Statistical Inverse Problems in Transportation Science

Past Lessons and Future Challenges

Martin Hazelton <martin.hazelton@otago.ac.nz>
Department of Mathematics & Statistics, University of Otago, NZ

Emerging Challenges in Statistical Modeling for Transportation Research, Banff International Research Station, 28 June – 3 July 2026

Modelling Travel Behaviour

Development of better road traffic systems depends fundamentally on understanding travel behaviour.

- What trips do travellers want to make?
- What mode will they select?
- What route? departure time? etc.

Existing traffic models address:

- Route choice – traffic assignment
- Travel demand – estimation of origin-destination matrices
- Traveller learning – dynamic traffic models
- etc.

The Role of Statistics

Traffic models exist at various levels of aggregation, and with various degrees of complexity

Common statistical tasks include:

- Model fitting – parameter estimation
- Quantification of uncertainty – confidence intervals, etc.
- Prediction – point and interval
- Hypothesis testing
- Model assessment and comparison

Data Sources for Learning about Travel Behaviour

	Link counts	Route flows	Surveys & Experiments
Availability	high	sporadic	sporadic
Cost	low	variable	high
Bias	no	probably	usually
Directness	low	high	mixed

Route flow data collected by e.g.

- Direct vehicle tracking – e.g. taxis (generalizability?)
- Number plate matching (coverage?)
- Sensor based vehicle trajectory reconstruction

Statistical Inverse Problem Formulation

Statistical inverse problems arise when a random variable of interest is observed only indirectly

- Ubiquitous when modelling travel behaviour

In many cases route traffic volumes are fundamental

- Highly informative about aggregate travel behaviour
- Sufficient statistics for many traffic models
- Seldom directly available

Statistical models relate observed data to route volumes

- Generates linear inverse problems

Linear Inverse Problems for Count Data

For count data, statistical linear inverse problems characterised by

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

- $\boldsymbol{x} \in \mathbb{Z}^r_{\geq 0}$ is count vector of interest;
- $\boldsymbol{y} \in \mathbb{Z}^n_{\geq 0}$ is vector of observed counts.
- **Configuration matrix** A is $n \times r$ and non-negative integer elements (often binary).

Typically $r > n$ so linear system will be (heavily) underdetermined.

Aim is to perform inference for $\boldsymbol{x}$ and/or parameter vector $\boldsymbol{\theta}$ describing underlying distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$

- $f(\boldsymbol{x}|\boldsymbol{\theta})$ may be explicit, or implicit from black box model
- Prior information or auxiliary data often critical.

Note linearity consequence of arithmetic, not model assumptions.

Network Tomography

Linear inverse problems for count data arise in many areas

In each case, observed data is either a summarised or corrupted version of target variables

A few examples:

- Mark-recapture surveys in ecology
- Inference for haplotypes in genomics
- Analysis of mail items for biosecurity surveillance
- Contingency table resampling

For network traffic data, inference in linear inverse problems is often referred to as **network tomography** (Vardi, 1996)

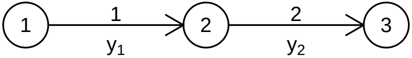
Vardi, Y. (1996). Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, **91(433)**, 365–377.

Inference from Link Count Data

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

- $\boldsymbol{y} \in \mathbb{Z}^n_{\geq 0}$ is vector of recorded link counts
- $\boldsymbol{x} \in \mathbb{Z}^r_{\geq 0}$ is vector of route volumes

Example



$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

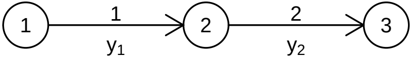
E.g. model route volumes $\boldsymbol{x} \sim \mathrm{Pois}(\boldsymbol{\theta})$

Inference from Link Count Data with Errors

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

- $\boldsymbol{y} \in \mathbb{Z}^n_{\geq 0}$ is vector of recorded link counts
- $\boldsymbol{x} \in \mathbb{Z}^r_{\geq 0}$ is vector of route volumes

Example: probability p of double count on link 1



$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_{1_1} \\ x_{1_2} \\ x_2 \\ x_{3_1} \\ x_{3_2} \end{bmatrix}$$

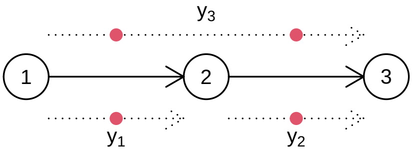
$$E[\boldsymbol{x}] = \begin{bmatrix} \theta_1(1-p) \\ \theta_1 p \\ \theta_2 \\ \theta_3(1-p) \\ \theta_3 p \end{bmatrix}$$

