Non-recurrent events in traffic prediction

Francisco Pereira
IPAM Workshop IV

17th November 2015

Collaborations: MIT, Singapore (SMART), Filipe Rodrigues, Stanislav Borysov, Aidan O’Sullivan, Haizheng Zhang, Constantinos Antoniou, Moshe Ben-Akiva
The cost of disruptions

Source: FHWA, 2005
The cost of disruptions

COUNTING THE FUTURE COST OF GRIDLOCK

The Economic Impact of Congestion in Europe and the US: 2013-2030

‘Cumulative’ cost of congestion between 2013 and 2030

$469B
$480B
$691B
$2.8T

FRANCE
UK
GERMANY
US

‘Annual’ cost of congestion for each country in 2030

Cost to nation
$30B (+31%)
$33B (+33%)
$44B (+31%)
$186B (+50%)

Cost per household
$3,163 (+23%)
$3,217 (+44%)
$2,927 (+34%)
$2,301 (+33%)

PARIS
LONDON
STUTTGART
LOS ANGELES

Cost to city
$18.7B (+60%)
$14.5B (+71%)
$4.28B (+34%)
$38.48B (+65%)

Cost per household
$5,525 (+51%)
$6,259 (+45%)
$5,552 (+35%)
$8,555 (+49%)

The ‘cumulative’ cost to each city by 2030

$267B
$204B
$65B
$559B

Cost of congestion in LA in 2013
$23.2B
– more than 10x the estimated value of the LA Clippers

Cost of congestion in London in 2030
$14.5B
– equivalent to the cost of the 2012 London Olympic Games

$187B/year in USA

$102.8B/yr in non-recurrent congestion!
Predictive traffic management

Recent data

Incident

Traffic estimation and prediction

Behavioral feedback

Predicted impact

Traffic management

Decision
...but do we really have a problem?

Figure 7a: Traffic flow (actual compared to the predicted value) at sensor station 805 EE for the first week of March 1995

From http://rodin.wustl.edu/cosc/doc1.html
Traffic prediction

In fact, the problem is **generally solved** (for a long time) for recurrent or habitual circumstances.

But…

- Under recurrent/habitual behavior, we already know what to expect!
- Help is welcome under unexpected circumstances…
Two approaches to traffic prediction

. Data driven

. Model based
Two approaches to traffic prediction

Data driven
- Mostly statistical (time series, neural networks,...)
- Based on historical data – trouble with *non-recurrent events*!
- Computationally efficient

![Diagram showing two approaches to traffic prediction.](image)
Two approaches to traffic prediction

Model based

- Mostly simulation based (mesoscopic level)
- Represent individual driving behavior (path choice, departure time choice)
- Computationally heavy – trouble with non-recurrent events!

[Diagram showing historical data, supply and demand parameters, state estimation, and state prediction with real-time data and predictions]
Best of both worlds?

- Historical Data
- Supply and demand parameters
- Statistical model
- State estimation
- State prediction
- Real-time data

Predictions
Which parameters?

Demand

. Origin/destination matrix
. Behavior models

Supply

. Capacity reduction
. Speed density functions
## Non-recurrent events

<table>
<thead>
<tr>
<th>Demand</th>
<th>Supply</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictable</td>
<td></td>
</tr>
<tr>
<td>Non-predictable</td>
<td></td>
</tr>
</tbody>
</table>
### Non-recurrent events

<table>
<thead>
<tr>
<th>Demand</th>
<th>Supply</th>
<th>Predictability</th>
</tr>
</thead>
<tbody>
<tr>
<td>special events, demonstrations</td>
<td>road works, closures, mega</td>
<td>Predictable</td>
</tr>
<tr>
<td>holidays</td>
<td>events, weather</td>
<td></td>
</tr>
<tr>
<td>crisis situations, unknown high</td>
<td>incidents, weather, crisis</td>
<td>Non predictable</td>
</tr>
<tr>
<td>fluctuations</td>
<td>situations</td>
<td></td>
</tr>
</tbody>
</table>
Predictable demand
<table>
<thead>
<tr>
<th>Demand</th>
<th>Supply</th>
</tr>
</thead>
<tbody>
<tr>
<td>special events, demonstrations</td>
<td>road works, closures, mega events, weather</td>
</tr>
<tr>
<td>holidays</td>
<td></td>
</tr>
<tr>
<td>crisis situations, unknown high</td>
<td>incidents, weather, crisis situations</td>
</tr>
<tr>
<td>fluctuations</td>
<td></td>
</tr>
</tbody>
</table>

**Predictable**

**Non predictable**
Predictable demand

Relevant events on the internet

. Announced by organizers (e.g. events websites)
. Communicated in social networks (e.g. twitter, facebook)
. Schools and official resources
. Religious venues websites/social networks

Lots of potential sources, but a lot in free form text
Predictable demand

- Special events, demonstrations, holidays
- Road works, closures, in age events, weather
- Crisis situations, unknown high fluctuations

<table>
<thead>
<tr>
<th>Demand</th>
<th>Supply</th>
<th>Non-predictable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special events, demonstration events, holidays</td>
<td>Road works, closures, in age events, weather</td>
<td>Crisis situations, unknown high fluctuations</td>
</tr>
</tbody>
</table>

- Web Mining
- Mobility traces

- Sensing data
  - PT tickets data
  - Traffic
  - Cell phones
  - Cameras
  -...

