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 $y \in \mathcal{N}$. 

Given $g \in L^2(\mathcal{N}, \tilde{q})$, $E_{Y|\Theta}[g(Y)] = \int_{\mathcal{N}} g(y) dP_{Y|\Theta}(y) = \int_{\mathcal{N}} \langle g, \tilde{K}(y, \cdot) \rangle_{\tilde{q}} dP_{Y|\Theta}(y) = \langle g, \mu_{Y|\Theta} \rangle_{\tilde{q}}$. 
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 $g \in L^2(\mathcal{N}, \tilde{q})$,

$$\mathbb{E}_{Y|\theta}[g(Y)] = \int_{\mathcal{N}} g(y) dP(y|\theta) = \int_{\mathcal{N}} \langle g, \tilde{K}(y, \cdot) \rangle_{\tilde{q}} dP(y|\theta) = \langle g, \int_{\mathcal{N}} \tilde{K}(y, \cdot) dP(y|\theta) \rangle_{\tilde{q}} = \langle g, \mu_{Y|\theta} \rangle_{\tilde{q}}.$$
Kernel embedding of conditional distribution

One can verify\(^1\) that

\[
\mu_{Y|\theta} = qC_{Y \Theta} C_{\Theta \Theta}^{-1} K(\theta, \cdot),
\]

where

\[
C_{\Theta Y} = \int_{\mathcal{M} \times \mathcal{N}} K(\theta, \cdot) \otimes \tilde{K}(y, \cdot) dP(\theta, y)
\]

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

\(^1\)Song et al. IEEE Signal Pross. Mag. 2013
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

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

$$\mu_{Y|\theta, k} = \langle p(\cdot|\theta), \psi_k \rangle = \mathbb{E}_{Y|\theta}[\psi_k] = \langle \mu_{Y|\theta}, \psi_k \rangle \tilde{q}$$

$$= \langle q C_{Y \Theta} C_{\Theta \Theta}^{-1} K(\theta, \cdot), \psi_k \rangle \tilde{q}$$

$$= \sum_j \varphi_j(\theta) \langle q C_{Y \Theta} C_{\Theta \Theta}^{-1} \varphi_j, \psi_k \rangle \tilde{q}$$

$$= \sum_j \varphi_j(\theta) \langle C_{Y \Theta} C_{\Theta \Theta}^{-1}, \varphi_j \otimes \psi_k \rangle \tilde{q} \otimes \tilde{q}$$
Nonparametric likelihood functions

Let

\[
\begin{align*}
[C_Y \Theta]_{jk} &= \langle C_Y \Theta, \psi_j \otimes \varphi_k \rangle \tilde{q} \otimes q \\
[C_\Theta \Theta]_{jk} &= \langle C_\Theta \Theta, \varphi_j \varphi_k \rangle_q.
\end{align*}
\]

Then one can show that,

\[
[C_Y \Theta C_\Theta^{-1}]_{kj} = \sum_\ell [C_Y \Theta]_{k\ell} [C_\Theta \Theta]_{\ell j}^{-1}
\]

\[
= \sum_\ell \langle C_Y \Theta, \psi_k \otimes \varphi_\ell \rangle \tilde{q} \otimes q \langle C_\Theta^{-1}_\Theta, \varphi_\ell \varphi_j \rangle_q
\]

\[
= \langle C_Y \Theta, \psi_k \otimes (\sum_\ell \langle C_\Theta^{-1}_\Theta, \varphi_\ell \varphi_j \rangle_q \varphi_\ell) \rangle \tilde{q} \otimes q
\]

\[
= \langle C_Y \Theta \Theta^{-1}, \psi_k \otimes \varphi_j \rangle \tilde{q} \otimes q
\]

Thus, the expansion coefficient is given as,

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

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

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

where,

$$[C_{\Theta}]_{jk} = \langle C_{\Theta}, \psi_j \otimes \varphi_k \rangle_{\tilde{q} \otimes q} = E_{\Theta}[\psi_j \otimes \varphi_k] \approx \frac{1}{N} \sum_{i=1}^N \varphi_j(y_i)\varphi_k(\theta_i),$$

$$[C_{\Theta}]_{jk} = \langle C_{\Theta}, \varphi_j \varphi_k \rangle_q = E_{\Theta}[\varphi_j \varphi_k] \approx \frac{1}{N} \sum_{i=1}^N \varphi_j(\theta_i)\varphi_k(\theta_i).$$

