Outlier detection in traffic data based on the Dirichlet process mixture model

NGAN, Henry Y T; Yung, Nelson H.C.; Yeh, Anthony G.O.

Published in:
IET Intelligent Transport Systems

DOI:
10.1049/iet-its.2014.0063

Published: 01/09/2015

Manuscript submitted to "**IET Intelligent Transport Systems**" on February 25, 2014.

Outlier Detection in Traffic Data Based on Dirichlet Process Mixture Model

Henry Y.T. Ngan*,
Department of Mathematics,
Hong Kong Baptist University,
Kowloon Tong, Hong Kong
Email: ngan.henry@gmail.com, Phone: +852-3411-2531

Nelson H.C. Yung,
Laboratory for Intelligent Transportation Systems Research,
Department of Electrical and Electronic Engineering
The University of Hong Kong, Pokfulam Road, Hong Kong
Email: nyung@eee.hku.hk

Anthony G.O. Yeh,
Geographic Information Center Research Center,
The University of Hong Kong, Pokfulam Road, Hong Kong
Email: hdxugoy@hkucc.hku.hk

*** Corresponding author**

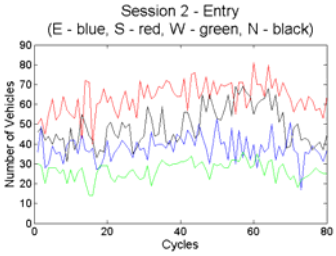
Abstract

Traffic data is exceedingly useful for road network management and is typically massive in size and full of errors, noise and abnormal traffic behaviors, which are regarded as outliers because they are inconsistent with the rest of the data. Hence the outlier detection (OD) problem is non-trivial. A novel method is presented for detecting outliers in large-scale traffic data by modeling it as a Dirichlet Process Mixture Model (DPMM). In essence, input traffic signals are truncated and mapped to a covariance signal descriptor, then its vector dimension is further reduced by Principal Component Analysis (PCA). This modified signal vector is then modeled by a DPMM. As traffic signals generally share heavy spatial-temporal similarities within signals and among various categories of traffic signals, classical OD methods are incapable to distinguish these similarities and to discern their differences. The contribution of this paper is to represent real-world traffic data (764,027 vehicles) by a generic DPMM-based method to perform an unsupervised OD to achieve a detection rate of 96.67% in a 10-fold cross validation.

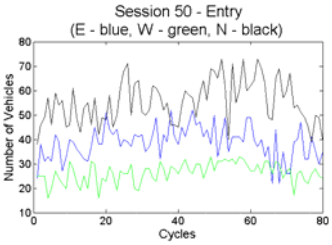
Index Terms—Dirichlet process mixture models, outlier detection, unsupervised learning, traffic flow analysis.

1 Introduction

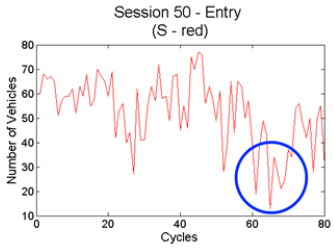
Traffic flow analysis (TFA) [1-3] based on massive traffic data, is to evaluate road dynamics in term of various statistics from highway [2, 4] and urban roads [2, 5]. These statistics include flow rates, volume, density, vehicle class, etc. to identify location of congestions and incidents in the network. TFA can be applied effectively in traffic forecasting [3, 5], control, design [2], incident detection [6] and management [7, 8]. Useful as it is, TFA often ignores the fact that massive traffic data may easily be contaminated by noise and errors (we call these 'outliers') during data acquisition, rendering the analytical results questionable. Specifically, an outlier may include all possible inconsistencies as anomaly found in traffic data can be due to one or more of the following conditions: (a) congestions or very small volume of vehicles, (b) incidents, (c) data capturing hardware failure, and (d) transmission errors. Moreover, due to the strong correlation between data captured from different parts of the road network, these outliers are not easy to be identified and removed. In this paper, a new method is proposed for cleaning up such outliers effectively. The goal of Outlier Detection (OD) [9] is to detect any data points appearing not consistent with the majority of the data (inliers). The problem is trivial if the features of the outliers and inliers are distinctive. Unfortunately, when the features between these two quantities are highly correlated, much more sophisticated mechanism is required to separate the outliers from the inliers. Fig. 1 depicts some real cases of normal and abnormal traffic patterns.



(a) Session 2 Entry: All signals are normal



(b) Session 50: Entries E, W, N are normal



(c) Session 50: Entry S is abnormal

Figure 1 Abnormal and normal traffic data: Session 2 for (a) 4 normal Entry (E, S, W, N) and (b) 4 normal Exit (E, S, W, N) signals; Session 50 for (c) 3 normal Entry (E, W, N). Blue circles in (e), (f) highlight the abnormal parts due to traffic congestions

In general, the captured data is assumed normally distributed which suits parametric modeling. Yet, not all data demonstrate such property in reality. These traffic signals have serious level of spatial-temporal similarities within themselves (Entry E signal in Fig. 1a seems repetitive) and between different signals (shapes of Exit signals in Fig. 1b look-alike). Furthermore, signals from different sessions would sometimes appear similar too (Entry N signal in Figs. 1a, 1c). However, abnormal signal behavior, due to traffic congestion highlighted by the blue circle (Signals of Entry S in Fig. 1c) may be very similar to other parts of a normal signal with minor differences. Such dissimilarities are almost unnoticeable and can be misclassified as normal variations quite easily. The major features of these spatial-temporal signals are random, casual, non-periodic and limited length. To overcome this serious level of spatial-temporal correlation, a generic data model is required.

According to the survey given in [9], OD methods cover areas including fault diagnosis, satellite image analysis, activity monitoring and others as well as some recent ones related to fabric defect detection [10] [11]. However, OD research straightly linked to TFA is not as popular [1, 6, 12-13], for which [1, 6, 13] are

supervised approaches for OD. Regarding OD algorithms, density-based, distance-based, and statistics-based algorithms were compared in [8] on outliers in traffic incidents or detection devices. Statistical based algorithms were found to be computationally quickest. Density-based algorithms give the best precision rate, but false negatives are high as well. Regarding OD's effectiveness, the original domain in presenting traffic data is usually employed to detect outliers [1, 8, 13] and only Cheng [12] transformed the data into another domain (e.g., wavelet-based). The detection rate may reach 100% with a manual control of selected cut-off value in [13], or 90% with a density-based approach in [8]. As much of spatial-temporal similarities exist in traffic data and the classical OD methods cannot effectively separate these similarities, exploiting traffic data in their original domain often arrives to unsatisfactory results. As such, it motivates us to investigate whether a transformation of traffic data into other domains for classification can improve OD. Also, there is plenty of room to investigate whether an unsupervised approach like DPMM can offer a better OD result.