- Events websites
  - Twitter
  - Facebook
  - News feeds
  - Wikipedia
  - Police feeds
  -...

- Prediction algorithm

- PT tickets data
- Traffic
- Cell phones
- Cameras
-...

- Mobility traces

- Web Mining

- Prediction algorithm

- Sensing data
  - PT tickets data
  - Traffic
  - Cell phones
  - Cameras
  -...
Predictable demand

Web Mining

Sensing data
- PT tickets data
- Traffic
- Cell phones
- Cameras
-...

Mobility traces

Prediction algorithm

category
location
time
venue size
topics
popularity
...

events websites
twitter
Facebook
news feeds
Wikipedia
police feeds
...

Demand
special events,
demonstrations,
holidays

Supply
road works,
disruptions in age
weathers, ... in

Non-predictable

Predictable

predictable

crisis situations,
unknown high
fluctuations

Incidences,
wether, crisis
situations
Predictable demand

- Events websites
  - Twitter
  - Facebook
  - News feeds
  - Wikipedia
  - Police feeds

- Sensing data
  - PT tickets data
  - Traffic
  - Cell phones
  - Cameras
  - Mobility traces

Web Mining

Mobility traces

Prediction algorithm

EZLink data
Airsage data
DynaMIT OD flows

Demand

<table>
<thead>
<tr>
<th>Spatial events</th>
<th>Demonstrations</th>
<th>Holidays</th>
<th>Predictable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Supply

<table>
<thead>
<tr>
<th>Road works</th>
<th>Disasters, inage events, weather events</th>
<th>Non-predictable</th>
</tr>
</thead>
</table>
Predictable demand

- Events websites (twitter, Facebook, news feeds, Wikipedia, police feeds)
- Sensing data (PT tickets data, Traffic, Cell phones, Cameras, ...)
- Web Mining
- Mobility traces
- Prediction algorithm

EZLink data

- Demand
  - Special events, demonstrations, holidays
  - Road works, closures, in age, severe, weather
- Supply
  - Crisis situations, unknown high fluctuations
  - Incidents, weather, crisis situations
  - Non-predictable

Predictable
Predictable demand

- Events websites
  - Twitter
  - Facebook
  - News feeds
  - Wikipedia
  - Police feeds

- Sensing data
  - PT tickets data
  - Traffic
  - Cell phones
  - Cameras
  -...

- Web Mining

- Mobility traces

- Prediction algorithm
- Gaussian Processes
- Artificial Neural Networks
- Graphical models
Modeling demand in planned events

Challenges:
- Multiple simultaneous events
- Trip purpose is not observed

Assumption:
Total demand, $h_n$, is sum of components

$$h_n = a_n + b_n + \epsilon$$

$n$ – time interval ($n<N$)
$a_n$ – routine component
$b_n$ – events component
$\epsilon$ – error term,
Modeling demand in planned events

\[ h_n = a_n + b_n + \epsilon \]

\[ a_n \sim \mathcal{N}(a_n | \eta_a^T x_n^a, \beta_a) \]

\[ b_n = \sum_{i=1}^{E_n} e_n^i, \text{ with } e_n^i \sim \mathcal{N}(e_n^i | \eta_e^T x_n^{e_i}, \beta_e) \]

\[ \epsilon \sim \mathcal{N}(\epsilon | 0, \nu) \]

with \( e_n^i \) - demand for event \( i \) \((i < E_n)\)
\( x_n^a \) - (vector of) routine variables (time of day, day of week…)
\( x_n^{e_i} \) - event variables (discussed shortly…)
\( \eta_a, \eta_e \) - model parameters
\( \nu, \beta_e, \beta_a \) - variances (externally defined)

**Linear model assumption is illustrative**
Modeling demand in planned events

As a generative (graphical) model

\[
p(h_n, a_n, e_n | \eta_a, \eta_e, X_n) = p(h_n | a_n, e_n) p(a_n | \eta_a, x_n^a) \left( \prod_{i=1}^{E_n} p(e^i_n | \eta_e, x_n^{e_i}) \right)
\]

First we need to determine $\eta = [\eta_a, \eta_e]$

Reminding Bayes’ rule (with $D = \{X_n, h_n\}_{n=1}^N$)…

$\text{Posterior} \quad p(\eta|D) = \frac{p(\eta,D)}{p(D)} = \frac{p(D|\eta)p(\eta)}{p(D)}$

likelihood \quad prior

evidence
First we need to determine first the “best” values for \( \eta = [\eta_a, \eta_e] \).

Reminding Bayes’ rule (with \( D = \{X_n, h_n\}_{n=1}^N \))…

\[
p(\eta|D) = \frac{p(\eta, D)}{p(D)} = \frac{p(D|\eta)p(\eta)}{p(D)}
\]

useful consequence:

\[
p(\eta|D) \propto p(D |\eta)p(\eta)
\]
Assuming a zero-mean Gaussian prior for $\eta$, we can do:

$$p(\eta|D) \propto \prod_{n=1}^{N} \int p(h_n, a_n, e_n | \eta, X_n) da_n de_n \ p(\eta)$$

Inference tools:

- Approximate inference – Expectation Propagation (EP), Variational Bayes $\Rightarrow$ Infer.NET
- Monte Carlo methods $\Rightarrow$ STAN package
Modeling demand in planned events

Extension:

. Gaussian Processes priors on routine and events components

. Truncated Gaussian constraints

Details in appendix
Experiments with PT arrivals

Data:

