Data-Driven Nonparametric Likelihood Functions

John Harlim
Department of Mathematics
Department of Meteorology & Atmospheric Science
Institute of CyberScience.
The Pennsylvania State University

March 19, 2019

Plan of the talk:

- A quick review of kernel embedding of conditional distribution.
- A Bayesian Inference application: Parameter estimation.
- A data assimilation application: An online estimation of observation model error.

Kernel embedding of conditional distribution

Let $L^2(\mathcal{N}, \tilde{q})$ denotes an RKHS with kernel $\tilde{K} : \mathcal{N} \times \mathcal{N} \to \mathbb{R}$,

$$g(y) = \langle g, \tilde{K}(y, \cdot) \rangle_{\tilde{q}},$$

for all $g \in L^2(\mathcal{N}, \tilde{q})$ and $y \in \mathcal{N}$.

Kernel embedding of conditional distribution

Let $L^2(\mathcal{N}, \tilde{q})$ denotes an RKHS with kernel $\tilde{K} : \mathcal{N} \times \mathcal{N} \to \mathbb{R}$,

$$g(y) = \langle g, \tilde{K}(y, \cdot) \rangle_{\tilde{q}},$$

for all $g \in L^2(\mathcal{N}, \tilde{q})$ and $y \in \mathcal{N}$.

Let $P(Y|\Theta)$ be distribution of random variable Y defined on $\mathcal{N}$. The kernel embedding of conditional distribution $P(Y|\Theta)$ is defined as,

$$\mu_{Y|\theta} := \mathbb{E}_{Y|\theta}[\tilde{K}(Y, \cdot)] = \int_{\mathcal{N}} \tilde{K}(y, \cdot) dP(y|\theta).$$

Kernel embedding of conditional distribution

Let $L^2(\mathcal{N}, \tilde{q})$ denotes an RKHS with kernel $\tilde{K} : \mathcal{N} \times \mathcal{N} \to \mathbb{R}$,

$$g(y) = \langle g, \tilde{K}(y, \cdot) \rangle_{\tilde{q}},$$

for all $g \in L^2(\mathcal{N}, \tilde{q})$ and $y \in \mathcal{N}$.

Let $P(Y|\Theta)$ be distribution of random variable Y defined on $\mathcal{N}$. The kernel embedding of conditional distribution $P(Y|\Theta)$ is defined as,

$$\mu_{Y|\theta} := \mathbb{E}_{Y|\theta}[\tilde{K}(Y, \cdot)] = \int_{\mathcal{N}} \tilde{K}(y, \cdot) dP(y|\theta).$$

Given $g \in L^2(\mathcal{N}, \tilde{q})$,

$$\begin{aligned} \mathbb{E}_{Y|\theta}[g(Y)] &= \int_{\mathcal{N}} g(y) dP(y|\theta) = \int_{\mathcal{N}} \langle g, \tilde{K}(y, \cdot) \rangle_{\tilde{q}} dP(y|\theta) \\ &= \left\langle g, \int_{\mathcal{N}} \tilde{K}(y, \cdot) dP(y|\theta) \right\rangle_{\tilde{q}} = \langle g, \mu_{Y|\theta} \rangle_{\tilde{q}}. \end{aligned}$$

Kernel embedding of conditional distribution

One can verify[1] that

$$\mu_{Y|\theta} = q\mathcal{C}_{Y\Theta}\mathcal{C}_{\Theta\Theta}^{-1}K(\theta,\cdot),$$

where

$$\mathcal{C}_{\Theta Y} = \int_{\mathcal{M}\times\mathcal{N}} K(\theta,\cdot)\otimes\tilde{K}(y,\cdot)\,dP(\theta,y)$$

is the kernel embedding of $P(\Theta, Y)$ on appropriate Hilbert spaces and $K : \mathcal{M}\times\mathcal{M}\to\mathbb{R}$ is the kernel of RKHS $L^2(\mathcal{M}, q)$.

[1] Song et al. IEEE Signal Pross. Mag. 2013

Nonparametric likelihood functions

Let $\varphi_j(\theta) \in L^2(\mathcal{M}, q)$ and $\psi_k(y) \in L^2(\mathcal{N}, \tilde{q})$ be the, respective, orthonormal bases. It is clear that

$$K(\theta, \tilde{\theta}) = \sum_k \langle K(\theta, \cdot), \varphi_j \rangle \varphi_j(\tilde{\theta}) = \sum_k \varphi_j(\theta)\varphi_j(\tilde{\theta}).$$

Let

$$p(y|\theta) = \sum_k \mu_{Y|\theta,k} \psi_k(y)\tilde{q}(y)$$

$$\begin{aligned}
\mu_{Y|\theta,k} &= \langle p(\cdot|\theta), \psi_k \rangle = \mathbb{E}_{Y|\theta}[\psi_k] = \langle \mu_{Y|\theta}, \psi_k \rangle_{\tilde{q}} \\
&= \langle q\mathcal{C}_{Y\Theta}\mathcal{C}_{\Theta\Theta}^{-1} K(\theta, \cdot), \psi_k \rangle_{\tilde{q}} \\
&= \sum_j \varphi_j(\theta) \langle q\mathcal{C}_{Y\Theta}\mathcal{C}_{\Theta\Theta}^{-1} \varphi_j, \psi_k \rangle_{\tilde{q}} \\
&= \sum_j \varphi_j(\theta) \langle \mathcal{C}_{Y\Theta}\mathcal{C}_{\Theta\Theta}^{-1}, \varphi_j \otimes \psi_k \rangle_{q\otimes\tilde{q}}
\end{aligned}$$

Nonparametric likelihood functions

Let

$$\begin{aligned}\left[C_{Y\Theta}\right]_{jk} &= \langle \mathcal{C}_{Y\Theta}, \psi_j \otimes \varphi_k \rangle_{\tilde{q}\otimes q} \\ \left[C_{\Theta\Theta}\right]_{jk} &= \langle \mathcal{C}_{\Theta\Theta}, \varphi_j \varphi_k \rangle_q .\end{aligned}$$