A Dirichlet process (DP), labeled as $DP(\alpha_0, G_0)$ and denoted by 2 parameters: one in base probability measure (G_0) and one in scaling ($\alpha_0 > 0$) is a distribution over probability distributions. Remarkably, in a sample's DP, its posterior distribution is still a DP. As DP has these characteristics, it is assumed if a DP generates a model of traffic flow from some traffic data, samplings of a fusion in these models of traffic flow are the corresponding observations which could be grouped as DPMM. DPMM is indeed composed of countable infinite mixture models and a DP could be developed from each model. Iteratively, the clusters

number in DPMM can be explored. By this property, DPMM has recently become a widely discussed stochastic model in the applications of unsupervised clustering [14-19]. DPMM has been applied to many applications such as abnormal activity detection [14, 18], scene categorization [16], tracking maneuvering targets [15], trajectory-based video retrieval [17], abnormal events [19] and motion segmentation [20]. In short, only one case [20] of DPMM was utilized to detect outliers in a problem of motion segmentation in large-scale video data. Among 6 real databases, the highest total rate in errors is 2.1% while the lowest one is 8.5% in evaluation. It supports the view that DPMM is promising to identify outliers in large-scale traffic data. Furthermore, in regard to literature review, it lacks an OD method for traffic data in model-based approach, especially using more generic approaches. The DPMM-based approach is shown to be generic and effective in segmenting outliers from a dataset by the unsupervised property in clustering. Therefore, DPMM is selected because of its elegance in mathematics for the stochastic data and its feature of unsupervised clustering.

In the proposed DPMM-based method, traffic signals (i.e., signals of Entry at Fig. 1a) are extracted from the database firstly. Secondly, the signals are truncated, and then a covariance signal descriptor would describe their characteristics and be input as the data of the conjugate priors to the DPMM afterward. Thirdly, a dimension reduction for the vector of each signal is carried out by a PCA. Next, the DPMM would model the dimension reduced signal vectors. Lastly, the collapsed Gibbs sampler would achieve the outlier detection. In summary, this paper has two major contributions: (1) real-world traffic data is modeled

by a generic DPMM-based method for unsupervised OD; (2) an average detecting outlier accuracy of 96.67% in a 10-fold cross validation is achieved on a large traffic database of 764,027 vehicles for 19 traffic signals. Detailed parameters testing are carried to examine the reliability of the performance of the proposed DPMM-based method.

This paper is organized as the following structure: The proposed DPMM method is described in details in Section 2. Data and measurement metrics are presented in Section 3. Results are given in Section 4, where Section 5 draws conclusion from the research.

2 Outlier Detection Based on DPMM

In this paper, a traffic video database taken at a four-arm junction in Hong Kong is used. This database was collected over 31 days, with two sessions per day: AM (07:00-10:00) and PM (17:00-20:00), giving a total of 62 sessions. Four motion patterns (MPs) characterize the junction in a cycle under the operations of traffic lights. The traffic flow data in an Entry/Exit is marked as a traffic signal of the Entry/Exit volume in one session. Fig. 2 illustrates the junction map. In each arm of the junction, Entry signals are the summation of entry vehicle volume from all 4 traffic MPs per each cycle and are expressed as $\{z^1, z^2, z^3, z^4\}$ and exit signals are the sum of exit vehicle volume expressed as $\{z^5, z^6, z^7, z^8\}$ for arm E, arm S, arm W and arm N, individually. The signal of Entry Direction Distribution (EDD) is denoted as the sum of entry vehicle volume in 4 MPs per each cycle travelling from one arm towards the junction then to either turn left, right or straight afterward. They are given as

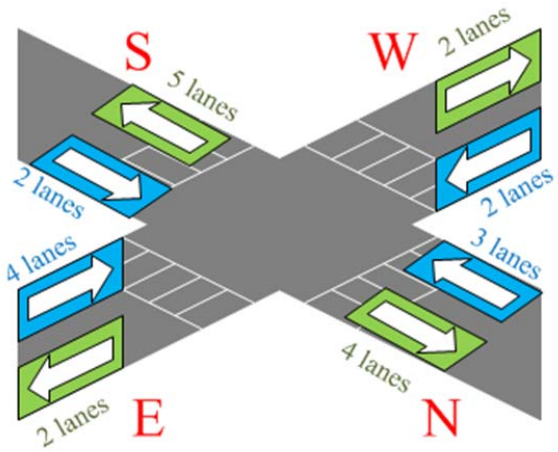


Figure 2 A 4-arm junction map of traffic flows in Exit and Entry directions of each arm marked by arrows in green and blue boxes, respectively. Number of lanes of each Entry/Exit is marked along the arrow signs

$\{z^9, z^{10}, z^{11}, z^{12}, z^{13}, z^{14}, z^{15}, z^{16}, z^{17}, z^{18}, z^{19}\}$ for EDDs E_l, E_r, E_s, S_l, S_r, S_s, W_l, W_s, N_l, N_r and N_s, individually. In the 4-arm junction, arm E, arm S, arm W and arm N are expressed by E, S, W and N, respectively while the subscripts l, r, s indicate left, right and straight directions, respectively. In the design of the 4-arm junction, there is no traffic flow for the EDD for W_r.

Mathematically, traffic signals may be represented as below:

Definition 1. Suppose that there are M sessions of observations (i.e. $M = 62$ in our database), a feature vector of an Entry, Exit, or EDD signal is expressed as $Z^i = (z_1^i, z_2^i, \cdots, z_M^i) \in \mathcal{R}^m$, where $i = 1, \cdots, 19$. A signal in one session with j cycles is denoted as

$$z_k^i = \{x_1^i, x_2^i, \cdots, x_m^i\} = \{x_r^i\}_{r=1,\cdots,m}, \quad (1)$$

where x_r^i is the vehicle number counted at the r^{th} cycle, $m = j_k$ is the cycle number at the k^{th} session.