- 5 months of Smartcard data from Singapore
- Focus on 2 areas (which have multiple venues)

Descriptive statistics:

<table>
<thead>
<tr>
<th>Area</th>
<th>Average daily arrivals ± std.</th>
<th>Average daily events ± std.</th>
<th>Maximum daily events</th>
<th>Num. days without events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stadium</td>
<td>4101 (± 925)</td>
<td>0.230 (± 0.554)</td>
<td>3</td>
<td>114 (82.014%)</td>
</tr>
<tr>
<td>Expo</td>
<td>15027 (± 5515)</td>
<td>2.446 (± 1.986)</td>
<td>8</td>
<td>23 (16.547%)</td>
</tr>
</tbody>
</table>
Experiments with PT arrivals

Data:

- Scraping and API based

Summary:

<table>
<thead>
<tr>
<th>Source</th>
<th>Num. events study areas</th>
<th>Number of categories</th>
<th>Text description size (± std. dev.)</th>
<th>Retrieval type</th>
</tr>
</thead>
<tbody>
<tr>
<td>eventful.com</td>
<td>1221</td>
<td>28</td>
<td>1112.3 (± 1337.1)</td>
<td>API</td>
</tr>
<tr>
<td>singaporeexpo.com.sg</td>
<td>58</td>
<td>28</td>
<td>124.9 (± 159.5)</td>
<td>scraper</td>
</tr>
<tr>
<td>last.fm</td>
<td>11</td>
<td>-</td>
<td>901.2 (± 1037.5)</td>
<td>API</td>
</tr>
<tr>
<td>timeoutsingapore.com</td>
<td>568</td>
<td>49</td>
<td>411.8 (± 866.6)</td>
<td>scraper</td>
</tr>
</tbody>
</table>
A quick note on text analysis

. Text represented as bag-of-words
  . Vector of word frequencies
  . Dimensionality of entire dictionary (N)

. **Topic modeling**
  . Dimensionality reduction
  . Text re-represented as a linear combination of $K<<N$ vectors (a.k.a.) topics

*Generative formulation and examples in appendix*
Experiments with PT arrivals

Attributes:

. Routine: time of day; day of week

. Events: topics; category; venue; start/end times; FCBK likes; Google hits; all day-multi-day/short

Different model for each area
Experiments with PT arrivals

Stadium:

<table>
<thead>
<tr>
<th>Model</th>
<th>CorrCoef</th>
<th>RAE</th>
<th>$R^2$</th>
<th>CorrCoef</th>
<th>RAE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Reg. (routine only)</td>
<td>0.646</td>
<td>0.642</td>
<td>0.417</td>
<td>0.649</td>
<td>0.735</td>
<td>0.092</td>
</tr>
<tr>
<td>Linear Reg. (routine + events)</td>
<td>0.739</td>
<td>0.620</td>
<td>0.543</td>
<td>0.709</td>
<td>0.620</td>
<td>0.502</td>
</tr>
<tr>
<td>GP (routine only)</td>
<td>0.667</td>
<td>0.616</td>
<td>0.445</td>
<td>0.654</td>
<td>0.707</td>
<td>0.117</td>
</tr>
<tr>
<td>GP (routine + events)</td>
<td>0.777</td>
<td>0.567</td>
<td>0.603</td>
<td>0.751</td>
<td>0.581</td>
<td>0.564</td>
</tr>
<tr>
<td>BAM-LR</td>
<td>0.737</td>
<td>0.605</td>
<td>0.544</td>
<td>0.694</td>
<td>0.582</td>
<td>0.474</td>
</tr>
<tr>
<td>BAM-GP</td>
<td>0.795</td>
<td>0.556</td>
<td>0.632</td>
<td>0.811</td>
<td>0.503</td>
<td>0.658</td>
</tr>
</tbody>
</table>

Expo:

<table>
<thead>
<tr>
<th>Model</th>
<th>CorrCoef</th>
<th>RAE</th>
<th>$R^2$</th>
<th>CorrCoef</th>
<th>RAE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Reg. (routine only)</td>
<td>0.581</td>
<td>0.723</td>
<td>0.338</td>
<td>0.390</td>
<td>0.816</td>
<td>0.098</td>
</tr>
<tr>
<td>Linear Reg. (routine + events)</td>
<td>0.707</td>
<td>0.617</td>
<td>0.500</td>
<td>0.557</td>
<td>0.743</td>
<td>0.300</td>
</tr>
<tr>
<td>GP (routine only)</td>
<td>0.718</td>
<td>0.576</td>
<td>0.514</td>
<td>0.621</td>
<td>0.670</td>
<td>0.341</td>
</tr>
<tr>
<td>GP (routine + events)</td>
<td>0.750</td>
<td>0.547</td>
<td>0.543</td>
<td>0.676</td>
<td>0.668</td>
<td>0.382</td>
</tr>
<tr>
<td>BAM-LR</td>
<td>0.661</td>
<td>0.652</td>
<td>0.436</td>
<td>0.484</td>
<td>0.772</td>
<td>0.229</td>
</tr>
<tr>
<td>BAM-GP</td>
<td>0.796</td>
<td>0.472</td>
<td>0.633</td>
<td>0.736</td>
<td>0.565</td>
<td>0.540</td>
</tr>
</tbody>
</table>
Experiments with PT arrivals

