基于贝叶斯网络推理的行驶时间分布估计

摘要

行驶时间估计是智能交通系统（ITS）的一个重要方面。在城市环境中，由于各种随机因素，行驶时间可能会出现很大的变化。因此，我们专注于估计行驶时间分布，**与大部分研究的平均预期行驶时间估计形成鲜明对比**。我们提出了从浮动汽车数据中推断行驶时间分布的算法，特别是从稀疏的GPS测量中推断行驶时间分布。该框架结合了**高斯耦合**和**网络推理**，以估计行驶时间的**边缘和联合分布**。我们对一个月的GPS轨迹进行了一系列广泛的数值实验。我们根据50个最常见的轨迹的Kullback–Leibler (KL)发散和海林格距离对所提出的模型进行了基准测试。与基线方法相比，结合高斯Coulas和贝叶斯推理的稀疏网络方法，KL发散减少4.9%，海林格距离减少2%。

1. 引言

智能交通系统（ITS）旨在减少交通拥堵，降低油耗，并改善交通的其他方面。行驶时间信息在不同的ITS应用中起着至关重要的作用，如随机路由[1]、拼车[2]和交通监控[3]。更好地估计行驶时间可以提高

此类系统的总体性能，为用户提供更准确的信息，包括更好的路线选择和交通选择[4]。ITS用户的另一个关键方面是行驶时间的变化，这会影响上述所有示例。然而，以前的大多数研究[5]–[9]试图估计预期的行驶时间。最近关于分布估计的研究[10]–[14]通常涉及源–目的地（OD）对或密集的GPS数据，而不是稀疏的GPS轨迹。

稀疏的GPS数据通常由车辆获取，因此，从这些数据中估计行驶时间分布是一个重要的主题。然而，关于这一主题的文献非常有限。

在本文中，我们提出了从稀疏GPS数据中推断行驶时间分布的新方法。我们的目标是开发一种能够从稀疏GPS数据中能够可靠地估计行驶时间分布的方法。这将允许利用更容易获得的数据为ITS应用程序提供好处。在我们的方法中，我们推断了**道路网中行驶时间的协方差结构**。因此，该方法能够可靠地学习任何源–目的对之间的分布。为了估计协方差结构，我们将**高斯耦合**与最近开发的**贝叶斯网络推理算法**[15]结合起来，与**图形套索**[16]相比，该算法产生了更准确的网络结构和精度矩阵。这种方法的主要优点是能够处理具有不同数量观察的变量。该方法不需要更昂贵的高分辨率数据，而且非常适合稀疏、低分辨率的GPS轨迹。我们对稀疏GPS记录的大型数据集进行了广泛的数值实验，以验证所提出的方法。有趣的是，在高分辨率数据被下采样为低分辨率数据后，拟议的方法也可以应用于高分辨率数据；当可用计算资源有限时，这种方法可能是可行的。

我们考虑了新加坡一家主流出租车公司提供的数据集。它载有2010年8月期间由15 000多辆出租车组成的车队抽样的GPS位置。数据集中总共包括约1200万次行驶。每个位置（经纬度）都附有时间戳和出租车的状态。每辆车以一定的采样间隔记录数据。**由于网络规模大（数千条链路）和采样间隔大（平均60秒），我们将数据集划分为1小时的间隔。对于每个间隔，我们从该间隔中随机选择的70%路径中推断行驶时间分布，然后通过计算Kullback–Leibler发散和海林格距离以及行驶时间的经验分布来评估推断的分布其余30%的路径**。我们提出的模型利用了稀疏网络的高斯耦合和贝叶斯推理（BISN,[15]）框架，在KL发散方面优于基线方法4.9%，在海林格距离方面优于基线方法2%。我们还通过在不同长度的时间间隔的数据上训练行驶时间模型，并将这些模型与以后时间间隔的经验行驶时间分布进行比较，来研究所提出模型的性能。

论文的其余部分组织如下。在第二节中，我们总结了关于估计行驶时间和行驶时间分布问题的文献。在第三节中，我们详细介绍了我们的方法。在第四节中，我们介绍并讨论了我们对新加坡出租车数据的实验结果。在第五节中，我们提出了结论性意见和今后研究的想法。

2. 相关工作

行驶时间估计是文献[5]–[9]、[17]–[23]中研究得很好的主题。以前的大多数研究都涉及行驶时间的估计和预测，但没有提供这些估计的置信区间[5]–[9]。最近，行驶时间分布估计受到了更多的关注[10]–[14]。估计行驶时间和行驶时间分布的问题在很大程度上取决于数据来源。这些来源包括感应回路探测器数据、探测车辆和交通摄像头的视频流。

