Statistical Inference for Transport Networks

Martin Hazelton
m.hazelton@massey.ac.nz
Massey University, New Zealand

Presented at Hong Kong University of Science & Technology, October 2011

Outline

1 The Modelling Process
- Objectives of Modelling
- Stages of Model Building

2 Challenges in Modelling and Inferences
- Complexity
- Dangers of Overfitting

3 Statistical Inference for Transport Networks
- Some General Thoughts on Inference
- Statistical Linear Inverse Problems
- Bayesian Inference from Link Counts

4 New Data Opportunities

5 Final Words
- Three Wishes
- Acknowledgements

The Modelling Process

1. Objectives of Modelling
2. Stages of Model Building

Some Objectives of Modelling

(i) **Understanding** transport processes.

- E.g. Improved insight into the mechanisms by which travellers select routes.

Some Objectives of Modelling

(i) **Understanding** transport processes.
- E.g. Improved insight into the mechanisms by which travellers select routes.

(ii) **Estimation** of model parameters.
- E.g. Estimation of elements of an OD trip matrix.

Some Objectives of Modelling

(i) **Understanding** transport processes.
- E.g. Improved insight into the mechanisms by which travellers select routes.

(ii) **Estimation** of model parameters.
- E.g. Estimation of elements of an OD trip matrix.

(iii) **Forecasting** future realized flows.
- E.g. Prediction of mean (and extreme) traffic flows levels on individual network links.

Model Assessment

- Quality/utility of model in terms of process understanding (i) and estimation (ii) can be difficult to judge.
- Prediction accuracy (iii) can be assessed.
- Ability to reproduce historical (observed) data (*flow reconstruction*) is not in itself an indication of a good model.
  - Beware model overfitting (more later...).

Model Assessment

- Quality/utility of model in terms of process understanding (i) and estimation (ii) can be difficult to judge.
- Prediction accuracy (iii) can be assessed.
- Ability to reproduce historical (observed) data (*flow reconstruction*) is not in itself an indication of a good model.
  - Beware model overfitting (more later...).

- All well known to experienced modellers, but keep in mind for later.

Two Stages of Model Building

1. Specification of mathematical model
   - Mathematical abstraction of system;
   - Examination of theoretical model properties;
   - Development of computational algorithms for implementation.
2. Statistical inference for model
   - Estimation of model parameters;
   - Specification of precision of parameter estimates;
   - Assessment of model fit;
   - Statistical comparison of competing models.

Two Stages of Model Building
continued

- Both stages, i.e. (1) mathematical modelling, (2) inference, are critical.
- Transport research literature has tended to focus more heavily on 1 than 2.

Two Stages of Model Building

continued

- Both stages, i.e. (1) mathematical modelling, (2) inference, are critical.
- Transport research literature has tended to focus more heavily on 1 than 2.

Papers in *Transportation Research Part B*

Year(s)	Math Modelling	Inference	Ratio
1989–1991	81	10	8.1:1
2010–2011	133	23	5.8:1

Methods of Inference for Traffic Models

Stochastic Models

- Traffic flows modelled as random variables.
- Standard statistical methodologies can be applied (in theory).
  - Estimation by maximum likelihood, least squares, method of moments estimation, Bayesian methods.
  - Model comparison by likelihood and Bayesian methods.
  - Goodness-of-fit testing for models.
  - Minimum-error prediction.

Methods of Inference for Traffic Models

Stochastic Models

- Traffic flows modelled as random variables.
- Standard statistical methodologies can be applied (in theory).
  - Estimation by maximum likelihood, least squares, method of moments estimation, Bayesian methods.
  - Model comparison by likelihood and Bayesian methods.
  - Goodness-of-fit testing for models.
  - Minimum-error prediction.

Deterministic Models

- More difficult to apply principled methodology
- One approach is to embed in stochastic model and fit that
  - E.g. SUE as approximate mean of Markov assignment process.

