Segmentation of well-log data

Danilo R. Velis, Facultad de Ciencias Astronómicas y Geofísicas, Universidad Nacional de La Plata, Argentina.

This paper was prepared for presentation at the 9th International Congress of the Brazilian Geophysical Society held in Salvador, Brazil, 11–14 September 2005.

Abstract

In this paper we present a method to automatically detect stationary segments in well-log data sequences. This process is carried out by searching for change points which correspond to abrupt changes in the statistical nature of the underlying process. For this purpose the we analyze the behavior of the probability density functions (pdf) of two adjacent sub-samples as they move along the original data sequence. A statistical test is used to set a significance level of the probability that the two distributions are the same, thus providing a means to decide how many segments comprise the data by keeping those change points that yield low probabilities. Examples using simulated and real well–log data show that the results are in good correspondence with what seems to be a reasonable segmentation.

Introduction

Segmentation is an important data mining process. One important application is the identification of locally stationary intervals, or, equivalently, the location of change points. In this context, segmentation (also known as zonation) is the dividing of a sequence into relatively homogeneous and stationary intervals, such that each segment is distinctive from the adjacent ones. Well logs can be subdivided into relatively uniform segments that represent zones of constant lithology (stratigraphic units and formations). Segment boundaries can be associated with abrupt changes in the layering, and conform the limits of relatively stable periods.

There are various strategies for addressing this segmentation problem. Classical approaches include the detection of abrupt changes in the mean or in the variance. For a brief description of these techniques see Davis (1986). Other strategies are based on the use of spectral analysis for identifying stationary intervals (Ligges et al, 2002). The method presented here takes into account both the mean and the variance, and also higher-order robust statistics such as certain non conventional skewness and kurtosis measures (Velis, 2003). Essentially, a split window is moved along the sequence and the probability density functions (pdf) of the two adjacent half-windows are compared. When a significant difference is detected, a change point is identified. Smooth pdfs are estimated using the maximum entropy method as described in Velis (2003), which guarantees robustness when dealing with short data sequences. Finally, a criterion for deciding which is the number of segments that comprise the data is proposed. The effectiveness of this strategy is supported by the analysis of various examples using simulated and real data sequences derived from well-logs.

The segmentation problem

Let $\vec{r} = (r_1, r_2, \cdots, r_N)$ be the sequence of well-log data. The objective of the segmentation process is to subdivide this sequence into smaller segments so that each interval is relatively locally stationary. That is, we look for the partition

$$\vec{r} = (\vec{s}_1, \vec{s}_2, \cdots, \vec{s}_M), \tag{1}$$

where $\vec{s}_j$, $j = 1, \cdots, M$, is the subset of $\vec{r}$ given by

$$\vec{s}_j = (r_{t_j}, r_{t_j+1}, \cdots, r_{t_{j+1}-1}), \tag{2}$$

and M, $M < N$, is the number of distinct segments that start at locations $(t_1, t_2, \cdots, t_{M-1})$, with $t_1 = 1$ and $t_{M+1} = N + 1$.

In practice the algorithm proceeds iteratively by searching successive change points $\{t_j\}$ based on the assumption that two adjacent intervals are distinct when the probability density functions (pdf) of the data on each side of t_j are significantly different. For this purpose, a split window of length $2W$ is centered at location t_j, and the corresponding pdfs are estimated and compared appropriately.

Here, W should be short enough to allow for the identification of short stationary intervals. Thus, a robust pdf estimation method that works well even for short data sequences is required. The maximum entropy (MaxEnt) method with moment constraints described in Velis (2003) produces smooth non-parametric pdfs which are consistent with the data. The approach utilizes robust statistics computed directly from the data to constrain the maximization of the pdf entropy.

