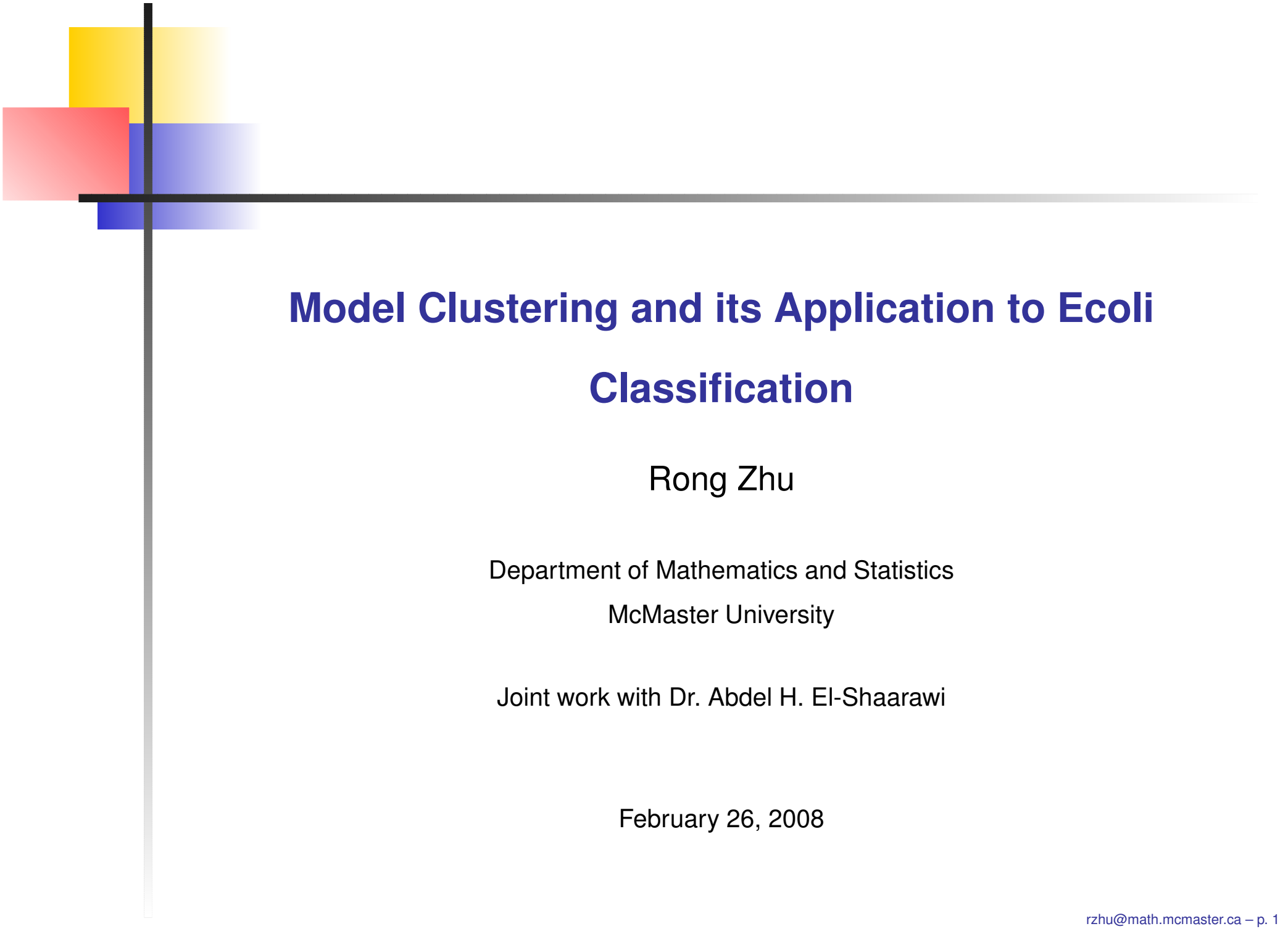


Model Clustering and its Application to Ecoli Classification

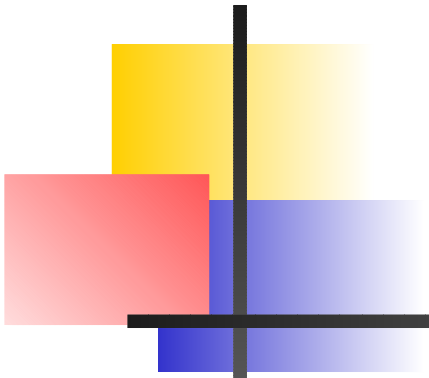
Rong Zhu

Department of Mathematics and Statistics

McMaster University

Joint work with Dr. Abdel H. El-Shaarawi

February 26, 2008



Outline

- Motivation:
  - Case Study: Ecoli Classification
  - Conventional Approach and Problems
- Model Clustering: Model Linking and Grouping Strategies
- Simulation Study
- Application to Ecoli Case Study
- Discussion

Motivation

- Case Study
  - **Data**
    - Temporal and Spatial Factors

      Experiments were conducted biweekly from twenty locations in three Canadian watersheds (Alberta, Ontario, Quebec).
    - Treatment Factors

      On each scheduled date, Ecoli isolates were treated separately by one of twelve antibiotics at different prescribed levels (11 had 3 levels, and one had four levels, so total is 37 levels).
    - Response

      Numbers of alive Ecoli bacteria before and after each antibiotic stress were recorded.

Motivation

- Case Study (Cont'd)
  - **Objective**
    - *Track the source of Ecoli bacteria (originated from either human or animals) of water to investigate the suitability of water for human use.*
    - *Statistically, classify Ecoli bacteria from those twenty locations according to their response behaviours to antibiotic treatments.*

      Reasoning: *Ecoli from the same source are assumed to have the same response behaviours to those antibiotics.*

    - *To this end, need to remove the temporal and spatial effects to the response behaviours.*

Motivation

- Conventional Approach
  - **Model Fitting**

    Each object associates with a data set, i.e., corresponding to many observations.

    Objects are characterized by underlying relationships between response variable and covariates. Specifically, they are characterized by a subset of parameters in the models of the same family.