Challenges in Modelling and Inferences

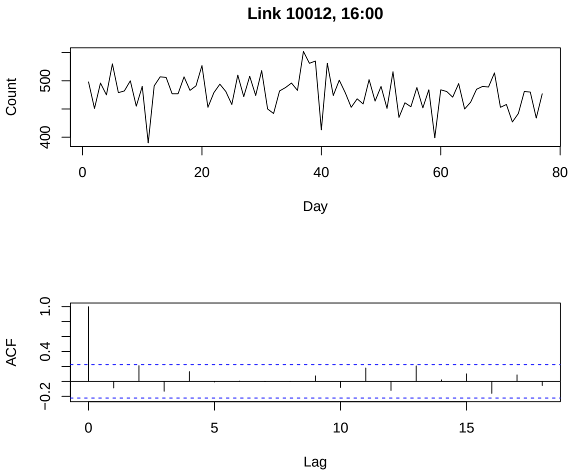
1. Complexity
2. Dangers of Overfitting

Modelling Traffic Networks is Complex

- Transport networks are prime examples of complex systems.
  - Macroscopic properties are intricate functions of multitudinous elements (individual vehicles; traffic signals; etc.).
  - Model specification challenging – what's important, what's not?
- Real-world applications can be very large scale.
  - Computational feasibility?
- Are suitable data available for effective model calibration?
  - More on that later...

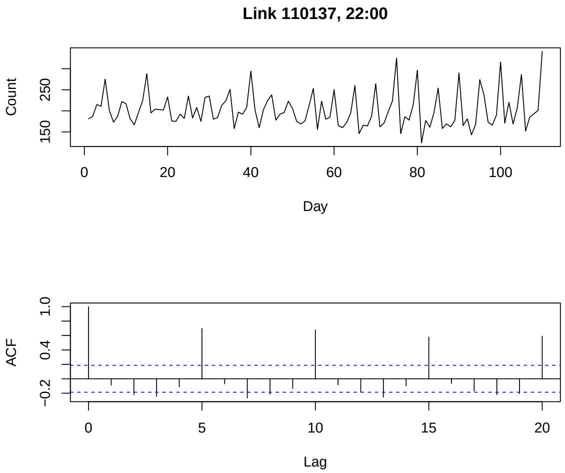
System Complexity: Example 1

Traffic Counts 16:00-17:00 northbound on Broadway, Bronx NY



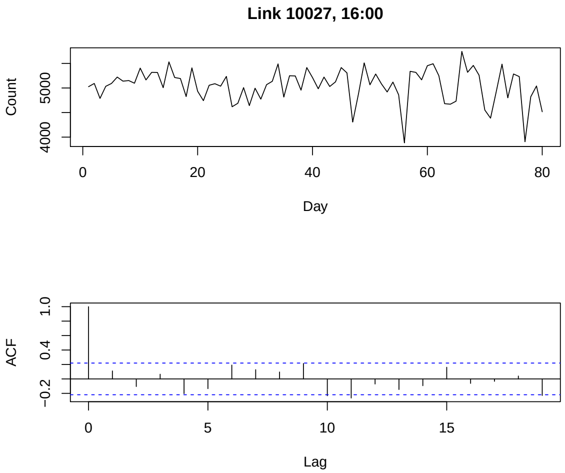
System Complexity: Example 2

Traffic Counts 22:00-23:00 northbound on Albany Post Road, Albany NY



System Complexity: Example 3

Traffic Counts 16:00-17:00 northbound on Major Deegan Expressway, Bronx NY



Examples of System Complexity

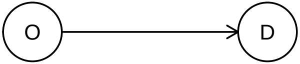
What are the Important Features?

- Example 1: flows more or less a sequence of independent random variables.
- Example 2: flows show periodicity – day-of-the-week effect.
- Example 3: flows show heteroscedasticity.
- Can our models capture all these characteristics?
- Do our models need to??

