Centile estimation using GAMLSS

Mikis Stasinopoulos\textsuperscript{1}  Bob Rigby\textsuperscript{1}

\textsuperscript{1}STORM, London Metropolitan University

GAMLSS: Short course Cordoba, Argentina, Oct 2012

1
1 Centiles
   - The Dutch boys data
   - The problem
   - The methods
   - The LMS model and its extensions
   - Estimation of the smoothing parameters

2 GAMLSS functions for centile estimation

3 Further Work

4 Conclusions
The Dutch boys data

head : the head circumference of 7040 boys
age  : the age in years

Source: Buuren and Fredriks (2001)
The Dutch boys data
The Dutch boys data with centiles
Centile estimation using GAMLSS

Centiles

The problem

The Dutch boys data: centiles
The methods

i) the non parametric approach of quantile regression (Koenker, 2005; Koenker and Bassett, 1978)

Centile estimation using GAMLSS

Centiles

The LMS model and its extensions

**Centiles: the model**

\[ Y \sim f_Y(y|\mu, \sigma, \nu, \tau) \text{ where } f_Y() \text{ is any distribution} \]

\[ Y \text{ = head circumference and } X = AGEx \]

\[ \mu = s(x, df_\mu) \]
\[ \log(\sigma) = s(x, df_\sigma) \]
\[ \nu = s(x, df_\nu) \]
\[ \log(\tau) = s(x, df_\tau) \]
The LMS method and extensions

Let \( Y \) be a random variable with range \( Y > 0 \) defined through the transformed variable \( Z \) given by:

\[
Z = \frac{1}{\sigma \nu} \left[ \left( \frac{Y}{\mu} \right)^\nu - 1 \right], \quad \text{if } \nu \neq 0
\]

\[
= \frac{1}{\sigma} \log \left( \frac{Y}{\mu} \right), \quad \text{if } \nu = 0.
\]

1. if \( Z \sim N(0, 1) \) then \( Y \sim BCCG(\mu, \sigma, \nu) = \text{LMS method} \)
2. if \( Z \sim t_\tau \) then \( Y \sim BCT(\mu, \sigma, \nu, \tau) = \text{LMST method} \)
3. if \( Z \sim PE(0, 1, \tau) \) then \( Y \sim BCPE(\mu, \sigma, \nu, \tau) = \text{LMSP method adopted by WHO} \)
We need to select the five values $df_{\mu}, df_{\sigma}, df_{\nu}, df_{\tau}, \xi$

- by trial and error
- minimize the generalized Akaike information criterion, $\text{GAIC}(\#)$
- minimize the validation global deviance $\text{VDG}$
- using local selection criteria, i.e. CV, ML

Diagnostics should be used in all above cases
For different values of $\#$ minimize GAIC($\#$) to find estimates for $df_\mu, df_\sigma, df_\nu, df_\tau, \xi$

- $\# = 2$, AIC
- $\# = 3$, our preference
- $\# = 3.84$, $\chi^2$
- $\# = \log(n)$, BIC, SBC

The `gamlss` functions `find.hyper()` can be used
Centile estimation using GAMLSS

Centiles
Estimation of the smoothing parameters

**Centiles: Validation Global Deviance**

Divide the data into two components 60% and 40%

Use the first component for training (fitting)
Use the second component for estimating the parameters

\[ df_\mu, df_\sigma, df_\nu, df_\tau, \xi \]

The gamlss functions \texttt{VDG()} in combination of the R function \texttt{optim()} can be used
Centile estimation using GAMLSS

Centiles

Estimation of the smoothing parameters

Centiles: Local selection criteria

- GCV
- GAIC
- ML
- EM

The gamlss smoothing functions pb() uses ML as default
The gamlss smoothing functions ga() is an interface for the gam() function of Simon Wood and therefore uses GCV.
GAMLSS functions for centile estimation

