A New Method for Analyzing Time Intensity Curves

Moyi Li[1], John C. Castura[2], Ryan P. Browne[1] and Paul D. McNicholas[1]

[1]Department of Mathematics & Statistics, University of Guelph, ON, Canada

[2]Conpusense Inc, Guelph, ON, Canada

July 12 2012

Outline

- 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.

Modelling Framework

- TI curves are monotonically increasing until time $T_{\max}$ and then monotonically decreasing thereafter.

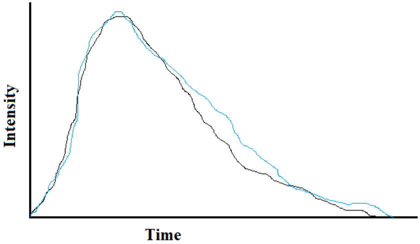

Figure 1: TI Curves

- We represent this dependence with a Markovian error term.

Modelling Framework

- Let z_i be the observed TI value and x_i be the latent TI value so that

$$z_i = \begin{cases} \max\{x_{i-1}, x_i\} & \text{for} \quad i = 2, \ldots, k, \\ \min\{x_{i-1}, x_i\} & \text{for} \quad i = k+1, k+2, \ldots, n. \end{cases}$$

where $X_i \backsim \mathcal{N}(\mu_i, \sigma_i^2)$, k is $T_{\max}$ 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.

Modelling Framework

- The complete-data log-likelihood function using the homoscedastic σ is

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

- The complete-data log-likelihood function using the heteroscedastic σ_i is

$$\mathcal{L}(\mu_1, \dots, \mu_n, \sigma_1, \dots, \sigma_n \mid z_1, \dots, z_n) = -\frac{np}{2}\log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2}\sum_{i=1}^{n}\sum_{j=1}^{p}((X_{ij} - \mu_i)^2 \mid \mathbf{Z}).$$

EM Algorithm

- 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

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

so, the expectation

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

and the variance

$$\mathbb{V}\mathrm{ar}(X \mid a < Z < b) = \sigma^2 \left[1 + \frac{\frac{a-\mu}{\sigma}\phi(\frac{a-\mu}{\sigma}) - \frac{b-\mu}{\sigma}\phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} - \left(\frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right)^2 \right]$$

E step

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

for $i = 2, \ldots, k$ and

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

for $i = k + 1, \ldots, n$.

M step

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

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

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

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

M step

For the second assumption-heteroscedastic σ, the expected value of the complete-data log-likelihood function is

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

where C is a constant.

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

Smoothing Spline: why

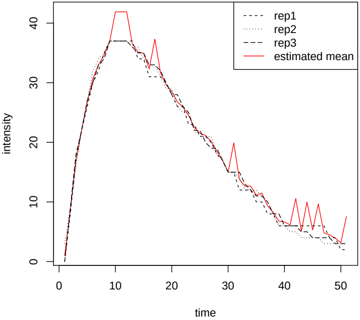

Figure 2: Fitted Curve

Smoothing Spline

- 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.

Smoothing Spline

- 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

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

- μ is any twice-differentiable function on $[a, b]$ and λ 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 μ is

$$\hat{\mu} = \underset{\mu \in \mu}{\arg\min}\ S(\mu).$$

Preparation

1. Randomly generate the latent TI x_i values which follow a normal distribution with parameters μ_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 $n = 51$.

