Clustering Time Course Microarray Data by Nonparametric fitting

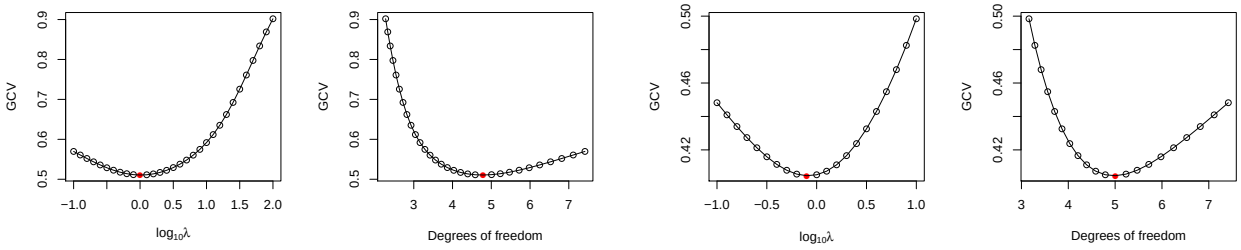
Xing Qiu

January 30, 2012

1 Notes

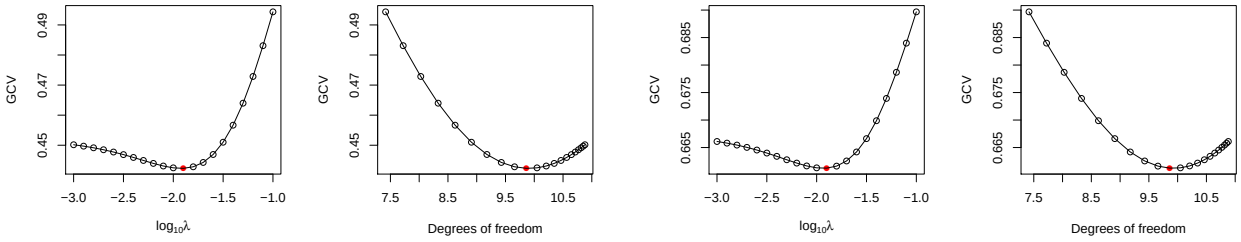
- (1/30/2012) Include `mikelung`, which is the 2,751 genes set selected by Mike Stover.
- (12/12/2011)
  1. Use Shuang's new data.
  2. blood and spleen.
  3. Now `ngenes` is a list.
- (11/17/2011)
  1. Analyzing the lymph node data.
  2. Many functions have been re-written and moved to `stepclust` package.
- (10/24/2011)
  1. Delay detection: use total variance instead of p-value/t-statistic as the criterion. This seems to give better results than the ones based on linear regression p-value visually.
- (10/06/2011) Hulin's new suggestions:
  - Detect regular/delayed genes first. The current *ad hoc* method: use stepwise linear regression and a suitable t-test for $\beta_1 = 0$.
  - For regular genes, basically re-run the two previous analyses:
    * Nonparametric fitting. Then classify them by activation days.
    * Parametric (damped oscillation) fitting. Then classify them by SK-means/XIC.
  - For delayed genes, use eyes to see their patterns.
- (9/30/2011) According to Hulin's new suggestions
  1. Nonparametric smoothing. Select smoothing parameter manually.
  2. Divide curves into two super-clusters: a) unimodal; b) bimodal/multi-modal.
  3. For unimodal ones, further divide them into 6 clusters: (early/middle/late) $\times$ (ups/downs).
  4. For bimodal-multi-modal ones, refit them by damped oscillations, then divide them into two super-clusters: a) damped and b) harmonic oscillations.
  5. For both damped and harmonic oscillations, I may further divide them into several clusters. I may have to manually select K, or I can also run XIC to find an *objective* K.

  Please see the earlier report for a brief introduction on the damped oscillation model.



(a) Blood

(b) Lung



(c) Lymph

(d) Spleen

Figure 1: GCV and degrees of freedom plots.

2 Data smoothing and quality control

Like before, curve fitting is done via B-spline with a smoothing parameter determined by the GCV criterion.