- `centiles()` to plot centile curves against an x-variable.
- `centiles.com()` to compare centiles curves for more than one object.
- `centiles.split()` as for `centiles()`, but splits the plot at specified values of x.
- `centiles.fan()` fan plot as in `centiles()`.
- `centiles.pred()` to predict and plot centile curves for new x-values.
- `fitted.plot()` to plot fitted values for all the parameters against an x-variable.
The head circumference data: results

<table>
<thead>
<tr>
<th>Method</th>
<th>$df_\mu$</th>
<th>$df_\sigma$</th>
<th>$df_\nu$</th>
<th>$df_\tau$</th>
<th>$\xi$</th>
<th>GAIC(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAIC(3)</td>
<td>12.3</td>
<td>5.7</td>
<td>2</td>
<td>2</td>
<td>0.33</td>
<td>26811.7</td>
</tr>
<tr>
<td>VGD</td>
<td>15.8</td>
<td>8.1</td>
<td>2</td>
<td>2</td>
<td>0.28</td>
<td>26815.6</td>
</tr>
<tr>
<td>local ML</td>
<td>13.0</td>
<td>4.9</td>
<td>2</td>
<td>2</td>
<td>0.30</td>
<td>26811.1</td>
</tr>
</tbody>
</table>

**Table:** The estimated degrees of freedom
The head circumference data: `fitted.plot()`

(a) 

(b) 

(c) 

(d)
The head circumference data: comparison of the centiles, centiles.com()
The head circumference data: The worm plots, \texttt{wp()}

![Graph showing worm plots for head circumference data.](image-url)
The head circumference data: `centiles()`
The head circumference data: `centiles.fan()`
The head circumference data: `centiles.split()`
## The head circumference data: The Q statistics

<table>
<thead>
<tr>
<th>Range</th>
<th>Z1</th>
<th>Z2</th>
<th>Z3</th>
<th>Z4</th>
<th>Agostino</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02500 to 0.24499</td>
<td>0.09</td>
<td>0.85</td>
<td>1.21</td>
<td>0.02</td>
<td>1.47</td>
<td>477</td>
</tr>
<tr>
<td>0.24499 to 0.75499</td>
<td>-0.48</td>
<td>-1.18</td>
<td>-0.58</td>
<td>-0.25</td>
<td>0.40</td>
<td>473</td>
</tr>
<tr>
<td>0.75499 to 1.255</td>
<td>0.88</td>
<td>-2.27</td>
<td>-0.24</td>
<td>-1.37</td>
<td>1.94</td>
<td>467</td>
</tr>
<tr>
<td>1.255 to 1.895</td>
<td>0.01</td>
<td>2.62</td>
<td>1.96</td>
<td>2.76</td>
<td>11.47</td>
<td>460</td>
</tr>
<tr>
<td>1.895 to 2.945</td>
<td>-0.29</td>
<td>0.72</td>
<td>-1.70</td>
<td>-0.01</td>
<td>2.90</td>
<td>473</td>
</tr>
<tr>
<td>2.945 to 5.535</td>
<td>-0.35</td>
<td>-0.40</td>
<td>-0.72</td>
<td>0.88</td>
<td>1.30</td>
<td>466</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.835 to 17.185</td>
<td>0.45</td>
<td>0.00</td>
<td>-0.59</td>
<td>-0.08</td>
<td>0.35</td>
<td>466</td>
</tr>
<tr>
<td>17.185 to 18.675</td>
<td>0.09</td>
<td>0.32</td>
<td>1.01</td>
<td>1.48</td>
<td>3.23</td>
<td>472</td>
</tr>
<tr>
<td>18.675 to 21.685</td>
<td>-0.54</td>
<td>-0.65</td>
<td>-0.88</td>
<td>-0.42</td>
<td>0.96</td>
<td>467</td>
</tr>
<tr>
<td>TOTAL Q stats</td>
<td>2.85</td>
<td>16.90</td>
<td>15.43</td>
<td>18.04</td>
<td>33.46</td>
<td>7040</td>
</tr>
<tr>
<td>df for Q stats</td>
<td>1.95</td>
<td>12.01</td>
<td>12.35</td>
<td>13.00</td>
<td>25.35</td>
<td>0</td>
</tr>
<tr>
<td>p-val for Q stats</td>
<td>0.23</td>
<td>0.15</td>
<td>0.24</td>
<td>0.16</td>
<td>0.13</td>
<td>0</td>
</tr>
</tbody>
</table>
The head circumference data: `Q.stats()`

