A New Method for Analyzing Time Intensity Curves

Moyi Li\textsuperscript{1}, John C. Castura\textsuperscript{2}, Ryan P. Browne\textsuperscript{1} and Paul D. McNicholas\textsuperscript{1}

\textsuperscript{1}Department of Mathematics & Statistics, University of Guelph, ON, Canada

\textsuperscript{2}Conpusense Inc, Guelph, ON, Canada

July 12 2012
A novel Modelling approach is introduced and parameters are estimated via an EM algorithm. Smoothing splines are also aggregated.

Four simulations are performed on simulated data; we obtain fitted curves based on the assumptions of homoscedastic and heteroscedastic error terms, respectively, at each time point.

Real fruit liqueur data are analyzed.

Discussion and suggestions for future work.
Aim

- To estimate underlying time intensity curves and cluster individuals.

- How it can help us to discover useful information about attributes.
TI curves are monotonically increasing until time $T_{\text{max}}$ and then monotonically decreasing thereafter.

We represent this dependence with a Markovian error term.

Figure 1: TI Curves
Modelling Framework

- Let $z_i$ be the observed TI value and $x_i$ be the latent TI value so that
  \[
  z_i = \begin{cases} 
  \max\{x_{i-1}, x_i\} & \text{for } i = 2, \ldots, k, \\
  \min\{x_{i-1}, x_i\} & \text{for } i = k + 1, k + 2, \ldots, n.
  \end{cases}
  \]

  where $X_i \sim N(\mu_i, \sigma_i^2)$, $k$ is $T_{\text{max}}$ and $n$ is total time points.

For the Markovian error term, we consider two options:

1. **Homoscedasticity**: there is a common standard deviation across all time points for all panellists, i.e., $\sigma_i^2 = \sigma^2$, for $i = 1, 2, \ldots, n$.

2. **Heteroscedasticity**: each time point has its own standard deviation.
The complete-data log-likelihood function using the homoscedastic $\sigma$ is

$$L(\mu_1, \ldots, \mu_n, \sigma | z_1, \ldots, z_n) = -\frac{np}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{p} (X_{ij} - \mu_i)^2 | Z),$$

The complete-data log-likelihood function using the heteroscedastic $\sigma_i$ is

$$L(\mu_1, \ldots, \mu_n, \sigma_1, \ldots, \sigma_n | z_1, \ldots, z_n) = -\frac{np}{2} \log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2} \sum_{i=1}^{n} \sum_{j=1}^{p} (X_{ij} - \mu_i)^2 | Z).$$
The EM algorithm is an iterative method for finding maximum likelihood estimates of parameters where there are unobserved or missing data.

An expectation (E-) step that computes the expectation of the complete-data log-likelihood given the current estimates is followed by a maximization (M-) step wherein the expectation of the complete-data log-likelihood is maximized with respect to the model parameters.

The E- and M-steps are iterated until convergence.
Truncated Normal Distribution

- $X \mid Z \sim \text{truncated } \mathcal{N}(\mu_i, \sigma^2)$
- $X \mid Z \sim \text{truncated } \mathcal{N}(\mu_i, \sigma_i^2)$

so, the expectation

$$
E(X \mid a < Z < b) = \mu + \frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \sigma,
$$

and the variance

$$
\text{Var}(X \mid a < Z < b) = \sigma^2 \left[ 1 + \frac{\frac{a-\mu}{\sigma} \phi\left(\frac{a-\mu}{\sigma}\right) - \frac{b-\mu}{\sigma} \phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} - \left(\frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}\right)^2 \right]
$$
E step

\[
E(X_i \mid z_{i-1}, z_i) = \begin{cases} 
Z_i & \text{if } z_i > z_{i-1}, \\
\mu_i - \frac{\phi(z_i - \mu_i)}{\Phi(z_i - \mu_i)} \sigma & \text{if } z_i = z_{i-1}, 
\end{cases}
\]

for \( i = 2, \ldots, k \) and

\[
E(X_i \mid z_{i-1}, z_i) = \begin{cases} 
Z_i & \text{if } z_{i-1} > z_i, \\
\mu_i + \frac{\phi(z_i - \mu_i)}{\Phi(z_i - \mu_i)} \sigma & \text{if } z_{i-1} = z_i, 
\end{cases}
\]

for \( i = k + 1, \ldots, n \).
M step

Under the homoscedastic assumption, the expected value of the complete-data log-likelihood is given by

$$Q_1(X, Z | \mu, \sigma^2) = -\frac{np}{2} \log(2\pi \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{p} E\{(X_{ij} - \mu_i)^2 | Z\},$$

where $p$ is number of repetitions, $n$ is number of time points and $\mu$ is a $n \times 1$ matrix.

