Estimation of mean and covariance of stochastic multi-class OD demands from classified traffic counts

Hu Shao\textsuperscript{a,b}, William H. K. Lam\textsuperscript{a,c,}\textsuperscript{*}, Agachai Sumalee\textsuperscript{a,d}, Martin L. Hazelton\textsuperscript{e}

\textsuperscript{a}Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
\textsuperscript{b}Department of Mathematics, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, China
\textsuperscript{c}School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China
\textsuperscript{d}Department of Civil Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand.
\textsuperscript{e}Institute of Fundamental Sciences, Massey University, Palmerston North 4442, New Zealand.

Abstract

This paper proposes a new model to estimate the mean and covariance of stochastic multi-class (multiple vehicle classes) origin-destination (OD) demands from hourly classified traffic counts throughout the whole year. It is usually assumed in the conventional OD demand estimation models that the OD demand by vehicle class is deterministic. Little attention is given on the estimation of the statistical properties of stochastic OD demands as well as their covariance between different vehicle classes. Also, the interactions between different vehicle classes in OD demand are ignored such as the change of modes between private car and taxi during a particular hourly period over the year. To fill these two gaps, the mean and covariance matrix of stochastic multi-class OD demands for the same hourly period over the year are simultaneously estimated by a modified lasso (least absolute shrinkage and selection operator) method. The estimated covariance matrix of stochastic multi-class OD demands can be used to capture the statistical dependency of traffic demands between different vehicle classes. In this paper, the proposed model is formulated as a non-linear constrained optimization problem. An exterior penalty algorithm is adapted to solve the proposed model. Numerical examples are presented to illustrate the applications of the proposed model together with some insightful findings on the importance of covariance of OD demand between different vehicle classes.

© 2015 The Authors. Published by Elsevier B.V.

Keywords: OD demand estimation; multiple vehicle class; covariance matrix; classified traffic counts.

* Corresponding author. Tel.: +852-27666045; fax: +852- 23659291.
E-mail address: william.lam@polyu.edu.hk

2214-241X © 2015 The Authors. Published by Elsevier B.V.
1. Introduction

Origin-destination (OD) traffic demand is one of the fundamental input data for transportation planning and traffic management. In the past decades, OD demand estimation from traffic counts has been an important topic in the field of transportation research so as to minimize the cost for data collection. However, most of the existing OD demand estimation models ignore two important features of the OD demands as follows.

- The interactions between different vehicle classes (or types) in OD demand, such as taxis, private cars and goods vehicles.
- The statistical characteristics of multi-class OD demands, such as the covariance of traffic demands between different vehicle classes.

This paper proposes a new model for estimation of the mean and covariance of stochastic multi-class (i.e. multiple vehicle classes) OD demands from hourly classified traffic counts throughout the whole year.

1.1. Covariance of OD demands

Due to daily and seasonal variations in activity patterns, the OD traffic demands of different vehicle classes during the same hourly period (e.g. morning peak, 8:00 am - 9:00 am) are stochastically varied from day to day over the whole year. This type of varying traffic demands is referred to as stochastic multi-class OD demands in this paper. Statistically, the random characteristics of the stochastic multi-class OD demands can be reflected by their mean and covariance. In the conventional OD demand estimation models, focus is usually put on the mean of the OD demands while the covariances of OD demands by vehicle class have not been considered. The covariance of stochastic multi-class OD demands would however reflect the correlations between OD demands by vehicle class. For instance, for the traffic demand of the same OD pair, the higher the private car usage, the less the taxis usage.

It should be pointed out that there are generally three categories of OD demand covariances, i.e. the spatial, temporal and vehicle class covariances. Firstly, spatial OD demand covariance refers to the correlation (or dependency to some extent) of the OD demands during the same hourly periods between different OD pairs in a spatial manner (Shao et al., 2014). Secondly, temporal OD demand covariance represents the correlation of the OD demands for the same OD pair between different time periods (e.g. 8:00 am - 9:00 am and 9:00 am - 10:00 am). Thirdly, the vehicle class OD demand covariance relates to the statistical dependency of traffic demand between different vehicle classes of the same OD pair. In order to facilitate the essential ideas on correlations between OD demands by vehicle class, this paper aims to estimate the third category of OD demand covariance using classified traffic counts for the same hourly period over the year.

In road transportation networks, the covariance of OD demands should not be ignored particularly for OD demand estimation from traffic counts. The ignorance of the correlation between random variables may lead to very different output of the models (Haas, 1999). For example, Waller et al. (2001) found that the correlation level of OD demands plays a major role in determining the degree of error in relation to the expected total network travel time. Zhao and Kockelman (2002) discussed the propagation of errors through the four-step traffic demand forecasting model. They stated that neglecting the correlation of data (e.g., OD demands) would ultimately reduce the reliability of the traffic forecasts, and in turn affect the policy-making and infrastructure decisions. Duthie et al. (2011) found that the assumption of independent demands when correlations do in fact exist could lead to errors in the estimation of system performance and result in poor policy decisions; for instance, building highways may not be able to meet a higher-than-expected demand. Shao et al. (2014) found that the spatial covariance has significant impact of network performance evaluation. In view of the above studies, it is shown that the correlation between OD demands should not be overlooked. This has a great effect on OD demand estimation problem particularly with the use of traffic counts.

The mean and covariance of the stochastic multi-class OD demands to be estimated in this paper can be used in the reliability-based traffic assignment models which have recently been developed for multi-modal transportation networks with uncertainty (Chen et al., 2002; Nakayama and Takayama, 2003; Clark and Watling, 2005; Shao et al., 2006; Lam et al., 2008; Chen and Zhou, 2010; Chen et al., 2011; Sumalee et al., 2011). These relevant studies...
demonstrated that increasing attention has been given on development of reliability-based network equilibrium models but it was assumed that the probability distributions of stochastic multi-class OD demands are known and given. However, to the best of our knowledge, less attention has been paid to estimation of the probability distribution of the stochastic multi-class OD demands, which is a necessary input for the application of the above models in networks with uncertainty. It is known that the mean and covariance matrix are the two key parameters to characterize the OD demand probability distribution. Therefore, the estimation of mean and covariance of the stochastic multi-class OD demands is an important extension of the current research work on reliability-based network equilibrium models.

1.2. Interactions between multiple vehicle classes in OD demand

Conventionally, the OD demands of different vehicle classes can be estimated by the combined model, which combines two steps of the conventional four-step trip-based model, i.e. mode and path (or route) choices. The combined model can be regarded as a “top-down” approach to reflect the interactions between multiple vehicle classes as shown in Figure 1. In the combined model, travel modes are classified by different vehicle classes, such as cars, taxis, buses, heavy trucks, light trucks, etc. The sum of OD demands for all modes is available, and the choice of each mode is a function of travel times between an OD pair using alternative travel modes. At equilibrium, no user can change his route to lessen his travel cost, but may do so by changing his mode (Lam and Huang, 1992a). The corresponding combined model is a process of integration of the modal split and traffic assignment, which can be used to estimate the OD demand of each travel mode (vehicle class). However, the accuracy of the combined model results depends on the value of model parameters which need to be calibrated by expensive household interview surveys (Frank, 1978; Lam and Huang, 1992b).

In comparison with the combined model, this paper proposes a new model which is a “bottom-up” approach to estimate the mean and covariance of stochastic multi-class OD demands so as to account for the interactions between multiple vehicle classes in OD demand. In practice, the observed link traffic counts (flows) can be classified by vehicle class with the use of advanced technologies. For example, the automatic vehicle identification (AVI) reader, which is being used in Hong Kong for electronic toll collection and journey time estimation purposes (Tam and Lam, 2008), can recognize the vehicle class. The observed classified link traffic counts during the same hourly period (e.g. morning peak, 8:00 am - 9:00 am) are stochastically fluctuated from day to day over the whole year. Thus, classified link traffic counts are considered as random variables. The sample mean and covariance of traffic counts for all vehicle classes can then be calculated using the observed classified link traffic counts. It should
be noted that the sample covariance of classified link traffic counts would capture the statistical dependence between different vehicle classes. It is a result of the interactions between multiple vehicle classes in traffic demand as well as the effect of the route or path choices. Such information and/or relationships are adopted in this paper to estimate the covariance of stochastic multi-class OD demands.