最常见的来源之一是安装在道路上的环路探测器和类似设备的数据。根据探测器的类型，环路探测器可以记录不同的流量参数，如流量和速度。对于此类数据，有几种常见的行驶时间估计技术。某些模型植根于交通流理论[5], [6]，而其他技术则是数据驱动的，不需要理论交通模型。时间序列分析是一个数据驱动技术系列[7]–[9]，包括SVR [7]和人工神经网络(AN)[18],[19]。**卡尔曼滤波**是另一种基于时间序列的方法，也被应用于行驶时间估计；例如，戴利[20]应用卡尔曼滤波器根据环路探测器的占用率和体积数据估计行驶时间。

探测车辆是行驶时间数据的另一个可行来源。这些车辆配备了GPS系统，可以记录位置数据和时间戳（有时还有其他信息，如速度），并具有一定的采样间隔，通常从几秒钟到一分钟以上。

GPS数据还对数据的性质提出了一些不同的挑战，包括位置数据错误，以及信号在采样间隔和探测车辆数量方面的稀疏性。由于GPS设备的普及，关于此类数据的研究越来越多[10]–[13]、[23]–[27]，旨在克服上述挑战。高斯混合[28]、耦合[12]、[13]和回归方法[23]等统计模型通常应用于GPS数据，此外，马尔可夫链[24]等图形模型，[28]和贝叶斯网络[25]。

在下文中，我们简要回顾了关于行驶时间分布估计的研究。使用探测车辆数据进行研究的一个重要考虑因素是采样间隔，即车辆报告其位置的频率。文献中的采样间隔从大约一秒[11]到仅报告来源和目的地[26]不等。在表I中，我们总结了一些更相关的研究。

Ramezani和Geroliminis[24]提出了用马尔可夫链建模行驶时间分布。对于每对连续链路，他们都会构建一个行驶时间的2D图。接下来，他们将启发式网格聚类应用于此图，以计算这些特定链接的行驶时间状态。他们应用马尔可夫链来计算过渡概率和行驶时间分布。这种方法即使对动脉链路也能产生良好的分布估计，然而，它需要高分辨率的行驶时间数据。亨特等人。[25]提议将贝叶斯网络应用于仅使用始发–目的地对作为输入数据的行驶时间估计。他们应用期望最大化（EM）算法来处理路径不确定性。该方法显示出了有希望的结果，然而，它假定链接行驶时间的独立性。作为改进，亨特等人。[11]提出了一种不同的行驶时间分布估计方法。他们开发了一种以1Hz采样的密集GPS数据的方法，可以推断车辆在特定路段的停车次数。在该研究中，使用马尔可夫模型估计了行程时间分布，该模型与高斯马尔可夫随机场耦合。所提出的模型比基线模型表现更好，然而，该方法需要高分辨率数据。

从探测车辆数据中估计行驶时间分布的另一种有希望的方法是图形模型[12]–[14], [25]。Wan和Kornhauser [13]应用图形套索和高斯耦合来估计行驶时间。它们通过一系列条件路径行驶时间分布的总

和来近似路径行驶时间分布。这些分布取决于滞后向量（车辆进入路径中每个链路的时间向量）。在这个框架中，不同边缘处的行驶时间之间的依赖考虑了行驶期间的时间差。为了解决观测数量可能较低的问题，通过图形套索方法估计滞后高斯耦合参数，以获得稀疏精度矩阵。在[12]中，万介绍了用于路径行驶时间估计的高斯Coula混合模型(GCMM)。在这种方法中，针对不同的场景对行驶时间进行建模。为了计算给定路径的总行驶时间，它们集成了一组路径方案上的条件路径行驶时间分布。每个场景代表一定的交通状况，每个场景中不同链路的行驶时间之间的统计依赖性是固定的。通过图形套索再次估计每个高斯耦合的参数。然而，[12]中的数值结果相当有限，因为只考虑了三条路径，每条路径都少于15个链接。

