Fitting Multiple Change-Point Models to a Multivariate Gaussian Model

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

[1] Department of Statistics, University of Central Florida,
4000 Central Florida Blvd,
Orlando, FL 32816, USA
emaboudo@mail.ucf.edu

[2] School of Statistics, University of Minnesota,
224 Church Street S.E. ,
Minneapolis, MN 55455, USA
dhawkins@umn.edu

Abstract. Statistical analysis of change point detection and estimation has received much attention recently. The problem of detecting a single change-point is not an easy task. Multiple change-point models arise when there are more than one change-point. In a amultiple change-point problem, the number of changes and their location are unknown. This paper develops an exact algorithm to detect, estimate the change-points for multivariate Gaussian model.

Keywords. Binary splitting, Dynamic programming, Separability.

1 Introduction

Change-point models are becoming important in numerous research fields and practical applications including economics, finance, medicine, and mulitmedia processing. The change detection problem has been studied for decades in many frameworks. One can refer to the books of Basseville and Nikiforov, Brodsky and Darkhovsky, and Carlstein et al. for a detailed bibliography. Dealing with a single breakpoint is relatively straightforward. Likelihood ratio tests for several common models are easy to define. Theoretical results and asymptotics are given in Csorgo and Horvath, 1987. The methods used are based on the availability of a sequential data and the detection of the change point uses the past observations as the only available information. Dealing with more than one change-point complicates matters considerably, from both the computational and inferential aspects. One approach is the hierarchic binary splitting algorithm proposed by Vostrikova (1981). She proved that the algorithm is consistent. The hierarchic solution makes choices that are optimal at each step, but not necessarily in terms of minimizing the overall residual sum of squares. Therefore, the hierarchical segmentation procedure does not guarantee the optimum splits if there are more than two segments.

The multiple change-point problem has several issues

1. choosing suitable parametric forms for the within-segment models, deciding whether there is any change (hypothesis testing problem),
2. locating the segment boundaries (estimation problem),
3. and deciding the appropriate number of change-point (model selection problem).

In this paper, we study the problem of simultaneous changes in the mean vector and covariance matrix of a multivariate Gaussian model.

2 Multivariate change point model

Once a model is specified, our task is to estimate the location of the change-points and the parameters within segments. In this estimation step, the number of segments has to be fixed. We study the problem of simultaneous changes in the mean vector and covariance matrix of a multivariate normal data. We start by listing the assumptions.

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

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

(all $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown).

Inferentially, 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 $\boldsymbol{\Sigma}_j \neq \boldsymbol{\Sigma}_{j-1}, j = 2, \ldots, k$.

2.1 The maximum likelihood method

Under H_0, the log likelihood is

$$\log L_0(\boldsymbol{\theta}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \boldsymbol{\mu}) \tag{1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $|\boldsymbol{\Sigma}|$ represents the determinant of $\boldsymbol{\Sigma}$

Under H_a, the k subsegments are assumed to be statistically independent as the vector $\mathbf{X}_i$ has a p-dimensional multivariate normal distribution. So, the log likelihood is

$$\log L_1(\boldsymbol{\theta}) = -\frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{k} r_j \log|\boldsymbol{\Sigma}_j| - \frac{1}{2}\sum_{j=1}^{k}\sum_{i=\tau_{j-1}+1}^{\tau_j}(\mathbf{X}_i - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_j) \tag{2}$$

where $r_j = \tau_j - \tau_{j-1}$ is the number of vectors in the j^{th} subsegment, $\boldsymbol{\mu}_j$ is the unknown mean vector of the j^{th} subsegment, $\boldsymbol{\Sigma}_j$ is the unknown covariance matrix of the j^{th} subsegment, and $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \ldots, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Maximization of the likelihood is a two-step procedure. If the changepoints τ_j are given, then the MLEs of the mean vectors and covariance matrices follow from the standard one-way Manova layout.

Result Under the assumptions A1-A3, -2 times the log-likelihood, from which the multiple change-point for the multivariate Gaussian model of $\mathbf{X}$ under H_a can be deduced, is

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

Proof. Let substitute the estimated θ from eq (4) in (3)

$$\begin{aligned}
-2\log L_1(\hat{\boldsymbol{\theta}}) &= np\log(2\pi) + \sum_{j=1}^{k+1} r_j \log|\hat{\boldsymbol{\Sigma}}_j| + \sum_{j=1}^{k+1}\sum_{i=\tau_{j-1}+1}^{\tau_j} (\mathbf{X}_i - \boldsymbol{\mu}_j)'\hat{\boldsymbol{\Sigma}}_j^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_j) \\
&= np\log(2\pi) + \sum_{j=1}^{k+1} r_j \log|\hat{\boldsymbol{\Sigma}}_j| + \sum_{j=1}^{k+1} tr(\hat{\boldsymbol{\Sigma}}_j^{-1} \sum_{i=\tau_{j-1}+1}^{\tau_j} (\mathbf{X}_i - \boldsymbol{\mu}_j)(\mathbf{X}_i - \boldsymbol{\mu}_j)') \\
&= np\log(2\pi) + \sum_{j=1}^{k+1} r_j \log|\hat{\boldsymbol{\Sigma}}_j| + \sum_{j=1}^{k+1} tr(\hat{\boldsymbol{\Sigma}}_j^{-1} r_j \hat{\boldsymbol{\Sigma}}_j) \\
&= np\log(2\pi) + \sum_{j=1}^{k+1} r_j \log|\hat{\boldsymbol{\Sigma}}_j| + \sum_{j=1}^{k+1} pr_j \\
&= np\log(2\pi) + \sum_{j=1}^{k+1} r_j \log|\hat{\boldsymbol{\Sigma}}_j| + np \\
&= np(\log(2\pi) + 1) + \sum_{j=1}^{k+1} r_j \log|\hat{\boldsymbol{\Sigma}}_j|
\end{aligned}$$

