Scenario-based path travel time aggregation with Gaussian copula estimated through Lasso

Ke Wan ¹ and Alain Kornhauser²

¹ Department of Operations Research and Financial Engineering
Princeton University, Princeton, NJ 08544 USA
Email:kwan@princeton.edu, alaink@princeton.edu

November 13, 2009

Abstract
In this paper, we highlight the characteristics of floating car data, and extend our research on single link to pathes, To carry out the minimum risk routing decision, we need to estimate the distribution especially the tails of the path travel time. By the law of total probability, we propose to approximate the total path travel time distribution by the probability-weighted sum of a series of scenario-specific conditional path travel time distributions. Each of them is characterized by the sequence of entering time to the links of the path and the sequence is expressed as its corresponding (entering-time) positioning vector and window width vector.Any of such conditional distribution is the sum distribution of a sequence of link travel time distribution with specific dependence structure. The dependent structure is modeled by a lagged Gaussian copula while the marginal distributions are estimated by kernel method. The L1-constrain-minimization(Lasso) method is utilized to obtain an invertible covariance matrix of the Gaussian copula even for limited data.

Compare to the iterative type procedure, this approach is efficient when the number of scenarios to visit is limited and it resolve the "conditional distribution puzzle" in the iterative formulas

Keyword: Path Travel Time Distribution, Sum of Dependent Random Variables, Gaussian Copula,Lasso, Lagged Copula

1 Introduction
The problem we try to deal with in this paper is what are the turn-by-turn directions that will allow one to travel to one’s destination in the best way given that costs throughout the network are neither constant nor deterministic. This is especially true for daily commuters who ”know” ”all” of the various ways to get to and from work but unfortunately have no means of knowing or anticipating what today’s traffic conditions as they are distributed throughout the various alternatives will play out to a ”best” way to go. Elements of this ”real” stochastic optimization problem are the focus of this paper. What makes this ”real” problem difficult inculdes: First, the cost (say, travel time) on individual arcs are not scalars but instead time varying probability distributions. Second, the time varying probability distributions on different arcs are not independent as is often assumed. Third, the cumulative cost of any path through the network is characterized by a probability distribution, thus any comparison of paths requires the comparison of probability distributions, not simple scalar values.

These difficulties place severe burdens on the information systems required to generate the solution procedure’s data needs and the computational procedures required to accommodate the additional reality. To address these difficulties we propose the following settings:

1. The probability law of links and dependence between links are in non-normal case. There is a marginal distribution F_i for each link travel time. The links are dependent with certain joint structure, not necessarily joint normal.
2. Only link data is stored. The data includes the travel time observations and the corresponding time point to make them. So the data is stored in links and restricted to O(N) where N is the number of links.

3. For each traveler, we have finite number of paths to consider. That is the method focus on the routing decision with finite subset of paths in a indefinite large network.

4. Stochastic decision rules/risk measures based on the travel time distribution are used for decision making, such as mean-variance rule, stochastic dominance rules etc.

5. Parameter estimation can be conducted through Lasso method, which tackle the ill-conditioned estimation when data is limited compared to the dimension.

Based on the work in Wan(2009)[22], this paper will mainly focus on aggregating the travel time distribution/utility of links to get that of the paths and possible decision rule selection. And the paper is organized as follows: In Section 2, A literature review on related theory is given, including copula theory and estimation of covariance matrix. In Section 3, Our Framework and models are given: dependence structure estimation, and interpretations of decision rules are presented. In Section 4, the aggregation of the link travel time is designed and compared. In Section 5, Conclusion and future work is discussed.

## 2 Methodology Clarification and Related Theory

In this section, we review different literature and offer the theoretical base for a solution to the fundamental problem.

### 2.1 Path travel time estimation

Link travel time estimation has been a hot topic in transportation research, and numerous studies have focused on the accurate prediction of travel time of highways. Including: density estimator Petty(1998)[15], cross-validation estimator, Zhang(2003)[23], time series in Vemuri(1998) [20], local linear models in Dailey(2000)[5] and CTM, Daganzo(1994)[4], Delay function in Nie(2005)[13]. Most of research is about link travel time estimation and the research about the distribution estimation is limited. Wan(2008)[22] proposes a copula-based link travel time distribution estimator, it reconstruct the joint distribution of the travel time between different links, and yields proper estimation.