在早期的工作[14]中，我们结合了高斯耦合和图形套索方法来推断新加坡网络部分地区的行驶时间分布。在本文中，我们使用更大的数据集进行了更深入的研究，包含更多的轨迹，跨越更大的时间段。我们还提出了一个贝叶斯框架，它超越了图形套索，能够处理大型网络中稀疏的GPS数据。稀疏GPS数据比高分辨率GPS数据更容易获得，也更常见。因此，拟议的框架可以更广泛地利用。为了解决图形套索方法[12]–[14]的缺点，我们提出了**应用一种新的稀疏网络贝叶斯推理(BISN)**[15]方法。图形套索方法没有考虑穿越不同链路的车辆数量的差异，而BISN允许我们准确地模拟一些链路比其他链路更频繁地穿越的通常情况下的行驶时间分布。同时，它具有较低的计算复杂度。我们研究了使用耦合器进行路径行驶时间建模的效果，以及估计性能如何取决于路径长度、一周中的一天和一天中的小时等各种因素。

上述先前的研究要么仅限于小型网络，要么限于时间框架[12]、[13]、[28]，需要难以获得高分辨率GPS数据[11]、[24]或两者兼而有之。此外，一些研究需要对行驶时间分布进行强有力的假设，如独立性、高斯性和其他。相比之下，我们所提出的方法能够基于稀疏数据估计大型交通网络中路径的行驶时间分布，而不假设链路行驶时间分布，同时也考虑到整个网络的可变链路覆盖。

3. 方法论

3.1 问题描述

接下来我们将概述我们的统计建模方法。我们的目标是从稀疏的GPS数据中估计交通网络中路径的行驶时间分布。我们从新加坡一家出租车公司获得了1个月（2011年8月）的出租车GPS数据。每个数据点包含地理坐标、出租车标识符、时间戳和当前状态。后者的可能值包括FREE（出租车正在寻找乘客）、ON CALL（出租车已收到预订，正在接客户的路上）、POB（车上有乘客）和不活动期。我们只对POB状态下的行驶感兴趣，因为它们通常代表真实的交通状况。在[17]中研究了这一特定数据集，其中表明

可以从出租车轨迹推断交通模式，只有700个（在超过15,000个中）获得70%的网络覆盖率。从这一系列GPS轨迹中，我们希望推断出网络中任何路径的行驶时间分布，间隔为一小时。

在下文中，我们将概述我们的行驶时间分布统计建模方法。**我们将与网络中路径相关的行驶时间建模为多元高斯或耦合高斯随机变量**。路径由多个链路组成，其行驶时间是单个链路行驶时间的总和，每个链路都建模为随机变量；在多变量耦合高斯模型中，单个链路上的行驶时间不一定是高斯分布的。为了确定路径行驶时间的分布，我们需要推断每个链路的行驶时间如何取决于其他链路的行驶时间。在高斯和Coula高斯模型中，这种统计关系被协方差矩阵 Σ 及其逆（精度矩阵 K ）完全捕获。精度矩阵以优雅的方式编码不同变量之间的依赖关系：如果两个变量 x_i 和 x_j 是条件独立的，则精度矩阵中对应的元素 $K_{i,j}$ 为零。

我们将网络建模为图 $G = (V, E)$ ，其中 V 是一组顶点，表示网络中的节点（如交叉口），E 是一组弧，表示道路连接（手头网络中的26972）。我们将**路径**定义为连续的道路连接序列，将**轨迹**定义为出租车穿过的路径。对于每个轨迹，我们知道出租车到达路径中第一个链路的时间戳和离开最后一个链路的时间戳。图1显示了我们提出的方法的总体概述。首先，**我们通过[29]中提出的基于HMM的算法，将原始GPS坐标数据转换为与图 G 匹配的路径**。

