Stochastic modelling of urban travel demand

MRes Project

Word count: 11331

Ioannis Zachos

Hughes Hall College

Academic supervisor: Professor Mark Girolami

Academic co-supervisor: Dr. Theodoros Damoulas

Industrial supervisor: Dr. Gerard Casey

This dissertation is submitted for the degree of

Masters of Research in Future Infrastructure and Built Environment

Department of Engineering

University of Cambridge

August 2020

Contents

Declaration

To my two beloved grandmothers, Sofia and Antonia, for their unconditional love.

Στις γιαγιάδες μου, Σοφία και Αντωνία, για την ανιδιοτελή τους αγάπη.

Acknowledgements

I would like to thank Professor Mark Girolami for helping me shape the research direction of this thesis and supporting me every step of the way. I would also like to thank Dr. Theodoros Damoulas for his invaluable contributions in steering this project in fruitful directions. I am also thankful to Dr. Gerard Casey for his valuable feedback throughout the duration of the project and for making me feel welcome in Arup's city modelling team. I would like to thank Dr. Louis Ellam, Professor Mark Girolami, Professor Grigoris Pavliotis and Professor Sir Alan Wilson for sharing their codebase on GitHub.

Additionally, I would like to express my gratitude towards my parents and aunt for believing in me and supporting me throughout my life. Finally, I would like to thank my love, Christina, for her patience and support.

Abstract

Modelling of evolution of urban travel demand is fundamental for urban planners and policy makers to assess the spatial demand for transportation capacity and decide on appropriate interventions. We follow the approach of (Ellam et al., 2018) and introduce a novel application of spatial interaction models and a mathematical evolution of their dynamics to urban travel demand. We exploit economic structure characteristics (e.g. employment) to inform travel demand between a set of origin and destination locations. The economic structural variables are described by a potential function defined in terms of utility and cost functions. We also use a system of stochastic differential equations to model temporal travel demand evolution. We calibrate our model using a Bayesian framework that formally incorporates uncertainty involved in the process due to random noise or unexplained events and propagates it into parameter inference. We apply our model to London's 2001 commuter flow data and find that we can adequately reconstruct the flow matrix only through the use of employment data as a latent force driving travel demand. We compute a Euclidean distance-based and transportation network-based cost matrices and find that the latter is marginally better at explaining travel demand. Finally, we overcome computational challenges arising from a doubly intractable posterior by applying appropriate Markov Chain Monte Carlo schemes for various noise regimes.

Keywords: stochastic travel demand modelling, spatial interaction modelling, origin-destination estimation, Bayesian inverse problems, stochastic differential equations, urban transportation

List of Figures

List of Tables

List of Algorithms

Nomenclature

In this thesis, we adopt standard notation and denote multidimensional objects in bold letters, say $\mathbf{x}$. Lowercase letters are used to denote scalars or vectors (x or $\mathbf{x}$, respectively) while capital letters denote matrices or random variables (the difference between the two is made clear in the context in which the mathematical objects are provided). The i-th element of a vector $\mathbf{x}$ is written as x_i while the (i, j)-th element of a matrix is indexed as X_{ij}. Additional explanations on abbreviations and notation is provided below.

Table 1: Acronyms and their corresponding descriptions.

Acronym	**Description**
AIS	Annealed importance sampling
DCM	Discrete choice model
HMC	Hamiltonian Monte Carlo
i.i.d.	Independent and identically distributed
L-BFGS	Limited memory Broyden-Flether-Goldfarb-Shanno
MCMC	Markov Chain Monte Carlo
MAP	Maximum a posteriori
MLE	Maximum likelihood estimate
n.c.	Normalising constant
OD	Origin-destination
ODE	Ordinary differential equation
pdf	Probability density function
p.f.	Potential function
PR	Poisson regression
PT	Parallel tempering
SDE	Stochastic differential equation

SIM	Spatial interaction model
SRMSE	Standardised root mean square error

Table 2: Commonly used notation and its corresponding description.

Symbol	**Description**
O_i	Origin supply from origin i
D_j	Destination demand from destination j
W_j	Latent destination size of destination j
X_j	Log latent destination size of destination j
c_{ij}	Cost of travelling from origin i to destination j
T_{ij}	Flow from origin i to destination j
U_{ij}	Utility function of travelling from origin i to destination j
N	Number of origin locations
M	Number of destination locations
α	Attractiveness parameter of spatial interaction model
β	Inconvenience of travel parameter of spatial interaction model
ϵ	Responsiveness/scaling parameter
λ	Standard deviation of observation noise or Lagrange multiplier
σ	Standard deviation
$\gamma = \frac{1}{2\sigma^2}$	Inverse temperature parameter of the SIM
κ	Job competitiveness (number of people per unit number of jobs) parameter
δ	additional utility parameter
$\boldsymbol{\theta} = (\alpha, \beta)$	Parameter vector of the SIM
ξ	Gumbel distributed random variable
ρ_∞	Stationary/equilibrium distribution of the Harris-Wilson SDE
$V_{\boldsymbol{\theta}}(\cdot)$	Potential function for given parameter choice
$\mathbf{J}$	Jacobian matrix
$\mathbf{H}$	Hessian matrix

$C^k(S)$	Set of k times continuously differentiable functions in S
$C(p)$	Cost constraint evaluated for a choice of parameter p
$\mathcal{H}$	Shannon's entropy function
$\mathcal{L}$	Lagrange multiplier objective function
$\mathcal{O}(\cdot)$	Big Oh notation for space and computational complexity
$\mathcal{Z}$	Normalising constant of Boltzmann-Gibbs measure
$\mathbb{E}[\cdot]$	Expectation
$\Gamma(\cdot)$	Gamma function
$\mathbb{R}^{\mathbb{M}}$	M-dimensional plane of real numbers
$\mathbb{R}^{\mathbb{M}}{}_{>0}$	positive M-dimensional plane of real numbers
$\mathbb{P}$	Probability or probability mass
$\sim$	Distributed as
$\forall$	For all
$\exists$	There exists
$\|\|$	'Given'. Used to denote conditionality
∞	Infinity
$\int$	Integral
∂	Partial derivative
d	Ordinary derivative
∇	Gradient of multidimensional object
Δ	Laplace operator of multidimensional object
log	Natural logarithm
lim	Limit
lim inf	Limit of infimum of a set

Chapter 1

Introduction

Transportation systems are a critical piece of infrastructure in any urban environment. They can stimulate economic growth by boosting the productivity of supply chains and provide access to people, goods and services. The UK Government is planning to spend around £90 billion on transport infrastructure in the next five years to accommodate some of the growing demand for travel (Marsden et al., 2018). In order for such investment to be impactful, it has to be targeted to communities who need it the most. One way of ensuring that appropriate investment decisions are made is to use a fine-grained view of **travel demand** updated on a frequent basis. Travel demand has been increased due to urbanisation, demographic change, and economic prosperity. At the same time, environmental restrictions on emissions imposed through legislation are likely to decelerate travel demand growth at least through certain modes of transport (e.g. single-occupancy cars).

Robust and scalable travel demand evolution modelling can therefore provide a useful decision making tool to urban planners and policy makers. A clear view of spatially distributed travel demand can unlock opportunities for governments to invest in building new or upgrading existing infrastructure to accommodate the needs of its citizens as well as increase availability of various modes of transport (e.g. schedule more frequent journeys from/to a given station). Purely increasing supply without precise knowledge of the expected demand can be a poor strategy as doing so is costly and can lead to vicious cycles of reasoning as the level of demand is influenced by the level of infrastructure supply (Kim and Oleson, 2007). Alternatively, localised intervention policies can be devised to accommodate the anticipated demand while relieving congestion and reducing emissions (e.g. introducing congestion charges). These policies can have profound impacts to social and economic development of certain regions and

therefore they have to be proposed after careful assessment of the risks involved. This type of assessment can be facilitated by modelling frameworks that account for known and unknown uncertainties.

1.1 Related work

In transportation planning and forecasting literature, the most-adopted framework of travel demand modelling is the four step model illustrated in Figure 1.1 or its variants. This framework treats the transportation network as a set of origins and destinations in-between which there is a flow of people. It consists of four steps executed iteratively until 'convergence' is reached. The first step is *trip generation*, which estimates the number of trips generated at each origin or destination. The second step is *trip distribution*, whose output is the matching of origins to destinations with an associated flow. What follows is modal split, which separates flows by mode of transport and finally *trip assignment*, which assigns a route of transport people use to reach their desired destinations.

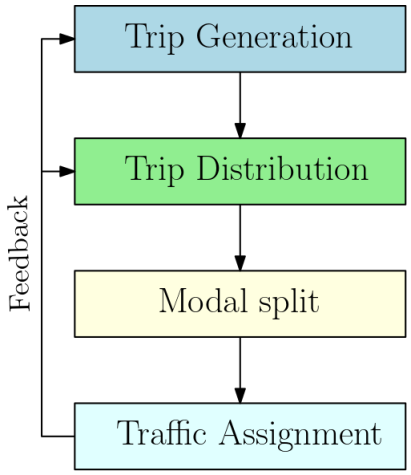


Figure 1.1: The four step travel demand forecasting modelling framework.

According to Great Britain's 2019 transport statistics there were "8.3 billion passenger journeys on public transport vehicles in 2018/19" (Department for Transport, 2019). There are broadly two ways of modelling such a large scale problem in the context of the four step transportation modelling framework. One is to adopt a microscopic view that examines decisions being made on an individual basis and another is to consider a macroscopic view of decisions being made on a more coarse level. Naturally, microscopic behaviour can be aggregated to derive macroscopic patterns.

Discrete choice models (DCM) is an important class of econometric models used to describe individual (agent) behaviour defined by a collection of mutually exclusive, exhaustive and finite set choices (Train, 2002, p.3, 16). The underlying assumption is that the continuum of choices an individual can make can be discretised with a small loss of information, which relies on effective partitioning of the choice space. The choice model is defined in terms of a random utility function, which allows for a probabilistic treatment of the problem. A widely accepted choice for the underlying probability distributions in econometric literature is the multinomial logit model (McFadden et al., 1973). This model suffers from its independent noise assumption, which is unrealistic in scenarios when choices made at time t are dependent on choices made at time $t - k$. Another example where this assumption fails is when there are common unobserved (random) factors affecting multiple choices (Train, 2002, p.23). For instance, an individual may dislike the use of the tube because they prefer avoiding public transport, which implies that the random effects of choice of mode are correlated for that individual. The generalised extreme value model has been proposed to alleviate the impact of the independence assumption. These models are based on the generalisation of the extreme value distribution and account for dependencies in the error structure (Small, 1994). Another popular choice of a more flexible DCM is the generalised random utility model (Walker and M. Ben-Akiva, 2002) that also allows for flexible error structures by introducing latent choice sets.

In the context of the four-step travel demand modelling framework, DCMs have been applied on trip, tour and activity-based approaches, the most commonly accepted of which is the last. Activity-based approaches model the number, purpose and sequence of time-ordered activities agents make while maximising their expected utility every time a decision is made (M. E. Ben-Akiva and Bowman, 1998). As a result, activity-based models generate a much larger choice set, which makes steps 2-4 in Figure 1.1 computationally intensive. DCMs also suffer from spatio-temporal aggregation biases (Baltas and Doyle, 2001) induced from aggregating individual choice over a large population. This bias is amplified if the independent error and/or choice set assumptions are violated. Moreover, parameter calibration of DCMs requires a large volume of origin-destination (OD) matrix (i.e. flow) data often available on

an individual basis (e.g. travel diaries), which is sensitive and not open-access and/or readily available. Unlike DCMs, conjoint preference and choice models rely on experimental design data for parameter calibration. However, these models do not overcome the problem of city-wide flow data availability and can introduce additional approximation errors.

We also note that sequentially and separately executing the four steps in Figure 1.1 may be unrealistic. The order of execution ignores the dynamic evolution of travel demand and a concurrent execution of some of these steps may be more represenative and efficient. Also, the evolution of travel demand is monitored by iteratively running the four steps until 'convergence' is achieved. Although this approach provides a temporal view of travel demand, it entails the computational costs of rerunning a computational cumbersome procedure multiple times, which is exacerbated if the equilibrium characteristics are unknown. Therefore, the conventional four step framework does not constitute a single unifying approach to travel demand modelling (Mladenovic and Trifunovic, 2014). Despite the fact that DMCs are interpretable due to the intuitive nature of the utility-maximising argument, their aforementioned shortcomings limits their scalability potential. Therefore, they cannot make a compelling case as a candidate decision making tool used by policy makers and urban planners.

A class of models that overcomes some of the shortcomings of disaggregate DCMs are the aggregated spatial interaction models (SIMs) first introduced by (A.G. Wilson, 1967). SIMs assume that flows of people are a result of production effects from origin locations, attraction effects from destination locations and the (in)convenience of travel between the two. These effects are described by the gravity model. Flow estimation is achieved through statistical optimisation procedures that are entropy-maximising or information-minimising in information theory terms. The theory developed in (Harris and A. G. Wilson, 1978) links the equilibrium position of the evolution of production effects to an extension of the Lotka-Volterra ordinary differential equations (ODEs) known as the Harris and Wilson model. The analogy arises from the fact that the production sizes (species populations) are competing for the same finite origin supplies (resources). Flows can be expressed in terms of urban demographic and economic features, such as population and employment data, and spatial interaction. However, the Harris and Wilson model is deterministic whose equilibrium is determined by its initial

condition and the discontinuities induced in the ODE due to small parameter changes (Dearden and Alan Wilson, 2011). The novel work of (Ellam et al., 2018) makes a stochastic treatment of the Harris and Wilson model by formulating a well-defined set of stochastic differential equations (SDEs) to account for uncertainty arising from the dynamic nature of urban systems.

1.2 Current approach

We extend the work of (Ellam et al., 2018) by applying the aggregated SIM and system of SDEs to travel demand and its evolution. The singly constrained SIM is well-posed for travel demand models of economic structure (Batten and Boyce, 1987) as it is often the case that the supply of people (population) at each origin is routinely available whereas the demand for destinations is unknown. Travel demand for a particular destination is driven by socio-economic features such as job availability whose inference is easier than that of travel demand. The aggregated SIM avoids aggregation biases and can be calibrated without flow data. By modelling the stochastic evolution of job availability (and therefore travel demand) in a unified framework allows us to avoid iteratively updating the four-step model and therefore achieve significant computational savings. The reconstructed OD matrix can also be updated efficiently every time the latent posterior is updated.

Moreover, the SDEs' drift functions are defined as the gradient flow of a potential function which in turn is a function of the travel demand and destination job availability. The potential function encodes constraints on people travelling, job availability and cost of travel by making use of the entropy-maximising argument (Alan Wilson, 2010). The Bayesian framework of SIM and SDE model calibration allows us to incorporate random unknown effects affecting travel demand and propagate the uncertainty into our inference. We exploit the limiting (equilibrium) distribution of the SDEs, which is a Boltzmann-Gibbs measure to define a likelihood over the latent sizes (job availability) driving travel demand. Several sophisticated MCMC schemes and computational tricks are employed to infer the joint SIM parameter and latent size posterior while dealing with a doubly intractable likelihood. We argue that this approach can be used as a basis for a decision-making tool for urban planners and policy makers.

1.2.1 Synopsis

This thesis aims to answer the following research question:

How well can the origin-destination matrix of Londoner commuter movements with unknown destination demand be reconstructed under the influence of uncertainty?

Chapter 1 introduced the societal need for travel demand monitoring and forecasting and provided a critical review of existing work in travel demand modelling. Limitations of previous approaches were identified and a new approach was suggested based on the work of (Ellam et al., 2018) that addressed the shortcomings of existing approaches. In Chapter 2 the spatial interaction model is mathematically formulated and its relation to the Harris and Wilson ODEs is established. The stochastic treatment of the systems of these ODEs is outlined and the equilibrium distribution of the state variable is derived. Additionally, a potential function is derived and modified to ensure a well-defined system of SDEs is defined. At the end of the chapter we provide a derivation of the Poisson regression model, which is used as a benchmark model in deterministic model calibration. In Chapter 3 we describe the Bayesian model calibration framework, the computational strategy we employ to simplify the inversion of the system of SDEs, and the posterior inference MCMC schemes. Chapter 4 is devoted to a London commuter case study that we apply our model on. We include implementation details such as cost matrix computation and discuss our key results when the model is applied to zero, low and high-noise scenarios. We also validate the reconstructed flow matrix with the actual origin-destination matrix. Finally, in Chapter 5 we conclude our findings and suggest directions for future research that will be carried out during the PhD. In Appendices, the reader can find relevant proofs of key theoretical results necessary to ensure our model is well-defined.

1.2.2 Contributions

We list the contributions made in this thesis. Specifically, we:

1. Introduce a novel application of stochastic spatial interaction modelling to travel demand;

2. Compute an informative travel cost matrix based on London's transportation network and benchmark its effect on flow matrix reconstruction accuracy against the conventional Euclidean distance-based cost matrix;

3. Employ the Poisson regression model to verify our model's results in the deterministic case;

4. Validate the estimated flow matrices for zero, low and high-noise regimes using the 2001 London commuter flow matrix;

5. Prove theoretical results to ensure that our model is well-defined (See Appendices A.1,A.2,C.3,C.4);

6. Derive estimates for the overall computational complexity of the calibration framework;

7. Develop a well-documented and extensible codebase of our model released on this GitHub repository.

Chapter 2

Urban travel demand modelling

2.1 Stochastic formulation of travel demand evolution

Define the number of jobs available in a transportation network as the vector of sizes $\mathbf{W} \coloneqq \{W_1, \ldots, W_M\} \in \mathbb{R}^M_{>0}$. For convenience, let the log-jobs be $\mathbf{X} \coloneqq \{X_1, \ldots, X_M\} \in \mathbb{R}^M_{>0}$, where $X_j = \exp(W_j)$ for each $j \in \{1, \ldots, M\}$.

Let the flow of people between origin i and destination j be denoted by $T_{ij} \geq 0$ (see Figure 2.1). It is assumed that there are N origins and M destinations and therefore a total of NM flows. For a singly constrained transportation system, the supplies generated by the N origins are

$$O_i = \sum_{j=1}^{M} T_{ij}, \quad i = 1, \ldots, N, \tag{2.1}$$

and the demands generated for the M destinations are

$$D_j = \sum_{i=1}^{N} T_{ij}, \quad j = 1, \ldots, M. \tag{2.2}$$

Origin supplies are known whereas the destination demands are unknown and have to be determined. The demand for destination zones is governed by the job availability in that destination; the more jobs are available the more people are travelling/commuting to that destination. Between two destinations zones of similar employability characteristics, people are assumed to prefer closer zones of lower transportation cost. Hence, a third constraint is added to reflect the finiteness of the total expenditure on transport:

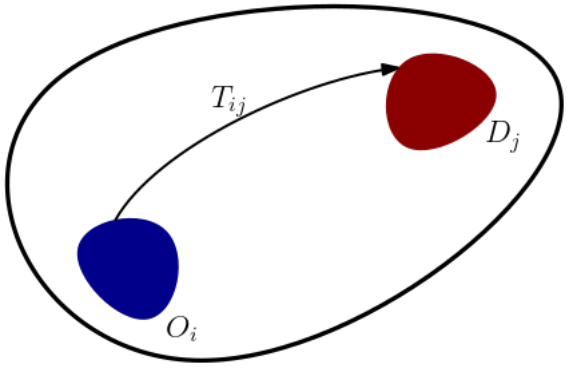


Figure 2.1: Illustration of spatial interaction in a transportation network occurring due to a flow T_{ij} of people between origin i of known supply and destination j of unknown size and demand.

$$C = \sum_{i=1}^{N} \sum_{j=1}^{M} T_{ij} c_{ij}, \quad i = 1, \ldots, N, \;\; j = 1, \ldots, M. \tag{2.3}$$

A final constraint is imposed on the finite 'benefit' (i.e. work) W_j people gain from travelling to destination j:

$$B = \sum_{i=1}^{N} \sum_{j=1}^{M} T_{ij} \log(W_j), \quad j = 1, \ldots, M. \tag{2.4}$$

In transportation applications, the origin supply can be assumed to be the residents of origin zones whereas the destination demands are often unknown and have to be inferred through other means i.e. variables. By following a maximum entropy argument subject to the constraints in (2.1)-(2.4) the resulting flows become

$$T_{ij} = O_i \frac{W_j^{\alpha} \exp(-\beta c_{ij})}{\sum_{k=1}^{M} W_k^{\alpha} \exp(-\beta c_{ik})}, \tag{2.5}$$

where $\alpha, \beta > 0$ are attractiveness and cost scaling parameters and $c_{ij} \geq 0$ denotes the cost of travelling[1]. The detailed derivation of (2.5) can be found in Appendix A.1. We use (2.5) to reconstruct the full origin-destination (flow) matrix.

[1]The cost of travelling is abstractly defined here. A Euclidean distance between origins and destinations is commonly used as a proxy for the cost function. In Chapter 4, we compute a shortest path distance between any origin-destination pair in London's transportation network.

The destination zone will grow or contract in size (i.e. job availability) depending on whether travel demand for them has been fulfilled or not, respectively. It is therefore reasonable to expect that the urban transportation's destination sizes will reach an equilibrium in some degree, after which their sizes will not change. A suitable model that models the evolution of these dynamics in time is the Harris and Wilson model (Harris and A. G. Wilson, 1978), which is described by the following system of M ordinary differential equations (ODEs)

$$\frac{\mathrm{d}W_j}{\mathrm{d}t} = \epsilon W_j(D_j - \kappa W_j), \quad \mathbf{W}(0) = \mathbf{w_0}, \tag{2.6}$$

where $\epsilon > 0$ is a responsiveness parameter and $\kappa > 0$ is the 'cost' of accommodating hiring/accommodating one more person in destination j and can be interpreted as a job competition term. The difference $D_j - \kappa W_j$ is the job capacity term and refers to a destination's ability to accommodate all its incoming demand.