    Use parametric models to summarize the underlying relationships for all objects.
  - **Cluster Analysis**

    Perform cluster analysis for the estimated values of the subset of parameters to partition all objects into groups.

Motivation

- Problems
  - **Informatical approach**

    In classic cluster analysis, each object is represented by one observation, and similarity between objects is usually measured by a type of distance.

    In hierarchical clustering, the similarity between two groups is indicated by a function of all pairwise similarity measures of objects between two groups, say the average method.

  - **Statistical approach**

    For our case study, each object (O_k) corresponds to a data set ($\mathbf{Z_k}$), thus represented by underlying relationships ($\mathbf{M_j}$).

    If two data sets have the same feature in their relationships on some aspect, models will be fitted by the pooled data, instead of using existing estimates obtained by separate date sets.

Motivation

- Problems (Cont'd)
  - **Conjecture**

    Metric closeness in parametric space, particularly the function of pairwise similarity measures for successive merged groups, may not well represent the closeness of underlying relationships for successive merged groups.

    Need to develop
    - *new similarity measure*
    - *new cluster analysis technique*

    for objects characterized by underlying relationships.

Model Clustering

- Model Linking
  - **Definition of Model Similarity**

    Models $M(\mathbf{Z}_k; \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$ *(*$k = 1, \ldots, l$*) are said to be similar if* $\boldsymbol{\beta}$*, the subset of parameters, are the same for all models, i.e.,* $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_l$*, where* $l \geq 2$*.*

    $\Rightarrow$ *all models are linked together by* $\boldsymbol{\beta}$*.*

    Special case: equality of all parameters leads to the same model.

