2102531 System Identification

Term Project
Semester 1/2015

Jitkomut Songsiri

November 23, 2015

Abstract

This project aims to apply an identification technique to real-world problems. There are 4 topics selected from various applications, namely, (i) structural equation modeling for fMRI data (ii) time series modeling for EEG data (iii) areal precipitation and (iv) river flow peak. Students spend six weeks on this project where each week a progress report must be submitted. The assignment

Instruction to students

Each week students are supposed to submit the weekly progress report to CourseVille in Assessment section, where *everyone* must submit the group report in typesetting and .pdf format. Any proposed ideas, comments, or practical considerations should be stated in the report to keep track of your concerns or problem found as the work progresses. The use of LaTeXto type the document is highly encouraged. The progress reports should be named as `sem_fmri_week1.pdf` `eeg_week1.pdf` `rainfall_week1.pdf` and `peakflow_week1.pdf` where the week number refers to the week of submitting the report.

Final report format

All groups should use the following formats:

- Use the report template from Rainfall group (Tanut, Pongsorn, Petchakrit)
- All vector graphics (such as MATLAB figures) must be in .eps or .pdf format.
- In the title, put "2102531 Term Project Report" "Semester 1/2015" and my name. For Rainfall and Flow peak groups, also put "Piyatida Hoisungwan". Put our affiliation "Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University". Tanut can adjust these elements first and other groups can just follow the template.
- The abstract of the report summarizes what this project is about. State the problem, the goal and the final results you found.
- All MATLAB codes must be in the appendix with proper comments explaining the codes. You can use `listings` package for source code listing in LaTeX. Mark the language (MATLAB) to highlight the codes. Each MATLAB file should be referred to the section where it is used, or you can explain which function is used to generate each figure.

1 Structural Equation Modeling for fMRI data

Student list:

1. Anupon Pruttiakaravanich
2. Tawan luprasong
3. Auaangkun Rangsikunpum
4. Pusit Suriyavejwongs

1.1 Progress deadlines

Wed Oct 21

- State problem description and relevant information
- Brief literature of fMRI data

Wed Oct 28

- Typesetting report of brief literature of fMRI data
- Example of fMRI data set, time plot, brain map
- Estimation formulation (nonlinear optimization)
- The proposed estimation problem in convex framework

Wed Nov 4

- Fix typos and LaTeXtypesetting in the report of last week. See my suggestions in the paper.
- Present some existing softwares to solve optimization problem in SEM. Pick 1-2 softwares and explain what they can do. What optimization techniques do they use in those softwares (in principle)?
- Find an example of fMRI data set and use the software you found to solve the problem. Report the estimated X, A, Ψ and possible present the path coefficient matrix A in a graphical model of SEM.

Wed Nov 11

- Reorganize the report. Create Appendix section and move some details of mathematical definitions into this section.
- Correct some LaTeXtypesetting. Rewrite some sentences that are currently not understandable. Elaborate with more details when presenting some new results or information.
- Select SEM softwares that allow to solve the problem with no constraint on the number of link in path diagram (A is fully dense). Compare the estimated Σ, Ψ, A with the ones obtained by convex framework.
- In SEM softwares, there seems to be many options on estimation methods (ML, weighted LS, and so on). For each of them, find the corresponding estimation formulation. Make sure that when comparing the estimated parameters from the software and from our formulation, we must solve the same problem (same cost objective).

Wed Nov 18

- Fix the experiment on comparing the estimated paramters by LISREL and the convex program. Change the value of α to be as large as possible.
- Estimate the path coefficients from several data sets. From each data set, you would get an estimated dependence structure by reading the zero pattern in A. Explain a way to detect its small values (and normalize the scale of A_{ij}.) Of course, we will get different dependence structures (meaning different models with different complexities). Explain how to design an experiment on model selection, so that you can select a good 'dependence structure' in this sense.
- Write efficient MATLAB codes and revise the report as suggested.
- Clean up and organize your codes. Explain in short and give comment to the code. Attach your code in the appendix. Use `verbatim` environment.

Wed Nov 25

- Consider more model candidates in the model selection problems to find a structure of path coefficient.
- Revise the report according to the comments.

