Bayesian P-splines and the estimation of uni- and bivariate densities from grouped data

Philippe Lambert
(Joint work with Paul Eilers, Utrecht Univ.)

http://www.statsoc.ulg.ac.be/

Institut des Sciences Humaines et Sociales,
Université de Liège
Motivation

- Descriptive statistics are easily computed when data are available with high precision.
- Unfortunately, this is not always the case:
  - Measured concentrations can be below the detection limit.
  - Instrument may have poor resolution.
  - Data are sometimes aggregated for privacy reasons.
- Then, the available data take the form of frequencies associated to 'large' intervals (rough histogram).
- Our goal: to estimate the underlying density.

Then, derived quantities such as moments or quantiles can be estimated.
Example: lead level in the blood

- Grouped data corresponding to concentrations of lead in the blood of New-Yorkers in the 1970-1976 period (Hasselblad et al., 1980).

- The dataset of interest concerns young Puerto Ricans aged 1-12 years in 1974:

<table>
<thead>
<tr>
<th>Wide interval</th>
<th>(0,15)</th>
<th>(15,25)</th>
<th>(25,35)</th>
<th>(35,45)</th>
<th>(45,55)</th>
<th>(55,65)</th>
<th>65+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq. $m_j$</td>
<td>27</td>
<td>71</td>
<td>32</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- The data were recorded in broad intervals for screening purposes.

- A blood lead level over 30 $\mu$g/dl is an unacceptable medical risk.

- One quantity of scientific interest: the proportion of persons with a blood lead level over that threshold.
- A B-spline $b_k(x)$ of degree $s$ is supported by $s+2$ knots.
- At given $x$, at most $(s + 1)$ B-splines are non-zero.
- $\sum_k b_k(x) = 1.$
Approximation of a curve using B-splines

- A smooth function can be approximated using a linear combination of B-splines:

\[ f(x) \approx \sum_k \theta_k b_k(x) \]

- The B-splines coefficients depict the same shape as the fitted curve.

Therefore, desired properties (such as monotonicity, convexity...) for the approximating curve can be obtained by forcing similar behaviours on the splines coefficients \( \theta_j \)'s.
The composite link model (Thompson & Baker, 1981)

- Let \( \pi \) be the latent distribution defined on a grid of \( I \) narrow intervals.
- We model these probabilities using polytomous logistic regression
  \[
  \pi_i = \frac{e^{\eta_i}}{e^{\eta_1} + \ldots + e^{\eta_I}} \quad \text{with} \quad \eta = B\phi.
  \]
- We cannot observe \( \pi \) itself, but only sums over \( J < I \) intervals. The expected values for these intervals are
  \[
  \gamma = C\pi
  \]
  where \( C \) is an indicator (0/1) matrix, such that \( c_{ji} = 1 \) if narrow interval \( i \) contributes to wide interval \( j \).
- The frequencies \( m \) associated to the big bins are such that
  \[
  (M_1, \ldots, M_J|\gamma) \sim \text{Mult}(m_+; \gamma_1, \ldots, \gamma_J) \quad \text{with} \quad \gamma = C\pi.
  \]
**Inference: frequentist setting**

- Thompson and Baker (1981) showed how to estimate the parameters of the CLM in a frequentist setting.

- It boils down to a polytomous logistic regression of \( m \) on a “working” matrix \( X = W^{-1}CHB \), with \( H = \text{diag}(m + \pi(1 - \pi)) \) and \( W = \text{diag}(m + \gamma(1 - \gamma)) \).

- Scoring algorithm at iteration \( t + 1 \):
  \[
  X_t'W_tX_t \phi_{t+1} = X_t'(m - m_t + W_tX_t\phi_t),
  \]

- We put a roughness penalty on \( \phi \), to force the estimated distribution to be smooth.

- The penalized log-likelihood is
  \[
  L^* = \sum m_j \log(\gamma_j) - \frac{\lambda}{2}||D\phi||^2 = \sum m_j \log(\gamma_j) - \frac{\lambda}{2} \phi'P\phi \quad \text{with} \quad P = D'D
  \]
  where \( D \) is the \( r \)th order differencing matrix such that \( D\phi = \Delta^r\phi \).