Preparation

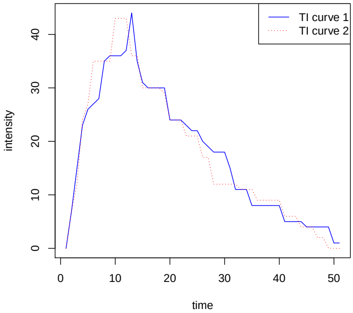

Figure 3: generate TI curves

Simulation Results:Homoscedastic Model

When $\sigma = 0.01$

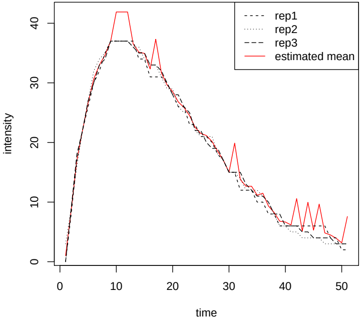

Figure 4: Fitted Curve

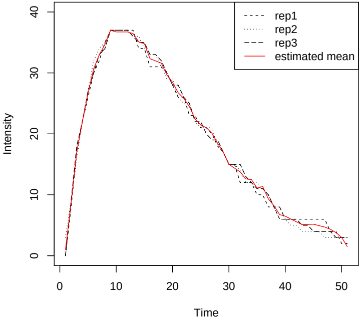

Figure 5: Smooth Curve

Simulation Results:Homoscedastic Model

When $\sigma = 0.03$

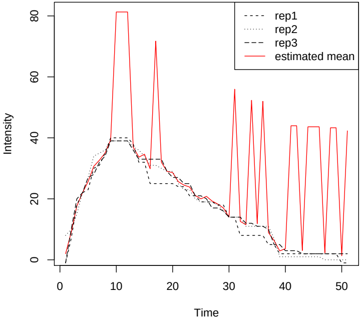

Figure 6: Fitted Curve

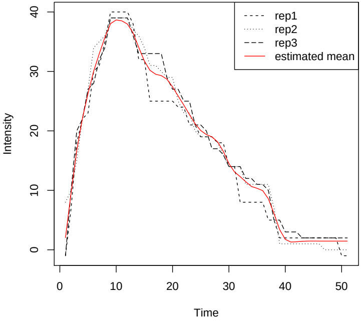

Figure 7: Smooth Curve

Simulation Results:Heteroscedastic Model

When $\sigma = 0.01$

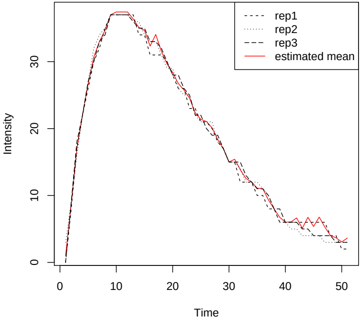

Figure 8: Fitted Curve

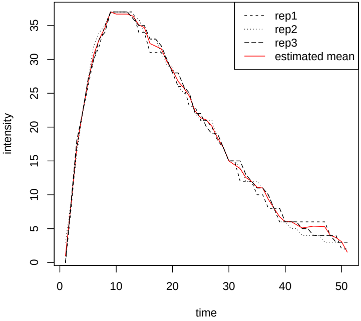

Figure 9: Smooth Curve

Simulation Results:Heteroscedastic Model

When $\sigma = 0.03$

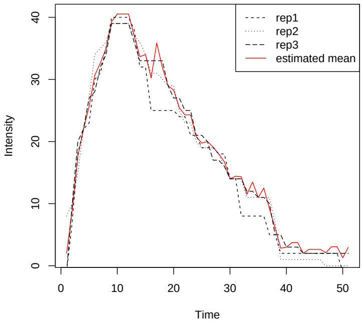

Figure 10: Fitted Curve

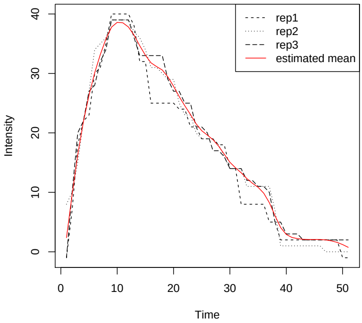

Figure 11: Smooth Curve

Results:Homoscedastic Model

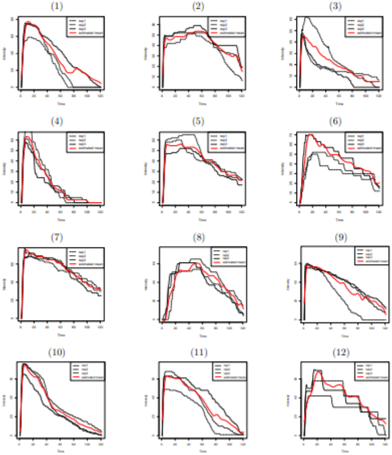

Figure 12: Smooth curves for product A

Results:Heteroscedastic Model

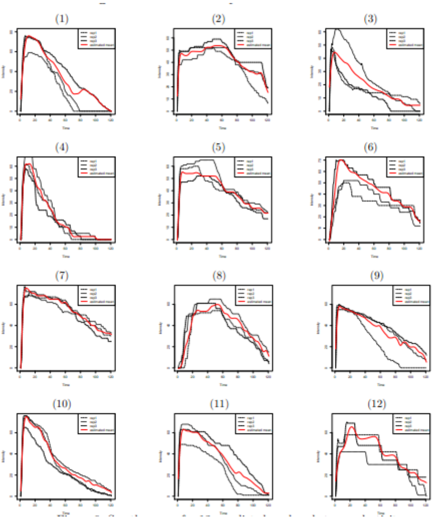

Figure 13: Smooth curves for product A

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

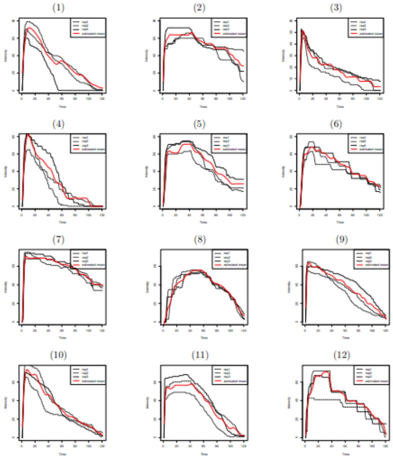

Figure 14: Smooth curves for product B

Product A vs. Product B: Homoscedastic Model

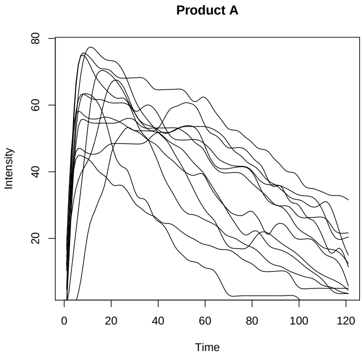