Then one can show that,

$$\begin{aligned}[C_{Y\Theta} C_{\Theta\Theta}^{-1}]_{kj} &= \sum_\ell [C_{Y\Theta}]_{k\ell} [C_{\Theta\Theta}]^{-1}_{\ell j} \\ &= \sum_\ell \langle \mathcal{C}_{Y\Theta}, \psi_k \otimes \varphi_\ell \rangle_{\tilde{q}\otimes q} \langle \mathcal{C}_{\Theta\Theta}^{-1}, \varphi_\ell \varphi_j \rangle_q \\ &= \langle \mathcal{C}_{Y\Theta}, \psi_k \otimes (\sum_\ell \langle \mathcal{C}_{\Theta\Theta}^{-1}, \varphi_\ell \varphi_j \rangle_q \varphi_\ell) \rangle_{\tilde{q}\otimes q} \\ &= \langle \mathcal{C}_{Y\Theta}, \psi_k \otimes \mathcal{C}_{\Theta\Theta}^{-1} \varphi_j \rangle_{\tilde{q}\otimes q} \\ &= \langle \mathcal{C}_{Y\Theta} \mathcal{C}_{\Theta\Theta}^{-1}, \psi_k \otimes \varphi_j \rangle_{\tilde{q}\otimes q}\end{aligned}$$

Thus, the expansion coefficient is given as,

$$\mu_{Y|\theta,k} = \sum_j \varphi_j(\theta) [C_{Y\Theta} C_{\Theta\Theta}^{-1}]_{kj} .$$

Nonparametric likelihood functions

To summarize, let $\varphi_j(\theta) \in L^2(\mathcal{M}, q)$ and $\psi_k(y) \in L^2(\mathcal{N}, \tilde{q})$ be the, respective, orthonormal bases. Then,

$$p(y|\theta) = \sum_{k,j} \varphi_j(\theta)[C_{Y\Theta} C_{\Theta\Theta}^{-1}]_{kj} \psi_k(y) \tilde{q}(y)$$

where,

$$\begin{aligned}
\left[C_{Y\Theta}\right]_{jk} &= \langle \mathcal{C}_{Y\Theta}, \psi_j \otimes \varphi_k \rangle_{\tilde{q}\otimes q} = \mathbb{E}_{Y\Theta}[\psi_j \otimes \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \tilde{\varphi}_j(y_i) \varphi_k(\theta_i), \\
\left[C_{\Theta\Theta}\right]_{jk} &= \langle \mathcal{C}_{\Theta\Theta}, \varphi_j \varphi_k \rangle_{q} = \mathbb{E}_{\Theta\Theta}[\varphi_j \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \varphi_j(\theta_i) \varphi_k(\theta_i).
\end{aligned}$$

Remarks:

Nonparametric likelihood functions

To summarize, let $\varphi_j(\theta) \in L^2(\mathcal{M}, q)$ and $\psi_k(y) \in L^2(\mathcal{N}, \tilde{q})$ be the, respective, orthonormal bases. Then,

$$p(y|\theta) = \sum_{k,j} \varphi_j(\theta)[C_{Y\Theta} C_{\Theta\Theta}^{-1}]_{kj} \psi_k(y) \tilde{q}(y)$$

where,

$$\begin{aligned}
\left[C_{Y\Theta}\right]_{jk} &= \langle \mathcal{C}_{Y\Theta}, \psi_j \otimes \varphi_k \rangle_{\tilde{q}\otimes q} = \mathbb{E}_{Y\Theta}[\psi_j \otimes \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \tilde{\varphi}_j(y_i) \varphi_k(\theta_i), \\
\left[C_{\Theta\Theta}\right]_{jk} &= \langle \mathcal{C}_{\Theta\Theta}, \varphi_j \varphi_k \rangle_q = \mathbb{E}_{\Theta\Theta}[\varphi_j \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \varphi_j(\theta_i) \varphi_k(\theta_i).
\end{aligned}$$

Remarks:

- This is a linear regression in the coordinates of Hilbert spaces.

Nonparametric likelihood functions

To summarize, let $\varphi_j(\theta) \in L^2(\mathcal{M}, q)$ and $\psi_k(y) \in L^2(\mathcal{N}, \tilde{q})$ be the, respective, orthonormal bases. Then,

$$p(y|\theta) = \sum_{k,j} \varphi_j(\theta)[C_{Y\Theta} C_{\Theta\Theta}^{-1}]_{kj} \psi_k(y) \tilde{q}(y)$$

where,

$$\begin{aligned}
\left[C_{Y\Theta}\right]_{jk} &= \langle \mathcal{C}_{Y\Theta}, \psi_j \otimes \varphi_k \rangle_{\tilde{q}\otimes q} = \mathbb{E}_{Y\Theta}[\psi_j \otimes \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \tilde{\varphi}_j(y_i) \varphi_k(\theta_i), \\
\left[C_{\Theta\Theta}\right]_{jk} &= \langle \mathcal{C}_{\Theta\Theta}, \varphi_j \varphi_k \rangle_{q} = \mathbb{E}_{\Theta\Theta}[\varphi_j \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \varphi_j(\theta_i) \varphi_k(\theta_i).
\end{aligned}$$

Remarks:

- This is a linear regression in the coordinates of Hilbert spaces.
- If $\theta_i \sim q$, it is clear that, $C_{\Theta\Theta} = \mathcal{I}$. Otherwise, $C_{\Theta\Theta}$ can be singular.