The strategy for carrying out the segmentation is based on the sliding window approach, which consists on moving the

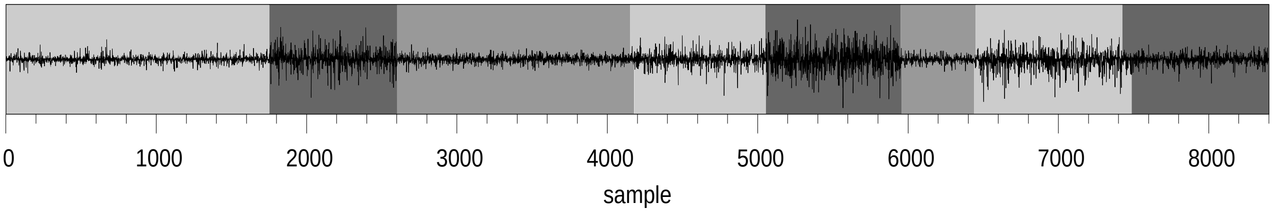

Figure 1: Simulated random sequence (8400 samples) comprised of eight statistical independent segments. The segmentation is indicated by gray blocks: true (top) and estimated (bottom). Table 1 shows the exact location of the change points.

analyzing window along the whole sequence, and assigning a change point when a significant difference between the pdfs is observed. To avoid the assigning of change points which are too close, we found it more appropriate to look for a single change point at a time. Starting with $j = 2$ (recall that $t_1 = 1$), we look for optimum change points until the next change point that is added does not yield a significant difference between the adjacent pdfs. These optimum change points correspond to the smallest probabilities along the whole sequence for the current iteration.

The segmentation algorithm

Let $\hat{t}_j$ be the current estimate of the $j-$th change point. Let $\vec{u} = (r_{\hat{t}-W}, r_{\hat{t}-W+1}, \cdots, r_{\hat{t}})$ and $\vec{v} = (r_{\hat{t}}, r_{\hat{t}+1}, \cdots, r_{\hat{t}+W})$ be the two subsets of $\vec{r}$ spanned by the split window, and $\hat{p}_u(\vec{u})$ and $\hat{p}_v(\vec{v})$ be the corresponding estimated pdf's. Rather than measuring the difference between $\hat{p}_u$ and $\hat{p}_v$, we measure the difference between their respective cumulative distribution functions (cdf), $\hat{P}_u$ and $\hat{P}_v$, using the Kuiper test. The Kuiper test, a variant of the well known Kolmogorov-Smirnov test (Press et al., 1992), quantifies the difference between two cdfs. The Kuiper statistics is

$$V = \max_{a \le r \le b} (\hat{P}_u - \hat{P}_v) + \max_{a \le r \le b} (\hat{P}_v - \hat{P}_u). \qquad (3)$$

where a and b define the region of support of the cdf (usually the minimum and maximum values in the data set). It turns out that the distribution in the case of the null hypothesis that the two data segments come from the same distribution can be calculated asymptotically, giving rise to a formula that allows one to compute the significance level (Press et al., 1992):

$$\text{Probability}(V > \text{observed}) = 2\sum_{i=1}^{\infty}(4i^2\lambda^2 - 1)e^{-2i^2\lambda^2}, \qquad (4)$$

where

$$\lambda = \left(0.155 + 0.24\sqrt{\frac{2}{W}} + \sqrt{\frac{W}{2}}\right) V \qquad (5)$$

The segmentation algorithm is a three stage process. In the first stage the probability (4) is calculated for every possible change point location throughout the whole sequence in the range $(W, N - W)$. In the second stage change points candidates are added according to the following strategy: at the beginning, the point with the smallest probability is selected as a candidate for the first change point, yielding t_2 and the new segmentation (t_1, t_2, t_3), which is comprised of two segments of lengths T_1 and T_2, respectively. Then, a new change point is added by selecting the smallest probability within the current longest segment (largest T_j), giving rise to a new partition (t_1, t_2, t_3, t_4). This process is repeated and new change points are added (within the longest segments obtained so far) until all segments are shorter than a given minimum length, T_{min}.

The third stage of the algorithm consists on discarding those change points whose associated probabilities are larger than a predefined threshold. Also, the change points with largest probabilities in excess of a predefined number of change points are deleted. Note that a large probability is indicative of a high degree of confidence on the null hypothesis that the two distributions are the same, so low values of probability are desired to obtain a high confidence on the hypothesis that the two distributions are different. Typical values are 95%-99%. To avoid too fine segmentations (i.e. two change points separated by a few samples), a minimum separation Δ between two consecutive change points is forced by adjusting the search range accordingly.