- The penalty modifies the scoring algorithm slightly (Eilers, 2007):
  \[
  (X_t'W_tX_t + \lambda P) \phi_{t+1} = X_t'(m - m_t + W_tX_t\phi_t)
  \]
Inference: Bayesian setting

• The roughness penalty translates into a smoothness prior for the spline coefficients

\[ p(\phi|\lambda) \propto \lambda^{R(P)/2} \exp \left\{ -\frac{\lambda}{2} \phi'P\phi \right\} \]

where \( P = D'D \) and \( R(P) \) is the rank of \( P \).

• A gamma prior with large variance for \( \lambda \) is a possible choice, although more robust results can be obtained with a mixture of gammas (Jullion and Lambert, 2007).

• Therefore, the full model is

\[ (M_1, \ldots, M_J|\phi) \sim \text{Mult}(m_+; \gamma_1(\phi), \ldots, \gamma_J(\phi)) \text{ with } \gamma = C\pi(\phi) \]

\[ p(\phi|\lambda) \propto \lambda^{R(P)/2} \exp \left\{ -\frac{\lambda}{2} \phi'P\phi \right\} \]

\[ \lambda \sim \mathcal{G}(a, b) \text{ with } a = b = .0001 \]

• Closed forms for the log-posterior of \((\phi, \lambda)\) and its gradient can be obtained.

• The Langevin-Hastings algorithm (Lambert and Eilers, 2005, 2006) can be used to draw a sample, \( \{(\phi^{(m)}, \lambda^{(m)}) : m = 1, \ldots, M\} \), from the posterior.
• To each element of the sample, \( \phi^{(m)} \), corresponds a density \( f^{(m)}(y) \) from which any summary measure \( \xi^{(m)} \) of interest such as the mean, the standard deviation or quantiles can be computed.

• Point estimates and credible intervals for \( \xi \) can be derived from the so-obtained sample \( \{\xi^{(m)}: m = 1, \ldots, M\} \).

• Specific properties such as unimodality or log-concavity can be imposed on the estimated density by excluding, through the prior, the configurations of \( \phi \) corresponding to non-desirable densities.
Sampling using the modified Langevin algorithm

- The posterior has many dimensions ⇒ a fast and efficient sampler is required.
- We propose to use the Langevin algorithm (Roberts & Tweedie, Bernoulli 1996).
- This is a MCMC sampler. Given a target density \( p(\theta|D) \) and state \( \theta^{(m-1)} \in \mathbb{R}^K \):

1. At iteration \( m \), generate a proposal \( \theta \) by sampling from

\[
N_K(\theta^{(m-1)} + 0.5 \delta \Sigma \nabla \log p(\theta^{(m-1)}|D), \delta \Sigma)
\]

with \( \Sigma = (X_t'W_tX_t + \lambda_0 P)^{-1} \)

2. Accept that proposal with probability

\[
\alpha(\theta^{(m-1)}, \theta) = \min \left\{ 1, \frac{p(\theta|D)}{p(\theta^{(m-1)}|D)} \frac{q(\theta, \theta^{(m-1)})}{q(\theta^{(m-1)}, \theta)} \right\}
\]

where

\[
\frac{q(\theta, \theta^{(m-1)})}{q(\theta^{(m-1)}, \theta)} = \exp \left\{ -\frac{1}{2} \left( G + G^{(m-1)} \right)' \left( (\theta - \theta^{(m-1)}) + \frac{\delta \Sigma}{4} (G - G^{(m-1)}) \right) \right\}
\]

with \( G = \nabla \log p(\theta|D) \); \( G^{(m-1)} = \nabla \log p(\theta^{(m-1)}|D) \).
Simulation study: setting

500 samples of size $n = 200$ or 1000 were simulated from a gamma(10,2).
Some unimodal densities generated by MCMC \((n = 200)\)

- For each of the \(S = 500\) datasets:

- **1-** A chain of length 2,000 was run to tune the estimation of the dependence structure of the posterior of the spline parameters.

- **2-** After a burn-in of length 500, a chain of length \(M = 3,000\) was run. To each of the generated \(\phi^{(m)}\) corresponds a density \(f^{(m)}(\cdot)\) (grey curves).