Nonparametric likelihood functions

To summarize, let $\varphi_j(\theta) \in L^2(\mathcal{M}, q)$ and $\psi_k(y) \in L^2(\mathcal{N}, \tilde{q})$ be the, respective, orthonormal bases. Then,

$$p(y|\theta) = \sum_{k,j} \varphi_j(\theta)[C_{Y\Theta} C_{\Theta\Theta}^{-1}]_{kj} \psi_k(y) \tilde{q}(y)$$

where,

$$\begin{aligned} \left[C_{Y\Theta}\right]_{jk} &= \langle \mathcal{C}_{Y\Theta}, \psi_j \otimes \varphi_k \rangle_{\tilde{q} \otimes q} = \mathbb{E}_{Y\Theta}[\psi_j \otimes \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \tilde{\varphi}_j(y_i) \varphi_k(\theta_i), \\ \left[C_{\Theta\Theta}\right]_{jk} &= \langle \mathcal{C}_{\Theta\Theta}, \varphi_j \varphi_k \rangle_q = \mathbb{E}_{\Theta\Theta}[\varphi_j \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \varphi_j(\theta_i) \varphi_k(\theta_i). \end{aligned}$$

Remarks:

- This is a linear regression in the coordinates of Hilbert spaces.
- If $\theta_i \sim q$, it is clear that, $C_{\Theta\Theta} = \mathcal{I}$. Otherwise, $C_{\Theta\Theta}$ can be singular.
- Given $\theta_i \sim q$, DM is a natural tool that estimates $\varphi_k \in L^2(\mathcal{N}, q)$.

Parameter estimation problem

Consider

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= f(\mathbf{x}, \theta). \\ \mathbf{y}_i(\theta) &= g(\mathbf{x}(t_i; \theta), \boldsymbol{\eta}_\mathbf{i}). \end{aligned}$$

and our goal is to estimate $p(\theta|\mathbf{y}^\dagger)$, given $\mathbf{y}^\dagger = \{\mathbf{y}_1^\dagger, \ldots, \mathbf{y}_T^\dagger\}$ are observations under a specific parameter value $\theta^\dagger$.

Parameter estimation problem

Consider

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= f(\mathbf{x}, \theta). \\ \mathbf{y}_i(\theta) &= g(\mathbf{x}(t_i; \theta), \boldsymbol{\eta}_\mathbf{i}). \end{aligned}$$

and our goal is to estimate $p(\theta|\mathbf{y}^\dagger)$, given $\mathbf{y}^\dagger = \{\mathbf{y}_1^\dagger, \ldots, \mathbf{y}_T^\dagger\}$ are observations under a specific parameter value $\theta^\dagger$.

A popular Bayesian inference is to use MCMC to sample,

$$p(\theta|\mathbf{y}^\dagger) \propto p(\mathbf{y}^\dagger|\theta)p(\theta)$$

where $p(\mathbf{y}^\dagger|\theta)$ denotes the likelihood function corresponding to the observation model above.

Parameter estimation problem

- In most applications, the explicit expression for likelihood function is not available.

[2] Tavaré et al, Genetics 1997, Turner & Van Zandt, J. Math. Psychology, 2012.

Parameter estimation problem

- In most applications, the explicit expression for likelihood function is not available.
- When $g(\mathbf{x}(t_i;\theta), \boldsymbol{\eta}_\mathbf{i}) = h(\mathbf{x}(t_i;\theta)) +, \boldsymbol{\eta}_\mathbf{i}$, where $, \boldsymbol{\eta}_\mathbf{i} \sim P(\eta_i)$ are i.i.d. noises, one can approximate the likelihood function as,

$$p(\mathbf{y}^\dagger|\theta) = \prod_i P(\boldsymbol{\eta}_i) = \prod_i P(\mathbf{y}_i^\dagger - h(\mathbf{x}(t_i;\theta)).$$

[2] Tavaré et al, Genetics 1997, Turner & Van Zandt, J. Math. Psychology, 2012.

Parameter estimation problem

- In most applications, the explicit expression for likelihood function is not available.
- When $g(\mathbf{x}(t_i;\theta), \boldsymbol{\eta}_\mathbf{i}) = h(\mathbf{x}(t_i;\theta)) +, \boldsymbol{\eta}_\mathbf{i}$, where $, \boldsymbol{\eta}_\mathbf{i} \sim P(\eta_i)$ are i.i.d. noises, one can approximate the likelihood function as,

$$p(\mathbf{y}^\dagger|\theta) = \prod_i P(\boldsymbol{\eta}_i) = \prod_i P(\mathbf{y}_i^\dagger - h(\mathbf{x}(t_i;\theta)).$$

- In the case where the evaluation of $p(\mathbf{y}^\dagger|\theta)$ is computationally feasible, one can apply direct MCMC (if the likelihood or approximate likelihood is available).

[2] Tavaré et al, Genetics 1997, Turner & Van Zandt, J. Math. Psychology, 2012.

Parameter estimation problem

- In most applications, the explicit expression for likelihood function is not available.
- When $g(\mathbf{x}(t_i;\theta), \boldsymbol{\eta}_\mathbf{i}) = h(\mathbf{x}(t_i;\theta))+, \boldsymbol{\eta}_\mathbf{i}$, where $, \boldsymbol{\eta}_\mathbf{i} \sim P(\eta_i)$ are i.i.d. noises, one can approximate the likelihood function as,

$$p(\mathbf{y}^\dagger|\theta) = \prod_i P(\boldsymbol{\eta}_i) = \prod_i P(\mathbf{y}_i^\dagger - h(\mathbf{x}(t_i;\theta)).$$

- In the case where the evaluation of $p(\mathbf{y}^\dagger|\theta)$ is computationally feasible, one can apply direct MCMC (if the likelihood or approximate likelihood is available).
- If such likelihood approximation is not available but evaluation of $\mathbf{y}_i(\theta)$ is computationally feasible, one can use Approximate Bayesian Computation (ABC).[2]

[2]Tavaré et al, Genetics 1997, Turner & Van Zandt, J. Math. Psychology, 2012.

Parameter estimation problem

- If the evaluation of $\mathbf{y}_i(\theta)$ is intractable for sequential sampling, one can either: Improve the sampling methodology, e.g., Hamiltonian MC[3], DRAM[4].