Briefly, there are feature vectors expressed as $Z^1 = E_{entry}$, $Z^2 = S_{entry}$, $Z^3 = W_{entry}$, $Z^4 = N_{entry}$, for

Entry volume each arm each cycle, $Z^5 = E_{exit}$, $Z^6 = S_{exit}$, $Z^7 = W_{exit}$, $Z^8 = N_{exit}$, for Exit volume each arm each cycle as well as $Z^9 = E_l$, $Z^{10} = E_r$, $Z^{11} = E_s$, $Z^{12} = S_l$, $Z^{13} = S_r$, $Z^{14} = S_s$, $Z^{15} = W_l$, $Z^{16} = W_s$, $Z^{17} = N_l$, $Z^{18} = N_r$ and $Z^{19} = N_s$ for E-D-D per arm per cycle. E, S, W, N, l, r and s have the same meaning as those mentioned above.

The details of the mathematical modeling can be referred to [21]. The procedures of the proposed DPMM-based method are as below.

1. **Truncate the 19 signals above.** A truncated feature vector, performed by choosing the lowest traffic cycle number among all sessions, is denoted as $\bar{Z}^i$ with 80 cycles in our case;
2. **Input $\bar{\mathbf{Z}}^{\boldsymbol{i}}$ in a covariance signal descriptor in the data $\boldsymbol{D} \in \boldsymbol{W_p^+}$,** where W_p^+ is a $p \times p$ symmetric positive definite (SPD) matrix and the mean feature vector is $u_{\bar{Z}^i}$.
3. **Reduce dimension by PCA.** The PCA coefficients are expressed as $\{y_i^r\}_{r=1,\cdots,d} = \bar{\gamma}_i$ where $y_i^1 > y_i^2 > \cdots > y_i^d$, the dimension $d = 2$ in our case due to its sufficiency to represent the original vector already in the experiments. The proposed DPMM method groups the signal that are represented low-dimensionally as, $S = \{\bar{\gamma}_1, \bar{\gamma}_2, \cdots, \bar{\gamma}_M\}$, for M sessions;
4. **Construct DPMM for the dimension reduced signals.** Some mathematical expressions employed in the modeling are described here: $\theta_b = \{\vec{\mu}_b, \sum_b\}$ are the mean and covariance for class b, $\vec{\pi} = \{\pi_b\}_{b=1}^K$, $\pi_b = P(c_i = b)$ are mixture weights. κ_0 means the number of pseudo-observations and c_i means the class of a signal f_k^i at the k^{th} session. Λ_0^{-1} and v_0 represent the mixture density

covariance. $\Lambda_0^{-1} = D$ is set from the covariance signal descriptor. S is the observed data and $C = \{c_i\}_{i=1}^{M}$ denotes the class c_i of the M sessions in which the signal f_k^i belongs to, $\bar{\gamma}_i$ is the PCA coefficients and $\mathcal{H} = \{\Lambda_0^{-1}, \vec{\mu}_b, \kappa_0, v_0\}$ is the hyper-parameters of the Normal Inverse-Wishart prior, to be studied in Section IV.

5. **Perform outlier detection by collapsed Gibbs sampler using Chinese restaurant process (CRP).** B_+ is obtained as the non-empty-group number in the clustering, $C_{-i} = \{c_1, \cdots, c_{i-1}, c_{i+1}, \cdots, c_M\}$.

A proof of the updates of a class label is offered below.

Proposition 2. The updates of a class label are computed as

$$P(c_i = b|C_{-i}, \alpha_0; \mathcal{H}) \propto P\left(\gamma_i|S^{(j)}\backslash\gamma_i; \mathcal{H}\right)P(c_i = j|C_{-i}, \alpha_0), \tag{2}$$

where $S^{(j)}\backslash\gamma_i$ denotes to ignore γ_i from the set $S^{(j)}$. As $P\left(\gamma_i|S^{(j)}\backslash\gamma_i; \mathcal{H}\right)$ is a multivariate Student-t distribution which can be expressed as

$$P\left(\gamma_i|S^{(j)}\backslash\gamma_i; \mathcal{H}\right) \sim t_{v_0-d+1}\left(\vec{\mu}_p, \frac{\Lambda_n^{-1}(\kappa_p+1)}{\kappa_p(v_p-d+1)}\right), \tag{3}$$

where $\vec{\mu}_p = \frac{\kappa_0}{\kappa_0+M}\vec{\mu}_0 + \frac{M}{\kappa_0+M}\vec{\gamma}$, $\kappa_p = \kappa_0 + M$, $v_p = v_0 + M$, $\Lambda_p = \Lambda_0 + Q + \frac{\kappa_0 p}{\kappa_0+M}(\vec{\gamma} - \vec{\mu}_0)(\vec{\gamma} - \vec{\mu}_0)^T$, d is the retained dimension in PCA coefficients γ_i, Q is the updates in the statistics and $v_p - d + 1$ is the number of the degree of freedom in the distribution. Both $v_0 - d + 1$ and $v_p - d + 1$ are 22 in our case. If κ_0 is chosen sufficiently small, then

$$\Lambda_n \approx \Lambda_0 + Q. \tag{4}$$

Proof. As the numerator is sufficiently small, i.e. $\kappa_0 \approx 0$, then $\frac{\kappa_0 p}{\kappa_0+M} \approx 0$ and

$$\Lambda_n = \Lambda_0 + Q + \frac{\kappa_0 p}{\kappa_0 + M}(\vec{\gamma} - \vec{\mu}_0)(\vec{\gamma} - \vec{\mu}_0)^T \approx \Lambda_0 + Q. \quad (5)$$

As a result, the process of the class label updates is stabilized. This will be further proven in Section 4.

To determine whether $z_k^i,\ i = \{1, \cdots, 19\}$ at the k^{th} session is outside or inside a normal group after using collapsed Gibbs sampler, two assumptions are proposed:

(1) $B_+ > 1$ as at least one clustering group exits; and

(2) The majority with the maximum number of elements is the normal group.

Deduced from the results of CRP and Proposition 2, the traffic signal can be classified in the following definition.

Definition 3. Provided that σ_ε signal classes after the collapsed Gibbs sampling, σ_ε comprises some class labels c_i. The normal signal group is decided by $arg\,max\{|\sigma_\varepsilon|\}$, for the data-point number of signals in class ε, $|\sigma_\varepsilon|$. In clustering of data points, the majority group is classified as inliers while all the remaining groups are regarded as outliers (abnormal data points). Samples of clustering results will be shown in Section 4 later.

