Efficient real time OD matrix estimation based on Principal Component Analysis

TRAIL Research School, October 2012

Authors
Tamara Djukic MSc, Gunnar Flötteröd PhD MSc, Hans van Lint PhD MSc and Prof. Serge Hoogendoorn PhD MSc
Faculty of Civiel Engineering and Geosciences, Department of Transport & Planning, Delft University of Technology, the Netherlands
Department of Traffic and Logistics, Royal Institute of Technology, KTH, Stockholm, Sweden

© 2012 by T. Djukic and TRAIL Research School
Contents

Abstract

1 INTRODUCTION .................................................................................................. 1

2 Reduced OD model formulation ............................................................................ 2
2.1 The idea of dimensionality reduction ................................................................. 2
2.2 The dimensionality reduction based on PCA ....................................................... 2

3 State space formulation of the model ................................................................. 4
3.1 Kalman filter formulation of the model .............................................................. 4
3.2 Colored noise Kalman filter ............................................................................. 4

4 Experiment: Synthetic data ................................................................................... 7
4.1 Simulating the OD demand .............................................................................. 7
4.2 The model performance .................................................................................. 8

5 Conclusion .......................................................................................................... 10

Acknowledgment .................................................................................................. 11

References ............................................................................................................ 11
Abstract

In this paper we explore the idea of dimensionality reduction and approximation of OD demand based on principal component analysis (PCA). First, we show how we can apply PCA to linearly transform the high dimensional OD matrices into the lower dimensional space without significant loss of accuracy. Next, we define a new transformed set of variables (demand principal components) that is used to represent the OD demand in lower dimensional space. We use these new variables as state variable in a novel reduced state space model for real time estimation of OD demand. Through an example we demonstrate the quality improvement of OD estimates using this new formulation and a so-called ‘colored’ Kalman filter over the standard Kalman filter approach for OD estimation, when correlated measurement noise is accounted due to reduction of variables in state vector.

Keywords
Dynamic OD matrix, Dimensionality reduction, Principal Demand Component, Kalman Filter
1 INTRODUCTION

Much of the work in OD matrix estimation and prediction so far has focused on improving estimation and prediction of OD matrices with more sophisticated and less time consuming algorithms (Bierlaire and Crittin 2004; Kattan and Abdulhai 2006; Zhou and Mahmassani 2007) and by including additional available data, ranging from traffic counts to automatic identification data (Asakura, Hato et al. 2000; Antoniou, Ben-Akiva et al. 2004; Dixon and Rilett 2005), and data from Bluetooth devices (Barcelo, Montero et al. 2010), to name a few. Lately, decomposition of network into smaller subareas has been proposed by authors (Lou and Yin 2010; Frederix 2011) to deal with high dimensional OD estimation problem. However, the OD estimation problem remains computationally intensive because these methods have to deal not only with high dimensional structures of OD matrices but also with the complexity of these methods.

One of the problems with high-dimensional datasets is that, in many cases, not all the measured variables are “important” for understanding the underlying phenomena of interest. In other words, we may believe that high-dimensional data are multiple, indirect measurements of an underlying source. While certain computationally expensive novel methods can construct predictive models with high accuracy from high-dimensional data (Zhou and Mahmassani 2007), it is still of interest in real time applications to reduce the dimension of the OD demand data prior to any modeling of the data.

Therefore, one possible solution approach to solve this “curse of dimensionality” is to map the high dimensional OD matrices into a space of lower dimensionality, such that most of the structural information about the demand is preserved. In this paper we apply the Principal Component Analysis (Jolliffe 2002) or PCA, also known as the Karhunen-Loeve transform, commonly used method for this purpose. PCA is a linear transformation technique for dimensionality reduction that performs a linear mapping of the data to a lower dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized.

