Density identification of distributions on bounded domain

By

Indrajith Wasala Mudiyanselage

11/29/2018

Outline

- Introduction
- Thesis focus
- Parametric density estimation
- Methodology
    - Estimating the derivatives
    - Method of polynomial approximation
    - Mode based density identification
- Brief introduction to Ecology
- Applications of the methods for ecological data sets
- Conclusion

Density estimation

- In density estimation we are interested in determining an unknown function *f*, given only random samples or observations distributed according to this function.
- The goal of density estimation is to infer the probability density function (PDF), from observed data of a random variable.
- Construction of the underlying unknown PDF is called probability density estimation.

There are two approaches to density estimation

Parametric Methods

- We assume that the data are drawn from a known parametric family of distributions.
- If either a model selection process justifies it, or it can be logically assumed that the observed data follows a well-known distribution, then parametric density estimation is possibly the simplest method for density approximation.

Non-Parametric Methods

- In nonparametric density estimation, we do not restrict the form of the density function with any parametric assumptions.
- We only assume that the density exists and is sufficiently smooth.
- The histogram is probably the oldest and simplest non-parametric density estimator.

Challenges in density estimation

- Existing techniques are computationally expensive for large datasets.
- Selection of an appropriate parametric family or a kernel function can be difficult.
- The underlying distribution may have an unknown truncation (ex. ecological data).
- Estimations near the boundaries of a distribution with a bounded domain can be difficult (ex. percentage data).
- Some methods only provide heuristic approximations without theoretical guarantees.

Focus of the thesis

The goal of the thesis is to construct some computationally efficient parametric density estimation methods for distributions with bounded domains. We focus on:

- Eliminating the need for model selection.
- Obtaining an estimation method for data with unknown bounded domain.
- Obtaining a method for estimating boundary derivatives when the distribution has known bounded domain.
- Avoiding heuristic approaches.

Parametric density estimation

- In parametric density estimation methods, we assume that we know the shape of the distribution, but not the parameters.
- We estimate the parameters using a method such as maximum likelihood estimation.

We define the likelihood function with parameter θ, $L_n(\mathbf{x}, \theta)$, as the joint density function of $X_1, X_2, \ldots, X_n$. Since $X_1, X_2, \ldots, X_n$ are iid, we have $L_n(\mathbf{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$.

Then the corresponding parameter estimation problem can be formulated as

$$\begin{aligned} \text{maximize} \quad & \prod_{i=1}^{n} f(X_i, \theta), \\ \text{subject to} \quad & \theta \in S. \end{aligned}$$

Uniform law of large numbers

If the data are i.i.d.,Θ is compact, $a(z_i, \theta)$ is continuous at each $\theta \in \Theta$ with probability one, and there is $d(z)$ with $\parallel a(z, \theta) \parallel \leq d(z)$ for all $\theta \in \Theta$ and $E[d(z)] < \infty$, then $E[a(z, \theta)]$ is continuous and $\sup_{\theta \in \Theta} \parallel n^{-1} \sum_{i=1}^{n} a(z_i, \theta) - E[a(z, \theta)] \parallel \xrightarrow{\text{P}} 0.$

Since $L_n(\boldsymbol{x}, \theta) \geq 0$ and the logarithmic function is monotone increasing on $(0, \infty)$, maximizing the likelihood function is equivalent to maximizing the average log likelihood function.

$$l_n(\mathbf{x}, \theta) = \frac{1}{n} \log(L_n(\mathbf{x}, \theta)) = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i, \theta).$$

Let $\hat{\theta}$ be the maximizer of $l_n(\boldsymbol{x}, \theta)$.

Let $L(\theta)$ be the expectation of $\log f(X,\theta)$ with respect to unknown parameter θ_0

$$L(\theta) = E_{\theta_0}[\log f(X,\theta)] = \int \log f(x,\theta) f(x,\theta_0) dx.$$

- θ_0 is the maximizer of $L(\theta)$ (We can show $L(\theta) \leq L(\theta_0)$).
- By uniform law of large numbers $l_n(\theta) \to L(\theta)$ almost surely for all θ.
- $\hat{\theta}$ is the maximizer of $l_n(\theta)$.

Based on above $\hat{\theta} \to \theta_0$ as $n \to \infty$.

Example

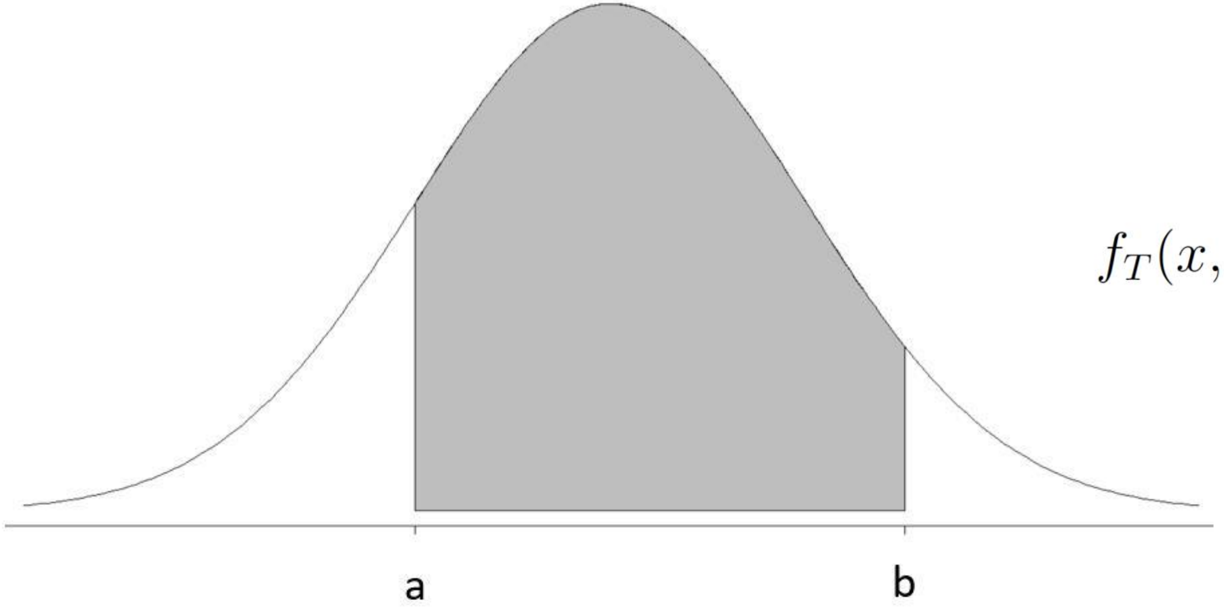
Let $X_1, X_2, \dots, X_n$ be a independent, identically distributed (i.i.d.), random variables such that $X_i \sim Normal(\mu, \sigma^2)$ for all i. By the uniform law of large numbers, the average log likelihood estimator

$$l_n(X, \mu, \sigma^2) = \frac{1}{n} \sum_{i=1}^{n} \left(-\frac{1}{2} \ln 2\pi\sigma^2 - \frac{(x_i - \mu)^2}{2\sigma^2} \right) = -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

converges almost surely to $-\dfrac{1}{2} \ln(2\pi\sigma^2) - \dfrac{1}{2}$. We have $\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{n} \to \sigma^2$ as desired.