[3] Neal et al. in Handbook of MCMC, 2011
[4] Haario et al, Stat. Comput. 2006
[5] Higdon et al., SIAM J. Sci. Comput. 2004.
[6] Marzouk and Xiu, Comm. Comput. Phys., 2009.
[7] Nagel and Sudret, J. Comput. Phys. 2016.

Parameter estimation problem

- If the evaluation of $\mathbf{y}_i(\theta)$ is intractable for sequential sampling, one can either: Improve the sampling methodology, e.g., Hamiltonian MC[3], DRAM[4].
- Or consider a surrogate modeling approach, e.g., Gaussian Process model[5], Polynomial chaos[6], spectral expansion[7] **Example:** The polynomial chaos is used to approximate $\mathbf{x}(t_i;\theta)$ in the parametric likelihood $p(\mathbf{y}^\dagger|\theta)$.

[3] Neal et al. in Handbook of MCMC, 2011

[4] Haario et al, Stat. Comput. 2006

[5] Higdon et al., SIAM J. Sci. Comput. 2004.

[6] Marzouk and Xiu, Comm. Comput. Phys., 2009.

[7] Nagel and Sudret, J. Comput. Phys. 2016.

Parameter estimation problem

- If the evaluation of $\mathbf{y}_i(\theta)$ is intractable for sequential sampling, one can either: Improve the sampling methodology, e.g., Hamiltonian MC[3], DRAM[4].
- Or consider a surrogate modeling approach, e.g., Gaussian Process model[5], Polynomial chaos[6], spectral expansion[7]
  Example: The polynomial chaos is used to approximate $\mathbf{x}(t_i;\theta)$ in the parametric likelihood $p(\mathbf{y}^\dagger|\theta)$.
- Our aim is to handle the situation where likelihood is intractable and the evaluation of $\mathbf{y}_i(\theta)$ is computationally expensive such that sequential sampling is not feasible.

---

[3] Neal et al. in Handbook of MCMC, 2011

[4] Haario et al, Stat. Comput. 2006

[5] Higdon et al., SIAM J. Sci. Comput. 2004.

[6] Marzouk and Xiu, Comm. Comput. Phys., 2009.

[7] Nagel and Sudret, J. Comput. Phys. 2016.

Metropolis-Hasting Scheme

At step i, suppose we are given sample at previous step, θ_i.

- Draw a proposal $\theta^* \sim q(\theta_{i-1}, \theta^*)$ where q denotes a transition kernel density.
- Compute an acceptance rate,

$$\alpha(\theta_{i-1}, \theta^*) = \frac{p(\theta^*|\mathbf{y}^\dagger)}{p(\theta_{i-1}|\mathbf{y}^\dagger)} = \frac{p(\mathbf{y}^\dagger|\theta^*)p(\theta^*)}{p(\mathbf{y}^\dagger|\theta_i)p(\theta_i)}$$

- Draw $z \sim U[0, 1]$ and let

$$\theta_i = \begin{cases} \theta^*, & \text{if } z < \alpha(\theta_{i-1}, \theta^*) \\ \theta_{i-1}, & \text{otherwise} \end{cases}$$

Application to the parameter estimation problem

Basically, our idea is to use a pair of training data set $\{\theta_j, \mathbf{y}_{i,j}\}_{i=1,\dots,N}^{j=1,\dots,M}$ to approximate the conditional density $p(\mathbf{y}|\theta)$.

Application to the parameter estimation problem

Basically, our idea is to use a pair of training data set $\{\theta_j, \mathbf{y}_{i,j}\}_{i=1,\ldots,N}^{j=1,\ldots,M}$ to approximate the conditional density $p(\mathbf{y}|\theta)$.

Let θ_j be uniformly distributed on a hyperrectangle $\mathcal{M}$. Then the cosine Fourier series $\varphi_l(\theta)$ form an orthonormal basis of $L^2(\mathcal{M})$.

Application to the parameter estimation problem

Basically, our idea is to use a pair of training data set $\{\theta_j, \mathbf{y}_{i,j}\}_{i=1,\dots,N}^{j=1,\dots,M}$ to approximate the conditional density $p(\mathbf{y}|\theta)$.

Let θ_j be uniformly distributed on a hyperrectangle $\mathcal{M}$. Then the cosine Fourier series $\varphi_l(\theta)$ form an orthonormal basis of $L^2(\mathcal{M})$.

Let $\mathbf{y}_j \in \mathcal{N} \subseteq \mathbb{R}^n$ distributed according to $q(\mathbf{y})$. Then applying the diffusion maps algorithm, we obtain $\psi_k(\mathbf{y}) \in L^2(\mathcal{N}, q)$. This choice of weights respect the geometry of the data and gives an improved error rate.

Application to the parameter estimation problem

Our nonparametric representation for $p(\mathbf{y}|\theta)$ [8] is given as,

$$
\begin{aligned}
\hat{p}(\mathbf{y}|\theta) &= \sum_{k=1}^{K_1} \sum_{\ell=1}^{K_2} (\mathbf{C}_{\mathbf{Y}\boldsymbol{\Theta}})_{kl} \varphi_l(\theta) \psi_k(\mathbf{y}) q(\mathbf{y}), \\
(\mathbf{C}_{\mathbf{Y}\boldsymbol{\Theta}})_{kl} &= \mathbb{E}_{\mathbf{Y}\boldsymbol{\Theta}}[\psi_k \varphi_l] \approx \frac{1}{MN} \sum_{j,i=1}^{M,N} \psi_k(\mathbf{y}_{i,j}) \varphi_l(\theta_j)
\end{aligned}
$$

[8] Jiang and H, arXiv:1804.03272

Application to the parameter estimation problem

Our nonparametric representation for $p(\mathbf{y}|\theta)$ [8] is given as,

$$
\begin{aligned}
\hat{p}(\mathbf{y}|\theta) &= \sum_{k=1}^{K_1}\sum_{\ell=1}^{K_2} (\mathbf{C}_{\mathbf{Y\Theta}})_{kl}\varphi_l(\theta)\psi_k(\mathbf{y})q(\mathbf{y}), \\
(\mathbf{C}_{\mathbf{Y\Theta}})_{kl} &= \mathbb{E}_{\mathbf{Y\Theta}}[\psi_k\varphi_l] \approx \frac{1}{MN}\sum_{j,i=1}^{M,N} \psi_k(\mathbf{y}_{i,j})\varphi_l(\theta_j)
\end{aligned}
$$