<table>
<thead>
<tr>
<th>Q–Statistics</th>
<th>0.02500 to 0.24499</th>
<th>0.24499 to 0.75499</th>
<th>0.75499 to 1.255</th>
<th>1.255 to 1.895</th>
<th>1.895 to 2.945</th>
<th>2.945 to 5.535</th>
<th>5.535 to 9.295</th>
<th>9.295 to 10.605</th>
<th>10.605 to 11.975</th>
<th>11.975 to 13.245</th>
<th>13.245 to 14.515</th>
<th>14.515 to 15.835</th>
<th>15.835 to 17.185</th>
<th>17.185 to 18.675</th>
<th>18.675 to 21.685</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z1</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Z2</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Z3</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Z4</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>
Comparing GAMLSS and QR: fitted centiles
Centile estimation using GAMLSS
GAMLSS functions for centile estimation

Discrete distributions: fitted centiles

Centile curves using SICHEL

![Graph showing centile curves using SICHEL.](image)
Further work

- model selection
- time series models
- peak oil
Conclusions: GAMLSS

- **Unified** framework for univariate regression type of models
- Allows **any** distribution for the response variable $Y$
- Models **all** the parameters of the distribution of $Y$
- Allows a variety of **additive terms** in the models for the distribution parameters
- The fitted **algorithm** is modular, where different components can be added easily
- Models can be fitted easily and fast
- Explanatory tool to find appropriate set of models
- It deals with **overdispersion, skewness and kurtosis**


End

for more information see

www.gamlss.org
Time series and GAMLSS

What is available:

- Modelling $\mu$
  - ARIMA (for normal errors)
  - Structural time series models (for normal errors)
  - GARMA and others (for exponential family)
  - Hidden Markov Chain models

- Modelling $\sigma$
  - GARCH and its by product (for normal and long tails)
  - Stochastic volatility models
  - Markov-Switching Multifractal models

What is suitable for GAMLSS?
What is our experience?

- We have expanded the GARMA model to accept any `gamlss.family` distributions
- Modelling $\mu$ only
- Not clear how lags should define in the case of $\nu$ and $\tau$
- No explicit stochastic model
A Latent-State Generalized Additive Model for Location, Scale and Shape for Multivariate Pollutant Concentrations

Antonello Maruotti

Università di Roma Tre, Roma, Italy.

Walter Zucchini

Institut für Statistik und Ökonometrie, Göttingen, Germany.

Summary. XXX YYY ZZZ

Keywords: GAMLSS, latent-state models, multivariate time series, pollutant concentrations.

1. Introduction

The model developed in this paper was motivated by the analysis of a daily multivariate time series of pollutant concentrations taken at a monitoring station in Rome, Italy. The objective is to quantify the relationship between pollutant levels and the meteorological factors that affect them.

The study and monitoring of environmental phenomena often involves the analysis of large multivariate time series. The growing interest in air quality has led to a substantial literature. One branch of this is concerned with synthetic environmental indices designed to summarize a series of measurements (Barnett and Brown, 2002; Sahana et al., 2005; Bellini, 2007).
Markov-Switching Multifractal Models

