}^{N} \sim \Pi$$

Tempering

- We work with the sequence of distributions:

$$\Pi_n(x) \propto \{\Pi(x)\}^{\phi_n} \ ,$$

for inverse temperatures

$$\phi_1 < \phi_2 < \cdots < \phi_n < \cdots < \phi_p \equiv 1$$

- We require sequence of Markov transition kernels to propagate particles $\{K_n\}_{n=1}^{p}$ such that:

$$\Pi_n K_n = \Pi_n$$

Towards a Stable Algorithm

- We make the temperature selections:

$$p = d + 1\,, \qquad \phi_{n+1} - \phi_n = \tfrac{1-\phi_1}{d}$$

- We consider the simplified scenario:

$$K_n(x_{n-1}, dx_n) = \prod_{j=1}^{d} k_n(x_{n-1,j}, dx_{n,j})\ ; \quad \pi_n k_n = \pi_n$$

Conditions for Stability

(A1) i. Minorisation condition uniformly in ϕ:

There exists set C, constant $\theta \in (0,1)$ and probability law ν so that C is $(1,\theta,\nu)$-small w.r.t. k_ϕ.

ii. Geometric Ergodicity uniformly in ϕ:

$$k_\phi V(x) \leq \lambda V(x) + b\mathbb{I}_C(x)\,,$$

with $\lambda < 1$, $b > 0$ and C as above, for all $\phi \in [\phi_1, 1]$.

(A2) Controlled Perturbations of $\{k_\phi\}$:

$$||k_\phi - k_{\phi'}||_V \leq M\,|\phi - \phi'|\ .$$

Statement of One of Results

- **Theorem:** Under the conditions, we have that as $d \to \infty$:

$$\log W_{\phi}^{(i)} \Rightarrow B_{\sigma^2_{\phi_1:\phi}}$$

where B is a Brownian motion.

- The asymptotic variance is:

$$\sigma^2_{\phi_1:\phi} = (1-\phi_1)\int_{\phi_1}^{\phi} \pi_s\{\, \widehat{g}_s^2 - k_s(\widehat{g}_s^2)\,\}\, ds\ .$$

- $\log W_1^{(i)}$ stabilise as $d \to \infty$ for fixed N.

Comments

- Recall that:

$$\sigma^2_{\phi_1:\phi} = (1-\phi_1)\int_{\phi_1}^{\phi} \pi_s\{\,\widehat{g}_s^2 - k_s(\widehat{g}_s^2)\,\}\,ds\;.$$

- Here, $\widehat{g}_s$ is the solution to the Poisson equation:

$$g(x) - \pi_s(g) = \widehat{g}_s(x) - k_s(\widehat{g}_s)(x)$$

- Note also that:

$$\pi\{\,\widehat{g}^2 - k(\widehat{g}^2)\,\}$$

  is the asymptotic variance in the standard CLT for geometric MCMC Markov chains.

Takehome Conclusions

- Ingredients for a potentially Stable SMC Algorithm:
  - Enough bridging steps to stabilise incremental weights.
  - MCMC steps uniformly effective over the sequence of bridging densities.
- Adaptation will be critical in practical applications.

Outline

Navier Stokes Dynamics

- Consider NS dynamics on $[0, L] \times [0, L]$, describing the evolution of the velocity $u = u(x, t)$ of incompressible fluid:

$$\frac{\partial u}{\partial t} - \nu \Delta u + (u \cdot \nabla) u + \nabla p = f$$

$$\nabla \cdot u = 0 \ , \quad \int_{[0,L]^2} u_i(x) dx = 0 \ , \ \ i = 1, 2$$

$$u(x, 0) = u_0(x)$$

  with ν the viscosity, p the pressure, f the forcing.

- $\Delta = \partial^2_{x_1} + \partial^2_{x_2}$ is the Laplacian operator.
- We will assume periodic boundary conditions: $u_i(0, t) = u_i(L, t)$ for $i = 1, 2$.

Spectral Domain

- Natural basis here is $\{\psi_k\}_{k\in\mathbb{Z}^2/\{0\}}$ such that:

$$\psi_k(x) = \frac{k^\perp}{|k|}\,\exp\{\,i\,\frac{2\pi}{L}\,k\cdot x\}$$

where $k^\perp = (-k_2, k_1)'$.

