Computationally efficient inference for latent position network models

Riccardo Rastelli^{1,*}, Florian Maire², and Nial Friel^{1,3}

School of Mathematics and Statistics, University College Dublin, Dublin, Ireland;
 Department of Mathematics and Statistics, University of Montreal, Montreal, Canada;
 Insight Centre for Data Analytics, Dublin, Ireland.

*riccardo.rastelli@ucd.ie

Abstract

Latent position models are widely used for the analysis of networks in a variety of research fields. In fact, these models possess a number of desirable theoretical properties, and are particularly easy to interpret. However, statistical methodologies to fit these models generally incur a computational cost which grows with the square of the number of nodes in the graph. This makes the analysis of large social networks impractical. In this paper, we propose a new method characterised by a linear computational complexity, which can be used to fit latent position models on networks of several tens of thousands nodes. Our approach relies on an approximation of the likelihood function, where the amount of noise introduced by the approximation can be arbitrarily reduced at the expense of computational efficiency. We establish several theoretical results that show how the likelihood error propagates to the invariant distribution of the Markov chain Monte Carlo sampler. In particular, we demonstrate that one can achieve a substantial reduction in computing time and still obtain a good estimate of the latent structure. Finally, we propose applications of our method to simulated networks and to a large coauthorships network, highlighting the usefulness of our approach.

Keywords: latent position models; noisy Markov chain Monte Carlo; social networks; Bayesian inference.

1 Introduction

In the last few decades, network data has become extremely common and readily available in a variety of fields, including the social sciences, biology, finance and technology. After the pioneering work of Hoff et al. (2002), latent position models (hereafter LPMs) have become one of the cornerstones in the statistical analysis of networks. LPMs are flexible models capable of capturing many salient features of realised networks while providing

results which can be easily interpreted. However, a crucial aspect in the statistical analyses of networks is scalability: the computational burden required when fitting LPMs generally grows with the square of the number of nodes. This seriously hinders their applicability, since estimation becomes impractical for networks larger than a few hundreds nodes. Here, we precisely address this issue by introducing a new methodology to fit LPMs which is characterised by a linear computational complexity in the number of nodes.

LPMs postulate that the nodes of an observed network are characterised by a unique random position in a latent space: in the most common setup, each node is mapped to a point of \mathbb{R}^2 . Additionally, the probability of observing an edge between two nodes is determined by the corresponding pairwise latent distance. A common assumption requires that closer nodes are more likely to connect than nodes farther apart, or, equivalently, that the probability of connection $\rho(d_{ij})$ is a non-increasing function of the distance d_{ij} between nodes i and j. Evidently, the aforementioned quadratic computing costs originate from the necessity of keeping track of all of the pairwise distances between the nodes.

In our approach, we construct a partition of the latent space, therefore inducing a partition on the nodes of the graph itself. This allows us to cluster together nodes that are expected to have approximately the same behaviour, with regard to their connections. In principle, this is similar to imposing a stochastic block model structure (Wang and Wong 1987), whereby the nodes belonging to the same block are assumed to be *stochastically equivalent* (Nowicki and Snijders 2001). The crucial advantage of our approach is that working with the aggregated information derived from the partitioning does not involve the calculation of all the pairwise distances at any stage, therefore decreasing the overall computational complexity.

Similarly to the original paper of Hoff et al. (2002), our approach also relies on Markov chain Monte Carlo (hereafter MCMC) to obtain a Bayesian posterior sample of the latent positions and other model parameters. However, in contrast to their approach, we replace the likelihood of the LPM with an approximate (hence noisy) counterpart that aggregates the latent position of nodes belonging to the same block. By construction, the cost of the calculation of this surrogate likelihood grows linearly in the number of nodes, hence giving a significant computational advantage to our method when compared to the approach of Hoff et al. (2002) or other subsequent related works.

Since the LPM likelihood is replaced by a proxy, our method broadly fits within the context of noisy Markov chain Monte Carlo (Alquier et al. 2016), a topic that has recently generated a noticeable interest within the field of computational statistics and beyond. The theoretical aspect of our paper relies and builds upon the core ideas of noisy MCMC. In particular, our methodology is supported by a collection of theoretical results showing that our approach leads to quantifiable gains in efficiency. More precisely, we show that the error in the MCMC output induced by the likelihood approximation can be arbitrarily bounded by refining the partition in the latent space. Besides, a finer partition also implies higher computational costs. As a consequence, our algorithm allows a trade-off between speed and accuracy that can be set according to the available computational budget, and the level of precision required for inference. In addition, our theoretical developments include a proposition that can be regarded as an extension of the results of Alquier et al. (2016) to the widely used Metropolis-within-Gibbs (MwG hereafter) algorithm, and which may thus have applications beyond the context of LPMs.

The theoretical results are established for a generic LPM framework: the assumptions we use are rather general and encompass most of the commonly used LPMs. In addition to these results, we propose applications of our method to both simulated and real datasets, whereby we focus on a more specific model which is equivalent to that of Hoff et al. (2002). Our simulation study demonstrates that the noisy algorithm succeeds in recovering the latent structure correctly, achieving the same qualitative results obtained with the currently available approaches. Crucially, the computing time required by our proposed approach is only a fraction of that of the non-noisy one.

To illustrate the usefulness of our method, we propose an application to a large social network representing coauthorships in the astrophysics category of the repository of electronic preprints, arXiv. This demonstration highlights the fact that our approach is capable of recovering the structure of the latent space at the macro level with just a small fraction of the actual computational cost, providing a useful bi-dimensional summary of the data.

The structure of the paper is as follows: in Section 2 we give an overview of the literature related to LPMs and noisy Markov chain Monte Carlo. In Section 3, we formally characterise the main features of the original LPM of Hoff et al. (2002), giving an overview of the MwG sampling strategy used to perform inference, highlighting some of its limitations. In Section 4, we lay the foundations for our theoretical results, by defining the

general assumptions that our LPMs must satisfy. In Section 5, we formally introduce the partitioning of the latent space and all of the associated notation. Section 6 introduces the novel noisy algorithm, whereas in Section 7 we expose the main theoretical results. Finally, Sections 8 and 9 illustrate the applications of our methodology to simulated and a real dataset, respectively.

2 Review of related literature

The study of the mathematical properties of LPMs dates back at least to Gilbert (1961). However, the first application of these models in the statistical analysis of social networks is due to Hoff et al. (2002), who introduced a feasible methodology to fit LPMs to interaction data. Since the work of Hoff et al. (2002), LPMs have been intensively studied and widely applied to a variety of contexts, becoming one of the prominent statistical models for network analyses. There are a number of reasons for this success. Most importantly, LPMs are particularly easy to interpret, and offer a clear and intuitive graphical representation of the results. In addition, LPMs are capable of capturing a number of features of interest such as transitivity, clustering, homophily and assortativity, which are often exhibited by observed social networks. An overview of the theoretical properties of realised LPMs is given in Rastelli et al. (2016).

In order to increase the flexibility of these models, a number of extensions of the basic framework have been considered. Handcock et al. (2007) introduce a more sophisticated prior on the latent point process to represent clustering in the network, that is, the presence of communities. Krivitsky et al. (2009) further extends the model to include nodal random effects, i.e. additional latent features on the nodes capable of tuning their in-degrees and out-degrees. Both of these extensions are implemented in the R package latentnet.

LPMs have also been extended to account for multiple network views (Gollini and Murphy 2014; Durante et al. 2017; Salter-Townshend and McCormick 2017), binary interactions evolving over time (Sarkar and Moore 2006; Sewell and Chen 2015b; Friel et al. 2016; Durante and Dunson 2016), ranking network data (Gormley and Murphy 2007; Sewell and Chen 2015a) and weighted networks (Sewell and Chen 2016). Review papers dealing with LPMs include Salter-Townshend et al. (2012), Matias and Robin (2014) and Raftery (2017).

Similarly to our contribution, three other papers address the issue of scalability for

the inference on LPMs. In Salter-Townshend and Murphy (2013), the authors propose a variational approximation (coupled with first order Taylor expansions to deal with various intractabilities) to perform posterior maximisation for the model described by Handcock et al. (2007). One drawback of this approach is that it is not possible to assess the magnitude of the error induced by the variational approximation. Also, the modelling assumptions are not flexible, since the variational framework can only be used with a restricted selection of parametric distributions.

In Ryan et al. (2017), the authors consider the same latent position clustering model, and propose a Gaussian finite mixture prior distribution on the latent point process that allows one to *collapse* the posterior distribution. This means that several model parameters can be analytically integrated out from the posterior distribution of the model, hence simplifying the sampling scheme and achieving better estimators with a smaller computational cost.