2 Time series modeling for EEG data

Student list:

1. Nuntanut Raksasri
2. Akasit Aupaiboon
3. Pawarisson Jaitahan

2.1 Progress deadlines

Wed Oct 21

- Brief literature of EEG data
- ML estimation formulation for estimating multivariate autoregressive model

Wed Oct 28

- Typesetting report of brief literature of EEG data
- ML estimation formulation for estimating multivariate autoregressive model
- Example of estimation results
- Brief literature of time series modeling for EEG data. What are other models (rather than AR) that have been used to explain EEG dynamics?

Wed Nov 4

- Finish the last week assignment:
  - Typesetting report of brief literature of EEG data
  - ML estimation formulation for estimating multivariate autoregressive model
  - Example of estimation results. Correct your code or any mistake because the result seems incorrect.
  - Brief literature of time series modeling for EEG data. What are other models (rather than AR) that have been used to explain EEG dynamics? Write a paragraph about it with complete reference list. Select references from journal papers only.
- Study Wald test to perform hypothesis test on estimated AR coefficients. To make it simple, consider LS problem and our goal is to test whether $\hat{x} = (A^T A)^{-1} A^T y$ is zero in some entries. Write a summary about Wald statistics and how to perform the test.

Wed Nov 11

- Add more references on time series modeling of EEG signals. See the purpose of using such models.
- Correct your code when the calculation involve singular matrices.
- Vary p (the model order) and plot the trade-off curve.
- Solve the model selection problem. Choose p from AIC and BIC. Consider two choices of n: 50 and 10. When using $n = 10$, sampling the channels that cover almost all area of the head.
- Perform a simple Wald test on LS problem. Consider $y = Ax + \nu$ where x is choosen to be zero in some entries. Generate y and estimate x using LS method. Perform Wald test on $\hat{x}$ and detect the zero entries of x. Consider the two types of errors and represent the results using ROC. The accuracy of detect the correct zero entries should depend on the number of sample size and noise variance.
- Change the typesetting by using LaTeX. Use the default font (Time News Roman).

Wed Nov 18

- Finish the last week assignment.
  - Give a clear explanation on modeling choice in the reference paper. More review on AR models.
  - Correct your code when the calculation involve singular matrices.
  - Vary p (the model order) and plot the trade-off curve.
  - Solve the model selection problem. Choose p from AIC and BIC. Consider two choices of n: 50 and 10. When using $n = 10$, sampling the channels that cover almost all area of the head.
  - Finish the experiment on Wald test applied to a simple LS problem.
- Apply Wald test on estimated AR coefficients.
- Revise the report sections as suggested.
- Clean up and organize your codes. Explain in short and give comment to the code. Attach your code in the appendix. Use `verbatim` environment.

Wed Nov 25

- Finish the last week assignment.
- Revise the report according to the comments.

3 Areal Precipitation

Student list:

1. Petchakrit Pinyopawasutthi
2. Tanut Aranchayanont
3. Pongsorn Keadtipod

3.1 Progress deadlines

Wed Oct 21

- Problem statement
- Draft of estimation formulation

Wed Oct 28

- Redefine the problem statement. We define *surrounding grids* of an (i, j) grid of interest. Our assumption is that the (i, j) grid should have measurement influence from surround cells only. The weight from any ground stations and from any satellite stations should be the same. This results in having only two parameters to be estimated.
- Draft of estimation formulation, explained in matrix format. Propose to use the notation I_{ij} refered to as the index set of surrounding grids of the (i, j) grid which has $4, 6$ or 9 grids depending on the location of the (i, j) grid of interest. Define the index set of grids having ground station and, I^g_{ij}.
- Figure out how to map the indices from 2D into linear indices.

Wed Nov 4

- Redefine the index sets I, I^g and I_{ij}.
- Consider the cost functions:
  - $f_1(a, b) = \sum_{(i,j)\in I^g}(\hat{y}_{ij} - x_{ij})^2 + \sum_{(i,j)\in I\setminus I^g}(\hat{y}_{ij} - z_{ij})^2$
  - $f_2(a, b) = f_1(a, b) + \text{TV}$ where TV is the penalty on the difference of rainfalls from adjacent grids.

