Symbolic Data Analysis

Boris Beranger

with

Jaslene Lin, Thomas Whitaker, Scott Sisson

ACEMS, Postdoc Workshop
Brisbane, 22nd March 2017

Motivation

- Why Symbolic Data Analysis? / What is it for?

  ⇒ When there is **'a lot'** of data

  ⇒ When the data ins't under the classical form

- **Aim:** Show how to use SDA.
- **Challenges:** Recent topic, very little known.

Motivation

- Why Symbolic Data Analysis? / What is it for?

  ⇒ When there is **'a lot'** of data

  ⇒ When the data ins't under the classical form

- **Aim:** Show how to use SDA.
- **Challenges:** Recent topic, very little known.

Motivation

- Why Symbolic Data Analysis? / What is it for?

  $\Rightarrow$ When there is **'a lot'** of data

  $\Rightarrow$ When the data ins't under the classical form

- **Aim:** Show how to use SDA.
- **Challenges:** Recent topic, very little known.

Motivation

- Why Symbolic Data Analysis? / What is it for?

  ⇒ When there is **'a lot'** of data

  ⇒ When the data ins't under the classical form

- **Aim:** Show how to use SDA.
- **Challenges:** Recent topic, very little known.

Motivation

- Why Symbolic Data Analysis? / What is it for?

  $\Rightarrow$ When there is **'a lot'** of data

  $\Rightarrow$ When the data ins't under the classical form

- **Aim:** Show how to use SDA.
- **Challenges:** Recent topic, very little known.

The setup

Notation:

- X: Classical random variable, $X \sim g_X(\cdot;\theta)$;
- S: Symbolic random variable, $S \sim f_S(\cdot;\vartheta)$;
- $\mathcal{L}(y;p)$: Likelihood evaluated at y for the parameter(s) p.

Result. The symbolic likelihood function can be obtained through

$$\mathcal{L}(s;\theta,\vartheta) \propto \int_x f_{S|X}(s|x;\vartheta) g_X(x;\theta)\mathrm{d}x,$$

where $x = (x_1, \ldots, x_m)$.

The setup

Notation:

- X: Classical random variable, $X \sim g_X(\cdot;\theta)$;
- S: Symbolic random variable, $S \sim f_S(\cdot;\vartheta)$;
- $\mathcal{L}(y;p)$: Likelihood evaluated at y for the parameter(s) p.

Result. The symbolic likelihood function can be obtained through

$$\mathcal{L}(s;\theta,\vartheta) \propto \int_x f_{S|X}(s|x;\vartheta) g_X(x;\theta)\mathrm{d}x,$$

where $x = (x_1, \ldots, x_m)$.

Some symbols

1. Interval-valued symbols $S = (\underline{X}, \bar{X})$,
   $\underline{X} = \min_i X_i$ and $\bar{X} = \max_i X_i$

$$\mathcal{L}(\underline{x}, \bar{x}; \theta, m) = m(m-1) \left[G_X(\bar{x}; \theta) - G_X(\underline{x}; \theta) \right]^{m-2} g_X(\bar{x}; \theta) g_X(\underline{x}; \theta),$$

2. Histogram-valued symbols $S = (S_1, \ldots, S_B)$ counts

$$\mathcal{L}(s_1, \ldots, s_B; \theta) = \frac{m!}{s_1! \cdots s_B!} \prod_{b=1}^{B} P_b(\theta)^{s_b},$$

3. Normal-valued symbols $X|S \sim \mathcal{N}_d(S_1, S_2)$

Some symbols

1. Interval-valued symbols $S = (\underline{X}, \bar{X})$,
   $\underline{X} = \min_i X_i$ and $\bar{X} = \max_i X_i$

$$\mathcal{L}(\underline{x}, \bar{x}; \theta, m) = m(m-1)\left[G_X(\bar{x}; \theta) - G_X(\underline{x}; \theta)\right]^{m-2} g_X(\bar{x}; \theta) g_X(\underline{x}; \theta),$$

2. Histogram-valued symbols $S = (S_1, \ldots, S_B)$ counts

$$\mathcal{L}(s_1, \ldots, s_B; \theta) = \frac{m!}{s_1! \cdots s_B!} \prod_{b=1}^{B} P_b(\theta)^{s_b},$$