Finally, Raftery et al. (2012) proposes a case-control likelihood approximation for the LPM with nodal random effects. In this paper, the authors argue that the majority of large social networks are sparse, hence, missing edges contribute the most to the LPM likelihood. By analogy with the case-control idea from epidemiology, they estimate the likelihood value using only a subset of the contributions given by the missing edges. We consider this approach similar to ours, since both methods rely on a noisy likelihood. We point out that our algorithm benefits from a series of theoretical results that guarantee its correctness and characterise the error induced by the approximation. In addition, our method may be applied to networks of potentially huge size regardless of the level of sparseness.

Regarding the theoretical analysis of our algorithm, the main reference that we relate to is Alquier et al. (2016). These authors argue that the computational problems arising when inferring large datasets can often be alleviated by introducing approximations in the MCMC schemes. These approaches are generally referred to as noisy MCMC, since one ends up sampling using a noisy transition kernel, rather than the correct one. In Alquier et al. (2016), the authors exploit a theoretical result from Mitrophanov (2005) to characterise the error induced by these approximations on the invariant distribution of the transition kernel. They also propose several applications based on the Metropolis-Hastings algorithm to a number of relevant statistical modelling frameworks. We also point out that, more recently, the noisy Monte Carlo framework has been adopted by

Boland et al. (2017) and Maire et al. (2018), as a means to speed up inference for Gibbs random fields and other general models. Even though the literature on noisy MCMC has been recently enriched by a number of relevant entries (Negrea and Rosenthal 2017; Johndrow and Mattingly 2017; Rudolf and Schweizer 2017), the theoretical framework developed in Alquier et al. (2016) proved sufficient to establish our results, as shown in Section 7.

3 Latent Position Models

3.1 Definition

A random graph is an object $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where $\mathcal{V} = \{1, ..., N\}$ is a fixed set of labels for the nodes and \mathcal{E} is a list of the randomly realised edges. In the social sciences, for example, random graphs are used to represent the social interactions within a set of actors. The values appearing on the undirected ties are modeled through the random variables:

$$\mathcal{Y} = \{Y_{ij} : i, j \in \mathcal{V}, \ i < j\}. \tag{1}$$

In this paper we only deal with undirected binary graphs, hence, the observed realisations are denoted as follows:

$$y_{ij} = \begin{cases} 1, & \text{if an edge between } i \text{ and } j \text{ appears;} \\ 0, & \text{otherwise;} \end{cases}$$
 (2)

for every $i \in \mathcal{V}$ and $j \in \mathcal{V}$ such that j > i. Note that, in the framework considered, self-edges are not modelled.

In LPMs the nodes are characterised by a latent position, generically denoted $\mathbf{z} \in \mathbb{R}^m$, which determines their social profile. The choice m=2 is the most common since it usually couples a good fit and a convenient framework to represent the results. Hence, we illustrate our methodology assuming that the number of latent dimensions is two, noting that extensions to other cases may be possible.

In the basic LPM, the probability of an edge appearing is determined by the positions of the nodes at its extremes and by some other global parameters (e.g. an intercept). This may be formally written as follows:

$$p(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\psi}) := \mathbb{P}(y_{ij} = 1 | \mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\psi}) = 1 - \mathbb{P}(y_{ij} = 0 | \mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\psi}). \tag{3}$$

Here ψ is a vector of global parameters with dimensions indexed by the labels $\mathcal{K} = \{1, \dots, K\}$. The parameter ψ is sometimes referred to as the static parameter of the

model, as opposed to the latent field $\mathcal{Z} := \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$. A number of possible formulations for the edge probabilities have been proposed. Within the statistical community, the most common choice is the logit link proposed by Hoff et al. (2002):

$$\log \left(\frac{p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \psi\right)}{1 - p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \psi\right)} \right) := \psi - d\left(\mathbf{z}_{i}, \mathbf{z}_{j}\right); \tag{4}$$

where $d(\mathbf{z}_i, \mathbf{z}_j)$ denotes the Euclidean distance between the two nodes, and $\psi \in \mathbb{R}$ is simply an intercept parameter (K = 1). Alternative formulations are used in Gollini and Murphy (2014) and Rastelli et al. (2016). In physics, a variety of edge probability functions have been proposed. A list of these can be found, for example, in Parsonage and Roughan (2017) and references therein. One feature that all of these formulations have in common is that the edge probability is a function of the distance between the two nodes, and that its value decreases as the latent distance increases, making long edges less likely to appear.

Since the data observations are conditionally independent given the latent positions, the likelihood of all undirected LPMs may be written as:

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{Z}, \boldsymbol{\psi}) = \mathbb{P}\left(\mathcal{Y}|\mathcal{Z}, \boldsymbol{\psi}\right) = \prod_{\{i \in \mathcal{V}\}} \prod_{\{j \in \mathcal{V} \setminus i\}} \left\{ \left[p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}\right) \right]^{y_{ij}} \left[1 - p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}\right) \right]^{1 - y_{ij}} \right\}^{1/2}$$
 (5)

where the square root is introduced to remedy the fact that each edge contributes twice to the likelihood of the undirected network (the motivation behind this particular formulation will be more clear in the following sections). We note that, for a given set of positions \mathcal{Z} and global parameters ψ , the computational cost for the likelihood evaluation is $\mathcal{O}(N^2)$, i.e. it grows with the square of the number of nodes.

3.2 Bayesian inference

Inference for LPMs is usually carried out in a Bayesian framework, using MCMC to obtain posterior samples of the model parameters (Hoff et al. 2002; Handcock et al. 2007; Krivitsky et al. 2009; Raftery et al. 2012). The posterior distribution of interest is:

$$\pi \left(\mathcal{Z}, \psi | \mathcal{Y} \right) \propto \mathcal{L}_{\mathcal{Y}} \left(\mathcal{Z}, \psi \right) \pi \left(\mathcal{Z} \right) \pi \left(\psi \right). \tag{6}$$

Assuming that the cost of the evaluation of the priors $\pi(\mathcal{Z})$ and $\pi(\psi)$ is $\mathcal{O}(N)$ or negligible, the computational cost required to evaluate the posterior value grows with N^2 , which corresponds to the bottleneck imposed by the likelihood term. A Markov

chain Monte Carlo sampler can be designed to sample each of the model parameters in turn, using the following full-conditional distributions:

$$\pi\left(\mathbf{z}_{i}|\mathcal{Z}_{-i},\boldsymbol{\psi},\mathcal{Y}\right) \propto \pi\left(\mathbf{z}_{i}\right) \prod_{\{j \in \mathcal{V}: j \neq i\}} \left[p\left(\mathbf{z}_{i},\mathbf{z}_{j};\boldsymbol{\psi}\right)\right]^{y_{ij}} \left[1 - p\left(\mathbf{z}_{i},\mathbf{z}_{j};\boldsymbol{\psi}\right)\right]^{1 - y_{ij}} \tag{7}$$

$$\pi\left(\psi_{k}\middle|\psi_{-k},\mathcal{Z},\mathcal{Y}\right)\propto\pi\left(\psi_{k}\right)\mathcal{L}_{\mathcal{Y}}\left(\mathcal{Z},\psi\right)\tag{8}$$

In the previous equations $i \in \mathcal{V}$, $k \in \mathcal{K}$, whereas $\mathcal{Z}_{-i} = \{\mathbf{z}_j\}_{j \in \mathcal{V} \setminus \{i\}}$ and $\psi_{-k} = \{\psi_{k'}\}_{k' \in \mathcal{K} \setminus \{k\}}$. Here we have assumed that the model parameters are all independent apriori: this is indeed very common and it will be formalised in the following sections. Since each evaluation of (7) requires $\mathcal{O}(N)$ calculations, the overall complexity of the sampler still grows with the square of N.

The full-conditionals (7) and (8) are generally not in standard form. Hence, new values for the model parameters are sampled through what is usually referred to as a Metropolis-within-Gibbs (MwG) type algorithm (see e.g. Gilks et al. 1995). More precisely, potential new parameters are drawn from proposal distributions $q_{\mathcal{Z}}(\mathbf{z}_i \to \mathbf{z}_i')$ and $q_{\psi}(\psi_k \to \psi_k')$ and are then accepted with probability:

$$\alpha_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right) := 1 \wedge \left\{ \frac{q_{\mathcal{Z}}\left(\mathbf{z}_{i}^{\prime} \to \mathbf{z}_{i}\right) \pi\left(\mathbf{z}_{i}^{\prime} | \mathcal{Z}_{-i}, \boldsymbol{\psi}, \mathcal{Y}\right)}{q_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right) \pi\left(\mathbf{z}_{i} | \mathcal{Z}_{-i}, \boldsymbol{\psi}, \mathcal{Y}\right)} \right\}$$
(9)

$$\alpha_{\psi} (\psi_{k} \to \psi'_{k}) := 1 \wedge \left\{ \frac{q_{\psi} (\psi'_{k} \to \psi_{k}) \pi (\psi'_{k} | \psi_{-k}, \mathcal{Z}, \mathcal{Y})}{q_{\psi} (\psi_{k} \to \psi'_{k}) \pi (\psi_{k} | \psi_{-k}, \mathcal{Z}, \mathcal{Y})} \right\}$$

$$(10)$$

for the latent positions and global parameters, respectively. In the previous equations, for two real number a and b, $a \wedge b$ stands for the minimum between the two numbers. Also, we point out that, as is common practice, the two dimensions of the latent positions are dealt with simultaneously, i.e. they are updated in block.