  - **Null and Alternative Hypotheses**

    $H_0 : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_l$ *vs* $H_a :$ $\boldsymbol{\beta}_k{}'s$ *are different from one another*

    Comparison between two extreme situations:

    - H_0 *implies that all of them are similar*
    - H_a *means that none of them is similar to each other*
    - $H_0 \subset H_a$

Model Clustering

- Model Linking (Cont'd)
  - **Testing**

Denote

$$\Theta_0 = \{(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_l, \boldsymbol{\beta}, \cdots, \boldsymbol{\beta})\} \quad \subset \quad \Theta = \{(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_l, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_l)\}$$

as the parameter space under H_0 and H_a respectively.

Use LRT, under H_0,

$$\begin{aligned} T &= 2 \max_{\Theta} \log L_a(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_l, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_l) - 2 \max_{\Theta_0} \log L_0(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_l, \boldsymbol{\beta}) \\ &= 2 \sum_{k=1}^{l} \max_{\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k} \log L(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k | \mathbf{Z}_k) - 2 \max_{\Theta_0} \sum_{k=1}^{l} \log L(\boldsymbol{\alpha}_k, \boldsymbol{\beta} | \mathbf{Z}_k) \\ &\xrightarrow{d} \chi^2_{(l-1)b}, \quad \text{as } n_k \to \infty, \quad \text{where } k = 1, \ldots, l. \end{aligned}$$

Remark: *Other general or specific testing approaches can be employed.*

Model Clustering

- Model Linking (Cont'd)
  - **Measure of Model Similarity**

    p-value can be obtained for the testing.

    If H_0 is likely to be true, p-value would be large. Otherwise, it would be small.

    Therefore, p-value indicates the degree of similarity of underlying models for given data sets $\mathbf{Z}_1, \ldots, \mathbf{Z}_l$.

    $\Rightarrow$ *adopt the p-value as the measure of model similarity*

    Remark*: Unlike most similarity measures in classic cluster analysis, the p-value is not a metric distance.*

Model Clustering

- Grouping Strategies

  Assume there are K objects, each associated with an underlying model from the same parametric family. α_0 is prescribed.

  - **Manual Grouping**

    - *Step 1: Calculate pairwise $K \times K$ p-value matrix for all K models.*
    - *Step 2: Select model pairs whose p-values are not smaller than α_0.*
    - *Step 3: Form rough clusters from selected pairs visually.*
    - *Step 4: Confirm rough clusters by model linking. If models not similar enough within a cluster, then adjust the formation and redo model linking.*

    K models will be partitioned as clusters obtained in Step 4 plus individual models not selected in Step 2.

    Remark: *Intuitive method used for small K.*

Model Clustering

- Grouping Strategies (Cont'd)
  - **Cluster Peeling**
    - *Step 1: Randomly select a model from the examined model pool as a cluster.*
    - *Step 2: Find a model most similar enough to the cluster from the remaining in the pool (i.e., having the largest p-value that is bigger or equal to α_0), and merge it into the cluster.*
    - *Step 3: Repeat Step 2 until no model in the remaining is similar to the newly formed cluster.*
    - *Step 4: Remove the cluster finalized at Step 3.*
    - *Step 5: Repeat Step 1 until the examined model pool is empty.*

Remark: *Save time by avoid calculation of p-value matrix. Used for large K. However, might assign a model to a less similar cluster first.*

Model Clustering

- Grouping Strategies (Cont'd)
  - **Pairwise Combining**
    - *Step 1: Calculate p-value matrix for the current cluster pool.*
    - *Step 2: Merge the most similar two clusters as one cluster. Thus, the size of cluster pool is reduced by one.*
    - *Step 3: Repeat Step 1 until the size of cluster pool is one.*

Remark: *Most time consuming. But can draw dendrogram plot and have a full picture.*

Model Clustering

- Grouping Strategies (Cont'd)
  - **Speeding Technique: Splitting and Binding**
    - *Splitting Step:*

      Split the entire set of K models into several subsets. Apply cluster peeling or pairwise combining strategy to each subset of models with a large threshold.

    - *Binding Step:*

      Bind all clusters obtained from each subsets in the splitting step to form a new cluster pool, and then apply cluster peeling or pairwise combining strategy again for this pool with a relative small threshold.

