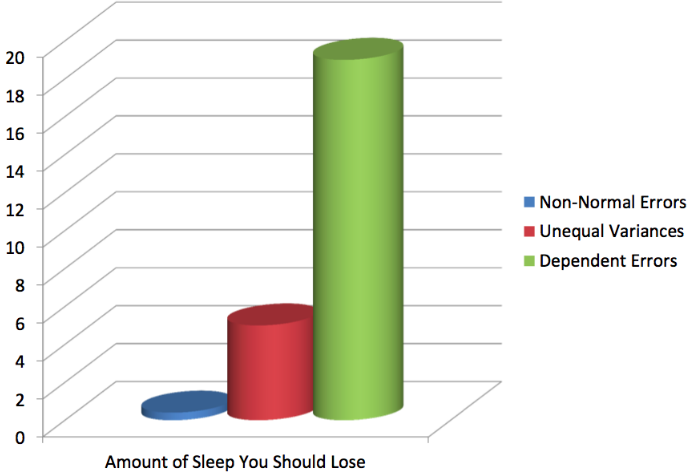
Separable covariance arrays via the Tucker product

by Peter Hoff

Kean Ming Tan

April 23, 2013

Correlated Errors are Bad!



Lecture slide from Biostat 533

Review of Multivariate Analysis

Multivariate normal model, $\mathbf{y} \in \mathbb{R}^m$:

$$\mathbf{z} = \{z_j : j = 1, \ldots, m\} \overset{\text{iid}}{\sim} \text{normal}(0, 1)$$
$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{A}\mathbf{z} \overset{\text{iid}}{\sim} \text{multivariate normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T)$$

Matrix-variate normal model, $\mathbf{Y} \in \mathbb{R}^{m_1 \times m_2}$:

$$\mathbf{Z} = \{z_{i,j}\}_{i=1}^{m_1}{}_{j=1}^{m_2} \overset{\text{iid}}{\sim} \text{normal}(0, 1)$$
$$\mathbf{Y} = \mathbf{M} + \mathbf{A}\mathbf{Z}\mathbf{B}^T \overset{\text{iid}}{\sim} \text{matrix normal}(\mathbf{M}, \boldsymbol{\Sigma}_1 = \mathbf{A}\mathbf{A}^T, \boldsymbol{\Sigma}_2 = \mathbf{B}\mathbf{B}^T)$$
$$\sim \text{matrix normal}(\mathbf{M}, \boldsymbol{\Sigma}_1 \circ \boldsymbol{\Sigma}_2)$$

Note that matrix-variate normal assumes separable covariance structure

What is separable covariance structure?

$$\mathbf{Y} \sim \text{matrix normal}(0, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$$

- Covariance is product of row covariance and column covariance
$$\text{Cov}(Y_{ij}, Y_{kl}) = \Sigma_{1\,ik} \times \Sigma_{2\,jl}$$
- Reduced number of parameters to be estimated
- From $\frac{(np)\times(np+1)}{2}$ to $\frac{p(p+1)}{2} + \frac{n(n+1)}{2}$

Made up motivation - linear regression model

$$\mathbf{Y} = \mathbf{M} + \mathbf{E}$$

- $\mathbf{M}$ is the mean structure (for instance, $\mathbf{X}\boldsymbol{\beta}$, or ANOVA model)
- $\mathbf{E}$ is the error term

Made up motivation - Example 1

Suppose $\mathbf{y}_i \in \mathbb{R}^{m_1}$ is the outcome variable obtained by repeatedly taking measurements from subject i across time $j = \{1, \ldots, m_1\}$.

Appropriate Model:

$$\mathbf{y}_i = \mathbf{x}_i^T \beta + \boldsymbol{\epsilon}_i$$

$\boldsymbol{\epsilon}_i \sim$ multivariate normal$(\mathbf{0}, \mathbf{\Sigma})$

Made up motivation - Example 2

Suppose $Y_i \in \mathbb{R}^{m_1 \times m_2}$ and $y_{i,j,k}$ is the ith outcome variable for location j at time k.
Naive Model: Assume that the locations are not correlated

$$\mathbf{y}_{ij} = \mathbf{x}_{ij}^T \beta + \boldsymbol{\epsilon}_{ij}$$

$\boldsymbol{\epsilon}_{ij} \sim$ multivariate normal$(\mathbf{0}, \mathbf{\Sigma})$

Made up motivation - Example 2

Suppose $Y_i \in \mathbb{R}^{m_1 \times m_2}$ and $y_{i,j,k}$ is the ith outcome variable for location j at time k.
Naive Model: Assume that the locations are not correlated

$$\mathbf{y}_{ij} = \mathbf{x}_{ij}^T \beta + \boldsymbol{\epsilon}_{ij}$$

$\boldsymbol{\epsilon}_{ij} \sim$ multivariate normal$(\mathbf{0}, \boldsymbol{\Sigma})$

Really? Ignoring dependent errors after taking Biostat571?