3 Data and Measurement Metrics

For evaluation, a corpus is produced for the traffic video data in details: vehicle types, traffic volume, jams and incidents were carefully identified in all sessions. The AM session and PM session are alternatively labeled in an ascending order, i.e. The AM part: Sessions 1, 3,…, 61; The PM part: Sessions 2, 4,…, 62. Totally, there are 764,027 vehicles in all 62 sessions, for which the AM and PM sessions have

312,333 and 451,694 vehicles, respectively. It composed of 3 groups: (A) 46 sessions (23 AM and 23 PM sessions) from Monday to Friday (shorthand as Mon-Fri), (B) 8 sessions from Saturday, and (C) 8 sessions from Sunday. In the following evaluation, major group (A) is chosen. Table 1 shows abnormal sessions and their respective anomalies of group (A) in details. Sample Exit N signals of the 23 PM sessions are shown in Fig. 3, in which Sessions 8, 28, and 50 are identified as having anomalies. Anomalies in 5 categories are as follows: Category 1: Failure of hardware; Category 2: Repeated jams in an Entry/Exit; Category 3: Vehicles obstructing an Entry/Exit; Category 4: Small volume in an Entry/Exit; Category 5: Jams in an Exit/Entry causing to small volume in other Entry/Exit.

The metrics below are employed to measure the accuracy of OD: detection success rate (DSR), negative

Table 1 List of Sessions with Their Respective Anomalies in Group (A)

	Session	Signals	Category	Anomalies
AM	15	Entry S, N,	1	Failure of hardware
		Exit W	1	Failure of hardware
		EDD Nr	1	Failure of hardware
	19	Entry W	2	Repeated jams in Exit E
		EDD Ws	2	Repeated jams in Exit E
PM	4	Entry W	3	Vehicles obstructing Entry W
		EDD Ws	3	Vehicles obstructing Entry W
	8	Entry S,	4	Small volume in Entry S
		Exit S,W,N	3	Vehicles obstructing Exit S
		EDD Ss	4	Small volume in Entry S
	28	Entry S,	3	Vehicles obstructing Entry S
		Exit E, N	5	Jams in Entry S causing small volume in Exit E, N
		EDD Sl, Sr, Ss	3	Vehicles obstructing Entry S
	30	Entry S	3	Vehicles obstructing Entry S
		EDD Ss	3	Vehicles obstructing Entry S
	36	Entry S	5	Jams in Exit W causing small volume in Entry S
	50	Entry S,	5	Jams in Exit W causing small volume in Entry S
		Exit N	5	Jams in Exit W causing small volume in Exit N
		EDD Ss	5	Jams in Exit W causing small volume in Entry S

Remark: EDD = Entry Direction Distribution

predictive value (NPV), true positive (TP), false positive (FP), true negative (TN) and false negative (FN),.

The DSR and NPV are defined as

$$DSR = (TP + TN)/(TP + FN + TN + FP), \quad (5)$$

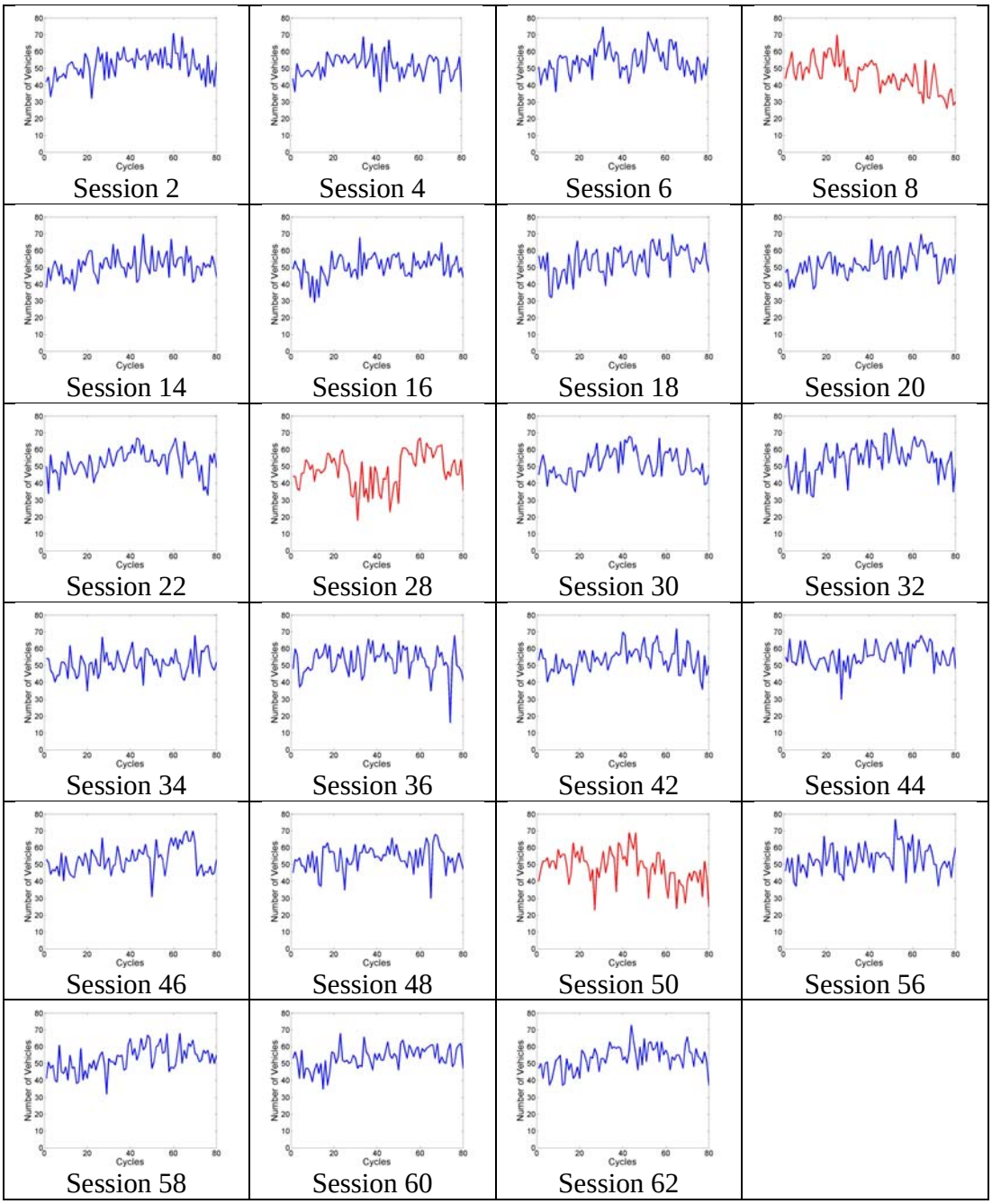


Figure 3 23 Signals of Exit N in the PM sessions. Session 8, Session 28 and Session 50 have anomalies in red). x-axis and y-axis are cycles and vehicle number, respectively

$$NPV = TN/(TN + FN) \quad (6)$$

Positive predictive value (PPV), $PPV = TP/(TP + FP)$, is not included because abnormal signals do not appeared in all sessions, as denominator of PPV goes to zero if no abnormal signal is present. As our literature survey review, there is no benchmark exists for outlier detection in traffic data. Therefore, the subsequent evaluation is focused on the performance and reliability of the proposed DPMM method.