The Dangers of Over-fitting

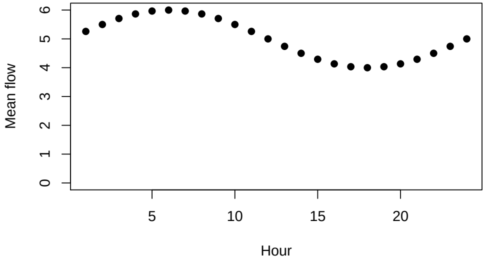
- Complexity of traffic systems may tempt us to build very complex models...
- ... but excessively complex models lead to over-fitting.
- Over-fitted models are deceptively 'realistic'.
- Excellent at reproducing yesterday.
- Poor at forecasting tomorrow.

An Illustrative Toy Example



Model

- Observed flow: $u^t \sim \text{Pois}(\mu^t)$, $t = 1, 2, \ldots, 24$.
- Mean flow: $\mu^t = 5 + \sin(\pi t/12)$, $t = 1, 2, \ldots, 24$.



Toy Example
continued

Aim

- To model hourly traffic flow from O to D.

Available Data

- Hourly counts: $\{u^t : t = 1, 2, \ldots, 24\}$ from one day only.

Toy Example: Two Possible Models

Complex Model

- Model correctly specified.
- Model complex: 24 unknown parameters, $\{\mu^t : t = 1, \ldots, 24\}$, to be estimated.

Simple Model

- Just model average hourly flow.
- I.e. assume that $\mu^t = \mu$ (constant) for all t.
- Model mis-specified.
- Model simple: just 1 parameter, μ, to be estimated.

Toy Example: Model Fitting

Parameter Estimation

- Complex model: $\hat{\mu}^t = u^t$.
- Simple model: $\hat{\mu}^t = \hat{\mu} = \frac{1}{24}\sum_{t=1}^{24} u^t$.

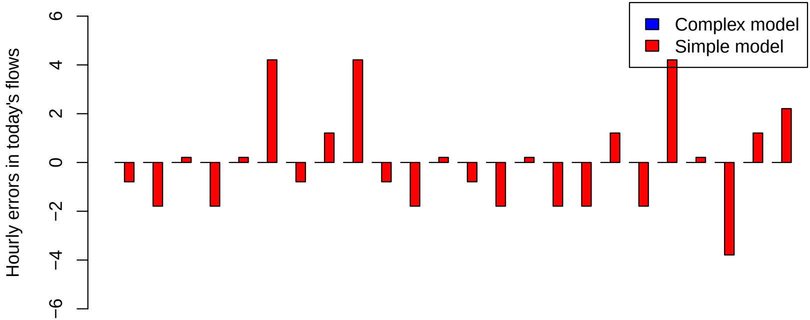
Toy Example: Model Performance

Hourly errors of form $e^t = \hat{\mu}^t - \breve{u}^t$.

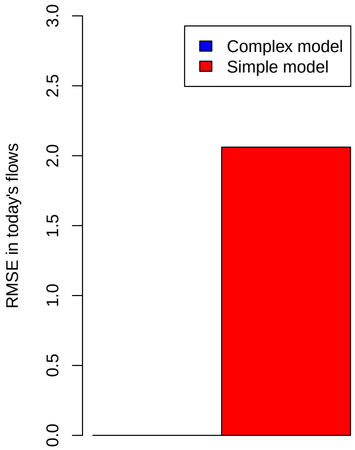
- Reconstruction: $\breve{u}^t = u^t$ is today's observed flow (used to fit model)
- Prediction: $\breve{u}^t$ is tomorrow's observed flow
- Weekly prediction: $\breve{u}^t$ is flow (for hour t) averaged over all days next week.