Apart from the covariance between different vehicle classes in OD demands, the mean OD demand for each vehicle class is also a kind of important information for transportation planning and traffic management. The mean OD demand by vehicle class can be used to assess the link choice proportion by vehicle class (i.e. the proportion of each vehicle class on links of the transportation network). Such information is useful for road traffic operation and control, traffic accident management and transportation policy evaluation. Although traffic flows of different vehicle classes can be uniformly converted into the passenger car units (pcu) for evaluation purpose, vehicle class is obviously an important factor that influences the operating capability and traffic accident severity in road transportation networks. For example, heavy vehicles have lower operating capabilities than private cars, particularly with respect to acceleration, deceleration, and the ability to maintain speed on hilly roads (Transportation Research Board, 2000). Involvement of motor cycles in multiple-vehicle accidents has a higher proportion of fatal or serious injuries than that of private cars and taxis (Yau, 2006). Therefore, estimation of mean value for different vehicle classes in OD demands can help the traffic manager effectively operate and evaluate the road transportation system particularly in congested network with uncertainty.

1.3. OD demand estimation models

Although various methods for OD demand estimation from traffic counts have been widely investigated in the past decades, the existing estimation models can not fully capture the statistical characteristics as well as the interactions between different vehicle classes of OD demands as shown in Table 1. These OD demand estimation models include entropy maximizing model (Van Zuylen and Willumsen, 1980), maximum likelihood model (Spiess, 1987; Watling, 1994), generalized least squares (GLS) model (Cascetta, 1984; Bell, 1991), stochastic mapping method (Ashok and Ben-Akiva, 2000 and 2002), Bayesian inference estimation model (Maher, 1983; Sun et al., 2006; Castillo et al., 2008a,b) and Markov chain model (Li, 2009). It is assumed in these estimation models that the OD demands are either deterministic variables or mutually independent random variables. The first outstanding feature of the existing OD demand estimation models is that some in-depth statistical properties were generally ignored. For example, few of them addressed the OD demand variations, such as the correlation between random traffic demands of different vehicle classes (covariance). In general, only the mean of the OD demand was estimated in most of the existing models. An exception was the research work by Shao et al. (2014), which considered the estimation of spatial covariance of OD demands. However, this work may suffer from the difficulty of the “overfitting” problem as there are too many parameters to be estimated in the covariance matrix of stochastic OD demands. Also, the path choice proportion and/or the link choice proportion is assumed to be a deterministic variable. Such an assumption can not account for the stochastic path choice behaviours under condition with uncertainty. The second outstanding feature of the existing models is that most of them only considers the OD demand estimation problem for single vehicle class except the work by Wong et al. (2005). In the latter, only the mean of multi-class OD demands were estimated. In view of the above two outstanding features of the existing OD demand estimation models, this paper proposes a new model to estimate simultaneously both the mean and covariance of multi-class OD demands while explicitly considering the stochastic mode and path choice behaviours.

<table>
<thead>
<tr>
<th>Statistical characteristics</th>
<th>Vehicle classes</th>
<th>Mean</th>
<th>Covariance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single</td>
<td>Most of the existing OD demand estimation models with the use of traffic counts</td>
<td>Shao et al. (2014)</td>
</tr>
<tr>
<td></td>
<td>Multiple</td>
<td>Wong et al. (2005)</td>
<td>This paper</td>
</tr>
</tbody>
</table>
1.4. Contribution statement

It is well-known that a common difficulty of the existing OD demand estimation models is the identifiability problem, i.e. it is impossible to identify the unique OD demand matrix from the observed traffic counts as the number of observed links (from which the traffic counts are available) is usually less than the number of parameters (or unknowns) to be estimated in the OD demand matrix (Hazarlton, 2003). For the problem of estimating the mean and covariance matrix of the stochastic multi-class OD demands, the identifiability problem may be exacerbated as more parameters (say for instance, the covariance matrix) are required to be estimated. In view of this, the shrinkage estimator, a technique that is useful for estimating large-dimensional parameters with comparatively fewer observations, has the potential to overcome this difficulty in practice.

The widely-used shrinkage estimator, lasso (least absolute shrinkage and selection operator) method (Tibshirani, 1996), is adopted to address the identifiability problem in this paper. There are a large number of elements (or parameters) in the covariance matrix of stochastic multi-class OD demands. For example, in a simple network with two OD pairs and three vehicle classes, there are 21 parameters needed to be estimated in the $6 \times 6$ covariance matrix of multi-class OD demands. It should be noted that the covariance matrix is symmetrical. Thus, for a $n \times n$ covariance matrix, there are actually $(n+1) \times n/2$ elements needed to be estimated. However, in reality, some elements in the multi-class OD demand covariance matrix may be zero. For instance, the traffic demands between private cars and goods vehicles of the same OD pair may be independent with each other. And the corresponding elements in the covariance matrix should be zero. Due to the identifiability difficulty, the reasonable solutions of these covariances (i.e. zero) may not be identified as multiple solutions may exist. To overcome this difficulty, the lasso method does variable selection and shrinkage by solving the L1-penalized least squares (or linear regression) (Tibshirani, 1996 and 2011). The lasso method is equivalent to minimization of the sum of squares with a constraint of L1-norm of estimated parameters. For the covariance between independent multi-class OD demands (e.g. private cars and goods vehicles), the adoption of the lasso method can make it shrunk towards zero. As such, it is expected that covariance matrix of the OD demands could be uniquely identified.

This paper proposes a new model for estimating the mean and covariance of stochastic OD demands from classified traffic counts for the same hourly period over the year. The proposed model extends the existing works with the following new features.

- Not only the mean but also covariance matrix of stochastic multi-class OD demands is estimated. Particularly, the estimated covariance of stochastic multi-class OD demands can statistically reflect the traffic demand interactions between vehicle classes under network uncertainty over the year. Meanwhile, the interactions of stochastic path and mode choices are explicitly considered in the proposed model, which may has great potential to help the transportation planner and/or traffic manager understand the complexity and randomness of the mode and path choice behaviors in congested network with uncertainty.
- Lasso method is firstly incorporated into the OD demand estimation problem for variable selection and shrinkage so as to address the exacerbated identifiability issue for estimation of stochastic multi-class OD demand covariance matrix.

An equivalent non-linear constrained optimization model is proposed and formulated for the multi-class stochastic OD demand estimation problem. A $n$-fold cross-validation procedure is adapted to determine the lasso parameter in the proposed model. A heuristic solution algorithm based on the penalty method is adapted to solve the proposed model. Numerical examples are shown below to demonstrate the applications of the proposed model and solution algorithm together with some insightful discussion on the problem concerned.

1.5. Paper organization

The rest of the paper is organized as follows. In Section 2, the model formulation for estimating the mean and covariance of multi-class OD demands is presented. Then, a heuristic solution algorithm is proposed in Section 3. Numerical examples are discussed in Section 4. Finally, conclusions and further studies are given in Section 5.
2. Model formulation

2.1. Notations and basic assumptions

The notations used throughout the paper are listed as follows unless otherwise specified. For notational consistency, the italic capital letters are used to denote random variables and the italic lower-case letters are used to denote deterministic variables throughout the paper.

