We are done! Some odds and ends

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 730: Multivariate Data Analysis

Review...

- Multivariate analysis immensely useful; anytime data are correlated we have a joint distribution. The mean may be of primary interest, the covariability, or both.
- We covered primarily classical multivariate analysis.
- Tried to throw in some more recent stuff along the way: bootstrap, general regression model, mixed models, functional data, infinite mixtures, discrimination beyond linear, LASSO, EM algorithm, MCMC, etc. Many things came up that post-date the book.
- Natural followup/companion courses listed in first set of notes.
- Elements of statistical learning should be offered Fall of 2015, hopefully by Edsel.

Review...

- Course webpage provides good review.
- Main topics: normal theory incl. small-sample results useful for testing, regression & MANOVA, PCA, MDS, canonical correlation, clustering, discrimination, and FA.
- PCA usually exploratory, but now being used to construct models. Very useful in "denoising" signals. Also useful: general basis expansions, splines, Gaussian processes.
- PCA provides main heuristic for MDS and FA. FA really a model-based PCA.
- In regression setting, two ways to model/allow for correlated $\mathbf{y}_i$: direct (marginal models, structured covariance matrices), and indirect or conditional (shared random effects leading to mixed models).

"Common components" models

- Briefly mentioned in MKB 2.6.2. We discussed in addendum to Chapter 6, mixed models.
- This is really the main approach toward inducing correlation among non-normal responses: generalized linear mixed models (GLMM).
- Poisson example

$$y_{ij}|\boldsymbol{\beta},\boldsymbol{\gamma}_i \overset{ind.}{\sim} \text{Pois}(\lambda_{ij}),\ \log(\lambda_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}+\mathbf{z}'_{ij}\boldsymbol{\gamma}_i,\ \boldsymbol{\gamma}_i \overset{ind.}{\sim} N_q(\mathbf{0},\boldsymbol{\Omega}).$$

- Bernoulli example

$$y_{ij}|\boldsymbol{\beta},\boldsymbol{\gamma}_i \overset{ind.}{\sim} \text{Bern}(\pi_{ij}),\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mathbf{x}'_{ij}\boldsymbol{\beta}+\mathbf{z}'_{ij}\boldsymbol{\gamma}_i,\ \boldsymbol{\gamma}_i \overset{ind.}{\sim} N_q(\mathbf{0},\boldsymbol{\Omega}).$$

"Common components" models

- Positive correlation induced among $\mathbf{y}_i' = (y_{i1}, \ldots, y_{in_i})$ through the common components γ_i.
- Simplest version: random intercept models. Correlation structure immediate for normal mixed model. Also termed "repeated measures" model in early literature.
- Factor analytic model $\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}_\mathbf{i} + \mathbf{u}_\mathbf{i}$ another example. Difference here is that loadings $\boldsymbol{\Lambda}$ are unknown; in mixed model $\mathbf{Z}_i$ is known.

Structured error

Mixed models induce the correlation among the components of $\mathbf{y}_i$. Another option is to model it directly.

- In STAT 704/705 one common assumption is independence. Often violated for data collected over time or space.
- For example, say y_{ij} is blood pressure of subject i taken at week j after starting medication. We expect the elements of $\mathbf{y}_i' = (y_{i1}, \ldots, y_{in_i})$ to be correlated.
- Assume $y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + u_{ij}$ (yielding $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + +\mathbf{u}_i$). Simple autoregressive structure posits

$$u_{ij}|u_{i1}, \ldots, u_{i,j-1} \sim N(\rho u_{i,j-1}, \sigma^2).$$

Can show then that

$$V(\mathbf{u}_i) = \boldsymbol{\Sigma}(\rho, \sigma^2) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n_i-1} \\ \rho & 1 & \rho & \cdots & \rho^{n_i-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & \cdots & 1 \end{bmatrix}.$$

AR(1) structure

- The covariance matrix has only two parameters, ρ and σ^2.
- Durbin-Watson test for independence is simply $H_0 : \rho = 0$ (using only one group).
- Other covariance structures: banded, Toeplitz, compound symmetry, many others.
- AR(1) generalized to ARMA(p, q), general time-series model (STAT 520 & STAT 720). Time series approach often "detrends" first; regression fits everything at once.
- How to allow for seeing individuals at differently spaced times? Markov temporal dependence, which leads to Gaussian process with exponential covariance function.

Spatial dependence

If observations are collected at different geographical locations, say at coordinates $\mathbf{s}_i = (s_{i1}, s_{i2})$ (often latitude/longitude), then spatial structure can be used for the residuals. This helps properly estimate regression effects and can also drastically improve prediction.