In this paper we show that any data set of observed OD flows or generated from detailed demand microsimulation system can be represented, without loss of generality, as a linear combination of a set of only a few orthonormal vectors (eigenvectors) and principal demand components. In our method, first we extract offline these Eigenvectors that capture the trip-making patterns and their spatial and temporal variations, whereas the principal demand components capture the contribution of each eigenvector to the realization of a particular OD demand. These principal demand components define the fixed structure of our OD matrices, which we then update on-line from traffic counts. They are used as state variables instead of the OD flows themselves, which leads to a simplified state space model that can be solved recursively using the so-called colored noise Kalman filter (Bryson 1968). Reducing the problem dimensionality through PCA replaces the usual approach of using prior OD matrices by structural information obtained either from data or from a detailed demand microsimulation system. The importance and originality of this approach lie in the possibility to capture the most important structural information without loss of accuracy and considerably decreasing the model complexity.

The paper is organized as follows. In the first part of the paper, we present the idea of dimensionality reduction and approximation of OD demand based on PCA. In the second part of the paper, we propose the new state space formulation of the OD
estimation model with principal demand components as the state variables. Next, we explore the properties of the colored noise Kalman filter to solve the proposed OD estimation method with time-correlated measurements. In the third part of paper, we investigate the quality improvement of estimates over standard Kalman filter when colored measurements are accounted due to degradation of variables in state vector on the synthetic network. The paper closes with a discussion on further application perspectives of the new model in estimation and prediction of OD demand and further research directions.

2 Reduced OD model formulation

2.1 The idea of dimensionality reduction

Since OD matrices are high dimensional multivariate data structures, the specification and estimation of OD matrices is both methodologically and computationally cumbersome for real time applications. There are three factors that increase the computational effort: the size of the state vector, the complexity of model components (e.g. assignment matrix, covariance matrices, etc.), and the number of measurements to be processed. For example, the Kalman filter algorithm is commonly used method to estimate and predict the OD matrices (Ashok, Ben-Akiva et al. 1993; Antoniou, Ben-Akiva et al. 2006; Barcelo 2010). Since the computational complexity of the Kalman filter is typically in the order of $O(n^3)$, where in the simplest case $n$ is the total number of the OD pairs in the network, this can represent a potential computational bottleneck. Clearly, reducing the dimensionality of the state vector, is a path to improve computational efficiency. For example, let us assume that OD flows have been estimated for several previous days or months. These flows subsume in them various kinds of information, about trip making patterns and their spatial and temporal variations. Therefor, the key idea in our approach is to reduce the dimensionality of the OD matrix, in such way that the structural and temporal patterns are preserved. With this approach the computational cost can be speeded up dramatically, without significant lost of accuracy. One commonly used method of dimensionality reduction is a linear transformation technique known as Principal Component Analysis (PCA). PCA has found application in traffic and transportation science before, for example for the dimensionality reduction in calibration of travel demand from traffic counts (Flötteröd 2009).

In the reminder of this paper, we will explain the main features of PCA method and how we use eigenvectors to define a fixed low dimensional structure of OD demand. Next, we will present a new state space model formulation that can be solved with a variation of the Kalman Filter algorithm to update OD flows on-line from traffic counts.

2.2 The dimensionality reduction based on PCA

Our goal is to map vectors of the OD demand $X \in \mathbb{R}^n$ onto the new vector in an $M$-dimensional space, where $M < n$. To this end we first demonstrate the remarkable fact that any OD matrix has a concise representation when expressed in terms of an orthonormal basis of $n \times 1$ vectors $e_i$, $i=1,2,...,n$ that can be derived using Principal Component Analysis or PCA, also known as the Karhunen-Loeve procedure and is discussed in detail in (Jolliffe 2002).
For example, suppose that we have used a microsimulation-based demand model to generate off-line a large sample of OD demand observations \( r \), each being a realization of the \( n \)-dimensional OD demand \( x = (x_1, x_2, \ldots, x_n) \). Thus, we have a \( r \times n \) observation matrix \( X \), where \( n \) represents the number of OD pairs in the following form:

\[
X = \begin{pmatrix}
x_1(1) & \cdots & x_n(1) \\
x_1(2) & \cdots & x_n(2) \\
\vdots & \ddots & \vdots \\
x_1(r) & \cdots & x_n(r)
\end{pmatrix}
\]

Once we generate the matrix \( X \), we apply off-line the PCA algorithm to extract the eigenvectors \( e_i \), \( i = 1, 2, \ldots, n \) and eigenvalues \( \lambda_i \), \( i = 1, 2, \ldots, n \).

