Fitting Multiple Change-Point Models to Multivariate Data

Edgard M. Maboudou[1]
Douglas M. Hawkins[2]

(1) Department of Statistics and Actuarial Sciences, U of Central Florida, Fl
(2) Department of Statistics, U of Minnesota, MN

June 4, 2009

Introduction Overview and Significance

- Change-point detection problem has been studied for decades in many frameworks
- Change-point models may be appropriate when data sets have natural ordering
- Several methods for the detection of change-points are available–likelihood ratio, nonparametric, and Bayesian
- Methods are based on the availability of sequential data

Ideas about Change-point Models

- In a change-point model, sequence of data can be broken down into segments
- Observations follow the same statistical model within each segment
- Different models in different segments
- As examples, data follow a normal distribution whose parameters change from one segement to another
- Best known application of change-point modeling in data analysis is regression trees (Breiman et al., 1984)

Change-point Models

- Working with two-segment models (one change-point) is straightforward
- Dealing with more than one change-point complicates matters considerably–computational and inferential aspects
- One solution is the hierarchic binary splitting algorithm proposed by Vostrikova, 1981
- Hierarchic solution makes choices that are optimal at each step but not necessarily in terms of minimizing the overall residual sum of squares
- Hierarchic binary splitting, though fast, usually fails to give the optimum splits if there are two or more subsequences

Complexities Surrounding Change-Point Problems

Issues

- Choice of suitable parametric forms for the within-segment models
- Choice of segment boundaries or change points
- Determination of the appropriate number of change-points to use in modeling the specific data
- Discussion focuses on the second of these questions
- Third question is outside the scope of this work. Future research probably.

Multiple Change-Point Gaussian Model

Our working model

A1 : The p component vector $\mathbf{X}_i$, $i = 1, 2, ..., n$ is a sequence of independent distributed p-dimensional normal random vectors.

A2 : We have k subsegments with $k - 1$ change-points, $\tau = (\tau_1, \tau_2, \ldots, \tau_{k-1})$. For notational convenience, we will also bracket the whole sequence with notional change-points $\tau_0 = 0, \tau_k = n$.

A3 : The data within subsegment j is identically and independently distributed (i.i.d) multivariate normal with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$, this means for

$$\tau_{j-1} < i \leq \tau_j, \quad \mathbf{X}_i \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

Generalized Likelihood Ratio

- We are interested in the following hypothesis test
  H_0: No change-point, that is $k = 1$
  H_a: k subsegments with $k-1$ change-points
  $\tau = (\tau_1, \tau_2, \ldots, \tau_{k-1})$ and $\boldsymbol{\mu}_j \neq \boldsymbol{\mu}_{j-1}$ or
  $\boldsymbol{\Sigma}_j \neq \boldsymbol{\Sigma}_{j-1}, j = 2, \ldots, k$.
- all $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown
- k denotes the unknown number of segments

Generalized Likelihood Ratio

- Consider a segment $\mathbf{X}_{h+1}, \ldots \mathbf{X}_m$
- Denote its mean vector by $\bar{\mathbf{X}}_{h,m} = \sum_{j=h+1}^{m} \mathbf{X}_j/(m-h)$
- Its cross products matrix by $\mathbf{A}_{h,m} = \sum_{j=h+1}^{m} (\mathbf{X}_j - \bar{\mathbf{X}}_{h,m})(\mathbf{X}_j - \bar{\mathbf{X}}_{h,m})'$

Computing the Sample Covariance Matrix

- Given the boundaries τ_{j-1}, τ_j, the MLEs of the j^{th} segment are
-
$$\hat{\mu}_j = \bar{\mathbf{X}}_{\tau_{j-1},\tau_j}, \tag{1}$$
- and
$$\hat{\boldsymbol{\Sigma}}_j = \mathbf{A}_{\tau_{j-1},\tau_j}/r_j \tag{2}$$

where r_j is the segment length $\tau_j - \tau_{j-1}$.

Multiple Change-Point Gaussian Model

- Let $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \ldots, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Under assumptions A1-A3,

$$-2 \log L_1(\hat{\boldsymbol{\theta}}) = np(\log(2\pi) + 1) + \sum_{j=1}^{k} r_j \log |\hat{\boldsymbol{\Sigma}}_j| \qquad (3)$$

Optimization

- Ignore the irrelevant constant $np(\log(2\pi) + 1)$ and focus our attention on $\sum_{j=1}^{k} r_j \log|\hat{\mathbf{\Sigma}}_j|$
- Likelihood function is not continuous in τ_j
- Conventional function optimization procedures can not be used
- Can formulate the problem as a partitioning problem where the goal is to obtain the best partition of the grid $\{1, \ldots, n\}$ into k segments

Principle of Optimality

- Dynamic programming, introduced by Bellman and Dreyfus (1962), is a recursive approach based on the Bellman's principle of optimality
- Bellman's principle of optimality: "subpaths of optimal paths are themselves optimal"
- Objective function $S_k = \sum_{j=1}^{k} r_j \log|\hat{\mathbf{\Sigma}}_j|$ has the additivity property, called separability by Bellman
- Separability of the objective function allows us to draw an analogy to the shortest path problem.

Dynamic Programming

- Objective function S_k can be seen as the total length of a path connecting point 1 to point n
- Our task is to find the shortest path to travel from point 1 to point n with $k-1$ steps
- The steps are the change-points $\tau_1, \ldots, \tau_{k-1}$
- Next, using Bellman's principle of optimality, the breakpoints obtained are guaranteed to be global maximum.