Error estimate: The first two moments of the error converges to 0 with convergence rates of order $M^{1/2}K_1^{1/2}N^{-1/2}$ and MK_1N^{-1}, respectively. **Independent to the variance of $\psi_k(\mathbf{Y})$.**

[8] Jiang and H, arXiv:1804.03272

Example 1: Fast-slow SDE on $\mathcal{N} = S^1 \times S^1 \subset \mathbb{R}^3$

Consider $(\theta, \phi) \in [0, 2\pi]^2$.

$$d(\theta, \phi) = a(\theta, \phi)\, dt + b(\theta, \phi, D)\, dW_t$$

where

$$b(\theta, \phi, D) = \begin{pmatrix} D + D\sin(\theta) & \frac{1}{4}\cos(\theta+\phi) \\ \frac{1}{4}\cos(\theta+\phi) & \frac{1}{40} + \frac{1}{40}\sin(\phi)\cos(\theta) \end{pmatrix}$$

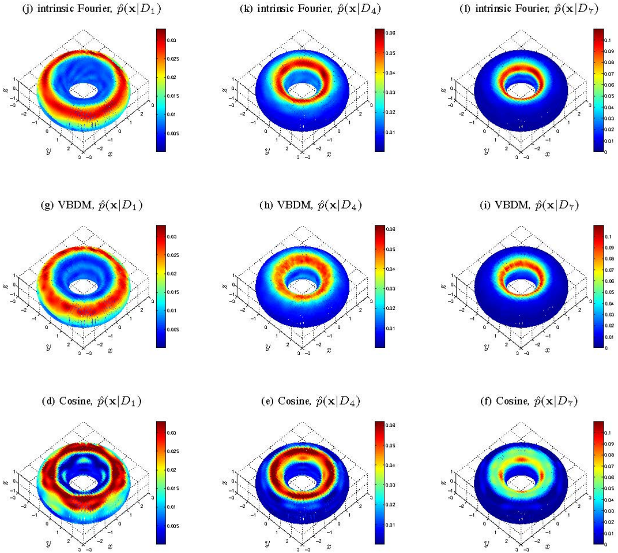
.

Let $\mathbf{x}_i \in \mathbb{R}^3$ be the observations defined via standard torus embedding. Our goal is to estimate $p(D|\mathbf{x}_{\mathbf{1}:\mathbf{T}})$, where $T = 10,000$.

We compare RKHS using:

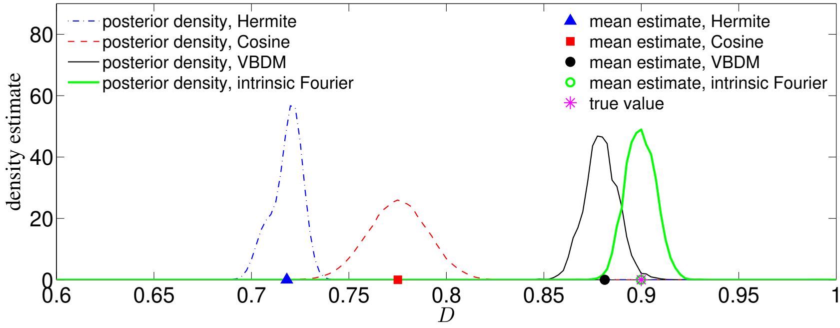
- Cosine (and Hermite) basis on $\mathbb{R}^3$.
- Variable Bandwidth Diffusion Maps (VBDM) basis obtained from ambient data.
- Fourier basis on intrinsic geometry.

Example 1: Likelihood function estimates



Example 1: Posterior density estimates

Posterior density estimates from MCMC.



Note: The likelihood function is trained on a wide range of parameter values $\{1/4, 2/4, 3/4, \ldots, 2\}$.

Example 2: Lorenz-96 model

Consider .

$$\begin{aligned} \frac{dx_j}{dt} &= x_{j-1}\left(x_{j+1} - x_{j-2}\right) - x_j + F, \quad j = 1, \ldots, 5, \\ y_j\left(t_m\right) &= x_j\left(t_m\right) + \epsilon_{m,j}, \quad \epsilon_{m,j} \sim \mathcal{N}\left(0, \sigma^2\right), \quad m = 1, \ldots, T, \end{aligned}$$

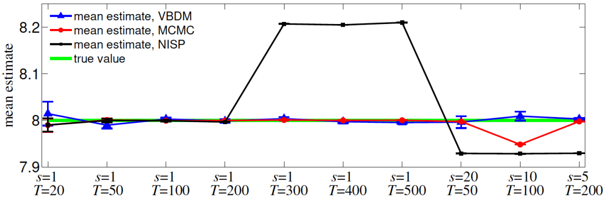
Here, the observation time $t_m = m(s\delta t)$, where δt denotes integration time step.

Our goal is to estimate $p(F|\mathbf{y}^\dagger)$ via MCMC. We compare:

- VBDM: $p(\mathbf{y}^\dagger|F) = \prod_i p(\mathbf{y}_i^\dagger|F)$, where $p(\mathbf{y}_i^\dagger|F)$ is estimated via RKHS.
- Direct Estimate: $p(\mathbf{y}^\dagger|F) = \exp\big(-\frac{1}{2\sigma^2}\sum_i(\mathbf{y}_i^\dagger - h(\mathbf{x}(t_i;F)))^2\big)$.
- NISP (Nonlinear Intrusive Spectral Projection) uses this Gaussian likelihood but approximate $\mathbf{x}$ with polynomial chaos expansion in F.

Example 2: Posterior mean estimates.