Since the covariance matrix of \( X \) is real and symmetric, its eigenvectors \( e_1, e_2, \ldots, e_n \) can be chosen as an orthonormal basis. Therefore, any vector \( x \), or actually \( (x - \bar{x}) \), can be represented, without loss of generality, as a linear combination of a set of \( n \) orthonormal vectors \( e_i \):

\[
x - \bar{x} = c_1 e_1 + c_2 e_2 + \ldots + c_n e_n = \sum_{i=1}^{n} c_i e_i
\]

in which explicit expressions for the coefficients \( c_i \) in (3) is given by

\[
c_i = e_i^T (x - \bar{x})
\]

which can be regarded as a simple rotation of the coordinate system from original \( x \)’s to a new set of coordinates given by \( c \)’s. Through sorting the eigenvectors in decreased order in the light of the size of the eigenvalue, we can retain the first \( M \) eigenvectors with the maximum data variance. However, since the covariance matrix of observed OD demand in general can be very large, it is inconvenient to evaluate and store it explicitly. To avoid this we can use efficient algorithms, which find the \( M \) largest eigenvectors of the covariance matrix such as the orthogonal iteration and power method (Golub and van Van Loan 1996).

Once the \( M \) largest eigenvectors \( e_1, e_2, \ldots, e_M \) are found, a new low dimensional representation of the OD demand can be expressed as follows

\[
\hat{x} - \bar{x} = \sum_{i=1}^{M} c_i e_i
\]

where \( \hat{x} \) is the approximated OD demand constructed using the first \( M \) eigenvectors. The representation of \( \hat{x} - \bar{x} \) on the orthonormal basis \( e_1, e_2, \ldots, e_M \) is thus given by principal demand components \( c_1, c_2, \ldots, c_M \). Thus, we define a new set of variables, principal demand components \( c_i \) that capture the contribution of each eigenvector \( e_i \) to the particular observations of OD demand. In turn, the eigenvectors \( e_i \) capture the common behavior of travelers over the all OD pairs. These eigenvectors then define
the fixed structure of our OD matrices, which we then update on-line from traffic
counts.

3 State space formulation of the model

3.1 Kalman filter formulation of the model

The OD demand state in the network at time $k$ is uniquely described by the vector of
the principal demand components $c_i$ in $M$-dimensional space, where $M<n$. Using this
reduced formulation we can now construct a new reduced state space model that
which can be solved using the Kalman filter (Kalman 1960). The transition equation
is given by

$$c_k = \phi_{k-1} c_{k-1} + \omega_{k-1}$$

(4)

where $c_i$ is the state vector at time $k$; $\phi_i$ is the state transition matrix that accounts
autoregressive process on the components, which, however, is omitted here for
simplicity. This implies that in our case the transition equation (4) with $\phi_i = I$
represents the random walk model. The process noise vector $\omega_i$ is assumed white
Gaussian noise vector with a $m \times m$ variance covariance matrix $Q_i$ with eigenvalues
on the diagonal stored in decreasing order.

The measurement equation can be expressed as follows:

$$y_k = A_k x_k + \nu_k$$

(5)

where $y_k \in R^n$ denotes the vector of traffic counts for time interval $k$, and $A_k$
denotes the assignment matrix. Following the lower dimensional representation of OD
demand and substituting (3) in (5), we can formulate the new measurement equation
as:

$$y_k = A_k (c_k E_k + \bar{x}) = H_k c_k + \bar{y}_k + \nu_k$$

(6)

where $H_k$ is a $l \times m$ matrix called observation matrix, mapping the principal demand
components during interval $k$ to traffic counts observed during interval $k$. Note that
the observation matrix $H_k$ is not the same as the assignment matrix $A_k$ given in (10).
Finally, the matrix $H_k$ is used for the linearization of the model; it equals the
transform of the assignment matrix $A_k$ to the orthonormal basis matrix of eigenvectors
$E_k$. The measurement noise vector $\nu_i$ is assumed white Gaussian noise vector with
$l \times l$ variance covariance matrix $R_i$. The standard Kalman filter algorithm to solve the
previously defined state space model, which consists of a prediction and a correction
step is given in (Kalman 1960).