Simulation Study

- Model Specifications and Sample Sizes

  Simulate six Poisson regression models of form:

  $$Y \sim \text{Poisson}(\lambda), \quad \textit{where } \log(\lambda) = \gamma + \alpha X_1 + \beta X_2.$$

  Sample sizes are $100, 100, 100, 300, 200$ *and* 200 *respectively.*

  Models are similar if they have the same values for parameter β*, the coefficient of* X_2*. Three clusters are purposely set:*

  - *Cluster 1: Models 1, 2 and 3*
  - *Cluster 2: Models 4*
  - *Cluster 3: Models 5 and 6*

Simulation Study

- Model Specifications and Sample Sizes (Cont'd)

Parameter Specifications and Estimates for Six Poisson Regression

Model	γ	α	β	*Sample size*	$\hat{\gamma}$	$\hat{\alpha}$	$\hat{\beta}$
1	*0.5*	*1.0*	*1.5*	*100*	*0.3628*	*0.9979*	*1.5502*
2	*1.0*	*2.0*	*1.5*	*100*	*0.9726*	*2.1254*	*1.5228*
3	*3.0*	*-1.0*	*1.5*	*100*	*2.9974*	*-0.9782*	*1.5077*
4	*1.0*	*2.5*	*3.0*	*300*	*0.9900*	*2.4966*	*3.0024*
5	*1.5*	*3.0*	*-2.0*	*200*	*2.2474*	*5.5614*	*-2.3457*
6	*2.5*	*-3.0*	*-2.0*	*200*	*2.5221*	*-3.1873*	*-2.1487*

Simulation Study

- Results

Pairwise p-value matrix among six fitted Poisson regression models

$$\begin{pmatrix} 1.000 & 0.554 & 0.246 & 0.000 & 0.000 & 0.000 \\ 0.554 & 1.000 & 0.619 & 0.000 & 0.000 & 0.000 \\ 0.246 & 0.619 & 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 & 0.498 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.498 & 1.000 \end{pmatrix}$$

(model 1, model 2), (model 2, model 3), (model 5, model 6) have p-values larger than prescribed $\alpha_0 = 0.45$.

Manual Grouping: *Cluster 1 =* $\{$*model 1, model 2, model 3*$\}$,

Cluster 2 = $\{$*model 4*$\}$, *Cluster 3 =* $\{$*model 5, model 6*$\}$

Simulation Study

- Results (Cont'd): *Cluster Peeling* ($\alpha_0 = 0.45$)

Cluster	*Models in cluster*	*Models in remaining pool*	*Model most similar to cluster*
1	*1*	*2, 3, 4, 5, 6*	*2 (0.554* $> \alpha_0$*)*
1	*1, 2*	*3, 4, 5, 6*	*3 (0.467* $> \alpha_0$*)*
1	*1, 2, 3*	*4, 5, 6*	*4, 5 or 6 (0.000* $< \alpha_0$*)*
	Remove Cluster 1 finalized in last row	*4, 5, 6*	–
2	*4*	*5, 6*	*5 or 6 (0.000* $< \alpha_0$*)*
	Remove Cluster 2 finalized in last row	*5, 6*	–
3	*5*	*6*	*6 (0.498* $> \alpha_0$*)*
3	*5, 6*	–	–
	Remove Cluster 3 finalized in last row	–	–

Application to Ecoli Case Study

- Objective and Data

Classify Ecoli strains from twenty locations in three Canadian watersheds according to their genetic differences, i.e., responses to antibiotic treatments.

Numbers of dates for experiments in each of 20 *locations.*

Watershed	*Location*								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
BH	*12*	*12*	*12*	*12*	*12*	*12*	–	–	–
LB	*9*	*9*	*9*	*9*	*9*	*9*	*9*	*7*	*9*
SN	*11*	*11*	*10*	*10*	*11*	–	–	–	–

On each scheduled date, Ecoli strain from one location had 37 trials to antibiotic treatments.