As the further improvement, the research to estimate path travel time has been developing for a long time, Chen(1968)[3], link-based model and path-based model are compared. And path based model is preferred. However there is not relation between link and path. The path based model is totally based on the observed total path travel time. For real ATIS system it is impossible to store the total path travel time for each path because the issue of combinatorial issue. so the further problem is how to aggregate link travel time distribution to get path travel time distribution.

The simplest assumption is to assume travel times on all the links along a path are generated by distributions that are statistically independent. And various modification is made to the results made by this assumption. Raha(2006)[16] uses coefficient of variation as $CV = \mu/\sigma$, to derive the estimates for the path travel time variance. And the assumption is statistically independence between arcs and path performance should be some basic statistics(mean/median/mode) of the link performance. Sherali(2006)[17] derived another formulation for estimating the path travel-time variance. They used the maximum and minimum segment travel-time CVs to construct bounds on the path CV given that the CV is independent of the length of the segment. In Fu(1998)[8], the assumption on the link travel time is that the travel times on individual links at a particular point in time are statistically independent.

The assumption is improved to the one-step dependence between arcs, in Waller(2002)[21]. In the paper Pattan(2003)[14], a Gaussian kernel is used to estimate the continuous mean travel time at a particular point in time t, a locally three-point polynomial approximation is used to estimate the mean link travel time as a function of time of day and a two factor model, in which stochastic travel time is caused by a systematic error and a vehicle error is used for error decomposition. when considering the case where travel times on two consecutive links are dynamic and stochastic but the joint probability is not estimated. Among the usual assumption on the dependent structure, joint normal distribution
is most usual one with respect to correlated arc travel time estimation. Dailey(2000)[5]. Gao(2006)[9] considers the simplified correlation model over time but not on different arcs and still does not go too far.

Meanwhile, the classical backward/forward dynamic programming can be used to estimate path travel time distribution, but there are two serious drawbacks: First, numerical error is large when we carried it out in continuous multidimensional integration. If using a discrete state enumeration, too many scenarios need to be enumerated and estimation error can be overwhelming with limited data. Second, The "Conditional distribution puzzle": The conditional distribution in the iteration formula is hard to estimate exactly. The assumption about the one step dependence of link travel time Waller[21] is just a usual way to simplify this conditional distribution structure and it brings great errors.

In all, to estimate dependent structure between stochastic link travel time is the essential problem in estimated path travel time and the corresponding models are improved tools to tackle the non-normal joint distribution is in need.

2.2 Dependence structure and Copula theory

Dependence structure is the dependent relationship between different random variables, mathematically, it is expressed in copula functions Nelsen(2006)[12], Cheruini(2004)[19]. The in previous section, is based on certain assumptions on dependence structure so it is a special case of the dependence structure theory in this section.

A copula-based approach allows a decomposition of a joint distribution into its marginal distributions and its copula. On the other hand marginal distributions may be combined to a joint distribution assuming a specific copula. The crucial point in using a copula-based approach is that it allows for a separate modeling of the marginal distributions (i.e. the univariate travel time distributions) and the dependence structure (the copula).

Sklar(1959) shows that an n-dimensional joint distribution function may be decomposed into its n marginal distributions and a copula, which completely describes the dependence between the n variables. The fundamental mathematical theorem here is as follows: Nelsen(2006) [12]

the conational copula we have the following theorem:

**Theorem 1. sklar’s theorem of conditional copula:**

sklar’s Theorem for continuous conditional distributions

Let $H$ be a conditional bivariate distribution function with continuous margins $F$ and $G$, and let $F$ be some conditioning set, Then there exists a unique conditional copula $C : [0,1] \times [0,1]$ such that

$$H(x, y|F) = C(F(x|F), g(y|F)|F), \forall x, y \in R$$

Conversely, if $C$ is a conditional copula and $F$ and $G$ are the conditional distribution functions of the two random variables $X$ and $Y$, then the function $H$ defined by the equation is a bivariate conditional distribution function with margins $F$ and $G$.

Different dependency measures such as person correlation, spearman’s $\rho$ and kendall’s $\tau$ can be unified under the copula theory, see Nelson(1995) [11] and Nelson Fredricks(2006) [6]. Copula theory enables us to go beyond the linear Pearson correlation coefficient and the joint Gaussian approximation, as for the practical traffic data, joint Gaussian is not always the case.