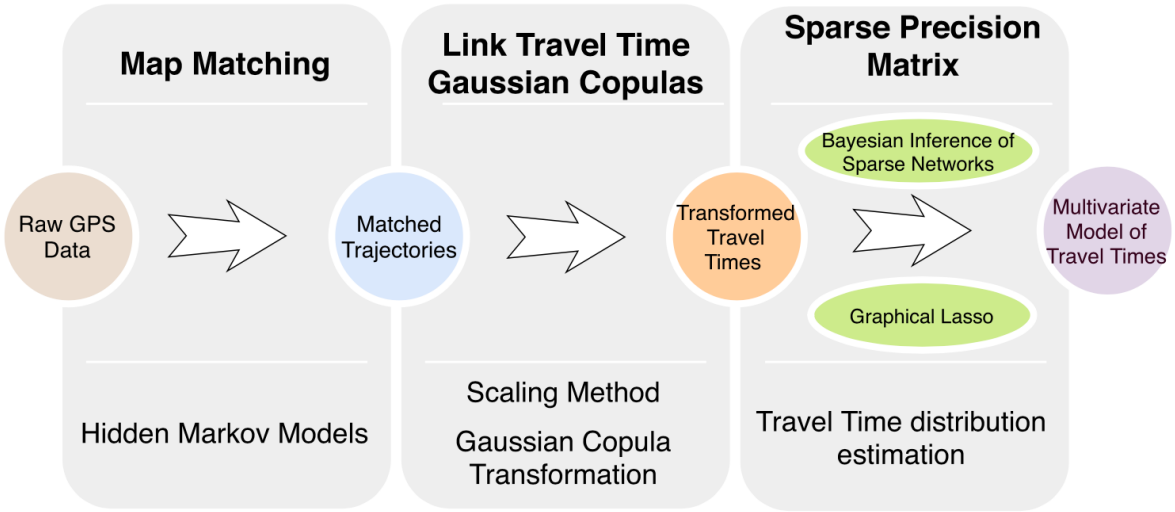

Fig. 1. Diagram of the proposed approach for travel time modeling.

接下来，根据这些投影路径以及报告的时间戳，我们计算每个单独链路的旅行时间。我们假设出租车在给定路径中的所有链路上都保持恒定的速度。然后，我们计算每个链接的旅行时间，与链接长度成比例。我们将把这种方法称为“缩放方法”（参见C节）。在处理稀疏GPS轨迹的研究中，通常考虑基于比例的方法[30], [31]。在手头的数据集中，所有道路连接都被大量不同的轨迹覆盖。*在某些情况下，在当*

前时间间隔内，给定路径只有一个链路子集拥塞。然而，当我们计算单个链接的旅行时间时，无论任何特定路径如何，我们都会这样做。这意味着每个拥塞链路的旅行时间将部分从仅覆盖拥塞链路的轨迹计算。由于数据量的原因，拥塞的链路将有更长的旅行时间，反之亦然。同时，所提出的分布估计方法（即将高斯耦合与贝叶斯网络推理相结合）可以与任何其他单个链路旅行时间计算方法一起应用。例如，可以使用基于张量的方法[10]。

为了模拟这些旅行时间估计的非高斯分布，我们应用高斯耦合，将非高斯旅行时间估计转换为高斯随机变量（参见F节）。

在这个转换之后，我们计算了**部分经验协方差矩阵（PECM；参见E节）**。它作为推断运输网络不同环节旅行时间协方差结构的基线方法。接下来，我们将图形套索算法[16]（图形套索；参见G节）应用于PECM，以便对精度矩阵施加稀疏性，产生旅行时间的协方差矩阵的更准确估计。除了图形套索方法，我们还应用了图形套索的替代方案，名为**“稀疏网络的贝叶斯推理”（BISN；参见H节）**，因为它已经被证明可以从数据中更准确地估计稀疏网络[15]。因此，这种方法能够考虑到数据点的数量（旅行时间）通常在链路上变化这一事实。相比之下，在图形套索方法中，人们隐式地假设每个链接具有相同数量的数据点（旅行时间）。事实上，图形套索直接应用于PECM，并且没有提供PECM元素的置信水平；这些置信水平取决于与每对链接关联的数据点的数量。BISN的另一个重要优势是其较低的计算复杂性；BISN的计算复杂性仅随变量的数量（运输网络中的链接）二次缩放，而图形套索具有立方计算复杂性。

根据缩放法获得的转换旅行时间，我们计算了平均旅行时间及其经验协方差矩阵；我们将图形套索和BISN应用于后者，以便获得协方差矩阵的更可靠估计。因此，我们得到了由旅行时间的平均向量和协方差矩阵指定的旅行时间的多元耦合模型；后者是通过图形套索或BISN获得的，我们将在本文中评估这两种方法。路径旅行时间估计值是通过从路径中链路的联合分布中采样计算的。该方法在H节和I节中描述。为了进行评估，我们选择了网络中最常见的50个轨迹，并将获得的分布与经验分布进行比较。具体来说，我们计算了两种不同的度量，即Kullback–Leibler (KL)发散和旅行时间经验分布与从不同旅行时间模型获得的边缘分布之间的海林格距离。

在下文中，我们将解释旅行时间建模管道中的 每个步骤。

3.2 数据预处理

来自探测车辆的数据往往有噪音；位置信息通常只有5米到10米的精度。因此，预处理数据并将GPS点投影到道路网上是很重要的。这个过程被称为地图匹配（MM），并为此目的提出了几种方法，包括几何、拓扑、概率和基于人工智能的方法[32]。由于本文所考虑的GPS数据是稀疏的，我们应用了一种基于隐马尔可夫模型的地图匹配算法，该算法特别适合这种稀疏数据[29]。对于该算法，每个可能的路段