- So that we can expand:

$$u(x) = \sum_{k\in\mathbb{Z}^2/\{0\}} u_k\,\psi_k(x)$$

for Fourier coefficients $u_k = \langle u, \psi_k\rangle$.

Example Dynamics

- Stationary regime, 2 videos:

$$L = 2\pi \ , \ \nu = \frac{1}{10} \ , \ f(x) = \nabla \cos\left((1,1)^{\perp} \cdot x\right)$$

- (Mildly) Chaotic regime, 2 videos:

$$L = 2\pi \ , \ \nu = \frac{1}{50} \ , \ f(x) = \nabla \cos\left((5,5)^{\perp} \cdot x\right)$$

Data Setting

- **Objective:** Learn about the initial condition u_0 of the PDE given available observations.
- We observe $u(x, t)$ with error:

$$y_{s,m} = u(x_m, s\,\delta) + N(0, \Sigma)$$

for indices $1 \leq s \leq T$, $1 \leq m \leq M$ and $\delta > 0$.

- We define the observation operator:

$$u_0 \mapsto G_{s,m}(u_0) = u(x_m, s\,\delta)$$

- This setting corresponds to *Eulerian* observations (there is also the *Lagrangian* set-up).

Prior Specification

- The parameter to be inferred (initial condition u_0) is in theory an infinite-dimensional object.
- Thus, a lot of care is need in terms of setting a prior, so that the posterior is well-posed.
- Following Stuart (10), we select a Gaussian prior:

$$\Pi_0 = N(0, \beta^2(-\Delta)^{-\alpha})$$

  for $\alpha > 1, \beta^2 > 0$.

- Such a choice allows a simple interpretation for the prior distribution of the Fourier coefficients:

$$Re(u_k), Im(u_k) \overset{i.i.d.}{\sim} N(0, \tfrac{1}{2}\beta^2(\tfrac{4\pi^2}{L^2}|k|^2)^{-\alpha})$$

Target Distribution

- We have the likelihood function (Y denotes all data):

$$L(Y \mid u_0) = e^{-\frac{1}{2}\sum_{s,m}|y_{s,m}-G_{s,m}(u_0)|_{\Sigma}^{2}}$$

- And the target posterior distribution:

$$\Pi(u_0|Y) \propto L(Y \mid u_0) \times \Pi_0(u_0)$$

- State space is Hilbert space $H = L^2([0,L]^2, \mathbb{R}^2)$.
- Target is in theory infinite-dimensional; in practice, a high-dimensional projection will be used.

Standard Approaches

- Kalman-Filter-type methods can give estimates of mean, uncertainty via linear approximation of PDE dynamics (Law & Stuart, 12)
- E.g. Ensemble KF (Evensen, 09).
- Such methods many times track well the mean but not the uncertainty.
- Efforts have recently been made to solve full Bayesian problem for non-linear dynamics.
- van Leeuwen (10), Law & Stuart (12)

Learning from Posterior

- Law & Stuart (12) propose a RWM-type MCMC algorithm.
- It proposes:

$$u_0^{pr} = \rho\, u_0 + \sqrt{1-\rho^2}\, Z$$

for noise $Z \sim \Pi_0$, accepted will probability:

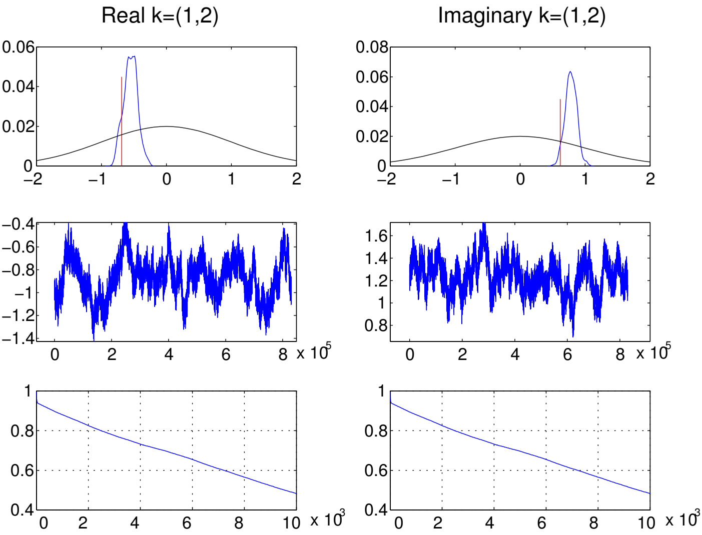
$$1 \wedge \frac{L(Y|u_0^{pr})}{L(Y|u_0)}$$

- This is relevant for off-line setup, and was used to check robustness of practical approximate algorithms.
- Algorithm needed $\rho \approx 1$ to give good acceptance probabilities, and could tackle some scenarios (state space made of 64^2 positions).

