Kernel Density Estimates in Marine Ecosystem Modeling and Parameter Optimization

Thomas Slawig, Joscha Reimer, Maria-Theresia Verwega – CAU Kiel
Markus Schartau – GEOMAR Helmholtz Centre for Ocean Research Kiel

CSE Workshop Plön March 2019
The carbon cycle: major driving force of climate
... and for global warming

IPCC AR5 The Physical Science Basis Ch. 12, FAQ 12.3, Figure 1
Reduction of model-data misfit

- Least-squares problem with (discretized) “model”:

\[
\min_{(y,u)} \frac{1}{2} \|y - y_{data}\|_Y^2 + \frac{1}{2} \|u - u_{guess}\|_U^2 \quad \text{s.t.} \quad e(y, u) = 0.
\]

- Classical approach of PDE-constrained optimal control
- Corresponds to maximum likelihood estimator

\[
\max_{(y,u)} \text{Prob}(u = u_{guess} \mid y = y_{data})
\]

- Assuming Gaussian error distribution with covariances \(\Sigma_y, \Sigma_u\):

\[
\sim \max_{(y,u)} \exp \left( -\frac{1}{2} (y - y_{data})^\top \Sigma_y^{-1} (y - y_{data}) - \frac{1}{2} (u - u_{guess})^\top \Sigma_u^{-1} (u - u_{guess}) \right)
\]

- Norms are then \(\|y\|_Y^2 := y^\top \Sigma_y^{-1} y\), similar for \(U\).
Getting covariance information for data is difficult

- Sparse data:

  # of $PO_4$ measurements in upper ocean layer per model grid-box
Considering biomes

- Biomes \( \equiv \) regions with “similar behavior”
- based on expert knowledge from ocean science

17 core biomes according to Fay, McKinley (2014, Earth Syst. Sci. Data)
Mean value is not enough

RMSE (Root Mean Square Error), splitted into mean and deviation:

\[
E^2 := \frac{1}{N} \| y - z \|_2^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - z_i)^2
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} (\bar{y} - \bar{z} + y'_i - z'_i)^2
\]

\[
= (\bar{y} - \bar{z})^2 + 2(\bar{y} - \bar{z}) \left( \overbrace{y'_i - z'_i}^{=0} \right) + \sum_{i=1}^{N} (y'_i - z'_i)^2
\]

\[
= (\bar{y} - \bar{z})^2 + \frac{1}{N} \sum_{i=1}^{N} (y'_i - z'_i)^2
\]

\[
= \bar{E}^2 \text{ (bias)} + E'^2 \text{ (pattern mean square error)}.
\]
Why choice of metric matters in climate science ...

- no bias within an ocean region between data and model results
- variability is the same, but ...
- nonzero pattern mean square error (representing uncertainty)
Constructing sub-samples of data in each biome

Two sub-samples ○ and ●
Constructing sub-samples of data in each biome

- “Decorrelation radius” = 400 km
- i.e., at neighbouring grid points subsamples are mutually disjoint in time (months)
- subsamples within depth ranges are random without replacement for same months.

Depth ranges:

1. 0 – 250 m (monthly mean concentrations)
2. 250 – 500 m (monthly mean concentrations)
3. 500 – 1000 m (annual mean concentrations)
4. 1000 – 2000 m (annual mean concentrations)
5. 2000 – 5000 m (annual mean concentrations)
Nonparametric statistics: Motivation

- Gaussian distribution is determined by two parameters (mean and variance) only
- Multi-modal distributions difficult or only inaccurately to represent
Nonparametric statistics: Kernel Density Estimators (KDE)

- Replace (here: Gaussian) probability density function (PDF)

\[ x \mapsto \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp \left( -\frac{1}{2} (x - X)^\top \Sigma^{-1} (x - X) \right) \]

for data \( X \in \mathbb{R}^n \) ...