(a)



Remarks:

- Direct MCMC simulation involves 40,000 model evaluations.
- Both NISP and the nonparametric VBDM likelihood functions involve only 8 model evaluations in training phase.

Example 3: 40D Lorenz-96 model

Let $\{\hat{x}_k(t_m, F)\}_{-J/2+1,\ldots,J/2}$ be the kth discrete Fourier coefficient of $\{x_j\}_{j=1\ldots,J}$ at time t_m and parameter value F.

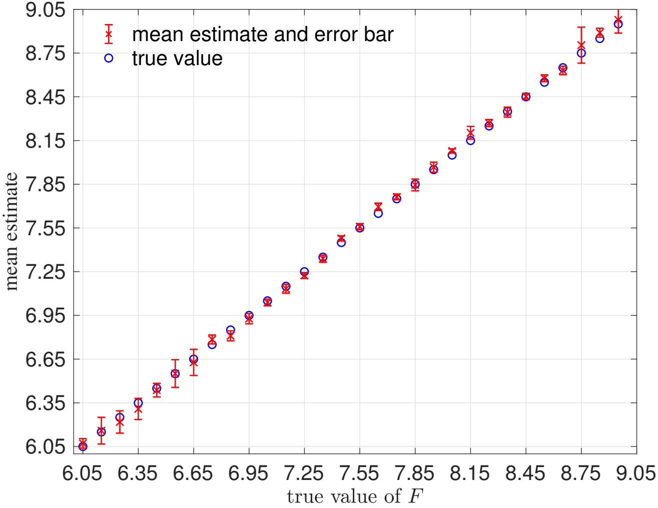
We consider Bayesian inference for estimating $P(F|\mathbf{y}_{0:T})$, where each component of $\mathbf{y}_m$ is an autocorrelation function of Fourier modes k_j:

$$\begin{aligned} y_{m,j}(F) &= \mathbb{E}[\hat{x}_{k_j}(t_m, F)\hat{x}_{k_j}(t_0, F)], \\ &\approx \frac{1}{L}\sum_{\ell=1}^{L} \hat{x}_{k_j}(t_\ell + m, F)\hat{x}_{k_j}(t_\ell, F) \end{aligned}$$

on energetic Fourier modes $k_j \in \{7, 8, 9, 14\}$ and $m = 0, \ldots, T$.

Remarks: We set $L = 10^6$ large enough to have small enough Monte-Carlo error. We set T to account for correlation up to model unit time 2.5.

Example 3: Estimates from chain of length 40,000.



Training parameter set: $\{6, 6.1, \ldots, 9\}$.

Verification parameter set: $\{6.05, 6.15, \ldots, 8.95\}$.

Biased observation model error problems in DA

The Kalman based DA formulation assumes unbiased observation model error, e.g.,

$$y_i = h(x_i) + \eta_i, \quad \eta_i \sim \mathcal{N}(0, R).$$

Suppose the operator h is un known. Instead, we are only given $\tilde{h}$, then

$$y_i = \tilde{h}(x_i) + b_i,$$

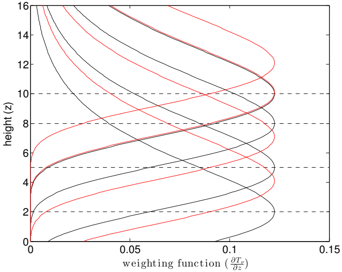
where we introduce a biased model error, $b_i = h(x_i) - \tilde{h}(x_i) + \eta_i$.

Example: Basic radiative transfer model

Consider solutions of the stochastic cloud model[9], $\{T(z), \theta_{eb}, q, f_d, f_s, f_c\}$. Based on this solutions, define a basic radiative transfer model as follows,

$$h_\nu(x) = \theta_{eb} T_\nu(0) + \int_0^\infty T(z) \frac{\partial T_\nu}{\partial z}(z)\, dz,$$

where T_ν is the transmission between heights z to ∞ that is defined to depend on q.
The weighting function, $\frac{\partial T_\nu}{\partial z}$ are defined as follows:



[9]Khouider, Biello, Majda 2010

Example: Basic radiative transfer model

Suppose the deep and stratiform cloud top height is $z_d = 12$km, while the cumulus cloud top height is $z_c = 3$km. Define $f = \{f_d, f_c, f_s\}$ and $x = \{T(z), \theta_{eb}, q\}$. Then the cloudy RTM is given by,

$$
\begin{aligned}
h_\nu(x, f) \quad &= \quad (1 - f_d - f_s)\left[\theta_{eb} T_\nu(0) + \int_0^{z_d} T(z) \frac{\partial T_\nu}{\partial z}(z)\, dz\right] \\
&\quad + (f_d + f_s) T(z_t) T_\nu(z_d) + \int_{z_d}^{\infty} T(z) \frac{\partial T_\nu}{\partial z}(z)\, dz
\end{aligned}
$$

Example: Basic radiative transfer model

Suppose the deep and stratiform cloud top height is $z_d = 12$km, while the cumulus cloud top height is $z_c = 3$km. Define $f = \{f_d, f_c, f_s\}$ and $x = \{T(z), \theta_{eb}, q\}$. Then the cloudy RTM is given by,

$$
\begin{aligned}
h_\nu(x, f) &= (1 - f_d - f_s)\Big[\theta_{eb} T_\nu(0) + \int_0^{z_d} T(z) \frac{\partial T_\nu}{\partial z}(z)\, dz\Big] \\
&\quad + (f_d + f_s) T(z_t) T_\nu(z_d) + \int_{z_d}^{\infty} T(z) \frac{\partial T_\nu}{\partial z}(z)\, dz \\
&= (1 - f_d - f_s)\Big[(1 - f_c)\big(\theta_{eb} T_\nu(0) + \int_0^{z_c} T(z) \frac{\partial T_\nu}{\partial z}(z)\, dz\big) \\
&\quad + f_c T(z_c) T_\nu(z_c) + \int_{z_c}^{z_d} T(z) \frac{\partial T_\nu}{\partial z}(z)\, dz\Big] \\
&\quad + (f_d + f_s) T(z_d) T_\nu(z_t) + \int_{z_d}^{\infty} T(z) \frac{\partial T_\nu}{\partial z}(z)\, dz
\end{aligned}
$$