Truncated distribution

A truncated distribution $f_T(x, \theta)$ is a conditional distribution that results from restricting the domain of some other probability distribution $f(x, \theta)$.



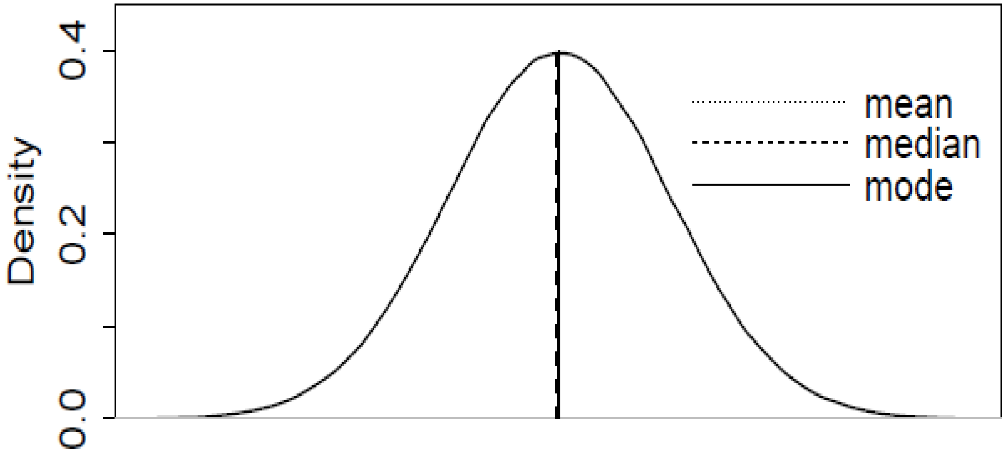
$$f_T(x,\theta) = \begin{cases} \dfrac{f(x,\theta)}{F(b,\theta) - F(a,\theta)}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

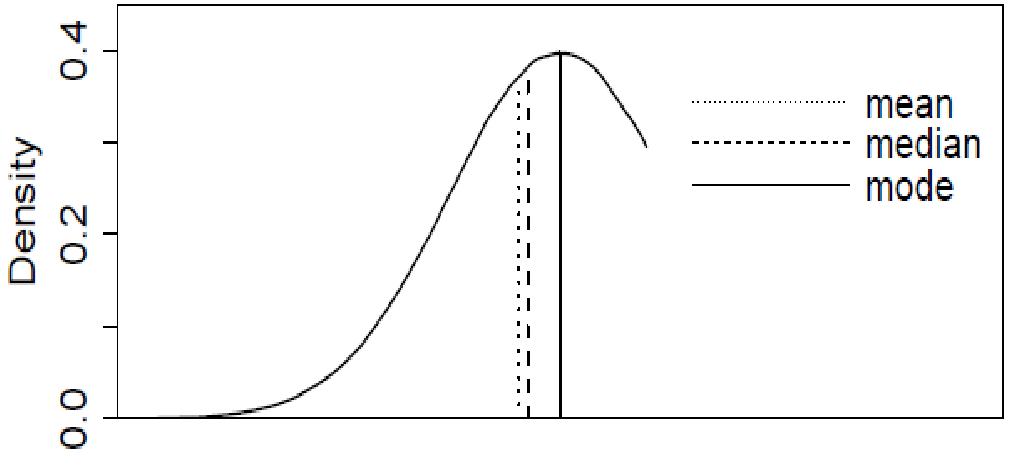
Let $\widetilde{L}(\theta)$ be the expectation of $\ln f(X,\theta)$ with respect to $f_T(X,\theta_0)$ and $L(\theta)$ be the expectation of $\ln f(X,\theta)$ with respect to $f(X,\theta_0)$. Then $\widetilde{L}(\theta) = \int_a^b \ln f(x,\theta) f_T(x,\theta_0)dx$ and $\text{Ł}(\theta) = \int_{-\infty}^{\infty} \ln f(x,\theta) f(x,\theta_0)dx$.

Then the error in convergence is $\|\widetilde{L}(\theta) - \text{Ł}(\theta)\| > 0$.

Visual Example

- If the observed data has a restricted domain then, the accuracy of parameters such as mean or median may drop. This will leads to inaccurate parametric density estimations (based on identification of population mean or median).





Example

For example, consider $X_i \sim N(\mu, \sigma^2)$ and domain is constrained by $[0, b]$. Let the probability density function of X be $f(X, \theta_0)$ and let $a(X, \theta) = -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}$.

By the uniform law of large numbers, the average loglikelihood estimator

$$l_n(X, \mu, \sigma^2) = \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{1}{2}\ln 2\pi\sigma^2 - \frac{(x_i-\mu)^2}{2\sigma^2}\right) = -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n}$$

converges almost surely to $-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2} - \frac{\alpha_0}{2}$. Then $\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n}$ converges to $(1-\alpha_0)\sigma^2$, resulting in errors in estimation. Here we introduce a method to solve above mentioned issue in identifying truncated distributions.

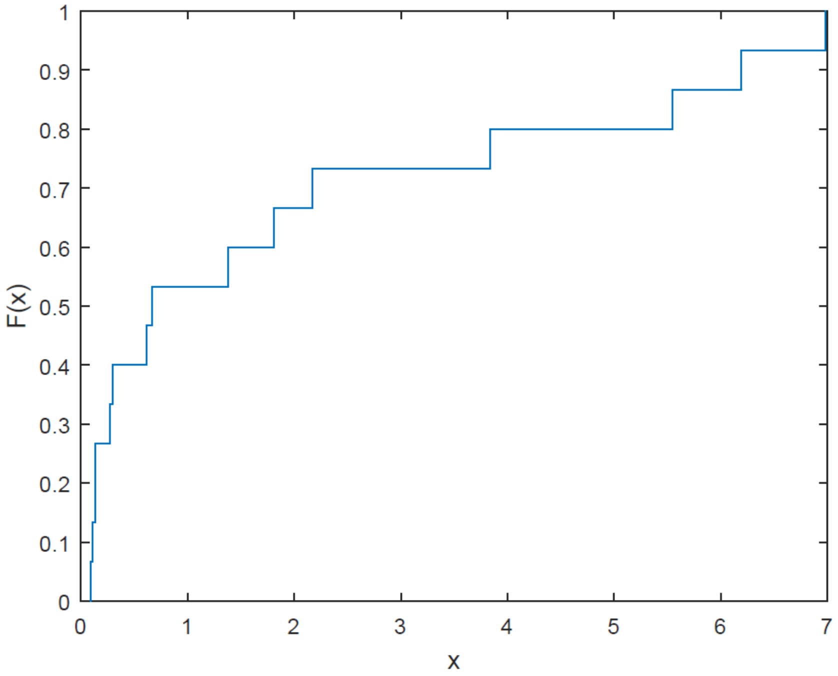
Methodology

- The density function of a continuous random variable can be obtained by the derivative of the cumulative distribution function (CDF). The empirical distribution function generated by the sample data is an estimate for the CDF.
- We introduce two approaches based on the estimated derivative of the empirical distribution.
    - Density estimate using a polynomial approximation of the empirical distribution function for distributions with a known bounded domain.
    - A mode-based density identification for distributions with unknown truncation.