2.3 Covariance matrix estimation and the Lasso method

The problem when estimates the covariance matrix or the gaussian matrix $P$, is how to deal with the limited non-synchronized data with different amounts on different link pairs. The characterics of data are: the marginal link travel time data is abundant, and there is some data for pair wise correlation between links. Furthermore there is limited synchronized data for all the links under consideration. The data matrix $X^T X$ is nearly ill-conditioned. Considering these characteristics, we review different ways of covariane matrix estimation as follows:

The first way based on relatively sufficient pairwise data is to estimate the pair-wise correlation and modify the term into 0 if it is not significant. The drawback is that it may lead to non-positive semi-definite covariance matrixes. This brings problem when we proceed to estimate the joint distribution.

The more reliable way is to estimate the covariance matrix directly using the observed travel time for all the links in the path. And the empirical estimator of covariance matrix is
\[ \Sigma = \frac{1}{n}X^TX - \frac{1}{n}X^T1^TX \]

The estimator leads to positive definite matrix but it yields nearly ill-conditioned covariance matrix when the number of independent data is approximately equal to the dimension, and this can lead to large error in estimation.

As a further improvement, two types of advanced version is introduced, shrinkage estimators and penalized estimator.

1. The shrinkage estimator is in a linear combination of the covariance matrix targeting to a prior.

\[ \Sigma_e = \alpha \Sigma_0 + (1 - \alpha) \Sigma \]

where \( \Sigma_0 \) is a prior positive definite covariance matrix. This estimator will always give good positive definite covariance matrix. And its error is subject to the prior matrix selected and might be large when data limited.

2. Since the floating car data may be limited for many paths, and the dimensionality is higher than the available date in some cases, we select another way:

In Meinshausen(2006)[10], the problem of inverse covariance matrix estimates can be formulated as an optimization problem with the objective:

\[ \max_{\Omega : \Omega \succeq 0} \log \text{det} \Omega + \text{tr}(S \Omega) + \rho \| \Omega \|_1 \]

where \( S \) is the sample variance matrix and \( \rho \) is the weight. Its dual problem can be formulated into a \( L - 1 \) constrained optimization problem, called Lasso. It also shows that neighborhood selection with the Lasso is a computationally attractive alternative to standard covariance selection for sparse high-dimensional graphs.

In Friedman(2008)[7], lasso is used to estimate the inverse covariance matrix, graph discovering problem in a similar way to get the solution sparse.

For the lasso algorithm, it is first proposed in Tibshirani[18]. It minimizes sum of square s subject to the sum of the absolute value of the coefficient being less than a constant.

\[ (\beta, \alpha) = \arg \min \sum_{i=1}^{n} (y_i - \alpha \sum_j \beta_j x_{ij})^2 \]

s.t

\[ \sum_j |\beta_j| \leq t \]

And a solution method is given in an iterative procedure which starts from overall least square estimates and solve constrained least square each step. No research has applied lasso to do parameter estimation for gaussian copula, nor applied it in the traffic research area. Our research tries to tackle this problem.

3 Decision based on path travel time estimation

The problem under consideration is the stochastic routing decision problem. In our research, we try to reconstruct the joint distribution of arc travel time using appropriate copula models models and further address the dependence structure with time series model if data is sufficient, and use factor models when data is not sufficient. Finally, several decision rules are given to yield a best routing choice. The settings are given as below:
3.1 Problem modeling and structuring

Based on the review, the objective function and model is discussed in this section:

Denote a network as $N, L$ described by a finite set $N$ of nodes and a set $L = \subseteq N \times N$ of links between those nodes. A Path is a sequence of links $p_{ab} = l_1, \ldots, l_n \in L & l_1 \in L_a \& l_n \in L_b$. and the path set for given OD pair $P_{ab} = p_{allp_{ab}}$. Then the Minimum risk decision problem is:

$$\min_p U(p)$$  \hspace{1cm} (1)

s.t

$$U(p) = \int f(p)\mu(dp)$$

where

$f(p)$ is the value of the decision statistics when path travel time equals a certain value. We do not use the expectation $E(U(p))$ here as some of the statistics can not be represented in this form.