都被表示为隐藏状态，其观测概率 p （观测）取决于轨迹点和路段之间的距离 d 、路段宽度 2ω 和GPS误差 δ_g 的标准偏差：

$$p(observation) = \frac{1}{2\omega}\int_{-\omega}^{\omega}\frac{1}{2\pi\delta_g^2}e^{-\frac{(1-d)^2}{2\delta_g^2}}dl$$

发射概率 $p(observation|r)$ 定义为：

$$p(observation|r) = \frac{v_{max}}{max(0, v_{obs} - v_{max}) + v_{max}}p(observation)$$

其中 r 是候选段， v_{obs} 是观察到的速度， v_{max} 是速度限制。(2)中的比率表示超速罚款系数。

Goh等人。[29]应用支持向量机计算状态之间的转换概率。它们通过改进的维特比算法确定最可能的路段序列。有关更多详细信息，我们请读者参考[29]。

3.3 基于标度法的链路旅行时间估计

在获得地图匹配数据后，我们计算每个链路的旅行时间。实现这一目标的最直接的方法之一是**假设出租车车辆在整个多链路路径上具有恒定的平均速度**。我们通过重新缩放路径旅行时间 T_p ，计算长度为 l_i 的链路 i 的旅行时间 t_i ，如下所示：

$$t_i = \frac{l_i T_P}{\sum_{j\in P} l_j}$$

我们在图2中提供了这种计算的说明。在计算所有轨迹（来自不同出租车）的这些估计后，我们对每个链路的平均值进行平均值，以计算平均旅行时间。

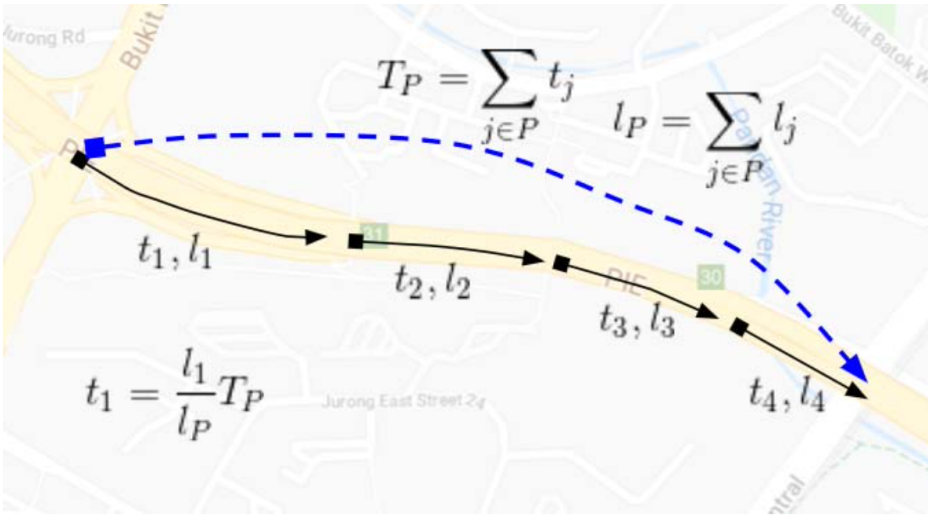

Fig. 2. Illustration of the scaling method (3) for computing link travel times from path travel times. In this example, the path (blue dashed line) contains four links (black arrows).

为了提高这些中间旅行时间估计的准确性，我们纳入了有关路段上轨迹点位置的信息。设第一段 $j_1 \in P$ 上的映射轨迹点为 A ，最后一段 $j_K \in P$ 上的点为 B ，其中 K 是沿路径 P 的GPS点总数。设 $\hat{l_1}$ 是 A 到 j_1 最后一点之间的距离， $\hat{l_K}$ 是 B 到 j_K 最后一点之间的距离。我们计算旅行时间估计 t_i 如下：

$$t_i = \frac{l_i T_P}{\hat{l_1} + \hat{l_K} + \sum_{j=2}^{K-1} l_j}$$

通过更仔细地考虑第一个和最后一个GPS点沿着第一个和最后一个链路的位置，估计(4)与（3）相比变得更准确。

3.4 高斯耦合

旅行时间分布的一个明显选择可能是高斯分布或对数正态分布。然而，如图3所示，对于最常见的轨迹，两种分布对经验行程时间数据的拟合度都很差。取而代之的是，我们应用了高斯连接函数，使数据能够适应更灵活的分布。具有p维的高斯copula定义为：