Inference Trajectory Data with Detection Errors

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

- $\boldsymbol{y} \in \mathbb{Z}^n_{\geq 0}$ is vector of recorded route trajectories
- $\boldsymbol{x} \in \mathbb{Z}^r_{\geq 0}$ is vector of true route volumes

Example: probability p that vehicle missed by sensor 1



$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{1_1} \\ x_{1_2} \\ x_2 \\ x_{3_1} \\ x_{3_2} \end{bmatrix}$$

$$\mathrm{E}[\boldsymbol{x}] = \begin{bmatrix} \theta_1(1-p) \\ \theta_1 p \\ \theta_2 \\ \theta_3(1-p) \\ \theta_3 p \end{bmatrix}$$

Principled Statistical Inference

Standard statistical theory provides practical guidance

Requires coherent statistical models

Likelihood-based methods typically optimal

- Optimal in frequentist setting
- Likelihood essential in Bayesian framework

Alternatives to likelihood most attractive when based on principled approximations

Likelihood for Discrete Traffic Models

Recall: probability mass function for route volumes is $f(\boldsymbol{x}|\boldsymbol{\theta})$

- **Example:** Independent Poisson model $f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{r} e^{-\theta_i}\theta_i^{x_i}/x_i!$

Likelihood is

$$\begin{aligned} L(\boldsymbol{\theta}) &= f(\boldsymbol{y}|\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{x}\in\mathbb{Z}^r_{\geq 0}} f(\boldsymbol{x}|\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{x}\in\mathcal{F}_{A,\boldsymbol{y}}} f(\boldsymbol{x}|\boldsymbol{\theta}) \end{aligned}$$

$\mathcal{F}_{A,\boldsymbol{y}} = \{\boldsymbol{x} \colon \boldsymbol{y} = A\boldsymbol{x}\} \cap \mathbb{Z}^r_{\geq 0}$

- Solution set to under-determined linear system
- Called the ***y*-fibre**

Likelihood-Based Inference in Practice

Recall: likelihood is $L(\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in \mathcal{F}_{A,\boldsymbol{y}}} f(\boldsymbol{x}|\boldsymbol{\theta})$

In most real problems, fibre $\mathcal{F}_{A,\boldsymbol{y}}$ will be too large to enumerate.

Example:

How many trip patterns consistent with zonal exit/entry counts?

	Destination				
Origin	1	2	3	4	Total
1	?	?	?	?	220
2	?	?	?	?	215
3	?	?	?	?	93
4	?	?	?	?	64
Total	108	286	71	127	592

Likelihood-Based Inference in Practice

Recall: likelihood is $L(\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in \mathcal{F}_{A,\boldsymbol{y}}} f(\boldsymbol{x}|\boldsymbol{\theta})$

In most real problems, fibre $\mathcal{F}_{A,\boldsymbol{y}}$ will be too large to enumerate.

Example:
How many trip patterns consistent with zonal exit/entry counts?

	Destination				
Origin	1	2	3	4	Total
1	?	?	?	?	220
2	?	?	?	?	215
3	?	?	?	?	93
4	?	?	?	?	64
Total	108	286	71	127	592

Answer: 1,225,914,276,276,768,514

Likelihood and OD Matrices

Estimation of origin-destination (OD) trip matrices is an important type of network tomography

Ideal approach:

- Develop coherent statistical model
  - E.g. account for uncertainty in route choice probabilities
- Identify target for inference
  - Estimation? Reconstruction? Prediction? (H, 2001)
- Apply likelihood-based methods for estimation and uncertainty specification
  - Search for 'most likely' $\boldsymbol{x} \in \mathcal{F}_{A,\boldsymbol{y}}$ common in transportation literature, but demonstrably sub-optimal

---

Hazelton, M.L. (2001). Inference for origin-destination matrices: estimation, reconstruction and prediction. *Transportation Research B*, **35**, 667–676.

Samplers for Likelihood-Based Inference

Recall: likelihood is $L(\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in \mathcal{F}_{A,\boldsymbol{y}}} f(\boldsymbol{x}|\boldsymbol{\theta})$

In essence, likelihood can be approximated by taking a sample $\{\boldsymbol{x}_{(i)} \in \mathcal{F}_{A,\boldsymbol{y}} \colon i = 1, \dots, s\}$

Frequentist likelihood-based inference (H, 2015)

- Employ stochastic EM algorithm
- Sample from $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$

Bayesian inference (Tebaldi & West, 1998)

- Employ e.g. Metropolis-Hastings algorithm
- Iterate sampling from $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ with sampling from $f(\boldsymbol{\theta}|\boldsymbol{x})$.