3. Normal-valued symbols $X|S \sim \mathcal{N}_d(S_1, S_2)$

Some symbols

1. Interval-valued symbols $S = (\underline{X}, \bar{X})$,
   $\underline{X} = \min_i X_i$ and $\bar{X} = \max_i X_i$

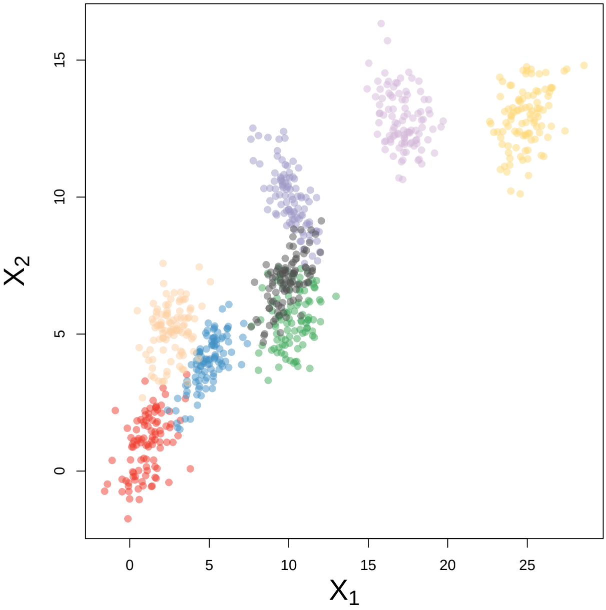
$$\mathcal{L}(\underline{x}, \bar{x}; \theta, m) = m(m-1)\left[G_X(\bar{x}; \theta) - G_X(\underline{x}; \theta)\right]^{m-2} g_X(\bar{x}; \theta) g_X(\underline{x}; \theta),$$

2. Histogram-valued symbols $S = (S_1, \ldots, S_B)$ counts

$$\mathcal{L}(s_1, \ldots, s_B; \theta) = \frac{m!}{s_1! \cdots s_B!} \prod_{b=1}^{B} P_b(\theta)^{s_b},$$

3. Normal-valued symbols $X | S \sim \mathcal{N}_d(S_1, S_2)$

E.g. 1



E.g. 1a: Biv hist symbols - $X_2 \sim \mathcal{N}(\beta_0 + \beta_1 X_1, \sigma^2)$

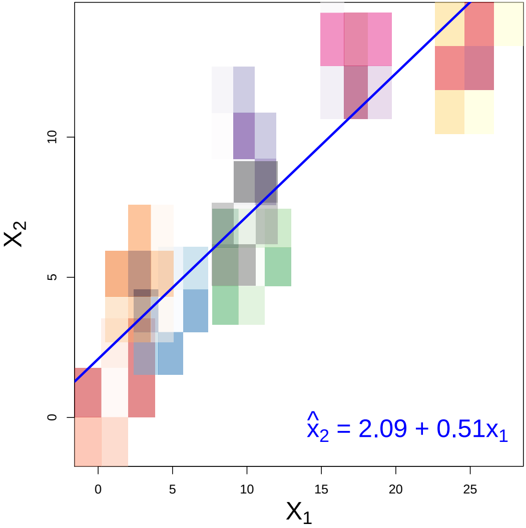


Figure: Linear regression for bivariate histograms with 3×3 bins.

E.g. 1a: Biv hist symbols - $X_2 \sim \mathcal{N}(\beta_0 + \beta_1 X_1, \sigma^2)$

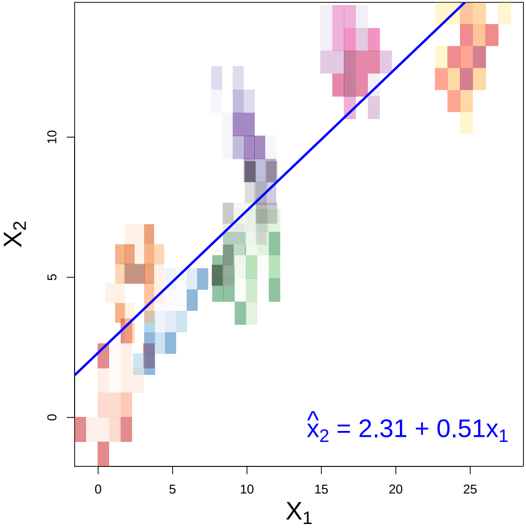


Figure: Linear regression for bivariate histograms with 'optimal' bins.

E.g. 1a: Biv hist symbols - $X_2 \sim \mathcal{N}(\beta_0 + \beta_1 X_1, \sigma^2)$

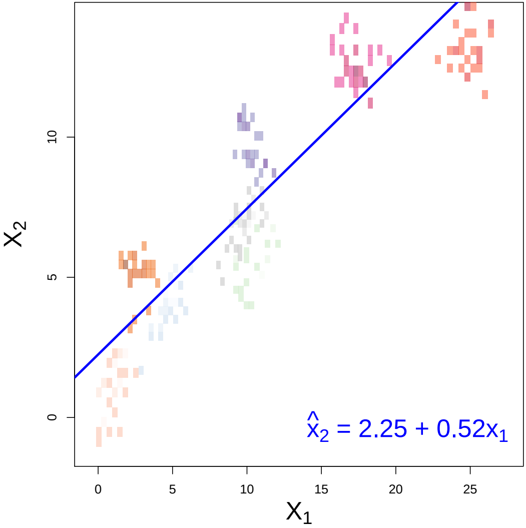


Figure: Linear regression for bivariate histograms with 15×15 bins.