4 Results

4.1 General Results

Initially, the mixture weights, as parameters of the DP prior, of the inliers (normal group) and outliers (abnormal group) are selected arbitrarily as $\pi_1 = 0.9$, $\pi_2 = 0.1$, respectively. This is because that normal traffic is assumed to be much more likely than abnormal traffic. The others are $\Lambda_0^{-1} = D$, $\vec{\mu}_b = [0\ 0]$, $\kappa_0 = 0.0001$, $v_0 = 3$. Using the collapsed Gibbs sampler, the iteration number is set to 300. In fact, any number over 100 is stable empirically for Gibbs sampling. This is depicted in Fig. 4, where the effect of the number of iterations versus $\log P$ (the model's log probability in the training data) is clearly shown. In PCA, the maintained dimension d is set to 2. The average DSR of 23 AM sessions (Mon-Fri), shown in Table 2, is 98.63% (99.08% in NPV) and the average DSR of 23 PM sessions, shown in Table 3, is 96.34% (97.80% in NPV). The overall average DSR of the AM sessions and PM sessions altogether is 97.49% (98.44% in NPV).

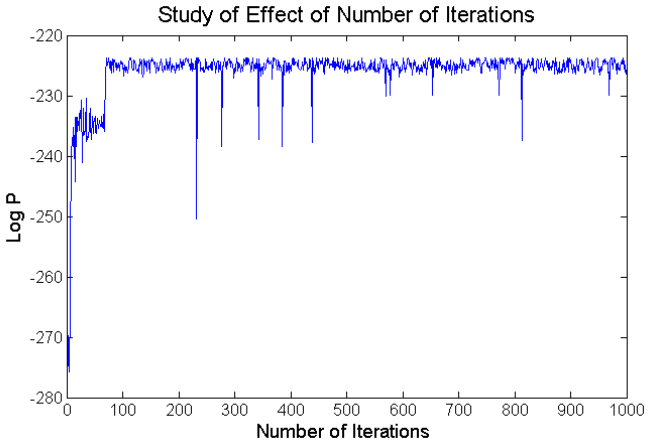


Figure 4 Number of iterations versus $\log P$ in collapsed Gibbs sampling. (Entry W signal from the AM sessions and setting of parameters: $\pi_1 = 0.9$, $\pi_2 = 0.1$, $\vec{\mu}_b = [0\ 0]$, $\kappa_0 = 0.0001$.)

Table 2 Outlier DSR of 23 AM Sessions (Mon-Fri)

		TP	FP	TN	FN	DSR	DSR in %
Entry	E	0	0	23	0	23/23	100%
	S	0	0	22	1	22/23	95.65%
	W	1	0	22	0	23/23	100%
	N	0	0	22	1	22/23	95.65%
Exit	E	0	0	23	0	23/23	100%
	S	0	0	23	0	23/23	100%
	W	0	0	22	1	22/23	95.65%
	N	0	0	23	0	23/23	100%
E-D-D	E_l	**0**	0	23	0	23/23	100%
	E_r	0	0	23	0	23/23	100%
	E_s	0	0	23	0	23/23	100%
	S_l	0	1	22	0	22/23	95.65%
	S_r	0	1	22	0	22/23	95.65%
	S_s	0	0	23	0	23/23	100%
	W_l	0	0	23	0	23/23	100%
	W_s	1	0	22	0	23/23	100%
	N_l	0	0	23	0	23/23	100%
	N_r	0	0	22	1	22/23	95.65%
	N_s	0	0	23	0	23/23	100%
Average							98.63%

Remark: Detection success rate (DSR), True positive (TP), False positive (FP), True negative (TN), False negative (FN), Entry Direction Distribution (EDD)

Table 3 Outlier DSR of 23 PM Sessions (Mon-Fri)

		TP	FP	TN	FN	DSR	DSR in %
Entry	E	0	0	23	0	23/23	100%
	S	3	0	18	2	21/23	91.3%
	W	0	2	20	1	20/23	87%
	N	0	0	23	0	23/23	100%
Exit	E	0	0	22	1	22/23	95.65%
	S	0	0	22	1	22/23	95.65%
	W	0	0	22	1	22/23	95.65%
	N	3	0	20	0	23/23	100%
E-D-D	E_l	0	0	23	0	23/23	100%
	E_r	0	2	21	0	21/23	91.3%
	E_s	0	2	23	0	23/23	100%
	S_l	0	0	22	1	22/23	95.65%
	S_r	1	0	22	0	23/23	100%
	S_s	3	0	19	1	22/23	95.65%
	W_l	0	0	23	0	23/23	100%
	W_s	0	1	21	1	21/23	91.3%
	N_l	0	0	23	0	23/23	100%
	N_r	0	2	21	0	21/23	91.3%
	N_s	0	0	23	0	23/23	100%
Average							96.34%

Remark: Detection success rate (DSR), True positive (TP), False positive (FP), True negative (TN), False negative (FN), Entry Direction Distribution (EDD)

Regarding computational complexity, the computer used for the testing was an Intel Core 2 Duo CPU P9700 2.8GHz, with 4GB of DDR3 SDRAM. The algorithm was implemented in MATLAB (Version 7.10). The execution time is evaluated with the same procedure in Section 2. The average computational time is 0.0832 second for 10 trials which is based on the same set of parameters on 23 signals of Entry W from the AM sessions (i.e. $\pi_1 = 0.9$, $\pi_2 = 0.1$, $v_0 = 3$, $k_0 = 0.0001$, number of iterations: 300).