Example: Basic radiative transfer model

Suppose the deep and stratiform cloud top height is $z_d = 12$km, while the cumulus cloud top height is $z_c = 3$km. Define $f = \{f_d, f_c, f_s\}$ and $x = \{T(z), \theta_{eb}, q\}$. Then the cloudy RTM is given by,

$$
\begin{aligned}
h_\nu(x,f) &= (1-f_d-f_s)\Big[\theta_{eb}T_\nu(0) + \int_0^{z_d} T(z)\frac{\partial T_\nu}{\partial z}(z)\,dz\Big] \\
&\quad +(f_d+f_s)T(z_t)T_\nu(z_d) + \int_{z_d}^{\infty} T(z)\frac{\partial T_\nu}{\partial z}(z)\,dz \\
&= (1-f_d-f_s)\Big[(1-f_c)\Big(\theta_{eb}T_\nu(0) + \int_0^{z_c} T(z)\frac{\partial T_\nu}{\partial z}(z)\,dz\Big) \\
&\quad +f_c T(z_c)T_\nu(z_c) + \int_{z_c}^{z_d} T(z)\frac{\partial T_\nu}{\partial z}(z)\,dz\Big] \\
&\quad +(f_d+f_s)T(z_d)T_\nu(z_t) + \int_{z_d}^{\infty} T(z)\frac{\partial T_\nu}{\partial z}(z)\,dz
\end{aligned}
$$

One can check that $h_\nu(x,0)$ corresponds to cloud-free RTM.

Observation model error in data assimilation

Suppose the observation is generated with

$$y_\nu = h_\nu(x, f) + \eta, \qquad \eta \sim \mathcal{N}(0, R)$$

The difficulty in estimating the cloud fractions, cloud top heights and (in reality we don't know precisely how many clouds under a column) induces model error.

Observation model error in data assimilation

Suppose the observation is generated with

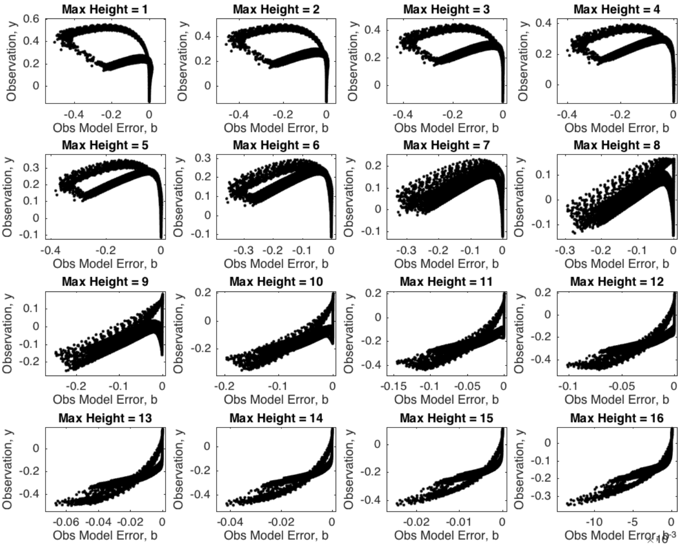
$$y_\nu = h_\nu(x, f) + \eta, \qquad \eta \sim \mathcal{N}(0, R)$$

The difficulty in estimating the cloud fractions, cloud top heights and (in reality we don't know precisely how many clouds under a column) induces model error.
In an extreme case, we consider filtering with a cloud-free RTM:

$$y_\nu = h_\nu(x, 0) + b_\nu$$

where $b_\nu = h_\nu(x, f) - h_\nu(x, 0) + \eta$ is model error with bias.

Observations (y_ν) v Model error (b_ν)



State estimation of the model error

We propose a secondary filter to estimate the statistics for b_i as follows:

$$
\begin{tikzcd}
\begin{matrix}\text{Prior}\\ p(x_i)\end{matrix} \arrow[r, "\text{Primary Filter}"] \arrow[d] & \begin{matrix}\text{Posterior}\\ p(x_i \mid y_i)\end{matrix} \\
\begin{matrix}\text{Error Prior}\\ p(b)\end{matrix} \arrow[r, "\text{Secondary Filter}"] & \begin{matrix}\text{Error Posterior}\\ p(b \mid y_i)\end{matrix} \arrow[u] \\
\begin{matrix}\text{Observation}\\ y_i\end{matrix} \arrow[u] \arrow[r, "\text{RKHS+Training Data}"] & \begin{matrix}\text{Likelihood}\\ p(y_i \mid b)\end{matrix} \arrow[u]
\end{tikzcd}
$$

We employ the RKHS theory to train a nonparametric likelihood function $p(y_i|b)$[10].

[10] Berry and H, Mon. Wea. Rev. 2017.

Recall that $b_i = h(x_i) - \tilde{h}(x_i) + \eta_i = y_i - \tilde{x}_i$.

Given the prior ensemble $\{x_i^{b,k}\}_{k=1\dots,K}$:

Recall that $b_i = h(x_i) - \tilde{h}(x_i) + \eta_i = y_i - \tilde{x}_i$.

Given the prior ensemble $\{x_i^{b,k}\}_{k=1\ldots,K}$:

- Compute $\bar{y}^b = \frac{1}{K}\sum_{k=1}^{K} \tilde{h}(x_i^{b,k})$.

Recall that $b_i = h(x_i) - \tilde{h}(x_i) + \eta_i = y_i - \tilde{x}_i$.

Given the prior ensemble $\{x_i^{b,k}\}_{k=1\ldots,K}$:

- Compute $\bar{y}^b = \frac{1}{K}\sum_{k=1}^{K} \tilde{h}(x_i^{b,k})$.
- Define $Y_i = [\tilde{h}(x_i^{b,K}), \ldots, \tilde{h}(x_i^{b,K})]$ and $P_{yy,i} = \frac{1}{K-1} Y_i Y_i^\top$.