Often there is repeated measures only over space, not within individuals too. That is, each subject i is measured only once yielding $(y_i, \mathbf{x}_i)$, but the location where y_i was collected $\mathbf{s}_i$ is known. However, it is possible to have spatial correlation for repeated mesures within a group and independence across groups.

Exponential decay of correlation

- The usual linear model can be assumed $\mathbf{Y} = \mathbf{X}\beta + \mathbf{u}$, but now $\mathbf{u}$ should have spatially-correlated elements.
- Different spatial correlation functions are available in `lme`, exponential is $\rho_\theta(\mathbf{s}_1, \mathbf{s}_2) = \exp(-||\mathbf{s}_1 - \mathbf{s}_2||/\theta)$ where θ is the "range." The ijth element of $V(\mathbf{u})$ is $\tau^2\rho_\theta(\mathbf{s}_i, \mathbf{s}_j) + I\{i = j\}\sigma^2$. Last part is called "nugget."
- Generalizes AR(1)! Can be fit to longitudinal data with irregularly observed observations.
- Other spatial correlation models: spherical, Matérn, Gaussian.
- Can fit correlogram to OLS residuals to see if there's spatial component; estimates $\rho(d) = \text{corr}(u(\mathbf{s}), u(\mathbf{s} + \mathbf{\Delta}))$ where $d = ||\mathbf{\Delta}||$. Since this is not a function of $\mathbf{s}$, stationarity is assumed.

Example in R

The "thick" dataset is from SAS documentation. Want to predict measured thickness of coal seams `thick` at different coordinates (`east` & `north`) as a function of `soil`, measuring soil quality.

```
library(ncf)
library(nlme)
d=read.table("http://www.ats.ucla.edu/stat/r/faq/thick.csv",header=T,sep = ",")
d # east and north are the spatial locations
# look at correlogram of residuals from OLS fit
f1=lm(thick~soil,data=d)
cr=spline.correlog(x=d$east,y=d$north,z=resid(f1))
plot(cr) # seems to be positive correlation for residuals close to each other

# default is nugget=F; need replication to estimate nugget effect
f2=lme(fixed=thick~soil,data=d,random=~1|dummy,     # dummy is all 1's (only on
 correlation=corExp(form=~east+north),method="ML")  # other spatial correlation
f3=gls(thick~soil,data=d) # Gaussian correlation also available
summary(f2) # much improved fit, soil quality no longer important after locatio
summary(f3) # same as lm fit, but gives AIC and BIC
```

Spatial random effects

The linear predictor in most models can be augmented to include spatial frailties.

The theoretical model is actually based on a Gaussian process, e.g.

$$\eta(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta} + u(\mathbf{s}),$$

where $u(\mathbf{s})$ is a mean-zero stationary Gaussian process. An exponential covariance function guarantees that $u(\mathbf{s})$ is infinitely differentiable, i.e. very smooth. Finite realizations, e.g. $\mathbf{u}' = (u(\mathbf{s}_1), \ldots, u(\mathbf{s}_n))$ are multivariate normal.

This $\eta(\mathbf{s})$ can be used in spatial Poisson regression, spatial logistic regression, spatial normal-errors regression, a spatially-varying proportional hazards model, etc.

Alternatively, a Gaussian process can be used in a marginal model through a copula.

Markov random fields

Sometimes geographic information is not in the form of exact locations $\mathbf{s}$, but rather is less exact, e.g. the county-of-residence a subject lives in. A conditionally autoregressive (CAR) prior smooths random effects $u_1, \ldots, u_m$ according to geographic location of the counties. Specifically,

$$u_j | u_{-j} \sim N\left(\rho \tilde{u}_j, \frac{\lambda}{n_j}\right),$$

where $\tilde{u}_j$ is the mean of the $\{u_i\}_{i \neq j}$ that share a border with u_j, and n_j is the number that share a border. One can show (not easy) that the joint distribution is then

$$\mathbf{u} \sim N_m(\mathbf{0}, \lambda(\mathbf{D} - \rho \mathbf{W})^{-1}),$$

where $\mathbf{W}$ is a "proximity matrix" and $\mathbf{D}$ is diagonal with elements $d_{ii} = w_{i+}$. $\rho = 1$ gives an intrinsic CAR, or ICAR; not proper; typcially $\rho \in [0, 1)$.

Directional data

MKB Chapter 15. First step in shape-related data analysis, statistical analysis of manifold data, etc. Ian Dryden used was here up until a few years ago.

Directional data is essentially the consideration of probability distributions on circles or spheres. McMillan et al. (2013) combined circular data, splines, and random effects to deal with multivariate repeated measures directional data across groups (a kind of MANOVA with lots of structure).