3.2 Colored noise Kalman filter

Reducing the state variables introduces additional uncertainty in the process, and this
noise increases as the reduced number of state variables increases. In order to explain
the potential reasons of the time correlation between measurements introduced by the
dimensionality reduction of the state vector, we analytically derive the measurement noise correlation.

The given measurement Eq. (5) for reduced number of state variables $m$ over time interval $k$ can be expressed as

$$ y_k = A_k \sum_{j=1}^{m} c_{j,k} e_{j,k} + A_k \sum_{m=1}^{n} c_{i,k} e_{i,k} + v_k $$

$$ y_k = A_k \sum_{j=1}^{m} c_{j,k} e_{j,k} + \xi_k $$

$$ \xi_k = A_k \sum_{m=1}^{n} c_{i,k} e_{i,k} + v_k $$

(7)

where, $\xi_k$ represent the measurement noise that consists of additional noise introduced by dropped state variables from $m+1$ till $n$ at time interval $k$.

Further, measurement Eq. (5) for reduced number of state variables $m$ for the next time interval $k+1$ can be expressed as

$$ y_{k+1} = A_{k+1} \sum_{j=1}^{m} c_{j,k+1} e_{j,k+1} + A_{k+1} \sum_{m=1}^{n} c_{i,k+1} e_{i,k+1} + v_{k+1} $$

$$ y_{k+1} = A_{k+1} \sum_{j=1}^{m} c_{j,k+1} e_{j,k+1} + \xi_{k+1} $$

$$ \xi_{k+1} = A_{k+1} \sum_{m=1}^{n} (c_{i,k} + \omega_k) e_{i,k+1} + v_{k+1} $$

$$ \xi_{k+1} = A_{k+1} \sum_{m=1}^{n} c_{i,k} e_{i,k+1} + A_{k+1} \sum_{m=1}^{n} \omega_k e_{i,k+1} + v_{k+1} $$

(8)

where, $\xi_{k+1}$ represent the measurement noise at time interval $k+1$ that consists of additional noise introduced by dropped state variables from $m+1$ till $n$ in previous time interval $k$. Therefore, $\xi_k$ and $\xi_{k+1}$ represent the temporal correlated measurement noise. It is well known that this condition destroys the assumption of independency between process and measurement noise that underlies the standard Kalman filter. The objective of this paper is to find an effective method to deal with this kind of correlation.

Now we can define a new system of measurement equations with temporally correlated measurement errors

$$ y_k = H_k c_k + \xi_k $$

$$ \xi_{k+1} = \psi \xi_k + v_k $$

(9) (10)

where correlation matrix $\psi$ is equivalent to the transition matrix $\phi_k$ for time correlated errors, and $v_k$ is a measurement noise vector assumed to be uncorrelated with the process noise vector $w_k$. 
Applying the time differencing approach on Eq. (9) and (10), which was first introduced in 1968 by Bryson and Henrikson (Bryson 1968), yields a new pseudo-measurement equation $z_k$ whose error is white in following form

$$z_k = y_k - \psi y_{k-1} = (H_k \phi_k - \psi_k H_k) c_{k-1} + H_k w_k + v_k$$

where $v_k = H_k w_k + u_k$ is zero mean white noise with covariance matrix $R_k$ and $H_k^* = H_k \phi_k - \psi_k H_k$.