<table>
<thead>
<tr>
<th>Nomenclature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Set of links in the network.</td>
</tr>
<tr>
<td>( \tilde{A} )</td>
<td>Observed link set, which is a subset of link set A. The traffic flows on link ( a \in \tilde{A} ) can be observed by the traffic sensor during the observed time period.</td>
</tr>
<tr>
<td>( a, a' )</td>
<td>Link index, ( a, a' \in A ).</td>
</tr>
<tr>
<td>D</td>
<td>The set of vehicle classes, ( i \in D ).</td>
</tr>
<tr>
<td>( F_{r,k,i} )</td>
<td>Random traffic flow of vehicle class ( i \in D ) on path (or route) ( k \in K_{rs} ) between OD ( rs \in R ).</td>
</tr>
<tr>
<td>( \mathbf{f} )</td>
<td>( [\mathbf{K}][\mathbf{D}] )-vector of random multi-class path flows ( (\cdots, F_{r,k,i}^k, \cdots)^T ) for all ( k \in K_{rs} ), ( rs \in R ) and ( i \in D ).</td>
</tr>
<tr>
<td>( f_{r,k,i} )</td>
<td>Mean traffic flow of vehicle class ( i \in D ) on path ( k \in K_{rs} ) between OD ( rs \in R ), ( f_{r,k,i}^k = E[F_{r,k,i}^k] ).</td>
</tr>
<tr>
<td>( \mathbf{G} = (\mathbf{N}, A) )</td>
<td>A road network, with ( \mathbf{N} ) being the set of nodes and ( A ) being the set of links, respectively.</td>
</tr>
<tr>
<td>( \mathbf{H} )</td>
<td>Sample of ( V_{a,i} ), ( H = {\tilde{v}<em>{a,i}^{(1)}, \tilde{v}</em>{a,i}^{(2)}, \cdots, \tilde{v}_{a,i}^{(h)}} ).</td>
</tr>
<tr>
<td>( H_n )</td>
<td>A subset of ( \mathbf{H} ). ( \mathbf{H} ) is randomly divided into 10 subsets with the same sample size ( \frac{1}{10} h ).</td>
</tr>
<tr>
<td>( i, i' )</td>
<td>Vehicle class index, ( i, i' \in D ).</td>
</tr>
<tr>
<td>( K )</td>
<td>Total path set of the network, ( K = \bigcup_{rs \in R} K_{rs} ).</td>
</tr>
<tr>
<td>( K_{rs} )</td>
<td>Path set between OD pair ( rs ).</td>
</tr>
<tr>
<td>( k, k' )</td>
<td>Path index, ( k, k' \in K_{rs} ).</td>
</tr>
<tr>
<td>( m )</td>
<td>A large integer used in cross-validation. In this paper, it is assumed that ( m = 100 ).</td>
</tr>
<tr>
<td>( p_{r,k,i} )</td>
<td>Random mode-path choice proportion of the traffic flow on path ( k \in K_{rs} ) between OD ( rs \in R ).</td>
</tr>
<tr>
<td>( p_{r,k,i}^k )</td>
<td>Mean path choice proportion of the traffic flow on path ( k \in K_{rs} ) of vehicle class ( i ) between OD pair ( rs ).</td>
</tr>
<tr>
<td>( q_{rs,i} )</td>
<td>Random traffic demand of vehicle class ( i ) between OD pair ( rs ).</td>
</tr>
<tr>
<td>( \mathbf{Q} )</td>
<td>( [\mathbf{R}][\mathbf{D}] )-vector of ( (\cdots, q_{rs,i}, \cdots)^T ) for all ( rs \in R ) and ( i \in D ).</td>
</tr>
<tr>
<td>( q_{rs,i} )</td>
<td>Mean OD demand between OD pair ( rs ) of vehicle type ( i ), ( E[q_{rs,i}] = q_{rs,i} ).</td>
</tr>
<tr>
<td>( q )</td>
<td>( [\mathbf{R}][\mathbf{D}] )-vector of ( (\cdots, q_{rs,i}, \cdots)^T ) for all ( rs \in R ) and ( i \in D ), ( q = E[\mathbf{Q}] ).</td>
</tr>
<tr>
<td>( q_- )</td>
<td>The lower boundary of the estimated mean OD demand.</td>
</tr>
<tr>
<td>( q^+ )</td>
<td>The upper boundary of the estimated mean OD demand.</td>
</tr>
</tbody>
</table>
\( \hat{q}_{rs,i} \) Actual mean OD demand of vehicle class \( i \) between OD pair \( rs \).
\( \mathbf{R} \) Set of OD pairs. \( \mathbf{R} \) is a subset of \( \mathbb{N} \times \mathbb{N} \).
\( rs, rs' \) OD pair index, \( rs, rs' \in \mathbf{R} \).
\( t \) Lasso parameter.
\( u \) Iteration number.
\( V_{a,i} \) Random traffic flow on link \( a \) of vehicle class \( i \).
\( v_{a,i} \) Mean traffic flow on link \( a \), \( v_{a,i} = E[V_{a,i}] \).
\( \bar{V}_{a,i}^{(t)} \) Observed traffic flow of vehicle class \( i \) on link \( a \in \mathbf{A} \) during the observed time period on day \( l \) (\( l = 1, 2, \ldots, h \)).
\( \mathbf{R} \) Sample mean of multi-class link flows, which is calculated using sample \( \mathbf{H} \).
\( \bar{V}_{a,i}^{H} \) Sample mean of multi-class link flows, which is calculated using sample \( \mathbf{H}_n \).
\( \bar{V}_{a}^{(t)} \) -vector of \( (\ldots, \bar{V}_{a,i}^{(t)}, \ldots)^T \) for \( a \in \mathbf{A} \) and \( i \in \mathbf{D} \) on day \( l \).
\( \bar{V}_{a} \) Sample mean of observed multi-class link flows, \( \bar{V}_{a} = (\ldots, \bar{V}_{a,i}, \ldots)^T = \frac{1}{h} \sum_{l=1}^{h} \bar{V}_{a}^{(t)} \).
\( \delta_{m,a}^{(t)} \) Element of link-path incidence matrix \( \Delta \).
\( \lambda_{min} \) Minimal eigenvalue of \( \Sigma^g \).
\( \mu_1^{(t)}, \ldots, \mu_t^{(t)} \) Four positive penalty coefficients at iteration \( u \).
\( \xi \) Enlarge parameter in the adapted penalty function solution algorithm, \( \xi > 1 \).
\( \Sigma^f \) Covariance matrix of multi-class path flows, \( \Sigma^f = \{ \sigma_{rs,i,k,k'} \}_{[\mathbf{R}]^2} \times [\mathbf{P}]^2 \).
\( \Sigma^g \) Covariance matrix of traffic demands of all OD pairs for all vehicle classes, \( \Sigma^g = \{ \sigma_{rs,i,k,k'} \}_{[\mathbf{R}]^2} \times [\mathbf{P}]^2 \).
\( \Sigma^z \) Sample covariance matrix of observed multi-class link flows, \( \Sigma^z = \{ \bar{Z}_{a,i}^{H} \}_{[\mathbf{R}]^2} \times [\mathbf{P}]^2 \).
\( \sigma_{rs,i,k,k'} \) Covariance between \( F_{rs,j}^{i} \) and \( F_{rs,j'}^{k} \), \( \sigma_{rs,i,k,k'} = \text{cov}[F_{rs,j}^{i}, F_{rs,j'}^{k}] \).
\( \sigma_{rs,i,k,k'}^{s} \) Covariance between OD demand \( Q_{rs,i} \) and \( Q_{rs,j'} \), \( \sigma_{rs,i,k,k'}^{s} = \text{cov}[Q_{rs,i}, Q_{rs,j'}] \).
\( \sigma_{a,i}^{H} \) Covariance between link flows \( V_{a,i} \) and \( V_{a,i}^{'} \), \( \sigma_{a,i}^{H} = \text{cov}[V_{a,i}, V_{a,i}^{'}] \).
\( \sigma_{a,i}^{H} \) Sample covariance of multi-class link flows, which is calculated using sample \( \mathbf{H} \).
\( \sigma_{a,i}^{H} \) Sample covariance of multi-class link flows, which is calculated using sample \( \mathbf{H}_n \).
\( \tau \) Stopping tolerance for the proposed solution algorithm.

To facilitate the presentation of the essential ideas without loss of generality, the following basic assumptions are made in this paper.

A1. It is assumed that the total OD demand of all vehicle classes in the same OD pair are identical very day. The randomness or stochasticity of observed day-to-day multi-class traffic counts are only resulted from the random individual path and mode choices of the travelers. To explicitly model the interactions between travel modes (in terms of vehicle classes) as well as path choices, other sources of randomness resulting in randomness of multi-class traffic are ignored in this paper. These sources include the day-to-day variations of the total OD demand, travelers’ departure times, measurement errors and so on.

A2. Following assumption A1, it is assumed that the multi-class traffic demands between different OD pairs are independent. This assumption would be reasonable. For example, the higher usage of private may only result in lower usage of taxi of the same OD pair. And it may not influence usage of any traffic mode of other OD pairs.
A3. To facilitate the presentation of essential ideas, the mode-path choice proportion is defined as \( P_{irj} \), which is a random variable to account for the stochastic mode choice and path choice. The mean of mode-path choice proportion is assumed to be known and fixed, as such information could be available from some probe-vehicle data (e.g., taxi GPS data). The covariance of between different \( P_{irj} \) is set as the decision variable in this paper.