Stadium area - 2012-11-24
- Observed arrivals
- GP (routine + events)
- BAM-GP (routine + events)

Events:
- Clash of Continents (14:00)
- Dance Drama Opera Marion (20:00)

Stadium area - 2012-12-01
- Observed arrivals
- GP (routine + events)
- BAM-GP (routine + events)

Events:
- 2011 New Zealand Global Tour (14:00)

Expo area - 2012-12-08
- Observed arrivals
- GP (routine + events)
- BAM-GP (routine + events)

Events:
- Mega Sports & Golf Expo Sale (13:00)
- Perfect Living 2012 (12:00)
- Celeste Live (18:00)
- Wellness Summit 2012 (09:00)

Expo area - 2012-12-28
- Observed arrivals
- GP (routine + events)
- BAM-GP (routine + events)

Events:
- Electronics Expo 2012 (11:00)
- Metro Expo Sale (11:00)
Experiments with PT arrivals

(b) BAM-LR

(c) BAM-GP
Predictable supply
Predictable supply

<table>
<thead>
<tr>
<th>Demand</th>
<th>Supply</th>
</tr>
</thead>
<tbody>
<tr>
<td>special events, demonstrations</td>
<td>road works, closures, mega events, weather</td>
</tr>
<tr>
<td>holidays</td>
<td></td>
</tr>
<tr>
<td>crisis situations, unknown</td>
<td>incidents, weather, crisis situations</td>
</tr>
<tr>
<td>high fluctuations</td>
<td></td>
</tr>
</tbody>
</table>

Predictable

Non predictable
Predictable supply

- Road works information feed
- Planned closures (e.g. national day)

Changes in capacity

→ The trick is to get the data!

Weather forecasts

Changes in speed density functions

$$u = u_0 + (u_f - u_0)(1 - \frac{k}{k_f})^\alpha$$

$$q = u k_f [1 - \left(\frac{u - u_0}{u_f - u_0}\right)^{1/\alpha}]$$

$$u_f(r) = ar^b + c$$

- Observed speed
- Min Speed
- Observed flow
Non predictable demand
### Non-predictable Demand

<table>
<thead>
<tr>
<th>Demand</th>
<th>Supply</th>
<th>Predictable</th>
<th>Non predictable</th>
</tr>
</thead>
<tbody>
<tr>
<td>special events, demonstrations, holidays</td>
<td>road works, closures, mega events, weather</td>
<td></td>
<td></td>
</tr>
<tr>
<td>crisis situations, unknown high fluctuations</td>
<td>incidents, weather, crisis situations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Non predictable demand

Lack of information
  - Where? When? What?

Can real-time feeds help?
  - Twitter, Facebook
  - Radio, TV
  - Traffic surveillance

On the semantic side, still an open challenge
Non predictable demand

DynaMIT’s approach

- Historical data
- A priori parameter values
- On-line calibration
- Surveillance data
- On-line calibrated parameters

Non predictable supply
### Non predictable supply

<table>
<thead>
<tr>
<th>Demand</th>
<th>Supply</th>
</tr>
</thead>
<tbody>
<tr>
<td>special events, demonstrations holidays</td>
<td>road works, closures, mega events, weather</td>
</tr>
<tr>
<td>crisis situations, unknown high fluctuations</td>
<td>incidents, weather, crisis situations</td>
</tr>
</tbody>
</table>

**Predictable**

**Non predictable**
Non predictable supply

Data

. Incident response feed (communications)
. Environmental sensing
. Traffic surveillance feed
. Incident detection

Challenges

. How serious (capacity reduction)?
. How long (incident duration)?
. How do things change (speed-density functions)?
Non predictable supply

Data

- Incident response feed (communications)
- Environmental sensing
- Traffic surveillance feed
- Incident detection

Challenges

- How serious (capacity reduction)?
- How long (incident duration)?
- How do things change (speed-density functions)?
Archive with 2 years (2010/11) of all incident records from Singapore. For each incident:

- Lanes blocked
- nr vehicles
- Location
- Time
- communication log.

Focus on **Singapore expressways** (total of 10139 events)
Example:

- id: 473586
- zone ID: 2
- Location (X, Y): 26266.6, 34916.9
- Road name: AYE
- Location type: 3
- Lane blockage: Lane 1, Shoulder
- Down point: 20.32
- Congestion status: 0
- Start time: 2010-08-20 22:50:01
- End time: 2010-08-20 23:31:45
- Number of vehicles: 2

Comm. log:
- 2250hrs - TP Joe X spots an accident. Car and bike involved.
- 2255hrs - Passers-by shift the bike to the shoulder.
- 2300hrs - Ambulance arrives at location. LTM arrives at location.
- 2309hrs - Ambulance conveys rider to National University Hospital.
- 2310hrs - TP arrives at location.
- 2311hrs - Notify by LTM the rider is seriously injured. The accident involves a car and bike.
- 2331hrs - TP requests RC and LTM to resume patrolling. All other vehicles move off. Shoulder clear

Note: This example does not correspond any specific incident in the database, but mimics the type of information available.
Methodology

- Incidents feed
- New data?
  - Yes: extract/update Link ID, location, num_veh, blockage
  - No: CR=blockage/Num_lanes
- Extract topics with LDA
- Incident duration:
  - RF classifier
  - Short durations GP regression
  - Long durations GP regression
- send updated info to DynaMIT supply
Results

$R^2 = 0.6$
$CC = 0.74$
Results

Figure 6: Mean Absolute Percentage Error with number of messages.