Empirical distribution function

- The proposed methods are based on the empirical distribution function.
- For a set of i.i.d. random variables X_1, X_2,… , X_n with distribution function $F_n(x) = P(X_i \leq x)$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1} \{X_i \leq x\}$$



Glivenko-Cantelli theorem

For any distribution, $F_n(x)$ *converges uniformly to* $F(x)$ *almost surely. That is,*

$\|F_n(x) - F(x)\|_\infty \to 0$ *almost surely.*

Estimating the derivative of $F_n(x)$

Inverse function theorem

If $f:(a,b) \to (c,d)$ is a differentiable surjection and $f'(x)$ is never zero then f is a homeomorphism, and its inverse is differentiable with derivative

$$(f^{-1})'(y) = \frac{1}{f'(x)}, \text{ where } y = f(x)$$

Let $\{x_k\}_{k=1}^{n} \subset \mathbb{R}$ be a data set arranged in ascending order and $S_i = \{x_{i-3}, x_{i-2}, \dots, x_{i+3}\} \subset \{x_k\}_{k=1}^{n}$ having $x_{i-3} < x_{i+3}$ for all $i \geq 4$. Assume S consists of distinct data points. Suppose F is differentiable strictly increasing smoothing approximation for the empirical distribution $F_n(x)$ on S such that $\mathrm{F}(x_i) = \frac{i}{n}$. By Inverse function theorem $\mathrm{F}'(x_i) = \frac{1}{(F^{-1}(i/n))'}$. Let $\mathrm{F}(x)$ be linear and, using the slope of the simple linear regression equation $F^{-1}(s) = ms + b$

$$F'(x_i) = \frac{1}{m} = \frac{\sum_{j=-3}^{3} (j/n)^2}{\sum_{j=-3}^{3} j\,(x_{i+j} - \overline{x})/n} = \frac{28/n}{3x_{i+3} + 2x_{i+2} + x_{i+1} - x_{i-1} - 2x_{i-2} - 3x_{i-3}},$$

where $\overline{x}$ is the mean of the data subset S.

Estimation of density near a boundary

- If $x_5 > x_1$, then $F'(x_3) \approx \dfrac{\sum_{j=-2}^{2} (j/n)^2}{\sum_{j=-2}^{2} j\,(x_{3+j} - \overline{x})\,/n} = \dfrac{10/n}{2x_5 + x_4 - x_2 - 2x_1}$. Otherwise $F'(x_3) \approx F'(x_4)$.

- If $x_3 > x_1$, then $F'(x_2) \approx \dfrac{\sum_{j=-1}^{1} (j/n)^2}{\sum_{j=-1}^{1} j\,(x_{1+j} - \overline{x})\,/n} = \dfrac{2/n}{x_3 - x_1}$. Otherwise $F'(x_2) \approx F'(x_3)$.

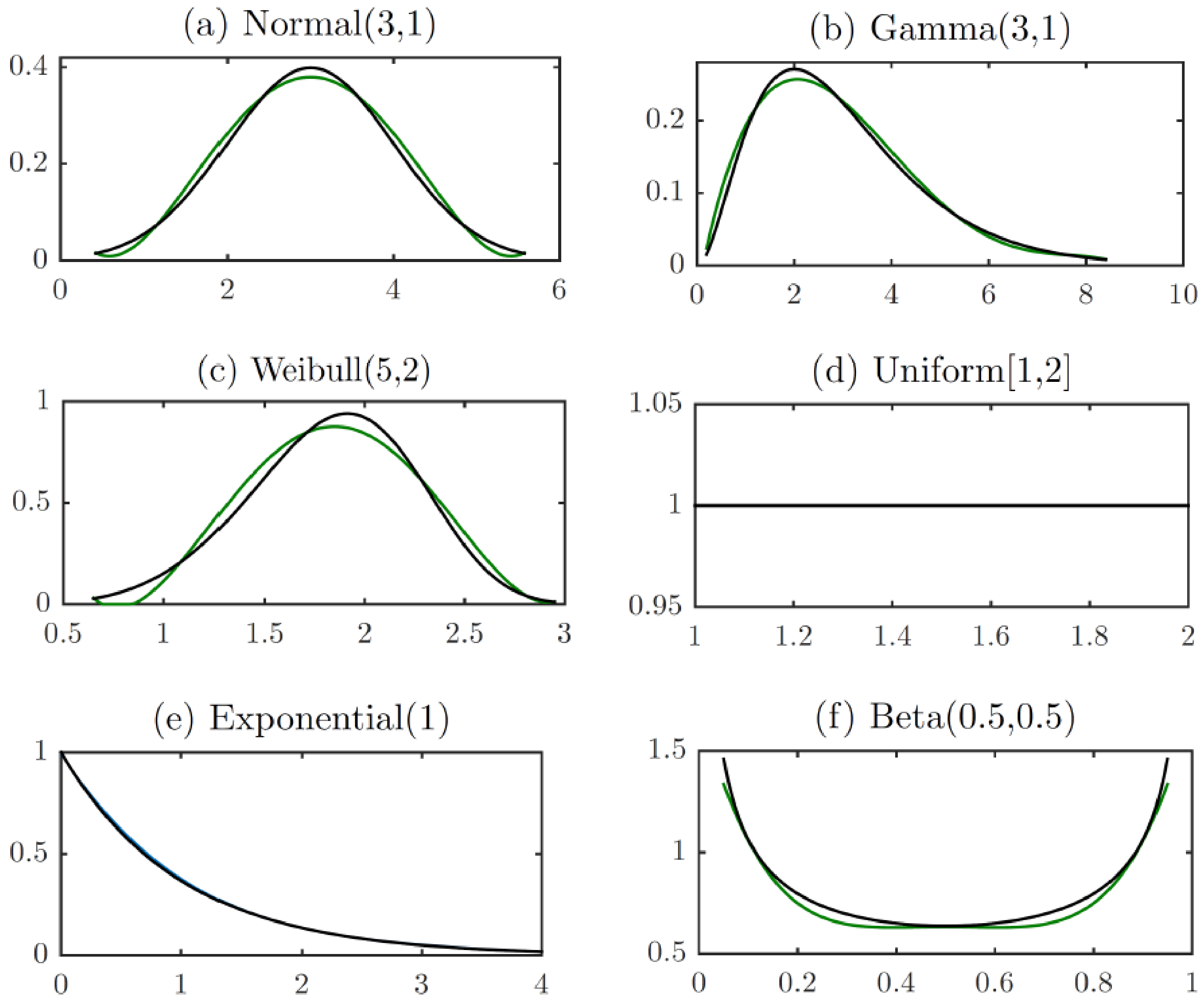
- If $x_2 > x_1$, then $F'(x_1) \approx \dfrac{1/n}{x_2 - x_1}$ (forward difference approximation). Otherwise $F'(x_1) \approx F'(x_2)$.