2.2. Covariance of stochastic multi-class OD demands

According to assumption A1, the daily total traffic demands of all vehicle classes during the same hourly period (e.g., morning peak, 8:00 am - 9:00 am) between OD pair \( rs \) is denoted as \( q_{rs} \), which is identical every day. \( q_{rs} \) is the summation of traffic demands of all vehicle classes as follows.

\[
q_{rs} = \sum_{i} q_{rs,i} \quad \forall \, rs \in R, \, i \in D
\]

where \( q_{rs,i} \) is the stochastic traffic demand of vehicle class \( i \), and \( E[q_{rs,i}] = q_{rs,i} \); \( R \) is the set of OD pairs; \( D \) is the set of vehicle classes. For convenience, denote \( Q \) and \( q \) as the \( |R||D| \)-vectors of \((\cdots, q_{rs,i}, \cdots)^T\) and \((\cdots, q_{rs,i}, \cdots)^T\) for all \( rs \in R \) and \( i \in D \), respectively. The covariance between OD demand \( q_{rs,i} \) and \( q_{rs,i'} \) is denoted as:

\[
\sigma_{rs,i,i'} = \text{cov}[Q_{rs,i}, Q_{rs,i'}] \quad \forall \, rs, rs' \in R, \, i, i' \in D
\]

The corresponding covariance matrix of stochastic multi-class OD demand can be expressed as:

\[
\Sigma^q = \sigma_{rs,i,i'} \quad \forall \, i, i' \in D
\]

**Remark 1:** It should be pointed out that although the traffic demand of each vehicle class is stochastic, the summation total OD demand of all vehicle classes is deterministic and fixed. Such feature differs the proposed model from the that of Shao et al. (2014). In their paper, the total OD demands is defined as random variables. The spatial covariance of OD demands between different OD pairs is to be estimated without consideration of stochastic traffic demand interactions between vehicle classes.

**Remark 2:** In line with assumption A2, it follows that:

\[
\sigma_{rs,i,i'} = 0, \quad \text{if} \quad rs \neq rs'
\]

This is because the assumption that multi-class traffic demands of different OD pairs are independent with each other. Thus, \( \Sigma^q \) is a block diagonal matrix.

**Remark 3:** The covariance matrix \( \Sigma^q \) describes the interactions of traffic demand between different vehicle classes of the same OD pair. It is generally believed that the higher usage of the one class of vehicle may not accompany with higher usage of the other classes of vehicle for the same OD pair. Thus, the value of \( \sigma^q_{rs,i,rs,i'} \) can be zero or negative but not positive.

**Remark 4:** The covariance matrix \( \Sigma^q \) has a special structural character. In view of assumption A1 and Equation (1), the variance of the total OD demand of all vehicle classes for each OD pair is zero. Then, the following condition should hold for \( \Sigma^q \).

\[
\sum_{i \in D} \sum_{i' \in D} \sigma^q_{rs,i,i'} = 0, \quad \forall \, rs \in R
\]

2.3. Conservation conditions between traffic flows in terms of mean and covariance

Let \( F^k_{rs} \) be the random traffic flow on path \( k \in K_{rs} \) of vehicle class \( i \) with its mean \( f^k_{rs} = E[F^k_{rs}] \), where \( K_{rs} \) is the path set between OD pair \( rs \) and \( K = \bigcup_{rs \in R} K_{rs} \). For convenience, \( F \) and \( f \) are denoted as the \( |K||D| \)-vectors of
\[(\cdots, F_{rs}^k, \cdots)^\top \text{ and } (\cdots, F_{rs}^{k'}, \cdots)^\top \text{ for all } k \in K_{rs}, rs \in R \
\] respectively. The path flows and OD demands satisfy the following flow conservation condition:

\[Q = \Lambda F \quad (6)\]

where \(\Lambda\) is the multi-class OD-path incidence matrix. Then, it follows that:

\[q = E[Q] = E[\Lambda F] = \Lambda f = \Lambda f \quad (7)\]

Equation (7) can be rewritten as:

\[q_{rs,i} = \sum_{k \in K_{rs}} f_{rs,i}^k \quad \forall rs \in R, i \in D \quad (8)\]

According to assumption A3, the path flow of vehicle class \(i\) is a product of the corresponding mode-path choice proportion and the OD demand as follows.

\[F_{rs,i}^k = P_{rs,i}^k q_{rs,i} \quad \forall k \in K_{rs}, rs \in R, i \in D \quad (9)\]

where \(P_{rs,i}^k\) is the random mode-path choice proportion of the traffic flow on path \(k \in K_{rs}\) of vehicle class \(i\). Then, it follows from Equation (9) and assumption A3 that:

\[f_{rs,i}^k = P_{rs,i}^k q_{rs,i} \quad \forall k \in K_{rs}, rs \in R, i \in D \quad (10)\]

According to assumption A3, the covariance between \(F_{rs,i}^k\) and \(F_{rs,i'}^{k'}\) can be deduced as:

\[\sigma_{rs,i,i'}^{k,k'} = \text{cov}[F_{rs,i}^k, F_{rs,i'}^{k'}] = \text{cov}[P_{rs,i}^k q_{rs,i}, P_{rs,i'}^{k'} q_{rs,i'}] = q_{rs,i} q_{rs,i'} \text{cov}[P_{rs,i}^k, P_{rs,i'}^{k'}] \quad \forall k \in K_{rs}, k' \in K_{rs}, rs', rs' \in R, i, i' \in D \quad (11)\]

The corresponding covariance matrix of multi-class path flows can be expressed as:

\[\Sigma = [\sigma_{rs,i,i'}^{k,k'}]_{rs,i,i' \in R, k,k' \in K} \quad (12)\]

According to Equation (6), the covariance conservation condition between multi-class path flows and multi-class OD demands is expressed as:

\[\sigma_{rs,i,i'}^{k,k'} = \sum_{k' \in K_{rs}} \sum_{k' \in K_{rs}} \sigma_{rs,i,i'}^{k,k'} = \sum_{k' \in K_{rs}} \sum_{k' \in K_{rs}} q_{rs,i} q_{rs,i'} \text{cov}[P_{rs,i}^k, P_{rs,i'}^{k'}] \quad \forall rs, rs' \in R, i, i' \in D \quad (13)\]

Denote \(\delta_{rs,a}^{i,i'}\) as the element of link-path incidence matrix \(\Delta\). If path \(k\) uses link \(a\), \(\delta_{rs,a}^{i,i'} = 1\). Otherwise, \(\delta_{rs,a}^{i,i'} = 0\). Then, the conservation condition of the estimated link and path flows for vehicle class \(i\) is expressed as:

\[V_{a,i} = \sum_{rs \in R} \sum_{k \in K_{rs}} \delta_{rs,a}^{i,i'} F_{rs,i}^k \quad \forall a \in A, i \in D \quad (14)\]

where \(V_{a,i}\) is the random traffic flow on link \(a\) of vehicle class \(i\). The mean link flow is denoted as \(v_{a,i} = E[V_{a,i}]\). It follows from Equation (14) that:

\[v_{a,i} = \sum_{rs \in R} \sum_{k \in K_{rs}} \delta_{rs,a}^{i,i'} q_{rs,i} = \sum_{rs \in R} \sum_{k \in K_{rs}} \delta_{rs,a}^{i,i'} F_{rs,i}^k \quad \forall a \in A, i \in D \quad (15)\]

where \(p_{rs,i}^k = E[P_{rs,i}^k]\), which is assumed to be known in this paper. It can be seen from Equation (15) that the mean of multi-class link flows is a linear function with respect to the mean of multi-class OD demands, which can be expressed as follows:

\[v_{a,i} = v_{a,i}(q) \quad \forall a \in A, i \in D \quad (16)\]

Also, the conservation condition of the estimated multi-class link and path flow covariances can be obtained as:

\[\sigma_{a,a',i,i'}^{i,i'} = \text{cov}[V_{a,i}, V_{a',i'}] = \text{cov}[\sum_{rs \in R} \sum_{k \in K_{rs}} \delta_{rs,a}^{i,i'} F_{rs,i}^k, \sum_{rs \in R} \sum_{k \in K_{rs}} \delta_{rs,a'}^{i,i'} F_{rs,i'}^{k'}] \]

\[= \sum_{rs \in R} \sum_{k \in K_{rs}} \sum_{k' \in K_{rs}} \sum_{k' \in K_{rs}} \delta_{rs,a}^{i,i'} \delta_{rs,a'}^{i,i'} \text{cov}[F_{rs,i}^k, F_{rs,i'}^{k'}] = \sum_{rs \in R} \sum_{k \in K_{rs}} \sum_{k' \in K_{rs}} \delta_{rs,a}^{i,i'} \delta_{rs,a'}^{i,i'} \sigma_{rs,i,i'}^{k,k'} \]

\[= \sum_{rs \in R} \sum_{k \in K_{rs}} \sum_{k' \in K_{rs}} \sum_{k' \in K_{rs}} \delta_{rs,a}^{i,i'} \delta_{rs,a'}^{i,i'} q_{rs,i} q_{rs,i'} \text{cov}[P_{rs,i}^k, P_{rs,i'}^{k'}] \quad \forall a,a' \in A, i,i' \in D \]
\[ = \sum_{n \in \mathbb{R}} \sum_{k \in \mathbb{R}} \sum_{k' \in \mathbb{R}} \sigma_{n,k,k'} \sigma_{n',k',k'} \left( \sum_{j \in \mathbb{D}} q_{n,j} \right) \left( \sum_{j' \in \mathbb{D}} q_{n',j'} \right) \text{cov} \left[ P_{n,j}, P_{n',j'} \right] \quad \forall \ a, a' \in \mathbb{A}, i, i' \in \mathbb{D} \]  