The acceptance probabilities may equivalently be written as follows:

$$\alpha_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right) = 1 \wedge \left\{ \frac{q_{\mathcal{Z}}\left(\mathbf{z}_{i}^{\prime} \to \mathbf{z}_{i}\right)}{q_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right)} \cdot \frac{\pi\left(\mathbf{z}_{i}^{\prime}\right)}{\pi\left(\mathbf{z}_{i}\right)} \cdot \mathcal{LR}_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right) \right\}$$

$$(11)$$

$$\alpha_{\psi} \left(\psi_k \to \psi_k' \right) = 1 \wedge \left\{ \frac{q_{\psi} \left(\psi_k' \to \psi_k \right)}{q_{\psi} \left(\psi_k \to \psi_k' \right)} \cdot \frac{\pi \left(\psi_k' \right)}{\pi \left(\psi_k \right)} \cdot \mathcal{LR}_{\psi} \left(\psi_k \to \psi_k' \right) \right\}$$
(12)

The quantities denoted with \mathcal{LR} indicate the *likelihood ratios*, and read as follows:

$$\mathcal{LR}_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right) = \prod_{j \in Y_{i}^{1}} \frac{p\left(\mathbf{z}_{i}^{\prime}, \mathbf{z}_{j}; \boldsymbol{\psi}\right)}{p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}\right)} \prod_{j \in Y_{i}^{0}} \frac{1 - p\left(\mathbf{z}_{i}^{\prime}, \mathbf{z}_{j}; \boldsymbol{\psi}\right)}{1 - p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}\right)}$$
(13)

$$\mathcal{LR}_{\psi}\left(\psi_{k} \to \psi_{k}'\right) = \left\{ \prod_{i \in \mathcal{V}} \left[\prod_{j \in Y_{i}^{1}} \frac{p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \psi'\right)}{p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \psi\right)} \prod_{j \in Y_{i}^{0}} \frac{1 - p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \psi'\right)}{1 - p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \psi\right)} \right] \right\}^{1/2}$$
(14)

where the new symbols indicate that:

$$Y_i^h := \{ j \in \mathcal{V} \mid j \neq i, \ y_{ij} = h \}, \qquad h = \{0, 1\},$$
 (15)

and

$$\psi' \in \mathbb{R}^K$$
 is such that $\psi'_{\ell} = \psi_{\ell}$ for all $\ell \neq k$, and $\psi'_{k} \neq \psi_{k}$ (16)

with some implicit dependence on the proposed parameter in (16).

The MwG sampler described above defines a Markov chain whose stationary distribution is the posterior of interest (6). As a consequence, provided that the Markov chain is ergodic, the samples obtained at stationarity can be used to fully characterise the posterior distribution of interest. In fact, the MwG chain is shown to be geometrically ergodic for a variety of proposal distributions and under some regulatory conditions on the invariant distribution π , see Roberts and Rosenthal 1998, Theorem 5.

3.3 Non-identifiability of the latent positions

LPMs are known to be non-identifiable with respect to translations, rotations, and reflections of the latent positions. This issue has no particular effect on the sampling itself, yet it may hinder the interpretation of the posterior samples. For this reason, the latent positions are usually post-processed using the so-called Procrustes' matching. This procedure consists of rotating and translating the configurations of points observed at the end of each iteration, to match a given reference layout. In this way, the trajectory of each node during the sampling may be properly assessed, since the overall rotation and translation effect has been removed. A detailed description of the method is given, for example, in Hoff et al. (2002) and Shortreed et al. (2006). In this paper, we adopt exactly the same strategy to solve the non-identifiability problem, using as reference either the true positions (if available) or the maximum a posteriori configuration.

4 Assumptions

The methodology we develop in this paper relies on several assumptions which are described in this section.

Assumption 1. All of the model parameters are defined on bounded sets, i.e.:

$$\forall k \in \mathcal{K} : \psi_k \in [\psi_k^{\mathcal{L}}, \psi_k^{\mathcal{U}}] =: \mathcal{S}_{\psi_k}, \tag{17}$$

$$\forall i \in \mathcal{V} : \mathbf{z}_i \in [-S, S] \times [-S, S] =: \mathcal{S}_{\mathcal{Z}}, \tag{18}$$

for some finite constants S, $\psi_k^{\mathcal{L}}$ and $\psi_k^{\mathcal{U}}$.

Remark 1. Assumption 1 is rather strong and contrasts with the usual LPM frameworks. However, we argue that, from a practical point of view, these imposed conditions do not change the essence of the model. In fact, very large LPM parameters normally lead to degenerate models, and hence to realised networks that are meaningless in this modelling context (e.g. full or empty graphs). In this perspective, there is in fact a necessity to constrain ψ to a bounded space in order to make the model more tractable.

Assumption 2. The model parameters are a priori independent and distributed according to the generic prior π . For all $k \in \mathcal{K}$, all $\psi_k, \psi'_k \in \mathcal{S}^2_{\psi_k}$, for all $i \in \mathcal{V}$ and all $\mathbf{z}_i, \mathbf{z}'_i \in \mathcal{Z}^2$, the prior satisfies

$$\frac{\pi\left(\psi_{k}'\right)}{\pi\left(\psi_{k}\right)} \le \kappa_{\pi} \qquad \frac{\pi\left(\mathbf{z}_{i}'\right)}{\pi\left(\mathbf{z}_{i}\right)} \le \varkappa_{\pi} \tag{19}$$

for some finite constants κ_{π} and \varkappa_{π} .

In the following, the letter κ will be used to designate constants related to the static parameter ψ and \varkappa to those referring to the latent field \mathcal{Z} .

Remark 2. In the applications sections of this paper, a spherical truncated Gaussian distribution is used as prior on the latent positions:

$$\pi\left(\mathbf{z}_{i}\right) = \prod_{m=1}^{2} \left\{ \frac{\phi\left(\frac{z_{im}}{\gamma}\right)}{\gamma\left[\Phi\left(\frac{S}{\gamma}\right) - \Phi\left(\frac{-S}{\gamma}\right)\right]} \right\}, \qquad \forall i \in \mathcal{V};$$
 (20)

where $\gamma > 0$, ϕ and Φ are the p.d.f. and c.d.f. of a standard Gaussian distribution, respectively. This prior specification satisfies Assumption 2, as shown in Appendix A.1.

Assumption 3. The edge probability function $p: \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^K \to [p^{\mathcal{L}}, p^{\mathcal{U}}] \subset (0, 1)$ satisfies the following properties:

• (p depends on the positions only through the latent distances) there exists a function $\rho: \mathbb{R}^+ \times \mathbb{R}^K \to [p^{\mathcal{L}}, p^{\mathcal{U}}]$ such that

$$\forall \mathbf{z}_{i}, \mathbf{z}_{j} \in \mathbb{R}^{2}, \ \forall \mathbf{\psi} \in \mathbb{R}^{K}: \ p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \mathbf{\psi}\right) = \rho\left(d\left(\mathbf{z}_{i}, \mathbf{z}_{j}\right), \mathbf{\psi}\right).$$

• (ρ is non-increasing w.r.t. distances) for any $\mathbf{z}_i \in \mathcal{S}_{\mathcal{Z}}$, i = 1, 2, 3, 4:

if
$$d(z_1, z_2) \ge d(z_3, z_4)$$
, then $p(z_1, z_2; \psi) \le p(z_3, z_4; \psi)$.

• (ρ is Lipschitz w.r.t. distances) for any $z_i \in \mathcal{S}_{\mathcal{Z}}$, i = 1, 2, 3, 4:

$$|p(z_1, z_2; \psi) - p(z_3, z_4; \psi)| \le \varkappa_p |d(z_1, z_2) - d(z_3, z_4)|$$

for some finite constant \varkappa_p .

Remark 3. Assumption 3 is satisfied for most link functions such as Eq. (4).