2.2 Stochastic evolution of travel demand

A generalisation of the Harris and Wilson model is the following stochastic differential equation (SDE) with multiplicative noise

$$\mathrm{d}W_j = \epsilon W_j(D_j - \kappa W_j)\,\mathrm{d}t + \sigma W_j \circ \mathrm{d}B_j, \quad \mathbf{W}(0) = \mathbf{w_0}, \tag{2.7}$$

where $\mathbf{B}$ is the M-dimensional Brownian motion and $\sigma > 0$ a volatility parameter. The '$\circ$' operator arises from a Stratonovich SDE formulation (Pavliotis, 2014, p.59). The stochastic dynamics described in (2.7) is an overdamped Langevin diffusion. In the noisy regime, the latent destination sizes' evolution is governed by the net capacity term in the drift function perturbed by Gaussian noise with standard deviation $\sigma\sqrt{\delta t}$. By applying the variable transformation $X_j = \log(W_j)$, the ODEs in (2.6) can be expressed as a gradient flow. Let $V : \mathbb{R}^M \to \mathbb{R}$ be a potential function and its gradient $\nabla V : \mathbb{R}^M \to \mathbb{R}^M$. Define $\gamma := 2\sigma^{-2}$ and reformulate the SDE as

$$\mathrm{d}\mathbf{X} = -\epsilon^{-1}\nabla V(\mathbf{X})\,\mathrm{d}t + \sqrt{2\gamma^{-1}}\,\mathrm{d}\mathbf{B}, \quad \mathbf{X}(0) = \mathbf{x_0}. \tag{2.8}$$

Exploiting the fact that $\frac{\mathrm{d}X_j}{\mathrm{d}t} = \frac{1}{W_j}\frac{\mathrm{d}W_j}{\mathrm{d}t}$ and substituting it together with (2.1)-(2.5) in (2.6) yields

$$\begin{aligned}
\epsilon^{-1}\nabla V(\mathbf{x}) &= -\sum_{i=1}^{N}\left(O_i\frac{\exp(\alpha x_j - \beta c_{ij})}{\sum_{k=1}^{M}\exp(\alpha x_k - \beta c_{ik})}\right) + \kappa\exp(x_j) \\
\therefore \epsilon^{-1}V(\mathbf{x}) &= -\alpha^{-1}\sum_{i=1}^{N}O_i\log\left(\sum_{j=1}^{M}\exp(\alpha x_j - \beta c_{ij})\right) + \kappa\sum_{j=1}^{M}\exp(x_j),
\end{aligned} \tag{2.9}$$

where $\mathbf{x}$ is a realisation of the random variable $\mathbf{X}$.

Assuming that $\mathbf{x_0}$ is a random variable with probability density function (pdf) $\rho_0(\mathbf{x})$, then $\rho(\mathbf{x}, t)$ is the pdf of $\mathbf{X}(t)$ and the solution to the initial value problem for the corresponding Fokker-Planck equation (Pavliotis, 2014, p.109-110):

$$\frac{\partial\boldsymbol{\rho}(\mathbf{x},t)}{\partial t} = \nabla\cdot\big(\rho(\mathbf{x},t)\nabla V(\mathbf{x})\big) + \gamma^{-1}\Delta\rho(\mathbf{x},t), \quad \rho(\mathbf{x},0) = \delta(||\mathbf{x}-\mathbf{x_0}||), \tag{2.10}$$

where $\rho(\mathbf{x}, t) \in \mathbb{R}^M \times (0, +\infty)$, and $\delta > 0$. The solution to (2.10) is very challenging to obtain especially in the given high dimensional setting. Under some smoothness conditions outlined in Appendix B (Pavliotis, 2014, p.110), the unique invariant distribution of $\rho(\mathbf{x}, t)$ is the Boltzmann-Gibbs measure given by

$$\rho_\infty(\mathbf{x}) = \frac{1}{Z}\exp\big(-\gamma V(\mathbf{x})\big) \quad Z \coloneqq \int_{\mathbb{R}^M}\exp\big(-\gamma V(\mathbf{x})\big)\,\mathrm{d}\mathbf{x}. \tag{2.11}$$

Note that the normalising constant Z in (2.11) is finite since the potential function in (2.9) satisfies the smoothness criterion (B.1) and therefore does not yield a well-defined probability. We follow the approach by (Ellam et al., 2018) and modify the potential function to include a confining term that ensures that the systems of SDEs in (2.11) is well-defined. Although the confining criterion is sufficient for the SDE to be well-defined, controlling the rate of convergence criterion in (B.2) is crucial in ensuring that the convergence occurs sufficiently fast.

2.3 Potential function decomposition

The Boltzmann-Gibbs measure can also be derived from a maximum entropy argument (Lasota and Mackey, 1994; Alan Wilson, 2010). The derivation can be found in Appendix A.2. This view allows for a natural interpretation of the potential function's terms as economic constraints. A potential function with three components is considered:

$$\epsilon^{-1}V(\mathbf{x}) = V_{\text{Utility}}(\mathbf{x}) + \kappa V_{\text{Cost}} + \delta V_{\text{Additional}}(\mathbf{x}), \tag{2.12}$$

where $\delta > 0$ is an additional parameter. The utility potential describes the financial incentives (jobs) emerging from utility-maximising choices, the cost potential imposes restrictions on potential benefit that can be gained from travelling and the additional potential can be interpreted as the effect of transportation network policies or background travel demand. Let $\mathbf{X} \in \mathbb{R}^M$ be a random variable that is subject to the following constraints

$$\left.\begin{aligned} \mathbb{E}[V_{\text{Utility}}(\mathbf{x})] &= C_{\text{Utility}}, \\ \mathbb{E}[V_{\text{Cost}}(\mathbf{x})] &= C_{\text{Cost}}, \\ \mathbb{E}[V_{\text{Additional}}(\mathbf{x})] &= C_{\text{Additional}}, \end{aligned}\right\} \tag{2.13}$$

where $C_i \in \mathbb{R}$. Then, the maximum entropy distribution of $\mathbf{X}$ can be written as the Boltzmann-Gibbs measure whose density is given by (2.11) and the corresponding potential function is (2.12).

2.3.1 Utility potential

The work of (Williams, 1977) highlighted the correspondence between the entropy-maximising or information-minimising argument and the utility-maximisation argument, which was popularised in econometric literature (Anas, 1983) and (Jong et al., 2007). Therefore, we motivate a utility-maximisation term defined by the utility potential, where each individual aims to maximise his/her access to a big job market while minimising the total cost they incur when travelling to that market (i.e. destination).

A suitable candidate for a utility potential which reflects a transportation user's benefit of travelling to destination zone j is given by consumer surplus. Consumer surplus is the area under the demand curve (Williams, 1977):

$$\begin{aligned} V_{\text{utility}}(\mathbf{x}) &= -\int_{\mathbf{x}_0}^{\mathbf{x}} \left(D_1(\mathbf{x}'), \ldots, D_M(\mathbf{x}')\right) \mathrm{d}\mathbf{x}' \\ &= -\alpha^{-1} \sum_{i=1}^{N} O_i \log\left(\sum_{j=1}^{M} \exp(U_{ij}(x_j))\right) + \text{constant} \end{aligned} \tag{2.14}$$

where the line of integration is defined along a path in log-size-space and the deterministic utility function is defined as

$$U_{ij} \coloneqq \alpha x_j - \beta c_{ij} \quad \forall\, (i,j) \in \{1,\ldots,N\} \times \{1,\ldots,M\}.$$

The generalised version of (2.14) is derived in Appendix C.1. The utility function U_{ij} can be obtained from random utility maximisation. Define the stochastic utility function for a choice made at origin i

$$\tilde{U}_{ij} = U_{ij} + \xi_{ij}, \tag{2.15}$$

where ξ_{ij} are independent and identically distributed Gumbel random variables with zero mean and scale one. Then, under the random utility maximisation framework the expected utility of a unit flowing from origin i to destination j is

$$\mathbb{E}\left[\max_{1\le j\le M} \tilde{U}_{ij}(x_j)\right] = \log\left(\sum_{j=1}^{M} \exp\left(U_{ij}(x_j)\right)\right) + c, \tag{2.16}$$

where $c \approx -\log(\log(2))$ is the Euler-Mascheroni constant (Chang, 2015). A detailed proof is provided in Appendix C.2. The utility potential then becomes

$$V_{\text{Utility}}(\mathbf{x}) = \alpha^{-1} \sum_{i=1}^{N} O_i \mathbb{E}\left[\max_{1\le j\le M} \tilde{U}_{ij}(x_j)\right] + c, \tag{2.17}$$

Equation (2.16) depicts a connection between the aggregated SIM and the disaggregated DCMs. Note also that in the limit of $\alpha \to 0$ it holds that $V_{\text{Utility}}(\mathbf{x}) \to \infty$ (i.e. the potential is

non-constant). Also, the utility potential is tightly bounded:

$$\alpha^{-1}\sum_{i=1}^{N} O_i \left\{ \max_{1\leq j\leq M} U_{ij}(\mathbf{x}) \right\} \leq V_{\text{Utility}}(\mathbf{x}) \leq \alpha^{-1}\sum_{i=1}^{N} O_i \left\{ \max_{1\leq j\leq M} U_{ij}(\mathbf{x}) + \log(M) \right\} \tag{2.18}$$

The derivation of the bounds is shown in Appendix C.3. The limit of the utility potential as $x_j \to -\infty$ for some $j \in \{1, \ldots, N\}$ may still be finite, which violates criterion (B.1) in Appendix B. Therefore, an additional term needs to be added to prevent destination "zones from collapsing from a lack of activity" (Ellam et al., 2018).

2.3.2 Cost potential

The purpose of the cost potential is to prevent each destination zone from growing uncontrollably and becoming too large in size. The cost potential has to be an increasing function to reflect the fact that the running costs of a destination increases with the number of people gathered at that destination. With the existing potential function (2.9) in mind, a suitable candidate for the cost potential is

$$V_{\text{Cost}}(\mathbf{x}) = \sum_{j=1}^{M} \exp(x_j), \tag{2.19}$$

with

$$\frac{\partial V_{\text{Cost}}(\mathbf{x})}{\partial x_j} = \exp(x_j).$$

2.3.3 Additional potential

The additional potential term must satisfy criteria (B.1) and (B.2) so that

$$\lim_{x_j \to -\infty} V(\mathbf{x}) = +\infty,$$

and the rate of growth at infinity must be sufficiently high. Contrary to the cost potential, the additional potential term ensures that every destination zone receives non-zero demand, which prevents the zone's size to collapse from lack of activity. A suitable and mathematically convenient potential is

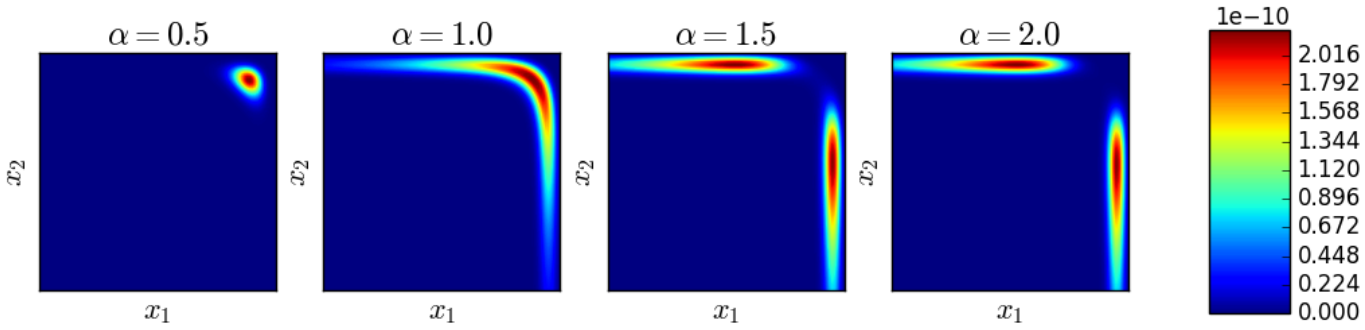


Figure 2.2: Illustration of potential function as appearing in the numerator of (2.11) for two competing destination zones ($\mathbf{x} \in \mathbb{R}^2_{<0}$) plotted for four different configurations of α and $\beta = 0.00143$.

$$V_{\text{Additional}}(\mathbf{x}) = \sum_{j=1}^{M} x_j, \tag{2.20}$$

with

$$\frac{\partial V_{\text{Additional}}(\mathbf{x})}{\partial x_j} = 1 \quad \forall\, j.$$

2.4 Model adjustments

Based on the potentials defined in the three previous sections, the modified potential function becomes

$$\epsilon^{-1} V(\mathbf{X}) = \underbrace{-\alpha^{-1} \sum_{i=1}^{N} O_i \log\left(\sum_{j=1}^{M} \exp(\alpha x_j - \beta c_{ij})\right)}_{\text{Utility}} + \underbrace{\kappa \sum_{j=1}^{M} \exp(x_j)}_{\text{Cost}} - \underbrace{\delta \sum_{j=1}^{M} x_j}_{\text{Additional}}, \tag{2.21}$$

which satisfies both criteria in Appendix 2.13. We note that the computational cost of evaluating (2.21) for a given choice of parameters and latent log-sizes is $\mathcal{O}(NM)$. The Stratonovich SDE is adjusted to

$$\mathrm{d}W_j = \epsilon W_j (D_j - \kappa W_j + \delta)\,\mathrm{d}t + \sigma W_j \circ \mathrm{d}B_j, \quad \mathbf{W}(0) = \mathbf{w_0}. \tag{2.22}$$

Equation (2.22) corresponds to the original SDE in (2.7) with a constant positive effect on the drift function. The original deterministic Harris and Wilson model can be obtained in the

limit $\delta, \sigma \to 0$. The stationary points of (2.22) are obtained by solving

$$\sum_{j=1}^{M} O_i \frac{\exp(\alpha x_j - \beta c_{ij})}{\sum_{k=1}^{M} \exp(\alpha x_k - \beta c_{ik})} = \kappa W_j - \delta \quad j = 1, \ldots, M, \tag{2.23}$$

which in the case of $\delta \to 0$ coincide with the stationary points of the deterministic ODEs in (2.6). Despite the similarities in behaviour between the ODE and SDE in low-noise regimes, their asymptotic behaviours differ significantly for large enough σ. The deterministic model will converge to a stable point that is heavily determined by the initial condition (Dearden and Alan Wilson, 2011). However, the uniqueness of the stable point is governed by the convexity of the potential function, as illustrated in (C.4). Also, the work of (Dearden and Alan Wilson, 2011) highlighted the fact that the potential function may exhibit irregularities caused due to discontinuities existing between even slightly different parameter configurations. On the contrary, the stochastic model will converge to a statistical equilibrium independent of the initial condition. In depth of time ($t \to +\infty$), the stochastic model will spend more time at lower values of $V(\mathbf{x})$ around the stable points determined by (2.11). This calls for theoretical results to be proven in order to better understand potential function convexity and discontinuity in noisy regimes.

In the limit $\gamma \to +\infty$ equation (2.11) collapses to a Dirac distribution centred around the global minimum of $V(\mathbf{x})$, i.e. the maximum a posteriori (MAP) estimate of (2.11). The MAP clearly does not constitute a good fit for the observed sizes. As $\gamma \to 0$, equation (2.11) converges to an improper uniform distribution (Ellam et al., 2018). The potential function appears to be sensitive to the values of α, β instead of the initial conditions, as illustrated in Figure 2.2. For small values of α, job availability is approximately the same for the two destination zones, whereas for larger values of α, job availability is more dispersed around the two zones.

We make another simplification by fixing the cost parameter κ. It follows from (2.22) that at equilibrium the total demand must match the total supply (Harris and A. G. Wilson, 1978).

Hence,

$$\begin{aligned}\kappa \sum_{j=1}^{M} \left(W_j - \delta\right) &= \sum_{j=1}^{M} D_j \\ &= \sum_{j=1}^{M} \sum_{i=1}^{N} T_{ij} \\ &= \sum_{i=1}^{N} O_i, \end{aligned} \tag{2.24}$$

where the coefficient of κ is assumed to be known. Therefore, κ can be obtained by solving (2.24):

$$\kappa = \frac{1}{K} \left(\sum_{i=1}^{N} O_i + \delta M \right), \tag{2.25}$$

where $K \coloneqq \sum_{j=1}^{M} W_j$. We set $K = 1$, although its choice is arbitrary. Parameter δ can be specified relative to the size $W_{j'}$ of the smallest zone j' since with no inward flows since $W_{j'} = \delta/\kappa$.

2.5 Benchmark model

In this section, we introduce a simpler model to benchmark our parameter estimates of α and β in a deterministic setting.

Assume that each T_{ij} is a Poisson-distributed random variable. Further assume that the table T_{ij} can be decomposed as $\hat{S_{ij}} = a_i b_j \quad \forall\ i, j$ such that $\hat{S}_{ij} \coloneqq \mathbb{E}[T_{ij}]$ and $\hat{S}_{ij}$ is the maximum likelihood estimate (MLE) of the expected flows. This assumption originates from a probabilistic view of the flows

$$T_{ij} = \pi_{ij} T, \tag{2.26}$$

where T is the total flow of the system and

$$\sum_{i=1}^{N} \sum_{j=1}^{M} \pi_{ij} = 1. \tag{2.27}$$

Under this view

$$\pi_{ij} = \pi_{(+,j)} \pi_{(i,+)} \coloneqq \left(\sum_{i=1}^{N} \pi_{i,j} \right) \left(\sum_{j=1}^{M} \pi_{i,j} \right), \tag{2.28}$$

which is known as the model independence assumption, and $a_i = \mathbb{E}[\sqrt{T}\pi_{(i,+)}]$, $b_j = \mathbb{E}[\sqrt{T}\pi_{(+,j)}]$ $\forall\, i, j$. Substituting (2.28) in (2.26) and taking logarithms yields

$$\log(T_{ij}) = \log(T) + \log(\pi_{(+,j)}) + \log(\pi_{(i,+)}) + z_{ij}, \tag{2.29}$$

where $\sum_{i=1}^{N} \log(\pi_{(i,+)}) = \sum_{j=1}^{M} \log(\pi_{(+,j)}) = 0$. The model in (2.29) resembles the kernel a Poisson generalised linear model (Oshan, 2016). Without loss of generality let $T = 1$, and T_{ij} be equal to (2.5). Then (2.29) takes the following form:

$$\log(T_{ij}) = \log(O_i) + \alpha \log(W_j) - \beta c_{ij} - \log\left(\sum_{k=1}^{M} W_k^{\alpha} \exp(-\beta c_{ik}\right), \tag{2.30}$$

where the first and last terms are constants for any given $i \in \{1, \ldots, N\}$. We use the Python library named `PySal` found in this GitHub repository to calibrate the model in (2.30). Model calibration is achieved efficiently by inducing sparsity on the large flow matrix (Oshan, 2016). We then compare the inferred parameters against the deterministic Harris and Wilson model. This helps us verify the validity of our claims in Chapter 4. However, our model is fundamentally different than the Poisson regression one in that we don't leverage the flow matrix to calibrate our model. Therefore, any inferred flow matrix constructions between the two models would be perplexing.

Chapter 3

Model calibration

This chapter is devoted to model calibration, which is also known as solving the inverse problem. Model calibration involves determining α and β from observational data. The scaling factor α reflects an individual's preference of popular destinations, while the scaling factor β depicts the individual sensitivity to high travel cost. As outlined in Chapter 1, traditional approaches for solving the inverse problem include the use of discrete choice models (McFadden, 1980; McFadden and Train, 2000), which are not computationally scalable due to the large volumes of flow data required for training.

3.1 Bayesian framework

We adopt a Bayesian approach to parameter and latent variable inference that incorporates uncertainty about observational data and parameter values by treating them as random variables with probability distributions. The probability distributions maintain a collection of beliefs about the true parameter/latent variable values, each weighted by a probability to reflect the uncertainty about involved about the knowledge of the true values. Prior to any data observation, a Bayesian model would only express *prior* beliefs about the true parameter/latent variable values. Under the influence of a likelihood, these beliefs are 'corrected' and updated into *posterior* beliefs. Posterior estimates can then be elicited from the posterior distribution's mean or mode and uncertainty can be expressed by the standard deviation of that posterior.

We now formalise the aforementioned arguments. Let $\mathbf{\Theta} = (\alpha, \beta) \in \mathbb{R}^2_{>0}$ and $\mathbf{X} = \log(\mathbf{W}) \in \mathbb{R}^M$ be random variables. We leverage the fact that the equilibrium position of the latent sizes follows the Boltzmann-Gibbs measure in (2.11), i.e. $\mathbf{X} \sim \rho_\infty$. Define $\mathbf{Y} \in \mathbb{R}^M_{>0}$ to

be the observational data on job availability for each destination zone. We then incorporate uncertainty through model discrepancy (error) terms $\mathbf{E}$ and express prior beliefs about the true parameters and latent variables. Therefore, we assume that the observed job market structure is informed by the latent destination sizes $\mathbf{W}$ subject to some multiplicative noise $\mathbf{E}$[1]:

$$\log(\mathbf{Y}) = \log(\mathbf{W}) + \log(\mathbf{E}), \tag{3.1}$$

where $\log(\mathbf{E}) \sim N(0, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{M \times M}$ is a positive definitive covariance matrix. The existence of a model discrepancy term in the latent sizes is motivated by the fact that the Boltzmann-Gibbs measure may provide a poor data fit to the data. In addition, we assign prior distribution over the parameters, which we denote by $\pi(\boldsymbol{\theta})$. We also define a prior over the latent sizes we denote by $\pi(\mathbf{x}|\boldsymbol{\theta})$. The uncertainty about the true (observed) sizes also induces a distribution $\pi(\mathbf{y}|\mathbf{x})$ over $\mathbf{Y}$,i.e.

$$\pi(\mathbf{y}|\mathbf{x}) = N(\mathbf{x}, \log(\mathbf{e})). \tag{3.2}$$

Note that Boltzmann-Gibbs measure has a dependence on $\boldsymbol{\theta}$ and therefore we make it explicit:

$$\pi(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x})\right) \qquad Z(\boldsymbol{\theta}) \coloneqq \int_{\mathbb{R}^M} \exp\left(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x})\right) \mathrm{d}\mathbf{x}. \tag{3.3}$$

We can now compute the joint posterior distribution over the parameters and latent variable. By Bayes rule it follows that

$$\begin{aligned}
\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) &= \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\pi(\mathbf{y})} \\
&= \frac{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}, \boldsymbol{\theta})}{\pi(\mathbf{y})} \\
&= \frac{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})} \\
&= \underbrace{\pi(\boldsymbol{\theta})}_{\text{parameter prior}} \quad \underbrace{\frac{1}{\pi(\mathbf{y})}}_{\text{marginal likelihood}} \quad \underbrace{\pi(\mathbf{y}|\mathbf{x})}_{\text{latent likelihood}} \underbrace{\frac{1}{Z(\boldsymbol{\theta})} \exp(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x}))}_{\text{parameter likelihood}}
\end{aligned} \tag{3.4}$$

where the last step follows from substitution of (3.3). Posterior inference using (3.4) is non-trivial as the normalising terms $\pi(\mathbf{y})$ and $\pi(Z(\boldsymbol{\theta}))$ are unknown. Specifically, the $\boldsymbol{\theta}$-dependence

[1]The error is multiplicative to preserve positivity of the latent sizes

of the function $Z(\boldsymbol{\theta})$ requires integration over a complex high dimensional space, which is a notoriously difficult problem (Murray, Ghahramani, and MacKay, 2012). The normalising constant in the parameter likelihood is necessary to penalise overly complex models that may lead to over-fitting and suboptimal parameter configurations. We devote the next section to devising a computational strategy to deal with this problem and compute the joint posterior density up to the normalising constant $\pi(\mathbf{y})$.