(17)

where \( \sigma_{n,k,k'} \) is the covariance between link flows of vehicle class \( n \) and \( k \) on link \( i \), and \( i', \mathbb{D} \) is the covariance matrix of link flows of vehicle class \( n \) and \( k \) on link \( i \). It can be seen from Equation (17) that the covariance of multi-class link flows is a linear function with respect to the covariance matrix of multi-class OD demands, which is expressed as follows.

\[ \sigma_{n,a,a'} = \sigma_{n,a,a'}(\mathbf{q}, \Sigma^p) \quad \forall \ a, a' \in \mathbb{A}, i, i' \in \mathbb{D} \]  

(18)

where \( \Sigma^p = \{ \sigma_{n,k,k'} \}_{0 \leq |k|,|k'| \leq t} = \text{cov} \left[ P_{n,j}, P_{n',j'} \right] \) is the covariance matrix of observed traffic counts.

2.4. Observed classified traffic counts

Denote \( \mathbb{A} \) as a subset of link set \( \mathbb{A} \). The traffic flows on link \( a \in \mathbb{A} \) can be observed by the traffic sensor during the observed time period. The link with (without) traffic sensor is called “observed link” (“unobserved link”) throughout this paper. Due to daily demand fluctuations, link flows of vehicle class \( n \) during the observed time periods vary from day to day. \( \tilde{v}^{(l)}_{n,a} \) is denoted as the observed traffic flow of vehicle class \( n \) on link \( a \in \mathbb{A} \) during the observed time period on day \( l \) \( (l = 1, 2, \cdots, h) \). Then, \( \mathbf{H} = \{ \tilde{v}^{(l)}_{n,a,\sigma} \tilde{v}^{(l)}_{a,\sigma} \cdots, \tilde{v}^{(h)}_{n,a,\sigma} \} \) is a sample of \( V_{n,a} \) with sample size \( h(h > 1) \). For convenience, \( \tilde{v}^{(l)}_{n,a} \) is denoted as the \( \mathbb{H} \)-vector of \( (\cdots, \tilde{v}^{(l)}_{n,a,\sigma}, \cdots)^T \) for all \( a \in \mathbb{A} \) and \( i, i' \in \mathbb{D} \) on day \( l \). Then, the sample mean of the observed multi-class link flows using sample \( \mathbf{H} \) can be calculated as:

\[ \tilde{v}_{a} = (\cdots, \tilde{v}^{H}_{a,\sigma}, \cdots)^T = \frac{1}{h} \sum_{l=1}^{h} \tilde{v}^{(l)}_{n,a} \]  

(19)

The sample covariance matrix of the observed multi-class link flows using sample \( \mathbf{H} \) can be calculated as:

\[ \tilde{\Sigma}^p = \left\{ \sigma_{n,a,a'} \right\}_{0 \leq |k|,|k'| \leq t} = \frac{1}{h-1} \sum_{l=1}^{h} \left( \tilde{v}^{(l)}_{n,a} - \tilde{v}_{a} \right) \left( \tilde{v}^{(l)}_{n,a} - \tilde{v}_{a} \right) \]  

(20)

where \( \tilde{\sigma}_{n,a,a'} \) is the sample covariance between observed traffic flows \( V_{n,a} \) and \( V_{a,a'} \) on link \( a \in \mathbb{A} \) and \( i, i' \in \mathbb{D} \).

2.5. Lasso method for multi-class stochastic OD demand estimation model

The lasso method for the estimation of mean and covariance of multi-class OD demands is formulated as follows.

\[ \min_{\mathbf{q}, \Sigma^p} \left\{ z = \sum_{a \in \mathbb{A}} \sum_{i,i' \in \mathbb{D}} \left( \tilde{v}^{H}_{n,a} - \tilde{v}_{a} \right)^2 + \sum_{a \in \mathbb{A}} \sum_{a' \in \mathbb{A}} \sum_{i,i' \in \mathbb{D}} \left( \tilde{\sigma}_{n,a,a'} - \sigma_{n,a,a'} \right)^2 (\mathbf{q}, \Sigma^p) \right\} \]  

(21a)

s.t.

\[ \sum_{a \in \mathbb{A}} \sum_{i,i' \in \mathbb{D}} \sum_{a' \in \mathbb{A}} \sum_{i,i' \in \mathbb{D}} \left| \sigma_{n,i,i',a,a'} \right| \leq t \]  

(21b)

\[ \mathbf{q} \geq \mathbf{q}^- \]  

(21c)

\[ \mathbf{q} \leq \mathbf{q}^+ \]  

(21d)

\[ \Sigma^p \succeq 0 \]  

(21e)

\[ \sum_{i,i' \in \mathbb{D}} \sigma_{n,i,i',a,a'} = \sum_{i,i' \in \mathbb{D}} \sum_{a' \in \mathbb{A}} \left( \sum_{j \in \mathbb{D}} q_{n,j} \right) \sigma_{n,a,a'} \succeq 0, \quad \forall r \in \mathbb{R} \]  

(21f)

where \( t \) is the lasso parameter, which can be determined by a \( n \)-fold cross-validation procedure. The lower and upper boundaries are attached to restrict the mean OD demands within an interval of \([ \mathbf{q}^-, \mathbf{q}^+ \]) and \([ \mathbf{q}^-, \mathbf{q}^+ \]) can be determined by prior or historical mean
OD demands. The boundary constraints mean that resultant mean multi-class OD demands should not be significantly changed in comparison with the prior one. $\Sigma^p \succeq 0$ represents that the covariance matrix of $\Sigma^p$ is a symmetric and positive semi-definite matrix. It can be seen that the optimization problem (21) is an extended model formulation of the conventional least squares (LS) method. The objective function is to minimize the squared difference between observed and estimated means and covariance matrices of link flows (Hazelton, 2003). The purpose of Equation (21a) is to make the estimated mean and covariance of link flows match (or approximate) the observed ones as closely as possible.

As mentioned above, there are too many elements in the covariance matrix to be estimated. However, the mode-path choice proportion covariance matrix has a special character. Some of the elements (say the covariances) in this matrix are actually zero. For example, it is assumed that the choices of private cars and goods vehicles are independent with each other, which means the corresponding covariance is zero. Due to the identifiability difficulty, the actual value (i.e. zero) of such covariance may not be obtained by the conventional OD estimation models (e.g. LS model). The outstanding feature of lasso method is exactly capable of overcoming this difficulty. The lasso method does variable selection and shrinkage by employing the L1-norm constraint in the conventional LS model. It causes shrinkage of estimated coefficients towards zero (Tibshirani, 1996). As such, lasso method shrinks some coefficients and sets others to zero. Regarding the multi-class OD demand covariance estimation herein, lasso method is adopted by the constraint (21b). It can be seen in the constraint (21b) that the covariances (excluding variances in the covariance matrix) are shrunk by the L1-norm constraint. It shrinks all the non-diagonal elements in the covariance matrix towards zero. The variances (diagonal elements in the covariance matrix) are not taken into account in the L1-norm constraint by setting $r_s \neq r'_s$ in Equation (21b). This is because the actual multi-class demand varies from day to day, which verifies that the corresponding variances are not zero. Thus, the variances should be not shrunk towards zero. An important issue of the lasso method is that the value of parameter $t$ controls the amount of shrinkage that is applied to the estimates (i.e. the covariances). Thus, it is important to determine the appropriate value of $t$. The cross-validation procedure is adopted and discussed in the following section to determine the value of lasso parameter $t$.

### 2.6. Cross-validation for determination of lasso parameter

To illustrate how to determine the lasso parameter $t$, a $n$-fold cross-validation procedure is proposed in this paper, where $n$ ($>1$) is an integer. For simplicity, it is assumed that $n = 10$ (i.e. tenfold) in the cross-validation. The observed multi-class link traffic flow data set $H$ is randomly divided into 10 subsets with the same sample size $\frac{1}{10}H$, $H = \bigcup_{i=1}^{10} H_i$. It is noted that for each subset $H_i$ is also a sample of the multi-class link flows. The corresponding sample mean and sample covariance of multi-class link flow can be calculated, which are denoted as $\tilde{\mu}_{a,j,o}^H$ and $\tilde{\sigma}_{a,j,o,a',o'}^H$, respectively. The lasso parameter $t$ can be determined by the following procedure.