$\mu(dp)$ is probability distribution of the path travel time it is in the following form:

To estimate such path travel time, the most intuitive way is to study the sequence of link travel time experienced by any specific user. This way needs the trip-by-trip data, and it will need too much storage. The improved way is to estimate the path travel time distribution based on link travel time data.

The way we propose in this paper is to do approximation of the path travel time based on link data is to estimate the conditional path distribution for each lag positioning vector and use continuous distribution to approximate the conditional path travel time distribution. Each conditional distribution is obtained by conditioning on the event the entering time to the links is within a certain range. For example, take a path A-B-C, we consider the path travel time on the condition that the entering time to link A-B is 0 and link B-C is within the range $[30, 60]$ (seconds). The definitions and formulas are as follows:

Denote: $i : 1 \ldots I$ the sequential number of the decision point in a network. $t_i$ the entering time to the links. vector $(i, t_i)$.

$$P(T_{t_1} + T_{t_2} + \ldots + T_{t_I} < t) = \int P(T_{t_1} + T_{t_2} + \ldots + T_{t_I} < t|V)\lambda(dV)$$ \hspace{1cm} (2)

where

$V = (t_1, t_2, \ldots, t_I)$ is the lag positioning vector, denote the sequence of entering time to the links in the path.

$\lambda(dV)$ is the probability corresponding to a given lag positioning vector $V$

In continuous scheme:

$$P(T_{t_1} + T_{t_2} + \ldots + T_{t_I} < t|V) = 1_{(t_1 + t_2 + \ldots + t_I < t)}$$ \hspace{1cm} (3)

is a conditional path travel time given the entering time to the links is a certain lag vector $V$.

$1_{(t_1 + t_2 + \ldots + t_I < t)}$ is the indicator function takes value 1 if $t_1 + t_2 + \ldots + t_I < t$, 0 otherwise.

In discrete scheme:

$$P(T_{t_1} + T_{t_2} + \ldots + T_{t_I} < t|V) = \int 1_{(T_{t_1} + T_{t_2} + \ldots + T_{t_I} < t)}\mu(V, \Delta, x)$$ \hspace{1cm} (4)

$$\mu(V, \Delta, x) = P(T_{t_1} + T_{t_2} + \ldots + T_{t_I} < t|V \in P_{V, \Delta}) = C(F_{T_{t_1}}(x_1), F_{T_{t_2}}(x_2), \ldots F_{T_{t_I}}(x_I)|\Delta)$$ \hspace{1cm} (5)

is the cumulative probability function for the link travel time, conditioning on a given lag positioning vector $V$ and window width vector $\Delta$. 

5
\( V = (V_1, V_2, \ldots, V_j) \) is the lag positioning vector. For the previous A-B-C example, \( V = (0, 45) \) and \( \Delta = (0, 15) \).

\( \Delta = (\Delta_1, \Delta_2, \ldots, \Delta_J) \) is the discrete window length vector, the \( i \)-th element is for the window length of entering time on Link \( i \). For the previous A-B-C example, \( \Delta = (0, 15) \).

\[ (V, \Delta) = [t_1 - \Delta_1/2, t_1 + \Delta_1/2] \times [t_2 - \Delta_2/2, t_2 + \Delta_2/2] \times \ldots \times [t_l - \Delta_N/2, t_t + \Delta_t/2] \]

is the sub space characterized by \( V \) and \( \Delta \).

\( C \) is a suitable copula between the link travel time distributions \( F_i \) for the random travel time \( T_i \).

\( F_{T_{i}}(x_{i}) \) is the value at \( x_{i} \) of the cumulative distribution function of the conditional link travel time for \( T_{i} \) corresponding to a given subspace \((V, \Delta)\).

After we get the aggregation procedure for fixed lags, the ultimate estimation will be a linear combination of several basic conditional distributions. The weights is in proportion to the probability of the occurrence of the corresponding subspace \((V_i, \Delta_i)\), i.e:

\[ \sum_{n=1}^{N} \lambda_{n} P(T_{t_1} + T_{t_2} + \ldots + T_{t_i} < t|V_n, \Delta) \]

where \( \lambda_{n} = \frac{P(V=V_n)}{\sum_{j=1}^{N} P(V=V_j)} \)

To clarify the probability assumptions made here, discussion about the choice of \( \Delta_i \) and the lag positioning Vector \( V \) is necessary:

1. The fundamental assumptions is that the Consistency of marginal distribution (CMS) and Constancy of dependence structure (CDS) in Wan(2009)[22] will be assumed to hold for fixed time intervals \( \Delta_i \).