Further, the decorrelation technique from (Bryson 1968) is applied on state transition function Eq.(7) to eliminate the correlation that now exists between the new measurement noise $v_k^*$ and the process noise $w_k$. A new transition equation can be written as

$$c_k = \phi_k^* c_{k-1} + w_{k-1} + J_{k-1} (z_{k-1} - H_k^* c_{k-1} - v_{k-1}^*)$$

where the new state transition matrix is expressed as $\phi_k^* = \phi_k - J_k^* H_k$ and $J_k^* z_{k-1}$ is the control item of the new system. The new process noise error $w_{k-1}^* = w_{k-1} - J_k^* v_{k-1}^*$ has zero mean with covariance matrix $Q_k^*$. The new process noise and measurement noise are uncorrelated with covariance matrix $E[w_k^* v_{k-1}^*] = M_k^*$.

For a more detailed derivation of colored noise Kalman filter, and derivation of covariance matrices $Q^*$, $R^*$ and $M^*$ we refer to (Bryson 2002). At this time, Eq. (11) and (12) satisfy the assumptions of standard Kalman filter, where the new process and measurements noise have a zero mean and they are independent of each other. A solution of such a system is summarized below:

Prediction step:

$$K_k = P_{k-1|k-1} H_k^* (H_k^* P_{k-1|k-1} H_k^* + R_k^*)^{-1}$$

$$c_{k-1|k} = c_{k-1|k-1} + K_k (z_k^* - H_k^* c_{k-1|k-1})$$

$$P_{k-1|k} = (I - K_k H_k^*) P_{k-1|k-1} (I - K_k H_k^*)^T + K_k R_k^* K_k^T$$

Update step:

$$c_{k|k} = \phi_k^* c_{k-1|k} + J_k z_k$$

$$P_{k|k} = \phi_k^* P_{k-1|k} \phi_k^T + Q_k^*$$

The given solution algorithm assumes that the a priori statistics $c_{0|k}$ and $P_{0|k}$ for time interval $k = 0$ are given. The result of the Kalman filter, the estimated a posterior state
vector $c_k$, can be used to estimate the OD demand by applying Eq. (6). In the next section, we will compare the performance of the standard Kalman filter and colored noise Kalman filter algorithm to solve the proposed OD estimation method.

4 Experiment: Synthetic data

Numerical experiments are presented to evaluate the performance of the proposed model and solution algorithms in the terms of degradation of number of principal demand components in state vector on academic network example and simulated data. Prior to method evaluation, we define a simplified synthetic network that consists of 5 nodes, 25 OD pairs with a single route between them and 16 corresponding links (Fig.1). This network was chosen because we could model and assume the availability of the “true” OD demand and assignment matrix to the analysts. The “true” assignment matrix is arbitrary derived assuming network is congested. The link flows resulting from the assignment of the true OD demand has been perturbed to obtain the traffic count data.

![Figure 1: The synthetic network](image)

4.1 Simulating the OD demand

A major problem with all model evaluations is obtaining meaningful evaluations of the algorithm results and performance, because the true sources of data are not available for comparison when working with real data. One solution is to use simulated OD demand data, where underlying sources and phenomena are known. To generate a simulated OD matrix dataset for this purpose requires us to define an arbitrary model for OD demand generation, which represents a common spatial and temporal behavior of travelers.

Here, we perform the Logit model in sequence in order to introduce the correlation in OD demand data. First, we defined the set of traveler’s decisions before making a trip, including decisions to make a trip or not, destination choice and departure time choice. Then, for each of these decisions we have defined the set of alternatives available to travelers. The activity and traveling intentions of traveler $n$ are presented in the Fig.2.
The main principle of this model is that a large number of simulations are performed for varying model inputs, reflecting the variability’s in the travelers behavior and consequently in OD demand based on Monte Carlo simulations. Subsequently, we generate 1000 observations, each being a realization of the 25-dimensional OD demand vector. Each generated OD demand vector without nonzero OD flows is stored in OD demand matrix $X$ where each row represents one observation of OD demand.

4.2 The model performance

To examine the effect of degradation of the number of principal demand components in state vector, we applied the PCA on the OD demand data matrix $X$. Once we perform the PCA, we obtain the set of eigenvectors $e_i$ for $i = 1, 2, \ldots, 20$ and eigenvalues $\lambda_i$ for $i = 1, 2, \ldots, 20$.