Root mean squared error: $\sqrt{\frac{1}{24}\sum_{t=1}^{24}(e^t)^2}$

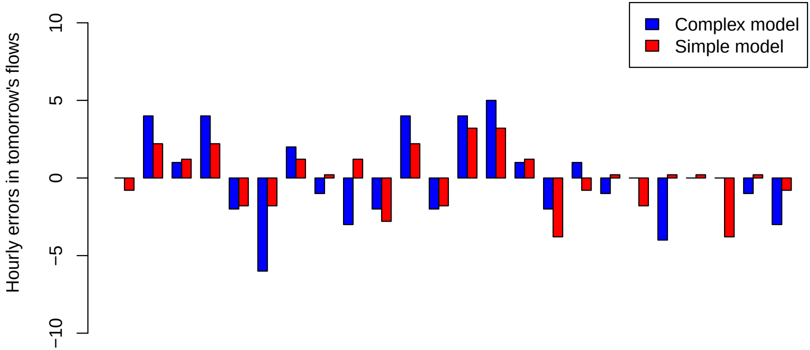
Toy Example: Hourly Errors in Reconstruction



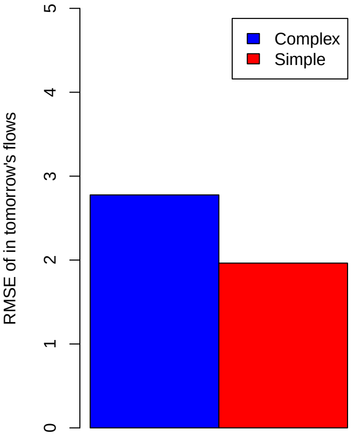
Toy Example: Aggregate Error in Reconstruction



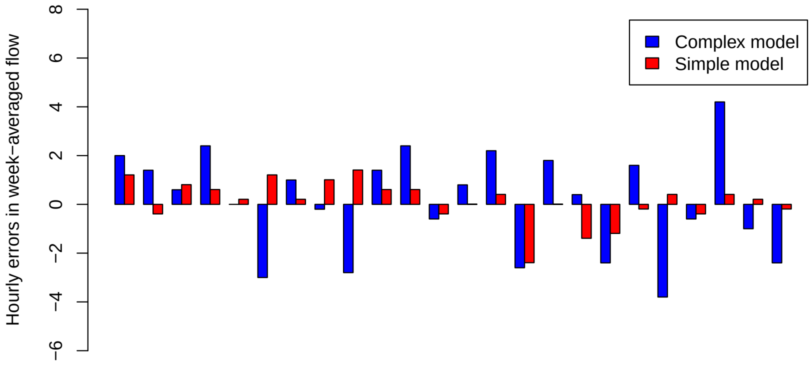
Toy Example: Hourly Errors for Tomorrow



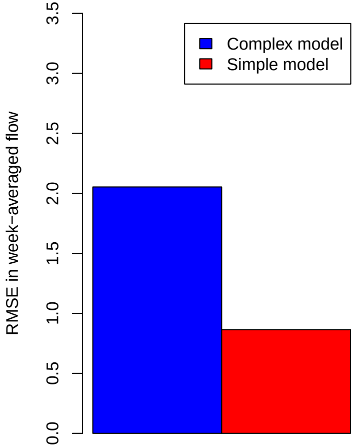
Toy Example: Aggregate Error for Tomorrow



Toy Example: Hourly Errors for Next Week



Toy Example: Aggregate Error for Next Week



Toy Example: Summary of Results

- Complex (hour-to-hour) model is great at forecasting yesterday.
- Simple (day-to-day) model is much better at predicting tomorrow.

Some Conclusions

- More realistic/complex models will not necessarily be better in practice.
- There is a trade off between bias (model mis-specification) and variance (precision in model fitting).
- We should develop models with an emphasis on estimation and prediction, not reconstruction.
- Model design should account for feasibility of good calibration.