2. The lag positioning vector and \( \Delta \) should match the choice of the marginal distribution. We choose the marginal distribution for link \( i \) such that \( \mu_i = t_{i+1} - t_i \) and \( \sigma = \frac{\Delta}{m} m \in R^+ \)

3. The set of all lag positioning vectors in the discrete scheme will be a set such that:

\[ \Theta = \left\{ V : \Omega = \bigcap P_v \text{ and } P_{v_i} \cap P_{v_j} = \Phi \text{ if } i \neq j \right\} \]

where \( \Omega = \{(t_1, \ldots, t_N) : t_i \in R^+, \forall i\} \). Conditional distribution conditioning on \( V \) is for the pathes which lie in \( P_v \). By definition, when \( \delta \rightarrow 0 \Theta \rightarrow \Omega \).

4. We approximate the conditional path travel time if each link travel time \( T_i \) is experienced by entering Link \( i \) in the time interval \([t_i - \Delta_i/2, t_i + \Delta_i/2]\). \( \Delta_i \) here acts as a tradeoff between discrete scheme and continuous process.

When \( \Delta T \rightarrow 0 \) the conditional distribution converges to a specific sequence of link travel time along a trip, when \( \Delta_i \approx \sigma T_i \), this is to assume the probability law of travel time hold constant in the \( \Delta_i \) and the procedure capture all the data in the discrete scenario to estimate the underlying relation with a continuous structure. This mixed design enables a good estimation of travel time in limited enumeration. Due to the sparsity of floating car data, this \( Delta T \) discrete scheme is efficient to find the data in need.

An intuitive explanation for the definition above is the conditional distribution for a given lag positioning vector will be the distribution of the experienced travel time for a user in a scenario which satisfies the following conditions:

1. The traveler enters the \( i \)-th link at \( t_i \), and experienced travel time \( T_{t_i} \), the conditional expectation of which is \( t_{i+1} - t_i \).

2. \( T_{t_i} \)s are constant in their joint probability law in the interval \( \Delta_i \) for the time intervals \([t_i - \Delta T/2, t_i + \Delta T/2]\).
3. The dependent structure between all the \( T_i \), satisfying 1 and 2 is determined by the lag positioning vector \( V \) and window width \( \Delta \) and is constant as the starting time \( t_1 \) shifts forwards.

This design is the second layer of the three-layer structure of routing decision framework based on copula methods, first given in Wan(2009)[22].

Compare to the iterative scheme, this yields better approximation when the scenarios visited is small. This is a great advantage for realtime systems, and the more scenarios are analyzed the better the approximation will be, by the law of total probability. In the next sections, we will focus on the construction of conditional path distribution w.r.t the following specific fixed lag positioning vector and the lags in the vector equal just the overall expected travel time in each link. We define this as the main scenario for the path, and \( V \) is

\[
t_i = \sum_{j=1}^{i-1} ET_{t_j} \text{ and } t_1 = 0
\]

The corresponding cumulative distribution used is the overall cumulative distribution \( F^X \) and the window width \( \Delta_i \) are set as fixed constants, \( C \) times standard deviation of the current link travel time distribution. Notice here, the overall cumulative travel time distribution for Link \( i \) satisfies the condition that \( E(T_i) = t_{i+1} - t_i \). For link \( i \), more specific distributions which satisfies the same condition can be used as a surrogate, such as the conditional distribution on link \( i \) at time \( t \) whose mean is \( t_{i+1} - t_i \).

### 4 Estimation of conditional path travel time

Follow the definition of discussion of previous sections, the estimation of the path travel time includes three steps: First, data organization considering time lag along a path; Second, joint distribution estimation based on organized data; Third, Monte Carlo simulation for the sum distribution. They will be explained in detail in the following subsections.