The optimum smoothing parameters and the corresponding degrees of freedom are summarized in Figure 1 and Table 1. The differences in the smoothing parameter as well as the corresponding

	blood	lung	lymph	spleen	mikelung
λ^*	1.00	0.80	0.01	0.01	0.80
DF	4.80	5.00	9.80	9.80	5.00

Table 1: Optimum smoothing parameter determined by GCV criterion and the corresponding degrees of freedom.

degrees of freedom among four datasets suggest that the lymph node and spleen data are less noisy therefore require less smoothing than the blood and lung data.

Figure 2 is a plot of residual sums of squares versus time. It shows that

1. Day 1 seems to be an outlier because of very high RSS for both data.
2. The RSS plot for lung and lymph node data are highly correlated. But the blood and spleen data have different RSS patterns.
3. The magnitude of these RSS are quite different. The blood and lung data have much (about 6

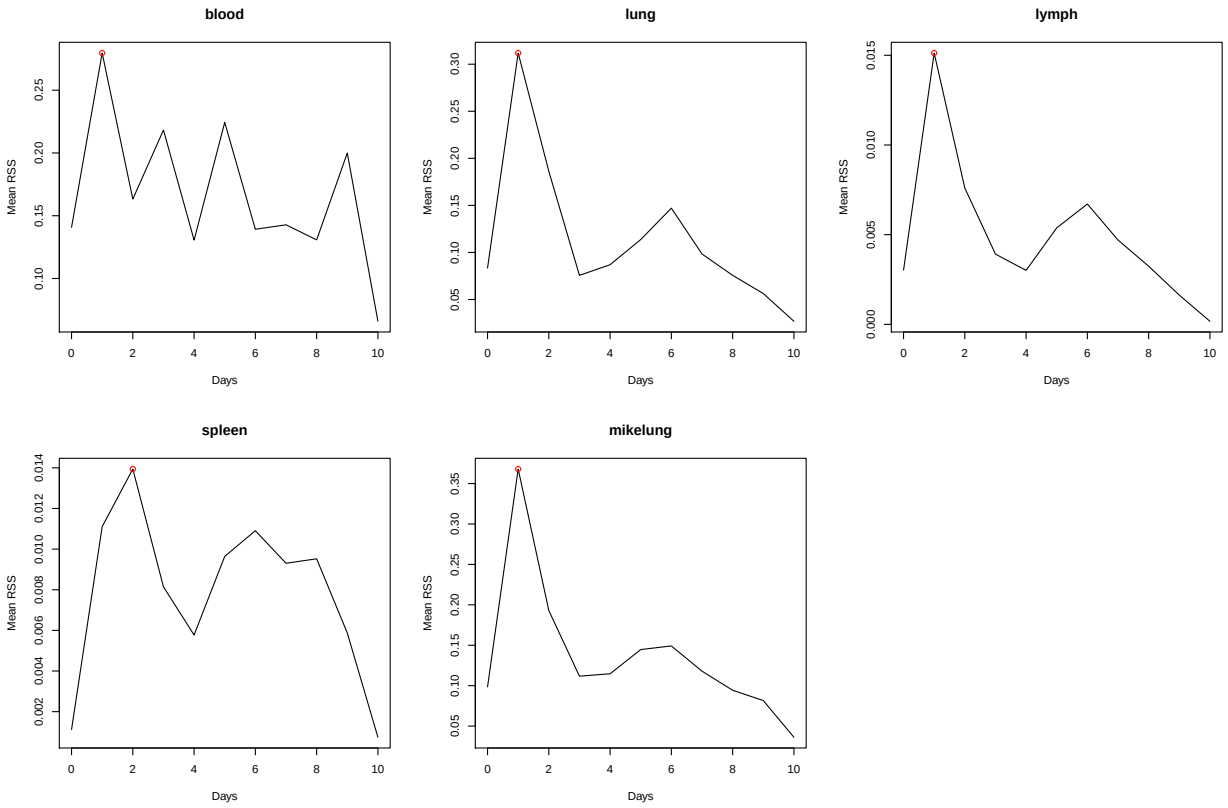


Figure 2: The mean residual sums of squares plot for each day. Residuals are defined as the difference between the observed data and the B-spline fits. Sample mean is taken over all genes.

to 7 times) larger RSS than the lymph node and spleen data. The true noise level between the two data are not that large because the blood and lung data receive much more smoothing.