$$C(t_1, ..., t_p) = \Phi_p(\Phi^{-1}(t_1), ..., \Phi^{-1}(t_p); \Sigma)$$

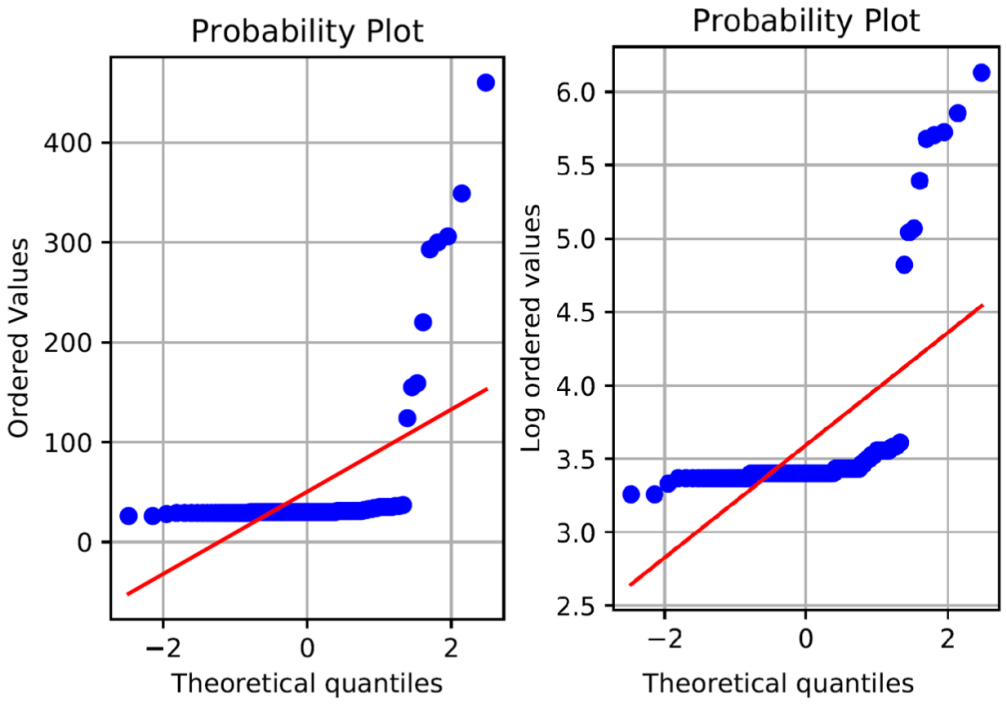

Fig. 3. QQ (quantile-quantile) plots of travel times and log travel times for the most traversed trajectory during one time interval.

其中 Φ 是标准正态分布的CDF， $\Phi_p(*;\Sigma)$ 是均值和协方差矩阵为 Σ 的p维多元高斯分布的CDF。我们通过概率积分变换将通过缩放方法（4）获得的链路 i 处的观测行程时间 t_i 转换为高斯分布值：

$$\hat{t_i} = \Psi^{-1}(F_i(t_i))$$

其中 F_i 是在链路 i 处的旅行时间的分布函数， Ψ 是反误差函数。分布 F_i 是从有限个样本中推断出来的。在实际应用中，我们通过先计算样本的累积直方图，然后对累积直方图的值进行线性插值来计算 F_i 。

3.5 经验协方差矩阵

在所提出的旅行时间建模流程中，我们估计了运输网络中链路处旅行时间的协方差。为此，我们计算经验协方差矩阵 S 。具体地说，我们计算部分经验协方差矩阵(PECM)[11]是通过分别计算链路的每对 (i,j) 的协方差来实现的：

$$\hat{S}_{i,j} = \beta_{ij} < t_i, t_j > - < t_i >< t_j >$$

其中， $< t_i >$ 是在包含链路 i 的所有路径 P 上计算的平均旅行时间（通过缩放方法（3）获得），而在包含链路 i 和 j 的所有路径 P 上计算的平均 $< t_i t_j >$ 。为了避免矩阵是负半定的，我们引入了标度系数 $\beta_{i,j}$ ，以便所有元素 $\hat{S}_{i,j}$ 具有与Hunter等人[11]提出的相同的方差：

$$\beta_{i,j} = \sqrt{\frac{< t_i^2 >< t_j^2 >}{< t_i^2 >_{i,j} < t_j^2 >_{i,j}}}$$