Numerical examples

As a consistency check, we applied the segmentation algorithm to the simulated sequence shown in Figure 1. This sequence was generated by concatenating samples drawn from eight different non-parametric distributions selected so as to simulate a realistic reflection coefficient series (Velis, 2003). In the segmentation process we set $W = 250$ and $\Delta = 200$, and change points were added until no segment was larger than $T_{min} = 200$ samples. At the end of the process, those change points with the associated probability larger than 0.01 were discarded. This significance level was chosen based on the inspection of Figure 2, where the probability (4) was plotted in ascending order for all the identified change points. For values larger than about 0.01, the probability of the null hypothesis that the two distributions are the same increases rapidly. The

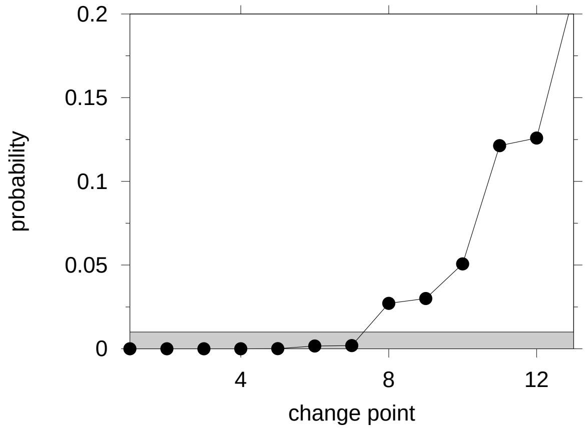

Figure 2: Probability of the null hypothesis that the two distributions are the same. The plot reveals an abrupt change at about 0.01, a value which is selected as a threshold to discard change points with high probabilities in the third stage of the segmentation process.

pdf	t_j	$\hat{t}_j$	V	Prob
1	1	-	-	-
2	1751	1751	0.360	0.00000
3	2601	2601	0.198	0.00162
4	4151	4179	0.196	0.00189
5	5051	5055	0.318	0.00000
6	5951	5957	0.469	0.00000
7	6451	6439	0.355	0.00000
8	7426	7487	0.230	0.00006

Table 1: The eight segments used to build the sequence shown in Figure 1 and their corresponding change points (true and estimated), Kuiper statistics and associated probability.

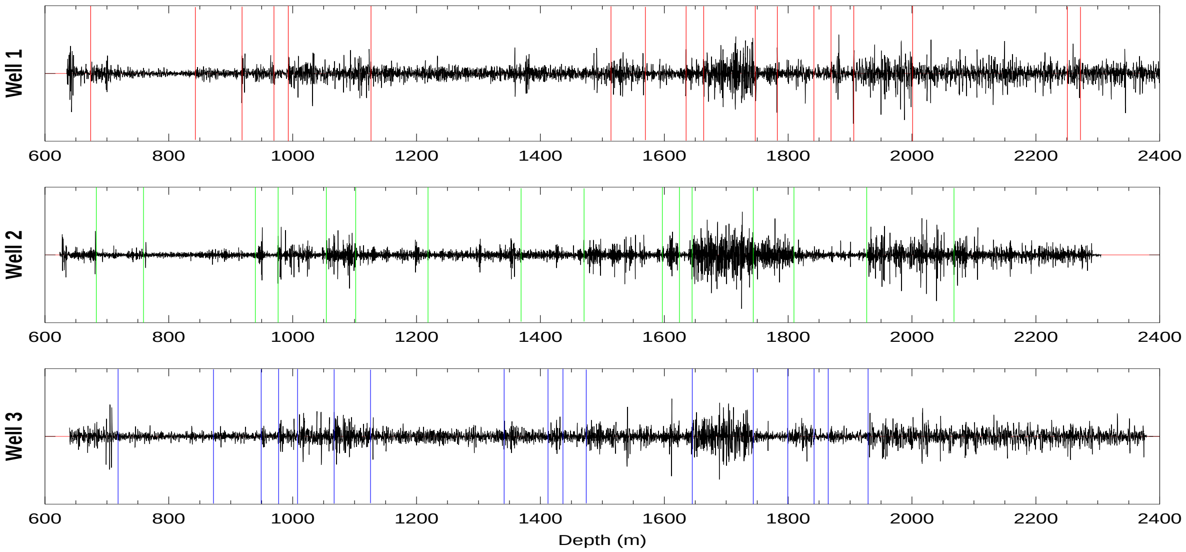

Figure 3: Reflectivity sequences. Vertical lines show the results of the segmentation algorithm.