So I decide to remove Day 1 from the data and refit the curves.

3 Regular v.s. delayed expressions

Gene expressions are first classified into the regular and delayed classes. I first transformed the data so that expression levels on day 0 are always zero. Then I conduct linear regressions passing through the origin for each gene from day 0 to days 3, 4, and 5.

$$x_{it} - x_{i0} = \beta_{1ik} t + \epsilon_{it}, \quad t = 0, 2, \ldots, k, \ k = 3, 4, 5. \tag{1}$$

The summation index start at Day 2 because Day 1 is already removed.

The t-statistic and the *adjusted* standard error of β_{1ik}, denoted as $s_{ik} := \sqrt{k-1}\widehat{\sigma}(\widehat{\beta}_1)$, are used for selecting regular (non-delayed) genes. I multiply the sample standard error of β_1 by $\sqrt{k-1}$ to make the three standard errors comparable. Without this standardization the SE on day 5 is much less than that on day 3.

I detect genes with delayed expressions based on the following *ad hoc* principles:

1. t-statistic less than 3 OR $\widehat{\beta}_{1ik} < 0.15$;

2. AND $s_{ik} := \sqrt{k-1}\widehat{\sigma}(\widehat{\beta}_{1ik}) < 0.12$, where $k = 3, 4, 5$ is the number of time points used in the regression.

Number of delayed genes detected:

Blood data: 0.

Lung data: 578.

Mike Lung data: 783.

Lymph node data: 55.

Spleen data: 0.

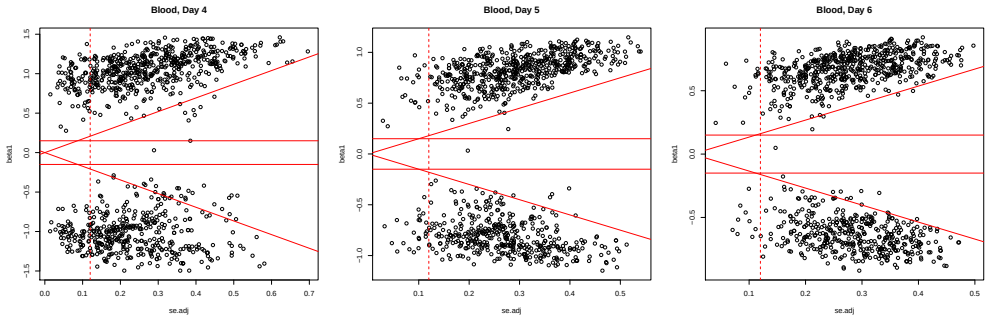
For the lung data, I also did manual adjustment to move genes with clear linear trend from the delayed genes to regular genes and *vice versa*. Originally, 63 were manually labeled as delayed genes and 3 were manual regular genes. By using this new criterion, only 8 out of 63 manual delayed genes needed manual adjustment. I still have do manually specify 3 genes as regular genes. After manual adjustment, 578 are labeled as delayed genes.

Figure 3 shows the scatter plot of adjusted residual standard deviation (s_{ik}) versus estimated slope coefficient ($\widehat{\beta}_{1ik}$). Manually adjusted genes (for the lung data) are highlighted as green (manual delay) or blue (manual regular) triangles. I think no manual adjustment is required for the blood, lymph node, and spleen data.

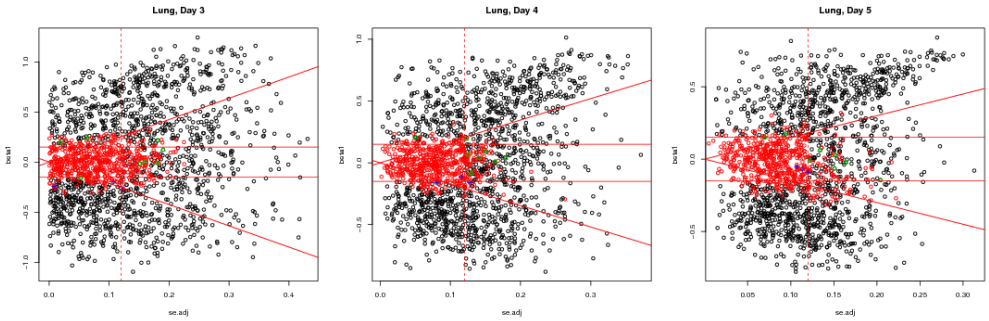
4 Number of modes (local minima/maxima)

The first clustering step is done using this heuristic algorithm.