As most outlier DSRs achieve high values in both AM and PM sessions, an analysis of a failure case is given below in order to further study the characteristic of the proposed method. The detection result of

Entry W signals in 23 PM sessions, 87%, is the worst among all signals. Entry W signals from Fig. 5 are the entry volume of vehicles which are different to Exit N signals from Fig. 4. Fig. 6a depicts the clustering result for the Entry W signals of all 23 PM sessions. The majority group, classified as inliers, is represented by blue circles whereas the minority group, classified as outliers, is represented by the red asterisk and square. Fig. 6b depicts the post-processing result that the points are further labeled by different shapes and colors and interpreted as true negative (TN), false positive (FP) and false negative (FN) cases with respect to their correctness in the OD. There is no true positive (TP) case (correctly detected as an outlier) in the OD for the Entry W signals. In Fig. 6b, blue circles represent the TN cases, purple stars represent the FP cases and red cross represents the FN case. Herein, the single red cross denotes Session 4 and is regarded as the FN case (misclassified as an inlier). The two green circles denote Sessions 28 and 60 are the false positive (FP) cases (misclassified as the outliers). The clustering result is compared with the original signals in Fig. 5. The signals from Sessions 28 and 60 appear to be normal in Fig. 5 at first glance. However, the data points of these 2 sessions in Fig. 6a are separated from the majority group. It is believed to be a reason why the collapsed Gibbs sampler classified them as abnormal. On the other hand, there is a deep valley of a sharp decrease of traffic volume near the middle part of the Session 4 signal, but the result shows that it is clustered into the majority group. Upon closer inspection of Fig. 5, the Session 4 signal is in fact somewhat similar to the signals of Sessions 14 and 18. This dilutes the abnormal appearance of the Session 4 signal and causes it to be classified incorrectly. On the other hand, as the signals from Sessions

28 and 60 are very similar to the signals from other normal sessions, it shows that the collapsed Gibbs sampler has its weakness sometimes and it certainly is an area for improvement of OD in the future.

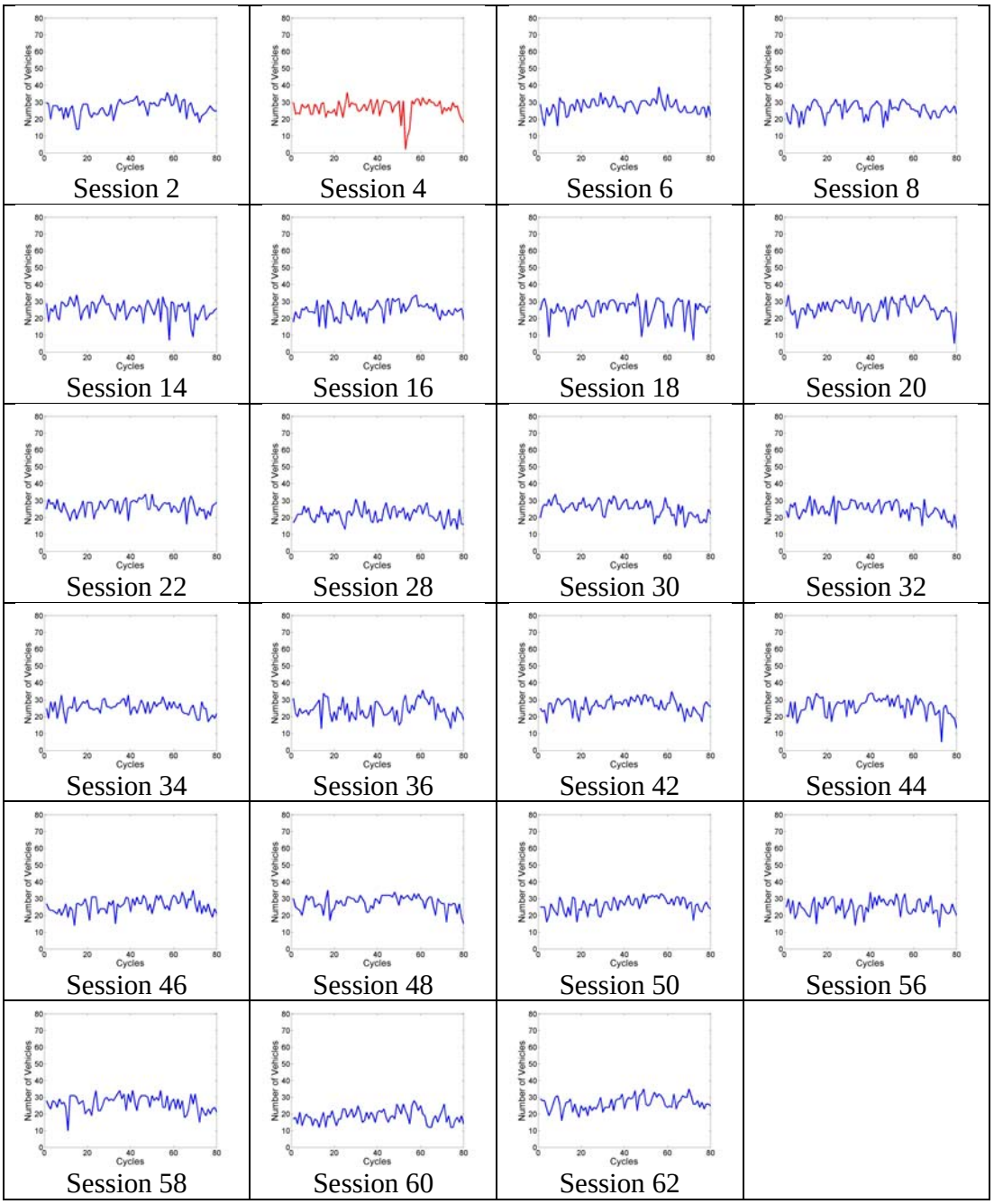
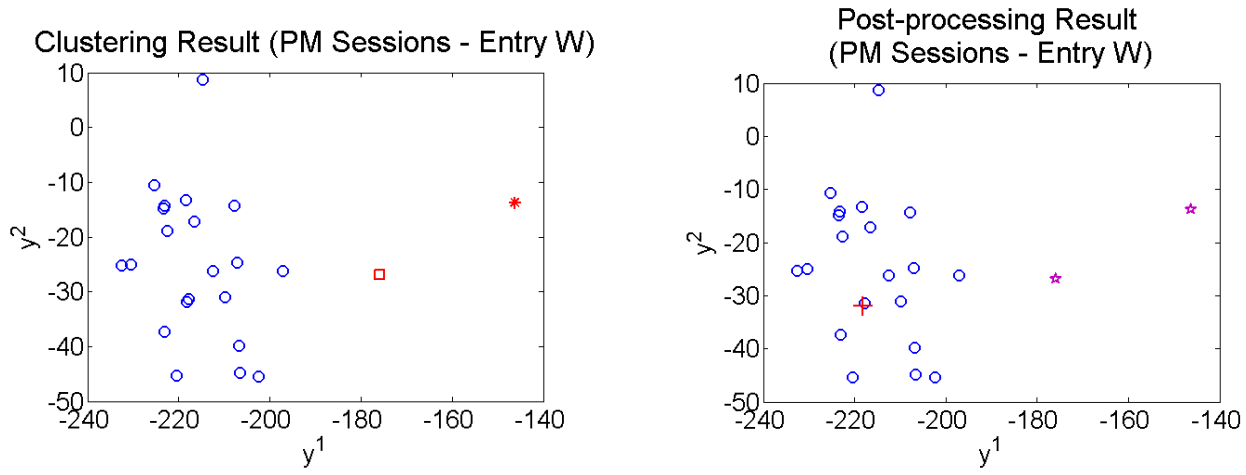