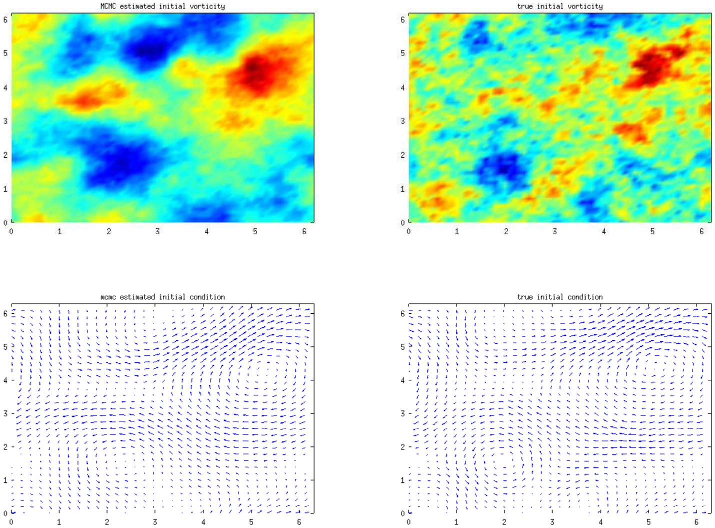
Example Application: Short-Time

- We considered the Chaotic Regime ($\nu = \frac{1}{50}$).
- Data: $M = 16, \;\; T = 5, \;\; \delta = 0.02, \;\; \Sigma = diag\{0.2, 0.2\}$.
- Prior: $\beta^2 = 5, \;\; \alpha = 2.2$.
- Kernel: $\rho = 0.9998, \;\; E[\,a\,] \approx 0.30$.
- True u_0: Sample from prior.
- Computational Time: 9 days ($dim = 64^2$, $dt = 0.002$)

MCMC Output


Real k=(1,2)
Imaginary k=(1,2)
0.06
0.04
0.02
0
–2
–1
0
1
2
0.08
0.06
0.04
0.02
0
–2
–1
0
1
2
–0.4
–0.6
–0.8
–1
–1.2
–1.4
0
2
4
6
8
x 10^5
1.6
1.4
1.2
1
0.8
0
2
4
6
8
x 10^5
1
0.8
0.6
0.4
0
2
4
6
8
10
x 10^3
1
0.8
0.6
0.4
0
2
4
6
8
10
x 10^3

MCMC Output


MCMC estimated initial vorticity
true initial vorticity
mcmc estimated initial condition
true initial condition
0
1
2
3
4
5
6

Mixing Issue for MCMC

- The proposal also writes as:

$$u^{pr}_{0,k} = \rho\, u_{0,k} + \sqrt{1-\rho^2}\, N(0, \tfrac{1}{2}\, (\tfrac{4\pi^2}{L^2}|k|^2)^{-\alpha})$$

- Scale of noise ideally tuned to the prior distribution, but badly tuned to the posterior.
- A-posteriori, low Fourier coeffs have much smaller variances and other means than a-priori, explaining $\rho \approx 1$.
- We would like to go on with as-global-as-possible MCMC steps, but greatly increase their efficiency.