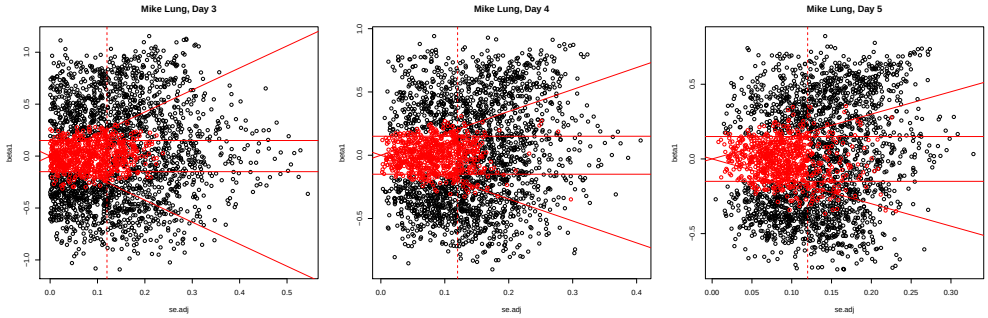
1. Smoothing by using penalized B-spline with $\lambda^* = 0.4$, which is selected manually.
2. For each *interior* time point ($t_i = 1, 2, \ldots, 9$), I call it a *turning point*[1] if for a given tolerance level $\delta = 0.1$,
$$\begin{cases} \widehat{g}(t_i) - \widehat{g}(t_{i-1}) > \delta, \quad \widehat{g}(t_i) - \widehat{g}(t_{i+1}) > \delta & \text{"V" case} \\ \widehat{g}(t_i) - \widehat{g}(t_{i-1}) < -\delta, \quad \widehat{g}(t_i) - \widehat{g}(t_{i+1}) < -\delta & \text{"}\Lambda\text{" case} \end{cases} \tag{2}$$
In other words, the smoothed expression values at (t_{i-1}, t_i, t_{i+1}) form either an "V" or a "Λ". Simple calculus shows that such a condition implies that there must exist *at least* one local minimizer ("V" case) or one local maximizer ("Λ" case) on the interval (t_{i-1}, t_{i+1}). Given our smoothing parameter, it is very safe to say that such a minimizer/maximizer is also *unique* on this interval. Furthermore, I think it is safe to say that no local maximizer/minimizer exists on the regular intervals.
3. I need to set a tolerance parameter because otherwise a very small up/down might be detected as a turning point.
4. Another exception: if $|\widehat{g}(t_{i+1}) - \widehat{g}(t_i)| < \delta$, we move to the next time point to check whether $|\widehat{g}(t_{i+2}) - \widehat{g}(t_i)| > \delta$.
5. For the lung data I consider those without turning points (the 0-mode group, 18 of them) as a special case of the 1-mode group. Furthermore, I consider the 3-mode group as a special case of the 2-mode group.
6. For the lymph node data the 0-modes group ($n = 40$) is merged to the 1-mode group; the 7-mode group ($n = 84$) is merged to the 6-mode group.