- **3-** The fitted density (in black) corresponds to the posterior mean.
Simulation: 10 out of the 500 fitted densities \((n = 200)\)
Simulation study: $\log \text{MISE}(\hat{f})$ when $n = 200$

$\text{MISE} = \int (f(x) - \hat{f}(x))^2 f(x) \, dx$

-1- and -2-: frequentist CLM with $\lambda$ selected using AIC or BIC.
-3-: Bayesian CLM.  
-4-: kernel density estimation (KDE).
Simulation study: $\log \text{MISE}(\hat{f})$ when $n = 1000$

- Kernel density estimate (computed from the ungrouped data !!!) less performant.
- A $\lambda$ selected using BIC is preferable to one selected using AIC.
- Bayes performs best when $bw \leq 1.2\sigma$. 
### Simulation study: $100 \times \sqrt{\text{MSE}}/\sigma$ ($n = 200$)

<table>
<thead>
<tr>
<th>bw</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6σ</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>1.4σ</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>1.2σ</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>13</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>1.0σ</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>0.8σ</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>0.6σ</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>0.4σ</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>KDE</td>
<td>13</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>SQ</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>17</td>
</tr>
</tbody>
</table>

- RMSE slowly $\uparrow$ with the amount of grouping with an acceleration when $bw > 1.2\sigma$.

- For ungrouped data, the Bayesian method performs
  - better than kernel: it is very marked in the lower tail.
  - better than the sample quantiles (SQ) thanks to the joint estimation of quantiles.
Coverage of credible intervals for some deciles \((n=200)\)

<table>
<thead>
<tr>
<th>bw (\sigma)</th>
<th>Credib.</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6σ</td>
<td>80%</td>
<td>78</td>
<td>75</td>
<td>74</td>
<td>75</td>
<td>76</td>
<td>78</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>89</td>
<td>87</td>
<td>85</td>
<td>85</td>
<td>87</td>
<td>89</td>
<td>89</td>
</tr>
<tr>
<td>1.4σ</td>
<td>80%</td>
<td>81</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>80</td>
<td>81</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>91</td>
<td>88</td>
<td>87</td>
<td>88</td>
<td>88</td>
<td>89</td>
<td>91</td>
</tr>
<tr>
<td>1.2σ</td>
<td>80%</td>
<td>82</td>
<td>77</td>
<td>77</td>
<td>77</td>
<td>76</td>
<td>78</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>91</td>
<td>88</td>
<td>88</td>
<td>87</td>
<td>88</td>
<td>89</td>
<td>89</td>
</tr>
<tr>
<td>1.0σ</td>
<td>80%</td>
<td>84</td>
<td>82</td>
<td>79</td>
<td>78</td>
<td>77</td>
<td>76</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>93</td>
<td>92</td>
<td>90</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>87</td>
</tr>
<tr>
<td>0.8σ</td>
<td>80%</td>
<td>82</td>
<td>79</td>
<td>76</td>
<td>77</td>
<td>78</td>
<td>76</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>91</td>
<td>88</td>
<td>87</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>89</td>
</tr>
<tr>
<td>0.6σ</td>
<td>80%</td>
<td>83</td>
<td>79</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>80</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>92</td>
<td>90</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>90</td>
</tr>
<tr>
<td>0.4σ</td>
<td>80%</td>
<td>84</td>
<td>79</td>
<td>77</td>
<td>79</td>
<td>80</td>
<td>78</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>92</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>89</td>
</tr>
</tbody>
</table>
Illustration: lead level in the blood

- 70 small bins; 20 cubic B-splines; 3rd order penalty.
- Thick solid: the Bayesian estimate.
- Thin solid: the 90% pointwise credible interval computed from a chain of length $M = 50,000$.
- Unimodality was forced in the Bayesian approach.
- Results:

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Pr($Y &gt; 30$)</th>
<th>$\mu_Y$</th>
<th>$\sigma_Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20</td>
<td>0.14</td>
<td>21.8</td>
<td>8.3</td>
</tr>
<tr>
<td>0.80</td>
<td>(0.10, 0.19)</td>
<td>(20.6, 23.0)</td>
<td>(7.3, 9.6)</td>
</tr>
</tbody>
</table>
Bivariate setting

