Outline

- Introduction
- Study Region
- Link Travel Time Estimation Model
  - Base Model
  - Probabilistic Model
- Numerical Results
- Conclusion
- Questions/Comments
Introduction

New York City has the largest market for taxis in North America:
- **12,779** yellow medallion (2006)
- Industrial revenue **$1.82 billion** (2005)
- Serving **240 million** passengers per year
- **71%** of all Manhattan residents’ trips

- GPS devices are installed in each taxicab
- Taxi data recorded by New York Taxi and Limousine Commission (NYTLC)

- **Massive amount of data!**
  - **450,000** to **550,000** daily trip records
  - More than **180 million** taxi trips a year
  - Providing a lot of opportunities!
Introduction

- Taxi trips in NYC

Trip Origin

Trip Destination
Estimating urban link travel times

• Traditional approaches:
  - Loop detector data
  - Automatic Vehicle Identification tags
  - Video camera data
  - Remote microwave traffic sensors

• Why taxicab data?
  - Novel large-scale data sources
  - Ideal probes monitoring traffic condition
  - Large coverage
  - Do not need fixed sensors
  - Cheap!
The data

- NYTLC records taxi GPS trajectory data, but not public
- Only trip basis data available
  - Contains only OD coordinate, trip travel time and distance, etc.
  - Path information not available
  - Large-scale data with partial information

The problem

- Given large-scale taxi OD trip data, estimate urban link travel times
- Sub-problems to solve:
  - Map data to the network
  - Path inference
  - Estimate link travel time based on OD data
Study Region

• 1370×1600m rectangle area in Midtown Manhattan
• Data records fall within the region are subtracted
Study Region

- **Test network**
  - Network contains:
    - 193 nodes
    - 381 directed links
Number of observations in the study region

- Day 1: Weekday (2010/03/15, Monday)
- Day 2: Weekend (2010/03/20, Saturday)
**Base Model**

- **Base link travel time estimation model**
  - Hourly average link travel time estimations
  - Direct optimization approach
  - Overall framework: four phases

---

**Data Mapping**
- Map origin-destination records on corresponding links

**Construct Reasonable Path Set**
- Shrink algorithm search space into a manageable scale

**Route Choice Model**
- MNL model based route choice considering trip time, distance

**Link Travel Time Estimation**
- Estimate link travel times by solving an optimization problem

---

Data mapping

- Mapping points to nearest links in the network
- Mapped point (blue) are used
- Identify intermediate origin/destination nodes
- $\alpha_1$, $\alpha_2$ are defined as distance proportions from mapped points to the intermediate origin/destination node
Base Model

- **Construct reasonable path sets**
  - Number of possible paths could be huge!
  - Need to shrink the size of possible path set
  - Use trip distance to eliminate unreasonable paths
  - K-shortest path algorithm* (k=20) is used to generate initial path sets
  - Filter out unreasonable paths (threshold: weekday 15%~25%, weekend 50%)

Base Model

- **Route choice model**
  
  - Assumption:
    - Each driver wants to minimize both trip time and distance to make more trips thus make more revenue
  
  - A MNL model based on utility maximization scheme
    \[
    P_m(\hat{t},d,\theta) = \frac{e^{-\theta c_m(\hat{t},d_m)}}{\sum_{j \in R_i} e^{-\theta c_j(\hat{t},d_j)}}
    \]
  
  - Path cost measured as a function of trip travel time and distance
    \[
    c_m(\hat{t},d_m) = \beta_1 \cdot g_m(\hat{t}) + \beta_2 \cdot d_m
    \]
    \[
    g_m(\hat{t}) = \alpha_1 t_O + \alpha_2 t_D + \sum_{l \in L} \delta_{ml} t_l
    \]
Link travel time estimation

- Minimizing the squared difference between expected ($E(Y_i|R_i)$) and observed ($Y_i$) path travel times

$$E(Y_i|R_i) = \sum_{m \in R_i} g_m(\hat{t}) P_m(\hat{t}, d, \theta)$$

$$\hat{t} = \arg \min_{\tilde{t}} \sum_{i \in D} (y_i - E(Y_i|R_i))^2$$

- Solve using Levenberg-Marquardt (LM) method
- Parallelized codes developed to estimate the model
- Entire optimization solved within 10 minutes on an intel i7 laptop
- Numerical results show in later section
Limitations of the base model

- Point estimate of hourly average travel time
- Not incorporating variability of link travel times
- Not utilizing historical data
- Problems of compensation effect
- Less robust

Solution: Adopt a probabilistic framework

- Accounting for variability in link travel times
- More robust
- Historical information can be incorporated as priors
Probabilistic Model

**Assumptions:**

1. Link travel time: $x_l \sim \mathcal{N}(\mu_l, \sigma_l^2)$

2. Path travel time is the summation of a set of link travel times

$$P(y_i|k, x) = P(y_i|k, \mu, \Sigma) = N \left( \alpha_1 \mu_0 + \alpha_2 \mu_D + \sum_{l \in k} \mu_l, (\alpha_1 \sigma_0)^2 + (\alpha_2 \sigma_D)^2 + \sum_{l \in k} \sigma_l^2 \right)$$

3. Route choice based on the perceived mean link travel times and distance

$$\pi^i_k(\mu, \beta, d_i) = \frac{\exp[-C^i_k(\mu, \beta, d_i)]}{\sum_{s \in R_i} \exp[-C^i_s(\mu, \beta, d_i)]}$$

