Bivariate Analysis Model for Incident Detection

Li & McDonald (2005) discovered that the joint distribution of travel time and travel time difference is bivariate normal in non-incident conditions. They developed a probe vehicle based algorithm using a bivariate analysis model to analyze travel time data for incident detection along four segments of motorways in UK. Travel time (T_i) and travel time difference ($\Delta T_i = T_i - T_{i-1}$) between adjacent time intervals are used. The bivariate normalT_i density function can be defined as,

$$f(T_i, \Delta T_i) = \left(2\pi\sigma_{T_i}\sigma_{\Delta T_i}\sqrt{(1-\rho_i^2)}\right)^{-1} e^{-k/2}$$

Where,

$$k = X^T \ \Sigma^{-1} X$$

$$X = \begin{bmatrix} (T_i - \mu_{T_i}) \\ (\Delta T_i - \mu_{\Delta T_i}) \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{T_i}^2 & \sigma_{T_i}\sigma_{\Delta T_i} \\ \sigma_{T_i}\sigma_{\Delta T_i} & \sigma_{\Delta T_i}^2 \end{bmatrix}$$

$$\text{Correlation coefficient, } \rho_i = \frac{Cov(T_i, \Delta T_i)}{\sigma_{T_i}\sigma_{\Delta T_i}} = \frac{\boldsymbol{E}\left[(T_i - \mu_{T_i})(\Delta T_i - \mu_{\Delta T_i})\right]}{\sigma_{T_i}\sigma_{\Delta T_i}}$$

The k value describes an ellipse in the $(T_i, \Delta T_i)$ plane with center at$\left(\mu_{T_i}, \mu_{\Delta T_i}\right)$. The k value is equal to Chi-Square value, $\chi^2(\alpha, 2)$. The elliptic contour will contain $100(1-\alpha)\%$ of the sample points on average. When a sample set of $(T_i, \Delta T_i)$ lies inside the contour, the following equation should be fulfilled and considered as non-incident data using 99% coverage (i.e., α=0.01).

$$X^T \ \Sigma^{-1} X \leq \chi^2(0.01, 2)$$

Similarly, we would like expand the above model to analyze travel time data for incident detection along equally divided roadway segments (e.g., 1 km). In addition, we proposed to use travel time (T_n) and travel time difference ($\Delta T_n = T_n - T_{n-1}$) between adjacent roadway segments where *n* is the n^{th} roadway segment.

Table C.1 Chi-Square Distribution Table

Chi-square	Probability					
Degrees of freedom (df)	0.01	0.05	0.1	0.9	0.95	0.99
1	0.00	0.00	0.02	2.71	3.84	6.63
2	0.02	0.10	0.21	4.61	5.99	**9.21**
3	0.11	0.35	0.58	6.25	7.81	11.34
4	0.30	0.71	1.06	7.78	9.49	13.28
5	0.55	1.15	1.61	9.24	11.07	15.09
6	0.87	1.64	2.20	10.64	12.59	16.81
7	1.24	2.17	2.83	12.02	14.07	18.48
8	1.65	2.73	3.49	13.36	15.51	20.09
9	2.09	3.33	4.17	14.68	16.92	21.67
10	2.56	3.94	4.87	15.99	18.31	23.21

Outlier Detection Using Hampel Identifier

Hampel (1971) introduced the concept of the breakdown point. The breakdown point is the smallest percentage of contaminated data (or outliers) that can cause an estimator to take arbitrary large aberrant values. The median and the median absolute deviation (MAD) are often recommended for robust estimations. For example, consider a data series, $\{x_i\}$, where i=1 to N. The MAD scale estimate (S) is defined as,

$$S = 1.482602\ median\{\ |x_i - x^m|\ \}$$

Where,

x^m is the median of a data series $\{x_i\}$,

Factor 1.4826 was chosen so the expected value of MAD is equal to the standard deviation (σ) for normally distributed data. That is,

$$\boldsymbol{E}[S(x_1 \ldots x_n)] = \sigma$$

For $\{x_i\}$ distributed as N (μ, σ^2) and large n.

If $|x_i - x^m| > t\ S$, x_i is considered as an outlier, where t is the rejection threshold often suggested to be around 2 to 5 as suggested by Pearson (2002).

We would like to use the Hampel identifier to remove potential speed outliers from GPS data processing.

References:

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematics Statistics*, 42, 1887–1896.

Li, Y. and McDonald, M., (2005). Motorway incident detection using probe vehicles, *proceedings of the Institute of Civil Engineers*, Transport 158, 11-15.

Perarson, R. K. (2002). Outliers in process modeling and identification. *IEEE Transactions On Control Systems Technology*, 10, 55–63.