Statistical Inference for Transport Networks

1. Some General Thoughts on Inference
2. Statistical Linear Inverse Problems in Transportation
3. Bayesian Inference from Link Counts

Preparatory Notation

Random Variables

$\boldsymbol{u} = (u_1, \ldots, u_L)^{\mathsf{T}}$	OD flows
$\boldsymbol{x} = (x_1, \ldots, x_M)^{\mathsf{T}}$	route (path) flows
$\boldsymbol{y} = (y_1, \ldots, y_N)^{\mathsf{T}}$	link (arc) flows

Model Parameters

$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)^{\mathsf{T}}$	mean OD flows
$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)^{\mathsf{T}}$	mean route flows

Data

- **Link count data**
  - Widely available
  - Typically unbiased
- **Vehicle routing information**
  - Availability varies
  - Can be biased
- **Other**
  - Surveys (bias? coverage?)
  - Experiments

Model Parameterization

- Some parameters can be estimated directly from link counts
  - E.g. link performance (cost) functions
- For many parameters, route flow data required to provide direct information.
  - E.g. 1: origin-destination matrix.
  - E.g. 2: Behavioural parameters control route choice/learning (such as logit route choice parameter).

Inference from Link Counts Alone

Link counts and indeterminism

Fundamental equation

$$\boldsymbol{y} = A\boldsymbol{x}$$

- $A = (a_{ij})$ is routing matrix.
  - $a_{ij} = 1$ if link i on route $j, 0$ otherwise.
- Number links $= N = \dim(\boldsymbol{y})$.
- Number routes $= M = \dim(\boldsymbol{x})$.
- Typically $N << M$ so equations hugely underdetermined.
- *Feasible route set* $\mathcal{X}_{\boldsymbol{y}} = \{\boldsymbol{x} : \boldsymbol{y} = A\boldsymbol{x}\}$ can defy enumeration.

The Importance of second order properties

Data

$\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots, \boldsymbol{y}^n$ sequence of link counts

First Order Statistical Properties

$$\bar{\boldsymbol{y}} = A\bar{\boldsymbol{x}}$$

- Mean link counts provide just N pieces of information.

The Importance of second order properties

Second Order Statistical Properties

$$S_{\boldsymbol{y}} = A^T S_{\boldsymbol{x}} A$$

- Sample variance provides $N(N-1)/2$ pieces of information.

The Importance of second order properties

Second Order Statistical Properties

$$S_{\boldsymbol{y}} = A^T S_{\boldsymbol{x}} A$$

- Sample variance provides $N(N-1)/2$ pieces of information.

Conclusion

Second order properties provide lots of additional information.

Linear Inverse Framework

Statistical Linear Inverse Problem

$$q(\boldsymbol{y}) = \int h(\boldsymbol{y}, \boldsymbol{x})\, dP(\boldsymbol{x})$$

- P is probability measure for latent variables $\boldsymbol{x}$
- h is *blurring* function
- q is density/mass function for observed variables $\boldsymbol{y}$.
- Examples:
  - Image deblurring
  - Decomposition of chemical spectra

Linear Inverse Problems in Transport

$$q(\boldsymbol{y}) = \int h(\boldsymbol{y}, \boldsymbol{x}) \, dP(\boldsymbol{x})$$

- $P(\boldsymbol{x})$ probability measure for route flows
  - possibly over multiple days
- $q(\boldsymbol{y})$ probability density/mass function for link flows.
- $h(\boldsymbol{y}, \boldsymbol{x}) = \mathbf{1}_{\boldsymbol{x} \in \mathcal{X}_{\boldsymbol{y}}}$ for error-free counts.
- E.g. $h(\boldsymbol{y}, \boldsymbol{x}) = f(\boldsymbol{y} - A\boldsymbol{x})$ for counts with measurement error.

Statistical Linear Inverse Problems (SLIPs)

- Puts inference for transport networks in wider context.
- Lots known about these problems ...
  - SLIPs are hard
  - Regularization typically necessary
  - Bayesian framework attractive
  - Each problem is different
- ... but much remains to be done.