#### 4.1 The lagged synchronization for path travel time aggregation

For different links and trips, the dependence structure definition accounts for the time-lag along the trip, as different geometric relationship brings different time lag. We therefore consider the dependence along a path as follows

**Definition 1.** The lagged correlation between two random travel time:

\[
\rho_{XY}^{t,s} = \frac{E(X(t)Y(t+s)) - E(X(t))E(Y(t+s))}{\sigma_X \sigma_Y^{t+s}}
\]

**Definition 2.** The lagged copula between the random travel time of two links.

\[
C_{t,t+s}^{X,Y}(x,y) = \{ C(a,b) : F_{t,s}^{X,Y}(x,y) = C(F_X(t), F_Y^{t+s}(y)) \}
\]

\[
C_{t_1,...,t_{n+1}}^{X_1,...,X_n} = \{ C(x_1,...,x_n) : F_{t_1,...,t_{n+1}}^{X_1,...,X_n}(t_1,...,t_n) = C(F_{t_1}^{X_1}(t_1),...,F_{t_n}^{X_n}(t_n)) \}
\]

Where \( s_i \) is the series of specified time instant for links in a consecutive path. We set \( s(n) = \sum_{i=1}^{n-1} E(T_{s_i}) \).

The definitions here is to reflex the time varying dependence of the two related travel time processes. The time difference is defined as the expected difference of starting time instant on each arc under consideration while traveling through them in a path. And the two arcs starts from the same nodes, the lagged structure reduce to normal dependence structure as the time lag is 0.

The definition above specifies the way to organize data. For paths, we select the second synchronization scheme as it is specified in previous sections. Then when doing estimation based on this lagged dependent structure, the observations obtained at the corresponding time intervals will form the observation vector \( X \). After the vector is formed, similar estimation procedure will be carried out for the whole vector \( X \).
4.2 Lasso estimation of the lagged gaussian copula

After we specify the type of the dependent structure model, we go for the parameter estimation. The goal of the procedure designed in this section is to give a method which can work for any path with different length and limited data.

The challenge here is that for a long path, the data in common with all the links might be limited. The methodology we proposed is to aggregate the link travel time based on lagged gaussian copula, with covariance matrix adjusted by the lasso method. We consider the whole covariance matrix $P$ together, and estimate the inverse of covariance matrix directly by maximizing an objective function penalizing by the L-1 norm of $P^{-1}$ by Lasso [7].

1. The gaussian copula parameter estimation based on the monotonic transformation is given by:

$$
\Sigma = \frac{1}{n}Z^T Z - \frac{1}{n}Z^T 1^T 1 Z
$$

where

$Z = \Psi^{-1}(F(X))$

$F = (F_1, F_2, \ldots, F_d)$ is the known marginal distribution

and

According to the definition of joint Gaussian distribution, it is Gaussian marginal distributions composed by a Gaussian copula. As we are considering the non-normal marginal distributions we take the monotone transform on data to do parameter estimation of gaussian copula. After the transformation, the data subject to a joint normal distribution with standard Gaussian margins.

2. Then the estimation of the covariance matrix for the Gaussian copula is conducted, the challenges here are,

a) $n \sim p$, $\Sigma$ cannot be considered a good estimate

b) $n \preceq p$, the empirical covariance $S$ is singular.

Using graphic Lasso method, we can generate a sparse, invertible estimate covariance matrix by solving the following problem:[2] and [1]:

$$
\max \log(\det X) - tr(\Sigma^T X) - v||X||_1
$$

(9)

where $|X|_1$ is the sum of the terms in $X$.

And the dual problem of this problem can be as follows:

As shown in, $L^1$ norm and $L^\infty$ are dual norm to each other

$|X|_1 = sup(X^T U ||U||_\infty \leq 1$

And then it can be decompose into solving sub problem the dual problems in the form of:

$$
\min y^T W_{j,j} y - S_{j} y + v||y||_1
$$

It is a L-1 constrained problem called lasso. The solver of lasso problem can be used to get the solution.

This procedure achieves the following goals: First, the $X$ is the inverse covariance matrix of the transferred data. And it is the parameter matrix for the gaussian copula. Second, The $X^{-1}$ is always positive definite while the initial sample covariance matrix $\Sigma$ is not necessarily when the number of common observations are smaller than the dimension. Third, the change of the parameter $v$ change the scarcity of the estimated covariance matrix. Then we can make trade off on the dependence between links far away from each other if data is limited for estimation. That is, there is little data for the pair and the links are far from each other, the covariance can be adjusted to near 0 while keeping the covariance matrix positive definite.
4.3 Aggregation of the link travel time distribution and Simulation