  Define the cost function based on the data from one year first.
- From your notation, $X = [x_{ij}]$ and $Z = [z_{ij}]$ and from the relation

$$\hat{y}_{ij} = a \sum_{(i,j)\in I^g\cap I_{ij}} x_{ij} + \sum_{(i,j)\in I_{ij}} z_{ij}$$

  Then we can write $\hat{y}_{ij} = a\mathbf{1}^T\mathbf{e}_i^T X\mathbf{e}_j\mathbf{1} + b\mathbf{1}^T\mathbf{e}_i^T Z\mathbf{e}_j\mathbf{1}$ where $\mathbf{e}_i = \begin{bmatrix}0 & 0 & \cdots & I_3 & 0 & 0\end{bmatrix}^T$ (standard unit block vector). Verify your matrix notation $\hat{Y} = T(aX + bZ)T^T$ where T is tridiagonal matrix. But I think it should be

$$\hat{Y} = U^T(aX + bZ)V$$

where $U \in \mathbf{R}^{m\times m}, V \in \mathbf{R}^{n\times n}, X \in \mathbf{R}^{m\times n}, Z \in \mathbf{R}^{m\times n}$ and

$$U^T = \begin{bmatrix} \mathbf{1}^T & & & \\ & \mathbf{1}^T & & \\ & & \ddots & \\ & & & \mathbf{1}^T \end{bmatrix}, \quad V = \begin{bmatrix} \mathbf{1} & & & \\ & \mathbf{1} & & \\ & & \ddots & \\ & & & \mathbf{1} \end{bmatrix}, \quad \mathbf{1} \in \mathbf{R}^3$$

(please check maybe i'm wrong)

- Solve the estimation problem with the cost function $f_1(a, b)$ (simplest one) first. Report the result.

- Present the rainfall data in grid with some color map. After you do the estimation, present $\hat{y}$ in grid and compare with the data as well.

Wed Nov 11

- We observe that the data from ground station and satellite are significantly different in magnitude scales. This results in a negative value of $\hat{a}$ (the weight from ground station) which is opposed to our intuition.

- The weight from ground station should take the number of surrounding stations into account.

- Change the cost objective so that $\hat{y}_{ij} = \sum a_{ij}x_{ij} + b\sum z_{ij}$. We allow the weights from ground station to be different in each grid.

- If the estimation result leads to a jump value from one grid to another grid, we may need to consider addting TV term.

Wed Nov 18

- Keep the current experiment results.

- Last time we have an example plot of rainfall data from gauge and satellite from a single day. Keep it in the report. Moreover, can you check the measurment difference between the two sources if they have the same trend almost every day? Ajarn Piyatida said the satellite tends to be underestimate during the rainy season. You found that Satellite data is overestimated during dry season. Maybe illustrate the plot of averaged X and Z over all days in a season and compare their values.

- Consider a revised cost objective where the term $\sum_{(i,j)\in I\setminus I_g}(\hat{y}_{ij} - z_{ij})^2$ is weighted by $\gamma < 1$, meaning that we put less penalty on fitting $\hat{y}$ to satellite measurments which are much less accurate than ground measurements. First, put no constraints on the positiveness of a_{ij} and see if the uncontrained solution satifies this.

- Revise the prediction equation. Vectorize the variables and use the same vectorizing order as MATLAB does (vectorizing along columns). You can say things like in our report, vectorization is always done in this manner.

$$\begin{aligned} \hat{y} &= (\hat{y}_{11}, \dots, \hat{y}_{m1}, \hat{y}_{12}, \dots, \hat{y}_{m2}, \dots, \hat{y}_{1n}, \dots, \hat{y}_{mn}) \\ z &= (z_{11}, \dots, z_{m1}, z_{12}, \dots, z_{m2}, \dots, z_{1n}, \dots, z_{mn}) \end{aligned}$$

The variable x is obtained by stacking x_{ij} for only $(i, j) \in I_g$ into a column vector. Since it's hard to write a clear definition of x, you can just explain by giving an example of I_g and x.