Remarks:

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

$$[C_{Y\Theta}]_{jk} = \langle C_{Y\Theta}, \psi_j \otimes \varphi_k \rangle_{\tilde{q} \otimes q} = \mathbb{E}_{Y\Theta}[\psi_j \otimes \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \psi_j(y_i)\varphi_k(\theta_i),$$

$$[C_{\Theta\Theta}]_{jk} = \langle C_{\Theta\Theta}, \varphi_j \varphi_k \rangle_q = \mathbb{E}_{\Theta\Theta}[\varphi_j \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \varphi_j(\theta_i)\varphi_k(\theta_i).$$

Remarks:

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

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

where,

$$[C_{Y\Theta}]_{jk} = \langle C_{Y\Theta}, \psi_j \otimes \varphi_k \rangle_{\tilde{q} \otimes q} = \mathbb{E}_{Y\Theta}[\psi_j \otimes \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \varphi_j(y_i) \varphi_k(\theta_i),$$

$$[C_{\Theta\Theta}]_{jk} = \langle C_{\Theta\Theta}, \varphi_j \varphi_k \rangle_q = \mathbb{E}_{\Theta\Theta}[\varphi_j \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \varphi_j(\theta_i) \varphi_k(\theta_i).$$

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,

$$[C_{Y\Theta}]_{jk} = \langle C_{Y\Theta}, \psi_j \otimes \varphi_k \rangle \tilde{q} \otimes q = \mathbb{E}_{\Theta}[\psi_j \otimes \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \tilde{\varphi}_j(y_i)\varphi_k(\theta_i),$$

$$[C_{\Theta\Theta}]_{jk} = \langle C_{\Theta\Theta}, \varphi_j \varphi_k \rangle_q = \mathbb{E}_{\Theta}[\varphi_j \varphi_k] \approx \frac{1}{N} \sum_{i=1}^{N} \varphi_j(\theta_i)\varphi_k(\theta_i).$$

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 $\varphi_k \in L^2(\mathcal{N}, q)$. 
Consider

\[
\frac{dx}{dt} = f(x, \theta).
\]

\[
y_i(\theta) = g(x(t_i; \theta), \eta_i).
\]

and our goal is to estimate \( p(\theta|y^\dagger) \), given \( y^\dagger = \{y_1^\dagger, \ldots, y_T^\dagger\} \) are observations under a specific parameter value \( \theta^\dagger \).
Consider

\[
\frac{dx}{dt} = f(x, \theta).
\]

\[
y_i(\theta) = g(x(t_i; \theta), \eta_i).
\]

and our goal is to estimate \( p(\theta|y^\dagger) \), given \( y^\dagger = \{y_1^\dagger, \ldots, y_T^\dagger\} \) are observations under a specific parameter value \( \theta^\dagger \).

A popular Bayesian inference is to use MCMC to sample,

\[
p(\theta|y^\dagger) \propto p(y^\dagger|\theta)p(\theta)
\]

where \( p(y^\dagger|\theta) \) 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.

\[ g(x(t_i; \theta), \eta_i) = h(x(t_i; \theta)) + \eta_i, \] where \( \eta_i \sim P(\eta_i) \) are i.i.d. noises, one can approximate the likelihood function as,

\[ p(y^*|\theta) = \prod_i P(\eta_i) = \prod_i P(y^*_i - h(x(t_i; \theta))). \]

- In the case where the evaluation of \( p(y^*|\theta) \) 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 \( y_i(\theta) \) is computationally feasible, one can use Approximate Bayesian Computation (ABC).

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

When \( g(\mathbf{x}(t_i; \theta), \eta_i) = h(\mathbf{x}(t_i; \theta)) + \eta_i \), where \( \eta_i \sim P(\eta_i) \) are i.i.d. noises, one can approximate the likelihood function as,

\[
p(\mathbf{y}^\dagger | \theta) = \prod_i P(\eta_i) = \prod_i P(y_i^\dagger - h(\mathbf{x}(t_i; \theta))).
\]
Parameter estimation problem

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

▶ When \( g(x(t_i; \theta), \eta_i) = h(x(t_i; \theta)) + \eta_i \), where \( \eta_i \sim P(\eta_i) \) are i.i.d. noises, one can approximate the likelihood function as,

\[
p(y^\dagger | \theta) = \prod_i P(\eta_i) = \prod_i P(y_i^\dagger - h(x(t_i; \theta))).
\]