Figure 5 23 PM sessions of Entry W Signal. Session 4 have anomalies (in red). x-axis and y-axis are cycles and vehicle number respectively



(a) (b)

Figure 6 Result of Clustering in Entry W signal from the PM sessions. (a) Clustering result: Blue circles, indicating the majority group, are inliers, while the red asterisk and square, indicating the minority groups, are outliers. (b) Post-processing result: Blue circles represent true negative (TN) cases, purple stars represent false positive (FP) cases and red cross represents the false negative (FN) case. The x-axis is coefficients y^1 and the y-axis is coefficient y^2 from PCA, where $d = 2$

4.2 10-fold cross validation

To ascertain the robustness of the proposed method, a 10-fold cross validation approach was used, producing altogether 380 tests of 19 signals in both AM and PM sessions. The parameter setting is the same as the preceding evaluation (i.e. $\pi_1 = 0.9$, $\pi_2 = 0.1$, $v_0 = 3$, $k_0 = 0.0001$, number of iterations: 300). The average DSRs of 23 AM and 23 PM sessions, as shown in Table 4, are 97.59% and 95.74%, respectively. The overall average of all sessions after 10-fold cross validation is 96.67% (98.32% in NPV). In general, every Entry, Exit and EDD signal achieves over 90% success rate. Moreover, compared to the DSR of 97.49% in the preceding evaluation, the result of 10-fold cross validation shows that the proposed DPMM-based method is reliable and robust.

Table 4 Outlier DSR of 10-fold Validation of 23 AM & 23 PM Sessions (Mon-Fri).

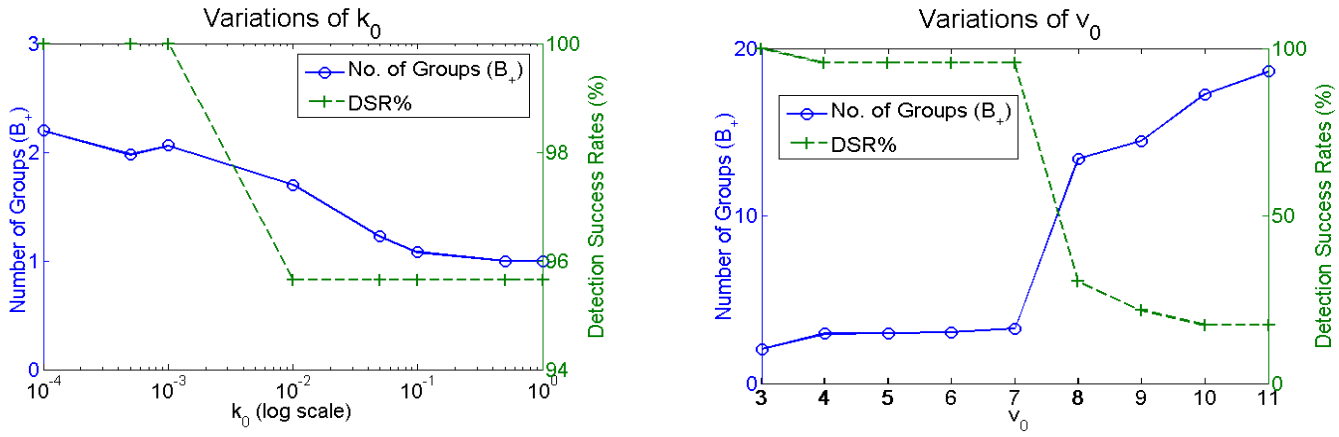
	AM	DSR in %	PM	DSR in %
Entry	E	100%	E	100%
	S	93.74%	S	91.29%
	W	99.05%	W	91.79%
	N	95.64%	N	99.52%
Exit	E	98.10%	E	96.12%
	S	100%	S	95.64%
	W	95.64%	W	92.74%
	N	98.05%	N	92.24%
E-D-D	E_l	98.05%	E_l	100%
	E_r	99.52%	E_r	92.76%
	E_s	96.67%	E_s	100%
	S_l	93.72%	S_l	95.17%
	S_r	99.5%	S_r	97.12%
	S_s	96.55%	S_s	95.64%
	W_l	98.55%	W_l	97.14%
	W_s	97.12%	W_s	91.29%
	N_l	99.05%	N_l	100%
	N_r	95.17%	N_r	98.05%
	N_s	100%	N_s	92.62%
Average		97.59%		95.74%
Final Average				96.67%

Remark: Detection success rate (DSR), Entry Direction Distribution (EDD)

4.3 Effect of different parameter setting

How the hyper-parameters $\mathcal{H}$ of the conjugate priors affect the OD is also investigated. The value of Λ_0^{-1} is found not to affect the result much because $\Lambda_0^{-1} = D$ relies on the dimension of d to be maintained in PCA and that the same detection results are acquired when d is set to larger than 2. This is understandable because the first several principal components have already contained significant information of the signal.

κ_0 and v_0 are further studied. First, 23 signals of Entry W from the AM sessions within Monday and Friday are utilized. To study the variations of κ_0, the other parameters are fixed as $\pi_1 = 0.9$, $\pi_2 = 0.1$, $\vec{\mu}_b = [0\ 0]$, $v_0 = 3$. Fig. 7a depicts how κ_0 varies from 0.0001 to 1 and its effect on the number of groups



(a) (b)

Figure 7 (a) Variation of κ_0 (in log scale) and its corresponding results in the number of groups B_+ (blue line with circles) and detection success rate (green dot-line with crosses); (b) Variation of v_0 and its corresponding results in the number of groups B_+ (blue line with circles) and detection success rate (green dot-line with crosses)

B_+ (blue line with circles) and the detection success rates (green dot-line with crosses). The best success rate of OD decreases from 100% to 95.65% when κ_0 is higher than 0.001, in which the number of groups clustered is also steadily reduced from 2.1967 ($\kappa_0 = 0.0001$) to 0.9967 ($\kappa_0 = 1$). From this empirical result, it is concluded that an appropriate κ_0 is smaller than 0.001. This also proves Proposition 2 in Section 2.