Impact in DynaMIT

- Numerical experiments with closed-loop framework (CLF), DynaMIT2.0 with a traffic microsimulator
Impact in DynaMIT

Sensor Counts: Incident affected segment

Error Comparison: only affected segments

<table>
<thead>
<tr>
<th>Scenario</th>
<th>RMSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without II</td>
<td>56%</td>
</tr>
<tr>
<td>Predicted II</td>
<td>36%</td>
</tr>
<tr>
<td>True II</td>
<td>24.6%</td>
</tr>
</tbody>
</table>

II=Incident Information
RMSN=Root Mean Squared error Normalized
Predictive traffic management: Case Study
Case Study

- Incidents/road works from September 15th, 2011
- Two Singapore expressways, 13 ERP gantries

- Simulation of morning peak traffic: 7:30 AM ~ 2:30 PM
- Toll rates changing at 5 min interval
Three scenarios

- Base case
- Guidance
- Guidance and toll optimization
- No guidance
- Predictive guidance
- guidance and optimized tolls
Reduction in Network Delay

**Affected vehicles are defined as vehicles passing incident locations**
Wrapping up
Conclusions

Non-recurrent events challenging for transport prediction
  . Scarce data, difficult to use
  . Statistical methods: too much reliance on historical patterns
  . Model-based approach: not efficient enough in real-time

Proposed approach:
  . Explore wide range data sources (including textual)
  . Combine statistical with model-based techniques

Collaborations: MIT, Singapore (SMART), Filipe Rodrigues, Stanislav Borysov, Haizheng Zhang, Constantinos Antoniou, Moshe Ben-Akiva
Appendix
On topic modeling

- Topic modeling – Latent Dirichlet Allocation (LDA) (Blei et al 2012)
On topic modeling

- Topic modeling – Latent Dirichlet Allocation (LDA) (Blei et al 2012)
- Based on assumptions on probabilistic word distributions by topics
  - Each document is distribution of topics, each topic is a distribution of words
On topic modeling

• Topic modeling – Latent Dirichlet Allocation (LDA) (Blei et al 2012)
• Based on assumptions on probabilistic word distributions by topics
  • Each document is distribution of topics, each topic is a distribution of words
• Generative algorithm (K=number of topics, \( \alpha \) and \( \eta \) are hyperparameters for Dirichlet priors)

1) Draw a topic \( \beta_k \) from \( \beta_k \sim \text{Dirichlet}(\eta) \) for \( k = 1 \ldots K \)

2) For each document \( d \):
   a) Draw topics proportions \( \theta_d \) such that \( \theta_d \sim \text{Dirichlet}(\alpha) \)
   b) For each word \( w_{d,n} \):
      i) Draw topic assignment \( z_{d,n} \sim \text{Multinomial}(\theta_d) \)
      ii) Draw word \( w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}}) \)
LDA graphical model
On topic modeling

• The previous algorithm translates into the following:

\[
p(\theta, \beta \mid w, \alpha, \eta) = \frac{p(w \mid \theta, \beta, \alpha, \eta)p(\theta, \beta \mid \alpha, \eta)}{p(w \mid \alpha, \eta)}
\]

with

\[
p(w \mid \theta, \beta, \alpha, \eta) = \sum_{k=1}^{K} \prod_{d=1}^{D} \prod_{n=1}^{N} p(z = k \mid \theta_d)p(w_{d,n} \mid \beta_z)
\]

\[
p(\theta, \beta \mid \alpha, \eta) = \prod_{k=1}^{K} p(\beta_k \mid \eta)\prod_{d=1}^{D} p(\theta_d \mid \alpha)
\]

• Only \( w_{d,n} \) are observed, so \( z \) are latent variables
• In general, direct inference is intractable. The common solutions are Expectation-Maximization (EM) or Gibbs sampling
An example with $K=6$, $\alpha$ and $\eta=1/6$ with our dataset that has 10139 with 3545 different words overall

- tp #0: $0.15*\text{noinjur} + 0.09*\text{veh} + 0.08*\text{noinfst} + 0.07*\text{came} + 0.06*\text{across} + 0.05*\text{selfdriven}$
- tp #1: $0.11*\text{open} + 0.09*\text{spill} + 0.09*\text{oil} + 0.09*\text{scdf} + 0.08*\text{close} + 0.06*\text{2ln}$
- tp #2: $0.06*\text{polic} + 0.06*\text{drive} + 0.04*\text{crew} + 0.04*\text{drink} + 0.03*\text{div} + 0.03*\text{driver}$
- tp #3: $0.07*\text{ab} + 0.07*\text{tow} + 0.06*\text{convei} + 0.06*\text{rider} + 0.06*\text{hosp} + 0.05*\text{bike}$
- tp #4: $0.08*\text{damag} + 0.07*\text{call} + 0.04*\text{tow} + 0.04*\text{veh} + 0.03*\text{tp} + 0.03*\text{vig}$
- tp #5: $0.06*\text{tp} + 0.05*\text{convei} + 0.05*\text{ab} + 0.04*\text{tow} + 0.04*\text{hosp} + 0.04*\text{veh}$

Our earlier example would have the following assignment (0.01, 0.02, 0.02, 0.32, 0.13, 0.50)