Method of polynomial approximation

$$\begin{aligned} p(x) &= x^3 + c_0\left(x - x^3\right) + (3 - 2c_0 - c_1)\left(x^2 - x^3\right) \\ &\quad + c_2\left(x^2 + x^4 - 2x^3\right) + c_3\left(2x^2 + x^5 - 3x^3\right) \end{aligned}$$

where $x \in [0,1]$ and $c_0, c_1, \ldots, c_3$ are model parameters. Then $p(0) = 0$, $p(1) = 1, p'(0) = c_0 \; and \; p'(1) = c_1$. We obtain c_2 and c_3 by least squares regression on the empirical distribution function.

Approximations of distributions on $y \in [a, b]$ using the derivative of the polynomial $p(x)$



Mode based density identification

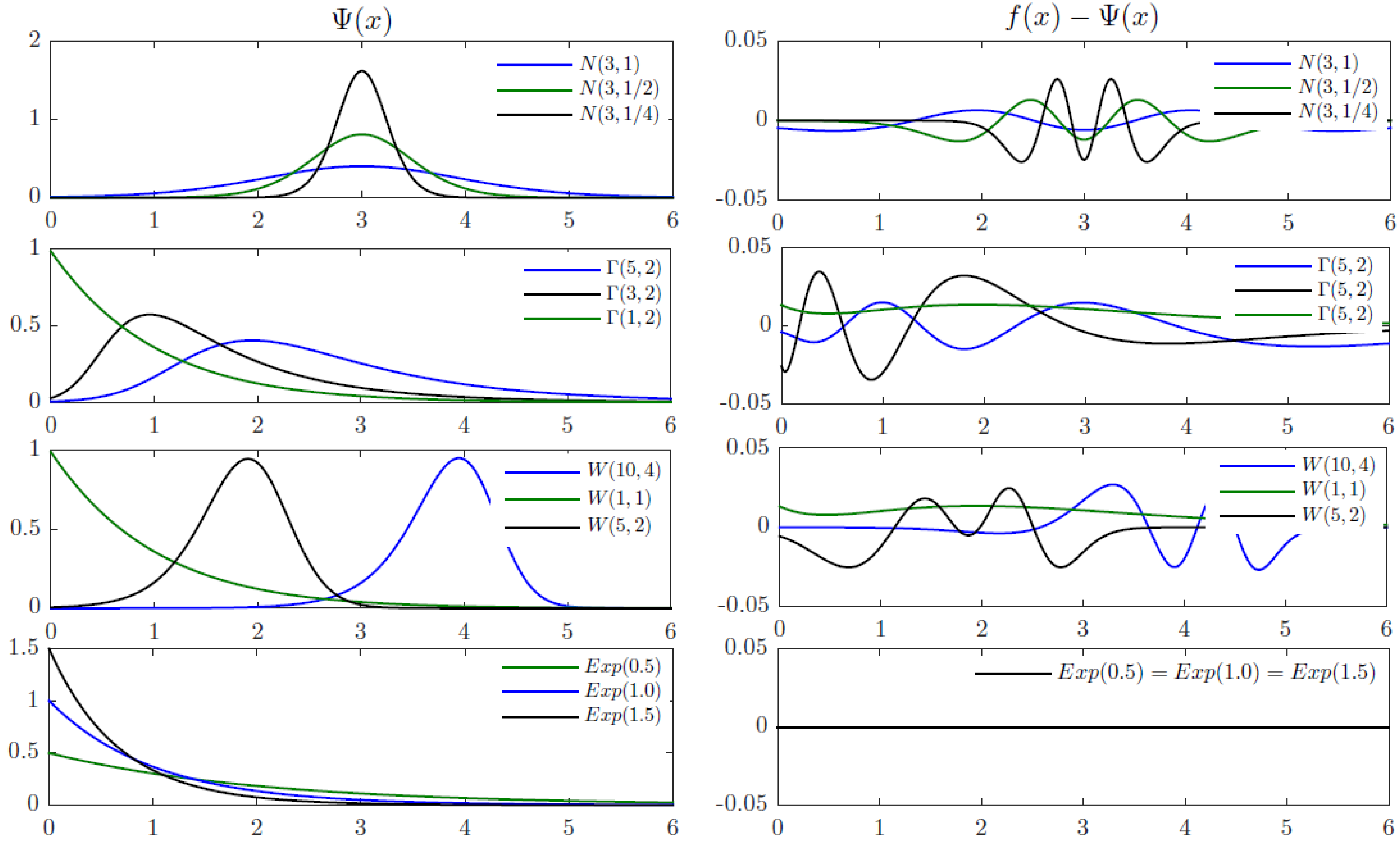
- An approximation of a distribution based on the location of the global maximum (the mode for a unimodal density) provides a robust density estimation method when the data range is restricted. The principle assumption for the proposed method is that the mode is contained in the data range.

We utilize the function

$\Psi(x_i) \triangleq \dfrac{\zeta_i e^{-4\alpha_i(x_i-\gamma_i)}}{\left(1+e^{-\beta_i(x_i-\gamma_i)}\right)^4}$ to estimate the density, where $\zeta_i > 0$, $\alpha_i, \beta_i \geq 0$, and $\gamma_i \in \mathbb{R}$.

For $\eta_i \triangleq \alpha_i/\beta_i \in (0,1)$, $\max \Psi(x_i) = \zeta_i \,(1-\eta_i)^{4(1-\eta_i)}\, \eta_i^{4\eta_i}$ at $x = \gamma + \ln\left(\dfrac{1-\eta_i}{\eta_i}\right)^{1/\beta_i}$.

Approximations of distributions in the exponential family by $\psi(x)$



Consider estimating the density function $f(x,\theta), x \in \mathbb{R}$ and Suppose that the data range is restricted to $[a,b]$. truncated density function is gi-ven by