For the optimization step, we omit the irrelevant constant $np(\log(2\pi) + 1)$ and focus our attention to $\sum_{j=1}^{k} r_j \log|\hat{\boldsymbol{\Sigma}}_j|$. The likelihood function is not continuous in τ_j so usual optimization techniques fail. Instead, the problem can be formulated as a partitioning problem where the goal is to obtain the best partition of the grid $\{1, \ldots, n\}$ into k segments. Once the breakpoints τ_j are found, we can estimate the corresponding $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$.

2.2 Dynamic programming solution

Dynamic programming was introduced by Bellman and Dreyfus (1962) and is a recursive approach based on the Bellman's principle of optimality. Auger and Lawrence (1989) were the first to use it in the context of segmentation problems. The objective function is $S_k = \sum_{j=1}^{k} r_j \log|\hat{\boldsymbol{\Sigma}}_j|$. The dynamic programming takes advantage of the additivity of the objective function. This additivity property was called separability by Bellman. The separability of the objective function allows us to draw an analogy to the shortest path problem. The objective function S_k can be seen as the total length of a path connecting point 1 to point n, so 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 $\tau_1, \ldots, \tau k - 1$.

Let $\sum_{j=1}^{k} r_j \log|\hat{\boldsymbol{\Sigma}}_j| = \sum_{j=1}^{k} Q(\tau_{j-1}, \tau_j)$, the dynamic programming recursion is defined as follows:

1. $F(1, m) = Q(0, m), \;\; m = p + 1, p + 2, \ldots, n$
2. $F(k, m) = \min_h F(k - 1, h) + Q(h, m), \;\; m = p(k - 1) + 1, \ldots, n.$

Each segment must contain at least $p+1$ readings to ensure that the covariance matrix is nonsingular. Next, using Bellman's principle of optimality "subpaths of optimal paths are themselves optimal", the breakpoints obtained are guaranteed to be global maximum.

3 Number of change-points

To be useful, the methodology also needs a rule for deciding how many segments are needed to model the data. While the objective is a maximized likelihood, and is in principle amenable to likelihood ratio testing, the problem does not satisfy Cramer regularity, and so conventional asymptotics are inapplicable. Several methods to select the number of change-points are available in the literature. Yao, 1988 proposed a method using the Schwarz's information criterion where the number of change-points k is chosen to minimize a penalized likelihood function. We propose using the SIC and adding the penalty term $p(p+3)(k-1)\log(n)/2$ proposed by Chen and Gupta, 2000 to -2 log-likelihood (equation **??**). Also, SIC gives an asymptotically consistent estimate of the order of the true model. Applying SIC to our problem yields

$$SIC(k) = np(\log(2\pi)+1) + \sum_{j=1}^{k} r_j \log|\hat{\boldsymbol{\Sigma}}_j| + \frac{p(p+3)(k-1)}{2}\log(n) \tag{4}$$

For each fixed k up to some maximal value K, we compute the SIC using equation **??** and the number of change present in the dataset is the value k^* that minimizes $SIC(k)$.

4 Example

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. Measurements are made on the following variables: fuel usage (Fuel), the amount of steam (Steam Flow) produced, the steam temperature (Steam Temp), the megawatt-hour production (MW) of the turbine, the coolant temperature (Cool Temp), and the absolute pressure (Pressure) observed from the condenser, so we have $p = 6$. Mason and Young claim that this baseline Phase I data is in control and then can be used to calibrate the control chart for the phase II monitoring.

This Phase I data set was followed by 16 Phase II vectors. We will ignore the distinction Mason and Young drew between the two sequences, and apply our methodology to the combined dataset of 44 observations.