SMC Samplers

- It seems sensible to apply an SMC sampler.
- We can build a bridging sequence of densities between prior Π_0 and posterior Π via Sequential Assimilation of data over observed locations and time instances.
- Assume data are ordered as y_n, for $0 \leq n \leq MT$.
- So we have the bridging densities:

$$\Pi_n = \Pi(\, u_0 \,|\, y_1, y_2, \ldots, y_n \,)\ , \quad 0 \leq n \leq MT$$

- We apply SMC sampler, starting from prior:

$$u_0^{(i)} \sim \Pi_0\ , \ldots ,\ (u_n^{(i)}, W_n^{(i)}) \sim \Pi_n\, , \ldots$$

Bridging Densities

- Are incremental weights stable?
- It turns out that some *tempering* might be needed.
- In-between Π_n and Π_{n+1} we introduce:

$$\Pi_{n,\phi} = \Pi_n \times \left(\frac{\Pi_{n+1}}{\Pi_n}\right)^{\phi}$$

- Incremental weights are equal to:

$$W_{n,\phi}^{(i)} = \left(\frac{\Pi_{n+1}}{\Pi_n}\right)^{\phi}(u_n^{(i)}) = (W_n^{(i)})^{\phi}$$

- **Adaptive Tempering:**

  Pick ϕ so that $ESS_{n,\phi} \approx N/3$, (Jasra et al., 11).

MCMC Kernel

- Are kernels K_n effective?
- Naive choice of K_{n+1} such that $\Pi_{n+1} K_{n+1} = \Pi_{n+1}$ would be to choose proposal:

$$u_{n+1,k}^{(i),pr} = \rho\, u_{n+1,k}^{(i)} + \sqrt{1-\rho^2}\, N(0, \tfrac{1}{2}\,(\tfrac{4\pi^2}{L^2}|k|^2)^{-\alpha})$$

- **SMC Adaptation:**

Assuming having $(u_n^{(i)}, W_n^{(i)}) \sim \Pi_n$, we estimate:

$$\hat{\mu}_k = \frac{\sum_{i=1}^N W_n^{(i)} u_{n,k}^{(i)}}{\sum_{i=1}^N W_n^{(i)}} \;, \quad \hat{\sigma}_\kappa^2 = \frac{\sum_{i=1}^N W_n^{(i)} (u_{n,k}^{(i)} - \hat{\mu}_k)^2}{\sum_{i=1}^N W_n^{(i)}}$$

and propose:

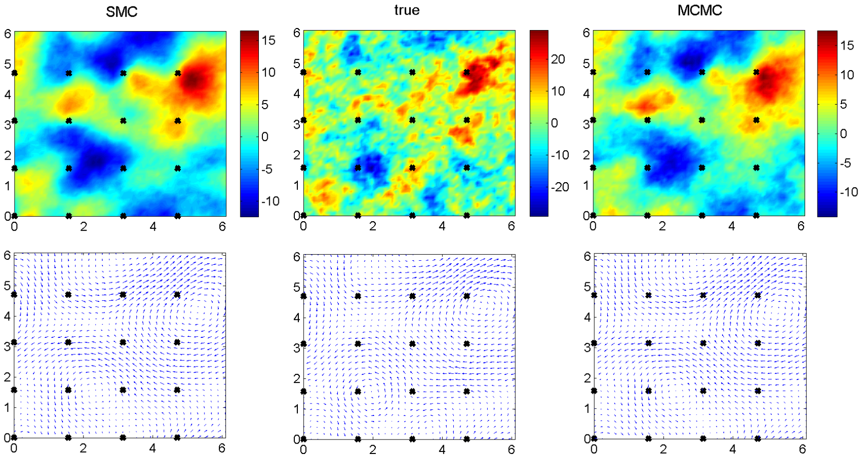
$$u_{n+1,k}^{(i),pr} = \hat{\mu}_k + \rho\,(\,u_{n,k}^{(i)} - \hat{\mu}_k\,) + \sqrt{1-\rho^2}\, N(0, \hat{\sigma}_k^2)$$

Example Application: Short Time

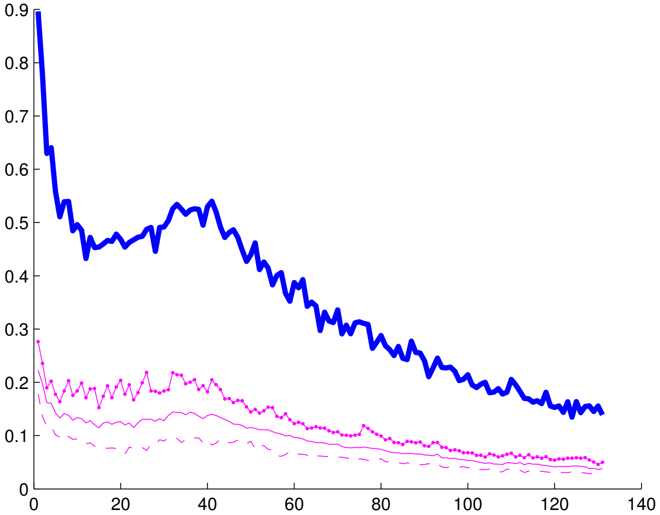
Algorithmic Specification:

- Distinguish between:
  - Low Frequencies: $|k_1| \leq 7, |k_2| \leq 7$, $dim_L \approx 200$.
  - High requencies: $dim_H = 64^2 - dim_L$.
- $\rho_H = 0.991$ for proposal tuned to prior in dim_H.
  $\rho_L = 0.99$ for advanced proposal in dim_L.
- Complete kernel synthesized 20 such steps.
- Used $N = 1,020$ particles.
- Computational time: 7.4h (used parallelisation).

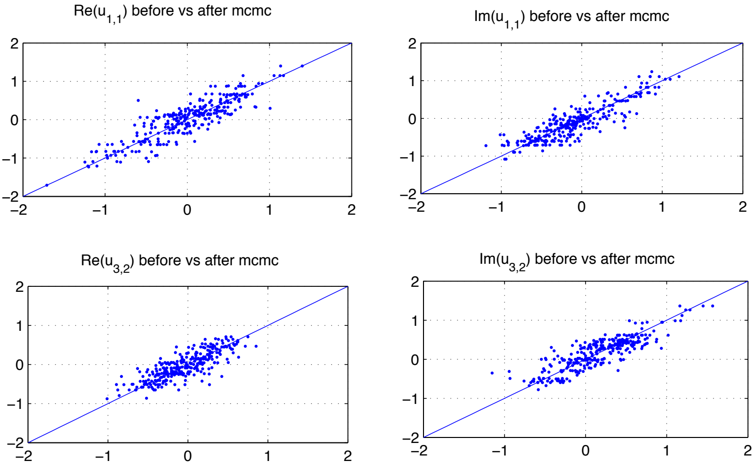
SMC Output



SMC Performance: Acceptance Probability



SMC Performance: Jittering



SMC Performance: Jittering

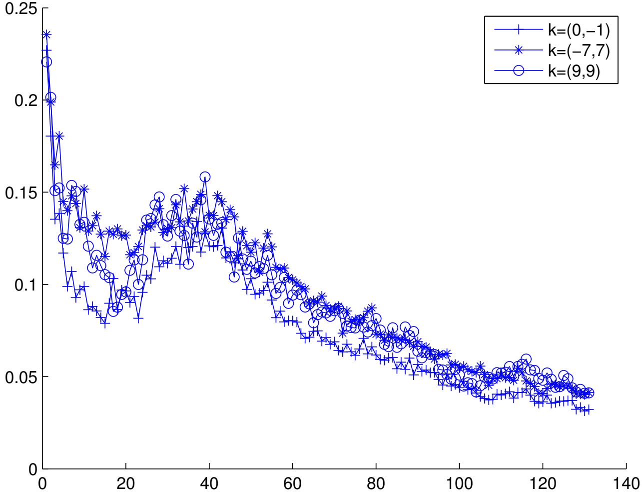
- We use one-dimensional summary:

$$J_k = \frac{\sum_{i=1}^{N} (u_k'^{(i)} - u_k^{(i)})^2}{2\sum_{i=1}^{N} |u_k^{(i)} - \overline{u}_k|^2} = 1 - \widehat{corr(u_k', u_k)}$$

  as a statistic to monitor amount of jittering for each frequency k.

- The closer J_k is to 1, the better.