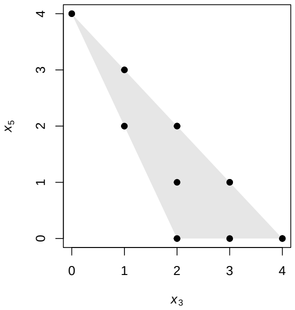
Tebaldi, C., & West, M. (1998). Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, **93(442)**, 557–573.

Hazelton, M.L. (2015). Network tomography for integer-valued traffic. *Annals of Applied Statistics*, **9(1)**, 474–506

Fibre Sampling

Sampling $\boldsymbol{x}$ from $\mathcal{F}_{A,\boldsymbol{y}}$ is difficult!

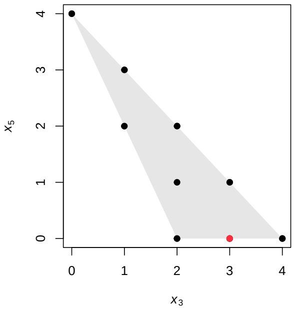
- Hard to ensure connectivity of sampler
- Hard to find helpful characterizition of fibre geometry ($\mathbb{Z}$-polytope)



Fibre Sampling

Sampling $\boldsymbol{x}$ from $\mathcal{F}_{A,\boldsymbol{y}}$ is difficult!

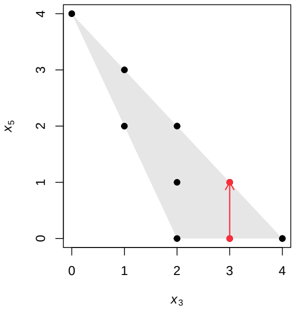
- Hard to ensure connectivity of sampler
- Hard to find helpful characterizition of fibre geometry ($\mathbb{Z}$-polytope)



Fibre Sampling

Sampling $\boldsymbol{x}$ from $\mathcal{F}_{A,\boldsymbol{y}}$ is difficult!

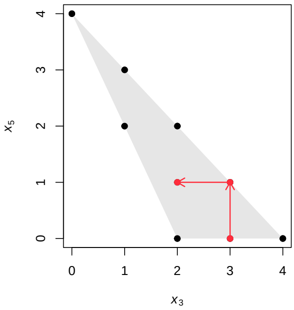
- Hard to ensure connectivity of sampler
- Hard to find helpful characterizition of fibre geometry ($\mathbb{Z}$-polytope)



Fibre Sampling

Sampling $\boldsymbol{x}$ from $\mathcal{F}_{A,\boldsymbol{y}}$ is difficult!

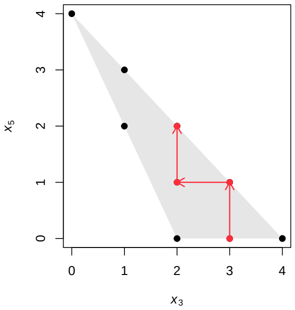
- Hard to ensure connectivity of sampler
- Hard to find helpful characterizition of fibre geometry ($\mathbb{Z}$-polytope)



Fibre Sampling

Sampling $\boldsymbol{x}$ from $\mathcal{F}_{A,\boldsymbol{y}}$ is difficult!

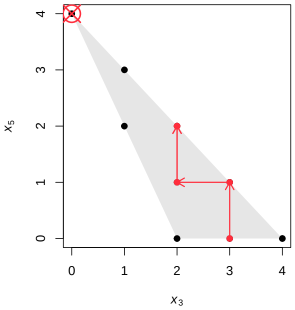
- Hard to ensure connectivity of sampler
- Hard to find helpful characterizition of fibre geometry ($\mathbb{Z}$-polytope)



Fibre Sampling

Sampling $\boldsymbol{x}$ from $\mathcal{F}_{A,\boldsymbol{y}}$ is difficult!

- Hard to ensure connectivity of sampler
- Hard to find helpful characterizition of fibre geometry ($\mathbb{Z}$-polytope)



Algebraic Statistics

Current state-of-the-art uses **Markov bases** found using **algebraic statistics** (Diaconnis & Sturmfels, 1998)

- Identifies $\boldsymbol{x}$ with monomial $\boldsymbol{t}^{\boldsymbol{x}} = t_1^{x_1} t_2^{x_2} \cdots t_r^{x_r}$, enabling application of computational tools from polynomial ring theory
- Beware – connectivity result in Tebaldi & West (1998) is wrong
- Promising practical methods use dynamic Markov bases (Dobra, 2012, H et al., 2021)

Tebaldi, C., & West, M. (1998). Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, **93(442)**, 557–573.