其中，在同时包含链路 i 和链路 j 的路径上计算的是参数 $<>_{i,j}$ 。如果包含这两个环节的轨迹少于5条，我们将 $\hat{S}_{i,j}$ 设为零。实际上，如果两个链接很少出现在同一路径上，那么沿着这些链接的旅行时间很可能是一个很好的近似独立的。

3.6 图形套索

在稀疏GPS数据的设置中，数据量有限，网络大。因此，很难可靠地推断出全精度矩阵。对于这样的场景，通常的做法是假设精度矩阵是稀疏的，并应用图形套索方法[16]:

$$\hat{K} = arg \min_{K} (TrSK - log\ det\ K + \alpha ||K||_1)$$

这个过程施加了一个 L_1 惩罚，结果是一个稀疏的精度矩阵 $\hat{K}$ 。估计值 $\hat{K}$ 的逆值是可靠估计。我们应用坐标下降实现[16]来解决优化问题(9)。为了估计精度矩阵，将该算法应用于前一节（7）中介绍的PECM。

3.7 高维稀疏网络的贝叶斯推理

到目前为止，该模型没有考虑到这样一个事实，即有些路段比其他路段更经常被出租车穿过。由于估计的不确定性不同，这影响了对各环节之间协方差的估计。例如，如果一个链路对有几个同时观测量，而另一个链路对有几倍多的观测量，在推断稀疏精度矩阵时，这些协方差估计应该被区分对待。为了解决这个问题，我们采用了高维稀疏网络框架(BISN)[15]的贝叶斯推理。BISN不是像图形套索那样建模精度矩阵，而是处理精度矩阵的LDU（下–上）分解。在这种情况下，具有精确矩阵 $K = LDL^T$ 的高斯模型的概率矩阵可以写成：

$$p(t|L,D) = \prod_{j=1}^{p} D_{j,j}\ exp(-\frac{1}{2}t^T LDL^T t)$$

其中 L 是下三角矩阵， D 是对角矩阵。

BISN在获得稀疏精度矩阵 K 之前施加一个尖峰和板条。LDU分解中元素的后验分布通过变分贝叶斯方法近似计算。BISN为 K 中的每个元素引入一个系数 γ ，并将 γ 视为随机变量。这使得BISN能够考虑到不同的变量可以有不同数量的观察这一事实。与图形套索相比，对应的 α 参数是矩阵的固定常数乘数。结果，BISN得到了 K 的稀疏模式的更可靠的估计，以及 K 中非零元素的更好的估计，最终导致协方差矩阵的更可靠的估计。与图形套索的三次复杂度相比，该算法仅具有二次计算复杂度；计算复杂度的大幅降低对于实时应用非常重要。采用基于矩阵随机化的随机逼近方法，降低了BISN的计算复杂度。

虽然图形套索应用于PECM，但BISN模型直接从数据（不同出租车行驶的每个路段的行程时间）推断稀疏精度矩阵。为了应用BISN，我们构造了一个矩阵 $X \in \mathbb{R}^{n \times p}$ ，其中， n 是观测次数（即滑行轨迹）， p 是链路的托亚数，每个元素 $X_{i,j}$ 是链路 j 的观测行程时间。如果一个特定的链路 j 不是轨迹 i 的一部分，则缺少相应的值 $X_{i,j}$ 。

为了控制 X 中的缺失数据量，我们假设不同的出租车在短时间内可以被看作一辆车。在此假设下，我们可以折叠X的行，以减少丢失值的比例。如果这些行中所有观察到的轨迹的开始时间都在两分钟的间隔内，并且没有一个链接出现在一个以上的轨迹中，我们折叠第 $i_1 ... i_t$ 行。因此每列在行 $i_1 ... i_t$ 上最多有一个不丢失的元素。

我们以两种不同的方式应用BISN。在第一种情况下，我们同时为50个轨迹构造矩阵 X 。这样，列数p等于跨越50个轨迹的唯一链接数。该算法将只对整个网络进行一次评估。我们把这种方法称为“BISN网络模型”。

在第二种方法中，我们将称为“BISN路径模型”，我们分别为每个路径 P 构造一个矩阵 X^P 。通过选择与出现在 P 中的链接相关联的列来获得矩阵 X^P 作为 X 的子矩阵。