- We have used them to model $\sigma$
- Very slow especially larger than 7 number of stages
- We have not find them to be an improvement compare to other models
- There is explicit stochastic model
Oil returns
Oil returns

a) GAMLSS

b) MSM

b) GARCH(1,1)
Structural/Stochastic Volatility models

- A proper stochastic model exist
- This is the most promising approach at the moment
- It uses sparse matrices (rather than Kalman filter)
- There are difficulties to estimate the smoothing parameters but we have done good progress
Simulated negative binomial type I

\[
g_1(\mu_t) = \eta_{\mu,t} = \eta_{\mu,t-1} + b_{\mu,t}
\]

\[
g_2(\sigma_t) = \eta_{\sigma,t} = \eta_{\sigma,t-1} + b_{\sigma,t}
\]

(1)
Negative binomial fit for the mean
Negative binomial fit for sigma
GEST = GAMLSS + structural models

The GEST approach looks very promising
"Peak Oil is the point in time when the maximum rate of petroleum extraction is reached, after which the rate of production is expected to enter terminal decline", (Wikipedia).
Figure 2: Discrete scenarios based on finite number of uncertainties\textsuperscript{23}
ACEGES models

- ACEGES Stands for Agent-based Computational Economics of the Global Energy System.
- It is an agent-based modelling simulation
- It tries to understand the outlook for oil production by accounting for uncertainties in resource estimates, demand growth, production growth and peak/decline point.
- People: Vlasis Voudouris, Ken’ichi Matsumoto, Bob Rigby, Paul Eilers, Michael Jefferson, John Sedgwick, Carlo Di Maio
- Interests: Oil, Gas, Electricity, Renewable energy
Peak Oil: scenario one
Peak Oil: scenario two
Peak Oil: estimated ultimate recovery EUR
ACEGES: What would have been the trajectory 30 years ago!
Peak Oil

- GAMLSS is used as a by product here
- ACEGES: statistical improvements with the simulations
- Can we use statistical inference to make the system self-learning?
### Selection of explanatory variables

#### How to select explanatory variables?

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\nu$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tau$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Strategy A**
- **Strategy B**
- **Other strategies?**
- **Boosting**
strategy A

Strategy A:

- It starts with a **forward stepwise** selection using GAIC.
- Each $x$ variables is set for selection first for $\mu$ then for $\sigma$, $\nu$ and $\tau$
- then it does a **backward** elimination for $\nu$, $\sigma$ and $\mu$.

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$\times$</td>
<td></td>
<td>$\times$</td>
<td>$\times$</td>
<td></td>
<td>$\times$</td>
</tr>
<tr>
<td>$\sigma$</td>
<td></td>
<td>$\times$</td>
<td>$\times$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\nu$</td>
<td>$\times$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tau$</td>
<td></td>
<td></td>
<td>$\times$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**strategy B**

**Strategy B:** forward stepwise selection using GAIC in which an $x$ variable is selected for all the parameters

**Table:** selecting explanatory variables

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$\times$</td>
<td></td>
<td>$\times$</td>
<td></td>
<td>$\times$</td>
<td></td>
</tr>
<tr>
<td>$\sigma$</td>
<td>$\times$</td>
<td></td>
<td>$\times$</td>
<td></td>
<td></td>
<td>$\times$</td>
</tr>
<tr>
<td>$\nu$</td>
<td>$\times$</td>
<td></td>
<td></td>
<td>$\times$</td>
<td></td>
<td>$\times$</td>
</tr>
<tr>
<td>$\tau$</td>
<td>$\times$</td>
<td></td>
<td>$\times$</td>
<td></td>
<td></td>
<td>$\times$</td>
</tr>
</tbody>
</table>
Question one: boosting

GAMLSS for high-dimensional data – a flexible approach based on boosting

Andreas Mayr\textsuperscript{1}, Nora Fenske\textsuperscript{2}, Benjamin Hofner\textsuperscript{1}, Thomas Kneib\textsuperscript{3}, Matthias Schmid\textsuperscript{1}

\textsuperscript{1} Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
\textsuperscript{2} Institut für Statistik, Ludwig-Maximilians-Universität München, Germany
\textsuperscript{3} Institut für Mathematik, Carl von Ossietzky Universität Oldenburg, Germany