Alive counts before and after experiment were recorded.

Application to Ecoli Case Study

- Model Fitting

Employ binomial regression

$$Y \sim \text{binomial}(n, p), \quad \textit{where } \text{logit}(p) = \log \frac{p}{1-p} = \gamma + \alpha_i + \beta_j,$$

subject to

$$\sum_i \alpha_i = 0, \qquad \sum_{j=1}^{37} \beta_j = 0.$$

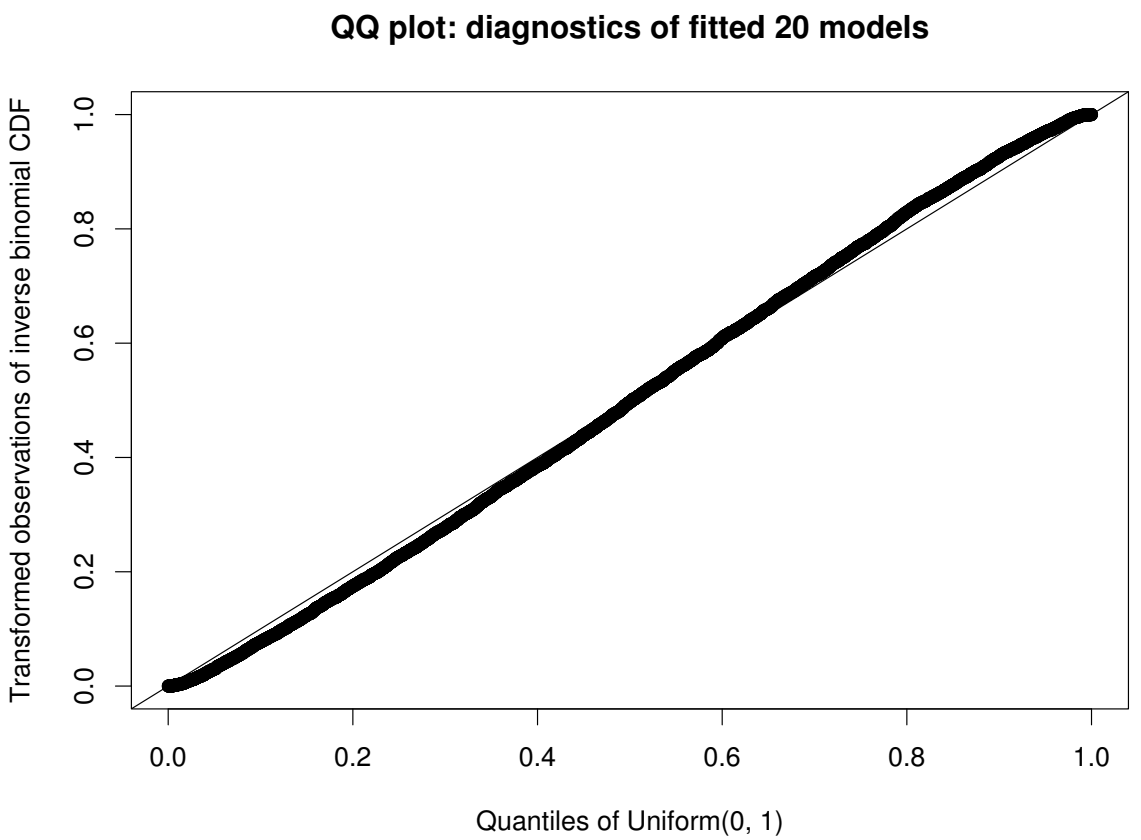
Here

- α_i*'s: temporal effects*
- β_j*'s: treatment effects*

Application to Ecoli Case Study

- Model Fitting (Cont'd)

Visual Diagnostics of Twenty Fitted Binomial Regression Models



Application to Ecoli Case Study

- Clustering using Conventional Approach

Denote $\hat{\boldsymbol{\beta}}(i) = \left(\hat{\beta}_2(i), \ldots, \hat{\beta}_{37}(i)\right)^T$ *for the* i*-th fitted model,*