The lasso modification enables the design to be invariant to the number of observations. And it is a promising way in aggregating link travel time based on sparse floating car data. After the estimation of the dependent structure, the choice of marginal distribution to generate the joint distribution is of several possible designs:

Based the copula method we can simulate the normal random variables according to the following algorithm:

1. Diagonalize the estimated covariance matrix, as $\Sigma = B'\Lambda B$
2. for $j = 1$ to $m$ and repeat 4-6 for $m$ times, $m$ is selected to be optimal to restrict the sample mean.
3. Generate the n independent normal variables $X$ with $\Psi(0, 1)$ and define $i$ as the random index in the set $1...n$;
4. Generate correlated normal variables as $Y = BX$
5. Take monotone transfer, as $Z_i = F_i^{-1}(\Psi_i(Y_i))$, as $F_i$ are selected by the marginal distribution discussed above, if necessary, we delete those data which is less than the shortest possible travel time(calculated from the speed limit).
6. define $Y(\omega_j) = \sum Z_i(\omega_j)$ as a realization of the stochastic path travel time.

This procedure is an alternative to the multidimensional integration for this complex high dimensional joint distribution. The experiments are carried out as validation in the experiment session.

5 Numerical examples

In this section, we first illustrate the basic procedure by estimate the overall path travel time distribution by one conditional path travel time distribution which satisfies (8). Routing decision are then made based on estimated distribution. A further aggregation of different scenario is given.

For the experiment, we select one twelve-link path between Allen town and Clinton through Highway 78 to do the estimation of path travel time while select two competing pathes near philadelphia to illustration of decision making process, the network is shown in Figure 1.

[Insert Figure 1 here]

First, we change Lasso penalty to do estimation while fixing $\Delta = 3\sigma$ and marginal distributions as the overall marginal link travel time distributions(due to space limitation the test for these choices are omitted). The estimates based on different lasso penalties shrinkage between the totally independent assumption and the the empirical estimation, as shown in Figure 2.The more penalty used, the more the estimated covariance structure tend to be independent and the worse the approximation is. This shows that the dependence between links is heavy and independent assumption does not work. Generally speaking, little penalty yields better estimation, and we will set $v$ as the smallest value 0.001 for accuracy, in later experiments.

[Insert Figure 2 here]

Then, we conduct a weighted sum of limited different scenarios, the scenarios are picked by setting $\Delta = 3\sigma$ fixed and changing the lag vector $V$. We take two other scenario where the travel time of the tenth link is within $\mu_{10} + 1.5*\sigma_{10}, \mu_{10} + 4.5*\sigma_{10}$ and the later link travel time are as before. Then the lag vector is $V_2 = [t_1, \ldots, t_{10}, t_{11} + 3\sigma_{10}, t_{12} + 3\sigma_{10}]$ and here, $t_i$ as defined in (8). $\sigma_{10} = \max_i \sigma$,so the two scenario will not intersect with each other. The relative probability of the two scenario is approximately estimated as 4.5:1, by applying Chebyshev's inequality to the travel time distribution of Link 1, then the new estimation for the path travel time is $P(x|\Delta) = \frac{4.5}{5.5}P(x|V_1, \Delta) + \frac{1}{5.5}P(x|V_2, \Delta)$, and the estimates yields larger P-value in 1.We can see the estimation is better than
the one only base on the main scenario. Actually, the more discrete scenario one take ,the better the final estimation is. The continuous approximation within the discrete scheme is an efficient and flexible way for path travel time distribution estimation.

We calculate the distance of estimated distribution from the empirical path travel time distribution, as shown in Table 1

[ Insert Table 1 here]

Finally, to show the comparison of the two competing paths, we use the overall marginal distribution with lasso penalty $v = 0.001, \Delta = 3\sigma$, and compare the two paths by their major scenarios. Based on our methods, we estimate the path travel time distribution as in figure and the decision statistics is calculated in Table 2. The path ends at different sides of a bridge, we do not count for the turning for the bridge as the data is no available for the section on the bridge. The estimated path travel time distribution is shown in the Fig ??

[ Insert Figure 3 here]

Then the decision statistics introduced in [22] are calculated to make routing decision based on the estimated path travel time distribution, as in 2. Although the two paths are not comparable under mean-variance decision rules, most other decision rules will prefer Path 2. The traveler can then select one rule to make his own decision.