**Note 1:** we show **only** top 6 words of each topic

**Note 2:** we skipped all the details in data preparation and cleaning!
Examples from events

Some topics

Live concerts, dance events, DJs…

topic #0: 0.163*exhibit + 0.103*film + 0.022*solo + 0.021*la + 0.019*best + 0.019*problem
topic #2: 0.059*expo + 0.055*train + 0.029*bar + 0.023*dinner + 0.023*youth + 0.022*wine
topic #3: 0.037*run + 0.032*littl + 0.031*2nd + 0.029*natur + 0.025*trail + 0.024*india + 0.022*sentosa
topic #4: 0.115*intern + 0.088*confer + 0.057*book + 0.053*nation + 0.049*annual + 0.037*food
topic #7: 0.096*parti + 0.085*concert + 0.040*christma + 0.039*launch + 0.038*gardenn + 0.033*xma
topic #9: 0.151*live + 0.062*danc + 0.038*life + 0.029*collect + 0.028*esplanad + 0.021*present
topic #13: 0.095*world + 0.054*present + 0.039*session + 0.034*zouk + 0.033*vs + 0.031*ft + 0.026*dj
topic #10: 0.056*storytel + 0.054*chines + 0.047*english + 0.040*friend + 0.024*publi + 0.023*season
topic #11: 0.059*seminar + 0.041*open + 0.040*asian + 0.032*celebr + 0.026*hous + 0.022*properti
topic #17: 0.058*trade + 0.033*manag + 0.031*onlin + 0.025*stream + 0.024*theatr + 0.023*preview
topic #15: 0.106*night + 0.060*museum + 0.031*famili + 0.022*chroam + 0.021*happi + 0.019*arm
topic #22: 0.213*art + 0.036*game + 0.028*certif + 0.028*football + 0.024*school + 0.019*rock
topic #19: 0.052*design + 0.047*forum + 0.036*park + 0.032*stock + 0.031*paint + 0.019*jurong
topic #20: 0.048*talk + 0.033*movi + 0.033*screen + 0.031*nu + 0.029*competit + 0.028*social
topic #21: 0.169*sale + 0.029*warehous + 0.025*librari + 0.022*bag + 0.022*exam + 0.021*mr25
topic #23: 0.270*festiv + 0.050*stori + 0.037*photographi + 0.030*tv + 0.029*introdutc + 0.019*write
topic #24: 0.167*asia + 0.043*invest + 0.040*meet + 0.032*anniversari + 0.025*comedi + 0.017*profession
Examples from events

Some topics

Cultural events, arts, theatre, cinema

topic #0: 0.163*exhibit + 0.103*film + 0.022*solo + 0.021*la + 0.019*best + 0.019*problem
topic #2: 0.059*expo + 0.055*train + 0.029*bar + 0.023*dinner + 0.023*youth + 0.022*wine
topic #3: 0.037*run + 0.032*littl + 0.031*2nd + 0.029*natur + 0.025*trail + 0.024*india + 0.022*sentosa
topic #4: 0.115*intern + 0.088*confer + 0.057*book + 0.053*nation + 0.049*annual + 0.037*food
topic #7: 0.096*parti + 0.085*concert + 0.040*christma + 0.039*launch + 0.038*garden + 0.033*xma
topic #9: 0.151*live + 0.062*danc + 0.038*life + 0.029*collect + 0.028*esplanad + 0.021*present
topic #13: 0.095*world + 0.054*present + 0.039*session + 0.034*zouk + 0.033*vs + 0.031*ft + 0.026*dj
topic #10: 0.056*storytel + 0.054*chines + 0.047*english + 0.040*friend + 0.024*public+ 0.023*season
topic #11: 0.059*seminar + 0.041*open + 0.040*asian + 0.032*celebr + 0.026*hous + 0.022*properti
topic #17: 0.058*trade + 0.033*manag + 0.031*onlin + 0.025*stream + 0.024*theatr + 0.023*preview
topic #15: 0.106*night + 0.060*museum + 0.031*famili + 0.022*chroam + 0.021*happi + 0.019*arm
itopic #22: 0.213*art + 0.036*game + 0.028*certif + 0.028*football + 0.024*school + 0.019*rock
topic #19: 0.052*design + 0.047*forum + 0.036*park + 0.032*stock + 0.031*paint + 0.019*jurong
topic #20: 0.048*talk + 0.033*movi + 0.033*screen + 0.031*nu + 0.029*competit + 0.028*social
topic #21: 0.169*sale + 0.029*wreahous + 0.025*librari + 0.022*bag + 0.022*exam + 0.021*mr25
topic #23: 0.270*festiv + 0.050*stori + 0.037*photographi + 0.030*tv + 0.029*intoduct + 0.019*write
topic #24: 0.167*asia + 0.043*invest + 0.040*meet + 0.032*anniversari + 0.025*comedi + 0.017*profession
Examples from events

Some topics

Outdoors events, sports

topic #0: 0.163*exhibit + 0.103*film + 0.022*solo + 0.021*la + 0.019*best + 0.019*problem
topic #2: 0.059*expo + 0.055*train + 0.029*bar + 0.023*dinner + 0.023*youth + 0.022*wine

topic #3: 0.037*run + 0.032*littl + 0.031*2nd + 0.029*natur + 0.025*trail + 0.024*india + 0.022*sentosa