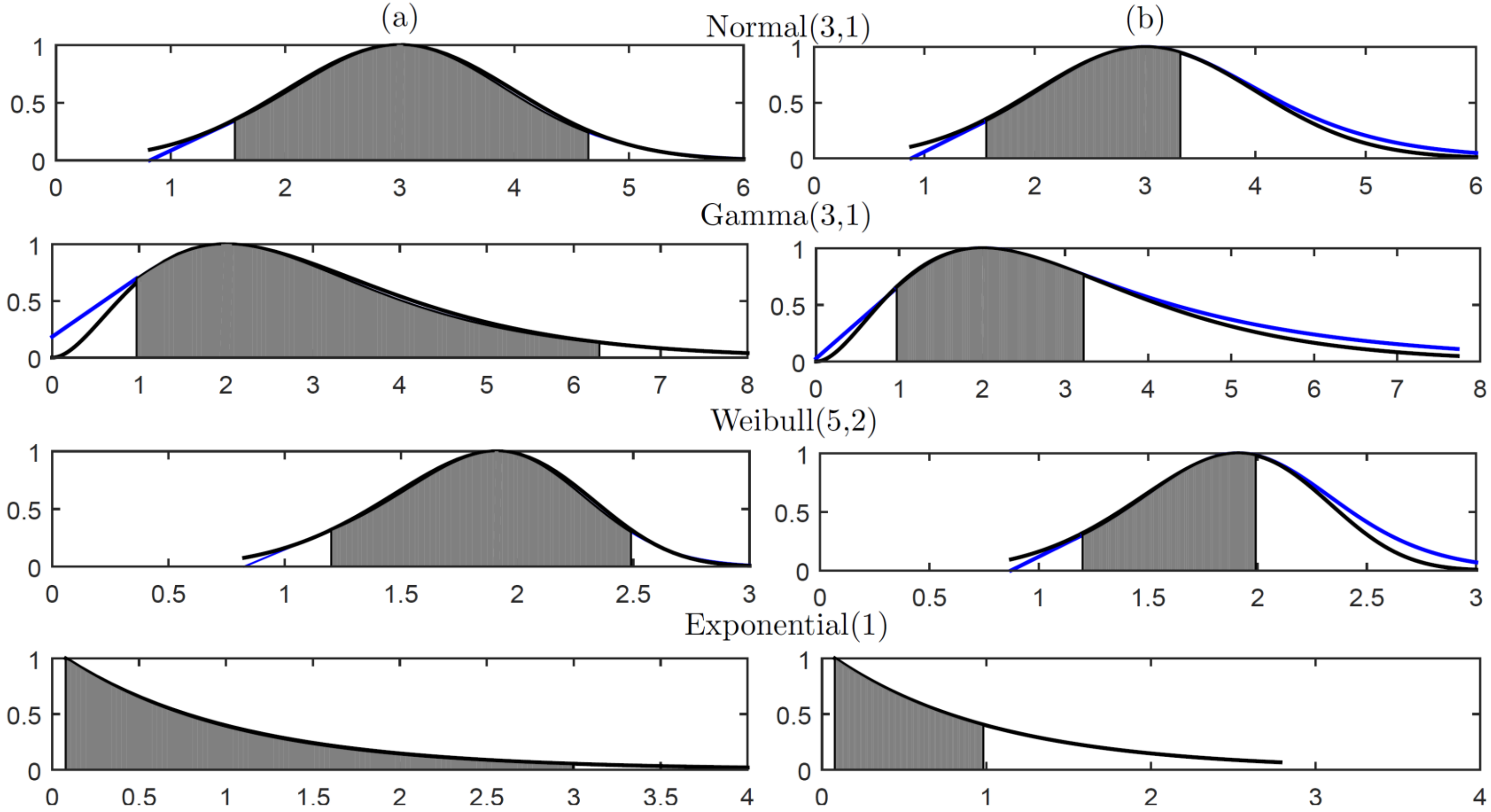
$$f_T(x,\theta) = \frac{f(x,\theta)}{F(b)-F(a)}, \text{ where } x \in [a,b]$$

Let $m = \arg max f(x,\theta)$ be the mode. Normalized density functions at m satisfy

$$\frac{f_T(x,\theta)}{f_T(m,\theta)} = \frac{f(x,\theta)}{f(m,\theta)} \quad and \max_{x\in[a,b]} \frac{f_T(x,\theta)}{f_T(m,\theta)} = 1$$

Estimate the normalized density function using the estimated derivatives of the empirical distribution function.

Density estimation with two different range restrictions. Shaded regions represent the range of the available data. Functions in column (a) and (b) are approximated from the corresponding cumulative distribution function within the 0.05-0.975 and 0.05-0.65 probability range respectively.



Ecology

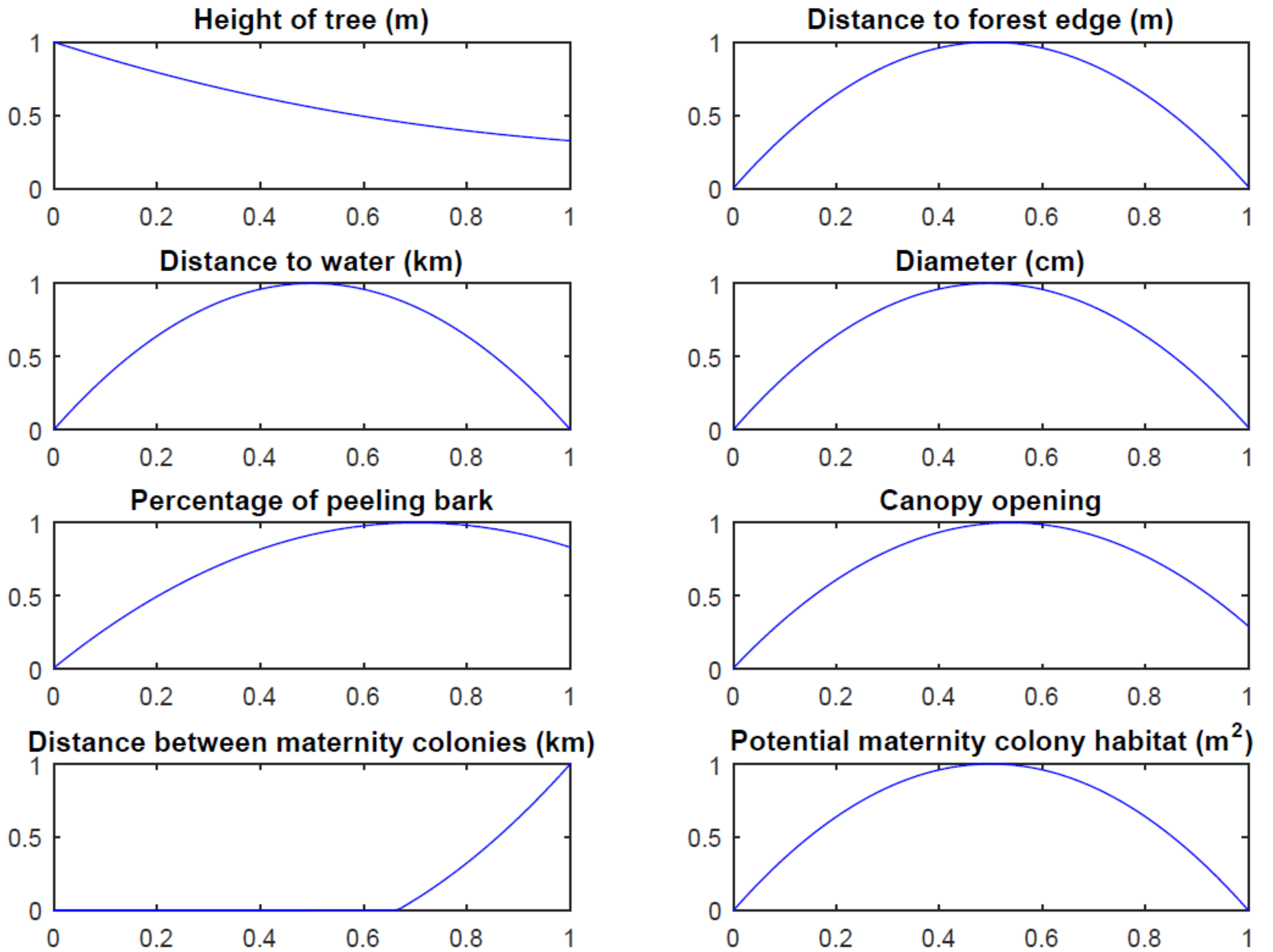
- An interdisciplinary science that includes biology and earth science.
- In ecological data, many data ranges are restricted by geographical or ecological constraints.
  - Measures of distances, which are restricted by landscape patterns.
  - Tree characteristic measures, which are restricted by tree species.
  - Some variables have well-defined bounded domains, like percentage measures.