3.2 Computational simplifications

We resort to numerical simulation procedures and use MCMC schemes to compute low-order summary statistics of the form

$$\mathbb{E}\left[g(\mathbf{X}, \boldsymbol{\Theta})\right] = \int_{\mathbb{R}^2_{>0}} \int_{\mathbb{R}^M} g(\mathbf{x}, \boldsymbol{\theta}) \pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\boldsymbol{\theta}, \tag{3.5}$$

where $g(\cdot, \cdot)$ is the integrable function of interest. Examples of low-order summary statistics include the mean, variance and density of the posterior marginals.

3.2.1 Laplace approximation

Our first strategy involves approximating the normalising constant $Z(\boldsymbol{\theta})$ using a second-order Taylor expansion of the potential function around the global minima $m_{\boldsymbol{\theta}}$. The quadratic Taylor approximation is given by (Rijk and Vorst, 1983, p.83):

$$V_{\boldsymbol{\theta}}(\mathbf{x}) \approx \hat{V}_{\boldsymbol{\theta}}(\mathbf{x}) = V_{\boldsymbol{\theta}}(m_{\boldsymbol{\theta}}) + \frac{1}{2} \left(\mathbf{x} - m_{\boldsymbol{\theta}}\right)^T \Delta V_{\boldsymbol{\theta}}(m_{\boldsymbol{\theta}}) \left(\mathbf{x} - m_{\boldsymbol{\theta}}\right). \tag{3.6}$$

We now leverage the fact that the integral in (3.3) has significant contributions in the neighbourhood of $m_{\boldsymbol{\theta}}$ since $\exp(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x})) > 0$ only when $V_{\boldsymbol{\theta}}(\mathbf{x}) < 0$. Therefore, we substitute (3.6) into the integral in (3.3) to obtain a sufficient approximation for $Z(\boldsymbol{\theta})$:

$$\begin{aligned} Z(\boldsymbol{\theta}) &\approx \int_{\mathbb{R}^M} \exp(-\gamma \hat{V}_{\boldsymbol{\theta}}(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &= \exp\left(-\gamma \hat{V}_{\boldsymbol{\theta}}(m_{\boldsymbol{\theta}})\right) \int_{\mathbb{R}^M} \exp\left(-\frac{\gamma}{2} \left(\mathbf{x} - m_{\boldsymbol{\theta}}\right)^T \Delta V_{\boldsymbol{\theta}}(m_{\boldsymbol{\theta}}) \left(\mathbf{x} - m_{\boldsymbol{\theta}}\right)\right) \mathrm{d}\mathbf{x}. \end{aligned} \tag{3.7}$$

We apply yet another approximation known as the Laplace or saddle point approximation (Butler, 2007, p.83) to evaluate the integral in (3.7). Therefore, (3.7) becomes

$$Z(\boldsymbol{\theta}) \approx \exp\left(-\gamma \hat{V}_{\boldsymbol{\theta}}(m_{\boldsymbol{\theta}})\right) \frac{\left(2\pi\gamma^{-1}\right)^{M/2}}{|\Delta V_{\boldsymbol{\theta}}(m_{\boldsymbol{\theta}})|^{1/2}}, \tag{3.8}$$

where the determinant of the Hessian of the potential in the denominator of (3.8) is given by (B.5). The approximation is asymptotically accurate as $\gamma \to +\infty$, which corresponds to low-noise regimes. The quality of the approximation depends on two factors. First, the peakedness of the exponential term outside the integral occurs at $m_{\boldsymbol{\theta}}$, which we have shown to hold true. Second, the degree to which the integrand is 'quadractic-looking'. The second condition holds if and only if the potential function is convex in the neighbourhood of $m_{\boldsymbol{\theta}}$. In Appendix C.4 we derive the necessary conditions for the potential function to be convex in the two dimensional case ($N = M = 2$). It is evident that the potential function is convex for specific choices of α and latent sizes $\mathbf{x}$. We argue that in all but special cases the potential function is convex and therefore has a unique global minimum which we can numerically compute.

A suitable and inexpensive numerical optimisation routine is the Newton-based limited memory Broyden-Flether-Goldfarb-Shanno (L-BFGS) (Nocedal and Wright, 2006, p.136), which however requires proper initialisation. We run the L-BFGS procedure for M different $\mathbf{x}$ initialisations and parameter configurations to robustly derive the global minima of $V_{\boldsymbol{\theta}}(\mathbf{x})$. We summarise the computation of Laplace approximation in Algorithm 1. The computational complexity of Algorithm 1 is $\mathcal{O}(M \times C_{LBFGS} + C_{Cholesky})$, where C_{LBFGS} is the unknown computational complexity of the L-BFGS algorithm. We assume that $C_{LBFGS} \approx M^2$ (Saputro and Widyaningsih, 2017) and $C_{Cholesky} \approx M^3$, which makes the asymptotic computational complexity $\mathcal{O}(M^3)$.

3.2.2 Unbiased importance weight estimation

In high-noise regimes, the Laplace approximation in (3.8) performs poorly and therefore we cannot obtain accurate joint posterior estimates. Hence, we seek a more robust to noise estimator of (3.5). In fact, an unbiased estimate of $Z(\boldsymbol{\theta})$ is given by an average of a batch of

Algorithm 1 Laplace approximation algorithm for the Boltzmann-Gibbs measure normalising constant in low-noise regimes.

Function call: `Laplace`($\boldsymbol{\theta}'$,$\mathbf{x}^{(0)}$,M)

Input: Augmented parameter vector $\boldsymbol{\theta}' = (\alpha, \beta, \gamma, \delta, \kappa, \epsilon)$, potential function $V_{\boldsymbol{\theta}}(\mathbf{x})$, number of destinations M.

Output: $\log(Z(\boldsymbol{\theta}))$ and its sign.

\# Start of algorithm

Initialise $m_{\boldsymbol{\theta}} \leftarrow \mathbf{x}^{(0)}$ and $V_{\boldsymbol{\theta}}(m_{\boldsymbol{\theta}}) \leftarrow V_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}$ using (2.21).

for $k \in \{1, \ldots, M\}$ **do**

  Initialise $x_{i0} \leftarrow \log(\delta)\ \forall\ i \neq k$ and $x_{i0} \leftarrow \log(1+\delta)$ for $i = k$.

  Use the L-BFGS optimiser to find $\mathbf{x}' \leftarrow \arg\min_{\mathbf{x} \in \mathbb{R}^M} V_{\boldsymbol{\theta}}(\mathbf{x})$ and $V_{\boldsymbol{\theta}}(\mathbf{x}')$ using $\mathbf{x}_0$ as the initialisation.

  if $V_{\boldsymbol{\theta}}(\mathbf{x}') < V_{\boldsymbol{\theta}}(m_{\boldsymbol{\theta}})$ **then**

    Update $m_{\boldsymbol{\theta}} \leftarrow \mathbf{x}'$ and $V_{\boldsymbol{\theta}}(m_{\boldsymbol{\theta}}) \leftarrow V_{\boldsymbol{\theta}}(\mathbf{x}')$ using (2.21).

  end if

end for

Compute the Hessian matrix $\mathbf{H} \leftarrow \nabla V_{\boldsymbol{\theta}}(m_{\boldsymbol{\theta}})$ using (C.10)-(C.11).

Find the Cholesky decomposition $\mathbf{L}$ that satisfies $\mathbf{H} \leftarrow \mathbf{L}\mathbf{L}^T$.

Compute the log-determinant $\log\left(\det(\mathbf{H})\right) \leftarrow \sum_{n=1}^{M} L_{nn}$.

Evaluate $Z(\boldsymbol{\theta}) \leftarrow -\gamma V_{\boldsymbol{\theta}}(m_{\boldsymbol{\theta}}) + \frac{M}{2}\log(2\pi\gamma^{-1}) - \frac{1}{2}\log\left(\det(\mathbf{H})\right)$.

importance weights

$$w(\mathbf{x}) = \frac{\exp(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x}))}{q(\mathbf{x})}, \tag{3.9}$$

where q is the density of the proposal distribution and $\mathbf{x}$ are draws from q. However, we note that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{w(\mathbf{x})}\right] &= \mathbb{E}\left[\frac{q(\mathbf{x})}{\exp(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x}))}\right] \\
&\geq \frac{1}{\mathbb{E}\left[\frac{\exp(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x}))}{q(\mathbf{x})}\right]} \\
&\neq \mathbb{E}\left[\frac{1}{Z(\mathbf{x})}\right],
\end{aligned} \tag{3.10}$$

where the last step follows from Jensen's inequality. Therefore, the reciprocal of our importance weight estimator in (3.9) is a biased estimated estimator of $Z(\boldsymbol{\theta})$. We follow the approaches by (Murray, Ghahramani, and MacKay, 2012),(Lyne et al., 2015), and (Wei and Murray, 2016) to debias this estimator. The Russian roulette truncation can be leveraged to obtain unbiased estimates of $1/Z(\boldsymbol{\theta})$. Define a sequence of estimators $\nu = \{\nu_i : i \geq 0\}$ such that $\lim_{i\to\infty} \mathbb{E}[\nu_i] = 1/Z(\boldsymbol{\theta})$. Draw a random integer T_s in independent of ν and then take the sum

$$S = \nu_0 + \sum_{i=1}^{K} \frac{\nu_i - \nu_{i-1}}{\mathbb{P}(T_s \geq i)}. \tag{3.11}$$

For the sake of simplicity we set $\mathbb{P}(T_s \geq i) \propto i^{-1.1}$. Then, $\mathbb{E}[S] = 1/Z(\boldsymbol{\theta})$ if and only if $\mathbb{E}\left[|\nu_0| + \sum_{i=1}^{\infty} |\nu_i - \nu_{i-1}|\right] \leq \infty$ (Wei and Murray, 2016). We advocate for the increasing averages estimator proposed by (Lyne et al., 2015)

$$\nu_i = \frac{i+1}{\sum_{k=0}^{i} w(\hat{\mathbf{x}}^{(k)})}, \tag{3.12}$$

where each $\mathbf{x}^{(k)} \quad \forall \; k = 0, \ldots, i$ is an independent draw from the q density. Although S constitutes an unbiased estimator of the reciprocal of the normalising constant, it has high variance since the sum in (3.11) does not always consist of positive terms. (Lyne et al., 2015) addresses the 'sign problem' by storing $|S|$ at every time step while monitoring $\text{sign}(S)$, where $\text{sign}(S) = -1$ if $S < 0$ and $\text{sign}(S) = +1$ if $S \geq 0$. The cases when S is negative introduce high variability in the estimator in (3.11), which renders its use in an MCMC posterior sampling

scheme impractical. This is because high variability in $1/Z(\boldsymbol{\theta})$ can lead to low acceptance rates and highly dependent posterior samples.

Fortunately, annealed importance sampling (AIS) can alleviate the estimator's high variance by augmenting the state-space from $(\mathbf{X}, \boldsymbol{\Theta})$ to $(\mathbf{X}, \boldsymbol{\Theta}, \Omega)$ (Neal, 1998). Therefore, we aim to construct a sequence $\{\mathbf{X}^{(i)}, \boldsymbol{\Theta}^{(i)}, \Omega^{(i)}\}_{i=1}^{n}$ such that

$$\mathbb{E}\left[g(\mathbf{X}, \boldsymbol{\Theta})|\mathbf{Y} = \mathbf{y}\right] = \lim_{n \to +\infty} \frac{\sum_{i=1}^{n} \Omega^{(i)} g(\mathbf{X}^{(i)}, \boldsymbol{\Theta}^{(i)})}{\sum_{k=1}^{n} \Omega^{(k)}}, \tag{3.13}$$

where $\Omega^{(i)} \in \{-1, 1\}$ is equal to the sign(S). The suitability of this estimator for the reciprocal of the normalising constant $Z(\boldsymbol{\theta})$ is limited to high-noise regimes. This is because in low-noise regimes the Boltzmann-Gibbs measure is asymptotically concentrated around a Dirac mass, which renders precise importance sampling challenging.

The expectation in (3.13) is defined with respect to the posterior density $p_0(\mathbf{x}, \boldsymbol{\theta}) \coloneqq \pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$. We define a sequence of other distributions $p_1(\mathbf{x}, \boldsymbol{\theta})$ to $p_n(\mathbf{x}, \boldsymbol{\theta})$ such that $p_j(\mathbf{x}, \boldsymbol{\theta}) \neq 0$ whenever $p_{j-1}(\mathbf{x}, \boldsymbol{\theta}) \neq 0$. We also need a function $f_j(\mathbf{x}, \boldsymbol{\theta})$ that is proportional to $p_j(\mathbf{x}, \boldsymbol{\theta})$ and a way to draw independent samples from p_n. Finally, we require $n-1$ Markov chain transitions T_j that leave p_j invariant. Fix $f_0 \coloneqq \exp(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x}))$ to give the function of interest and f_n to be a function proportional to f_0 whose distribution we can sample from. We then let

$$f_j(\mathbf{x}, \boldsymbol{\theta}) = f_0(\mathbf{x}, \boldsymbol{\theta})^{t_j} f_n(\mathbf{x}, \boldsymbol{\theta})^{1-t_j}, \tag{3.14}$$

where $1 = t_0 > t_1 > \ldots, t_n = 0$ are the temperatures we use to anneal the densities. In Appendix C.5 we show that a suitable initialisation for AIS that generates independent samples of $\mathbf{x}$ is

$$x_j \sim \log\left(\Gamma\left(\gamma(\delta + 1/M), 1/\gamma\kappa\right)\right) \quad \forall\, j \in \{1, \ldots, M\} \tag{3.15}$$

and a suitable corresponding function $\log(f_n)$ is

$$-\gamma V'_{\boldsymbol{\theta}}(\mathbf{x}) \coloneqq \lim_{\alpha \to 0, \beta \to 0} -\gamma V_{\boldsymbol{\theta}}(\mathbf{x}) = \gamma\kappa \sum_{j=1}^{M} \exp(x_j) - \gamma(\delta + 1/M) \sum_{j=1}^{M} x_j \tag{3.16}$$

According to (Neal, 1998), the i-th importance weight is given by

$$\begin{aligned}\mathbf{w}^{(i)} &= \frac{\prod_{j=0}^{n-1} f_j(\mathbf{x}^{(j)}, \boldsymbol{\theta})}{\prod_{k=1}^{n} f_k(\mathbf{x}^{(k-1)}, \boldsymbol{\theta})} \\ &= \frac{\prod_{j=0}^{n-1} f_0(\mathbf{x}^{(j)}, \boldsymbol{\theta})^{t_j} f_n(\mathbf{x}^{(j)}, \boldsymbol{\theta})^{1-t_j}}{\prod_{k=1}^{n} f_0(\mathbf{x}^{(k-1)}, \boldsymbol{\theta})^{t_k} f_n(\mathbf{x}^{(k-1)}, \boldsymbol{\theta})^{1-t_k}},\end{aligned} \tag{3.17}$$

where the last step follows by substitution of (3.14). The Markov chain transitions are chosen to be the transition kernel of the Hamiltonian Monte Carlo (HMC) scheme, which we explore in the next section. We summarise the AIS scheme in Algorithm 2. Overflow problems are avoided by performing computations in log-space. We note that line 11 of the algorithm follows by substitution of (2.21) and (3.16) in (3.17). By inspection, it is clear that the asymptotic complexity of AIS is $\mathcal{O}(LNMn_pn_t)$, where n_p is the number of particles/weights generated and n_t is the number of temperatures used in annealing. The complexity of $\texttt{HMC}(\mathbf{x}^{(0)},\boldsymbol{\theta}',\epsilon,L,M)$ in step 12 will be shown to be $\mathcal{O}(LNM)$.

We can now devise an algorithm to obtain unbiased, variance-reducing estimates of $1/Z(\boldsymbol{\theta})$. The pseudo-code is provided in Algorithm 3. The resulting computational complexity is $\mathcal{O}(KLNMn_pn_t)$, where K is the number of stopping times/truncations of the (in)finite series in (3.11).

3.3 Latent size posterior sampling

Posterior inference in the high-dimensional setting we have described in the previous sections of this chapter is challenging. In order to avoid slow exploration of the state space achieved by random-walk behaviour we resort to efficient sampling schemes. The latent log-size posterior marginal in (3.3) lends itself to HMC sampling. The reason is that the Boltzmann-Gibbs measure resembles the canonical distribution from statistical mechanics, which itself resembles the joint distribution of the Hamiltonian (Neal, 2012):

$$P(\mathbf{q},\mathbf{p}) = \frac{1}{Z}\exp(-H(\mathbf{q},\mathbf{p})/T) = \frac{1}{Z}\exp(-U(\mathbf{q})/T)\exp(-K(\mathbf{p})/T), \tag{3.18}$$

where $H(\mathbf{q},\mathbf{p}) \coloneqq U(\mathbf{q}) + K(\mathbf{p})$ is the total Hamiltonian energy, $U(q)$ is the potential energy at a given position $\mathbf{q}$ and $K(\mathbf{p})$ is the kinetic energy for a given momentum $\mathbf{p}$. The variable

Algorithm 2 Annealed importance sampling algorithm used to reduce the variance of the unbiased increasing averages estimator for the Boltzmann-Gibbs measure normalising constant in high-noise regimes.

Function call: `AnnealedImportanceSampling`($\mathbf{x}^{(0)}$,$\boldsymbol{\theta}'$,n_p,n_t,ϵ,L,M)

Input: Latent log-size initialisation $\mathbf{x}^{(0)}$, Augmented parameter vector $\boldsymbol{\theta}' = (\alpha, \beta, \gamma, \delta, \kappa, \epsilon)$, potential function $V_{\boldsymbol{\theta}}(\mathbf{x})$, modified potential function $V'_{\boldsymbol{\theta}}(\mathbf{x})$, number of particles/weights n_p, number of temperatures for annealing n_t, leapfrog step size ϵ, number of leapfrog steps L, number of destinations M.

Output: $\log(\mathbf{W})$.

\# Start of algorithm

Define temperatures $\mathbf{t}$ for annealing by generating n_t equally spaced scalars in the range $[0, 1]$.

Initialise AIS weights $\log(w_j^{(0)}) \leftarrow -\log(n_p) \;\; \forall j \in \{1, \ldots, n_p\}$.

for $i \in \{1, \ldots, n_p\}$ **do**

  Sample $\mathbf{x}^{(0)}$ from $\log\left(\Gamma\left(\gamma(\delta + 1/M), 1/\gamma\delta\right)\right)$.

  Compute $V_0 \leftarrow V'_{\boldsymbol{\theta}}(\mathbf{x}^{(0)})$ and $V_n \leftarrow V_{\boldsymbol{\theta}}(\mathbf{x}^{(0)})$ using (2.21),(3.16).

  for $j \in \{1, \ldots, n_t\}$ **do**

    Update $\log(\mathbf{w}^{(j)}) \leftarrow \log(\mathbf{w}^{(j)}) + (t_j - t_{j-1})(V_0 - V_n)$ using (3.17).

    Generate proposal $\mathbf{x}^{(j+1)} \leftarrow$ `HMC`($\mathbf{x}^{(0)}$,$\boldsymbol{\theta}'$,ϵ,L,M) using 4.

  end for

end for

Algorithm 3 Algorithm for Russian roulette random series truncation of an importance weight estimator for the normalising constant of the Boltzmann-Gibbs measure in high-noise regimes.

Function call: `ImportanceWeightUnbiasedEstimator`($\mathbf{x}^{(0)}$,$\boldsymbol{\theta}'$,K,n_p,n_t,ϵ,L,M)

Input: Latent log-size initialisation $\mathbf{x}^{(0)}$, augmented parameter vector $\boldsymbol{\theta}' = (\alpha, \beta, \gamma, \delta, \kappa, \epsilon)$, random stopping time K, potential function $V_{\boldsymbol{\theta}}(\mathbf{x})$, modified potential function $V'_{\boldsymbol{\theta}}(\mathbf{x})$, number of particles/weights n_p, number of temperatures for annealing n_t, leapfrog step size ϵ, number of leapfrog steps L, number of destinations M.

Output: $\log(\mathbb{E}[S])$ and $\text{sign}(\mathbb{E}[S])$).

\# Start of algorithm

Initialise K dimensional $w^{(j)}$ and $\log(\nu)$ sequences.

for $i \in \{1, \ldots, K\}$ **do**

  Compute $\log(w^{(j)}) \leftarrow$ `AnnealedImportanceSampling`($\mathbf{x}^{(0)}$,$\boldsymbol{\theta}'$,n_p,n_t,ϵ,L,M).

end for

for $i \in \{1, \ldots, K\}$ **do**

  Compute log of increasing averages estimator $\log(\nu_i) \leftarrow \log(i+1) - \log(\sum_{j=0}^{i} w^{(j)}))$.

end for

for $i \in \{1, \ldots, K\}$ **do**

  Recursively update estimator $\mathbb{E}[S] \leftarrow \nu_0 + \log(\sum_{j=0}^{K} \exp\left(\log(\nu_j) - \log(\nu_j) + 1.1\log(j)\right)$ and store its sign.

end for

of interest (in our case $\mathbf{x}$) is represented by $\mathbf{q}$ and $U(\mathbf{q})$ defines the negative log likelihood of that variable. The kinetic energy of the system is defined by

$$K(p) = \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}/2, \tag{3.19}$$

where $\mathbf{M}$ is a positive semi-definite 'mass matrix' of the particle moving in the physical system. We set $\mathbf{M} = \mathbf{I}$ for simplicity. Momentum variables are drawn from $N(\mathbf{0}, \mathbf{I})$. The Hamiltonian energy is derived from the Hamiltonian dynamics of a physical system that describes the motions of particles inside a multidimensional plane. Therefore, HMC has a natural and intuitive interpretation which fits many problem areas.

We focus on two fundamental properties of HMC sampling. First, the Hamiltonian energy is (approximately) conserved regardless of any movement in the $(\mathbf{q}, \mathbf{p})$-space. This results to an acceptance probability of any given proposal being very close to one. Contrary to traditional MCMC schemes such as Metropolis-Hastings, a high acceptance rate for an efficient exploration of the posterior provided that the proposals are carefully constructed. Secondly, the volume in $(\mathbf{q}, \mathbf{p})$-space is preserved. This implies that any space transformation will leave the volume invariant. Volume preservation prevents the acceptance rate from being affected by changes in volume.