- Setting: observation of pairs of grouped data for \((X,Y)\)
  \[ \Rightarrow \text{contingency table with observed freq.} \{m_{j_1j_2}\}_{1..J_1}^{1..J_2} \text{ and unknown prob. } \Gamma = \{\gamma_{j_1j_2}\}_{1..J_1}^{1..J_2}. \]

What are the bivariate densities compatible with this 3D histogram?
Observed big cells and latent distribution

Latent (small cell) probabilities: $\Pi = \{\pi_{i_1i_2}\}$; Marginal (big bin) probabilities: $\{\gamma_{j_1+}\}, \{\gamma_{+j_2}\}$
Bayesian P-splines and CLM for the margins

• Let \( \Pi = \{\pi_{i_1i_2}\} \) be the latent distribution for a grid of \( I_1 \times I_2 \) narrow rectangles.

• We use Bayesian P-splines for the marginal latent distributions:

\[
\begin{align*}
\pi_{i_1+} &= \frac{e^{\eta_{i_1}}}{(e^{\eta_1} + \ldots + e^{\eta_{I_1}})} \quad \text{with} \quad \eta = B^{(1)} \phi^{(1)} \\
\pi_{+i_2} &= \frac{e^{\eta_{i_2}}}{(e^{\eta_1} + \ldots + e^{\eta_{I_2}})} \quad \text{with} \quad \eta = B^{(2)} \phi^{(2)}
\end{align*}
\]

where \( B^{(1)} \) and \( B^{(2)} \) denote the B-splines basis for margins 1 and 2.

• The marginal big bin probabilities are related to the latent margins with the CLM:

\[
\begin{pmatrix}
\gamma_{1+} \\
\vdots \\
\gamma_{J_1+}
\end{pmatrix}
= C_1 \begin{pmatrix}
\pi_{1+} \\
\vdots \\
\pi_{I_1+}
\end{pmatrix}
= C_1 \Pi \begin{pmatrix}
1 \\
\vdots \\
1
\end{pmatrix}
\]

where \( \Pi = \begin{pmatrix}
\pi_{11} & \ldots & \pi_{1I_2} \\
\vdots & \vdots & \vdots \\
\pi_{I_11} & \ldots & \pi_{I_1I_2}
\end{pmatrix} \quad \Pi \quad \begin{pmatrix}
\pi_{1+} & \ldots & \pi_{+I_2}
\end{pmatrix} \quad C_2' = \begin{pmatrix}
1 & \ldots & 1
\end{pmatrix} \Pi \quad C_2'
\]

\[
\begin{pmatrix}
\gamma_{1+} & \ldots & \gamma_{J_1+}
\end{pmatrix}
= \begin{pmatrix}
\pi_{1+} & \ldots & \pi_{+I_2}
\end{pmatrix} \quad C_2' = \begin{pmatrix}
1 & \ldots & 1
\end{pmatrix} \quad C_2'
\]
Dependence structure: copulas

• Fact: if \( F(x) \) and \( G(y) \) are the marginal c.d.f.’s of \( X \) and \( Y \), then there exists an (unknown) copula \( C \) such that

\[
\text{Joint c.d.f. : } H(x, y) = \Pr(X \leq x, Y \leq y) = C(F(x), G(y))
\]

\[
\text{Joint density : } h(x, y) = C''(F(x), G(y)) f(x) g(y)
\]

• Assumption: the unknown copula \( C \) can be reasonably approximated by a parametric copula \( C_\tau \) (Frank, Clayton, Gumbel, Gaussian, \ldots ).
Model and likelihood

• Probability to be within the ‘big’ rectangle \((k, l)\):

\[
\gamma_{kl} = C_\tau \left( \sum_{s=1}^{k} \gamma_{s+}, \sum_{t=1}^{l} \gamma_{+t} \right) + C_\tau \left( \sum_{s=1}^{k-1} \gamma_{s+}, \sum_{t=1}^{l-1} \gamma_{+t} \right) - C_\tau \left( \sum_{s=1}^{k} \gamma_{s+}, \sum_{t=1}^{l} \gamma_{+t} \right) - C_\tau \left( \sum_{s=1}^{k-1} \gamma_{s+}, \sum_{t=1}^{l-1} \gamma_{+t} \right)
\]

• Sampling distribution:

\[
(M_{11}, \ldots, M_{J1J2} | \gamma) \sim \text{Mult}(m++; \gamma_{11}, \ldots, \gamma_{J1J2})
\]

• Log-likelihood:

\[
\log L(\phi^{(1)}, \phi^{(2)}, \tau) = \sum_{j_1} \sum_{j_2} m_{j_1j_2} \log \gamma_{j_1j_2}
\]

• The posterior further involves one roughness penalty prior per margin.