Dobra, A. (2012). Dynamic Markov bases. *Journal of Computational and Graphical Statistics*, **21(2)**, 496–517.

Hazelton, M. L., Mcveagh, M. R., & Van Brunt, B. (2021). Geometrically aware dynamic Markov bases for statistical linear inverse problems. *Biometrika*, **108(3)**, 609–26.

Continuous Traffic Models

Approximate traffic volumes as continuous variables

- Discretization errors worse in highly disaggregate settings
- E.g. dynamic OD matrix estimation with thin time slices

Normal models have been heavily used

- $\boldsymbol{x} \sim \mathrm{N}(\mu(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}))$
  - Often not recognized in the literature that μ and Σ should be linked
  - E.g. normal approximation to Poisson: $\mathrm{E}[\boldsymbol{x}] = \boldsymbol{\theta}$, $\mathrm{Var}(\boldsymbol{x}) = \mathrm{diag}(\boldsymbol{\theta})$
- $\boldsymbol{y} = A\boldsymbol{x} \Rightarrow \boldsymbol{y} \sim \mathrm{N}(A\mu, A\Sigma A^T)$

Normal Models Are Not a Panacea

If $\boldsymbol{y} \sim \mathrm{N}(A\mu(\boldsymbol{\theta}), A\Sigma(\boldsymbol{\theta})A^T)$ then likelihood function available

Log-likelihood typically highly non-linear in $\boldsymbol{\theta}$

- No closed form expression for maximum likelihood estimates (MLEs)
- MLEs may not be unique theoretically, or up to machine precision

Prior information incorporated via prior $f(\boldsymbol{\theta})$ regularizes, but...

- Integral $\int f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta}$ typically not available in closed form
- So must resort to MCMC methods
- MCMC samplers can struggle because of highly ridged posterior

Don’t Aggregate your Data!

Linear system $\boldsymbol{y} = A\boldsymbol{x}$ underdetermined in practice

Has led to unfortunate folklore about potential for statistical inference

- E.g. impossible to estimate OD matrix without informative prior

In practice we usually observe sequences of data through time

- $\boldsymbol{y}^{(t)} = A\boldsymbol{x}^{(t)}$ observed over periods $t = 1, 2, \ldots, N$

Assuming independence between observation windows, likelihood given by

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \prod_{t=1}^{N} f(\boldsymbol{y}^{(t)}|\boldsymbol{\theta}) \\
&= \prod_{t=1}^{N} \sum_{\boldsymbol{x}^{(t)} \in \mathcal{F}_{A\boldsymbol{y}^{(t)}}} f(\boldsymbol{x}^{(t)}|\boldsymbol{\theta})
\end{aligned}
$$

Don't Aggregate your Data! (part 2)

Example: Likelihood for data sequence using Poisson model is

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{N} \sum_{\boldsymbol{x}^{(t)} \in \mathcal{F}_{A\boldsymbol{y}^{(t)}}} \prod_{i=1}^{r} e^{-\theta_i} \theta_i^{x_i^{(t)}} / x_i^{(t)}!$$

In this case, and more generally, there is no sufficient statistic for $\boldsymbol{\theta}$ other than the full dataset $\{\boldsymbol{y}^{(t)} : t = 1, \ldots, N\}$ (Vardi, 1998)

Aggregating data to totals or means involves critical loss of information

---

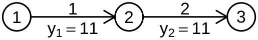
Vardi, Y. (1998) Bayesian inference on network traffic using link count data: Comment. *Journal of the American Statistical Association*, **93(442)**, 573–574.

Learning from the Dependence Structure

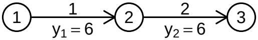
Example: Suppose $\{\boldsymbol{x}^{(t)} : t = 1, 2, \ldots, N\}$ are independent $\mathrm{Pois}(\boldsymbol{\theta})$ random variables for toy network from earlier example.

Scenario 1

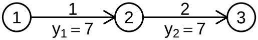
Observation 1



Observation 2

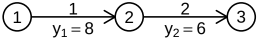


Observation 3

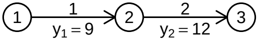


Scenario 2

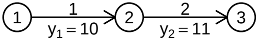
Observation 1



Observation 2



Observation 3

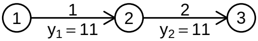


Learning from the Dependence Structure

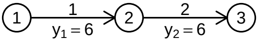
Example: Suppose $\{\boldsymbol{x}^{(t)} : t = 1, 2, \ldots, N\}$ are independent $\text{Pois}(\boldsymbol{\theta})$ random variables for toy network from earlier example.