**Step 0: Initialization.**

Set $t = 0$. Let $m$ be a large integer (In is paper, it is assumed that $m = 100$). The conventional LS estimates of $q$ and $\Sigma^p$ of can be obtained using the observed data set $H$, which are denoted as follows.

\[
(q^H, \Sigma^p_H) = \arg \min_{q^H, \Sigma^p_H} \left\{ \sum_{a \in A} \sum_{j \in J} \left( \tilde{\mu}_{a,j,o}^H - v_{a,j} q^H \right)^2 + \sum_{a \in A} \sum_{j \in J} \sum_{a' \in A} \sum_{j' \in J} \left( \tilde{\sigma}_{a,j,o,a',o'}^H - \sigma_{a,j,o,a',o'}^p \right) \right\} \quad (22a)
\]

s.t.

\[
q^H \geq q^- \quad (22b)
\]

\[
q^H \leq q^+ \quad (22c)
\]

\[
\Sigma^p_H \succeq 0 \quad (22d)
\]
\[
\sum_{i \in D} \sum_{j, k, s, r, s' \in K_k} \sigma_{i,j,s}^{H} = \sum_{s \in K_k} \sum_{j, r, s'} \left( \sum_{i \in D} q_{i,j,s} \right)^2 \sigma_{i,j,s}^{H} = 0, \quad \forall rs \in R
\]  
(21f)

where \( q^H \) is denoted as a vector \((\cdots, q_{m,s}^H, \cdots)^T \) all \( rs \in R \) and \( i \in D \), \( \Sigma^{p,H} \) is denoted as the matrix \[
\Sigma^{p,H} = \left\{ \sigma_{i,j,s}^{H} \right\}_{|p||p|}. \]

Step 1: Estimation errors for possible values of \( t \)

For each \( t = \frac{1}{m} \), \( \frac{2}{m} \), \( \cdots, \frac{m-1}{m} \), \( s \) (i.e. \( t_j = j \times \Delta t \), where \( \Delta t = \frac{1}{m} \)), repeat Steps 1.1 and 1.2.

Step 1.1 For each \( H_s \), repeat Steps 1.1.1 and 1.1.2.

Step 1.1.1 Use data \( H_s = (H_1, H_2, \cdots, H_{s-1}, H_{s+1}, \cdots, H_{10}) \), to solve the following optimization problem.

\[
(q^{H,H}, \Sigma^{p,H-H}) = \arg \min_{q^H, \Sigma} \left\{ \sum_{i,j,s} \left( v_{i,j,s} - (q^{H,H})_{i,j,s} \right)^2 + \sum_{i,j,s} \sum_{i,j,s} \sum_{i,j,s} \left( \sigma_{i,j,s}^{n} - \sigma_{i,j,s}^{n} \right)^2 (q^{H,H}, \Sigma^{p,H-H}) \right\}
\]  
(23a)

s.t.
\[
\sum_{i,j,s} \sum_{i,j,s} \sum_{i,j,s} \sum_{i,j,s} \sum_{i,j,s} \sum_{i,j,s} \left( \sigma_{i,j,s}^{n} - \sigma_{i,j,s}^{n} \right)^2 (q^{H,H}, \Sigma^{p,H-H}) \leq t_j
\]  
(23b)
\[
q^{H,H} \geq q^H
\]  
(23c)
\[
q^{H,H} \leq q^H
\]  
(23d)
\[
\Sigma^{p,H-H} \geq 0
\]  
(23e)

Step 1.1.2 Use \( H_s \) to calculate the estimation error
\[
\left( \sum_{i,j,s} \sum_{i,j,s} \sum_{i,j,s} \sum_{i,j,s} \sum_{i,j,s} \sum_{i,j,s} \left( \sigma_{i,j,s}^{n} - \sigma_{i,j,s}^{n} \right)^2 (q^{H,H}, \Sigma^{p,H-H}) \right)^{1/2} = \epsilon_{H_s}.
\]  
(24)

Step 1.2 Calculate the mean estimation error of the lasso parameter \( t_j \)
\[
\epsilon(t_j) = \frac{1}{10} \sum_{i=j}^{10} \epsilon_{H_s}
\]  
(25)

Step 2: Determine \( t \).
\[
t = \arg \min_{j=0,1,2,\cdots, m} \epsilon(t_j)
\]  
(26)

At each iteration of the cross-validation procedure, 90% observed data is used for calibration while 10% observed data is used for validation. It can be seen from the above cross-validation procedure that the lasso parameter is determined to obtain a “best” shrinkage of the estimates with minimum estimation errors on the basis of the whole observed data. It should be noted that the determined value of lasso parameter \( t \) is strictly less than value of \( s \) (the L1-norm of covariance matrix by conventional LS method). \( t < s \) will cause shrinkage of the estimated parameters of the conventional LS method towards zero, and some parameters may be exactly equal to zero (Tibshirani, 1996). In such a way, the covariance, which is actually zero, is shrunk towards zero or even driven to be zero.

Remark 5: The lasso parameter \( t \) is determined on the basis of the observed data, the values of \( t \) may be different for different set of the observed data even if for the same network.

Remark 6: On the one hand, the larger the value of \( m \), the more accurate the lasso parameter \( t \). On the other hand, the larger the value of \( m \), the more computational time. Thus, how to select the value of \( m \) remains interesting investigations in further study.
3. Solution algorithm

If the lasso parameter $t$ is known, the mean and covariance of multi-class OD demands can be obtained by solving the constrained optimization problem (21). It should be noted that the problem (21) is a non-linear constrained optimization problem. The main difficulty for solving it comes from the non-linear constraint (21e). In the literature of optimization, the non-linear constrained optimization problem is usually transformed into an unconstrained one.

Penalty method is one of the most widely used approaches for transforming the constrained optimization problem into the unconstrained one. In this paper, the exterior penalty method is employed due to its simplicity and efficiency. In connection to the constraints (21b)-(21f), the following penalty terms are proposed to be added in the unconstrained one.

The objective function for optimization problem (21) can be expressed as following.

$$\sum_{r \in R, s \in S} \sum_{i \in K_r, k \in K_s} \sum_{r' \in R, s' \in S} \sum_{i' \in K_r', k' \in K_s'} \sigma_{r,r',s,s'}^{p,K,K'} \leq t \Rightarrow \mu_1^{(a)} \left( \sum_{r \in R, s \in S} \sum_{i \in K_r, k \in K_s} \sum_{r' \in R, s' \in S} \sum_{i' \in K_r', k' \in K_s'} \sigma_{r,r',s,s'}^{p,K,K'} \right)^2$$

$$q \geq q \Rightarrow \mu_2^{(a)} \left( \min \{ q - q \cdot 0 \} \right)^2$$

$$q^* \geq q \Rightarrow \mu_3^{(a)} \left( \min \{ q^\prime - q \cdot 0 \} \right)^2$$

$$\Sigma^p \geq 0 \Rightarrow \mu_4^{(a)} \left( \min \{ \lambda_{\min} \cdot 0 \} \right)^2$$

$$\Sigma = \sum_{r \in R, s \in S} \sum_{i \in K_r, k \in K_s, i' \in K_r', k' \in K_s'} \left( \sum_{q \in B} q \right)^2 \sigma_{r,r',s,s'}^{p,K,K'} = 0 \Rightarrow \mu_5^{(a)} \left( \sum_{r \in R, s \in S} \sum_{i \in K_r, k \in K_s} \sum_{r' \in R, s' \in S} \sum_{i' \in K_r', k' \in K_s'} \sum_{q \in B} q \sigma_{r,r',s,s'}^{p,K,K'} \right)^2$$

where $\mu_1^{(a)}$, $\mu_2^{(a)}$, $\mu_3^{(a)}$, $\mu_4^{(a)}$ and $\mu_5^{(a)}$ are the five positive penalty coefficients at iteration $u$; $\lambda_{\min}$ is the minimal eigenvalue of $\Sigma^p$. Equation (30) means that the minimal eigenvalue of $\Sigma^p$ should be non-negative. It implies that all the eigenvalues of $\Sigma^p$ are non-negative. Otherwise, a penalty is given in the objective function. Then, the penalized objective function for optimization problem (21) can be expressed as following.