- where $x, \mu, \Sigma$ are the vector of link travel times, their mean and variance
**Probabilistic Model**

- **Mixture model**
  - A Mixture model is developed to model the posterior probability of the observed taxi trip travel times given link travel time parameters $\mu, \Sigma$

$$H(y|\mu, \Sigma, D) = \prod_{i=1}^{n} \sum_{k \in \mathcal{R}^i} \pi_k^i(\mu, \beta, d_i)P(y_i|k, \mu, \Sigma)$$

- Introducing $z_k^i$ as the latent variable indicating if path $k$ is used by observation $i$

---

Plate notation

---

Urban Link Travel Time Estimation Using Large-scale Taxi Data with Partial Information
Bayesian Mixture model

- Incorporating historical information:
  - Prior on $\mu$:
  
  $H(y|\mu, \Sigma, D) = \prod_{i=1}^{n} \sum_{k \in R^i} \pi^i_k(\mu, \beta, d_i)P(y_i|k, \mu, \Sigma) \cdot \prod_{j \in L} p(\mu_j)$

  - Priors on $\mu$ and variance $\Sigma$
  
  $H(y|\mu, \Sigma, D) = \prod_{i=1}^{n} \sum_{k \in R^i} \pi^i_k(\mu, \beta, d_i)P(y_i|k, \mu, \Sigma) \cdot \prod_{j \in L} p(\mu_j)p(\sigma_j^2)$
Solution approach

- An EM algorithm is proposed for estimation
- A iterative procedure of two steps:
  - E-step:
    \[
    \mathbb{E}(z_k^i) = \frac{\sum z_k^i \left[ \pi_k^i (\mu, \beta, d_i) P(y_i | k, \mu, \Sigma) \right] z_k^i}{\sum z_k^i \sum_{s \in R^i} \left[ \pi_s^i (\mu, \beta, d_i) P(y_i | s, \mu, \Sigma) \right] z_s^i} = \gamma(z_k^i)
    \]
  - M-step: Let $\tau = \sigma^2$, $\Sigma = \Sigma$, 
    \[
    Q(\mu, \tau) = \mathbb{E}_z [\ln P(y, z | \mu, \tau)] = \sum_{i=1}^{n} \sum_{k \in R^i} \gamma(z_k^i) [\ln \pi_k^i (\mu, \beta, d_i) + \ln P(y_i | k, \mu, \tau)]
    \]
    \[
    (\mu^{\text{new}}, \tau^{\text{new}}) = \arg \max_{\mu, \tau} Q(\mu, \tau)
    \]
Solving for large-scale data and large networks

- The M-step involves a large-scale optimization problem
- Our goal:
  - Solve for large-scale data input
  - Solve for large network
  - Short term link travel time estimation (say 15min)

Solution: parallelize the computation!
- **Alternating Direction Method of Multiplier** (ADMM) to decouple the problem into smaller sub-problems
- Solve decomposed sub-problems in parallel
- Deals with large size of network and data
- Faster model estimation
Model results for base model

• Validation metrics

• Root mean square error

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (T_{i}^{Pr} - T_{i}^{Ob})^2}
\]

• Mean absolute percentage error

\[
\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{T_{i}^{Pr} - T_{i}^{Ob}}{T_{i}^{Ob}} \right| \times 100\%
\]
### Model results for base model

- Test data: 3/15/2010 ~ 3/21/2010

<table>
<thead>
<tr>
<th>Day</th>
<th>Error</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>9:00-10:00</td>
</tr>
<tr>
<td>Monday</td>
<td>RMSE (min)</td>
<td>2.614</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>29.51%</td>
</tr>
<tr>
<td>Tuesday</td>
<td>RMSE (min)</td>
<td>2.461</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>29.63%</td>
</tr>
<tr>
<td>Wednesday</td>
<td>RMSE (min)</td>
<td>3.827*</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>41.32%*</td>
</tr>
<tr>
<td>Thursday</td>
<td>RMSE (min)</td>
<td>2.468</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>27.28%</td>
</tr>
<tr>
<td>Friday</td>
<td>RMSE (min)</td>
<td>2.26</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>27.76%</td>
</tr>
<tr>
<td>Saturday</td>
<td>RMSE (min)</td>
<td>1.034</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>16.84%</td>
</tr>
<tr>
<td>Sunday</td>
<td>RMSE (min)</td>
<td>2.041</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>25.44%</td>
</tr>
</tbody>
</table>

* Traffic disturbance caused by Patrick's Day Parade.
Numerical Results

Monday 9:00-10:00
Tuesday 9:00-10:00
Wednesday 9:00-10:00
Saturday 9:00-10:00

Monday 13:00-14:00
Tuesday 13:00-14:00
Wednesday 13:00-14:00
Saturday 13:00-14:00

Monday 19:00-20:00
Tuesday 19:00-20:00
Wednesday 19:00-20:00
Saturday 19:00-20:00

Monday 21:00-22:00
Tuesday 21:00-22:00
Wednesday 21:00-22:00
Saturday 21:00-22:00

Urban Link Travel Time Estimation Using Large-scale Taxi Data with Partial Information

Xianyuan Zhan

MPE 2013+
Two new models are proposed to estimate urban link travel times

Utilizing data with only partial information

Efficiently estimation using base model with reasonable accuracy

Mixture models are proposed to get more robust and accurate estimations

Applicable to trajectory data, can provide more accurate estimations

Future work

Test the mixture models for larger network

Efficient implementation using distributed computing technique

Result validation
Thank you!

Questions / Comments?