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}(1) \\ \vdots \\ \hat{\boldsymbol{\beta}}(20) \end{pmatrix} = \begin{pmatrix} \hat{\beta}_2(1) & \cdots & \hat{\beta}_{37}(1) \\ \vdots & & \vdots \\ \hat{\beta}_2(20) & \cdots & \hat{\beta}_{37}(20) \end{pmatrix}$$

Classify

$$\hat{\boldsymbol{\beta}}(1), \quad \hat{\boldsymbol{\beta}}(2), \quad \ldots, \quad \hat{\boldsymbol{\beta}}(20)$$

using Euclidean distance in a 36*-dimensional space and hierarchical clustering with the average agglomeration method.*

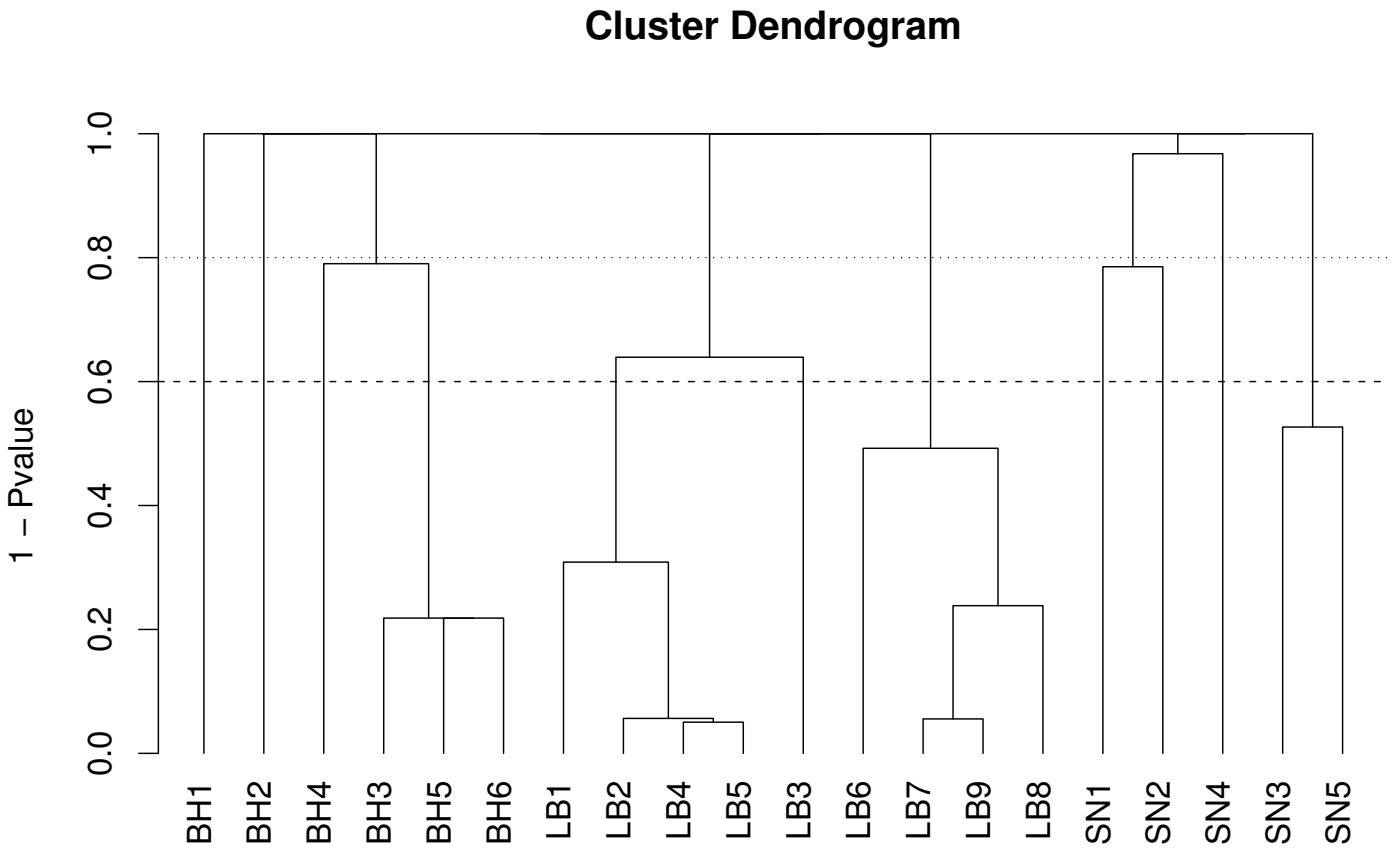
Application to Ecoli Case Study

- Clustering using Conventional Approach (Cont'd)

Application to Ecoli Case Study

- Model Clustering

Model clustering using pairwise combining strategy



Application to Ecoli Case Study

- Findings
  - *Using model clustering, Ecoli strains from the same watershed seem to group together, although they further form different small groups within each watershed.*

    Possible explanation: neighbours share the same environment in a large but closed area.
  - *Using conventional approach, Ecoli strains from different watersheds mix one another.*

    Potential risk: may lead to wrong classification.