$$\min_{q \in \mathbb{R}} \left\{ z = z(q, \Sigma^p) + z_1(q, \Sigma^p, \mu_1^{(a)}, \mu_2^{(a)}, \mu_3^{(a)}, \mu_4^{(a)}, \mu_5^{(a)}) \right\}$$

where

$$z(q, \Sigma^p) = \left( \sum_{a \in A} \sum_{\alpha \in A} \sum_{\alpha' \in A} \sum_{\beta \in B} \sum_{\beta' \in B} (v_{\alpha, \alpha'}^H - v_{\alpha, \alpha'}(q))^2 + \sum_{a \in A, \alpha \in A, \beta \in B} \sum_{\alpha' \in A, \beta' \in B} (g_{\alpha, \alpha', \beta, \beta'}^H - g_{\alpha, \alpha', \beta, \beta'}(q, \Sigma^p))^2 \right)$$

$$z_1(q, \Sigma^p, \mu_1^{(a)}, \mu_2^{(a)}, \mu_3^{(a)}, \mu_4^{(a)}, \mu_5^{(a)}) = \mu_1^{(a)} \left( \sum_{r \in R, s \in S} \sum_{i \in K_r, k \in K_s, i' \in K_r', k' \in K_s'} \sum_{q \in B} q \sigma_{r,r',s,s'}^{p,K,K'} \right)^2 + \mu_2^{(a)} \left( \min \{ q - q \cdot 0 \} \right)^2 + \mu_3^{(a)} \left( \min \{ q^\prime - q \cdot 0 \} \right)^2$$

Based on the exterior penalty method, the flowchart of the proposed solution algorithm is presented in Figure 2. It should be noted that the gradient of the objective function (32) is difficult to obtain due to the penalty term corresponding to the non-linear inequality constraint (Equation (30)). Therefore, some derivative-free optimization methods should be employed for solving the minimization problem in the upper level, such as the simplex search method (Lagarias et al., 1998) and generalized pattern search methods (Torczon, 1997; Audet and Dennis, 2003). Details on the derivative-free optimization methods can be referred to Conn et al. (2009). In this paper, the simplex search method (Lagarias et al., 1998) is used to solve the unconstrained optimization problem in Step 2, which is available in the Matlab optimization toolbox by the subroutine “fminsearch”.

Details on the derivative-free optimization methods can be referred to Conn et al. (2009). In this paper, the simplex search method (Lagarias et al., 1998) is used to solve the unconstrained optimization problem in Step 2, which is available in the Matlab optimization toolbox by the subroutine “fminsearch”.

Details on the derivative-free optimization methods can be referred to Conn et al. (2009). In this paper, the simplex search method (Lagarias et al., 1998) is used to solve the unconstrained optimization problem in Step 2, which is available in the Matlab optimization toolbox by the subroutine “fminsearch”.
The convergence of the employed exterior penalty solution algorithm can be guaranteed by the following proposition.

**Proposition 1**: Suppose that each \( \{q^{(n+1)}, \Sigma^{p,(n+1)}\} \) is the exact global minimizer of minimization problem (30), then, every limit point \( \{q^*, \Sigma^{p,*}\} \) of the sequence \( \{q^{(n+1)}, \Sigma^{p,(n+1)}\} \) is a global solution of the optimization problem (21) (Nocedal and Wright, 2006).

---

Fig. 2. The flowchart of the proposed solution algorithm
4. Numerical examples

4.1. Preliminary

To demonstrate the properties of the proposed model, a small tractable network is chosen. A simple network with six nodes, seven links, two OD pairs (1→3 and 2→4) and four paths, as shown in Figure 3, is used to illustrate the applications of the proposed stochastic multi-class OD demand estimation model. It is assumed that there are three classes of vehicles, i.e. private car (pc), taxi (tx) and goods vehicle (gv). That is to say, \( D = \{pc, tx, gv\} \). The corresponding pcu for these three vehicles are set as: \( 1 \text{ pc} = 1 \text{ pcu}, 1 \text{ tx} = 1 \text{ pcu} \) and \( 1 \text{ gv} = 2.0 \text{ pcu} \). The mean path choice proportions are all set to be \( 6/1 \), i.e. \( p_{i,r,s}^k = 6/1, \forall k \in K, rs \in R, i \in D \). Suppose that the classified traffic counts during the same hourly period of weekdays have been collected over the whole year (about 300 days). Thus, the sample size is set to be \( h = 300 \). The actual mean and covariance of the multi-class OD demands are given in Tables 2 and 3 respectively, while the observed link flows by vehicle class and their covariances are shown in Tables 4 and 5, respectively.

![Fig. 3. A small test network](image)

Table 2. The actual mean multi-class OD demands (pcu/hour)

<table>
<thead>
<tr>
<th>OD pair</th>
<th>Private car ( (\hat{q}_{rs,pc}) )</th>
<th>Taxi ( (\hat{q}_{rs,tx}) )</th>
<th>Goods vehicle ( (\hat{q}_{rs,gv}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>1200</td>
<td>1200</td>
<td>1200</td>
</tr>
<tr>
<td>24</td>
<td>1200</td>
<td>1200</td>
<td>1200</td>
</tr>
</tbody>
</table>

Table 3. The actual multi-class OD demand covariance matrix (pcu/hour)

<table>
<thead>
<tr>
<th>( \text{cov}(Q_{rs,\cdot},Q_{rs',\cdot}) )</th>
<th>OD Pair 1→3</th>
<th>OD Pair 2→4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Private car</td>
<td>Taxi</td>
</tr>
<tr>
<td>Private car</td>
<td>216000</td>
<td>-216000</td>
</tr>
<tr>
<td>Taxi</td>
<td>-216000</td>
<td>216000</td>
</tr>
<tr>
<td>Goods vehicle</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OD Pair 2→4</td>
<td>Private car</td>
<td>0</td>
</tr>
<tr>
<td>Taxi</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Goods vehicle</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4. The observed classified link flows (pcu/hour)

<table>
<thead>
<tr>
<th>Link no</th>
<th>Private car ( (\tilde{V}_{rs,pc}) )</th>
<th>Taxi ( (\tilde{V}_{rs,tx}) )</th>
<th>Goods vehicle ( (\tilde{V}_{rs,gv}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>600</td>
<td>600</td>
<td>600</td>
</tr>
<tr>
<td>2</td>
<td>600</td>
<td>600</td>
<td>600</td>
</tr>
<tr>
<td>4</td>
<td>1200</td>
<td>1200</td>
<td>1200</td>
</tr>
</tbody>
</table>
4.2. Stochastic interactions of mode and path choices

One of the outstanding features of the proposed model is capability of figuring out the interactions between vehicle classes by a statistical method. Specifically, the mode-path choice proportion \( k_{irsP} \) is used to capture the stochastic interactions of mode and path choices. The expected value of \( k_{irsP} \) is assumed to be known in this example. However, the covariance matrix of all \( \text{cov}[k_{irsP}] \) in the whole network is treated as decision variable.

The resultant mode-path choice proportion covariance matrix for this example is shown in Table 6. The covariance information in Table 6 can reflect the stochastic interactions between vehicle classes as well as the path choice proportions. For example, the covariance between traffic flows of paths 1 and 2 using private car is negative, i.e. -0.0054, which means the more private cars use path 1 accompanies with less private cars use path 2. Such path choice behavior is reasonable. The reason is that paths 1 and 2 are two parallel paths of the same OD pair. The more amount of traffic flow of the same vehicle class on one path will result in the less amount on the other one. The stochastic interactions between vehicle classes can also be demonstrated by covariance information in Table 6. For path 1, the covariance between private car and taxi traffic flows is also negative, i.e. -0.0107. Such negative covariance means that for the same path the higher usage of private car leads to the lower usage of taxi. Meanwhile, the for path 1, the covariance of private car and goods vehicle traffic flows is zero. This demonstrates that the traffic flows of private car and goods vehicle is not linearly dependent. They may be independent and not have interactions. In reality, the mode-path choice proportion covariance information provides an alternative simple way to capture the interactions between path choices and mode choices from a statistical viewpoint.

According to Equation (11), the corresponding multi-class path flows can be calculated easily, which is shown in Table 7. It should be noted that the multi-class OD demand covariance matrix can be calculated using the results in Table 7 according to Equation (13). For example, the summation of the four italic covariance numbers (-138341 + 37415 + 37415 - 138341) in Table 7 can obtain the OD demand covariance between private car and taxi (-201852) for OD pair 1→3. Such covariance information of stochastic multi-class OD demands can be used to evaluate the network performance with respect to different vehicle classes for network with uncertainty. Due to the length limitation of this paper, similar illustration can be found in section 4.1.3 in Shao et al. (2014).