Recall that $b_i = h(x_i) - \tilde{h}(x_i) + \eta_i = y_i - \tilde{x}_i$.

Given the prior ensemble $\{x_i^{b,k}\}_{k=1\dots,K}$:

- Compute $\bar{y}^b = \frac{1}{K}\sum_{k=1}^{K} \tilde{h}(x_i^{b,k})$.
- Define $Y_i = [\tilde{h}(x_i^{b,K}), \dots, \tilde{h}(x_i^{b,K})]$ and $P_{yy,i} = \frac{1}{K-1} Y_i Y_i^\top$.
- Assume that

$$p(b) = \mathcal{N}(y_i - \bar{y}_i^b, P_{yy,i} + R)$$

Recall that $b_i = h(x_i) - \tilde{h}(x_i) + \eta_i = y_i - \tilde{x}_i$.

Given the prior ensemble $\{x_i^{b,k}\}_{k=1\ldots,K}$:

- Compute $\bar{y}^b = \frac{1}{K}\sum_{k=1}^{K} \tilde{h}(x_i^{b,k})$.
- Define $Y_i = [\tilde{h}(x_i^{b,K}), \ldots, \tilde{h}(x_i^{b,K})]$ and $P_{yy,i} = \frac{1}{K-1} Y_i Y_i^\top$.
- Assume that

$$p(b) = \mathcal{N}(y_i - \bar{y}_i^b, P_{yy,i} + R)$$

- Then apply the secondary Bayesian on training data set b_ℓ with nonparametric likelihood from RKHS.

$$p(b|y_i) \propto p(b)p(y_i|b)$$

Recall that $b_i = h(x_i) - \tilde{h}(x_i) + \eta_i = y_i - \tilde{x}_i$.

Given the prior ensemble $\{x_i^{b,k}\}_{k=1\ldots,K}$:

- Compute $\bar{y}^b = \frac{1}{K}\sum_{k=1}^{K} \tilde{h}(x_i^{b,k})$.
- Define $Y_i = [\tilde{h}(x_i^{b,K}), \ldots, \tilde{h}(x_i^{b,K})]$ and $P_{yy,i} = \frac{1}{K-1} Y_i Y_i^\top$.
- Assume that

$$p(b) = \mathcal{N}(y_i - \bar{y}_i^b, P_{yy,i} + R)$$

- Then apply the secondary Bayesian on training data set b_ℓ with nonparametric likelihood from RKHS.

$$p(b|y_i) \propto p(b)p(y_i|b)$$

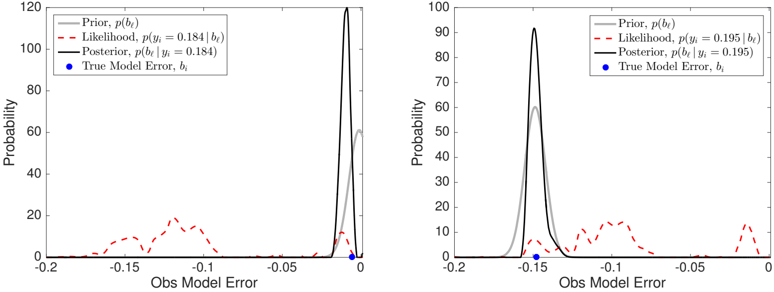
- Compute the mean and variance of the observation model error,

$$\begin{aligned} \hat{\mu}_{b_i} &= \mathbb{E}[b|y_i] \\ \hat{R}_{b_i} &= Var[b|y_i] \end{aligned}$$

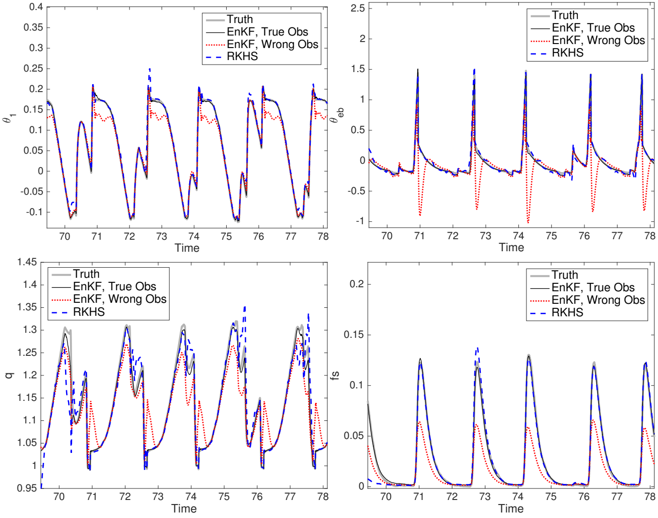
and use these terms to compensate for bias and variance of the observation model error in the primary filter.

Secondary Bayesian filter

$$p(b|y_i) \propto p(b)p(y_i|b)$$



Filter estimates (with adaptive tuning of R and Q).



Discussion

For junior participants: How to extend $p(y|\theta)$ on observations $y^\dagger$ that do not belong to the training data set?

General Research problems:

For Parameter Estimation Problem:

- How to pick data set for training of the likelihood function? This issue will be problematic for high-dimensional parameter space.
- Identifiability of high-dimensional parameter space.

For DA Problem:

- How to use this RKHS supervised learning algorithm in other DA context?

References:

- S. Jiang and J. Harlim, *Parameter Estimation with Data-Driven Nonparametric Likelihood Functions*, arXiv:1804.03272.
- T. Berry and J. Harlim, *Correcting biased observation model error in data assimilation*, Mon Wea. Rev. 145(7), 2833-2853, 2017.

Collaborators:

- Shixiao Jiang, Postdoc at Department of Mathematics, The Pennsylvania State University.
- Tyrus Berry, Assistant Professor at Department of Mathematical Sciences, George Mason University.