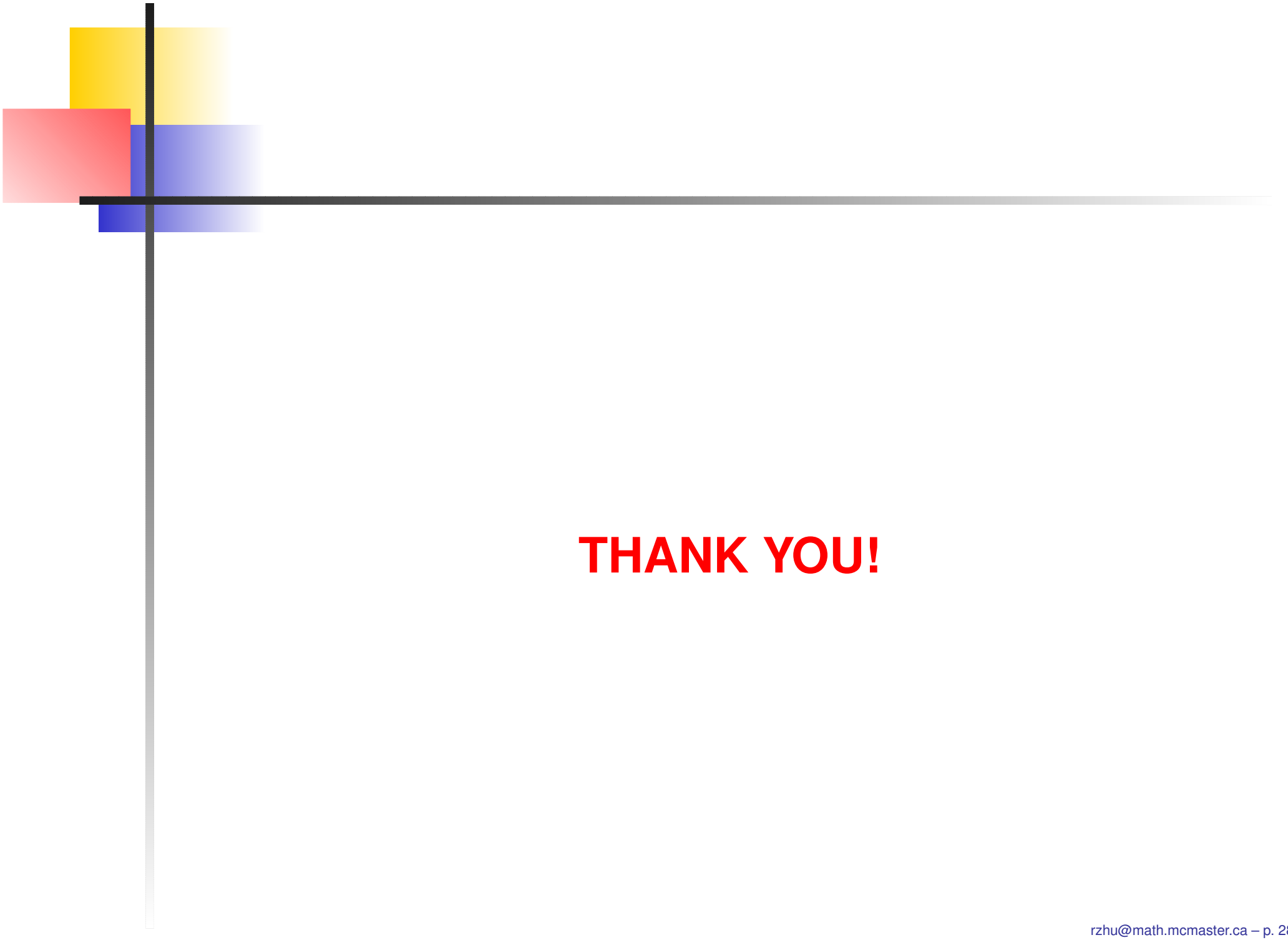
Application to Ecoli Case Study

- Comments
  - *Conventional cluster analysis groups observations or points, while model clustering classifies underlying parametric models*
  - *Metric closeness in parameter space may not be consistent with likelihood closeness in model space, leading to different classification results*

    Eg, (BH5, LB4) are very close in parameter space, but very different in model space
  - *When classifying models, similarity between clusters in conventional hierarchical clustering using agglomeration method does not have a clear explanation*
  - *p-value is more straightforward and appropriate to measures the likelihood closeness in model spaces, i.e., model difference for given data sets.*

Discussion

- *Model clustering classifies objects according to the equality relationship of partial parameters among their underlying parametric models.*
- *Each object is associated with a data set and characterized by a parametric model from the same family.*
  Hence, model fitting is crucial in model clustering.
- *Conventional approach uses metric closeness in parametric space. Model clustering extends the similarity measure to likelihood closeness in model space. This approach is better to find the model differences under given data sets.*
- *Model clustering usually requires more computational time. Proper speeding technique can accelerate the clustering.*
- *Further investigation and applications are expected.*



THANK YOU!

References

1. Chow, G.C. (1960). Tests of Equality between Sets of Coefficients in Two Linear Regressions. *Econometrica*, 28, 591-605.