The Hamiltonian equations describe the continuous-time dynamics of a physical system and therefore have to be discretised. The discretisation scheme is known as the leapfrog integrator where the state of the Hamiltonian is computed for L leapfrog steps of step size ϵ. Typically, "the HMC algorithm will not get trapped in some subset of the state space, and hence will asymptotically converge to its (unique) invariant distribution" (Neal, 1998). However, in practise the leapfrog steps L and step size ϵ will heavily impact the ergodicity of the HMC and have to be fine-tuned. For the purposes of this thesis we tune these parameters by trial and error. The acceptance probability for the latent log-size posterior marginal is

$$a_{\mathbf{X}}(\mathbf{x}', \mathbf{p}'|\mathbf{x}, \mathbf{p}) = \min \left\{ 1, \frac{\pi(\mathbf{y}|\mathbf{x}', \boldsymbol{\theta}) \exp(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x}') - (1/2)|\mathbf{p}'|^2)}{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \exp(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x}) - (1/2)|\mathbf{p}|^2)} \right\} \tag{3.20}$$

The computational complexity of HMC sampling is $\mathcal{O}(LNM)$ based on the pseudo-code in Algorithm 4.

Algorithm 4 Hamiltonian Monte Carlo sampling algorithm used in latent log-size posterior marginal inference.

Function call: HMC($\mathbf{x}^{(0)}$,$\boldsymbol{\theta}'$,ϵ,L,M)

Input: Augmented parameter vector $\boldsymbol{\theta}' = (\alpha, \beta, \gamma, \delta, \kappa, \epsilon)$, potential function $V_{\boldsymbol{\theta}}(\mathbf{x})$, leapfrog step size ϵ, number of leapfrog steps L, number of destinations M.

Output: Latent posterior samples $\hat{\mathbf{X}}$.

\# Start of algorithm

Initialise latent samples $\hat{\mathbf{X}}^{(0)} \leftarrow \mathbf{x}^{(0)}$ and momentum $\mathbf{p}^{(0)} \sim N(\mathbf{0}, \mathbf{I})$.

Compute $\log(\pi(\mathbf{y}|\hat{\mathbf{X}}^{(0)}))$ using (3.2).

Compute initial log potential function $V \leftarrow -\gamma V_{\boldsymbol{\theta}}(\hat{\mathbf{X}}^{(0)})$ and its gradient $gradV \leftarrow \nabla V_{\boldsymbol{\theta}}(\hat{\mathbf{X}}^{(0)})$ using (C.9) .

Compute initial log potential energy $U \leftarrow V + \log(\pi(\mathbf{y}|\hat{\mathbf{X}}^{(0)}))$ and its gradient $gradU \leftarrow \nabla U(\hat{\mathbf{X}}^{(0)})$ using (C.9).

Compute initial log Hamiltonian energy $H \leftarrow 0.5\mathbf{p}^T\mathbf{p} + U$.

Set $current_q \leftarrow \hat{\mathbf{X}}^{(0)}$, $current_p \leftarrow \mathbf{p}^{(0)}$.

for $i \in \{1, \ldots, L\}$ **do**

  Use previous position $\hat{\mathbf{X}}^{(i)} \leftarrow \hat{\mathbf{X}}^{(i-1)}$

  Make a half step for momentum $current_p \leftarrow current_p - 0.5\epsilon\ gradU_p$.

  Make a full step for the position $current_q \leftarrow current_q + \epsilon\ current_p$.

  Update log potential energy and its gradient $U_p \leftarrow \log(\pi(\mathbf{y}|current_q)) - \gamma V_{\boldsymbol{\theta}}(current_q)$ and $gradU_p$ using (C.9) .

  Make a half step for the momentum $current_p \leftarrow current_p - 0.5\epsilon\ gradU_p$.

end for

Update Hamiltonian energy $H_p \leftarrow 0.5(current_p)^T(current_p) + U(current_q)$

Draw sample $u \sim Uniform(0, 1)$.

if $\log(u) < H - H_p$ **then**

  Update position $\mathbf{x}^{(i)} \leftarrow current_q$

  Update log potential function and its gradient $V \leftarrow V_p$, $gradV \leftarrow gradV_p$.

end if

3.4 Joint posterior sampling

In the previous sections, we have described two methods for obtaining estimates of the normalising constant $Z(\boldsymbol{\theta})$ in high and low-noise regimes and a latent posterior marginal sampling scheme. Our overall aim is to construct an ergodic Markov chain (Robert and Casella, 2013, p.231) of latent log-size and parameter samples whose time average (3.5) and (3.13) converge to their true posterior mean. By deriving expressions for the posterior marginals, we are able to use a Metropolis-within-Gibbs scheme with reflective boundaries to iteratively perform $\mathbf{X}$ and $\boldsymbol{\Theta}$ updates, as outlined in (Ellam et al., 2018). Asymptotically, these updates comprise a Markov chain that in theory leaves the joint posterior $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ invariant. The parameter proposals are generated using a random walk transition kernel. Reflective boundaries allow us to constrain these proposals in the $[0, 2]$ range by 'reflecting' proposals that lie outside that range.

3.4.1 Low-noise regimes

In low-noise regimes, we leverage the Laplace approximation in (3.8) to accept/reject $\boldsymbol{\Theta}$ samples. The Metropolis-Hastings acceptance probability for the $\boldsymbol{\Theta}$ updates in low-noise settings is

$$a_{\boldsymbol{\Theta}}(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \min\left\{1, \frac{\pi(\mathbf{y}|\mathbf{x}', \boldsymbol{\theta}')Z(\boldsymbol{\theta}')\exp(-\gamma V_{\boldsymbol{\theta}'}(\mathbf{x}'))\pi(\boldsymbol{\theta}')}{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})Z(\boldsymbol{\theta})\exp(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x}))\pi(\boldsymbol{\theta})}\right\} \tag{3.21}$$

The pseudo-code for the low-noise MCMC posterior sampling scheme is provided in Algorithm 5 and its computational complexity is $\mathcal{O}(n_{mcmc}LNM + n_{mcmc}M^3)$ for large enough N, M.

3.4.2 High-noise regimes

In high-noise regimes, the deterministic Laplace approximation of the normalising constant fails. As illustrated in Section 3.2.2, we obtain an unbiased variance-reducing estimator of the reciprocal of the normalising constant using Algorithm 3. Since the likelihood is intractable we use an we approximate it using a pseudo-likelihood, which is defined as a product of conditional probabilities. For this reason, the MCMC scheme is known as pseudo-marginal

Algorithm 5 Metropolis-within-Gibbs posterior sampling algorithm with reflective boundaries using the Laplace approximation for the normalising constant in low-noise regimes.

Function call: `MCMC-LOW`($\mathbf{x}^{(0)}$,$\boldsymbol{\theta}'^{(0)}$,$\epsilon_\theta$,$\epsilon_x$,$L$,$M$,$\mathbf{C_p}$,$n_{\text{mcmc}}$)

Input: Augmented parameter initialisation $\boldsymbol{\theta}'^{(0)} = (\alpha_0, \beta_0, \gamma, \delta, \kappa, \epsilon)$, potential function $V_{\boldsymbol{\theta}}(\mathbf{x})$, parameter step size ϵ_θ,leapfrog step size ϵ_x, number of leapfrog steps L, number of destinations M, random walk covariance $\mathbf{C_p}$, number of MCMC iterations n_{mcmc}.

Output: Parameter posterior samples $\hat{\boldsymbol{\Theta}}$, latent posterior samples $\hat{\mathbf{X}}$.

\# Start of algorithm

Initialise latent samples $\hat{\mathbf{X}}^{(0)} \leftarrow \mathbf{x}^{(0)}$ and parameter samples $\hat{\boldsymbol{\Theta}}^{(0)} \leftarrow \boldsymbol{\theta}^{(0)}$.

Compute initial log reciprocal of n.c. $logZ_{inv} \leftarrow$ `Laplace`($\boldsymbol{\theta}'^{(0)}$,$\mathbf{x}^{(0)}$,$M$) using 1.

Evaluate initial p.f. $V \leftarrow V_{\boldsymbol{\theta}'^{(0)}}(\mathbf{x}^{(0)})$ using (2.21) and its gradient $gradV$ using (C.9).

Store $xx \leftarrow \mathbf{x}^{(0)}$ and $tt \leftarrow \boldsymbol{\theta}^{(0)}$.

for $i \in \{1, \ldots, n_{mcmc}\}$ **do**

  \# Perform $\boldsymbol{\Theta}$ update.

  Generate two-dimensional sample $\mathbf{s} \sim N(\mathbf{0}, \mathbf{I})$.

  Compute theta proposal $tt_p \leftarrow tt + \epsilon_\theta \mathbf{C_p} \mathbf{s}^T$.

  if $tt_p \in \mathbb{R}^2 \setminus [0,2]^2$ **then**

    Reflect tt_p off the $[0,2]$ boundary.

  end if

  if $tt_p \in [0,2]^2$ **then**

    Set $\theta'^{(i)}[0] \leftarrow tt_p[0]$ and $\theta'^{(i)}[1] \leftarrow tt_p[1]$.

    Compute updated log reciprocal of n.c. $logZ_{inv_p} \leftarrow$ `Laplace`($\boldsymbol{\theta}'^{(i)}$,$\mathbf{x}^{(i)}$,$M$) using 1.

    Compute updated p.f. $V_p \leftarrow V_{\boldsymbol{\theta}'^{(i)}}(\mathbf{x}^{(i)})$ using (2.21) and its gradient $gradV_p$ using (C.9).

    Compute log posterior marginal for initial parameter choice $pp \leftarrow logZ_{inv} - V$ and updated parameter choice $pp_p \leftarrow logZ_{inv_p} - V_p$.

Draw sample $u \sim Uniform(0,1)$.
if $\log(u) < pp_p - pp$ **then**
Accept parameter proposal $tt \leftarrow tt_p$.
Update initial potential function $V \leftarrow V_p$ and its gradient $gradV \leftarrow gradV$.
Update initial log reciprocal of normalising constant $logZ_{inv} \leftarrow logZ_{inv_p}$.
end if
end if
# Perform $\mathbf{X}$ update.
Reset parameters for latent update $\theta'^{(i)}[0] \leftarrow tt[0]$ and $\theta'^{(i)}[1] \leftarrow tt[1]$.
Generate posterior latent log-size sample $xx \leftarrow \texttt{HMC}(xx, \boldsymbol{\theta}', \epsilon_x, L, M)$ using 4.
Store posterior samples $\hat{\boldsymbol{\Theta}}^{(i)} \leftarrow tt$ and $\hat{\mathbf{X}}^{(i)} \leftarrow xx$.
end for

MCMC (Andrieu and Roberts, 2009) and its applicability relies on positive estimates $\mathbf{S}^{(i)}$ of the $1/Z(\boldsymbol{\theta})$. For that reason we cache $\mathbf{S}^{(i)}$ and monitor its percentage of positive-signs. The resulting MCMC chain consists of samples $\{\Omega^{(i)}, \mathbf{X}^{(\mathbf{i})}, \boldsymbol{\Theta}^{(i)}\}_{i=1}^{n_{mcmc}}$ and used in (3.13) to compute the posterior mean, where $\Omega^{(i)} := \text{sign}(\mathbf{S}^{(i)})$. The Metropolis-Hastings acceptance probability for the $\boldsymbol{\Theta}$ updates is adjusted to

$$a_{\boldsymbol{\Theta}}(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \min\left\{1, \frac{\pi(\mathbf{y}|\mathbf{x}', \boldsymbol{\theta}')|\mathbf{S}'| \exp(-\gamma V_{\boldsymbol{\theta}'}(\mathbf{x}'))\pi(\boldsymbol{\theta}')}{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})|\mathbf{S}| \exp(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x}))\pi(\boldsymbol{\theta})}\right\} \tag{3.22}$$

The high-noise MCMC posterior sampling scheme is outlined in Algorithm 5 and its computational complexity is $\mathcal{O}(n_{mcmc} K L N M n_p n_t)$.

3.5 Implementation details

The model calibration framework presented in this Chapter is tailored to learning parameters α and β. Learning parameters in complex noisy high-dimensional settings is challenging and over-parametrisation can lead to identifiability and tuning problems, especially in the estimation of $Z(\boldsymbol{\theta})$. To alleviate this problem, we restrict parameter inference in the $[0,2]^2$ range, as

Algorithm 5 Metropolis-within-Gibbs posterior sampling algorithm with reflective boundaries using an unbiased variance reducing estimator for the normalising constant in high-noise regimes.

Function call: `MCMC-HIGH`($\mathbf{x}^{(0)}$,$\boldsymbol{\theta}'^{(0)}$,$K$,$n_p$,$n_t$,$\sigma_{rw}$,$\epsilon$,$L$,$M$,$n_{\text{mcmc}}$)

Input: Augmented parameter initialisation $\boldsymbol{\theta}'^{(0)} = (\alpha_0, \beta_0, \gamma, \delta, \kappa, \epsilon)$, potential function $V_{\boldsymbol{\theta}}(\mathbf{x})$, random stopping time K, number of AIS samples n_p, number of annealing temperatures n_t, random walk standard deviation σ_{rw}, leapfrog step size ϵ, number of leapfrog steps L, number of destinations M, number of MCMC iterations n_{mcmc}.

Output: Parameter posterior samples $\hat{\boldsymbol{\Theta}}$, latent posterior samples $\hat{\mathbf{X}}$.

\# Start of algorithm

Initialise latent samples $\hat{\mathbf{X}}^{(0)} \leftarrow \mathbf{x}^{(0)}$ and parameter samples $\hat{\boldsymbol{\Theta}}^{(0)} \leftarrow \boldsymbol{\theta}^{(0)}$.

Compute initial log reciprocal of normalising constant $logZ_{inv} \leftarrow$ `ImportanceWeightUnbiasedEstimator`($\boldsymbol{\theta}'$,K,n_p,n_t,ϵ,L,M) using 3.

Evaluate initial p.f. $V \leftarrow V_{\boldsymbol{\theta}'^{(0)}}(\mathbf{x}^{(0)})$ using (2.21) and its gradient $gradV$ using (C.9).

Store $xx \leftarrow \mathbf{x}^{(0)}$ and $tt \leftarrow \boldsymbol{\theta}^{(0)}$.

for $i \in \{1, \ldots, n_{mcmc}\}$ **do**

  \# Perform $\boldsymbol{\Theta}$ update.

  Generate two-dimensional sample $\mathbf{s} \sim N(\mathbf{0}, \sigma_{rw}\mathbf{I})$.

  Compute theta proposal $tt_p \leftarrow tt + \mathbf{s}$.

  if $tt_p \in \mathbb{R}^2 \setminus [0,2]^2$ **then**

    Reflect tt_p off the $[0,2]$ boundary.

  end if

  if $tt_p \in [0,2]^2$ **then**

    Set $\theta'^{(i)}[0] \leftarrow tt_p[0]$ and $\theta'^{(i)}[1] \leftarrow tt_p[1]$.

    Compute updated log reciprocal of normalising constant $logZ_{inv_p} \leftarrow$ `ImportanceWeightUnbiasedEstimator`($\mathbf{x}^{(i)}$, $\boldsymbol{\theta}'^{(i)}$,$K$,$n_p$,$n_t$,$\epsilon$,$L$,$M$) using 3.

    Compute updated p.f. $V_p \leftarrow V_{\boldsymbol{\theta}'^{(i)}}(\mathbf{x}^{(i)})$ using (2.21) and its gradient $gradV_p$ using (C.9).

Compute log posterior marginal for initial parameter choice $pp \leftarrow logZ_{inv} - V$ and updated parameter choice $pp_p \leftarrow logZ_{inv_p} - V_p$.
Draw sample $u \sim Uniform(0, 1)$.
if $\log(u) < pp_p - pp$ **then**
Accept parameter proposal $tt \leftarrow tt_p$.
Update initial potential function $V \leftarrow V_p$ and its gradient $gradV \leftarrow gradV$.
Update initial log reciprocal of normalising constant $logZ_{inv} \leftarrow logZ_{inv_p}$.
end if
end if
# Perform $\mathbf{X}$ update.
Reset parameters for latent update $\theta'^{(i)}[0] \leftarrow tt[0]$ and $\theta'^{(i)}[1] \leftarrow tt[1]$.
Generate posterior latent log-size sample $xx \leftarrow \texttt{HMC}(xx, \boldsymbol{\theta}', \epsilon_x, L, M)$ using 4.
Store posterior samples $\hat{\boldsymbol{\Theta}}^{(i)} \leftarrow tt$ and $\hat{\mathbf{X}}^{(i)} \leftarrow xx$.
end for

advocated by (Dearden and Alan Wilson, 2011) and (Ellam et al., 2018). Therefore, a weakly informative uniform distribution is used as a parameter prior $\pi(\boldsymbol{\theta})$. We fix the responsiveness parameter ϵ of the potential function in (2.21) to be equal to one in an attempt to speed up convergence to the equilibrium Boltzmann-Gibbs measure. We also set $\delta \coloneqq \min_{1 \leq j \leq M} \exp(x_j)$, which is justified by a non-collapsing zone with no inward flows argument. Equation (2.25) can be leveraged to obtain a value for the job competition term or cost of assigning a person to a job. We normalise origin supplies O_i and destination sizes W_j to 1 and the cost matrix $\mathbf{C}_{ij}$ to 7×10^5, as in (Ellam et al., 2018). This implies that the estimated flow matrix is also normalised to 1, i.e. it expresses the transition probability from each origin to each destination. Parameter β is scaled appropriately as determined by a preliminary study (Dearden and Alan Wilson, 2011). Finally, γ is set to 10^4 for low-noise regimes and 10^2 for high-noise regimes.

In the deterministic setting, parameters are inferred heuristically using a grid search over $[0, 2]^2$ where the most suitable parameter pair maximises the coefficient of determination R^2 between the actual and estimates destination sizes. For each pair of parameters the estimated

sizes are the ones minimising the potential function in (2.9). We then use the Poisson regression (PR) model (which requires the entire flow matrix for calibration) to compare the inferred parameters between the two methods.

In noisy regimes, we specify independent and homogeneous observation noise covariance as $\boldsymbol{\Sigma} = \lambda \mathbf{I}$, where λ is the standard deviation of the noise. Initially, we apply `Laplace` estimation (See Algorithm 1) in both low and high-noise settings to obtain maximum likelihood estimates of the parameters. Then, the Metropolis-within-Gibbs with reflective boundaries posterior sampling scheme is employed to solve the inverse problem in a Bayesian manner. In all scenarios, we monitor the acceptance rate of the parameter and latent size samples to be between 40%-70% and above 90%, respectively. To ensure that we obtain independent posterior samples, we check that the sample auto-correlation drops sufficiently fast. Moreover, in the high-noise scenario we also monitor the signs of the unbiased estimator $\mathbf{S}$ for $Z(\boldsymbol{\theta})$ to be as high as possible. Both normalising constant estimators summarised in Algorithms 1 and 3 are suitably fine-tuned. We initialise the destination sizes with the true latent sizes and the parameters with the parameter MAP estimates obtained using a Laplace approximation. In AIS estimation we set $n_p = 10$, $n_t = 50$, $L_p = 10$ and $\epsilon_p = 0.1$. Algorithm 5 is run with $\epsilon_\theta = 1$, $\epsilon_x = 0.02$, $L = 50$, and $n_{mcmc} = 20,000$ while algorithm 5 is executed with K randomly generated positive integers, $\sigma_{rw} = 0.3$, $\epsilon = 0.02$, $L = 50$, and $n_{mcmc} = 10000$.

Chapter 4

London commuting case study

This chapter is devoted to illustrating the proposed methodology on a real-world travel demand dataset; namely, the 2001 London commuter pattern dataset provided by the Office of National Statistics [1] In this context, we demonstrate how the Boltzmann-Gibbs measure in (2.11) can be used to simulate travel demand scenarios for any given level of uncertainty (noise) and set of parameters α and β. The attractiveness parameter α reflects the benefit (i.e. job) an individual enjoys when travelling to destination while parameter β encodes information about the inconvenience of travel. Aggregated travel flow data for multiple modes of transportation (cars, buses, trains etc.) is often scarce and commercially licensed when available, as argued in Chapter 1. Our calibration approach does not utilise flow data and provides a stochastic evolution of latent travel demand. To our knowledge spatial interaction models have been applied to travel demand driven by economic factors (Batten and Boyce, 1987), but these models do not monitor its stochastic evolution.

4.1 Data overview

The commuter dataset comprises of flows of commuters between London Boroughs and is illustrated in Figure 4.1. We treat this flow matrix as 'missing data' and validate our inferred flow matrix against it at the end of this Chapter. Instead, we use London ward-level population data as origin supplies from the Greater London Authority. Borough-level job availability data are used to represent true destinations sizes. There are 628 Wards ($N = 628$) and 33 Boroughs

[1] We choose this dataset because it is the only available London-wide origin-destination flow dataset publicly accessible, which allows us to validate our methodology.

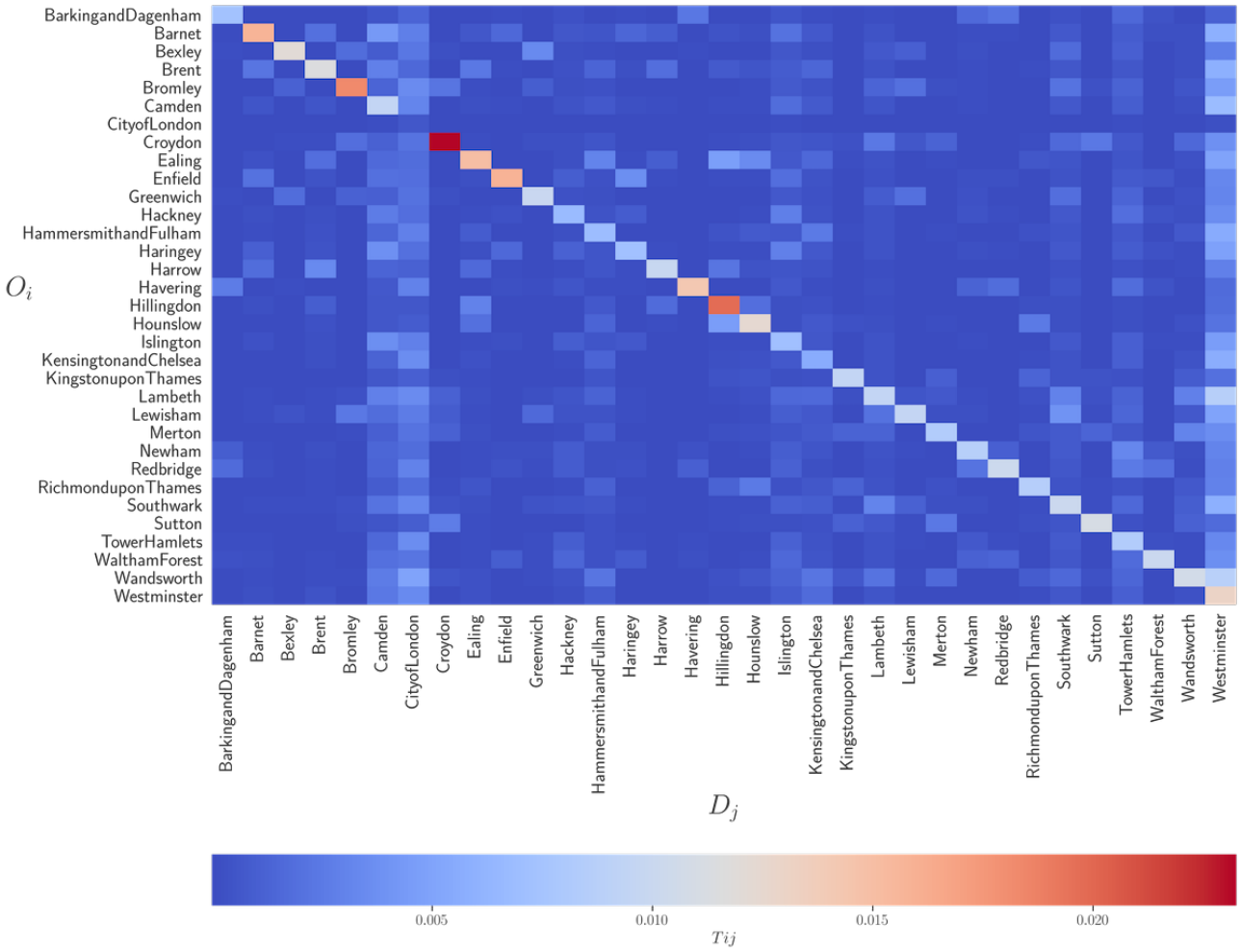


Figure 4.1: Two-dimensional heat-map of commuter flow between London Boroughs in 2001.