### Table 5. The observed classified link flow covariance matrix (pcu/hour)²

<table>
<thead>
<tr>
<th>Link 1</th>
<th>Link 2</th>
<th>Link 3</th>
<th>Link 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>OD Pair 1→3</td>
<td>OD Pair 2→4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private car</td>
<td>Taxi</td>
<td>Goods vehicle</td>
<td>Private car</td>
</tr>
<tr>
<td>Path 1</td>
<td>Path 2</td>
<td>Path 1</td>
<td>Path 2</td>
</tr>
<tr>
<td>0.0143</td>
<td>-0.0054</td>
<td>0.0029</td>
<td>0</td>
</tr>
<tr>
<td>-0.0054</td>
<td>0.00143</td>
<td>0.0029</td>
<td>-0.0107</td>
</tr>
<tr>
<td>-0.0107</td>
<td>0.0029</td>
<td>0.156</td>
<td>-0.0089</td>
</tr>
<tr>
<td>0.0029</td>
<td>-0.0107</td>
<td>-0.0089</td>
<td>0.156</td>
</tr>
<tr>
<td>-0.0056</td>
<td>0.0029</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.0056</td>
<td>-0.0056</td>
<td>-0.0056</td>
<td>0</td>
</tr>
</tbody>
</table>
Path 3 0 0 0 0 0 0 0.0143 -0.0054 -0.0107 0.0029 0 0
Private car
Path 4 0 0 0 0 0 0 -0.0054 0.0143 0.0029 -0.0107 0 0
Taxi
Path 3 0 0 0 0 0 0 -0.0107 0.0029 0.0156 -0.0089 0 0
Goods vehicle
Path 4 0 0 0 0 0 0 0 0 0 0 0.0056 -0.0056
OD Pair 2 → 4
Path 3 0 0 0 0 0 0 -0.0054 0.0143 0.0029 -0.0107 0 0
Private car
Path 4 -69929 185178 37415 -138341 0 0 0 0 0 0 0 0
Taxi
Path 3 -138341 37415 201722 -115345 0 0 0 0 0 0 0 0
Goods vehicle
Path 4 0 0 0 0 -72000 72000 0 0 0 0 0 0
OD Pair 1 → 3
Table 7. The resultant multi-class path flow covariance matrix (pcu/hour)

<table>
<thead>
<tr>
<th>OD Pair</th>
<th>Path 1</th>
<th>Path 2</th>
<th>Path 3</th>
<th>Path 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private car</td>
<td>185178</td>
<td>-69929</td>
<td>37415</td>
<td>0</td>
</tr>
<tr>
<td>Taxi</td>
<td>-138341</td>
<td>37415</td>
<td>-138341</td>
<td>0</td>
</tr>
<tr>
<td>Goods vehicle</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

4.3. Lasso vs LS methods

To test the performance of lasso method in comparison to the conventional LS methods, three scenarios are used and shown in Table 9. In Scenario A, the observed traffic data contains the “full information” of the multi-class OD demands. It is easy to verify that the traffic flow information of links 1, 2 and 4 covers all the needed information for estimation of multi-class OD demand. That is to say, the observed traffic data from links 1, 2 and 4 is sufficient to estimate the multi-class OD demands. Thus, Scenario A is called the “full information” scenario in this example. In Scenarios B and C, the observed traffic data only contains the “partial information” of the multi-class OD demands. In Scenario B, the observed traffic data only includes the data of private car and taxi but not the goods vehicles. In Scenario C, the observed traffic data only includes the traffic flow information for OD pair 13. In this example, root mean squared error (RMSE) is used to show the estimation errors for the three scenarios. The corresponding formulae of RMSEs for mean and covariance matrix of multi-class OD demands are given as below.

$$RMSE_q = \sqrt{\frac{1}{|\mathbf{R}|} \sum_{i \in \mathbf{R}} \sum_{j \in \mathbf{R}} (\mathbf{q}_{ij} - \hat{\mathbf{q}}_{ij})^2}$$  (35)
\[
\text{RMSE}_{\sigma_{ij}} = \sqrt{\frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \sum_{c \in R} \sum_{r \in R} \left( \sigma_{r,i,c} - \hat{\sigma}_{r,i,c} \right)^2} \tag{36}
\]

It can be seen from Table 9 that under full information condition (Scenario A) both lasso and LS methods have the same estimation results. As the RMSEs are small (RMSE$_{q}=0$ (pcu/hour) and RMSE$_{\sigma}=6752$ (pcu/hour)$^2$), the estimated mean and covariance matrix are close to the actual ones. Thus, if the observed traffic data is sufficient, there is no difference between lasso and conventional LS method. However, if the traffic count data is insufficient (i.e. overfitting occurs), the lasso method could outperform the LS method, which can be evidenced by the results in Scenarios B and C. For instance, if the traffic data of good vehicles is unavailable, the lasso method could obtain a better estimation of multi-class OD demand covariance matrix than the conventional LS method. For lasso method in Scenario C, the estimated results are better than that of LS method. For example, in lasso method RMSE$_{\sigma}=28461$ (pcu/hour)$^2$ which is smaller than in LS method (RMSE$_{\sigma}=45588$ (pcu/hour)$^2$). In Scenario B, the performance of lasso method is still better than that of LS method. It is shown in this example, that the lasso method is an extension of the LS method. The estimation error of lasso method is smaller than that of LS method if the observed data is insufficient. That is to say the proposed lasso method is more suitable for handling the overfitting issue than the LS method. And the lasso method could outperform the conventional LS method on overcoming the identifiability difficulty for estimating the mean and covariance of multi-class OD demands.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Observed links</th>
<th>Descriptions</th>
<th>LS method</th>
<th>Lasso Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(pcu/hour)</td>
<td>(pcu/hour)$^2$</td>
</tr>
<tr>
<td>A</td>
<td>1, 2, 4</td>
<td>Full information</td>
<td>$0$ (RMSE$_{q}$)</td>
<td>$6752$ (RMSE$_{\sigma}$)</td>
</tr>
<tr>
<td>B</td>
<td>1, 2, 4 (only two vehicle classes)</td>
<td>Partial information</td>
<td>$326$ (RMSE$_{q}$)</td>
<td>$9033$ (RMSE$_{\sigma}$)</td>
</tr>
<tr>
<td>C</td>
<td>1, 3 (only cover OD pair 1-3)</td>
<td>Partial information</td>
<td>$936$ (RMSE$_{q}$)</td>
<td>$45588$ (RMSE$_{\sigma}$)</td>
</tr>
</tbody>
</table>

5. Conclusions and further studies

This paper proposed a new model for estimating the mean and covariance of stochastic multi-class OD demands based on the classified traffic counts for the same hourly period throughout the year. Different from the conventional OD demand estimation models, the proposed model utilized the statistical properties of the observed hourly traffic count data by vehicle class (or type) over the whole year. Also, the stochastic mode and path choices were explicitly considered in the proposed model using a random mode-path choice proportions. To overcome the identifiability difficulty in this paper, the lasso (least absolute shrinkage and selection operator) method was incorporated into the proposed model. A cross-validation procedure was proposed to determine the lasso parameter. An equivalent non-linear constrained optimization model was proposed and formulated as the corresponding OD demand estimation problem shown in this paper. A heuristic solution algorithm based on exterior penalty function method was adapted to solve the proposed model.

It was found in the numerical examples that (i) the estimated covariance of mode-path choice proportions can reflect the interactions of path and mode choices between different vehicle classes; (ii) the proposed model based on the lasso method could outperform the conventional least squares (LS) model on overcoming the identifiability difficulty for model application.

On the basis of the model proposed in this paper, some further extensions can be envisaged as follows.

- This paper only estimates the vehicle class covariance of stochastic multi-class traffic demands during the same hourly period over the year. Further investigation should be carried out on how to extend the proposed model to simultaneously estimate the spatial and temporal covariance of the multi-class traffic demands between different hourly periods in dynamic or time-dependent network models.
- Although the lasso method could overcome the identifiability difficulty to some extent in the proposed OD demand estimation model, multiple solutions may still exist. Also, the estimation errors still existed according the
results of the numerical example shown in this paper. How to address this issue in practice still reveals further investigations.

- As the mean of mode-path choice proportion is assumed to be known in this paper. How to relax this assumption by taking into account the traveler’s path choice behavior in network with uncertainty would be worthwhile for further study.

Acknowledgements

The work described in this paper was jointly supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. 5187/12E and 5181/13E), Research Committee of The Hong Kong Polytechnic University (Project No. G-YJ93), and National Natural Science Foundation of China (Project Nos. 70901073, 71271205 and 71171013).

References