- Use the expression $\hat{y} = Ax + bz$ where A, b are the variables. Explain the zero structure of A. We should have an efficient MATLAB function that takes I_g, m, n as inputs and returns the matrix A with binary entries as the output where 1 is the free entry of A and 0 represents the zero constraint on A. We note that the zero structure of A is very important and should be defined as some index set I_A.
- Define a permutation matrix P such that $P\hat{y}$ is arranged into

$$P\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix}$$

  where $\hat{y}_1$ is a vector consisting $\hat{y}_{ij}$ for $(i,j) \in I_g$ and $\hat{y}_2$ contains $\hat{y}_{ij}$ for $(i,j) \notin I_g$. Note that the length of $\hat{y}_1$ is the cardinality of I_g. This goes the same for $Pz = (z_1, z_2)$.
- Transform the equation by

$$\begin{aligned} \hat{y} &= Ax + bz \\ P\hat{y} &= PAx + bPz = \tilde{A}x + bPz = \begin{bmatrix} \tilde{A}_1 \\ \tilde{A}_2 \end{bmatrix} x + b \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} \end{aligned}$$

  The cost objective can be expressed as

$$\begin{aligned} f(A,b) &= \|\hat{y}_1 - x\|_2^2 + \gamma \|\hat{y}_2 - z_2\|_2^2 \\ &= \|\tilde{A}_1 x + bz_1 - x\|_2^2 + \gamma \|\tilde{A}_2 x + bz_2 - z_2\|_2^2 \end{aligned}$$

  Now the zero structure of $\tilde{A}$ is also changed. We index it by the set $I_{\tilde{A}}$. This initiates a simple MATLAB function that transform an index set to another under a permutation.
- Suppose I_1 is a *linear* index set specifying the sparse structure of a matrix $A \in \mathbf{R}^{m \times n}$. Let P be any permutation matrix of size $mn \times mn$ (that permute rows of A). Let I_2 be a linear index set specifying the sparse structure of PA. Write a MATLAB function that maps I_1 into I_2.
- Revise the report according to the comments in the report.
- Clean up and organize your codes. Explain in short and give comment to the code. Attach your code in the appendix. Use `verbatim` environment.

Wed Nov 25

- Add a detailed example of the expression of $\hat{y}$ as a linear mapping from x and z.
- Include the plots of X and Z in the introduction part.
- Include more results on varying γ and illustrate the plots as γ varies.
- The interpolation on y can be expressed as

$$\hat{y}_{ij} = \sum_{(k,l) \in I_g \cap I_{ij}} a_{ij,m} x_{kl} + b \sum_{(k,l) \in I_{ij}} z_{ij}$$

  where $a_{ij,1}, a_{ij,2}, \ldots, a_{ij,m}$ are unknown coefficients that depends on (i,j) and m is the cardinality of $I_g \cap I_{ij}$.
- Revise the report according to the comments.

4 River flow peak

Student list:

1. Jitin Khemvong
2. Tiwat Boonyaviwat
3. Tanakorn Kriengkomol

4.1 Progress deadlines

Wed Oct 21

- Problem statement
- Draft of estimation formulation

Wed Oct 28

- Check the formula for μ and α in Gumbel distribution (that Prof. Piyatida has shown us) whether it comes from the principle of Maximum likelihood or other concepts. From the paper by Yue, The Gumbel mixed model for flood frequency analysis, I think the formula is obtained from the method of moment (MM), which is another estimation method that is not difficult to understand but we did not cover this in class.
- Given the peak flow data, use ML estimation to estimate μ and α in Gumbel pdf: $f(x) = e^{-e^{-(x-\mu)/\alpha}}$. Explain if the optimality conditions can lead to analytical expression of μ and α.
- Use samples of peak flow data to computer a numerical solution of ML estimates. If ML estimates are not obtained in closed-form, check out commands `fminunc` to solve a nonlinear optimization or `fsolve` to find roots of nonlinear equations.
- Let p_1, p_2, p_3, p_4 be peak flows at Ping, Wang, Yom and Nan rivers, and p_N be the peak flow at C_2 station (at Nakhon Sawan). Suppose we would like to find a joint PDF of these 5 random variables, so that we can obtain the condition PDF $f(p_N|p_1, p_2, p_3, p_4)$. As data p_1, p_2, p_3, p_4 are measured, we can estimate the probabilities of any events in p_N. Check out multivariate Gumbel distribution if it's feasible to consider this approach.