Figure 15: Smooth Curves for 12 panelists

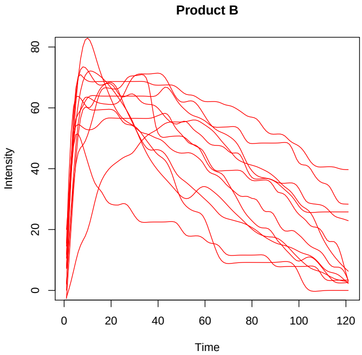

Figure 16: Smooth Curves for 12 panelists

Product A vs. Product B: Homoscedastic Model

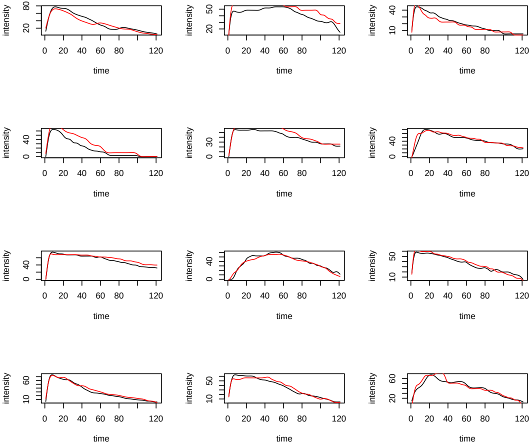

Figure 17: Smooth Curves between product A and B for each panelist

Conclusion

- Using different assumptions, the smoothing curves have similar shapes and are a representation of 3 TI curves.
- Recommending using homoscedastic σ 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.

Bibliography

Amerine, M. A., Pangborn, R. M. and Roessler, E. B. (1965), *The principles of sensory evaluation of food. In: Food Science and Technology Monographs*, Academic Press, New York.

Bloom, K. and Duizer, Land Findlay, C. (1995), 'An objective numerical method of assessing the reliability of time-intensity panelists', *Sensory Studies* **10**, 285–294.

Chaya, C., Perez-Hugalde, C., Judez, L., Wee, C. and Guinard, J. (2004), 'Use of the statis method to analyze time-intensity profiling data', *Food Quality and Preference* **15**(1), 3 – 12.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Royal Statistical Society* **39**(1), 1–38.

Dijksterhuis, G. and Eilers, P. (1997), 'Modelling time-intensity curves using prototype curves', *Food Quality and Preference* **8**(2), 131 – 140.

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