Example 1

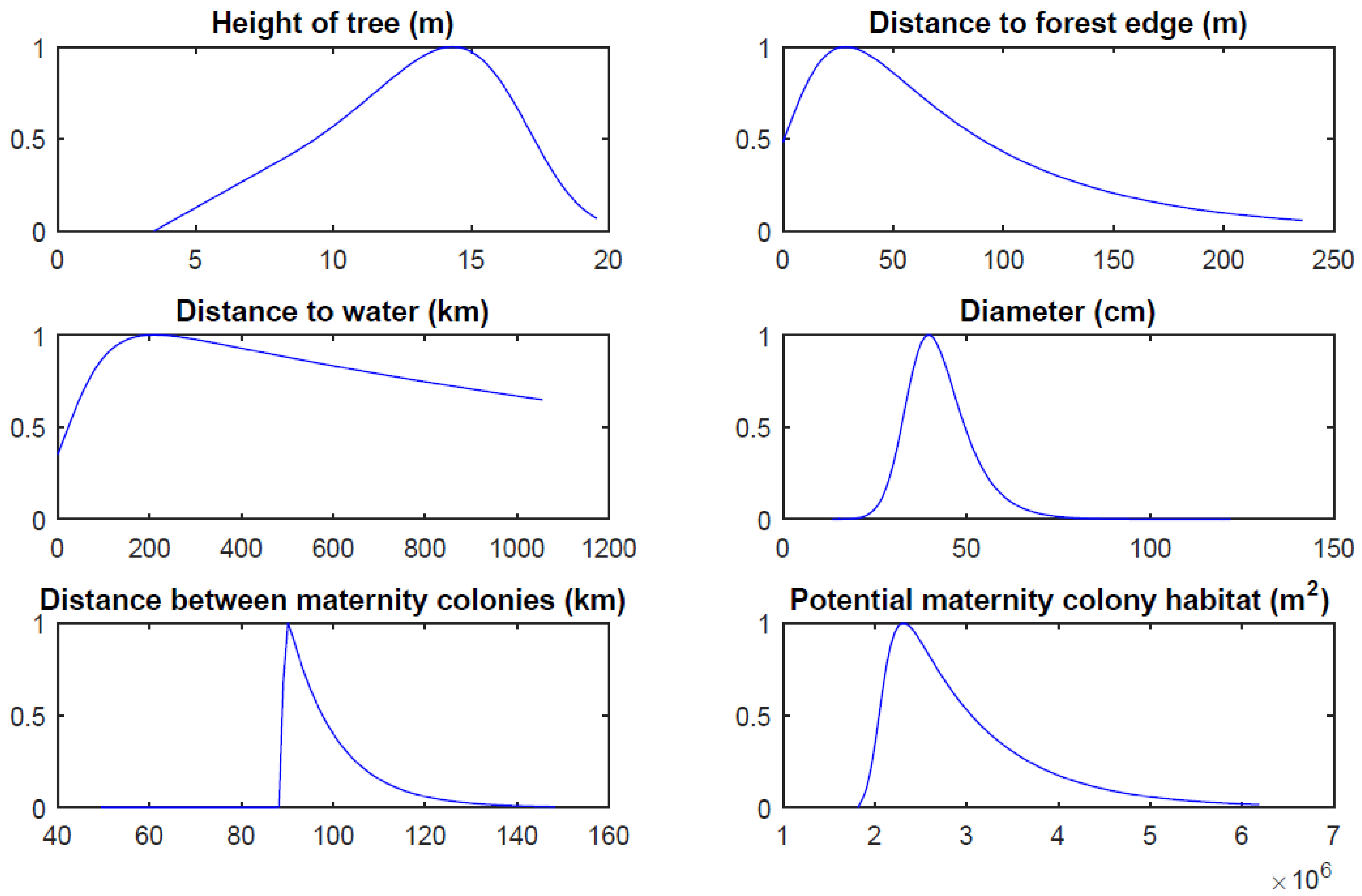
In the first example, we estimate densities for an Indiana bat maternity roost selection dataset (Schroder et al., 2018)

Variables:

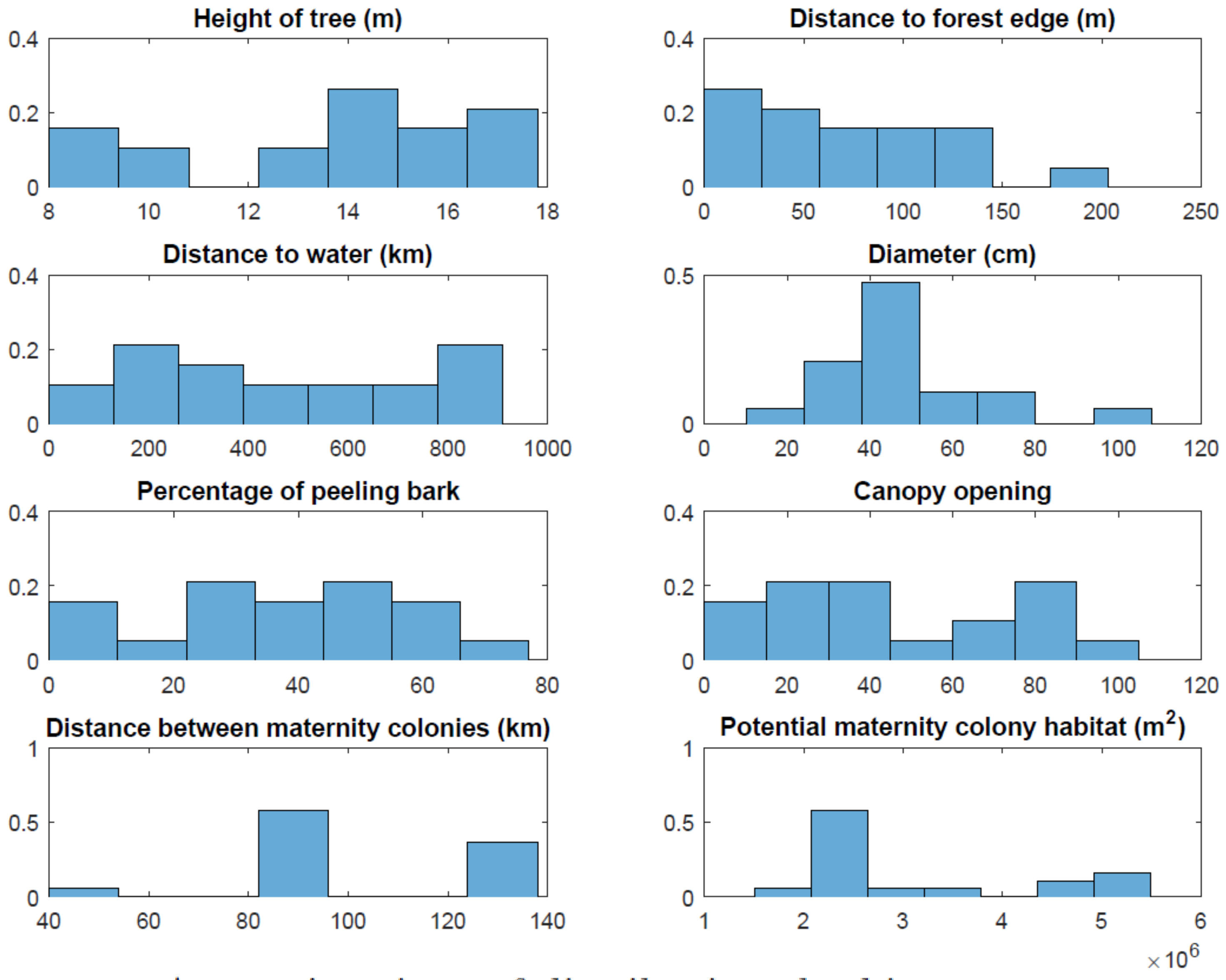
- Tree height (m) .
- Distance to forest edge (m),
- distance to water(km).
- Tree diameter (cm).
- Percentage of peeling bark.
- Canopy opening (percentage variable).
- Distance between maternity colonies (km) .
- Potential maternity colony habitat (m^2).