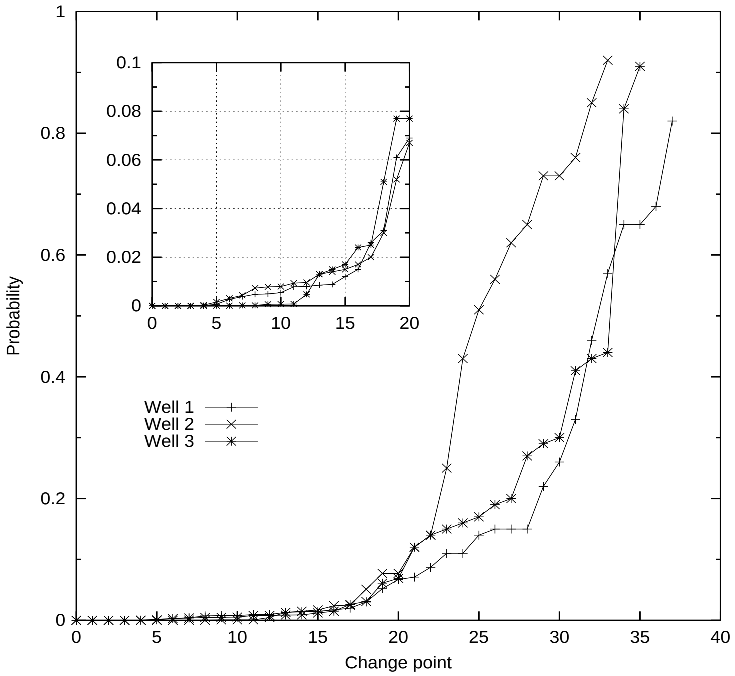

Figure 4: Probability of the null hypothesis that the two distributions are the same.

estimated change points are shown in Figure 1 and in Table 1, along with the correct change points. All eight segments were identified correctly.

Figure 3 shows various reflectivity sequences obtained from real density and sonic logs (sampling interval = 0.2 m). Clearly, several distinct segments can be distinguished. But it is very difficult to identify the exact location of the change points, and to decide which is the number of locally stationary segments that comprise each of the sequences.

The described algorithm was applied to carry out the segmentation using fixed parameters. In all cases $W = 100$ and $\Delta = 200$. Results are shown in the same figure, where 19, 17 and 18 segments were identified for wells 1, 2 and 3, respectively. Change points were added until no segment was larger than $T_{min} = 200$ samples. At the end of the process, those change points with the associated probability larger than 0.02 were deleted. This significance level was chosen based on the inspection of Figure 4, where the probability (4) was plotted in ascending order for all the identified change points. For values larger than about 0.02, the probability of the null hypothesis that the two distributions are the same increases rapidly.

Conclusions

The detection of stationary segments in well-log data sequences can be carried out in a quasi-unsupervised mode by searching for change points in the data. The MaxEnt method using robust non-conventional statistics that measure shape provides an appropriate technique to estimate the distributions that are to be compared. After estimating the distributions of the two halves of a moving window, abrupt changes can be identified based on the analysis of the probability of the null hypothesis that the two distributions are the same. The Kuiper test proved to be a useful criterion to decide which change points lead to significant differences between adjacent distributions. This provides a means of choosing the appropriate number of locally stationary segments that the data sequence can be subdivided into.

References

Davis, J., 1986, *Statistics and data analysis in Geology*, J. Wiley & Sons, Inc., New York.

Ligges, U., Weihs, C. and Hasse-Becker, P., 2002, Detection of locally stationary segments in time series – algorithms and applications: in *Proceedings in Computational Statistics (COMPSTAT2002)*, W. Härdle and B. Rönz (Eds.), Berlin, Germany.

Press, W., Teukolsky, S., Vetterling, W., and Flannery B., 1992, *Numerical Recipes in FORTRAN*, 2nd. Edition, Cambridge University Press.

Velis, D., 2003, Estimating the distribution of primary reflection coefficients: Geophysics, **68**, 1417–1422.