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

---

Parameter estimation problem

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

- When $g(x(t_i; \theta), \eta_i) = h(x(t_i; \theta)) + \eta_i$, where $\eta_i \sim P(\eta_i)$ are i.i.d. noises, one can approximate the likelihood function as,

$$p(y^\dagger | \theta) = \prod_i P(\eta_i) = \prod_i P(y_i^\dagger - h(x(t_i; \theta))).$$

- In the case where the evaluation of $p(y^\dagger | \theta)$ 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 $y_i(\theta)$ is computationally feasible, one can use Approximate Bayesian Computation (ABC).\(^2\)

---

Parameter estimation problem

- If the evaluation of $y_i(\theta)$ 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
Parameter estimation problem

- If the evaluation of $y_i(\theta)$ 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 $x(t_i; \theta)$ in the parametric likelihood $p(y^\dagger|\theta)$.

---

$^3$Neal et al. in Handbook of MCMC, 2011


Parameter estimation problem

- If the evaluation of $y_i(\theta)$ 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 $x(t_i; \theta)$ in the parametric likelihood $p(y^\dagger|\theta)$.

- Our aim is to handle the situation where likelihood is intractable and the evaluation of $y_i(\theta)$ is computationally expensive such that sequential sampling is not feasible.

---

$^3$Neal et al. in Handbook of MCMC, 2011
At step $i$, suppose we are given sample at previous step, $\theta_i$.

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

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

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

$$\theta_i = \begin{cases} 
    \theta^*, & \text{if } z < \alpha(\theta_{i-1}, \theta^*) \\
    \theta_{i-1}, & \text{otherwise}
\end{cases}$$
Basically, our idea is to use a pair of training data set \( \{ \theta_j, y_{i,j} \}_{j=1,\ldots,M}^{i=1,\ldots,N} \) to approximate the conditional density \( p(y|\theta) \).
Application to the parameter estimation problem

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

Let \( \theta_j \) 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, y_{i,j}\}_{i=1,\ldots,N}^{j=1,\ldots,M} to approximate the conditional density \(p(y|\theta)\).

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

Let \(y_j \in \mathcal{N} \subseteq \mathbb{R}^n\) distributed according to \(q(y)\). Then applying the diffusion maps algorithm, we obtain \(\psi_k(y) \in L^2(\mathcal{N}, q)\). 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 $p(y|\theta)^8$ is given as,

$$\hat{p}(y|\theta) = \sum_{k=1}^{K_1} \sum_{\ell=1}^{K_2} (C_{Y\Theta})_{kl} \varphi_{l}(\theta) \psi_{k}(y) q(y),$$

$$(C_{Y\Theta})_{kl} = \mathbb{E}_{Y\Theta}[\psi_{k} \varphi_{l}] \approx \frac{1}{MN} \sum_{j,i=1}^{M,N} \psi_{k}(y_{i,j}) \varphi_{l}(\theta_{j}).$$

---

$^8$Jiang and H, arXiv:1804.03272
Our nonparametric representation for $p(y|\theta)$ \(^8\) is given as,

$$
\hat{p}(y|\theta) = \sum_{k=1}^{K_1} \sum_{\ell=1}^{K_2} (C_{Y\Theta})_{k\ell} \varphi_\ell(\theta) \psi_k(y) q(y),
$$

$$
(C_{Y\Theta})_{k\ell} = \mathbb{E}_{Y\Theta}[\psi_k \varphi_\ell] \approx \frac{1}{MN} \sum_{j,i=1}^{M,N} \psi_k(y_{i,j}) \varphi_\ell(\theta_j)
$$

**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** $\psi_k(Y)$.  

\(^8\) 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 $x_i \in \mathbb{R}^3$ be the observations defined via standard torus embedding. Our goal is to estimate $p(D|x_{1:T})$, 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 .

\[
\frac{dx_j}{dt} = x_{j-1}(x_{j+1} - x_{j-2}) - x_j + F, \quad j = 1, \ldots, 5, \\
y_j(t_m) = x_j(t_m) + \epsilon_{m,j}, \quad \epsilon_{m,j} \sim \mathcal{N}(0, \sigma^2), \quad m = 1, \ldots, T,
\]

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

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

- **VBDM:** \( p(y^\dagger|F) = \prod_i p(y_i^\dagger|F) \), where \( p(y_i^\dagger|F) \) is estimated via RKHS.