More complicated models

Look at Laina Mercer's slides, or alternatively,

More complicated models

Look at Laina Mercer's slides, or alternatively,

$$\mathbf{Y}_i = \mathbf{\Theta}\mathbf{X}_i + \mathbf{E}_i$$

$\mathbf{E}_i \overset{iid}{\sim}$ matrix normal$(\mathbf{0}, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$

Closely related to (Knorr-Held and Besag, 1998), it does not allow for space $\times$ time interactions.

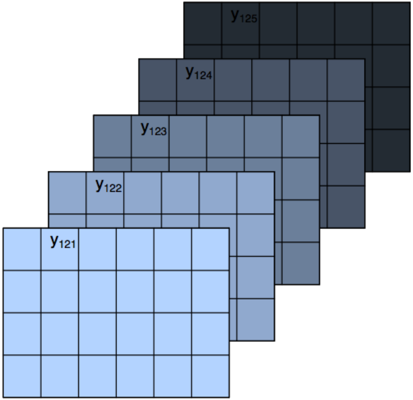
Citation: On matrix-variate regression analysis by Cinzia Viroli (2012)

What is an array?

Gene expression data set
$\mathbf{Y} = \{y_{i,j,k}\}$.

- i indexes the ith subject
- j indexes the jth gene
- k indexes the kth repeated measurement

Then, $y_{i,j,k}$ is the gene expression level for the jth gene of the ith subject, measured at time k.



Citation: Are a set of microarrays independent of each other by Brad Efron (2009)

Motivation - Example 3

Yearly change in log trade value (in 2000 dollars): $\mathbf{Y} = \{y_{i,j,k,l}\}$

- $i \in \{1, \ldots, 30\}$ indexes the exporting nation
- $j \in \{1, \ldots, 30\}$ indexes the importing nation
- $k \in \{1, \ldots, 6\}$ indexes the commodity type
- $t \in \{1, \ldots, 10\}$ indexes the year

Interested in modeling the mean $M_{ijk} = \mu_{i,j,k}$ across t measurements

What can we do?

Motivation - Example 3 cont

Interested in the model

$$y_{i,j,k,l} = \mu_{i,j,k} + \epsilon_{i,j,k,l}$$

- iid error model: $\epsilon_{i,j,k,l} \sim \text{normal}(0, \sigma^2)$
- multivariate error model: $\boldsymbol{\epsilon}_{i,j,k} \sim \text{multivariate normal}(\mathbf{0}, \boldsymbol{\Sigma})$
- matrix-variate error model: $\boldsymbol{\epsilon}_{i,j} \sim \text{matrix normal}(\mathbf{0}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$

Motivation - Example 3 cont

Interested in the model

$$y_{i,j,k,l} = \mu_{i,j,k} + \epsilon_{i,j,k,l}$$

- iid error model: $\epsilon_{i,j,k,l} \sim \text{normal}(0, \sigma^2)$
- multivariate error model: $\boldsymbol{\epsilon}_{i,j,k} \sim \text{multivariate normal}(\mathbf{0}, \boldsymbol{\Sigma})$
- matrix-variate error model: $\boldsymbol{\epsilon}_{i,j} \sim \text{matrix normal}(\mathbf{0}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$

But all four dimensions are correlated!

$$\mathbf{E} \sim ???$$

Goal of the paper

Propose the Array Normal distribution for array data

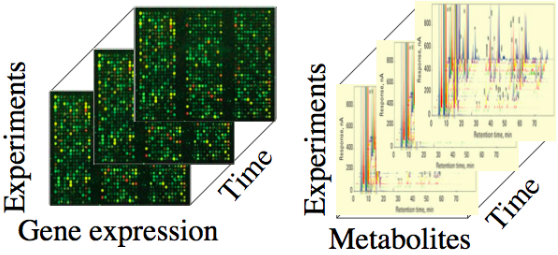
- model mean structure
- model covariance structure

Suppose $\mathbf{Y} \in \mathbb{R}^{m_1 \times \ldots \times m_k}$

$$\mathbf{Y} \sim \text{ array normal}(\mathbf{M}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_k)$$

Take home message until the next talk

- Array data is everywhere



- Most people assume certain dimensions are independent
- Maybe it is a good idea to model the dependencies after all!

Questions?