  Hazelton, M.L. (2010). Bayesian inference for network-based modes with a linear inverse structure. *Transportation Research Part B*, **44**. 674-785

Overview of Bayesian Inference

- Suppose we have model parameterised by $\boldsymbol{\psi}$.
- In Bayesian paradigm, $\boldsymbol{\psi}$ is a random vector.
  - Distribution of $\boldsymbol{\psi}$ represents our knowledge/beliefs about it.
- Any existing knowledge expressed by *prior*, density $f(\boldsymbol{\psi})$.
  - Provides principled method for incorporating additional information to counter indeterminism problems.
- After observing link counts $\boldsymbol{y}$, distribution updates to *posterior*: $f(\boldsymbol{\psi}|\boldsymbol{y})$.
- Posterior mode or mean can be used as point estimate of $\boldsymbol{\psi}$.
- Being likelihood-based, Bayesian methods automatically incorporate second (and higher) order properties of the data.

The Posterior Distribution

- Posterior related to prior by

$$f(\boldsymbol{\psi}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\psi})f(\boldsymbol{\psi})}{f(\boldsymbol{y})}$$

  - $f(\boldsymbol{y}|\boldsymbol{\psi}) = L(\boldsymbol{\psi})$ is model *likelihood*.
- $f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{\psi})f(\boldsymbol{\psi})\, d\boldsymbol{\psi}$ is marginal density of $\boldsymbol{y}$ (just a normalizing constant).

The Posterior Distribution

- Posterior related to prior by

$$f(\boldsymbol{\psi}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\psi})f(\boldsymbol{\psi})}{f(\boldsymbol{y})}$$

  - $f(\boldsymbol{y}|\boldsymbol{\psi}) = L(\boldsymbol{\psi})$ is model *likelihood*.
- $f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{\psi})f(\boldsymbol{\psi})\, d\boldsymbol{\psi}$ is marginal density of $\boldsymbol{y}$ (just a normalizing constant).

- To compute Bayesian posterior exactly, need:
  - Prior and likelihood in closed form;
  - To be able to evaluate high-dimensional integral for $f(\boldsymbol{y})$ (can be hard).
- Alternative is to simulate from posterior.
  - Markov chain Monte Carlo methods; MCMC.

Likelihood Functions

- Recall, to implement Bayesian approach we must compute model likelihood.
- Likelihood (for static, discrete traffic models):

$$L(\psi) = \sum_{\boldsymbol{x}} f(\boldsymbol{y}|\boldsymbol{x}, \psi) f(\boldsymbol{x}|\psi) = \sum_{\boldsymbol{y} \in \mathcal{X}(\boldsymbol{y})} f(\boldsymbol{x}|\psi)$$

  - $\mathcal{X}(\boldsymbol{y}) = \{\boldsymbol{x} : \boldsymbol{y} = A\boldsymbol{x}\}$ is set feasible route flows.
  - ψ model parameters

Likelihood Functions

- Recall, to implement Bayesian approach we must compute model likelihood.
- Likelihood (for static, discrete traffic models):

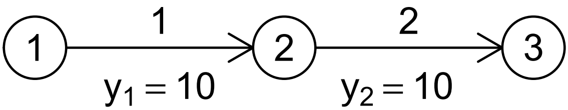
$$L(\psi) = \sum_{\boldsymbol{x}} f(\boldsymbol{y}|\boldsymbol{x}, \psi) f(\boldsymbol{x}|\psi) = \sum_{\boldsymbol{y} \in \mathcal{X}(\boldsymbol{y})} f(\boldsymbol{x}|\psi)$$

  - $\mathcal{X}(\boldsymbol{y}) = \{\boldsymbol{x} : \boldsymbol{y} = A\boldsymbol{x}\}$ is set feasible route flows.
  - ψ model parameters

Passing Comment

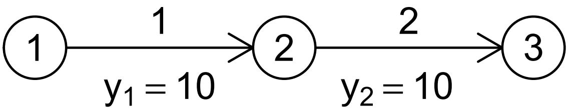
- Many *ad hoc* methods of inference look for one ('optimal') element of $\mathcal{X}(\boldsymbol{y})$.
- Above demonstrates that optimal inference needs information from *all* possible route flows (in theory at least).