Wed Nov 4

- Check the optimality condition for ML estimation of Gumbel distribution. I doubt the nonlinear equation in α you have derived is correct. When you compute the ML estimates using `fsolve`, did you check the Hessian matrix for local maximum?
- Try using several starting points when solving the nonlinear equation. Compare the likelihood functions computed from ML estimates and MM (method of moment) estimates. Compare the plots of Flood peak VS Reduced variate from the two methods (see the example of this plot in Yue (1999) page 92.
- Explain the concept of method of moments. Also explain the equation of $\mathbf{E}[X]$ of $\mathbf{var}(X)$ to see how the parameter γ arises in the equation.
- Rewrite the whole report. Make it readable and complete.

Wed Nov 11

- Reorganize the report sections
  - Introduction: Give the problem background. Explain the related variables: flow, flow peak and their notations. Insert an example of hydograph plot, routes of four rivers (can be drawn neatly by hand), scatter plot of flow peaks from the five station with correlation coefficients.
  - Problem description: Given the data during year XXX to year XXX, you would like to estimate i) marginal density functions ii) bivariate density function of two variables iii) joint density function of five variables. Explain available distribution functions used by the literature (see paper Kidson and Richards, "Flood frequency analysis: assumptions and altervatives")
  - Parameter estimation: Report the estimation results of MM and ML (in different subsections).
  - Appendix: The method of moments (explain basic concept for any random variable). Definition of gamma function, Euler's constant, some detail about the constant γ (no need to explain everything but you select to report the information that the reader can further read from the given reference.) The definition of multivariate Gumbel distribution.

- Your writing needs a huge improvement. A good report does not mean you have to type it with LaTeX. Your report lacks of overview explanation, transitions from one idea to another, and proper usage of English sentences. Everything that is not obvious to the reader (that you found or search it from the Internet) must be cited by a reference list. Refer to an equation by its equation number.

- Check LaTeXtypesetting of EP, $\mathbf{Cov}, \mathbf{Var}, \mathbf{E}$. Use $(,)$ properly. (See the comments in the report.

- Rewrite the reference list. Use BibTeX to help manage the paper list.

- Consider ML estimation of parameters in bivariate Gumbel distribution (see "The Gumbel mixed model for flood frequency analysis"). Consider Ping & Nan (by ajarn Piyatida suggestion) and Ping & Wang (because they have a high correlation coefficient). Compare the result with 'joint non-exceedance probabilities' as in Table 2 in the paper. Find a way to represent the results.

Wed Nov 18

- Keep the current estimation results.

- Perform ML estimation of five parameters, $\alpha_x, \alpha_y, \mu_x, \mu_y, \theta$ of the joint CDF Gumbel distribution. Compare the estimation result with the estimated parameters given by the formula in the paper.

- Ajarn Piyatida suggestion: Illustrate the plots of return period. Consider two cases: the return period computed from the marginal CDF of Chao Praya and the *conditional* return period of Chao Praya given one of the four rivers. The Cha Praya's capacity of the flow peak is $3,500$ m/s^2, and we would like to evaluate the peak value when $T = 100$, for example. Explain if we know the peak of Ping, Wang, Yom or Nan, how the conditional return perioud changed from the marginal value.

- Revise the appendix of MM and ML. Move the detail of these methods when applying to Gumbel distribution to the main section.

- Give short explantion on the summary of methods explored in the survey paper.
- Highlight the high correlation values in the scatter plot.
- Elaborate more on the problem description.
- Collapse many tables together and revise the sections as suggested in the report.
- Clean up and organize your codes. Explain in short and give comment to the code. Attach your code in the appendix. Use `verbatim` environment.

Wed Nov 25

- Extend the plot of marginal return period to 500 years.
- Compute the conditional return period using θ, α, μ from the formula and include the conditional return period plots.
- Is the conditional probability always less than the marginal probability in Gumbel variable?
- Revise the report according to the comments.