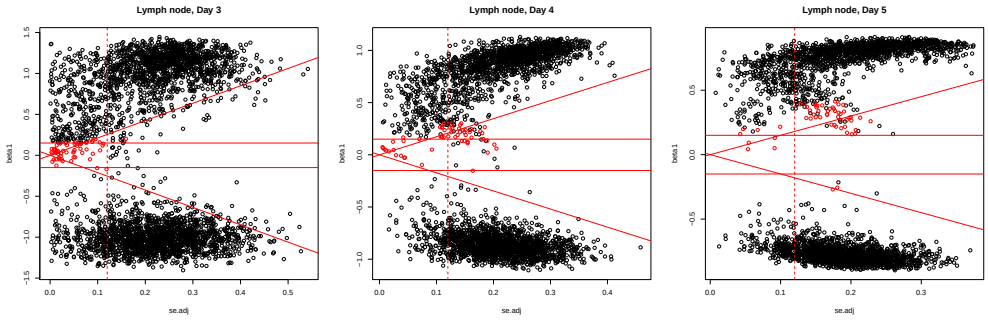
(a) Blood data



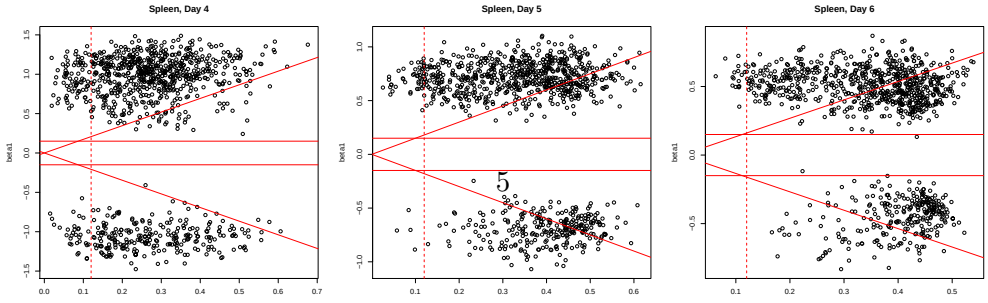
(b) Lung data



(c) Mike Lung data



(d) Lymph node data



	0-mode	1-mode	2-mode	3-mode	4-mode	5-mode	6-mode	7-mode
blood	103	722	125	50	0	0	0	0
lung	0	1431	475	0	0	0	0	0
lymph	0	319	733	702	900	548	262	84
spleen	0	0	102	331	270	297	0	0
mikelung	87	1699	902	63	0	0	0	0

Table 2: Number of genes in each mode group.

Number of genes in each mode-group is summarized in Table 2.

5 Classification based on the activation time

For each gene, I define the activation time as the time when it reaches reach 50% of its maximum/minimum expression levels.

- For the blood data, most genes (947 out of 1000) are activated on Day 2. I group genes activated on and after day 3 as one group because there are only 19 of them.
- For the lung data, I group genes activated on and after day 8 as one group because there are only 23 of them.
- For the lymph node data, I group the activation group on days 6 ($n = 21$) into Day 5.
- For the spleen data, only three activation days are possible: Days 2, 3, and 4. I group the activation group on days 4 ($n = 2$) into Day 3.

I made Table 3 to summarize these results. And the table of up/down regulated genes

	Day1	Day2	Day3	Day4	Day5	Day6	Day7	Day8
blood	190	750	60	0	0	0	0	0
lung	0	470	296	192	318	446	127	57
lymph	0	3175	85	198	90	0	0	0
spleen	127	743	130	0	0	0	0	0
mikelung	0	693	424	252	400	601	298	83

Table 3: Number of genes activated on each day.

5.1 Merge smaller clusters into larger clusters

I would like to reiterate that genes are classified into unique clusters based on the following information

1. Regular or delayed;

[1] I do not call them as critical points or local maxima/minima because they are only discrete approximations to the true critical points.

	Up	Down
blood	572	428
lung	980	926
lymph	1742	1806
spleen	721	279
mikelung	1598	1153

Table 4: Number of up/down regulated genes.

2. Up or Down regulated;
3. Activation day.
4. Number of modes of the temporal curve.

In theory, there can be 56 clusters for the lung data and 96 clusters for the lymph node data, etc. In reality I noticed there are some empty clusters and many clusters with only a handful of genes. So I decide to merge every small cluster ($n < 10$) to a large cluster ($n \geq 10$) which is most *similar* to it in terms of the shape of curves.

More specifically, I decide to use the L^2-distance between mean curves of each cluster as the measurement of similarity. A small cluster is merged to one of the large clusters with which such distance is the smallest. After merging I end up with fewer unique clusters for each data. The results are shown in Table 5.

	blood	lung	lymph	spleen	mikelung
Original	30	72	65	37	88
Merged	12	32	24	17	17

Table 5: Number of unique clusters in each compartment.

5.2 Other results

The following results are presented in separate files.

- Spaghetti plots for genes according to their clusters.
- Individual plots for regular/delayed genes.

References