E.g. 1a: Biv hist symbols - $X_2 \sim \mathcal{N}(\beta_0 + \beta_1 X_1, \sigma^2)$

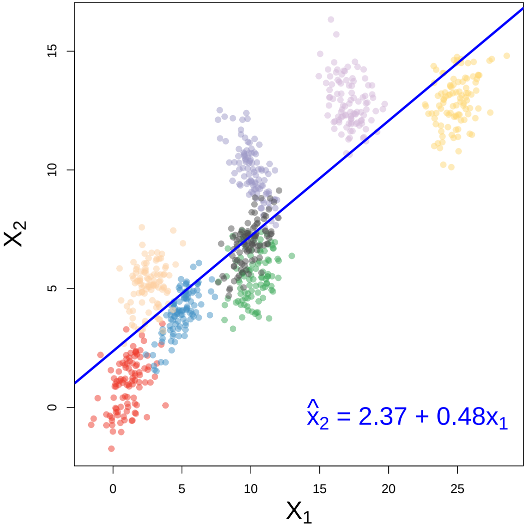


Figure: Linear regression for classical data.

E.g. 1b: Distrib. symbols - $f_{S|X}(s;\theta) \sim \mathcal{N}_2(\mu, \Sigma)$

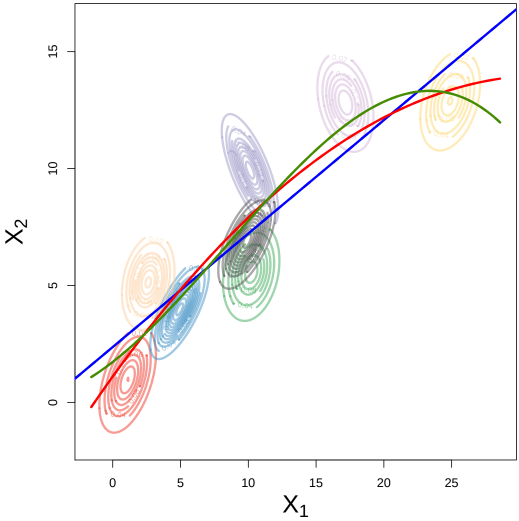


Figure: Normal distribution-valued symbols.

E.g. 2: Distrib. symbols - $f_{S|X}(s;\theta) \sim \mathcal{N}_2(\mu, \Sigma)$

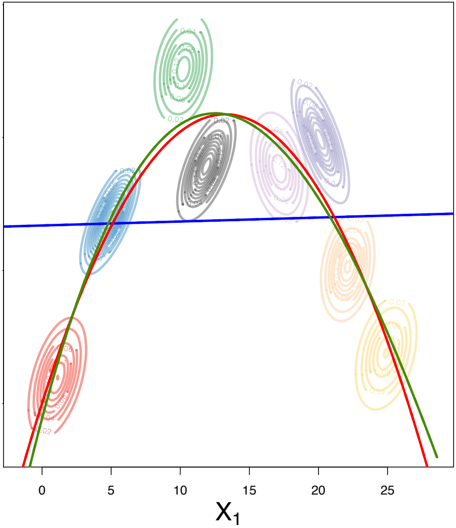


Figure: Normal distribution-valued symbols.

E.g. 2: Distrib. symbols - $f_{S|X}(s;\theta) \sim \mathcal{N}_2(\mu, \Sigma)$

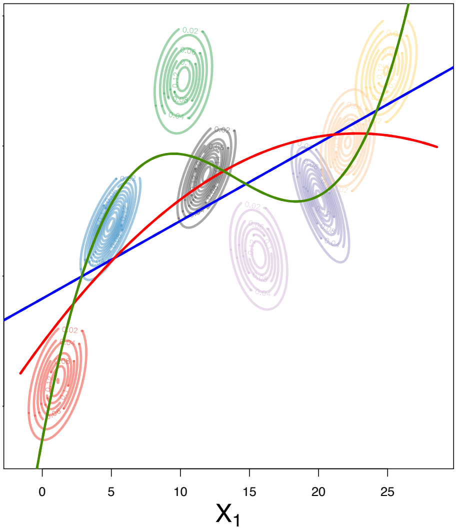


Figure: Fit classical density using normal symbols.

Impact

- Use in **Spatial Extremes**: Application to daily max temp in Aus.
- Likelihood is intractable $\rightarrow$ Composite likelihood methods
- Lot of data at many locations in space $\rightarrow$ **Composite likelihood methods for histogram data**

Thank you!

Impact

- Use in **Spatial Extremes**: Application to daily max temp in Aus.
- Likelihood is intractable → Composite likelihood methods
- Lot of data at many locations in space → **Composite likelihood methods for histogram data**

Thank you!

Impact

- Use in **Spatial Extremes**: Application to daily max temp in Aus.
- Likelihood is intractable $\rightarrow$ Composite likelihood methods
- Lot of data at many locations in space $\rightarrow$ **Composite likelihood methods for histogram data**

Thank you!

Impact

- Use in **Spatial Extremes**: Application to daily max temp in Aus.
- Likelihood is intractable $\rightarrow$ Composite likelihood methods
- Lot of data at many locations in space $\rightarrow$ **Composite likelihood methods for histogram data**

Thank you!