$$\hat{\mu}_i = \frac{1}{p} \sum_{i=1}^{p} E\{X_{ij} | Z\} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} E\{(X_{ij} - \mu_i)^2 | Z\}.$$
For the second assumption-heteroscedastic $\sigma$, the expected value of the complete-data log-likelihood function is

$$Q_2(\mathbf{X}, \mathbf{Z} \mid \mu, \sigma^2) = -\frac{p}{2} \sum_{i=1}^{n} \log \sigma_i^2 - \sum_{i=1}^{n} \frac{1}{2\sigma_i^2} \sum_{j=1}^{p} E \left\{ (X_{ij} - \mu_i)^2 \mid \mathbf{Z} \right\} + C,$$

where $C$ is a constant.

$$\hat{\mu}_i = \frac{1}{p} \sum_{i=1}^{p} E\{X_{ij} \mid \mathbf{Z}\} \quad \text{and} \quad \hat{\sigma}_i^2 = \frac{1}{p} \sum_{j=1}^{p} E \left\{ (X_{ij} - \mu_i)^2 \mid \mathbf{Z} \right\}.$$
Smoothing Spline: why

Figure 2: Fitted Curve
The penalized spline smoothing was introduced by O’Sullivan (1986).

This smoothing method with flexible choice of bases and penalties can be viewed as a compromise between regression and smoothing splines which are piecewise polynomials with pieces smoothly connected together.
Let \((x_i, Y_i)\), so that \(x_1 < x_2 < \cdots < x_n\), be a sequence of observations modelled by the relation \(Y_i = \mu(x_i)\). The penalized sum of squares is

\[
S(\mu) = \sum_{i=1}^{n} (Y_i - \mu(x_i))^2 + \lambda \int_{a}^{b} \mu''(x)^2 dx,
\]

\(\mu\) is any twice-differentiable function on \([a, b]\) and \(\lambda\) is a smoothing parameter.

The first term measures the closeness of the fitted function to the data, while the second penalizes the curvature in the function.

The smoothing spline estimate \(\hat{\mu}\) of the function \(\mu\) is

\[
\hat{\mu} = \arg \min_{\mu \in \mu} S(\mu).
\]
Preparation

1. Randomly generate the latent TI $x_i$ values which follow a normal distribution with parameters $\mu_i$ and $\sigma = 0.01$.

2. A straightforward method to generate observed data $z_1, \ldots, z_n$ is given below:

$$z_1 = x_1, z_2 = \max(x_1, x_2), \ldots, z_{k-1} = \max(x_{k-2}, x_{k-1}), z_k = \max(x_{k-1}, x_k),$$

$$z_{k+1} = \min(x_k, x_{k+1}), \ldots, z_{n-1} = \min(x_{n-2}, x_{n-1}), z_n = \min(x_{n-1}, x_n),$$

where $n = 51$. 
Figure 3: generate TI curves
Simulation Study

Simulation Results: Homoscedastic Model

When $\sigma = 0.01$

Figure 4: Fitted Curve

Figure 5: Smooth Curve
Simulation Results: Homoscedastic Model

When $\sigma = 0.03$

**Figure 6: Fitted Curve**

**Figure 7: Smooth Curve**
Simulation Results: Heteroscedastic Model

When $\sigma = 0.01$

*Figure 8: Fitted Curve*

*Figure 9: Smooth Curve*
Simulation Results: Heteroscedastic Model

When $\sigma = 0.03$

Figure 10: Fitted Curve

Figure 11: Smooth Curve
Results: Homoscedastic Model

Figure 12: Smooth curves for product A
Figure 13: Smooth curves for product A
Clustering

- Group 1: panelist 1, 3, 10
- Group 2: panelist 2, 5, 6, 7, 12
- Group 3: panelist 9, 11
- Group 4: panelist 4
- Group 5: panelist 8
Results: Homoscedastic Model

Figure 14: Smooth curves for product B
Product A vs. Product B: Homoscedastic Model

Figure 15: Smooth Curves for 12 panelists

Figure 16: Smooth Curves for 12 panelists
Fruit Liqueur Data

Product A vs. Product B: Homoscedastic Model

Figure 17: Smooth Curves between product A and B for each panelist
Using different assumptions, the smoothing curves have similar shapes and are a representation of 3 TI curves.

Recommending using homoscedastic $\sigma$ obtain smooth TI curves.

There is variation among the panelists for product A and product B.

Panelists give very similar smoothing curves between product A and B.
Future Work

In the future, the problem of dealing with $T_{max}$. Because it is the crucial part of conducting a accurately fitted curve.


Echols, S., Lakshmanan, A., Mueller, S., Rossi, F. and Thomas, A. (2003), ‘Parametric modeling of time intensity data collected on product...