Scenario 1

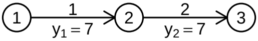
Observation 1



Observation 2



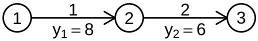
Observation 3



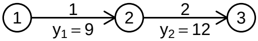
$$\boldsymbol{\theta} = (0, 10, 0)^{\mathsf{T}}$$

Scenario 2

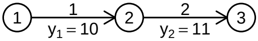
Observation 1



Observation 2



Observation 3



$$\boldsymbol{\theta} = (10, 0, 10)^{\mathsf{T}}$$

Identifiability in Network Tomography Problems

Even if system $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ is hugely underdetermined, parameters $\boldsymbol{\theta}$ typically statistically identifiable: i.e. $f(\boldsymbol{y}|\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta}')$ only if $\boldsymbol{\theta} = \boldsymbol{\theta}'$

Vardi (1996) – route flows $\boldsymbol{\theta}$ identifiable for Poisson models

H(2015) extension – if $f(\boldsymbol{x}|\boldsymbol{\theta})$ is identifiable then so is $f(\boldsymbol{y}|\boldsymbol{\theta})$ for any distribution supported on $\mathbb{Z}^r_{\geq 0}$

Singhal & Michailidis (2007) – identifiability based on first two moments

---

Vardi, Y. (1996). Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, **91(433)**, 365–377.

Singhal, H., & Michailidis, G. (2007). Identifiability of flow distributions from link measurements with applications to computer networks. *Inverse Problems*, **23(5)**, 1821–1849.

Hazelton, M.L. (2015). Network tomography for integer-valued traffic. *Annals of Applied Statistics* **9(1)**: 474–506.

Method of Moments for Network Tomography Problems

Can often base estimation on method of moments using just mean vector and covariance matrix

- Vardi (1996) studied estimation using this methodology
- Sufficient to ensure identifiability in sensible normal models
- Lev-Ari et al. (2023) extended to account for higher moments

Theoretically method of moments often inferior to likelihood methods

Practical performance not encouraging in my experience

Vardi, Y. (1996). Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, **91(433)**, 365–377.

Lev-Ari, H., Ephraim, Y., & Mark, B. L. (2023). Traffic rate network tomography with higher-order cumulants. *Networks*, **81(2)**, 220–234.

R packages for Network Tomography

`github.com/awblocker/networkTomography`

Developed originally for electronic communication networks; see Airoldi and Blocker (2012)

`github.com/MartinLHazelton/LinInvCount`

General package for linear inverse problems with count data; see H et al. (2021)

Airoldi, E. M., & Blocker, A. W. (2013). Estimating latent processes on a network from indirect measurements. *Journal of the American Statistical Association*, **108(501)**, 149–164.

Hazelton, M. L., Mcveagh, M. R., & Van Brunt, B. (2021). Geometrically aware dynamic Markov bases for statistical linear inverse problems. *Biometrika*, **108(3)**, 609–26.

Past Lessons

Statistical linear inverse problems are ubiquitous when conducting inference for network-based traffic models

- Route traffic volumes pivotal, but never directly observed

Transfer of fundamentally common ideas between statistics, transport, and electronic engineering, has often been slow

- Plenty to learn from each other

Standard statistical theory applies, and provides useful guidance regarding practical performance of methods

- Theory beats anecdote or folklore!

Limited assessment and comparison of competing methods

- In part due to lack of software and publicly shared data

Future Challenges

Computational implementation and testing:

- Existing R packages for network tomography are limited
- Catalogue of publicly available real data sets would be great

Informative priors and data synthesis:

- Model based, e.g. H. & Najim (2024).
- Correcting bias in e.g. trajectory data (Parry & H, 2012)

Efficient MCMC for Bayesian inference in normal models

- Huge literature on samplers for awkward high-dimensional normal distributions
- Might enable inference for large network tomography problems

---

Parry, K. and Hazelton, M.L. (2012). Estimation of origin-destination matrices from link counts and sporadic routing data. *Transportation Research Part B*, **46**, 175–188.

Hazelton, M.L., & Najim, L. (2024). Using traffic assignment models to assist Bayesian inference for origin-destination matrices. *Transportation Research Part B* **186**, 103019.

Thanks to ...

Collaborators

David Watling (U. Leeds)
Lara Najim (former summer student, U. Otago)
Linus Fromm (current PhD, U. Otago)
Mike McVeagh (former PhD, Massey U.)
Rina Parry (former PhD, Massey U.)
Sam Lee (former summer student, U. Otago)
Timothy Bilton (former summer student, Massey U.)

Funding

MARSDEN FUND
TE PŪTEA RANGAHAU A MARSDEN

ROYAL SOCIETY TE APĀRANGI