Toy Example



$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_{12} \\ x_{13} \\ x_{23} \end{bmatrix} = A\boldsymbol{x}$$

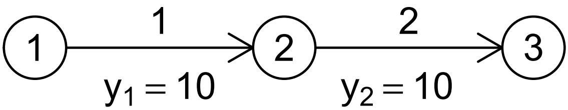
Toy Example



$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_{12} \\ x_{13} \\ x_{23} \end{bmatrix} = A\boldsymbol{x}$$

Conditional on $\boldsymbol{y} = (10, 10)^\mathsf{T}$, latent route flows might be $\boldsymbol{x} = (0, 10, 0)^\mathsf{T}$ or $\boldsymbol{x} = (5, 5, 5)^\mathsf{T}$, or ...

Toy Example continued



$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_{12} \\ x_{13} \\ x_{23} \end{bmatrix} = A\boldsymbol{x}$$

$$\begin{aligned} \mathcal{X}(\boldsymbol{y}) &= \{(x_{12}, x_{13}, x_{23})^{\mathsf{T}} : x_{12} + x_{13} = y_1, x_{23} + x_{13} = y_2\} \\ &= \{(0, 10, 0)^{\mathsf{T}}, (1, 9, 1)^{\mathsf{T}}, \ldots, (10, 0, 10)^{T}\}. \end{aligned}$$

Computation of the Likelihood

- Recall
$$L(\boldsymbol{\psi}) = \sum_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{y})} f(\boldsymbol{x}|\boldsymbol{\psi})$$
- Full evaluation of $L(\boldsymbol{\psi})$ requires all solutions to $\boldsymbol{y} = A\boldsymbol{x}$.
- Solution set $\mathcal{X}(\boldsymbol{y})$ can be huge.
- Enumeration computationally infeasible.

Sampling Based Methods

- Recall

$$L(\boldsymbol{\psi}) = \sum_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{y})} f(\boldsymbol{x}|\boldsymbol{\psi})$$

- Approximate by sampling from $\mathcal{X}(\boldsymbol{y})$.

Sampling Based Methods

- Recall
$$L(\boldsymbol{\psi}) = \sum_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{y})} f(\boldsymbol{x}|\boldsymbol{\psi})$$
- Approximate by sampling from $\mathcal{X}(\boldsymbol{y})$.
- Problem – typically can't enumerate set $\mathcal{X}(\boldsymbol{y})$.
- So how to generate candidates from $\mathcal{X}(\boldsymbol{y})$ (as part of MCMC)?

Sampling Based Methods

- Recall

$$L(\boldsymbol{\psi}) = \sum_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{y})} f(\boldsymbol{x}|\boldsymbol{\psi})$$

- Approximate by sampling from $\mathcal{X}(\boldsymbol{y})$.

- Problem – typically can't enumerate set $\mathcal{X}(\boldsymbol{y})$.
- So how to generate candidates from $\mathcal{X}(\boldsymbol{y})$ (as part of MCMC)?

- Some solutions proposed, but not entirely reliable/general.
  Tebaldi, C. and M. West (1998). Bayesian inference on network traffic using link count data (with discussion). *JASA* **93**, 557–576.
  Hazelton, M.L. (2010). Statistical inference for transit system origin-destination matrices. *Technometrics*, **52** 221-230.

Conclusions on Bayesian Inference

- Bayesian inference a nice approach in principle
- In practice, likelihood and hence posterior can only be approximated by sampling-based methods.
- Efficient algorithms difficult.
- Work in progress with PhD student Katharina Parry.

New Data Opportunities

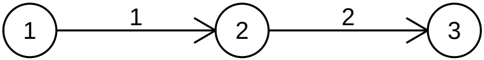
- Problems with inference arise from linear inverse lack of identifiability when using link counts only.
- Modern data collection methods provide scope for resolution.
- Just a little routing information can help a lot.