Assumption 4. The proposal distributions $q_{\mathcal{Z}}(z \to z')$ and $q_{\psi}(\psi \to \psi')$ are such that:

$$\frac{q_{\mathcal{Z}}(\mathbf{z}' \to \mathbf{z})}{q_{\mathcal{Z}}(\mathbf{z} \to \mathbf{z}')} \le \varkappa_{q} \qquad \frac{q_{\psi}(\psi' \to \psi)}{q_{\psi}(\psi \to \psi')} \le \kappa_{q}$$
 (21)

for some finite constants \varkappa_q , κ_q , for all z, z' in a compact set $S \subset \mathbb{R}^2$, and for all ψ , ψ' in a compact subset of \mathbb{R} .

Remark 4. In the applications which follow, a truncated Gaussian proposal for the latent positions is advocated. In such case Assumption 4 is satisfied, as shown in Appendix A.2.

5 Grid approximation of the latent distances

Hereafter, we consider a generic LPM satisfying Assumptions 1, 2, 3 and 4, and we illustrate an estimation procedure based on a grid partitioning of the latent space. Following an approach similar to that of Parsonage and Roughan (2017), we create a partitioning of the latent positions \mathcal{Z} using a grid in \mathbb{R}^2 . The grid is made of adjacent squares (called boxes hereafter) of side length b > 0, each having both sides aligned to the axes. A generic box B[g,h] has corners located in (bg-b,bh-b), (bg-b,bh), (bg,bh) and (bg,bh-b), where the indexes g and h are positive or negative but non-null integers, i.e. $g,h \in \mathbb{Z} \setminus 0$. Figure 1 shows the latent space with the partitioning given by these boxes.

We denote with N[g, h] the number of points located in a generic box:

$$N[g,h] = |\{i \in \mathcal{V} : \mathbf{z}_i \in B[g,h]\}|,$$
 (22)

where |H| denotes the cardinality of the set H.

It is also useful to introduce the centre of a generic box $\mathbf{c}[g,h] := (bg - b/2, bh - b/2)$. Given a node $j \in \mathcal{V}$ such that $\mathbf{z}_j \in B[g,h]$, we also indicate the centre of B[g,h] with \mathbf{c}_j , representing the centre of the box containing j. An essential aspect of our proposed approach is determined by the fact that the distance $d(\mathbf{z}_i, \mathbf{z}_j)$ between any two nodes may be approximated by $d(\mathbf{z}_i, \mathbf{c}_j)$, i.e. the distance between node i and the centre of the box containing j.

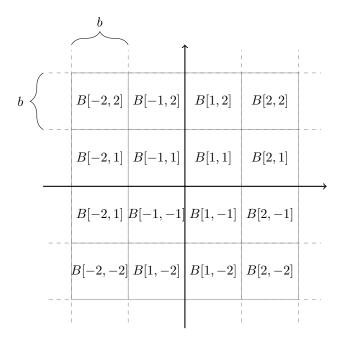


Figure 1: Grid partitioning the latent space.

Finally, we denote with $\xi_i[g,h]$ the number of edges between node i and the nodes allocated in B[g,h], i.e.:

$$\xi_i[g,h] = \sum_{\{j \in \mathcal{V}: \ \mathbf{z}_j \in B[g,h]\}} y_{ij}; \tag{23}$$

and by $\zeta_i[g,h]$ the number of missing edges:

$$\zeta_i[g,h] = N[g,h] - \xi_i[g,h] - \mathbb{1}\left(\{z_i \in B[g,h]\}\right); \tag{24}$$

where $\mathbb{1}(A)$ is 1 if A is true or 0 otherwise. Also, the degree of node $i \in \mathcal{V}$, i.e. the number of edges incident to it, is indicated by D_i .

These quantities introduced are exploited in the following sections to illustrate a new way of carrying out Bayesian inference for LPMs, requiring a dramatically reduced computational cost.

6 Noisy MCMC

As explained in the previous section, the distance from node i to the centre of a generic box $\mathbf{c}[g,h]$ can be used as a proxy for the true distances between i and all of the points contained in B[g,h], for all g and h. This in turn allows one to approximate the edge probability $p(\mathbf{z}_i,\mathbf{z}_j;\boldsymbol{\psi})$ using $p(\mathbf{z}_i,\mathbf{c}_j;\boldsymbol{\psi})$, for all $j \in \mathcal{V}$ such that $\mathbf{z}_j \in B[g,h]$. This fact

may be exploited in a number of ways. For example, the likelihood defined in (5) may be replaced by the following noisy likelihood:

$$\tilde{\mathcal{L}}_{\mathcal{Y}}(\mathcal{Z}, \boldsymbol{\psi}) := \left\{ \prod_{i=1}^{N} \prod_{g,h} \left[p\left(\mathbf{z}_{i}, \mathbf{c}[g, h]; \boldsymbol{\psi}\right) \right]^{\xi_{i}[g, h]} \left[1 - p\left(\mathbf{z}_{i}, \mathbf{c}[g, h]; \boldsymbol{\psi}\right) \right]^{\zeta_{i}[g, h]} \right\}^{1/2}; \tag{25}$$

where each edge contribution is essentially replaced by its noisy counterpart. Here, by counting each edge contribution twice and then correcting with the square root, one has the possibility to use the noisy approximation in a symmetric way, with respect to any pair of nodes i and j. We point out that a number of alternative estimators are available for the likelihood value using the grid approximation: the estimator proposed in (25) is one that generally works well in practice and that makes our theoretical developments easier to follow.

With NoisyLPM, we refer to a MwG sampler that relies on the approximate edge probabilities rather than the true ones, or, equivalently, that uses the noisy likelihood $\tilde{\mathcal{L}}_{\mathcal{Y}}$ instead of the true likelihood $\mathcal{L}_{\mathcal{Y}}$. In NoisyLPM the likelihood ratios introduced in (13) and (14) can be approximated as follows:

$$\widetilde{\mathcal{LR}}_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right) = \prod_{g,h} \left\{ \left[\frac{p\left(\mathbf{z}_{i}^{\prime}, \mathbf{c}[g,h]; \boldsymbol{\psi}\right)}{p\left(\mathbf{z}_{i}, \mathbf{c}[g,h]; \boldsymbol{\psi}\right)} \right]^{\xi_{i}[g,h]} \left[\frac{1 - p\left(\mathbf{z}_{i}^{\prime}, \mathbf{c}[g,h]; \boldsymbol{\psi}\right)}{1 - p\left(\mathbf{z}_{i}, \mathbf{c}[g,h]; \boldsymbol{\psi}\right)} \right]^{\zeta_{i}[g,h]} \right\}$$
(26)

$$\widetilde{\mathcal{LR}}_{\psi}\left(\psi_{k} \to \psi_{k}'\right) = \prod_{i \in \mathcal{V}} \prod_{g,h} \left\{ \left[\frac{p\left(\mathbf{z}_{i}, \mathbf{c}[g,h]; \boldsymbol{\psi}'\right)}{p\left(\mathbf{z}_{i}, \mathbf{c}[g,h]; \boldsymbol{\psi}\right)} \right]^{\xi_{i}[g,h]} \left[\frac{1 - p\left(\mathbf{z}_{i}, \mathbf{c}[g,h]; \boldsymbol{\psi}'\right)}{1 - p\left(\mathbf{z}_{i}, \mathbf{c}[g,h]; \boldsymbol{\psi}\right)} \right]^{\zeta_{i}[g,h]} \right\}^{1/2}$$
(27)

It is apparent that the computational cost of one evaluation of the approximate likelihood ratios is much smaller than that of the true counterpart. In fact, the complexity of a noisy MwG update becomes $\mathcal{O}(1)$ and $\mathcal{O}(N)$ for latent positions and global parameters, respectively. Overall, this makes the computational complexity of the NoisyLPM procedure of an order smaller than $\mathcal{O}(N^2)$.

7 Theoretical guarantees

This section provides theoretical results that characterise the error induced by our approximation. Indeed, replacing $\mathcal{L}_{\mathcal{Y}}(\mathcal{Z}, \boldsymbol{\psi})$ with $\tilde{\mathcal{L}}_{\mathcal{Y}}(\mathcal{Z}, \boldsymbol{\psi})$ in the MwG acceptance ratio implies that the stationary distribution of the Markov chain may not coincide anymore with the posterior distribution of interest described in Section 3.2. Here, our main goal is

to show that a noisy MwG sampler, such as the NoisyLPM, generates a sequence of random variables whose distribution can be made arbitrarily close to the posterior π (\cdot | \mathcal{Y}).