Polynomial approximations of distributions on $y \in [a, b]$ using the derivative of the polynomial $p(x) = x^3 + c_0 (x - x^3) + (3 - 2c_0 - c_1)(x^2 - x^3) + c_2 (x^2 + x^4 - 2x^3) + c_3 (2x^2 + x^5 - 3x^3)$, where $x = \dfrac{y-a}{b-a}$, $c_0, c_1 \geq 0$, $c_2, c_3 \in \mathbb{R}$.



Approximations of distributions in the exponential family by $\Psi(x)$ for every variable that is not a percentage.



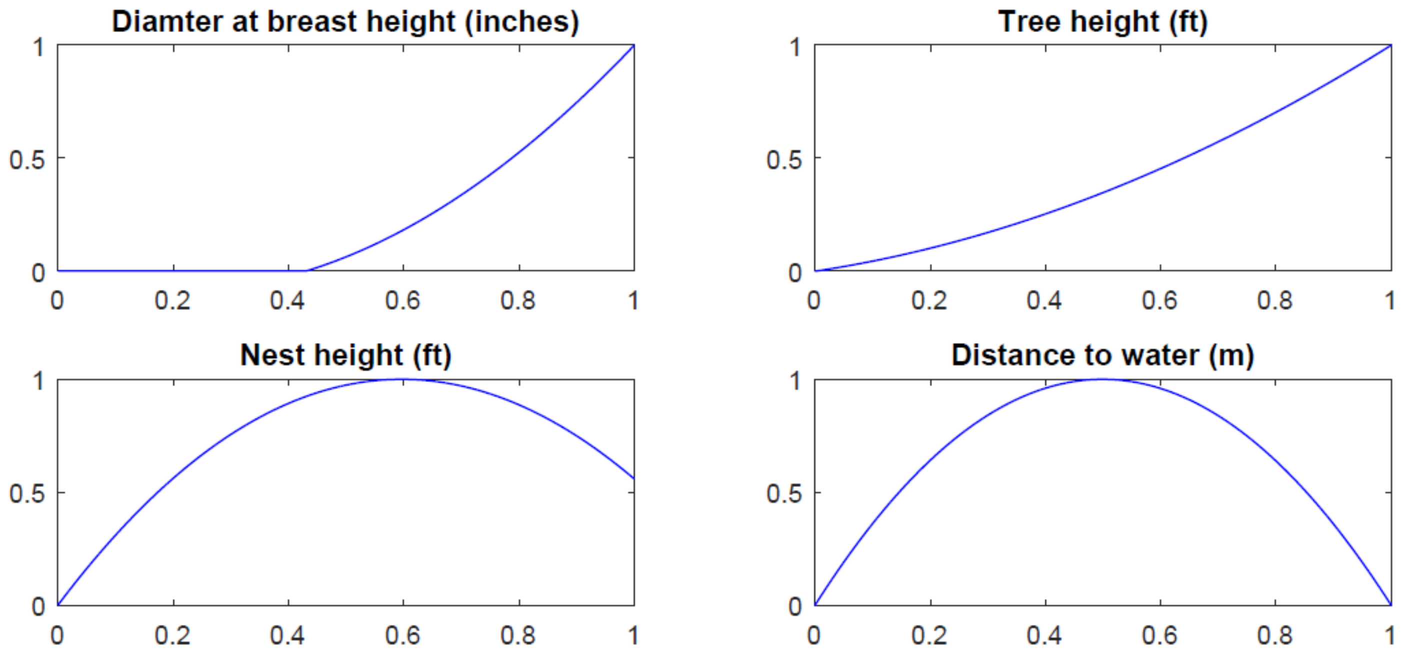
Approximations of distributions by histogram.