Schwarz Information Criterion

- To be useful, the methodology needs a rule for deciding how many segments are needed to model the data
- Applying SIC to our problem yields

$$SIC(k) = -2\log L_1(\hat{\theta}) + \frac{p(p+3)(k-1)}{2}\log(n) \qquad (4)$$

- For each fixed k up to some maximal value K, compute the SIC using equation 4
- Number of change present in the dataset is the value k^* that minimizes $SIC(k)$

Example

Presentation of the data

- The steam turbine system data set is presented and discussed in Mason and Young, 2002
- It consists of a Phase I data set of 28 observations on a steam turbine system
- This Phase I data set was followed by 16 Phase II vectors
- Will ignore the distinction Mason and Young drew between the two sequences, and apply our methodology to the combined dataset of 44 observations

Example

- Measurements are made on the following variables: Fuel, Steam Flow, Steam Temp, MW, Cool Temp, and Pressure, so $p = 6$
- Want to investigate if all the 44 observations follow the same statistical distribution
- or if there are some change-points where the data follow different statistical distributions
- Our methodology is well suited to answer this question

Example

Results

- Use $K = 5$ and results of fit show below
-
k	SIC(k)	$\hat{\tau}$
1	1881.1	No change
2	1846.4	14
3	1873.2	14, 28
4	1936.6	15, 24, 31
5	2004.6	7, 15, 24, 31

Table: Estimated Change-points With All Data

Example

Analysis

- Minimum *SIC* occurs for $k = 2$, indicating that the series contains two natural segments – 1 to 14 and 15 to 44
- According to Mason and Young, observations 1 to 28 are assumed to be in-control
- If this claim is valid, there should not be break points between these 2 observations.
- The indicated two-segment model however split in the middle of the "in-control" segment
- This change-point is quite stable, with either observation 14 or 15 appearing in all four segmentations

Example

Estimation

- SIC suggests a single change-point model for this data ($SIC(k)$ attains its minimum for k =2) and the change-point is located at case 14.
- The estimation portion can be solved as follow. Given a change-point at case 14, segment 1 which consists of cases 1 to 14 has mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$
- Segment 2 consists of cases 15 to 44 and has mean vector $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$
- To see the actual difference between the two halves, we will look at the mean vector and covariance matrix of the studentized values

Example

- For segment 1, an estimate of the studentized mean vector is

$$\hat{\mu}_{z1} = (-0.28, -0.62, 1.06, -0.53, -0.62, 0.55)'$$

and an estimate of the covariance matrix is

$$\hat{\boldsymbol{\Sigma}}_{z1} = \begin{pmatrix} 0.21 & -0.002 & -0.01 & 0.01 & 0.07 & -0.01 \\ -0.002 & 0.006 & 0.02 & 0.00 & 0.008 & 0.002 \\ -0.01 & 0.02 & 0.15 & 0.00 & 0.03 & -0.008 \\ 0.01 & 0.00 & 0.00 & 0.003 & 0.01 & 0.006 \\ 0.07 & 0.008 & 0.03 & 0.01 & 0.27 & 0.17 \\ -0.01 & 0.002 & -0.008 & 0.006 & 0.17 & 0.43 \end{pmatrix}$$

Example

- For segment 2, an estimate of the studentized mean vector is

$$\hat{\mu}_{z2} = (0.13, 0.29, -0.49, 0.25, 0.29, -0.25)'$$

and an estimate of the covariance matrix is

$$\hat{\mathbf{\Sigma}}_{z1} = \begin{pmatrix} 1.33 & 1.11 & -0.38 & 1.09 & -0.041 & -0.51 \\ 1.11 & 1.203 & -0.38 & 1.15 & 0.25 & -0.76 \\ -0.38 & -0.38 & 0.61 & -0.37 & 0.09 & -0.19 \\ 1.09 & 1.15 & -0.37 & 1.28 & 0.18 & -0.65 \\ -0.041 & 0.25 & 0.09 & 0.18 & 1.09 & -0.78 \\ -0.51 & -0.76 & -0.19 & -0.65 & -0.78 & 1.08 \end{pmatrix}$$

Example

Comments:

- The numbers in the segment mean vectors and covariance matrices are dimensionless and so much easier to comprehend and compare
- The major distinction though is in the covariance matrix, where the second and fourth variables (Steam Flow and MW) show much less variance in the first segment than the second
- The covariances also show interesting structure

Conclusion

- This work presents an effective and fast method to solve the problem when the data can be represented by a Gaussian model
- Unlike the repeated hierarchic algorithm, our method finds the global optimum of the likelihood function
- Computational efficiency is achieved by using fast numerically stable update of the Cholesky factorization

Some References

[1] Bellman, R.E., Dreyfus, S.E., 1962. Applied Dynamic Programming. Princeton University Press. Princeton

[2] Chambers, J.M. (1971), "Regression Updating" JASA, 66, 744-748

[3] Hawkins, D. M., Maboudou-Tchao, E. M., 2007 Self-starting multivariate exponentially weighted moving average control charting for location. Technometrics, 49, 199-209

[4]Mason, R. L., Young, J. C., 2002. Multivariate Statistical Process Control with Industrial Application. ASA-SIAM series on Statistics and Applied Probability

[5] Yao, Y. C., 1988 Estimating the number of change-points via Schwarz's criterion. Statistics and Probability Letters, 6, 181-189