In fact, one can note that, by construction, our noisy MwG sampler admits the approximate posterior $\tilde{\pi}$ ($\cdot | \mathcal{Y}$) as stationary distribution. Hence, the approximation error is directly, and globally, measured by $\|\pi - \tilde{\pi}\|$, i.e. the total variation distance between the two posteriors. However, obtaining an explicit expression or an upper bound of $\|\pi - \tilde{\pi}\|$ is challenging. In addition, since $\tilde{\pi}$ is the limiting distribution, evaluating $\|\pi - \tilde{\pi}\|$ would only be meaningful for an analysis in the asymptotic regime. For these reasons, we propose a series of results whose final goal is to quantify the error in a non-asymptotic framework, by aggregating the elementary errors that are generated by the noisy sampler at each iteration.

The theoretical framework is the analysis of the perturbation of uniformly ergodic Markov chains, initiated in Mitrophanov 2005 and refined for the noisy Metropolis-Hastings case in Alquier et al. 2016. We first recall the uniform ergodicity assumption.

Assumption 5. A π -invariant Markov kernel P operating on a state space S is uniformly ergodic if after $t \in \mathbb{N}$ iterations, the distance between the chain distribution and the stationary distribution is bounded as follows:

$$\sup_{u \in \mathcal{S}} \|P^t(u, \,\cdot\,) - \pi\| \le C\tau^t,\tag{28}$$

for some $C < \infty$ and $\tau < 1$.

The section is divided in two parts: in the spirit of Alquier et al. 2016, we first derive an extension of their theoretical framework to include the analysis of noisy Metropolis-within-Gibbs algorithms in a generic setup, that is, beyond the LPM context. In the second part we give a series of theoretical results that are specific to LPMs, and that aim to characterise the magnitude of the approximation error in the likelihood ratios and MwG acceptance probabilities, in preparation for applying our general result. In particular, we show that the distance between the exact algorithm and the NoisyLPM can be arbitrarily reduced by refining the latent grid.

7.1 Noisy MwG aggregated errors

This paper deals with an approximation of a MwG Markov chain, where the parameters of the model are updated in turn. Perturbations of uniformly ergodic Metropolis-Hastings

Markov chains have been studied in Alquier et al. 2016. We show, here, that a similar analysis can be carried out in a generic MwG sampler framework.

We introduce the following notation. The model parameters are indexed with $r \in \mathcal{R} = \{1, \ldots, R\}$. An arbitrary sigma-algebra on the compact parameter space \mathcal{S} is denoted by \mathcal{A} . For any signed measure μ on $(\mathcal{S}, \mathcal{A})$, we denote the total variation distance of μ by $\|\mu\| := \sup_{A \in \mathcal{A}} |\mu(A)|$. For any Markov kernel P operating on $\mathcal{S} \times \mathcal{A}$, we denote the operator norm of P as:

$$||P|| := \sup_{u \in \mathcal{S}} ||P(u, \cdot)|| = \sup_{u \in \mathcal{S}} \sup_{A \in \mathcal{S}} |P(u, A)|$$
 (29)

Finally, let μP be the measure on $(\mathcal{S}, \mathcal{A})$ defined as $\mu P := \int_{\mathcal{S}} \mu(\mathrm{d}x) P(x, \cdot)$. The following proposition is the building block of Theorem 1. It shows that the distance between the one step transition of an elementary MwG update and its noisy counterpart is uniformly bounded.

Proposition 1. Let P_r and \tilde{P}_r be an exact and noisy transition kernels, respectively, for the MwG update of the model parameter $r \in \mathcal{R}$. Let α_r and $\tilde{\alpha}_r$ the corresponding exact and noisy acceptance probabilities, respectively. Then, if

$$|\alpha_r - \tilde{\alpha}_r| \le \mathfrak{R}_r \tag{30}$$

for some finite constant $\Re_r > 0$, there exists a finite constant $\nu > 0$ such that:

$$||P_r - \tilde{P}_r|| \le \nu \mathfrak{K}_r \,, \tag{31}$$

where ν is independent of α_r and $\tilde{\alpha}_r$.

The proof of this proposition is given in Appendix B.2.

Now, we show that the error introduced by a full sweep of the MwG sampler over all of the model parameters is also bounded. We denote with $P_{[R]}$ (resp. $\tilde{P}_{[R]}$) the kernel corresponding to a sequential update of all of the model parameters using an exact (resp. approximate) acceptance probability:

$$P_{[R]}(u,\cdot) := P_1 \cdots P_R(u,\cdot) ,$$

$$= \int \cdots \int P_1(u,du_1) \cdots P_{R-1}(u_{R-2},du_{R-1}) P_R(u_{R-1},\cdot) ,$$

$$\tilde{P}_{[R]}(u,\cdot) := \tilde{P}_1 \cdots \tilde{P}_R(u,\cdot) .$$
(32)

This corresponds to the composition of the R elementary kernels, each characterising the update of one model parameter. The following theorem proves that if the errors on

the elementary transition kernels are bounded, then the error on the kernel for the full iteration is bounded as well.

Theorem 1. Let P_1, \ldots, P_R and $\tilde{P}_1, \ldots, \tilde{P}_R$ be a finite number of transition kernels defined on the compact set S and a constant $\Re < 1$ such that:

$$\sup_{r \le R} \|P_r - \tilde{P}_r\| \le \mathfrak{K}. \tag{33}$$

Then, the composite kernels defined in (32) satisfy:

$$||P_{[R]} - \tilde{P}_{[R]}|| \le R\mathfrak{K}. \tag{34}$$

Remark 5. In Theorem 1 the deterministic-scan assumption (see the definition of $P_{[R]}$ with Eq. (32)), is in fact not necessary: the result holds for any MwG sampler where component updates are performed sequentially, in any (even random) order.

The proof of this theorem is provided in Appendix B.3. Finally, as in Alquier et al. (2016), we rely on Corollary 3.1 of Mitrophanov (2005) to give our main result for the NoisyLPM algorithm.

Corollary 1. Let $P_{[R]}$ (resp. $\tilde{P}_{[R]}$) be the transition kernel for the exact MwG sampler (resp. noisy) described in Eq. (32). Assume that the Markov chain with kernel P is uniformly ergodic (Assumption 5). Then, for any t > 0 and for any starting point $u \in \mathcal{S}$:

$$\|\delta_u P_{[R]}^t - \delta_u \tilde{P}_{[R]}^t\| \le \left(\lambda + \frac{C\tau^{\lambda}}{1-\tau}\right) R\mathfrak{K}, \tag{35}$$

where $\lambda = \lceil \log(1/C)/\log(\tau) \rceil$.

7.2 LPM likelihood errors

We now report a series of theoretical results, specific to LPM, in preparation of applying Corollary 1 to this context. In the following theorem, we show that the likelihood error is bounded, and that it can be arbitrarily reduced by refining the latent grid partition.

Theorem 2. Under Assumptions 1, 2 and 3, the error on the noisy likelihood ratios for a latent position update (see Eq.(26)) satisfies for all $i \in V$

$$\left| \mathcal{LR}_{\mathcal{Z}} \left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime} \right) - \widetilde{\mathcal{LR}}_{\mathcal{Z}} \left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime} \right) \right| \leq \left\{ \frac{1 - 1/p^{\mathcal{L}}}{1 - 1/p^{\mathcal{U}}} \right\}^{N - 1} \left\{ 1 - e^{-\eta(b)} \right\}$$
(36)

and for a static parameter update (see Eq. (27)) for all $k \in \mathcal{K}$:

$$\left| \mathcal{L}\mathcal{R}_{\psi} \left(\psi_k \to \psi_k' \right) - \widetilde{\mathcal{L}\mathcal{R}}_{\psi} \left(\psi_k \to \psi_k' \right) \right| \le \left\{ \frac{1 - 1/p^{\mathcal{L}}}{1 - 1/p^{\mathcal{U}}} \right\}^{N(N-1)/2} \left\{ 1 - e^{-(N/2)\eta(b)} \right\} , \quad (37)$$

where

$$\eta(b) := \chi_1 b + \chi_2 \log(1 + \chi_3 b)$$
(38)

for suitable positive finite constants χ_1 , χ_2 and χ_3 defined at Eq. (68) which depend on N, $p^{\mathcal{U}}$, $p^{\mathcal{L}}$ and \varkappa_p but not on b.

The proof of Theorem 2 is given in Appendix B.4.

Remark 6. The upper bounds provided by Theorem 2 clearly converge to zero as the grid parameter b, denoting the sidelength of the boxes, goes to zero.

In particular, it can be proven that the bounds in Eq. (36) and Eq. (37) go to zero at the same rate as the functions

$$b \mapsto \left\{ 1 - \frac{1}{\left(1 + \varkappa_p \sqrt{2}b/(1 - p^{\mathcal{U}})\right)^{2(N-1)}} \right\} \text{ and } b \mapsto \left\{ 1 - \frac{1}{\left(1 + \varkappa_p \sqrt{2}b/(1 - p^{\mathcal{U}})\right)^{N(N-1)}} \right\}$$

converge, respectively.