[ Insert Table 2 here]

For long pathes usually we do not have enough historical travel time observations, in this case there are just about 10 observations for either whole path while hundreds more observations for the links. The direct aggregation scheme based on the link data saves the cost of state enumeration and enables the decision based on path travel time distribution for realtime systems. Note we can not do any effective estimation with the dynamic programming approach shown in Section 3.1, as the conditional distribution between links is hard to estimated and the multivariate integration is hard to evaluate due to error from estimation and numerical procedures.

6 Conclusion and Future research

In the paper, we propose a continuous approximation for the conditional path travel time distribution within a discrete state space. For each scenario characterized by a lag vector $V$ and window width $\Delta$, a gaussian copula based estimator is proposed with Lasso method to estimate the copula parameter matrix. The whole estimation framework, the introduction of copula method to path travel time distribution and the introduction of Lasso for Gaussian copula estimation are the new development both in theory and applications.

With respect to the decision making aspect, non-normal dependent path travel time can be estimated based on floating car data and stochastic routing decision can be made according appropriate decision rules. Decision rules are shown and given interpretation in the transportation context, which make it possible to a better objective decision for the travelers.

In future research, The challenges mainly lie in:

1. $\Delta$ acts as the trade off the specific realization of path travel time (when $\Delta = 0$) and the conditional path travel time distribution in the scenario (when $\Delta > 0$). The larger $\Delta$ is, the more theoretical error there is, as we assume probability constancy in longer time while more data to do estimation so less data error. An optimal value for $\Delta$ is in need.

2. Other copula family or nonparametric method for copula can be studied. The gaussian copula is restricted by its specific parametric form and may not be able to fully describe the high dimensional dependent structure. Kernel method can be used.

3. More data is in need for the research. As the scheme includes a detailed division of link travel time data set based on the entering time and the short term mean (characterized by different scenarios). More dense data set for vehicle travel time is in need for further validation.
References


<table>
<thead>
<tr>
<th>Δ</th>
<th>2σ</th>
<th>3σ</th>
<th>4σ</th>
<th>6σ</th>
<th>Combine Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2</td>
<td>0.0115</td>
<td>0.0155</td>
<td>0.0111</td>
<td>0.0133</td>
<td>0.0154</td>
</tr>
<tr>
<td>P-value</td>
<td>1.43e-008</td>
<td>1.78e-007</td>
<td>6.16e-008</td>
<td>3.43e-008</td>
<td>1.2955e-007</td>
</tr>
<tr>
<td>Marginal</td>
<td>Overall</td>
<td>Realtime</td>
<td>Conditional 1</td>
<td>Conditional 2</td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td>0.0176</td>
<td>0.0320</td>
<td>0.0782</td>
<td>0.0275</td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>1.70e-008</td>
<td>3.51e-021</td>
<td>1.52e-055</td>
<td>1.36e-017</td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>v = 0.001</td>
<td>v = 0.01</td>
<td>v = 0.05</td>
<td>v = 0.1</td>
<td>v = 0.5</td>
</tr>
<tr>
<td>L2</td>
<td>0.0176</td>
<td>0.0131</td>
<td>0.0171</td>
<td>0.0189</td>
<td>0.0125</td>
</tr>
<tr>
<td>P-value</td>
<td>1.70e-008</td>
<td>3.80e-008</td>
<td>7.61e-009</td>
<td>3.08e-008</td>
<td>7.5282e-009</td>
</tr>
</tbody>
</table>

Table 1: Distance measures for different estimation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>(μ, σ)</th>
<th>μ + rσ</th>
<th>FSD</th>
<th>SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Path 1</td>
<td>0.2476</td>
<td>251.5</td>
<td>1832</td>
<td></td>
</tr>
<tr>
<td>Preference</td>
<td>Path 2</td>
<td>Path 2</td>
<td>Path 2</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Decision statistics for different rules
Figure 1: Experiment Network Graph

Figure 2: Change of estimation as the Lasso penalty changes. (Empirical:red dot; Independent: yellow; \(\lambda=0.001\): Cyan; \(\lambda=0.01\): red; \(\lambda=0.05\): Blue; \(\lambda=0.1\): Black)
Figure 3: 12 Path travel time based on the approximation Lasso penalty changes, red Path 1 and blue Path 2