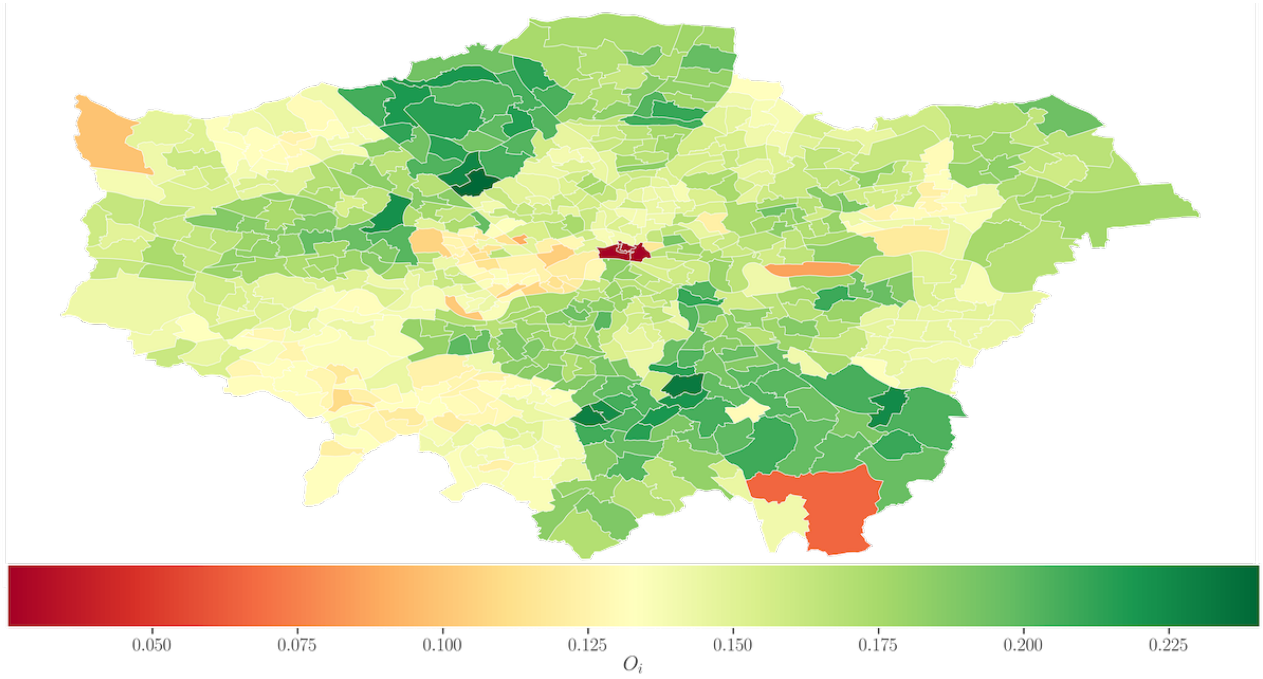


Figure 4.2: Normalised London ward-level population data from 2001 used to quantify available supply O_i at each origin i.

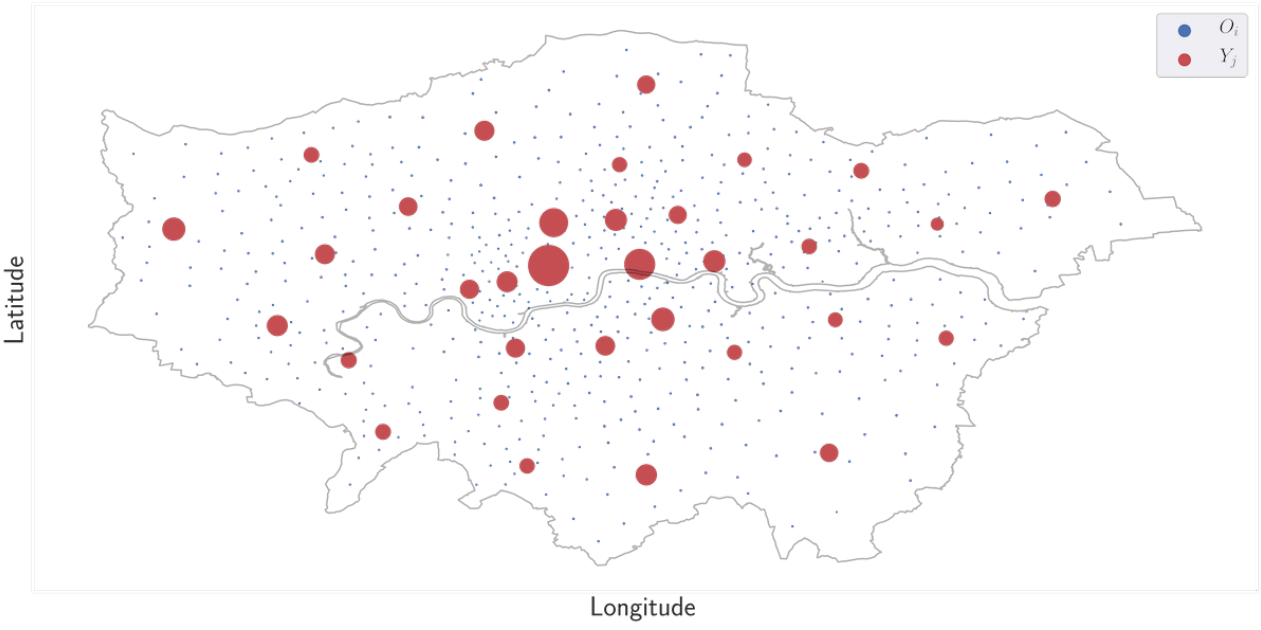
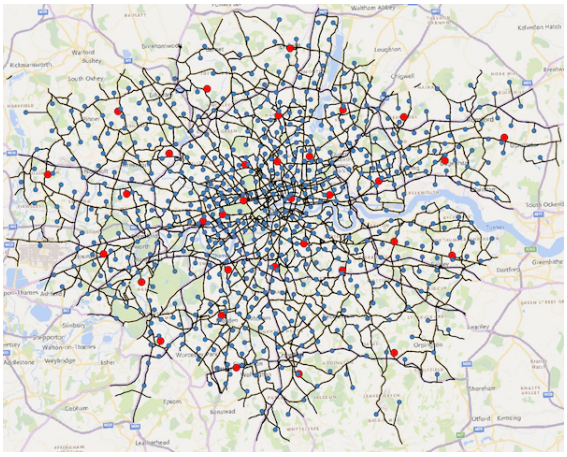
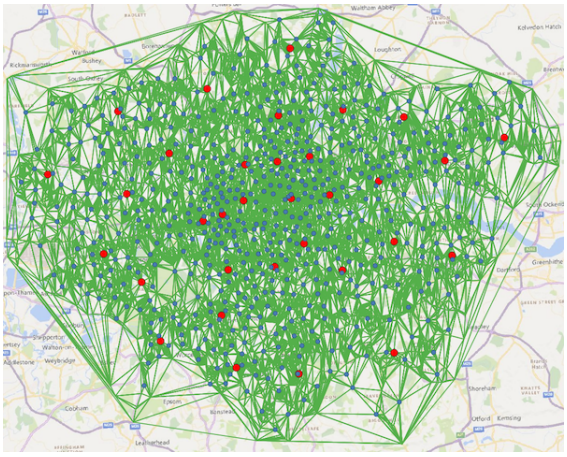


Figure 4.3: Normalised London Borough-level job availability data from 2001 used to quantify the true equilibrium destination sizes Y_j at each destination j.



(a) Greater London Area transportation network comprising of ward origins (blue dots), destination boroughs (red dots), and A,B roads (black lines).



(b) Greater London Area transportation network origins (blue dots) and destinations (red dots) connected by triangulated edges (green lines).

($M = 33$) in London. Origin supplies and destination sizes are depicted in Figures 4.2 and 4.3. During validation we aggregate our origins to Borough-level to obtain a 33×33 flow matrix. We note that job availability from 2001 is used to drive the equilibrium travel demand. At equilibrium, travel demand is $D_j = \kappa Y_j - \delta$.

4.1.1 Cost matrix

The inconvenience of travel quantified by the cost matrix $\mathbf{C}_{ij}$ is estimated in two ways. First, we follow the approach by (Dearden and Alan Wilson, 2011) and (Ellam et al., 2018) and compute a Euclidean distance-based cost matrix for all NM origin-destination pairs. We recognise that people do not travel in straight lines, but use a transportation network. Following the approach outlined in (Dearden and Alan Wilson, 2015, p.43), we construct a graph representing London's transportation network. Transportation costs are obtained from Meridian 2 data. Our constructed transportation network comprises of A and B roads inside the Greater London Area, as shown in Figure 4.4a. The vertices in the graph is the set of origins, destinations, and vertices of each road appearing in the network. To ensure that there is a path in the graph between every origin-destination pair, we first connect every origin and destination to

its nearest road vertex and compute the Euclidean distance to that edge in meters. Then, we triangulate all the vertices using the Delaunay algorithm (Chen and Xu, 2004), as illustrated in Figure 4.4b. Finally, we apply Djikstra's shortest path algorithm using Python's `networkx` library to estimate travel costs. The shortest path cost choice is justified by a cost minimisation argument, where individuals aim to find the fastest route to their destination. We assume that ticket and fuel costs are negligible. The computational cost of evaluating a transportation network-based cost matrix is $\mathcal{O}(|V|^2 \log(|V|)+|V||E|)$, where $|V| = N+M+2R$ are the number of vertices, $|E| \approx NMR$ are the number of edges, and R is the number of A and B roads in the network. No computational improvement is made on the described algorithm since the cost matrix is only computed once. Both the Euclidean-based ($\mathbf{C}_{\text{euclidean}}$) and transportation-based ($\mathbf{C}_{\text{transport}}$) are used in inference and their discrepancy in terms of inferred parameters and latent destination sizes is examined in the next sections. We refer to the Euclidean-based cost matrix as 'naive' cost matrix and the transportation-based one as 'informative' cost matrix.

4.2 Deterministic analysis

First, we solve the inverse problem in a deterministic setting by performing an R^2 analysis for a 100×100 grid of (α, β) pairs as advocated by (Dearden and Alan Wilson, 2011). The coefficient of determination is defined as

$$R^2 = 1 - \frac{SS_{\text{residual}}}{SS_{total}}, \tag{4.1}$$

where SS_{residual} and SS_{total} are the residual and total sum of squares, respectively. The equilibrium sizes are obtained from the system of SDEs in (2.8) in the limit of $\sigma \to 0$ with initialisation $\mathbf{w}_0 = \mathbf{y}$. Our results are presented in Table 4.1. The informative cost matrix yields a negligibly higher $R^2 = 0.75$, which corresponds to a slightly smaller attraction effects and higher travel cost effects. This is expected because the transportation network-based cost matrix is expected to be more informative than the the Euclidean-based one. The magnitude of discrepancy tells us that the former is not significantly more informative of travel inconvenience than the latter in deterministic settings. By calibrating the deterministic Poisson regression

Method	Cost matrix	α	β	$y_{\text{intercept}}$	R^2
SIM	$\mathbf{C}_{\text{euclidean}}$	1.144	0.08	-	0.74
	$\mathbf{C}_{\text{transport}}$	1.122	0.068	-	**0.75**
PR	$\mathbf{C}_{\text{euclidean}}$	1.39	0.00765	-10.380	-
	$\mathbf{C}_{\text{transport}}$	1.421	0.00529	-11.063	-

Table 4.1: Inferred parameters from an R^2 and Poisson regression analyses using the Euclidean distance-based and transportation network-based cost matrices in deterministic settings.

Noise level	Cost matrix	α	β	$\log(\pi(\boldsymbol{\theta}\|\mathbf{y}))$
Low	$\mathbf{C}_{\text{euclidean}}$	1.14	0.02	**-177**
	$\mathbf{C}_{\text{transport}}$	1.12	0.02	-183.7
High	$\mathbf{C}_{\text{euclidean}}$	0.24	0.08	-26
	$\mathbf{C}_{\text{transport}}$	0.24	0.06	**-25**

Table 4.2: Maximum a posterior estimates using a Laplace approximation for the normalising constant in zero observation noise settings.

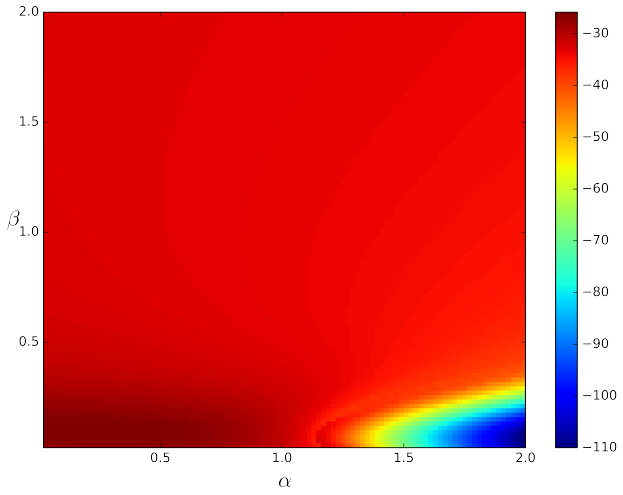
benchmark model in (2.29) using the flow matrix in Figure 4.1, we get that the transport cost matrix had higher travel cost effects and smaller attraction effects than the naive cost matrix. The existence of the $y_{\text{intercept}}$ does not allow for explicit parameter comparisons. However, the Poisson regression model agrees with our model about the fact that good data fits are found for $\alpha > 1$ and small enough β. The MAP estimates are provided in Table 4.2.

4.3 Maximum a posteriori estimation

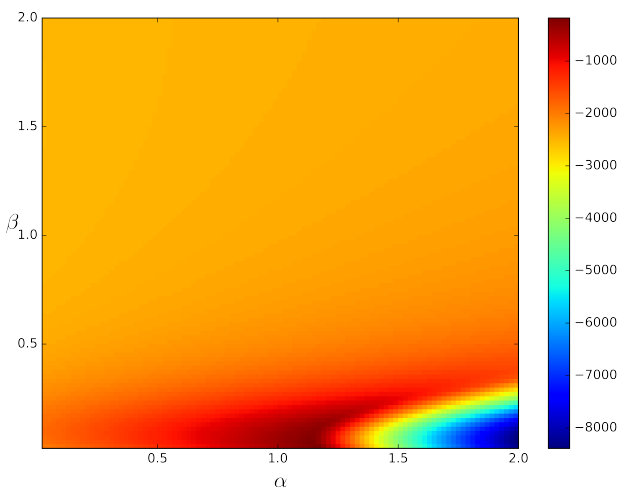
In the limit of observation noise $\lambda \to 0$, the parameter posterior marginal becomes

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) \frac{1}{Z(\boldsymbol{\theta})} \exp(-\gamma V_{\boldsymbol{\theta}}(\mathbf{x})). \tag{4.2}$$

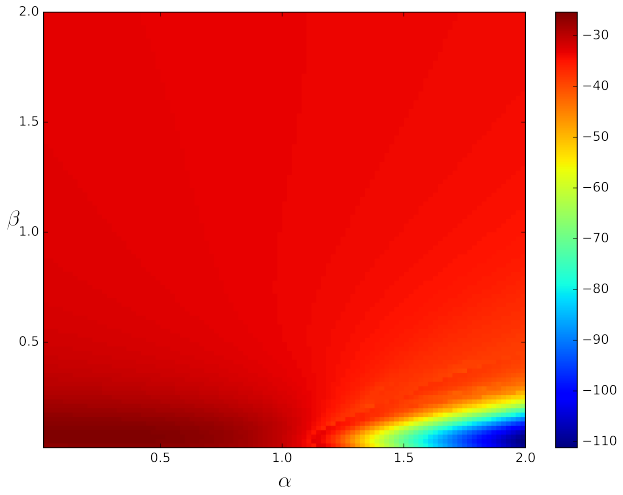
With this simplification, marginal posterior probabilities are evaluated using the Laplace approximation in Algorithm 1 for a 100×100 grid of (α, β) pairs, $\gamma = 10^2$ (high-noise regime) and



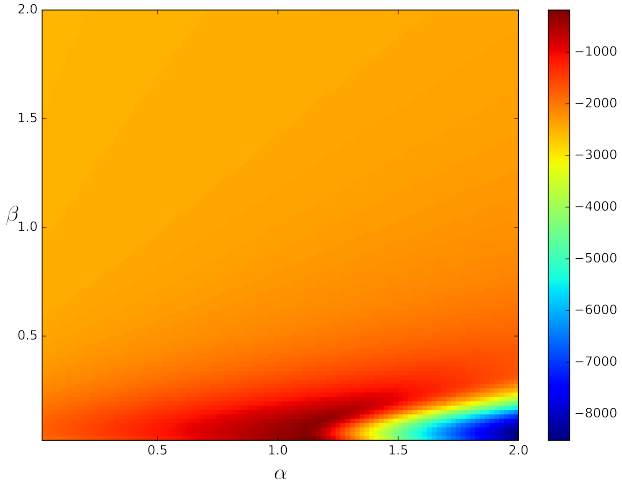
(a) high-noise maximum a posteriori estimation using $\mathbf{C}_{\text{euclidean}}$.



(b) low-noise maximum a posteriori estimation using $\mathbf{C}_{\text{euclidean}}$.



(c) high-noise maximum a posteriori estimation using $\mathbf{C}_{\text{transport}}$.



(d) low-noise maximum a posteriori estimation using $\mathbf{C}_{\text{transport}}$.

Figure 4.4: Maximum a posterior estimates for high (left) and low (right) regimes based on the Euclidean distance (top) and transportation network-based (bottom) cost matrices.

$\gamma = 10^4$ (low-noise regime) using the two cost matrices outlined in Section 4.1.1. As justified in previous chapters, we fix $\delta = 0.0118$ and $\kappa = 1.389$. We also obtain the maximum a posteriori estimate of (4.2). The results are presented in Figures 4.4a - 4.4d and Table 4.2. The low-noise models cannot explain stochastic growth, which results in inflated attraction effects. That inflation is slightly smaller for the informative cost matrix which can better explain some of this stochasticity through higher travel cost effects, as expected. Figure 4.2 depicts competing effects in the utility potential in (2.14), which indicates that the parameters α and β are correlated. We also note that in the low-noise regime the Euclidean cost matrix achieves a significantly higher likelihood than the transport cost matrix whereas in the high-noise regimes the latter matrix achieves a modestly higher likelihood. This can be attributed to the fact that the explanation of stochastic growth coupled with a better explanation of the travel effects yields a better data fit. However, the likelihood discrepancy is not that large to make this statement with certainty. In both low and high-noise scenarios, the generated samples for high values of α display a more concentrated size structure, as expected.

4.4 Latent priors

Figure 4.5: Low-noise latent size prior samples for $\alpha = 0.5, 1, 1.5, 2$ and $\beta = 0.5$ using numerically obtained estimates of the global minimum of the potential function. Origin supplies and destination sizes are shown in blue and red dots, respectively.

Next, we draw samples from the latent size prior for $\alpha = 0.5, 1, 1.5, 2$, $\beta = 0.5$ and $\gamma = 10^2, 10^4$ to verify that we producing suitable samples. The Euclidean distance-based cost matrix is chosen to generate samples since both cost matrices yield similar samples. For the low-noise regime, samples are generated for the latent sizes globally minimising the potential

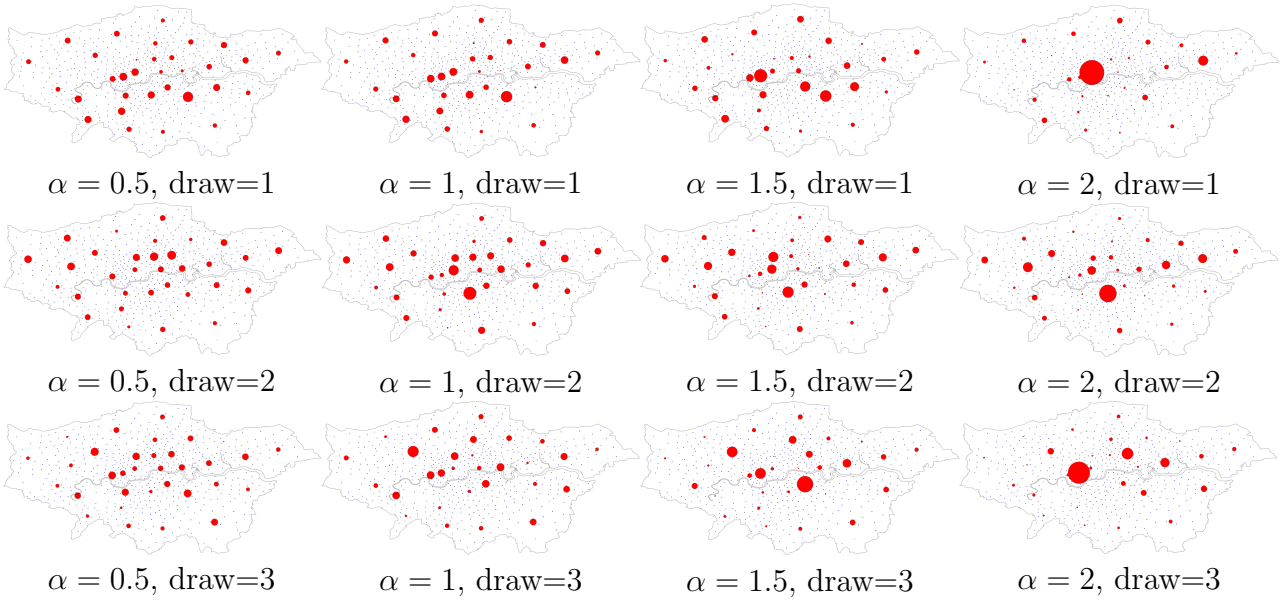


Figure 4.6: High-noise latent size prior samples for $\alpha = 0.5, 1, 1.5, 2$ and $\beta = 0.5$ using HMC sampling with parallel tempering. Origin supplies and destination sizes are shown in blue and red dots, respectively.

function $V_{\boldsymbol{\theta}}(\mathbf{x})$. The global minima are obtained numerically using the L-BFGS algorithm at $M = 33$ different initialisations. This is because in low-noise high-gamma regimes the most significant contributions in (4.2) are attained around the potential function's global minimum, as argued in Section 3.2.1. For high-noise regimes, samples are drawn approximately from the Markov chain resulting from running the HMC algorithm in 4 with parallel tempering (PT) (Liu, 2008, p.212) for $n_{mcmc} = 10000$ and five different temperatures $T = 1, 1/2, 1/4, 1/8, 1/16$. Our generated samples are illustrated in Figures 4.5 and 4.6.