• Sampling from the joint posterior: blockwise Metropolis-Hastings.

• Next slides: simulation study:
  
  → Margin 1: Beta(3,5) ; Margin 2: 0.6 Beta(3,10) + 0.4 Beta(12,8)
  
  → Copula with small, medium or large Kendall’s tau.
  
  → \(S = 100\) datasets of size \(n = 1000\) ; 5 or 10 big bins.

  → Joint estimation of the spline and dependence parameters.
5 big bins – **CLAYTON** – Moderate dependence

**True copula: clayton(τ=0.5)**

- **MISE**
- **DIC**
- **Estimated Kendall’s τ**

Clayton Frank Gumbel

0.42 0.44 0.46 0.48 0.50 0.52 0.54

- **Assumed copula**

P. Lambert © - Université de Liège, Belgique

IAP Ghent, September 2008 - Slide 25
5 big bins – FRANK – Moderate dependence

True copula: Frank(\(\tau=0.5\))

- **MISE**
  - Clayton
  - Frank
  - Gumbel

- **DIC**

- **Estimated Kendall's \(\tau\)**
  - Clayton
  - Frank
  - Gumbel

- **True copula: Frank(\(\tau=0.5\))**

- **Assumed copula**

P. Lambert © - Université de Liège, Belgique

IAP Ghent, September 2008 - Slide 26
5 big bins – GUMBEL – Moderate dependence

True copula: gumbel(\(\tau=0.5\))

- MISE
- DIC
- Estimated Kendall's \(\tau\)

Assumed copula

Clayton Frank Gumbel

0 1 2 3 4 5 6

True copula: gumbel(\(\tau=0.5\))

Assumed copula

Clayton Frank Gumbel

0.42 0.44 0.46 0.48 0.50 0.52 0.54

True copula: gumbel(\(\tau=0.5\))

Assumed copula

Clayton Frank Gumbel

0.42 0.44 0.46 0.48 0.50 0.52 0.54

True copula: gumbel(\(\tau=0.5\))

Assumed copula

Clayton Frank Gumbel

0.42 0.44 0.46 0.48 0.50 0.52 0.54

P. Lambert © - Université de Liège, Belgique

IAP Ghent, September 2008 - Slide 27
5 big bins – CLAYTON – Large dependence

True copula: clayton(τ=0.7)

MISE
Assumed copula

DIC
Assumed copula

Estimated Kendall’s τ
Assumed copula

P. Lambert © - Université de Liège, Belgique

IAP Ghent, September 2008 - Slide 28
5 big bins – **FRANK** – Large dependence

**True copula: Frank(\(tau=0.7\))**

- MISE
- DIC
- Estimated Kendall's \(\tau\)

Assumed copula:
- Clayton
- Frank
- Gumbel

---

P. Lambert © - Université de Liège, Belgique  
IAP Ghent, September 2008 - Slide 29
5 big bins – GUMBEL – Large dependence

**True copula: gumbel(\(\tau=0.7\))**

- **MISE**
  - Clayton
  - Frank
  - Gumbel

- **DIC**
  - Clayton
  - Frank
  - Gumbel

- **Estimated Kendall's \(\tau\)**
  - Clayton
  - Frank
  - Gumbel

---

P. Lambert © - Université de Liège, Belgique

IAP Ghent, September 2008 - Slide 30
Final comments

- The Bayesian approach enables to quantify all sources of uncertainty (cfr. choice of the penalty parameters, joint estimation of marginal and copula parameters...).

This is a major advantage over the frequentist approach.

- Simulations suggest that credible intervals for quantiles have a coverage close to their nominal level.

- The method is not limited to “histogram” data: it can also deal with interval-censored data (joint work with Aysun Çetinyürek at ULg).

- We have also developed a Bayesian 2-D composite link model.

  - It does not require to choose an arbitrary parametric copula.
  - It involves many more parameters.
Some references


http://www.statsoc.ulg.ac.be/