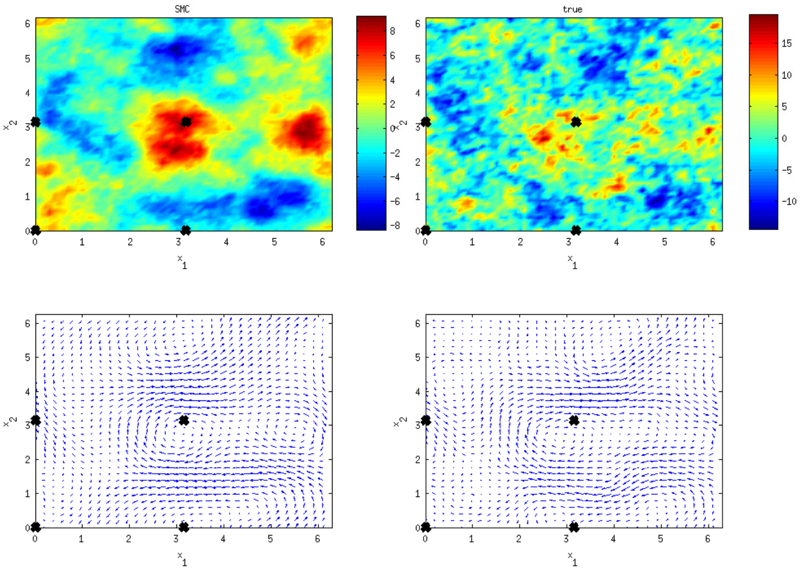
SMC Performance: Jittering



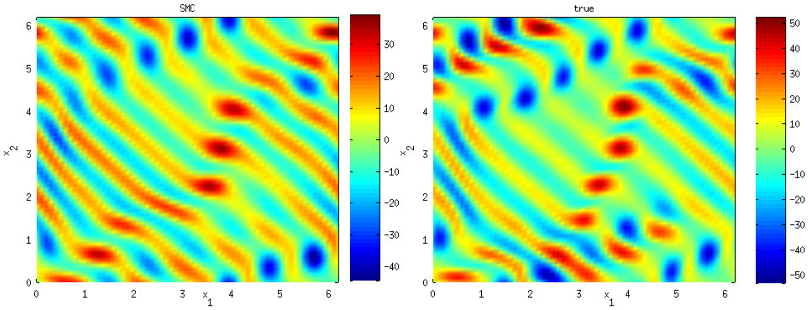
Example Application: Long-Time

- We considered the Chaotic Regime ($\nu = \frac{1}{50}$).
- Data: $M = 4, \;\; T = 20, \;\; \delta = 0.2, \;\; \Sigma = diag\{0.2, 0.2\}$.
- Prior: $\beta^2 = 1, \;\; \alpha = 2$.
- Computational Time: SMC 3.5 days, MCMC ∞.

SMC Output: Initial Field



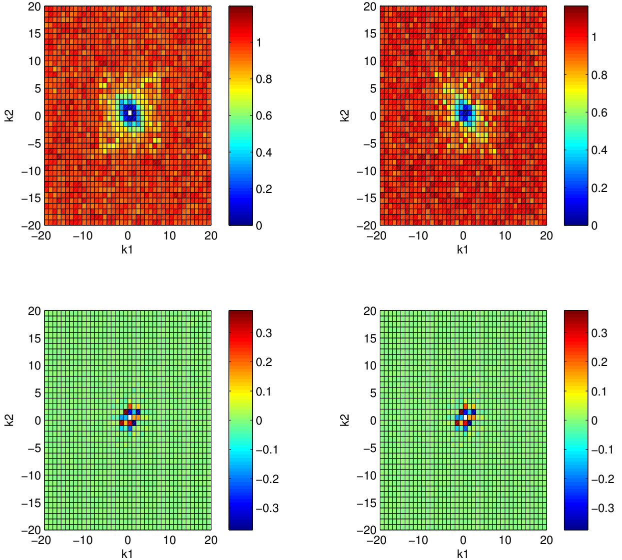
SMC Output: Final Field



SMC Output

- Video No1.
- Video No2.

SMC Output: Prior vs Posterior



Outline

Discussion

- PDE solver run over $\approx 10^4$ dimensions in our examples.
- To move to dimensions great than 10^4, some possible directions could be as follows.
  - Upgrade to Online Algorithm:
    Described algorithm for NS is of cost $\mathcal{O}(T^2)$ as at every calculation of a particle weight, or every MCMC step, PDE dynamics have to run from time $t = 0$ to current time.
  - Improve Development of MCMC steps.
- Shown MCMC has subsequently being greatly improved (Law and collaborators).

Discussion

- We have treated the case of deterministic signal.
- The case of stochastic signal (i.e. signal driven by SPDE) is also very important for applications (e.g. stochastic Navier-Stokes model).
- We are currently working on the development of an algorithm in this direction, which will be online, and (hopefully) could be the state-of-the-art for algorithms that try to tackle the full Bayesian problem for SPDE models.

- Thanks!