Example 2

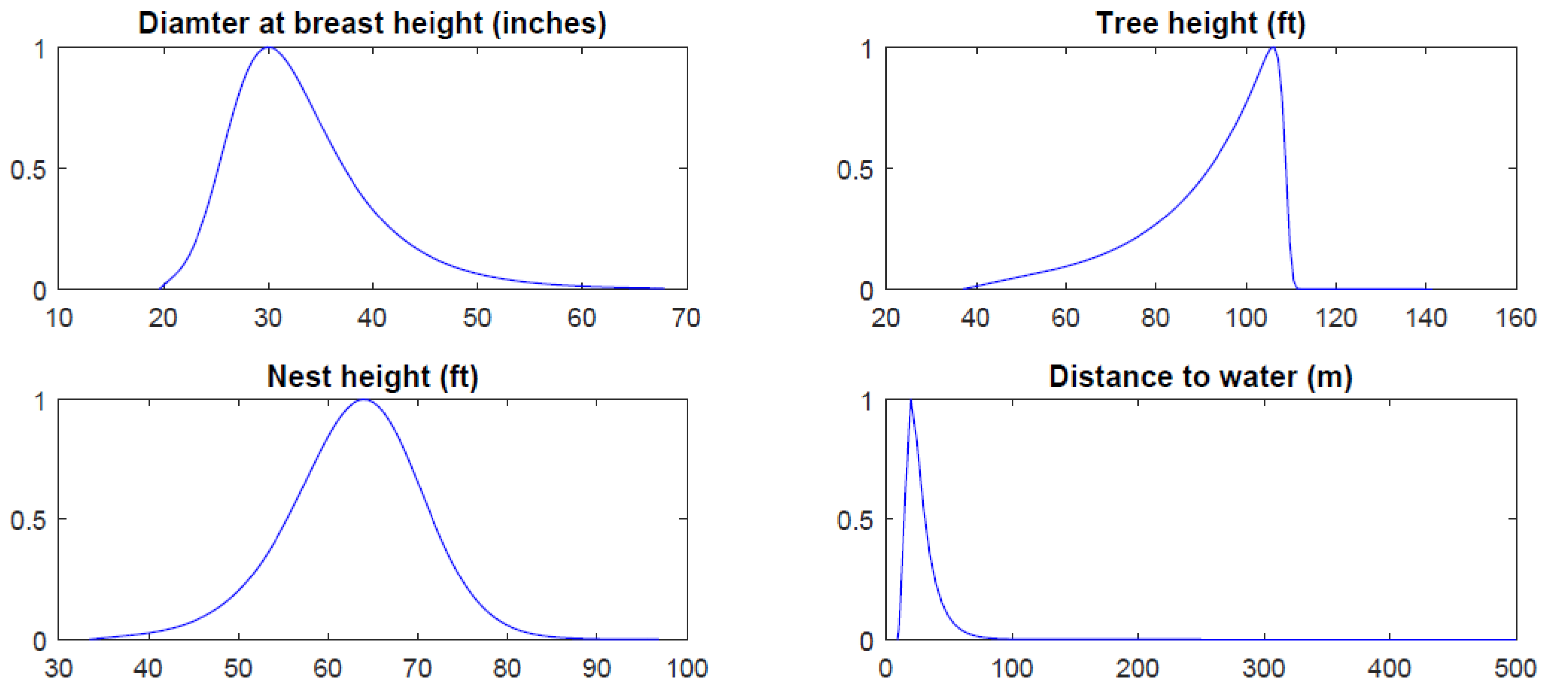
In the second example, we construct the densities for bald eagle nesting habitats in the Upper Mississippi River National Wildlife and Fish Refuge data (Mundahl et al., 2013).

Variables:

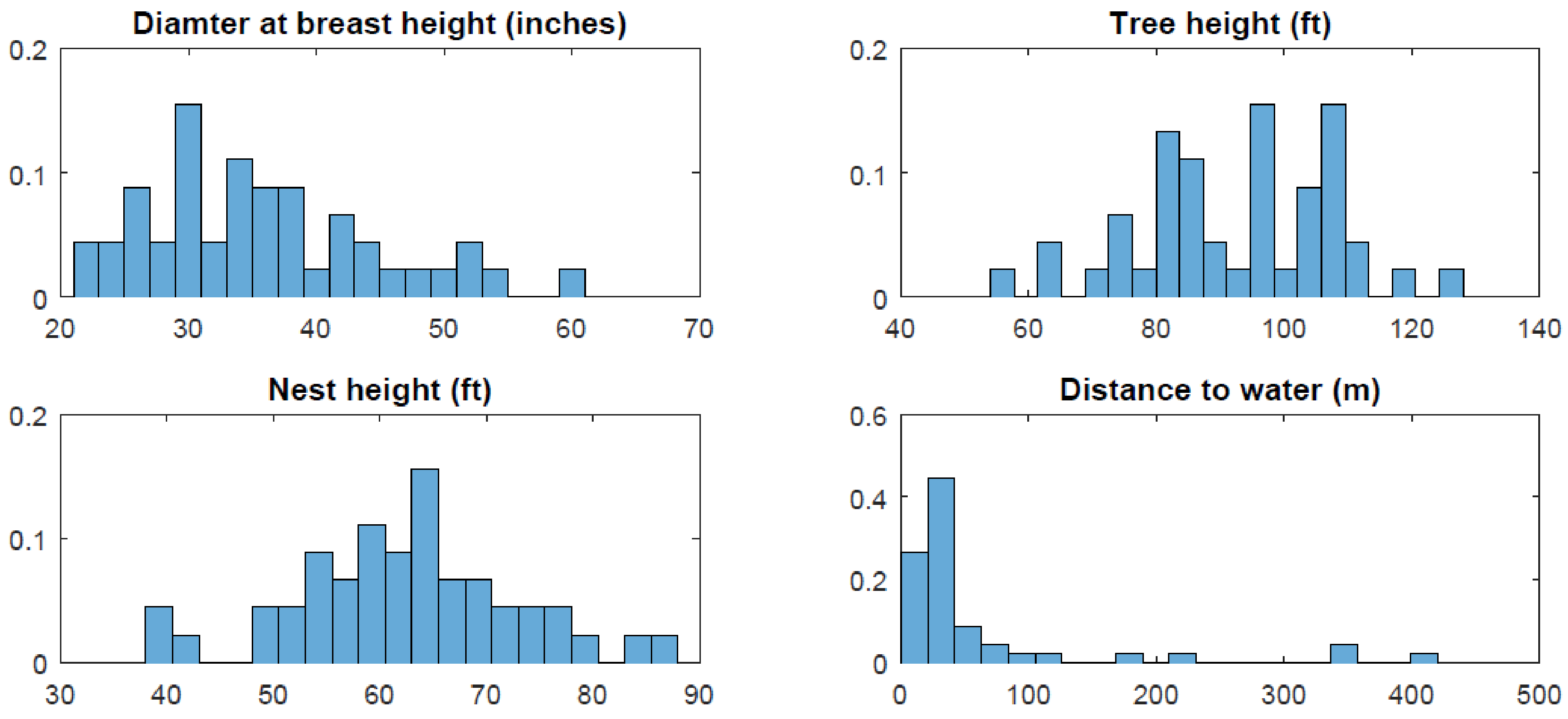
- Tree diameter at breast height (inches).
- Tree height (ft).
- Nest height (ft).
- Distance to water (m).



Polynomial approximations of distributions on $y \in [a, b]$ using the derivative of the polynomial $p(x) = x^3 + c_0\,(x - x^3) + (3 - 2c_0 - c_1)\,(x^2 - x^3) + c_2\,(x^2 + x^4 - 2x^3) + c_3\,(2x^2 + x^5 - 3x^3)$, where $x = \dfrac{y-a}{b-a}$, $c_0, c_1 \geq 0$, $c_2, c_3 \in \mathbb{R}$.



Approximations of distributions in the exponential family by $\Psi(x)$.



Approximations of distributions by histogram.

Conclusion

- The Polynomial approximation is well suited to approximate any unimodal distribution with known truncation.
- The approximations of distributions in the exponential family by $\psi(x)$ is appropriate to model various unimodal density functions for data with restricted ranges.
- Unlike traditional sample-based methods, such as histograms, the proposed density estimation methods are well suited when data are drawn from a truncated distribution.
- Since many ecological data ranges are restricted by geographical or ecological constraints. Therefore, the methods proposed here will be useful in ecological data analysis.

Thank you…