SAMSI Workshop

Derived variables

D.R. Cox

Nuffield College, Oxford, UK

david.cox@nuffield.ox.ac.uk

Measurement

Classification by

- mathematical structure
- by purpose
  - response (or outcome)
  - explanatory
    - primary explanatory variable (treatment)
    - intrinsic
    - non-specific
- primary measure (pointer reading) or derived variable

Strong implications for construction of graphical models

Some examples

- true-false answers to questions on xxxx
- body-mass index
- autopsy of diseased animals , 18 sites grouped by 5 parts of body, scored 0,1,2,3,4 for presence of visible lesions; summarization by severity score
- fluid velocity, density, absolute viscosity and chstic length leading to Reynolds number for laminar versus turbulent flow

In the last example, atypically, dimensional analysis leads to a virtually unique answer.

Some approaches

Indices based on variables not all on an equal footing. Often essentially *a priori* assumptions about regression coefficients.

Two versions of body mass index

- height intrinsic, weight an outcome or primary explanatory variable
- in children variables on an equal footing (ponderosity index)

More detail

- latent variable in simplest case with single factor structure of simple graphical form. May be somewhat plausible or purely a device for formalizing an analysis
- internal analysis specific to each set of data
- internal analysis leading to consensus
- external analysis; canonical analysis

Index to detect departures in a particular direction

Let Y be a vector of p components with mean and covariance matrix Σ. Suppose we can specify a direction l such that it is desired to detect departures from μ to $\mu + al$. Apply Lagrange multipliers to justify index $s^T Y$ where the vector s of scores is such that

$$s \propto \Sigma^{-1} l.$$

Provided all components point in the same direction might take

$$l = [\mathrm{diag}\Sigma]^{1/2}.$$

Advantages to taking a simple sum score, $l = 1$, interpretability and communicability. Reasonable if the unit vector is close to the dominant eigenvector of Σ.

Role of graphical Markov models in determining indexes

Suppose all variables on an equal footing.

Question:

Does the concentration graph or covariance graph of the variables give much guidance on how to achieve dimension reduction?

Question:

Often there is prior subject-matter grouping of the variables within a block of variables. Is it useful to incorporate this information in the graph? Or is that too late?

Transformation of dependence to a canonical form

Consider regression of Y on X, these being vectors assumed in the first place both to be $p \times 1$. Transform Y to $Y^* = AY$. Then regrssion matrix of Y^* on X is

$$B_{Y^*X} = \Sigma_{Y^*X}\Sigma_{XX}^{-1} = AB_{YX}$$

so that a standardized version of SER is obtained if

$$A = B_{YX}^{-1} = \Sigma_{XX}\Sigma_{YX}^{-1}.$$

Produces simply interpretable set of regressions of Y^* on X in which conditional correlations among the components of Y^* re induced from the error terms.

Time series version

Vector time series Y_t. Consider $Y_t^* = AY_t$ regressed on $Y_{t-1}^* = AY_{t-1}$ and possibly other explanatory variables.

Regression matrices related by

$$B_{t,t-1}^* = AB_{t,t-1}A^{-1}.$$

Simple structure achieved if for diagonal D

$$AB_{t,t-1} = DA.$$

All the cross-covariance between components is forced into the innovation process.

Many possible generalizations

REFERENCES

J. Multivariate Analysis **42** (1992), 162-170.

Proc. Nat. Acad. Sci. **96** (1999), 13273-13274.