For v_0, the other parameters $\pi_1 = 0.9, \pi_2 = 0.1, \vec{\mu}_b = [0\ 0],\ \kappa_0 = 0.0001$ are fixed. In Fig. 7b, the best success rate at 100% is found when $v_0 = 3$, then the success rate (green dot-line with crosses) starts to decrease to 30.43% when $v_0 = 11$. At the same time, the number of groups B_+ (blue line with circles) increases from 2.0533 to 18.62 versus v_0 from 3 to 11. It implies that v_0 should be set as small as possible. v_0 cannot be set to be smaller than 3 as the value of v_0 is determined by the dimension, d, in the PCA plus 1, in which $d = 2$ is the basic dimension of Λ_0^{-1} and $\vec{\mu}_0$.

5 Conclusion

From of the performance evaluation of the proposed method, it can be concluded that it achieves a high 96.67% detection accuracy from a 10-fold cross validation on a set of real-world traffic data. The proposed DPMM-based OD method is generic by nature and the result so far is positive. It is found that the proposed method performs outlier detection quite well in 19 real traffic signals generated from one of the busiest junctions in Hong Kong, especially from the study of the effect of different parameter settings. DPMM is a promising choice to model large-scale traffic data, which is challenging in itself because of the serious level of spatial-temporal similarities in such data. The DPMM-based OD method can be further developed as follows: A detailed evaluation of OD methods including supervised and unsupervised methods will be carried out. The proposed DPMM method may be extended to multiple junctions in scale, online detection and outlier classification in algorithm.

6 Acknowledgment

Two grants support this research: the Hong Kong Special Administrative Region's Research Grant Council, China: Project HKU754109, and the Hong Kong Baptist University FRG: FRG1/12-13/075. A preliminary study of this research appears in [21].

6 References

[1] Barria, J.A., Thajchayapong, S.: "Detection and Classification of Traffic Anomalies Using Microscopic Traffic Variables", *IEEE Trans. Intelligent Transportation Systems*, 2011. 12(3), pp. 695-704

[2] Hu, J., Wang, Y., Zhang, Z., Li, D.: "Analysis on Traffic Flow Data and Extraction of Nonlinear Characteristic Quantities", *IEEE 13th Int'l Conf. ITS,* 2010, pp. 712-717

[3] Zhang, G., Zhou, Z., Zhou, H.: "The High Frequency Traffic Flow Analysis", *IEEE 2nd Int'l Sym. Computational Intelligence and Design*, 2009, pp. 221-224

[4] Morris, B., Trivedi, M.: "Real-time Video Based Highway Traffic Measurement and Performance Monitoring", *Proc. IEEE Conf. ITS*, 2007, pp. 59-64

[5] Weng, X-X., Tan, Y., Du, G., Hong, Q.: "Prediction and Identification of Urban Traffic Flow Based on Features", *IEEE 9th Int'l Conf. Control, Automation, Robotics and Vision*, 2006, pp. 1-6

[6] Thomas T., van Berkum E.C.: "Detection of incidents and events in urban networks", *IET Intell.Transp. Syst.*, 2009, 3(2), pp. 198-205

[7] Tang, S., Gao, H.: "Traffic-incident Detection-algorithm based on Non-parametric Regression. *IEEE. Trans. ITS*, 2005, 6(1), pp. 38-42

[8] Chen, S-C., Shyu, M-L., Peeta, S., Zhang, C.: "Learning-based Spatio-temporal Vehicle Tracking and Indexing for Transportation Multimedia Database Systems", *IEEE. Trans. ITS*, 2003, 4(3), pp. 154-167

[9] Hodge, V.J.: "A Survey of Outlier Detection Methodologies", *Artificial Intelligence Review*, 2004, 22(2), pp. 85-126

[10] Ngan, H.Y.T., Pang, G.K.H., Yung, N.H.C.: "Ellipsoidal Decision Regions for Motif-based Patterned Fabric Defect Detection" *Pattern Recognition*, 2010, 43(6), pp. 2132-2144

[11] Ngan, H.Y.T., Pang, G.K.H., Yung, N.H.C.: "Performance Evaluation for Motif-based Patterned Texture Defect Detection", *IEEE Trans. Automation Science & Engineering*, 2010, 7(1), pp. 58-72

[12] Cheng, Y., Zhang, Y., Hu, J., Li, L.: "Mining for Similarities in Urban Traffic Flow Using Wavelets", *Proc. IEEE Conf. ITS*, 2007, pp. 119-124

[13] Park, E.S., Turner, S., Spiegelman, C.H.: "Empirical Approaches to Outlier Detection in Intelligent Transportation Systems Data", *Transportation Research Record*, 2003, 03-2990, pp. 21-30

[14] Ihler, A.T., Smyth P,: "Learning Time-Intensity Profiles of Human Activity using Non-parametric Bayesian Models", *Proc. NIPS*, 2006, pp. 625-632

[15] Fox, E.B., Sudderth, E.B., Willsky, A.S.: “Hierarchical Dirichlet Processes for Tracking Maneuvering Targets”, *10th IEEE Int’l Conf. Information Fusion*, 2007, pp. 1-8

[16] Kivinen, J., Sudderth, E.B., Jordan, M.I.: “Learning Multiscale Representations of Natural Scenes Using Dirichlet Processes”, *IEEE 11th ICCV*, 2007, pp. 1-8

[17] Li, X., Hu, W., Zhang, Z., Zhang, X., Luo, G.: “Trajectory-based Video Retrieval Using Dirichlet Process Mixture Models”, *Proc. BMVC*, 2008

[18] Hu, D.H., Zhang, X-X., Yin, J., Zheng, V.W., Yang, Q.: “Abnormal Activity Recognition Based on HDP-HMM Models”, *Proc. 21st Int’l Joint Conf. Artificial Intelligence*, 2009, pp. 1715-1720

[19] Zhang, X-X., Liu, H., Gao, Y., Hu, D. H., “Detecting Abnormal Events via Hierarchical Dirichlet Processes”, *PAKDD 2009, LNAI 5476*, 2009, pp. 278-289

[20] Jian, Y-D., Chen, C-S.: “Two-View Motion Segmentation with Model Selection and Outlier Removal by RANSAC-Enhanced Dirichlet Process Mixture Models”, *IJCV*, 2010, 88, pp. 489-501

[21] Ngan, H.Y.T., Yung, N.H.C., Yeh, A.G.O.: “Modeling of Traffic Data Characteristics by Dirichlet Process Mixtures”, *8th IEEE Int’l CASE*, 2012, pp. 224-229