Incorporating Sporadic Routing Information in OD Matrix Estimation

- Suppose we have some routing information from e.g. tracking GPS equipped vehicles.
- Let $\boldsymbol{p}$ be vector of probabilities of vehicle tracking for each route.
- If exogenous estimates of $\boldsymbol{p} > 0$ are available, then identifiability problems in theory addressed.
- Where available, important to include both sporadic routing *and* link count data.
- When collected contemporaneously, creates two part likelihood:

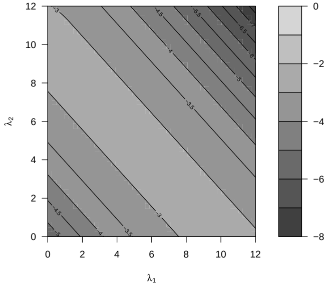
$$L(\boldsymbol{\lambda}, \boldsymbol{p}) = f(\boldsymbol{y}_{not} \,|\, \boldsymbol{\lambda}, \boldsymbol{p}) \cdot f(\boldsymbol{x}_{trk} \,|\, \boldsymbol{\lambda}, \boldsymbol{p})$$

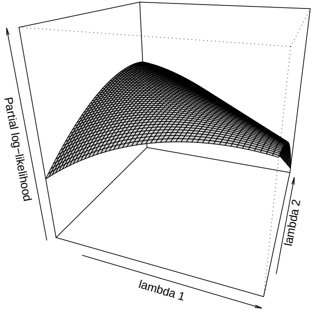
Example



Observe data $\boldsymbol{y} = (10, 10)^\mathsf{T}$ and $\boldsymbol{x}_{trk} = (1, 1, 1)$, so that $\boldsymbol{y}_{not} = (8, 8)^\mathsf{T}$.

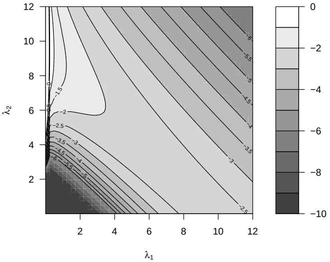
Example (Profile) Log-Likelihood Without Routing Information

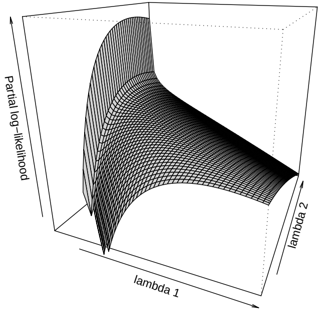




Ridged – complete lack of identifiability.

Example (Profile) Log-Likelihood With Routing Information





Curvature introduced, and hence unique maximum likelihood estimate obtained.

Further Comments on Incorporation of Routing Information

- Problems much harder when $\boldsymbol{p}$ is not known.
- In extreme case, routing information is then of no help at all!
- In practice we can get somewhere by using simple (perhaps crude) models for $\boldsymbol{p}$.
- Analysis done in collaboration with Katharina Parry.
  Parry, K. and M.L. Hazelton (2011). Estimation of origin-destination matrices from link counts and sporadic routing data. *Transportation Research Part B*, in press.

A Statistician's Three Wishes

A Statistician's Three Wishes

1. More emphasis on principled model assessment and comparison.

A Statistician's Three Wishes

1. More emphasis on principled model assessment and comparison.
2. Establishment of public warehouse of real case studies (including all available data; network topology etc.).

A Statistician's Three Wishes

1. More emphasis on principled model assessment and comparison.
2. Establishment of public warehouse of real case studies (including all available data; network topology etc.).
3. Development of better methods for sampling from set of feasible route flows.

Acknowledgements

Support from the Royal Society of New Zealand (Marsden fund) gratefully acknowledged.

David Watling (ITS, University of Leeds, United Kingdom)

Katharina Parry (IIMS, Massey University, New Zealand)

For a copy of these slides...

http://www.massey.ac.nz/~mhazelto/seminars