The results exposed in the above theorem are exploited in the following corollary to bound the errors on the acceptance probabilities, for any elementary parameter update.

Corollary 2. Under Assumptions 1, 2, 3 and 4, the noisy acceptance probabilities, obtained by replacing the likelihood ratios in (11) and (12) with the noisy counterparts, satisfy the following:

$$\left|\alpha_{\psi}\left(\psi_{k} \to \psi_{k}'\right) - \tilde{\alpha}_{\psi}\left(\psi_{k} \to \psi_{k}'\right)\right| \le \kappa_{\pi} \kappa_{q} \left\{\frac{1 - 1/p^{\mathcal{L}}}{1 - 1/p^{\mathcal{U}}}\right\}^{N - 1} \left\{1 - e^{-\eta(b)}\right\}, \quad (39)$$

$$\left|\alpha_{\mathcal{Z}}\left(\boldsymbol{z}_{i} \rightarrow \boldsymbol{z}_{i}^{\prime}\right) - \tilde{\alpha}_{\mathcal{Z}}\left(\boldsymbol{z}_{i} \rightarrow \boldsymbol{z}_{i}^{\prime}\right)\right| \leq \varkappa_{\pi}\varkappa_{q} \left\{\frac{1 - 1/p^{\mathcal{L}}}{1 - 1/p^{\mathcal{U}}}\right\}^{N(N-1)/2} \left\{1 - e^{-(N/2)\eta(b)}\right\}$$
(40)

where $(\kappa_{\pi}, \varkappa_{\pi})$ are constants from Assumption 2, $(\kappa_{q}, \varkappa_{q})$ from Assumption 4 and η the function defined in Eq. (38).

The proof of this corollary is provided in Appendix B.5.

Theorem 3. Let $P_{[R]}$ be the exact MwG kernel which operates on $S = S_{\psi} \times S_{\mathcal{Z}}$ and $\tilde{P}_{[R]}$ be the kernel of NoisyLPM. If the model satisfies Assumptions 1, 2, 3, the random-walk proposal satisfies 4 and $P_{[R]}$ is uniformly ergodic (Assumption 5) then for any starting point $u \in S_{\psi} \times S_{\mathcal{Z}}$ and any t > 0:

$$\|\delta_{u}P_{[R]}^{t} - \delta_{u}\tilde{P}_{[R]}^{t}\|$$

$$\leq \left(\lambda + \frac{C\tau^{\lambda}}{1-\tau}\right)\left\{\kappa_{q}\kappa_{\pi} \vee \varkappa_{q}\varkappa_{\pi}\right\} \left\{\frac{1-1/p^{\mathcal{L}}}{1-1/p^{\mathcal{U}}}\right\}^{(N-1)} \left\{1 - e^{-(N/2)\eta(b)}\right\}, \quad (41)$$

where $\lambda = \lceil \log(1/C)/\log(\tau) \rceil$ depends on the exact sampler convergence properties, $\kappa_q, \kappa_\pi, \varkappa_q, \varkappa_\pi$ are constants related to Assumptions 2 and 4 and η is the function defined in Eq. (38).

Proof. Assumptions 1–4 guarantee that Corollary 2 holds. Therefore, Theorem 1 holds for the LPM sampler and its noisy version. Combining Assumption 5 and Theorem 1 yields that a LPM version of Corollary 1 exists, hence Eq. (41) holds true. □

In our context, this corollary gives a bound on the error for the transition kernel after an arbitrary number of iterations. This proves that the bias introduced by our noisy approximation is controlled by the level of refinement of the grid, which is in turn regulated by the arbitrarily chosen parameter b.

7.3 Note on the uniform convergence assumption

Assumption 5 is usually strong in the context of MCMC algorithms. However, since the state space is compact (see Assumption 1), it is easy to show that the convergence of the Gibbs kernel $P_{[R]}$ to π is uniform. Even though this result is not surprising, we could not identify a specific entry in the literature providing a rigorous proof of this fact. For completeness, we include Theorem 4 in Appendix C.

8 Experiments

In this section we propose two simulation studies to characterise the bias introduced by our approximation, and to gauge the gain in computing time achieved. We consider a LPM characterised by two global parameters $\psi = (\beta, \theta)$ which determine the edge probabilities as follows:

$$\log \left(\frac{p(\mathbf{z}_i, \mathbf{z}_j; \beta, \theta)}{1 - p(\mathbf{z}_i, \mathbf{z}_j; \beta, \theta)} \right) := \beta - e^{\theta} d(\mathbf{z}_i, \mathbf{z}_j). \tag{42}$$

Here, $\beta \in \mathbb{R}$, $\theta \in \mathbb{R}$, and d denotes the Euclidean distance between the two latent positions.

A priori, the latent positions are IID variables distributed according to a truncated Gaussian, as shown in (20). We fix both the threshold parameter S and the standard deviation γ to 1. This choice does not hinder the flexibility of the model; in fact, the likelihood parameter θ directly regulates the contribution given by the latent space. In other words, e^{θ} may simply be considered as the standard deviation for the latent positions. The likelihood parameters β and θ are assumed to be independent a priori, and both distributed according to non-informative Gaussian priors with fixed large standard deviations.

We note that the model specification considered does not completely satisfy 1 and 2, since, for example, the supports of β and θ are not bounded. However, we argue that large values of these parameters correspond to degenerate LPMs, that are of little interest in practical situations. In other words, the extreme values of the LPM parameters do not play a role and do not affect the MCMC estimation unless the observed graph is degenerate or near-degenerate.

8.1 Study 1: likelihood approximation

In the first study, we focus only on the approximation of the log-likelihood, i.e. we analyse the error introduced when (5) is replaced with the noisy counterpart in (25).

First, we generate random LPMs with global parameters set to $\beta = 0.5$ and $\theta = \log(3)$, and with latent positions drawn uniformly in the rectangle $\mathcal{S}_{\mathcal{Z}}$. This combination of parameters yields realised networks where about 10% of the possible edges appear. We simulate 1000 networks for each value of N varying in the set $\{200, 400, 600, 800, 1000\}$. For each of these realised networks, we evaluate the exact log-likelihood function derived from (5) for the true parameter values.

On each axis, the interval [-1,1] is segmented in M=8 adjacent intervals of the same length, hence obtaining a grid of 64 squared boxes of side length 1/4. The noisy log-likelihood derived from (25) is thus evaluated using such grid. The same procedure is repeated on the same networks using two finer grids determined by M=12 and M=16, respectively. Note that the finest grid contains 256 boxes; hence, it may give a computational advantage only when N is particularly large.

Figure 2 shows the error introduced by the noisy approximation, for some combina-

tions of N and M. Evidently, regardless of N, the bias of the approximation diminishes

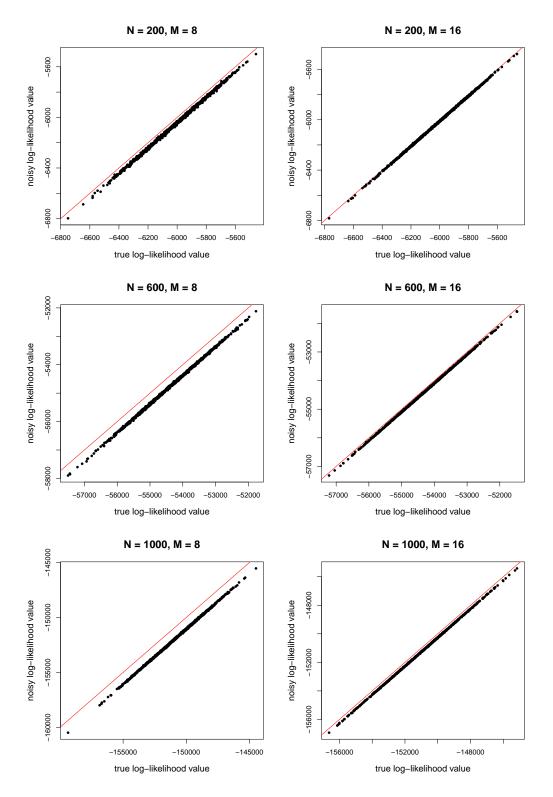


Figure 2: Simulation study 1. Comparison between true and noisy log-likelihood values for each of the artificial networks.

as the number of boxes increases. Across all of the cases considered, the approximation

is biased towards an underestimation of the log-likelihood value. The number of nodes N seems also to play a role in the magnitude of the bias, in that larger networks exhibit a larger approximation error.

Remark 7. The persistent underestimation of the log-likelihood value originates from the log transformation. Essentially, this is a demonstration of Jensen's inequality, as shown in Appendix B.6.