- ... by Kernel density estimator:

\[ \hat{f}(x, h) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right), \quad x \in \mathbb{R}. \]

- data \( X_i \in \mathbb{R}, i = 1, \ldots, N \), treated as random variables
- assumed to be independent with common PDF
- Kernel function \( K : \mathbb{R} \to \mathbb{R}_{\geq 0} \)
- Kernel bandwidth \( h > 0 \).
“Standard”: Gaussian KDE

- $X_i \in \mathbb{R}, i = 1, \ldots, N$: independent random variables with common PDF $f$ (represents the data).
- Gaussian KDE:

$$\hat{f}_N(x, t) = \frac{1}{N} \sum_{i=1}^{N} \Phi(x, X_i, t)$$

with Gauss kernels

$$\Phi(x, X_i, t) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-X_i)^2}{2t}}$$

and bandwidth ($\approx$ standard deviation) $\sqrt{t}$.

- Gaussian KDE $\hat{f}_N$ is a PDF for all $t \in \mathbb{R}_{>0}$.
- Disadvantage: support is whole real axis
- $\rightsquigarrow$ difficulties for data in bounded intervals.
KDE as sum of individual kernels

shape strongly determined by bandwidth parameter $h \in \mathbb{R}_{>0}$
Gauss KDE as solution of diffusion equation

- **Gauss kernel**

\[
\Phi(x, X_i, t) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-X_i)^2}{2t}}
\]

solves 1-dim. diffusion equation

\[
y_t = \frac{1}{2} y_{xx}, \quad \text{in } \mathbb{R} \times \mathbb{R}_{>0}
\]

\[
y(x, 0) = \delta(x - X_i), \quad x \in \mathbb{R}.
\]

- **Gaussian KDE**

\[
\hat{f}_N(x, t) = \frac{1}{N} \sum_{i=1}^{N} \Phi(x, X_i, t)
\]

solves same linear equation with sum of deltas as initial value:

\[
y(x, 0) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - X_i), \quad x \in \mathbb{R}.
\]

- **bandwidth** $\sqrt{t}$ as tuning parameter
Data bounded $\implies$ Gaussian KDE on bounded domain $(0, 1)$

- KDE

$$
\hat{f}_N(x, t) = \frac{1}{N} \sum_{i=1}^{N} \kappa(x, X_i, t)
$$

with kernels

$$
\kappa(x, X_i, t) = \sum_{k=-\infty}^{\infty} \left( \Phi(x, 2k + X_i, t) + \Phi(x, 2k - X_i, t) \right)
$$

Gauss kernels as above

- solves same PDE in $(0, 1) \times \mathbb{R}_{>0}$ with same initial values and homogeneous Neumann conditions

$$
y_x(0, t) = y_x(1, t), \quad t \in \mathbb{R}_{>0}.
$$

- $\hat{f}_N$ is consistent at the boundaries, i.e., here for $x = 0$:

$$
\mathbb{E} \hat{f}_N(x_N, t_N) = f(x_N) + \mathcal{O}(\sqrt{t_N}), \quad N \to \infty, \; t_N \downarrow 0, \; x_N = \alpha t_N, \; \alpha \in [0, 1]
$$

- In the limit $t \to \infty$:

$$
\kappa(\cdot, X_i, t), \hat{f}_N(\cdot, t) \sim \chi_{(0,1)}, \text{ uniform distribution.}
$$
Model-Data Comparison $\delta^{13}C_{DIC}$ Gaussian KDE

Gaussian KDE of $\delta^{13}C_{DIC}$ in the whole Ocean, 1980s - 2000s

- Blue line: Model Output
- Orange line: Field Data
- Dashed line: Full Model Output

Error = 1.792591
2D KDE of model results

Distribution of $\delta^{13}$C_{POC} and $\delta^{13}$C_{DIC} model output in the euphotic zone, 1990s

2D Gaussian KDE of $\delta^{13}$C_{POC} and $\delta^{13}$C_{DIC} model output in the euphotic zone, 1990s

Density
Metrics for non-parametric statistics

- **Hellinger distance:**
  Given KDEs (= PDFs) $f_{dat}$, $f_{mod}$ for data and model.
  \[
  d_H(f_{dat}, f_{mod}) = \sqrt{\frac{1}{2} \int_{\Omega} \left( \sqrt{f_{dat}(x)} - \sqrt{f_{mod}(x)} \right)^2 dx}
  \]