We have seen that we can use eigenvalues to explore the data reduction potential, for instance by considering the total (cumulative) percentage of total variation explained (e.g. 95%), Fig.3. We can observe that the 90% of the variance of the data is captured by first 5 eigenvectors.
We have performed the experiments such that in every experiment run we omit the one state variable (principal demand component) from the state vector. Since the principal demand components in the state vector are arranged in decreasing order of an eigenvalues, in every experiment run we remove the principal demand component that captures the lowest variance. In every experiment, we have performed the 1000 iteration runs.

The performance of the OD estimation method is evaluated by comparing the error covariance estimates of the standard Kalman filter and correlated noise Kalman filter algorithms from the obtained synthetic data in section IV.A. For a close up look, we compare the error covariance estimates of the first principal demand component that was included in all states.

![Figure 4: Distribution of estimated standard deviation by the standard Kalman filter over dimensionality reduction of state variables](image)

In Fig. 4, the average standard deviation over all variables in the state vector is shown for different numbers of leading principal demand components in the standard Kalman filter. For the standard deviation obtained by Kalman filter, the standard deviation stabilizes around 12 principal components in state vector. This feature is due to the fact that from this number of state variables system becomes observable. For small number of the variables in state vector, the proposed method is not very efficient and standard deviation obtained from empirical data is overestimated. We can observe that sharp elbow appears around state vector with 5 principal demand components, showing that omitting the principal demand components with highest captured variance in OD demand will lead to non-effective dimensionality reduction of the state vector. This observation is consistent with cumulative percentage of the total variation explained by the eigenvalues, given in Fig. 3.

The performance of the standard Kalman filter for standard deviation of principal demand component that captures the highest variance in data is given in Fig. 5. These plots demonstrate that under the experimental conditions, our proposed reduced OD estimation model solved by standard Kalman filter deteriorate in accuracy with the reduction of dimensionality. Typically, when we omit the state variables we introduce more correlated noise in measurements that is not captured by the initial assumption of the Kalman filter. This finding is consistent with the general assumptions of the standard Kalman filter in which measurements noise are independent, zero mean, Gaussian noise processes.
Figure 5: Distribution of standard deviation for the PC variable with highest variance over dimensionality degradation of state variables

The performance of the proposed reduced OD estimation model solved by colored noise Kalman filter algorithm is presented in Fig.6.

Figure 6: Distribution of estimated standard deviation by the colored noise Kalman filter over dimensionality reduction of state variables

We can observe that the correlation between measurements introduced by the dimensionality reduction of the state vector is very well captured leading to significant improvement in accuracy of the proposed model. Also, the standard deviation stabilizes around the 12th principal component due to achieved system observability. The improved identifiability of our lower dimensional state space model is very important for the convergence of the Kalman filter.

5 Conclusion

From the results presented here we can conclude that PCA can be used to linearly transform high dimensional OD matrices into the lower dimensional space without
significant loss of estimation accuracy. We have proposed a new OD estimation method that uses the eigenvectors and principal demand components as state variables instead of OD flows. These variables can be used to construct a state space model that can be solved with recursive solution approaches such as the Kalman filter. The proposed state space model, however, appears to be sensitive to the reduction of the dimensionality due to the induced temporal measurement correlation. We have explored and derived an analytical solution for the so-called colored noise Kalman filter algorithm that accounts for temporal correlated measurement noise to avoid this limitation. The presented results are still academic in nature, and must be interpreted as a proof of concept. More results in more realistic settings are part of current research to ascertain that the method performs well in practice.

Acknowledgment

This research is partly funded by the COST Action TU0903 – MULTITUDE project in collaboration with Delft University of Technology and KTH University.

References


Barcelo, J. e. a. (2010). "A Kalman-filter approach for dynamic OD estimation in corridors based on bluetooth and Wi-Fi data collection." 12th World Conference on Transportation Research WCTR.