Figure 3 shows the average log-likelihood computing times for all of the combinations of N and M. This plot clearly shows that the order of complexity is smaller for the

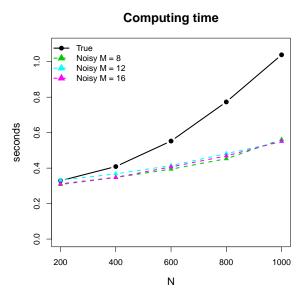


Figure 3: Simulation study 1. Average (across 1000 networks) computing time for the log-likelihood evaluations.

noisy log-likelihood, since the its computing time grows slower as N increases. For small networks, there seems to be no difference between noisy and non-noisy methods, since the construction of the grids generally requires a number of additional computations. However, for networks of 1000 nodes, the overall computing cost is approximately halved in the noisy method. The difference between the three grid approaches appears to be not particularly relevant and is mainly due to the randomness of the approach.

8.2 Study 2: Metropolis-within-Gibbs samplers

We focus now on the estimation of the LPM characterised by the edge probabilities in (42). For this task, we use and compare two MwG samplers: one corresponds to a standard implementation of the MwG algorithm as described in Section 3.2 and in Hoff

et al. (2002), whereas the other is our NoisyLPM introduced in Section 6. Our goal is to show that, when we use the non-noisy MwG as ground truth, the NoisyLPM achieves approximately the same results using only a fraction of the computing time.

As data, we use three artificial networks, which are generated following the same procedure of the previous study; hence, each of them has edge density close to 10%. One difference with the previous setup is that, in this study, node 1 is assumed to be located exactly at the origin of the space, for comparison purposes. The number of nodes N of the networks is set to 200, 400 and 600, respectively. For the NoisyLPM, we consider three different grid structures: as in the previous study, the number of intervals M in each axis varies in the set $\{8, 12, 16\}$.

The non-noisy MwG sampler and the NoisyLPM are run on each dataset for a total of 200,000 iterations. The first 100,000 iterations are discarded as burn-in, and only one draw every 10-th is stored to be kept in the final sample. Eventually, all of the algorithms are bound to return a collection of 10,000 draws for each model parameter.

Figure 4 shows the posterior densities for the node located in the centre of the space. The two NoisyLPM posterior densities shown are extremely similar to the ground truth,

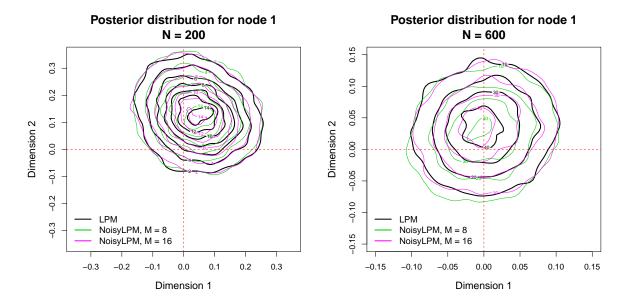


Figure 4: Simulation study 2. Posterior densities for the node in the centre.

proving that the uncertainty in the positioning is not necessarily amplified by the approximation.

Figure 5 compares instead the (posterior) average position of all nodes between ground truth and two NoisyLPM samples. Again, the approximation appears to have very limited

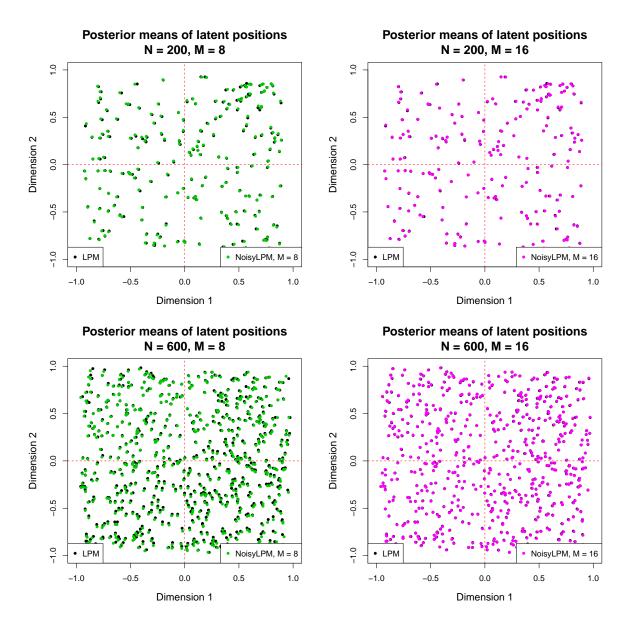


Figure 5: Simulation study 2. Comparison between ground truth and noisy estimates of the positions. The black circles correspond to the posterior means of the positions in the ground truth configuration, whereas the green and pink nodes correspond to the noisy counterparts.

consequences on the correctness of the results. In particular, the estimation error is almost non-existent when M=16.

Figures 6 and 7 illustrate the posterior densities for the global parameters β and θ . Note that, in both figures, the horizontal axes of the plots are on different scales. In fact, these plots confirm that the uncertainty on global parameters tends to vanish as N increases, for both non-noisy and noisy algorithms. As expected, a larger M gives results closer to the ground truth.

We further analyse the results by comparing the estimated edge probabilities, in Figure

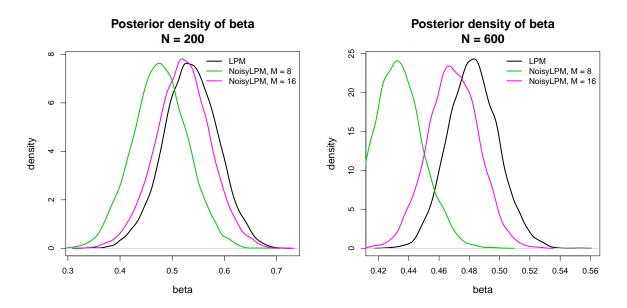


Figure 6: Simulation study 2. Posterior densities for β . Note the different scaling in the horizontal axis.

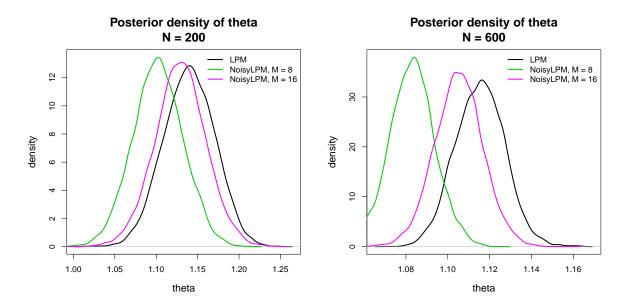


Figure 7: Simulation study 2. Posterior densities for θ . Note the different scaling in the horizontal axis.

8. These plots also confirm the correctness of the noisy procedure, and the limited effects of the approximation on the results.

Finally, in Table 1 we show the computing time required for each sampler. The highest gain is achieved for M=8 and N=600, where the NoisyLPM is roughly three times faster than the benchmark. As we will show in the next section, the gain can become substantial when larger networks are considered.

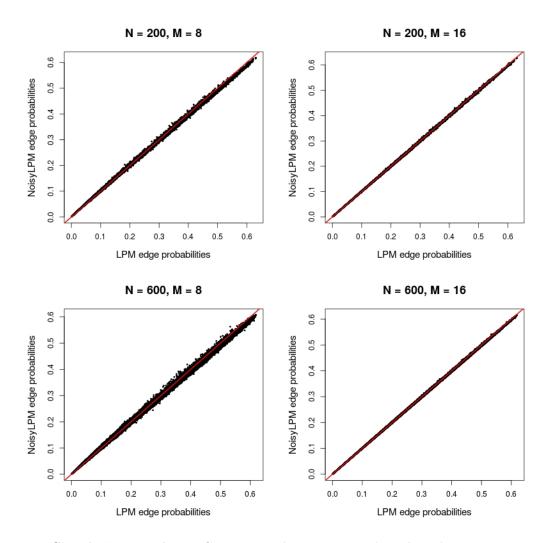


Figure 8: Simulation study 2. Comparison between ground truth and noisy estimates of the edge probabilities. These estimates are obtained by pluggin-in the posterior mean estimates of the model parameters in (42).

9 Coauthorship in astrophysics

The coauthorship network studied in this section was first analysed by Leskovec et al. (2007). The nodes correspond to authors, whereas the presence of an edge between two nodes means that the two researchers appear as coauthors on a paper submitted to arXiv, in the astrophysics category. The network is by construction undirected and without self-edges. The number of nodes is 18,872, whereas the number of edges is 198,110, corresponding to an average degree of about 21.

We fit the LPM of Section 8 to this data using the NoisyLPM with M=16. First, we let the algorithm run for a large number of iterations. We use this phase as burn-in, and to tune the proposal variances individually for each parameter until the corresponding

Table 1: Simulation study 2. Seconds (rounded value) required to obtain 200,000 iterations from each of the networks, for both algorithms.