- **Direct Estimate:**
  \[
  p(y^\dagger|F) = \exp \left( - \frac{1}{2\sigma^2} \sum_i (y_i^\dagger - h(x(t_i; F)))^2 \right).
  \]

- **NISP (Nonlinear Intrusive Spectral Projection)** uses this Gaussian likelihood but approximate \( x \) with polynomial chaos expansion in \( F \).
Example 2: Posterior mean estimates.

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 \( k \)th discrete Fourier coefficient of \( \{ x_j \}_{j=1, \ldots, J} \) at time \( t_m \) and parameter value \( F \).

We consider Bayesian inference for estimating \( P(F|y_0:T) \), where each component of \( y_m \) is an autocorrelation function of Fourier modes \( k_j \):

\[
y_{m,j}(F) = \mathbb{E}[\hat{x}_{k_j}(t_m, F)\hat{x}_{k_j}(t_0, F)],
\]

\[
\approx \frac{1}{L} \sum_{\ell=1}^{L} \hat{x}_{k_j}(t_\ell + m, F)\hat{x}_{k_j}(t_\ell, F)
\]

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\}.
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 unknown. 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 \(T_\nu\) is the transmission between heights \(z\) to \(\infty\) 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 $z_d = 12\text{km}$, while the cumulus cloud top height is $z_c = 3\text{km}$. Define $f = \{f_d, f_c, f_s\}$ and $x = \{T(z), \theta_{eb}, q\}$. Then the cloudy RTM is given by,

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

One can check that $h_\nu(x, f)$ corresponds to cloud-free RTM.
Suppose the deep and stratiform cloud top height is $z_d = 12\text{km}$, while the cumulus cloud top height is $z_c = 3\text{km}$. Define $f = \{f_d, f_c, f_s\}$ and $x = \{T(z), \theta_{eb}, q\}$. Then the cloudy RTM is given by,

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

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

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

One can check that $h_{\nu}(x, 0)$ corresponds to cloud-free RTM.
Suppose the observation is generated with

\[ y_\nu = h_\nu(x, f) + \eta, \quad \eta \sim \mathcal{N}(0, R) \]

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.
Suppose the observation is generated with

\[ y_\nu = h_\nu(x, f) + \eta, \quad \eta \sim \mathcal{N}(0, R) \]

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 \((y_\nu)\) v Model error \((b_\nu)\)
State estimation of the model error

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

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

---

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

Given the prior ensemble \( \{ x_i^{b,k} \}_{k=1,...,K} \)
Recall that \( b_i = h(x_i) - \tilde{h}(x_i) + \eta_i = y_i - \tilde{x}_i \).

Given the prior ensemble \( \{ x_{i}^{b,k} \}_{k=1,...,K} \):

- 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 $\{x^{b,k}_i\}_{k=1,...,K}$:

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

Assume that $p(b) = N(y_i - \bar{y}^b_i, P_{yy,i} + R_b)$.

Then apply the secondary Bayesian on training data set $b_\ell$ with nonparametric likelihood from RKHS.

Compute the mean and variance of the observation model error, $\hat{\mu}_b^i = E[b_i | y_i]$ and $\hat{\sigma}_b^i = Var[b_i | y_i]$ and use these terms to compensate for bias and variance of the observation model error in the primary filter.
Recall that \( b_i = h(x_i) - \tilde{h}(x_i) + \eta_i = y_i - \tilde{\eta}_i \).

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

- 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 \( p(b) = \mathcal{N}(y_i - \bar{y}_i^{b}, P_{yy,i} + R) \).
Recall that $b_i = h(x_i) - \tilde{h}(x_i) + \eta_i = y_i - \tilde{x}_i$.

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

- 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

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

- Then apply the secondary Bayesian on training data set $b_\ell$ 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 $\{x_i^{b,k}\}_{k=1}^{K}$:

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

- Then apply the secondary Bayesian on training data set $b_\ell$ 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,
  $$\hat{\mu}_{b_i} = \mathbb{E}[b|y_i]$$
  $$\hat{R}_{b_i} = \text{Var}[b|y_i]$$

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 $Q$).
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:


Collaborators:

- Shixiao Jiang, Postdoc at Department of Mathematics, The Pennsylvania State University.

- Tyrus Berry, Assistant Professor at Department of Mathematical Sciences, George Mason University.