2. El-Shaarawi, A. H. and Kwiatkowski, R. E. (1977). A Model to Describe the Inherent Spatial and Temporal Variability of Parameters in Lake Ontario 1974. *Journal of Great Lakes Research, The International Association for Great Lakes Research*, 3(3-4), 177-183.
3. El-Shaarawi, A. H. and Shah, K. R. (1978). Statistical Procedures for Classification of a Lake. *Inland Waters Directorate, Environment Canada, Scientific Series No.* 86.
4. Hartigan, J.A. (1975). *Clustering algorithms.* Wiley, New York.
5. Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis (Fourth Edition).* Prentice Hall, New Jersey.

References

6. Kaufman, L. and Rousseeuw, P.J. (1990). *Finding groups in data.* Wiley, New York.
7. Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005). *Applied Linear Statistical Models (Fifth Edition).* McGraw-Hill/Irwin, New York.
8. Lehmann, E.L. (1986). *Testing Statistical Hypotheses (second edition).* John Wiley and Sons, New York.
9. Liao, T.F. (2002). *Statistical Group Comparison.* Wiley, New York.
10. Liao, T.F. (2004). Comparing Social Groups: Wald Statistics for Testing Equality Among Multiple Logit Models *International Journal of Comparative Sociology*, 45, 3-16.
11. Liu W., Jamshidian M. and Zhang Y. (2004). Multiple Comparison of Several Linear Regression Models *JASA*, 99, 395-403.

References

12. Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26, 354-359.

13. McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models (Second Edition).* Chapman and Hall, London.

14. Self, S.G. and Liang, K.-Y. (1987). Asymptotic properties of the maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.