Some topics

Outdoors events, sports

topic #4: 0.115*intern + 0.088*confer + 0.057*book + 0.053*nation + 0.049*annual + 0.037*food

topic #7: 0.096*parti + 0.085*concert + 0.040*christma + 0.039*launch + 0.038*garde + 0.033*xma

topic #9: 0.151*live + 0.062*danc + 0.038*life + 0.029*collect + 0.028*esplanad + 0.021*present

topic #10: 0.056*storytel + 0.054*chines + 0.047*english + 0.040*friend + 0.024*puclic+ 0.023*season

topic #11: 0.059*seminar + 0.041*open + 0.040*asian + 0.032*celebr + 0.026*hous + 0.022*properti

topic #17: 0.058*trade + 0.033*manag + 0.031*onlin + 0.025*stream + 0.024*theatr + 0.023*preview

topic #15: 0.106*night + 0.060*museum + 0.031*famili + 0.022*criom + 0.021*happi + 0.019*arm

topic #22: 0.213*art + 0.036*game + 0.028*certif + 0.028*footbal + 0.024*school + 0.019*rock

topic #19: 0.052*design + 0.047*forum + 0.036*park + 0.032*stock + 0.031*paint + 0.019*jurong

topic #20: 0.048*tal + 0.033*movi + 0.033*screen + 0.031*nu + 0.029*competit + 0.028*social

topic #21: 0.169*sale + 0.029*warehous + 0.025*librari + 0.022*bag + 0.022*exam + 0.021*mr25

topic #23: 0.270*festiv + 0.050*stori + 0.037*photographi + 0.030*tv + 0.029*introdut + 0.019*write

topic #24: 0.167*asia + 0.043*invest + 0.040*meet + 0.032*anniversari + 0.025*comedi + 0.017*profession


Examples from events

Some topics

Festivals, exhibitions, sales

topic #0: 0.163*exhibit + 0.103*film + 0.022*solo + 0.021*la + 0.019*best + 0.019*problem

topic #2: 0.059*expo + 0.055*train + 0.029*bar + 0.023*dinner + 0.023*youth + 0.022*wine

topic #3: 0.037*run + 0.032*littl + 0.031*2nd + 0.029*natur + 0.025*trail + 0.024*india + 0.022*sentosa

topic #4: 0.115*intern + 0.088*confer + 0.057*book + 0.053*nation + 0.049*annual + 0.037*food

topic #7: 0.096*parti + 0.085*concert + 0.040*christma + 0.039*launch + 0.038*gard en + 0.033*xma

topic #9: 0.151*live + 0.062*danc + 0.038*life + 0.029*collect + 0.028*esplanad + 0.021*present

topic #13: 0.095*world + 0.054*present + 0.039*session + 0.034*zouk + 0.033*vs + 0.031*ft + 0.026*dj

topic #10: 0.056*storytel + 0.054*chines + 0.047*english + 0.040*friend + 0.024*public+ 0.023*season

topic #11: 0.059*seminar + 0.041*open + 0.040*asian + 0.032*celebr + 0.026*hous + 0.022*properti

topic #17: 0.058*trade + 0.033*manag + 0.031*onlin + 0.025*stream + 0.024*theatr + 0.023*preview

topic #15: 0.106*night + 0.060*museum + 0.031*famili + 0.022*chroam + 0.021*happi + 0.019*arm

topic #22: 0.213*art + 0.036*game + 0.028*certif + 0.028*football + 0.024*school + 0.019*rock

topic #19: 0.052*design + 0.047*forum + 0.036*park + 0.032*stock + 0.031*paint + 0.019*jurong

topic #20: 0.048*talk + 0.033*movi + 0.033*screen + 0.031*nu + 0.029*competit + 0.028*social

topic #21: 0.169*sale + 0.029*warehous + 0.025*librari + 0.022*bag + 0.022*exam + 0.021*mr25

topic #23: 0.270*festiv + 0.050*stori + 0.037*photographi + 0.030*tv + 0.029*introdut + 0.019*write

topic #24: 0.167*asia + 0.043*invest + 0.040*meet + 0.032*anniversari + 0.025*comedi + 0.017*profession
Examples from events

Some topics

Seminars, conferences, storytelling

topic #0: 0.163*exhibit + 0.103*film + 0.022*solo + 0.021*la + 0.019*best + 0.019*problem
topic #2: 0.059*expo + 0.055*train + 0.029*bar + 0.023*dinner + 0.023*youth + 0.022*wine
topic #3: 0.037*run + 0.032*littl + 0.031*2nd + 0.029*natur + 0.025*trail + 0.024*india + 0.022*sentosa
topic #4: 0.115*intern + 0.088*confer + 0.057*book + 0.053*nation + 0.049*annual + 0.037*food
topic #7: 0.096*parti + 0.085*concert + 0.040*christma + 0.039*launch + 0.038*garden + 0.033*xma
topic #9: 0.151*live + 0.062*danc + 0.038*life + 0.029*collect + 0.028*esplanad + 0.021*present
topic #11: 0.059*seminar + 0.041*open + 0.040*asian + 0.032*celebr + 0.026*hous + 0.022*properti
topic #17: 0.058*trade + 0.033*manag + 0.031*onlin + 0.025*stream + 0.024*theatr + 0.023*preview
topic #15: 0.106*night + 0.060*museum + 0.031*famili + 0.022*chroam + 0.021*happi + 0.019*arm
itopic #22: 0.213*art + 0.036*game + 0.028*certif + 0.028*football + 0.024*school + 0.019*rock
topic #19: 0.052*design + 0.047*forum + 0.036*park + 0.032*stock + 0.031*paint + 0.019*jurong
topic #20: 0.048*talk + 0.033*movi + 0.033*screen + 0.031*nu + 0.029*competit + 0.028*social
topic #21: 0.169*sale + 0.029*warehous + 0.025*librari + 0.022*bag + 0.022*exam + 0.021*mr25
topic #23: 0.270*festiv + 0.050*stori + 0.037*photographi + 0.030*tv + 0.029*introduct + 0.019*write
topic #24: 0.167*asia + 0.043*invest + 0.040*meet + 0.032*anniversari + 0.025*comedi + 0.017*profession
Extension to Gaussian Processes