- **Integrated quadratic difference:**
  Given corresponding CDFs $F_{dat}$, $F_{mod}$:
  \[
  d_{IQ}(F_{dat}, F_{mod}) = \int_{\Omega} (F_{dat}(x) - F_{mod}(x))^2 dx
  \]
  with
  \[
  F_*(x) := \int_{s \leq x} f_*(s) ds.
  \]
Example Hellinger distance

WG4.1: A model calibration framework exemplified for the ocean biogeochemical component
Comparison of nonparametric probability densities

Hellinger Distance ($d_{HD}$) as a measure for the difference between two probability densities.
Example Integrated Square Distance

Integrated Quadratic Distance ($d_{IQD}$) as measure for the difference between probability distributions (cumulative probability densities)

Comparison of nonparametric probability densities

Variable (x, e.g. nitrate concentration in mmol N m$^{-3}$)
Example for model-data comparison: biome

OXYGEN: NP_SPSS (Biome #2); depth range: 0 m - 120 m

Model results - histogram (HST1, 80 bins)
WOA13 data - histogram
Model results - kernel density estimate (KDE)
WOA13 data - KDE

0.284 = HD-HST1 (80 bins)
0.273 = HD-HST2 (40 bins)
0.273 = HD-KDE
2.937 = IQD-HST1
2.98 = IQD-HST2
2.901 = IQD-KDE
Generalization: Diffusion KDE

- KDE $g_N$ as solution to generalized diffusion equation

\[
y_t = Ly := \frac{1}{2} \nabla \cdot \left( a \nabla \left( \frac{y}{p} \right) \right) \quad \text{in } \Omega \times \mathbb{R}_{>0}
\]

\[
\nabla \left( \frac{y}{p} \right) \cdot \eta = 0 \quad \text{on } \partial \Omega \times \mathbb{R}_{>0}
\]

\[
y(x, 0) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - X_i), \quad x \in \Omega
\]

on a bounded domain $\Omega$.

- Coefficient functions $a, p \in C^2(\overline{\Omega}, \mathbb{R}_{>0})$
- large bandwidth KDE ($t \to \infty$): $g_N(\cdot, t) \to p$ in $\Omega$
- small bandwidth KDE close to $f$
- Choice of $a, p$ allows for different smoothing in interval $\Omega$: greater/smaller smoothing in areas with sparse/dense data
- determined by diffusion $\sqrt{a(x)/p(x)}$
Errors in KDE

▶ $f$ PDF, $\hat{f}$ estimator of $f$

▶ Mean Integrated Squared Error:

$$\text{MISE}(\hat{f}, h) := \mathbb{E} \left( \int_{\mathbb{R}} (\hat{f}(x, h) - f(x))^2 \, dx \right)$$

▶ Asymptotic Mean Squared (Integrated) Error $z := \text{AMISE}(\hat{f})$:

$$z : \mathbb{R} \to \mathbb{R}_{>0} : \lim_{h \to h_0} \frac{\text{MISE}(\hat{f}, h)}{z(h)} = 1, \quad h_0 \in \mathbb{R} \cup \{\infty\}.$$ 

▶ Diffusion KDE [Botev et al. 2010]:

$$\text{AMISE}(g, t_N) = \frac{t_N^2}{4} \| Lf \|^2_{L^2(\Omega)} + \frac{\mathbb{E}(\sqrt{p/a})}{2N \sqrt{\pi t_N}}, \quad t_N \downarrow 0$$
Optimal bandwidth

- Optimal bandwidth [Botev et al. 2010]:

\[
t^*_N = \arg\min_{t_N \in \mathbb{R}_>0} \text{AMISE}(g_N, t_N) = \left( \frac{\mathbb{E}(1/\sigma)}{2N\sqrt{\pi} \| Lf \|_{L^2(\Omega)}^2} \right)^{\frac{2}{5}}
\]

- needs estimate for for (unknown) data PDF \( f \).
Summary and Outlook

- KDEs as non-parametric statistical approach
- for data representation
- and model-data comparison
- model assessment and improvement
- Gaussian KDE standard
- Different bandwidth selection methods
- Cost: Hellinger (PDF) or Integrated Quadratic Error (CDF)
- Generalization: KDE based on diffusion equation

Our future work:
- Diffusion KDEs for optimization
- Project starts in MARData Helmholtz graduate school