3.8 提议的模型

接下来，我们描述了所提出的整体模型。我们用标度法（4）计算出网络中每个链路的一组行程时间值。然后，我们计算经验协方差矩阵(PECM)，并应用图形套索算法。作为替代，我们将BISN应用于数据矩阵 X （网络模型）和 X^P （路径模型）。图形套索和BISN方法都可以得到稀疏精度矩阵。高斯

模型中的协方差矩阵被构造为稀疏精度矩阵的逆。对于高斯耦合模型，我们将同样的计算应用于积分变换的旅行时间（参见(9)），得到协方差矩阵 $\hat{\Sigma}^C$ 。提出了高斯模型作为比较，然而，主要提出的模型是高斯耦合模型。我们得到了旅行时间的多元高斯模型 TT_P 和高斯耦合模型 TT_P^C ：

$$TT_p \sim \mathcal{N}([t_1 ... t_{|P|}], \hat{\Sigma})$$
$$TT_P^{\mathcal{C}} \sim \mathcal{CN}([\hat{t}_1 ... \hat{t}_{|P|}], \hat{\Sigma^{\mathcal{C}}})$$

其中， t_i 是由方程（4）得到的链路行程时间， $\hat{t}_i$ 是用积分变换（8）变换的行程时间，是用图形套索或BISN得到的协方差矩阵。关于非高斯变量和高斯连接函数变换的进一步阅读，请参见[33]。

根据这些模型，我们生成路径旅行时间的样本，如下所示。首先，我们从链路行程时间分布中提取样本（3）。具体地说，我们从与包含在路径中的链接相关联的（多变量）边际分布中取样。换句话说，我们不是从与每个环节相关的单个边际分布中取样，这是不合适的，而是从路径中所有环节的联合行程时间分布中取样，这是模型（3）的边际分布。

对于高斯模型，我们可以通过一个标准过程生成这样的样本：首先对独立标准正态变量的向量 z 进行采样，然后对 Σ 进行谱分解： $\Sigma = U \Lambda U^T$ 。然后，所需样品x可计算为 $x = \mu + Az$ ，其中 $A = U\Lambda^{\frac{1}{2}}$ ， μ 是平均向量。为了从高斯耦合模型中采样，我们对高斯样本应用逆变换。最后，通过对链路行程时间的样本求和，得到路径行程时间的样本，得到多元链路行程时间分布的每个样本的路径行程时间样本。

3.9 评价

我们通过将旅行时间模型与经验路径旅行时间分布进行比较来评估旅行时间模型。为了评估模型与经验分布之间的偏差，我们计算Kullback Leibler散度：

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) log \frac{p(x)}{q(x)} dx$$

其中 Q 是从模型中产生的分布， P 是经验分布。我们可以解析地描述 Q 。对于轨迹 P ，设 $t_1 ... t_{|p|}$ 为轨迹中各环节的平均旅行时间， $\hat{\Sigma} = \{\hat{\sigma}_{i,j}\}$ 为估计的协方差矩阵。然后，对于高斯模型，轨迹 P 的行程时间分布可以描述为：

$$TT_P \sim \mathbb{N}(\sum_{i=1}^{|P|} t_i, \sum_{i=1}^{|P|} \sum_{j=1}^{|P|} \hat{\sigma}_{i,j})$$

然而，对于copula模型和经验分布，还没有分析形式。因此，我们考虑离散分布的Kullback Leibler散度公式：

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$$

其中 $P = \sum_i P(i)$ ， $Q = \sum_i Q(i)$ 是两个离散的概率分布。为了应用这个公式，我们首先通过计算直方图来离散这两个分布。由于公式中存在 $\frac{P(i)}{Q(i)}$ ，离散 KL 散度对离散化很敏感。我们用直方图的方法计算KL散度，并对不同数量的直方图箱进行了测试。我们对每个直方图使用固定数量的均匀分类，当（14）未定义时将它们折叠。为了确定我们实验的箱数，我们使用了两个标准估计量之间的一个最大值，即Sturgs方法和Freedman–Diaconis规则。我们将默认箱数设置为11，因为这是估计器在不同输入中计算的最常见箱数。尽管选择了适当数量的箱，但由于（14）中的比率，KL发散度的数值评估可能会导致不稳定性。因此，我们还考虑了一个替代方案，即Hellinger距离：

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_i (\sqrt{P(i)} - \sqrt{Q(i)})^2}$$

它不需要计算概率比，从而避免了上述的KL发散问题。离散分布的Hellinger距离类似于向量平方根之间的均方根差。