\[ y = f_r(x^r) + \sum_{i=1}^{E} f_e(x^{ei}) + \epsilon \]

\[ y^r \sim \mathbb{I}(y^r > 0) \mathcal{N}(y^r | f_r(x^r), \beta_r), \]
\[ y^{ei} \sim \mathbb{I}(y^{ei} > 0) \mathcal{N}(y^{ei} | f_e(x^{ei}), \beta_e) \]

Defining \( \mathbf{f} \) as the vector with function \( f(x) \) evaluated on all vectors \( \mathbf{x} \)

\[ \mathbf{f}^r \sim \mathcal{GP}(m_r(x^r) \equiv 0, k_r(x^r, x^{r'})) \]
\[ \mathbf{f}^e \sim \mathcal{GP}(m_e(x^e) \equiv 0, k_e(x^e, x^{e'})) \]
Figure 7.2: Factor graph of the proposed Bayesian additive model with Gaussian process components. The outer plate represents the observations, while the inner plate represents the events associated with each observation. The blue arrows represent the message-passing algorithm for performing approximate Bayesian inference. The second flow of messages starting from the GP factor for the events component that goes in the opposite direction is not shown.
GA based solution approach

Control Strategy Optimization
- Store the current state from the Estimation Module

Parallel Evaluation of N Control Strategies of initial/parent population
- Single Cluster of CPUs, Batch size = N
  - CPU 1: Prediction Module, Control Strategy 1; Demand → Supply
  - CPU 2: Prediction Module, Control Strategy 2; Demand → Supply
  - CPU N: Prediction Module, Control Strategy N; Demand → Supply

Generate new set of Control Strategies using crossover & mutation to form child population

Parallel Evaluation of Control Strategies in the child population
- K Cluster of CPUs, Batch size = N/K, K batches evaluated in a distributed manner with N/K control strategies evaluated in parallel on a single cluster
  - Cluster 1
  - Cluster k
  - ... (clusters)

Select best N Control Strategies from (parent + child) population

Terminate
- No

Select the best Strategy from final population
- Yes
Numerical Experiments using DynaMIT2.0

• Application to the problem of real time toll optimization with the objective of minimizing total network travel times

OBJECTIVES

• Examine the performance of the parallel genetic algorithm on the Singapore expressway network

• Examine the impact of providing predictive guidance and optimal tolls in mitigating network congestion under incident scenarios
Performance of Parallel GA

- Singapore Expressway network: 948 nodes, 1150 links, 3891 segments, 4121 OD pairs, 16 toll gantries

- Period between 7:30 AM and 8:30 PM simulated in DynaMIT2.0 using a 5 minute state estimation interval and 15 minute prediction/optimization horizon

- GA parameters - Population size: 60 | cross-over probability: 0.7 | mutation probability: 0.05 | Number of generations: 30 | 25 replications with different initial populations
## Performance of Parallel GA

<table>
<thead>
<tr>
<th>Maximum Toll (SGD)</th>
<th>Travel Time * with static toll</th>
<th>Travel Time with Toll optimization</th>
<th>Percentage Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>480</td>
<td>475</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>480</td>
<td>473</td>
<td>1.32</td>
</tr>
<tr>
<td>3.5</td>
<td>480</td>
<td>472</td>
<td>1.60</td>
</tr>
<tr>
<td>4</td>
<td>480</td>
<td>466</td>
<td>2.70</td>
</tr>
<tr>
<td>4.5</td>
<td>480</td>
<td>466</td>
<td>2.70</td>
</tr>
</tbody>
</table>

*seconds/vehicle

- Significant travel time savings of up to **2.7%** even under non-recurrent scenarios
Some error formulas

Mean absolute percentage error ("text analysis paper")

\[
\text{MAPE} = \frac{1}{|M(t)|} \sum_{i \in M(t)} \left| \frac{\hat{x}_i - x_i}{x_i} \right| \times 100\
\]

where \(M(t)\) corresponds to the set of measurements that occurred in time \(t\) and \(x_i\) and \(\hat{x}_i\) the observed and predicted values, respectively. Figure 2 shows

Mean absolute error (Wikipedia)

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |f_i - y_i| = \frac{1}{n} \sum_{i=1}^{n} |e_i|.
\]

As the name suggests, the mean absolute error is an average of the absolute errors \(e_i = |f_i - y_i|\), where \(f_i\) is the prediction and \(y_i\) the true value. Note that alternative formulations may include relative frequencies as weight factors.

Correlation Coefficient (Wikipedia)

Pearson's correlation coefficient when applied to a sample is commonly represented by the letter \(r\) and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient. We can obtain a formula for \(r\) by substituting estimates of the covariances and variances based on a sample into the formula above. That formula for \(r\) is:

\[
r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}
\]