4.5 Posterior marginals

We proceed by specifying observation noise to be equal to $\lambda = 0.1$, which corresponds to $\lambda/\log(M) = 6\%$ relative noise for a zone of size $1/M$. The concentration of measure in the low-noise regime does not allow us to obtain accurate importance sampling estimates. To alleviate this problem, we apply the MCMC sampling scheme in Algorithm 5 with $n_{mcmc} = 20000$ as

Noise level	Cost matrix	Parameter	μ	σ
Low	$\mathbf{C}_{\text{euclidean}}$	α	1.14	0.003
		β	0.025	0.003
High	$\mathbf{C}_{\text{euclidean}}$	α	0.26	0.2
		β	0.24	0.36
Low	$\mathbf{C}_{\text{transport}}$	α	1.135	0.004
		β	0.016	0.002
High	$\mathbf{C}_{\text{transport}}$	α	0.233	0.125
		β	0.184	0.204

Table 4.3: Inferred parameter means and standard deviations using MCMC sampling in low and high-noise regimes for Euclidean distance-based and transportation network-based cost matrices.

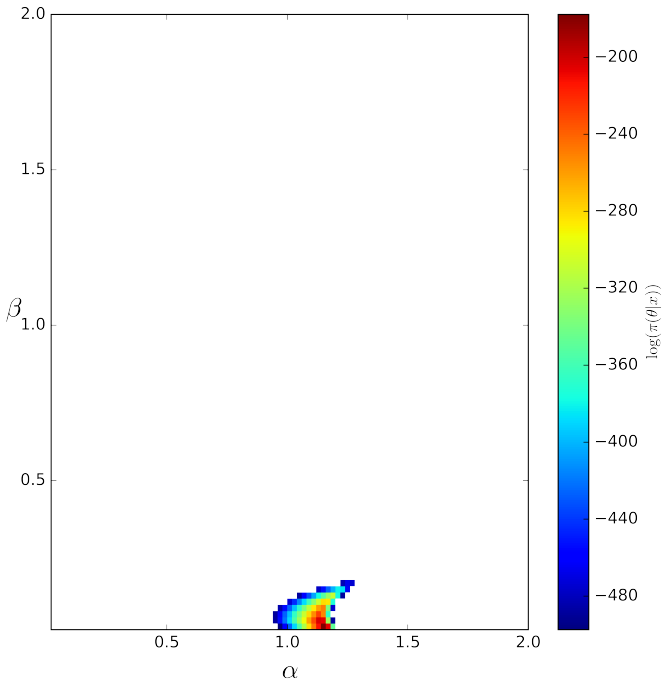


Figure 4.7: Two dimensional parameter posterior the Boltzmann-Gibbs measure collapses to in the low-noise regime.

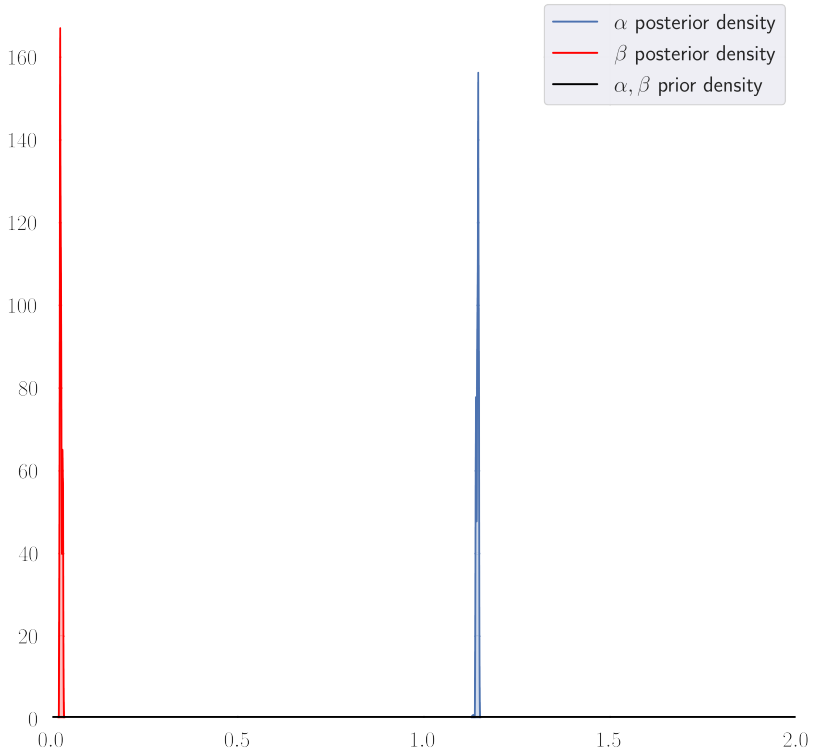


Figure 4.8: Low-noise parameter posterior empirical distributions using the Euclidean distance-based cost matrix.

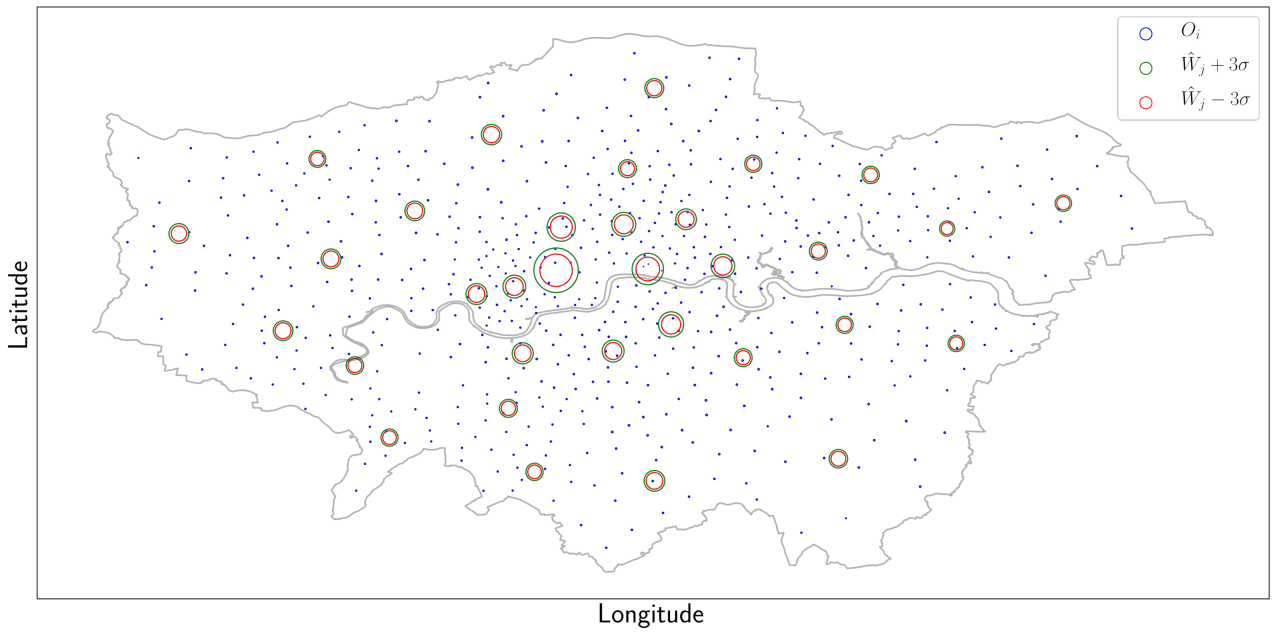
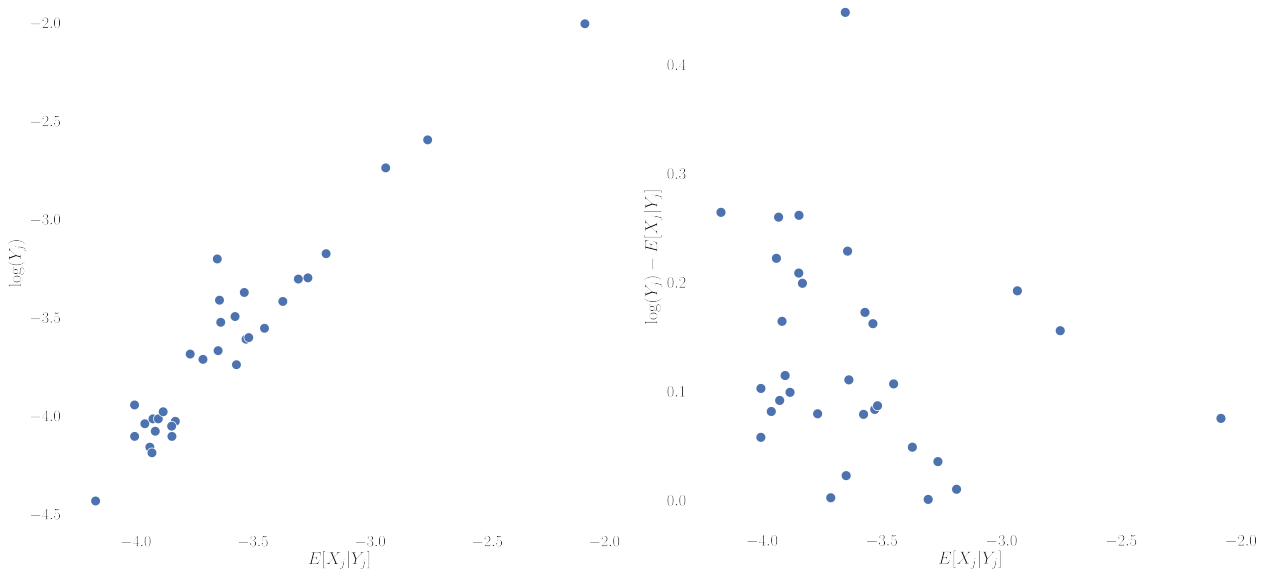


Figure 4.9: Low-noise latent destination size posterior visualisation using the Euclidean distance-based cost matrix. Upper $(\mu + 3\sigma)$ and lower $(\mu - 3\sigma)$ credible interval bounds are represented by green and blue rings, respectively.



(a) True destination sizes versus latent size posterior mean predictions plot for the low-noise regime.

(b) Posterior mean latent size residuals versus latent size posterior mean predictions plot for the low-noise regime.

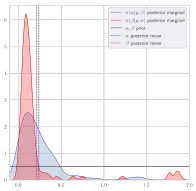


Figure 4.10: High-noise parameter posterior empirical distributions using the Euclidean distance-based cost matrix.

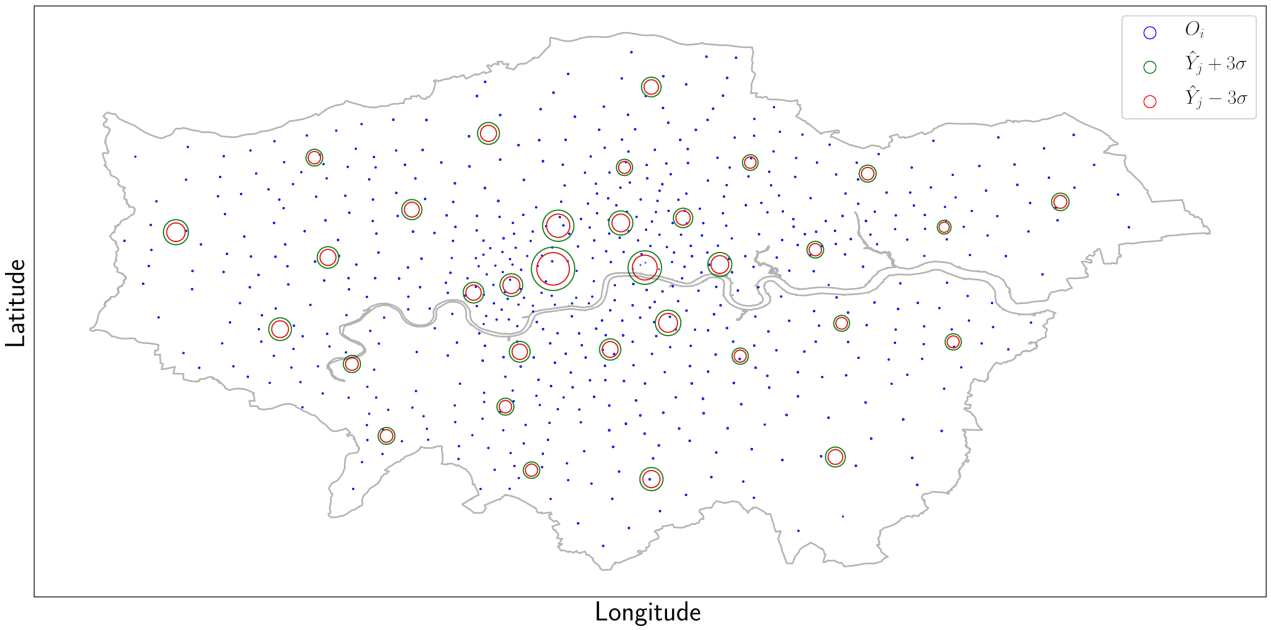
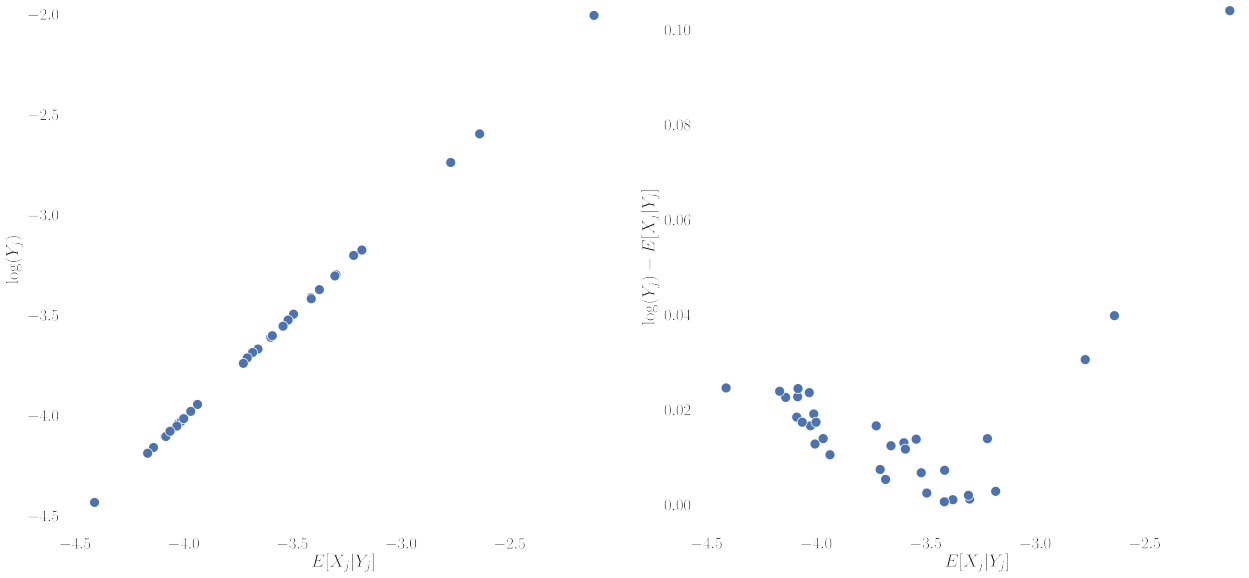


Figure 4.11: High-noise latent destination size posterior visualisation using the Euclidean distance-based cost matrix. Upper ($\mu + 3\sigma$) and lower ($\mu - 3\sigma$) credible interval bounds are represented by green and blue rings, respectively.



(a) True destination sizes versus latent size posterior mean predictions plot for the high-noise regime.

(b) Posterior mean latent size residuals versus latent size posterior mean predictions plot for the high-noise regime.

summarised in Section 3.5. However, we did not manage to get enough independent posterior samples using either cost matrices as the auto-correlation did not decay sufficiently fast. In fact, the auto-correlation dropped below 0.2 after more than 100 steps, giving us at most 200 independent samples. The small sample size cannot guarantee convergence in (3.5). The reason behind this shortcoming is because of the irregular shape of the measure the Boltzmann-Gibbs measure collapses to (see Figure 4.7). Learning a parameter posterior with a 'boomerang'-like shape such as the one in Figure 4.7 requires a more sophisticated transition kernel than the asymmetric random-walk proposed in Algorithm 5. Although in theory we would expect the mass in Figure 4.7 to resemble a Dirac mass, this is not case due to the existence of parameter discontinuities (Dearden and Alan Wilson, 2011). As explained before, small changes in the parameters can often lead to sudden changes in the potential function, which induces local discontinuities and is evidenced by potential function shape irregularities. In MCMC sampling the parameter and latent size sample acceptance rates ranges from 43-46% and above 96%, respectively.

We present the posterior marginals for the low-noise regime using the Euclidean cost matrix in Figures 4.8 and 4.9 as well as the latent size prediction and residual plots in Figures 4.10a and 4.10b. The inferred parameter values are in line with the maximum a posterior estimates in Figures 4.4b and 4.4d. Despite the lack of observation noise in the model, the low-noise HMC sample summary statistics of latent posteriors provides a reasonable fit to the data. The residual plot in 4.10b indicates a degree of heteroskedasticity, which is natural since the model's low error structure cannot explain all of the stochastic growth. The attractiveness effects are much more dominant in the low-noise model as its inflexibility does not provide an explanation of the stochastic growth of the destination sizes. The assumption of homogeneous noise is not severely violated, although the predictions plot in Figure 4.10a is significantly dispersed. Moreover, the informative cost matrix mildly suppresses the attraction effects and suggests a hypothesis under which travel inconvenience is more indicative of the flows and travel demand. Despite the small degree of uncertainty, parameter posterior marginals in Figure 4.8 end up being overconfident (due to the concentration of measure) by not adequately accounting for the uncertainty involved in the data generating process.

In the high-noise regime, we leverage the pseudo-marginal MCMC sampling scheme outlined in Algorithm 5 with $n_{mcmc} = 10000$. We obtain parameter and latent size sample acceptance rates ranging from 37-44% and above 96%, respectively. The sign of the normalising constant in (2.11) is positive 98% of the time. The empirical auto-correlation was below 0.2 after about 20 steps, generating about 500 independent posterior samples. The posterior marginals for the high-noise regime using the Euclidean cost matrix in Figures 4.10 and 4.11 as well as the latent size prediction and residual plots in Figures 4.12a and 4.12b. The sampling procedures are presented only for the naive cost matrix as there are not major discrepancies with running the same procedures using the informative cost matrix. In terms of the inferred parameters, they also agree with the MAP estimates in Figures 4.4a and 4.4c. Parameter uncertainty is much higher in the high-noise regime than in the noise regime. This is because the large noise standard deviation makes more data fit hypotheses plausible, inducing variability in the empirical posterior marginals in Figure 4.10. The transportation network-based cost matrix absorbs some of that variability because of the informative nature of travel cost effects.

γ	λ	Cost matrix	Method	Algorithm reference	SRMSE
∞	0	$\mathbf{C}_{\text{euclidean}}$	R^2	-	2.236
10^4	0	$\mathbf{C}_{\text{euclidean}}$	Laplace MAP	1	2.297
10^4	0.1	$\mathbf{C}_{\text{euclidean}}$	MCMC	5	**2.204**
10^2	0	$\mathbf{C}_{\text{euclidean}}$	Laplace MAP	1	2.552
10^2	0.1	$\mathbf{C}_{\text{euclidean}}$	MCMC	5	2.292
∞	0	$\mathbf{C}_{\text{transport}}$	R^2	-	2.233
10^4	0	$\mathbf{C}_{\text{transport}}$	Laplace MAP	1	2.276
10^4	0.1	$\mathbf{C}_{\text{transport}}$	MCMC	5	**2.208**
10^2	0	$\mathbf{C}_{\text{transport}}$	Laplace MAP	1	2.326
10^2	0.1	$\mathbf{C}_{\text{transport}}$	MCMC	5	2.295

Table 4.4: Inferred flow matrix SRMSEs computed for various γ, observation noises λ, cost matrices and methods.

However, the large noise to signal ratio makes high-noise MAP estimates provide less insights in model calibration. A stronger prior on parameters and a more informative specification of the error structure is required to reduce uncertainty and improve the accuracy of parameter inference. Finally, latent size predictions are much better in the high-noise regime and there is a smaller degree of heteroskedasticity involved, which is justified by the additional level of uncertainty accounted for.

4.6 Flow matrix validation

In order to provide a complete answer to the research question formulated in Chapter 1, we reconstruct the entire flow matrix by substituting the inferred parameters in (2.5), aggregate the matrix on a borough-to-borough level ($N = M = 33$) and validate it against the 2001 commuter flow matrix in Figure 4.1. We use the standardised root mean square error (SRMSE)

(Oshan, 2016) as a validation metric:

$$SRMSE = \frac{\frac{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M}(T_{ij}-\hat{T}_{ij})^2}}{NM}}{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}T_{ij}}{NM}}, \tag{4.3}$$

where $\hat{T}_{ij}$ and T_{ij} are the inferred and actual flow matrices. The mean of the true (observed) flows is used to standardise the root mean square error in the numerator. Aggregating a ward-to-borough matrix to a borough-to-borough matrix entails biases. However, we do not concern ourselves with the absolute SRMSE and are only interested in the relative discrepancies between SRMSEs using different methods under different noise regimes and cost matrices. We summarise the SRMSE of each flow matrix constructed using each method outlined in the previous Chapter in Table 4.4.

