Alternative Methods and Models for Longitudinal Data

Further Statistical Methods, Lecture 10
HT 2007

Steffen Lauritzen, University of Oxford; February 21, 2007

Types of longitudinal data

There are many cases where the 'standard model' from last lecture is inadequate, i.e. when the data are not well described as the sum of three components: a general trend, a (stationary) component with serial correlation, and random noise.

This is for example true for such cases as

- *Biokinetics:* A substance is introduced into a person and the concentration level of one or more components is measured at selected time intervals over a period.

  The 'substance' can e.g. be one or more specific drugs or types of food.

The purpose of such analysis may be to understand the *shape of the curve*, to get a grip of the *duration* of a transient phenomenon, or e.g. the variation in the *maximally achieved value.*

- *Cucumber plants* are grown in greenhouses. One would like to know how different watering/fertilization/treatment schemes affect the growth. Cucumbers are picked daily from each plant and recorded.

  Cucumbers have a season. It takes a while before they develop, then they give a lot of cucumbers for a while, and then stop. The farmer would like to have a lot of cucumbers when others don't, so the price is high.

- *Event history data* follow individuals over time and record when events happen.

- *Flowers* under different conditions. They develop buds, the buds become flowers, and then die. Different treatments make the plants develop differently.

  Plants that have lots of buds and some flowers are selling best.

  This can be seen as a type of event history data.

- *Panel data* follow a group of individuals (panel members) over time. From time to time the members are filling questionnaires, for example on their political or consumer preferences.

- *Growth models.* It is not always reasonable to assume this to be trend plus stationary error. Typically growth can be high in some periods and low in others, with some random variation.

- *Speech analysis.* Frequency properties of speech is recorded at dense discrete time points (millisecond intervals). One is interested in describing the behaviour as different phonemes are pronounced, e.g. for automatic speech recognition and -understanding.

Descriptive methods

Transform an observed curve to a some *features*, e.g.

- The area A under the curve, representing the total amount of something;
- The maximal value M reached of the curve;
- The total duration D of a signal, i.e. the time spent above a certain level.
- A set of Fourier- or wavelet coefficients F;
- etc...

Now use your favourite (multivariate) technique to analyse (part of) the vector A, M, D, F.

Differential equations

If the phenomenon observed is well understood, there might be a relevant differential equation explaining the main features of the observations.

An example from insulin kinetics postulates the following relation between the plasma glucose concentration $G(t)$, insulin concentration $I(t)$, and the insulin's effect on the net glucose disappearance $X(t)$:

$$\begin{aligned}
\dot{G}(t) &= -p_1\{G(t) - G_b\} - X(t)G(t), \quad G(0) = 0, \\
\dot{X}(t) &= -p_2 X(t) + p_3\{I(t) - I_b\}, \quad X(0) = 0, \\
\dot{I}(t) &= -n\{I(t) - I_b\} + \gamma\{G(t) - h\}^{+} t, \quad I(0) = 0.
\end{aligned}$$

This is known as *Bergman's minimal model*.

The parameters are *individual* and to be determined from observations. The important quantities are

- Insulin sensitivity: $S_I = p_1/p_2$;
- Glucose effectiveness: $S_G = p_1$;
- Pancreatic responsiveness: (ϕ_1, ϕ_2) where $\phi_1 = (I_{\max} - I_b)/\{n(G_0 - G_b)\}$, $\phi_2 = \gamma \times 10^4$.

This is generally difficult, as only $G(t), I(t)$ can be observed, and only at discrete time points. Using graphical models and MCMC in the right way, it is possible.

This general area is known as PK/PD for pharmaco-kinetics/-dynamics.

Dynamic models

These models, also known as *state-space models* (SSM) are similar in spirit to differential equation models.

Typically they have two levels, but sometimes more. One level describes the development of an unobserved (hidden) *state* X_t, typically using a Markov model with e.g.

$$\mathcal{L}(X_{t+1} \,|\, X_s = x_s, s \leq t, \theta) \sim \mathcal{N}\{A_t(\theta)x_t, \sigma_t^2(\theta)\}$$

and an *observational model* for Y_t with

$$\mathcal{L}(Y_t \,|\, X, \eta) = \mathcal{N}\{B_t(\eta)x_t, \tau^2(\eta)\},$$

where $Y_t, t = 1, \ldots, T$ are observed.

Parameters are then estimated by using a variant of the EM algorithm. The E-step can be performed elegantly using a recursive algorithm known as the *Kalman Filter*.

MCMC is also a viable alternative and a hot research topic is that of *particle filters* which can be seen as MCMC variants of the Kalman filter.

Generalisations include replacing each of the models above with *generalised linear models*.

For example, in the cucumber example it is natural to consider Poisson model for the observed number of cucumbers on a plant.

In speech analysis, Y is typically a *feature vector* of the signal and the state space equation should depend on what the individual is saying. Hence another level is typically

introduced with Z_t discrete taking values in possible *phonemes* and and following a Markov model so that

$$P(Z_{t+1} = z_{t+1} \,|\, Z_s = z_s, s \leq t) = q(z_{t+1} \,|\, z_t, \theta),$$

and

$$\mathcal{L}(X_{t+1} \,|\, (X_t = x_s, Z_t = z_s), s \leq t, \theta) \sim \mathcal{N}\{A_t(\theta, z_t)x_t, \sigma_t^2(\theta, z_t)\}.$$

and

$$\mathcal{L}(Y_t \,|\, X, \eta) = \mathcal{N}\{B_t(\eta)x_t, \tau^2(\eta)\},$$

where $Y_t, t = 1, \ldots, T$ are observed.

Such models are *switching state space* models (SSM).

If the middle level is missing, it is also called a *hidden Markov model* (HMM).