We execute the segmentation fitting $K = 5$ segments, obtaining as by-products the optimal change-points for 2 and 3 segments. The results of our fit in the table **??** 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 1. Estimated Change-points With All Data

1. The SIC suggests a single change-point model for this data ($SIC(k)$ attains its minimum for k =2, cf table) and the change-point is located at case 14.
2. 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 $\boldsymbol{\Sigma}_1$. An estimate for $\boldsymbol{\mu}_1$ is

$$\hat{\boldsymbol{\mu}}_1 = (236900, 178000, 849, 21, 54, 29)'$$

and an estimate of $\boldsymbol{\Sigma}_1$ is

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 10232560 & -72202.73 & -204.41 & 41.77 & 178.82 & -8.59 \\ -72202.73 & 126635.1 & 245.47 & -0.045 & 12.55 & 1.09 \\ -204.41 & 245.47 & 1.19 & 0.00 & 0.033 & -0.0022 \\ 41.77 & -0.045 & 0.00 & 0.0009 & 0.002 & 0.0003 \\ 178.82 & 12.55 & 0.033 & 0.002 & 0.031 & 0.0055 \\ -8.59 & 1.09 & -0.0022 & 0.0003 & 0.005 & 0.0037 \end{pmatrix}$$

Segment 2 consists of cases 15 to 44 and has mean vector $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$. An estimate for $\boldsymbol{\mu}_2$ is

$$\hat{\boldsymbol{\mu}}_2 = (240000, 182200, 844, 21, 54, 29)'$$

and an estimate of $\boldsymbol{\Sigma}_2$ is

$$\hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 65274200 & 35746000 & -7403.54 & 4129.98 & -98.34 & -334.33 \\ 35746000 & 25336201.28 & -4828.76 & 2852.75 & 391.22 & -328.24 \\ -7403.54 & -4828.76 & 4.74 & -0.57 & 0.08 & 0.05 \\ 4129.98 & 2852.75 & -0.57 & 0.37 & 0.03 & -0.03 \\ -98.34 & 391.22 & 0.08 & 0.03 & 0.13 & -0.02 \\ -334.33 & -328.24 & 0.05 & -0.03 & 0.02 & 0.009 \end{pmatrix}$$

5 Conclusion

Multiple change-point problems are of interest in different areas of data analysis. This paper presents an effective and fast algorithm to solve the problem when the data can be represented by a Gaussian model. It solves the algorithmic problem of finding the optimal heteroscedastic segmentation, and is helpful for retrospective segmentation of multivariate sequences.

References

Abramowitz, M., Stegun, I. (1970). *Handbook of Mathematical Functions. Dover Publications.* Inc., New York

Bellman, R.E. (2003). *Dynamic Programming.* Courier Dover Publications.

Birge, L., Massart, P., (2001). Gaussian model selection. *Journal of the European Mathematics Society*, **3**, 203–268.

Braun, J. V., and Hans-Georg Muller, H.-G., (1998). Statistical Methods for DNA Sequence Segmentation. *Statistical Science*, **13**, 142–162.

Braun, J. V., Braun, R. K., and Muller, H. -G., (2000). Multiple Changepoint Fitting via Quasilikelihood, with Application to DNA Sequence Segmentation. *Biometrika*, **87**, 301–314.

Chambers, J. M., (1971). Regression Updating. *Journal of the American Statistical Association*, **66**, 744–748.

Chen, J., Gupta, A. K., (2000). *Parametric Statistical Statistical Change-point Analysis.* Birkhauser Publications

Csörgö, M., Horvath, L., (1997) *Limit Theorems in Change-Point Analysis.* Wiley, New York

Gu, C., Wang, J., (2003). Penalized likelihood density estimation: Direct cross-validation and scalable approximation. *Statistica Sinica*, **13**, 811–826.

Johnson, N. L., Kotz, S., (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* Wiley, New York.

Hawkins, D. M., Ten Krooden, J. A., (1979). Zonation of sequences of heteroscedastic multivariate data. *Computers in Geoscience*, **5**, 189–194.

Hawkins, D. M., (1976). Point estimation of the parameters of piecewise regression models. *Applied Statistics*, **25**, 51–58.

Hawkins, D. M., (2001) Fitting multiple change-point models to data. *Computational Statistics & Data Analysis*, **37**, 323–341.

Hawkins, D. M., Maboudou-Tchao, E. M., (2007) Self-starting multivariate exponentially weighted moving average control charting for location. *Technometrics*, **49**, 199–209.

Lavielle, M., (2005) Using penalized contrasts for the change-point problem. *Signal Processing*, **85**, 1501–1510.

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

Srivastava, M. S. and Worsley, K. J., (1986). Likelihood ratio tests for a change in the multivariate normal mean. *Journal of American Statistical Association*, **81**, 199.

Sen, A., Srivastava, M. S., (1973). On multivariate tests for detecting change in mean. *Sankhya*, A **35** 173–186.

Tibshirani, R., (1986) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.

Venter, J. H. and Steel, S. J., (1996). Finding multiple abrupt change points. *Computational Statistics and Data Analysis*, **22**, 481–504

Vostrikova, L. J., (1981). Detecting disorder in multidimensional random processes. *Soviet Math. Dokl*, **24**, 55–59.

Worsley, K. J., (1982). An improved Bonferroni inequality and applications. *Biometrika*, **69**, 297–302.

Yao, Y. C., (1988) Estimating the number of change-points via Schwarz's criterion. *Statistics and Probability Letters*, **6**, 181–189.