The transportation network-based cost matrix yields smaller SRMSEs than the Euclidean distance-based cost matrix except for the latent sizes and parameters inferred from MCMC sampling which are similar for both cost matrices. This indicates that the travel cost effect are better explained by the informative cost matrix to some extent. The low-noise regime MCMC-inferred flow matrix seems to produce the smallest SRMSE, but is only marginally better than its high-noise counterpart. This may be attributed to the biases induced from matrix aggregation, so it is imperative that a ward-to-borough matrix is obtained or the analysis is performed for a borough-to-borough OD matrix. However, the SRMSE discrepancies are not highly significant except for the high-noise Laplace approximated MAP estimated flow matrix, which performs poorly in matrix reconstruction for both cost matrices.

Chapter 5

Discussion

5.1 Conclusions

In summary, we have introduced a novel application of stochastic spatial interaction modelling to urban travel demand modelling. We have successfully simulated travel demand scenarios generated from employment characteristics in London's Boroughs and have calibrated our models under various noise regimes. Our approach has also provided the basis for a unified framework of travel demand evolution modelling which addresses some of the limitations of the four step travel demand modelling framework in Figure 1.1, such as its computationally expensive iterative procedure. Socioeconomic data together stochastic modelling of aggregate spatial interactions has enables us to tackle trip generation and distribution while also modelling the evolution of the latter (i.e. steps one and two of the four-step model in Figure 1.1). However, we have not addressed mode and route choice. Fortunately, dis-aggregating the spatial interaction model to compute OD flows by mode of transport and choice of route is fully compatible with the framework we have laid out.

Our proposed modelling framework provides an attractive alternative to discrete choice modelling, which relies on large volumes of commercially licensed flow data. The computational complexities of model calibration in the presence of observation noise are linear in the number of origins N while the low-noise MCMC is cubic in the number of destinations M and the high-noise regime is linear in M. This is a reasonable computational cost for computing the evolution of travel demand since a single iteration of trip distribution using the multinomial probit DCM takes $\mathcal{O}(FNM^2)$, where F is the sum of other factors affecting the computational complexity of the multinomial probit model (Daganzo, Bouthelier, and Sheffi, 1977). Finally, the effect

of the informative cost matrix has not been substantial as both the inferred parameters and SRMSEs of inferred flow matrices have only been marginally better than the ones obtained using a naive cost matrix.

5.2 Further research

The time-varying and stochastic nature of transportation problems calls for a variety of further research directions some of which will be pursued at the beginning of the PhD. First, there is a number of theoretical considerations that have to be made, such as convexity of multidimensional potential function maps and the presence of discontinuities (Dearden and Alan Wilson, 2011) in stochastic settings. These theoretical aspects will shed light on the stochastic model's behaviour and allow us to enrich/adjust our computational inference toolbox. Secondly, the scalability of our modelling framework can improved by seeking computational savings and ensuring the model does not suffer from the curse of dimensionality. For example, as the number of destinations M grows, the dimensionality of the intractable term $Z(\boldsymbol{\theta})$ increases, which makes it harder to accurately estimate. Therefore, more tractable methods should be developed in the future to ensure our model scales well.

Although the case has been made for economic structure driving travel demand (Batten and Boyce, 1987), employability is not the only latent force driving doing that. Incorporating additional latent forces and encoding them in an informative fashion in our modelling framework can help our model explain more of the stochastic variation currently attributed to noise. Moreover, the comparison between SIM and DCMs should be made more explicit. Although the computational complexities of the two model classes were compared, flow validation and inferred equilibrium demand was not. Such a comparison would be necessary to formally assess and compare the effectiveness of the two approaches. In addition, we only cover steps one and two of the four-step model in Figure 1.1. The works of (A.G. Wilson, 1967) and (A. G. Wilson, 1971) have illustrated that we can dis-aggregate the SIM to account for mode of transport and route choices. Regarding the informative cost matrix, we used London's A and B roads to compute a shortest-path based cost matrix. However, commuters make frequent use of the

tube in the transport across London. Therefore, it would be more realistic to include London underground's map with empirical cost estimates that are adjusted for speed of transport by taking into account average travel times.

Our proposed model is able to encode spatial but not temporal interactions of travellers. The effect of time on travel demand is crucial when estimating travel demand patterns and their evolution in time. Our assumptions of static origin supplies, travel costs and destination sizes are therefore limiting in dynamic travel demand modelling. One fundamental way to address this issue is to solve the filtering problem in (2.8) by using time series data instead of cross sectional. Also, the notion of a fixed stationary equilibrium seems unrealistic in the context of transportation. Everyday policy changes and other types of political, economic and business interventions introduce disturbances to the dynamical system of transportation networks. Hence, the Harris and Wilson SDEs whose stationary equilibrium distribution (i.e. the Boltzmann-Gibbs measure in (2.11)) is only available in closed form are not suitable for modelling multiple dynamically changing equilibria. To account for the dynamic nature of one or multiple equilibria, a different set of stochastic differential equations should be used (Tahmasbi and Hashemi, 2014) (Polson and Sokolov, 2015). The state variables of SDEs such as the Hull-White SDE (Tahmasbi and Hashemi, 2014) have closed form probability distributions at each time step, which allows us to make exact and computationally efficient inference potentially in real-time.

Bibliography

Aggarwal, Charu C. (2020). "Linear Algebra and Optimization: An Introduction". In: *Linear Algebra and Optimization for Machine Learning*. Cham: Springer International Publishing, pp. 1–40. DOI: 10.1007/978-3-030-40344-7{_}1. URL: http://link.springer.com/10.1007/978-3-030-40344-7_1 (cit. on p. 74).

Anas, Alex (Feb. 1983). "Discrete choice theory, information theory and the multinomial logit and gravity models". In: *Transportation Research Part B* 17.1, pp. 13–23. ISSN: 01912615. DOI: 10.1016/0191-2615(83)90023-1 (cit. on pp. 13, 67).

Andrieu, Christophe and Gareth O. Roberts (Apr. 2009). "The pseudo-marginal approach for efficient Monte Carlo computations". In: *The Annals of Statistics* 37.2, pp. 697–725. ISSN: 0090-5364. DOI: 10.1214/07-AOS574. URL: https://projecteuclid.org/euclid.aos/1236693147 (cit. on p. 34).

Baltas, George and Peter Doyle (Feb. 2001). "Random utility models in marketing research: a survey". In: *Journal of Business Research* 51.2, pp. 115–125. ISSN: 01482963. DOI: 10.1016/S0148-2963(99)00058-2. URL: https://linkinghub.elsevier.com/retrieve/pii/S0148296399000582 (cit. on p. 4).

Batten, David F. and David E. Boyce (Jan. 1987). "Spatial interaction, transportation, and interregional commodity flow models". In: *Handbook of Regional and Urban Economics*. Vol. 1. Elsevier. Chap. 9, pp. 357–406. DOI: 10.1016/S1574-0080(00)80012-7. URL: https://linkinghub.elsevier.com/retrieve/pii/S1574008000800127 (cit. on pp. 6, 38, 56).

Ben-Akiva, Moshe E. and John L. Bowman (1998). "Activity Based Travel Demand Model Systems". In: *Equilibrium and Advanced Transportation Modelling*. Boston, MA: Springer US, pp. 27–46. DOI: 10.1007/978-1-4615-5757-9{_}2. URL: http://link.springer.com/10.1007/978-1-4615-5757-9_2 (cit. on p. 4).

Butler, Ronald W. (Jan. 2007). *Saddlepoint approximations with applications*. Cambridge University Press, pp. 1–564. ISBN: 9780511619083. DOI: 10.1017/CBO9780511619083 (cit. on p. 23).

Chang, Kuang-Hua (Jan. 2015). "Reliability Analysis". In: *e-Design*. Elsevier, pp. 523–595. DOI: 10.1016/B978-0-12-382038-9.00010-7. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780123820389000107 (cit. on pp. 14, 74).

Chen, Long and Jin-chao Xu (2004). "Optimal Delaunay Triangulations". In: *Journal of Computational Mathematics* 22.2, pp. 299–308. ISSN: 02549409, 19917139. URL: http://www.jstor.org/stable/43693155 (cit. on p. 42).

Crooks, Gavin E (2006). *Beyond Boltzmann-Gibbs statistics: Maximum entropy hyperensembles out-of-equilibrium*. Tech. rep. (cit. on p. 68).

Daganzo, Carlos F, Fernando Bouthelier, and Yosef Sheffi (1977). "Multinomial Probit and Qualitative Choice: A Computationally Efficient Algorithm". In: *Transportation Science* 11.4, pp. 338–358 (cit. on p. 55).

Dearden, Joel and Alan Wilson (Apr. 2011). "A Framework for Exploring Urban Retail Discontinuities". In: *Geographical Analysis* 43.2, pp. 172–187. ISSN: 00167363. DOI: 10.1111/j.1538-4632.2011.00812.x. URL: http://doi.wiley.com/10.1111/j.1538-4632.2011.00812.x (cit. on pp. 6, 17, 36, 41, 42, 51, 56).

— (Mar. 2015). *Explorations in Urban and Regional Dynamics*. Routledge. ISBN: 9781315779126. DOI: 10.4324/9781315779126. URL: https://www.taylorfrancis.com/books/9781317698531 (cit. on p. 41).

Department for Transport (2019). *Transport Statistics Great Britain: 2019*. Tech. rep. URL: www.gov.uk/dft (cit. on p. 3).

Ellam, L. et al. (May 2018). "Stochastic modelling of urban structure". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 474.2213, p. 20170700. ISSN: 1364-5021. DOI: 10.1098/rspa.2017.0700. URL: https://royalsocietypublishing.org/doi/10.1098/rspa.2017.0700 (cit. on pp. vii, 6, 7, 12, 15, 17, 32, 36, 41).

Feller, William (1957). *An introduction to probability theory and its applications*. John Wiley & Sons, Inc, pp. 50–53. ISBN: 9780471257097 (cit. on p. 66).

Harris, B and A G Wilson (Apr. 1978). "Equilibrium Values and Dynamics of Attractiveness Terms in Production-Constrained Spatial-Interaction Models". In: *Environment and Planning A: Economy and Space* 10.4, pp. 371–388. ISSN: 0308-518X. DOI: 10.1068/a100371. URL: http://journals.sagepub.com/doi/10.1068/a100371 (cit. on pp. 5, 11, 17).

Jaynes, E. T. (May 1957). "Information Theory and Statistical Mechanics". In: *Physical Review* 106.4, pp. 620–630. ISSN: 0031-899X. DOI: 10.1103/PhysRev.106.620. URL: https://link.aps.org/doi/10.1103/PhysRev.106.620 (cit. on p. 67).

Jong, Gerard de et al. (Nov. 2007). "The logsum as an evaluation measure: Review of the literature and new results". In: *Transportation Research Part A: Policy and Practice* 41.9, pp. 874–889. ISSN: 09658564. DOI: 10.1016/j.tra.2006.10.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0965856407000316 (cit. on p. 13).

Kim, Hoon and Jacob J. Oleson (Dec. 2007). "A Bayesian Dynamic Spatio-Temporal Interaction Model: An Application to Prostate Cancer Incidence". In: *Geographical Analysis* 40.1, pp. 77–96. ISSN: 00167363. DOI: 10.1111/j.0016-7363.2007.00714.x. URL: http://doi.wiley.com/10.1111/j.0016-7363.2007.00714.x (cit. on p. 2).

Lasota, Andrzej and Michael C. Mackey (1994). *Chaos, Fractals, and Noise*. Vol. 97. Applied Mathematical Sciences. New York, NY: Springer New York. ISBN: 978-1-4612-8723-0. DOI: 10.1007/978-1-4612-4286-4. URL: http://link.springer.com/10.1007/978-1-4612-4286-4 (cit. on p. 13).

Liu, Jun S (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media (cit. on p. 46).

Lyne, Anne-Marie et al. (Nov. 2015). "On Russian Roulette Estimates for Bayesian Inference with Doubly-Intractable Likelihoods". In: *Statistical Science* 30.4, pp. 443–467. ISSN: 0883-4237. DOI: 10.1214/15-STS523. URL: http://arxiv.org/abs/1306.4032%20http://dx.doi.org/10.1214/15-STS523%20http://projecteuclid.org/euclid.ss/1449670853 (cit. on p. 25).

Marsden, Greg et al. (2018). *All change? The First Report of the Commission on Travel Demand*. ISBN: 978-1-899650-83-5. URL: www.demand.ac.uk/commission- (cit. on p. 2).

McFadden, Daniel (1980). *Econometric Models for Probabilistic Choice Among Products*. DOI: 10.2307/2352205. URL: https://www.jstor.org/stable/pdf/2352205.pdf (cit. on p. 20).

McFadden, Daniel et al. (1973). "Conditional logit analysis of qualitative choice behavior". In: (cit. on p. 4).

McFadden, Daniel and Kenneth Train (Sept. 2000). "Mixed MNL models for discrete response". In: *Journal of Applied Econometrics* 15.5, pp. 447–470. ISSN: 0883-7252. DOI: 10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/1099-1255%28200009/10%2915%3A5%3C447%3A%3AAID-JAE570%3E3.0.CO%3B2-1%20https://onlinelibrary.wiley.com/doi/abs/10.1002/1099-1255%28200009/10%2915%3A5%3C447%3A%3AAID-JAE570%3E3.0.CO%3B2-1%20https://onlinelibrary. (cit. on p. 20).

Mladenovic, Milos and Aleksandar Trifunovic (2014). "The Shortcomings of the Conventional Four Step Travel Demand Forecasting Process". In: *Journal of Road and Traffic Engineering* (cit. on p. 5).

Murphy, Kevin P (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press. ISBN: 0262018020 (cit. on p. 76).

Murray, Iain, Zoubin Ghahramani, and David MacKay (June 2012). "MCMC for doubly-intractable distributions". In: *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence, UAI 2006*, pp. 359–366. URL: http://arxiv.org/abs/1206.6848 (cit. on pp. 22, 25).

Neal, Radford M. (Mar. 1998). "Annealed Importance Sampling". In: *Statistics and Computing* 11.2, pp. 125–139. URL: http://arxiv.org/abs/physics/9803008 (cit. on pp. 26, 27, 30).

— (June 2012). "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo*, pp. 113–162. URL: http://arxiv.org/abs/1206.1901 (cit. on p. 27).

Nielsen, Frank and Ke Sun (Dec. 2016). "Guaranteed Bounds on Information-Theoretic Measures of Univariate Mixtures Using Piecewise Log-Sum-Exp Inequalities". In: *Entropy* 18.12,

p. 442. ISSN: 1099-4300. DOI: 10.3390/e18120442. URL: http://www.mdpi.com/1099-4300/18/12/442 (cit. on p. 74).

Nocedal, Jorge and Stephen Wright (2006). *Numerical optimization*. Springer Science & Business Media (cit. on p. 23).

Oshan, Taylor M (Dec. 2016). "A primer for working with the Spatial Interaction modeling (SpInt) module in the python spatial analysis library (PySAL)". In: *REGION* 3.2, p. 11. ISSN: 2409-5370. DOI: 10.18335/region.v3i2.175. URL: http://dx.doi.org/10.18335/region.v3i2.175%20http://openjournals.wu.ac.at/ojs/index.php/region/article/view/175 (cit. on pp. 19, 54).

Pavliotis, Grigorios A. (2014). *Stochastic Processes and Applications*. Vol. 60, pp. 1–345. ISBN: 978-1-4939-1322-0. DOI: 10.1007/978-1-4939-1323-7 (cit. on pp. 11, 12, 70).

Polson, Nicholas and Vadim Sokolov (Dec. 2015). "Bayesian analysis of traffic flow on interstate I-55: The LWR model". In: *The Annals of Applied Statistics* 9.4, pp. 1864–1888. ISSN: 1932-6157. DOI: 10.1214/15-AOAS853. URL: https://projecteuclid.org/euclid.aoas/1453993096%20http://projecteuclid.org/euclid.aoas/1453993096 (cit. on p. 57).

Rijk, F J A and A C F Vorst (Apr. 1983). "On the Uniqueness and Existence of Equilibrium Points in an Urban Retail Model". In: *Environment and Planning A: Economy and Space* 15.4, pp. 475–482. ISSN: 0308-518X. DOI: 10.1068/a150475. URL: http://journals.sagepub.com/doi/10.1068/a150475 (cit. on p. 22).

Robert, Christian and George Casella (2013). *Monte Carlo statistical methods*. Springer Science & Business Media (cit. on p. 32).

Roberts, Gareth O. and Richard L. Tweedie (Dec. 1996). "Exponential Convergence of Langevin Distributions and Their Discrete Approximations". In: *Bernoulli* 2.4, p. 341. ISSN: 13507265. DOI: 10.2307/3318418. URL: https://www.jstor.org/stable/3318418?origin=crossref (cit. on p. 70).

Saputro, Dewi Retno Sari and Purnami Widyaningsih (2017). "Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method for the parameter estimation on geographically weighted ordinal logistic regression model (GWOLR)". In: vol. 1868, p. 040009. ISBN: 9780735415485. DOI: 10.1063/1.4995124. URL: https://doi.org/10.1063/

1.4887583https://doi.org/10.1063/1.4995125https://doi.org/10.1063/1.4995126%20http://aip.scitation.org/doi/abs/10.1063/1.4995124 (cit. on p. 23).

Small, Kenneth A. (June 1994). "Approximate generalized extreme value models of discrete choice". In: *Journal of Econometrics* 62.2, pp. 351–382. ISSN: 03044076. DOI: 10.1016/0304-4076(94)90028-0. URL: https://linkinghub.elsevier.com/retrieve/pii/0304407694900280 (cit. on p. 4).

Tahmasbi, Rasool and S. Mehdi Hashemi (Feb. 2014). "Modeling and Forecasting the Urban Volume Using Stochastic Differential Equations". In: *IEEE Transactions on Intelligent Transportation Systems* 15.1, pp. 250–259. ISSN: 1524-9050. DOI: 10.1109/TITS.2013.2278614. URL: http://ieeexplore.ieee.org/document/6588933/ (cit. on p. 57).

Train, Kenneth (2002). *Discrete Choice Methods with Simulation*. Cambridge University Press, p. 388 (cit. on p. 4).

Walker, Joan and Moshe Ben-Akiva (July 2002). "Generalized random utility model". In: *Mathematical Social Sciences* 43.3, pp. 303–343. ISSN: 01654896. DOI: 10.1016/S0165-4896(02)00023-9. URL: https://linkinghub.elsevier.com/retrieve/pii/S0165489602000239 (cit. on p. 4).

Wei, Colin and Iain Murray (Oct. 2016). "Markov Chain Truncation for Doubly-Intractable Inference". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*. URL: http://arxiv.org/abs/1610.05672 (cit. on p. 25).

Williams, H.C.W.L. (1977). "On the formation of travel demand models and economic evaluation measures of user benefit". In: *Environment and Planning A* 9, pp. 285–344 (cit. on pp. 13, 14).

Wilson, A G (1971). "A family of spatial interaction models, and associated developments t". In: 3, pp. 1–32 (cit. on p. 56).

— (Nov. 1967). "A statistical theory of spatial distribution models". In: *Transportation Research* 1.3, pp. 253–269. ISSN: 00411647. DOI: 10.1016/0041-1647(67)90035-4. URL: https://linkinghub.elsevier.com/retrieve/pii/0041164767900354 (cit. on pp. 5, 56).

Wilson, Alan (Oct. 2010). "Entropy in Urban and Regional Modelling: Retrospect and Prospect". In: *Geographical Analysis* 42.4, pp. 364–394. ISSN: 00167363. DOI: 10.1111/j.1538-4632.2010.00799.x. URL: http://doi.wiley.com/10.1111/j.1538-4632.2010.00799.x (cit. on pp. 6, 13).

— (Jan. 2013). *Entropy in Urban and Regional Modelling (Routledge Revivals)*. Routledge. ISBN: 9780203142608. DOI: 10.4324/9780203142608. URL: https://www.taylorfrancis.com/books/9780203142608 (cit. on p. 66).

Appendices

Appendix A

Maximum entropy derivations

A.1 Spatial interaction model flow

Consider the total number of ways W of assigning T identical people/vehicles in total from N origin zones to M destinations, where $T \coloneqq \sum_{i=1}^{N} \sum_{j=1}^{M} T_{ij}$. This is equivalent to assigning people to "boxes" (Alan Wilson, 2013), where each box corresponds to each one of the NM origin-destination pairs. By rules of permutations it follows that

$$W' = \frac{T!}{\prod_{i=1}^{N} \prod_{j=1}^{M} T_{ij}!}. \tag{A.1}$$

The most probable assignment T_{ij} can be found by maximising W' subject to constraints (2.1)-(2.4). Taking logs[1] and using Stirling's approximation[2] (Feller, 1957, p.50-53), maximising W' is equivalent to maximising

$$\begin{aligned}
\log(W') &= \log(T!) - \sum_{i=1}^{N} \sum_{j=1}^{M} \log(T_{ij}!) \\
&\approx T \log(T) - T - \sum_{i=1}^{N} \sum_{j=1}^{M} \left(T_{ij} \log(T_{ij}) - T_{ij} \right) \\
&\overset{\text{def}}{=} T \log(T) - \sum_{i=1}^{N} \sum_{j=1}^{M} T_{ij} \log(T_{ij})
\end{aligned} \tag{A.2}$$

Since T is constant, only the second term of (A.2) needs to be maximised. This term is known as Shannon's entropy in information theory and is denoted by $\mathcal{H}$. Therefore, the

[1] Any monotonic transformation of W' will in fact yield the same result.

[2] There are a few variants of this approximation. We use $\log(n!) \approx n \log(n) - n$.

entropy-maximising argument is equivalent to the information-minimising argument (Anas, 1983). Therefore, using Lagrange multipliers the objective function becomes

$$\mathcal{L}_{ij}(T_{ij}) = -\sum_{i=1}^{N}\sum_{j=1}^{M} T_{ij}\log(T_{ij}) - \sum_{i=1}^{N}\lambda_i\left(\sum_{j=1}^{M} T_{ij} - O_i\right) - \beta\left(\sum_{i=1}^{N}\sum_{j=1}^{M} T_{ij}c_{ij} - C\right) - \alpha\left(\sum_{i=1}^{N}\sum_{j=1}^{M} T_{ij}\log(W_j) - B\right) \tag{A.3}$$

Setting $\frac{\mathrm{d}\mathcal{L}_{ij}(T_{ij})}{\mathrm{d}T_{ij}} = 0 \quad \forall\,(i,j) \in \{1,\dots,N\}\times\{1,\dots,M\}$ yields

$$\begin{aligned} 0 &= -\log(T_{ij}) - \lambda_i - \beta c_{ij} - \alpha'\log(W_j) \\ \therefore T_{ij} &= \exp\left(-\lambda_i - \beta c_{ij} - \alpha'\log(W_j)\right). \end{aligned} \tag{A.4}$$

Letting $\alpha \coloneqq -\alpha'$ and substituting (2.1) in (A.4) yields

$$\begin{aligned} O_i &= \exp(-\lambda_i)\sum_{k=1}^{M} W_k^{\alpha}\exp\left(-\beta c_{ik}\right) \\ \therefore \exp(-\lambda_i) &= \frac{O_i}{\sum_{k=1}^{M} W_k^{\alpha}\exp\left(-\beta c_{ik}\right)} \end{aligned}$$

and therefore by (A.4)

$$T_{ij} = \frac{O_i W_j^{\alpha}\exp\left(-\beta c_{ij}\right)}{\sum_{k=1}^{M} W_k^{\alpha}\exp\left(-\beta c_{ik}\right)}. \tag{A.5}$$

The second derivative is

$$\frac{\mathrm{d}^2\,\mathcal{L}_{ij}(T_{ij})}{\mathrm{d}T_{ij}^2} = -\frac{1}{T_{ij}} < 0 \quad \forall\, 1 \le i \le N, 1 \le j \le M$$

and therefore (A.5) is indeed a maximum. □

A.2 Steady-state distribution

Let the probability density of the state $\mathbf{X}(t)$ of the SDE in (2.8) be denoted as $\rho(\mathbf{x},t)$ and let $\rho_\infty(\mathbf{x})$ denote the steady-state distribution of the state vector $\mathbf{X}$. The maximum entropy probability distribution (Jaynes, 1957) for $\rho_\infty(\mathbf{x})$ can be found be maximising

$$\mathcal{H}\left(\rho_\infty(\mathbf{x})\right) = -\int_{\mathcal{R}^M} \rho_\infty(\mathbf{x}') \log\left(\frac{\rho_\infty(\mathbf{x}')}{m(\mathbf{x}')}\right) \mathrm{d}\mathbf{x}', \tag{A.6}$$

subject to the following constraints:

$$\int_{\mathcal{R}^M} \rho_\infty(\mathbf{x}')\, \mathrm{d}\mathbf{x}' = 1, \tag{A.7}$$

and

$$\int_{\mathcal{R}^M} \rho_\infty(\mathbf{x}') V(\mathbf{x}')\, \mathrm{d}\mathbf{x} = V(\mathbf{x}), \tag{A.8}$$

where $\mathcal{H}(\cdot)$ is the entropy function, $m(\cdot)$ is the invariant measure and $V(\mathbf{x})$ is assumed to be constant. Constraint (A.7) guarantees that the probability density is well-defined while constraint (A.8) guarantees finiteness of the mean potential function (Crooks, 2006).

The Lagrange multipliers objective function to be maximised is equal to

$$\begin{aligned}\mathcal{L}(\mathbf{x}) = -\int_{\mathcal{R}^M} \rho_\infty(\mathbf{x}') \log\left(\frac{\rho_\infty(\mathbf{x}')}{m(\mathbf{x}')}\right) \mathrm{d}\mathbf{x}' - \lambda^{(1)} \left(\int_{\mathcal{R}^M} \rho_\infty(\mathbf{x}')\, \mathrm{d}\mathbf{x}' - 1\right) \\ - \lambda^{(2)} \left(\int_{\mathcal{R}^M} \rho_\infty(\mathbf{x}') V(\mathbf{x}')\, \mathrm{d}\mathbf{x}' - V(\mathbf{x})\right)\end{aligned} \tag{A.9}$$

Setting first derivative to zero yields

$$\begin{aligned}\rho_\infty(\mathbf{x}') \log\left(\frac{\rho_\infty(\mathbf{x}')}{m(\mathbf{x}')}\right) &= -\lambda^{(1)} \rho_\infty(\mathbf{x}') - \lambda^{(2)} \rho_\infty(\mathbf{x}') V(\mathbf{x}') \\ \therefore \log\left(\frac{\rho_\infty(\mathbf{x}')}{m(\mathbf{x}')}\right) &= -\lambda^{(1)} - \lambda^{(2)} V(\mathbf{x}') \\ \therefore \rho_\infty(\mathbf{x}') &= m(\mathbf{x}') \exp\left(-\lambda^{(1)} - \lambda^{(2)} V(\mathbf{x}')\right).\end{aligned} \tag{A.10}$$

Let $\lambda^{(1)} \coloneqq \log(\mathcal{Z})$ and $\lambda^{(2)} \coloneqq \gamma = \sigma^{-2} > 0$. Assuming a uniform prior over probabilities implies that $m(\mathbf{x}') \propto$ constant. Then, (A.10) becomes

$$\rho_\infty(\mathbf{x}') = \frac{1}{\mathcal{Z}} \exp\left(-\gamma V(\mathbf{x}')\right),$$

where $\mathcal{Z} \coloneqq \int_{\mathcal{R}^M} \exp(-\gamma V(\mathbf{x}')\, \mathrm{d}\mathbf{x}'$ to ensure that $\rho_\infty(\mathbf{x}')$ integrates to 1. The above expression is the Boltzmann-Gibbs measure.

The second derivative is negative since $\rho_\infty(\mathbf{x}'), V(\mathbf{x}') > 0 \quad \forall\ \mathbf{x}' \in \mathcal{R}^M_{>0}$ and therefore the Boltzmann-Gibbs measure is the candidate density with the highest entropy. $\square$

Appendix B

Potential function assumptions

Let the potential function in (2.11) $V(\mathbf{x}) \in C^2(\mathbb{R}^M)$, that is $V(\mathbf{x})$ is a smooth function that is twice differentiable and continuous in $\mathbb{R}^M$. Assumptions (B.1) (Pavliotis, 2014, p.110) and (B.2) (Roberts and Tweedie, 1996) are made about the potential function $V(\mathbf{x})$ appearing in the Langevin diffusion (2.8) and the Boltzmann-Gibbs measure (2.11).

(i) $V(\mathbf{x})$ is *confining*, that is $\lim_{|\mathbf{x}|\to\infty} V(\mathbf{x})$ and

$$\exp(-\gamma V(\mathbf{x})) \in L^1(\mathbb{R}^M), \quad \forall\, \gamma > 0. \tag{B.1}$$

(ii) Let $|V(\mathbf{x})|$ be bounded for $|\mathbf{x}| \geq S$, for some $S > 0$. Then, $\exists\ 0 < d < 1$ such that $V(\mathbf{x})$ satisfies

$$\liminf_{|\mathbf{x}|\to\infty} \left\{(1-d)|\nabla V(\mathbf{x})|^2 - \gamma^{-1}\Delta V(\mathbf{x})\right\} > 0 \tag{B.2}$$

Note that $L^1(\mathbb{R}^M)$ is the Banach space of measurable functions on Ω with L-1 norm (Pavliotis, 2014, p.308) and $\Delta V(\mathbf{x}) = \nabla^2 V(\mathbf{x})$ is the Laplace operator. Assumption (B.1) guarantees that the normalising constant in the Boltzmann - Gibbs measure (2.11) is finite according to Proposition 4.2 of (Pavliotis, 2014, p.110). The two assumptions (B.1) and (B.2) are sufficient to show that the distribution of the state $\mathbf{X}(t)$ of the Langevin diffusion in (2.8) converges exponentially fast to its stationary Boltzmann-Gibbs measure (2.11), i.e. the diffusion is *exponentially ergodic* according to Theorem 2.3 of (Roberts and Tweedie, 1996).

The potential function defined in (2.9) is visibly twice differentiable and continuous because of its exponential term. For notational convenience define

$$\Lambda_{ij} = \frac{\exp(\alpha x_j - \beta c_{ij})}{\sum_{k=1}^{M} \exp(\alpha x_k - \beta c_{ik})}. \tag{B.3}$$

It follows that

$$|\nabla V(\mathbf{x})|^2 = \sum_{j=1}^{M} \left| \sum_{i=1}^{N} O_i \Lambda_{ij} - \kappa \exp(x_j) + \delta \right|^2, \tag{B.4}$$

and

$$\Delta V(\mathbf{x}) = \sum_{j=1}^{M} \left[\kappa \exp(x_j) - \alpha \sum_{i=1}^{N} O_i \Lambda_{ij} \left(1 - \Lambda_{ij}\right) \right]. \tag{B.5}$$

For $0 < d < 1$ we have

$$\lim_{x_j \to +\infty} \left\{ (1-d)|\nabla V(\mathbf{x})|^2 - \gamma^{-1} \Delta V(\mathbf{x}) \right\} = +\infty, \tag{B.6}$$

and

$$\lim_{x_j \to -\infty} \left\{ (1-d)|\nabla V(\mathbf{x})|^2 - \gamma^{-1} \Delta V(\mathbf{x}) \right\} = \underbrace{\sum_{j=1}^{M} (1-d) \left| \sum_{i=1}^{N} O_i \Lambda_{ij} + \delta \right|^2}_{\geq 0} + \underbrace{\gamma^{-1} \sum_{j=1} O_i \Lambda_{ij} (1 - \Lambda_{ij})}_{\geq 0}$$

$$\geq \delta^2 > 0,$$

since $0 \leq \Lambda_{ij}(1 - \Lambda_{ij}) \leq 1$ by definition of (B.3).

Appendix C

Potential function derivations

C.1 Utility potential derivation

Based on the gradient flow formulation, the utility potential potential can be obtained by solving

$$D_j = -\frac{\partial V_{\text{Utility}(\mathbf{x})}}{\partial x_j}, \tag{C.1}$$

subject to the constraint in (2.2). This is achieved whenever flows satisfy

$$T_{ij}(\mathbf{x}) = O_i v_{ij}(\mathbf{x}), \quad \sum_{j=1}^{M} v_{ij}(\mathbf{x}) = 1, \quad v_{ij}(\mathbf{x}) \geq 0. \tag{C.2}$$

Equivalently, given a positive function $\phi(\cdot)$, (C.2) can be expressed as

$$T_{ij}(\mathbf{x}) = O_i \frac{\phi\left(U_{ij}(\mathbf{x})\right)}{\sum_{k=1}^{M} \phi\left(U_{ik}(\mathbf{x})\right)} \tag{C.3}$$

By inspection the utility potential must be of the form

$$V_{\text{Utility}}(\mathbf{x}) = -\sum_{i=1}^{N} O_i \left\{ f_i(\mathbf{x}) \log\left(\sum_{k=1}^{M} \phi\left(U_{ij}\right) \right) \right\}, \tag{C.4}$$

for some functions $f_i(\mathbf{x}) \geq 0$. By taking the gradient in (C.4) and substituting in (C.1) we get

$$\frac{\phi\left(U_{ij}(\mathbf{x})\right)}{\sum_{k=1}^{M} \phi\left(U_{ik}(\mathbf{x})\right)} = \frac{\partial f_i(\mathbf{x})}{\partial x_j} \log\left(\sum_{k=1}^{M} \phi\left(U_{ij}\right) \right) + f_i(\mathbf{x}) \frac{\partial \phi\left(U_{ij}\right)}{\partial x_j} \left(\sum_{k=1}^{M} \phi\left(U_{ik}\right) \right)^{-1},$$

$\forall\ i \in \{1, \ldots, N\}$. The above equation holds for $\phi(\cdot) \coloneqq \exp(\cdot)$ and $f_i(\mathbf{x}) \coloneqq \alpha_i^{-1}$ with each $\alpha_i \neq 0$. By (2.9), the utility function is

$$U_{ij} = \alpha x_j + \beta c_{ij} \ \ \forall\ i, j.$$

This gives a utility potential equal to

$$V_{\text{Utility}}(\mathbf{x}) = -\sum_{i=1}^{N} O_i \left\{ \alpha^{-1} \log \left(\sum_{k=1}^{M} \exp\left(U_{ij}\right) \right) \right\},$$

where in the case of (2.14) $\alpha_i \coloneqq \alpha \ \ \forall\, i$.

C.2 Random utility maximisation of stochastic utility potential

Fix $i \in \{1, \ldots, N\}$ and let $\xi_{ij} \sim_{\text{i.i.d.}} \text{Gumbel}(0, 1)$. It follows that

$$\begin{aligned}
\mathbb{P}\left(\max_{1 \leq j \leq M} \tilde{U}_{ij}(x_j) \leq y_i \right) &= \mathbb{P}\left(\bigcup_{j=1}^{M} \left\{ \tilde{U}_{ij}(x_j) \leq y_i \right\} \right) \\
&\overset{\text{iid}}{=} \prod_{j=1}^{M} \mathbb{P}\left(\tilde{U}_{ij}(x_j) \leq y_i \right) \\
&= \prod_{j=1}^{M} \mathbb{P}\left(\xi_{ij} \leq y_i + U_{ij}(x_j) \right) \\
&= \prod_{j=1}^{M} \exp\left(-\exp\left(y_i + U_{ij}(x_j) \right) \right) \\
&= \exp\left(-\sum_{j=1}^{M} \exp\left(y_i + U_{ij}(x_j), \right) \right) \\
&= \exp\left(-\exp\left(y_i + \log\left(\sum_{j=1}^{M} \exp\left(U_{ij}(x_j) \right) \right) \right) \right) \qquad \text{(C.5)}
\end{aligned}$$

where equality 3 follows from (2.15). Therefore, $\max_{1 \leq j \leq M} \tilde{U}_{ij}(x_j)$ is a Gumbel random variable with scale one and mean $\mu = \log\left(\sum_{j=1}^{M} \exp\left(U_{ij}(x_j) \right) \right)$. The expectation of the Gumbel random variable is equal to

$$\mathbb{E}\left[\max_{1 \leq j \leq M} \tilde{U}_{ij}(x_j) \right] = \mu + c = \log\left(\sum_{j=1}^{M} \exp\left(U_{ij}(x_j) \right) \right) + c,$$

where c is the Euler-Mascheroni constant (Chang, 2015). □

C.3 Utility potential bound derivation

Fix $i \in \{1, \ldots, N\}$. Given that $\exp(\cdot)$ is a positive function, it follows that (Nielsen and Sun, 2016)

$$\begin{aligned}
\log\left(\sum_{j=1}^{M}\exp(U_{ij}(x_j))\right) &\geq \log\left(\max_{1\leq j\leq M}\left\{\exp\left(U_{ij}(x_j)\right)\right\}\right) \\
&= \log\left(\exp\left(\max_{1\leq j\leq M}\left\{U_{ij}(x_j)\right\}\right)\right) \\
&= \max_{1\leq j\leq M}\left\{U_{ij}(x_j)\right\}, \qquad \text{(C.6)}
\end{aligned}$$

and that

$$\begin{aligned}
\log\left(\sum_{j=1}^{M}\exp(U_{ij}(x_j))\right) &\leq \log\left(M\max_{1\leq j\leq M}\left\{\exp\left(U_{ij}(x_j)\right)\right\}\right) \\
&= \log\left(\exp\left(\max_{1\leq j\leq M}\left\{U_{ij}(x_j)\right\}\right)\right) + \log\left(M\right) \\
&= \max_{1\leq j\leq M}\left\{U_{ij}(x_j)\right\} + \log\left(M\right). \qquad \text{(C.7)}
\end{aligned}$$

Inequalities (C.6) and (C.7) constitute the lower and upper bounds of (2.18), respectively. □

C.4 Potential function convexity

By Lemma 4.3.5 of (Aggarwal, 2020, p.158), we know that if the potential function is strictly convex, then it has a unique global minimum, i.e. its local minimum is also its global minimum. By Lemma 4.3.4 (Aggarwal, 2020, p.154) a twice differentiable map $V : \mathbb{R}^M \to \mathbb{R}$ is convex if and only if it has a positive semi-definite Hessian $\forall\, \mathbf{x} \in \mathbb{R}^M$. Define

$$\Lambda_{i,j} = \frac{\exp(\alpha x_j - \beta c_{ij})}{\sum_{k=1}^{M}\exp(\alpha x_k - \beta c_{ik})}, \qquad \text{(C.8)}$$

$\forall\ (i,j) \in \{1,\ldots,N\} \times \{1,\ldots,M\}$, where $0 \leq \Lambda_{i,j}^2 \leq \Lambda_{i,j} \leq 1$. The Jacobian matrix of the potential function in (2.21) at the stationary points of $V_{\boldsymbol{\theta}}$ is equal to

$$J_j(\boldsymbol{\theta}) = \frac{\partial V_{\boldsymbol{\theta}}(\mathbf{x})}{\partial x_j} = -\sum_{i=1}^{N} O_i \Lambda_{i,j} + \kappa \exp(x_j) - \delta = 0, \tag{C.9}$$

$\forall\ j \in \{1,\ldots,M\}$. The Hessian matrix of the potential function in (2.21) is

$$H_{j,j}(\boldsymbol{\theta}) = \frac{\partial V_{\boldsymbol{\theta}}(\mathbf{x})}{\partial x_j^2} = \alpha \sum_{i=1}^{N} O_i \left(\Lambda_{i,j}^2 - \Lambda_{i,j}\right) + \kappa \exp(x_j) \geq 0, \tag{C.10}$$

$\forall\ j \in \{1,\ldots,M\}$ and

$$H_{j,k}(\boldsymbol{\theta}) = \frac{\partial V_{\boldsymbol{\theta}}(\mathbf{x})}{\partial x_j x_k} = \alpha \sum_{i=1}^{N} O_i \Lambda_{i,j} \Lambda_{i,k} \geq 0, \tag{C.11}$$

$\forall\ k \neq j \in \{1,\ldots,M\}$.

Consider the case for $N = M = 2$. Then,

$$\begin{aligned}
\alpha^{-2} \det(\mathbf{H}) = & \underbrace{\sum_{i=1}^{2} O_i(\Lambda_{i,1}^2 - \Lambda_{i,1}) \sum_{i=1}^{2} O_i(\Lambda_{i,2}^2 - \Lambda_{i,2})}_{\text{term1}} \\
& + \underbrace{\frac{\kappa}{\alpha}\left(\sum_{i=1}^{2} O_i(\Lambda_{i,1}^2 - \Lambda_{i,1}) \exp(x_1) + \sum_{i=1}^{2} O_i(\Lambda_{i,2}^2 - \Lambda_{i,2}) \exp(x_2)\right)}_{\text{term2}} \\
& + \underbrace{\frac{\kappa^2}{\alpha^2} \exp(x_1) \exp(x_2)}_{\text{term3}} \\
& - \underbrace{\left(\sum_{i=1}^{2} O_i \Lambda_{i,1} \Lambda_{i,2}\right)^2}_{\text{term4}},
\end{aligned} \tag{C.12}$$

By definition of (C.8), $\Lambda_{i,1} = 1 - \Lambda_{i,2}\ \forall\ i = 1, 2$. Therefore, term1 and term4 in (C.12) cancel each other out. Given that $0 \geq \Lambda_{i,1}^2 - \Lambda_{i,1} \geq -\Lambda_{i,1}$, it follows that

$$\begin{aligned}
\det(\mathbf{H}) &\geq \alpha\kappa \left(-\sum_{i=1}^{2} O_i \Lambda_{i,1} \exp(x_1) - \sum_{i=1}^{2} O_i \Lambda_{i,2} \exp(x_2)\right) + \kappa^2 \exp(x_1) \exp(x_2), \\
&= -\alpha\kappa^2 \left(\exp(2x_1) + \exp(2x_2) + \delta/\kappa(\exp(x_1) + \exp(x_2))\right) + \kappa^2 \exp(x_1) \exp(x_2) \\
&= \kappa^2 \left(\exp(x_1 + x_2) - \alpha \left(\exp(2x_1) + \exp(2x_2) + \delta/\kappa(\exp(x_1) + \exp(x_2))\right)\right),
\end{aligned} \tag{C.13}$$

where the second step follows by the Jacobian in (C.9) being zero. Consequently, the matrix positive semi-definite (i.e. the potential function is convex) if and only if

$$\alpha \leq \frac{\exp(x_1 + x_2)}{\exp(2x_1) + \exp(2x_2) + \delta/\kappa(\exp(x_1) + \exp(x_2)}, \tag{C.14}$$

where $\delta = \min_{j=1,\dots,M} x_j$ and $\kappa = 1 + \delta M$. The proof for arbitrary choices of N and M follows by induction by leveraging (C.9) and is left to the reader an an exercise.

C.5 Potential function as a candidate proposal distribution

By inspection of (2.21), the cost and additional potential resemble the log of a Gamma distribution kernel. To make the utility potential vanish, we take the following limit of the log Boltmann-Gibbs measure numerator

$$\begin{aligned}
\lim_{\alpha\to 0,\beta\to 0} -\gamma V_{\text{Utility}}(\mathbf{x}) &= \lim_{\alpha\to 0,\beta\to 0} \gamma\alpha^{-1} \sum_{i=1}^{N} O_i \log\left(\sum_{j=1}^{M} \exp(\alpha x_j - \beta c_{ij})\right) \\
&= \lim_{\alpha\to 0} \gamma\alpha^{-1} \sum_{i=1}^{N} O_i \log\left(\sum_{j=1}^{M} \exp(\alpha x_j)\right) \\
&\approx \lim_{\alpha\to 0} \gamma\alpha^{-1} \max_{1\leq j\leq M}(\alpha x_j) \sum_{i=1}^{N} O_i \\
&= \gamma \max_{1\leq j\leq M}(x_j) \\
&\approx \frac{\gamma}{M} \sum_{j=1}^{M} x_j.
\end{aligned} \tag{C.15}$$

The approximation in step 3 is the approximation of the log sum of exp function (Murphy, 2012, p.86) while the approximation in step 5 is a crude approximation of the maximum. It

follows that

$$\begin{aligned}\lim_{\alpha\to 0,\beta\to 0} -\gamma V_{\boldsymbol{\theta}}(\mathbf{x}) &\approx \frac{\gamma}{M}\sum_{j=1}^{M} x_j - \gamma\kappa\sum_{j=1}^{M}\exp(x_j) + \gamma\delta\sum_{j=1}^{M} x_j \\ &= -\gamma\kappa\sum_{j=1}^{M}\exp(x_j) + \gamma(\delta + 1/M)\sum_{j=1}^{M} x_j \\ &\propto \prod_{j=1}^{M}\log\Big(\Gamma\left(\gamma(\delta+1/M), 1/\gamma\kappa\right)\Big). \end{aligned} \tag{C.16}$$

Therefore, we have derived an approximation for the log-Gamma distribution using the potential function in (2.21).