N	Ground truth	NoisyLPM		
		M = 8	M = 12	M = 16
200	2,310	1,669	2,252	2,767
400	7,242	3,515	5,458	7,673
600	14,347	4,718	7,501	11,825

acceptance probability lies between 20% and 50%. Then, we run the NoisyLPM for 50,000 iterations, storing only one draw every 10-th. Trace plots and other standard convergence diagnostics suggest good mixing and good convergence of the chain to its stationary distribution. In summary, for each latent position and global parameter, we obtain 5,000 random draws that can be used to characterise the distribution of interest.

Figure 9 shows the average latent positions for all of the nodes in the network. We point out that the nodes have a tendency to be distributed close to the centre of each box. This is reasonably a natural consequence of our construction, since the centre of the boxes is used as a proxy to calculate the latent distances. For example, if a node with a low degree is connected only to nodes allocated to the same box, it will tend to move towards the centre of the same box, since that would maximise the likelihood of those edges appearing. More generally, we argue that, while the overall macro-structure of the network (i.e. the association of nodes to boxes, or the association of nodes to subregions of the space) is properly recovered, the micro-structure, given by the positions of the nodes within each box, may not necessarily be accurate.

Figure 10 shows instead the posterior densities for the global parameters β and θ . We find the parameter θ to be rather large, signalling that the heterogeneity of the graph is well captured by the latent space.

The computing time required to obtain the sample was about 46 hours (3.3 seconds per iteration). After convergence of the Markov chain, we also ran the non-noisy MwG sampler for 50 additional iterations, to compare the computational efficiency of the two methods. The non-noisy MwG sampler required an average of 453 seconds per iteration, corresponding to a theoretical 262 days of computations for the full sample. The vast difference between the two computing times highlights the linear computational complexity of our method, which extends the applicability of LPMs to networks of much larger sizes.

Posterior means of latent positions 0.5 Dimension 2 0.0 -0.5 0.0 0.5 1.0 -1.0 Dimension 1

Figure 9: **Astrophysics**. Average latent positions of the nodes with circle size proportional to node degree. The grid in dashed red line corresponds to the partitioning imposed.

10 Conclusions

In this paper, we have introduced a new methodology to perform inference on latent position models. Our approach specifically addresses a crucial issue: the scalability of the method with respect to the size of the network. By taking advantage of a discretisation of the latent space, our proposed approach is characterised by a computational complexity which is linear with respect to the number of nodes of the network.

The framework introduced heavily relies on several important results introduced in the

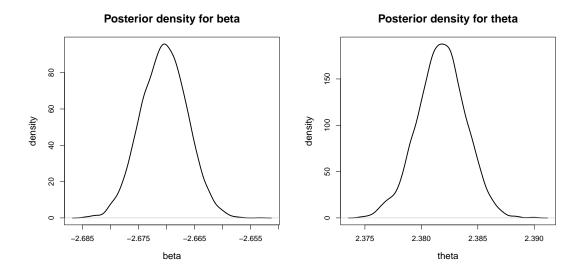


Figure 10: **Astrophysics**. Posterior densities for the global parameters β and θ .

context of noisy MCMC. We have followed the core ideas of such strand of literature, and adapted the main results to the latent position model context, thereby giving theoretical guarantees for our proposed approximate method. In particular, our results underline the existence of a trade-off between the speed and the bias of the noisy algorithm, whereby the user can arbitrarily increase the accuracy at the expense of speed. One possible extension of our results would include a characterisation of the bounds proposed as the number of nodes of the network increases. In fact, in the current formulation, the bounds given refer to a constant network size, and are not useful in the asymptotic scenario.

Additionally, we have proposed applications to both simulated and real datasets. When compared to the non-noisy algorithm, the noisy results did not show any relevant qualitative difference, yet they were obtained with a substantially smaller computing time.

A limitation of our work is that it does not cope well with an increasing number of latent dimensions. For example, introducing nodal random effects would increase the number of boxes to consider, hence dramatically reducing the computational gain. For the same reason, the introduction of covariates or any other node-specific features may not be practical. However, our work can be easily extended to include different distributions on the latent point process, such as Gaussian mixture models (Handcock et al. 2007). Also, the grid approximation may be extended to other types of edge probabilities, or, more generally, factor models, such as the projection models of Hoff et al. (2002).

Software

The method has been implemented in C++, and it uses parallel computing through the library OpenMPI. All of the computations described in the paper have been performed on a 8-cores (2.2 GHz) Debian machine. The code for both the non-noisy sampler and NoisyLPM are available from the corresponding author upon request.

Acknowledgements

Part of this research has been carried out while R. R. was affiliated with the Institute of Statistics and Mathematics, Vienna University of Economics and Business, Vienna, Austria; and funded through the Vienna Science and Technology Fund (WWTF) Project MA14-031. This research was also supported by the Insight Centre for Data Analytics through Science Foundation Ireland grant SFI/12/RC/2289.

References

- Alquier, P., N. Friel, R. Everitt, and A. Boland (2016). "Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels". In: *Statistics and Computing* 26.1-2, pp. 29–47.
- Boland, A., N. Friel, and F. Maire (2017). "Efficient MCMC for Gibbs Random Fields using pre-computation". In: arXiv preprint arXiv:1710.04093.
- Durante, D. and D. B. Dunson (2016). "Locally adaptive dynamic networks". In: *The Annals of Applied Statistics* 10.4, pp. 2203–2232.
- Durante, D., D. B. Dunson, and J. T. Vogelstein (2017). "Nonparametric Bayes modeling of populations of networks". In: *Journal of the American Statistical Association*, pp. 1–15.
- Friel, N., R. Rastelli, J. Wyse, and A. E. Raftery (2016). "Interlocking directorates in Irish companies using a latent space model for bipartite networks". In: *Proceedings of the National Academy of Sciences* 113.24, pp. 6629–6634.
- Gilbert, E. N. (1961). "Random plane networks". In: Journal of the Society for Industrial and Applied Mathematics 9.4, pp. 533–543.
- Gilks, W. R., N. G. Best, and K. K. C. Tan (1995). "Adaptive rejection Metropolis sampling within Gibbs sampling". In: *Applied Statistics*, pp. 455–472.
- Gollini, I. and T. B. Murphy (2014). "Joint modelling of multiple network views". In: Journal of Computational and Graphical Statistics.
- Gormley, I. C. and T. B. Murphy (2007). "A latent space model for rank data". In: Statistical Network Analysis: Models, Issues, and New Directions. Springer, pp. 90–102.
- Handcock, M. S., A. E. Raftery, and J. M. Tantrum (2007). "Model-based clustering for social networks". In: Journal of the Royal Statistical Society: Series A (Statistics in Society) 170.2, pp. 301–354.

- Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). "Latent space approaches to social network analysis". In: *Journal of the American Statistical Association* 97.460, pp. 1090–1098.
- Johndrow, J. E. and J. C. Mattingly (2017). "Error bounds for approximations of Markov chains". In: arXiv preprint arXiv:1711.05382.
- Krivitsky, P. N., M. S. Handcock, A. E. Raftery, and P. D. Hoff (2009). "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models". In: *Social networks* 31.3, pp. 204–213.
- Leskovec, J., J. Kleinberg, and C. Faloutsos (2007). "Graph evolution: Densification and shrinking diameters". In: ACM Transactions on Knowledge Discovery from Data (TKDD) 1.1, p. 2.
- Maire, F., N. Friel, and P. Alquier (2018). "Informed sub-sampling MCMC: approximate Bayesian inference for large datasets". In: *Statistics and Computing*, pp. 1–34. ISSN: 1573-1375. DOI: 10.1007/s11222-018-9817-3. URL: https://doi.org/10.1007/s11222-018-9817-3.
- Matias, C. and S. Robin (2014). "Modeling heterogeneity in random graphs through latent space models: a selective review". In: *ESAIM: Proceedings and Surveys* 47, pp. 55–74.
- Mitrophanov, A. Y. (2005). "Sensitivity and convergence of uniformly ergodic Markov chains". In: *Journal of Applied Probability* 42.4, pp. 1003–1014.
- Negrea, J. and J. S. Rosenthal (2017). "Error bounds for approximations of geometrically ergodic Markov chains". In: arXiv preprint arXiv:1702.07441.
- Nowicki, K. and T. A. B. Snijders (2001). "Estimation and prediction for stochastic blockstructures". In: *Journal of the American Statistical Association* 96.455, pp. 1077–1087.
- Parsonage, E. and M. Roughan (2017). "Fast generation of spatially embedded random networks". In: *IEEE Transactions on Network Science and Engineering* 4.2, pp. 112–119.
- Raftery, A. E. (2017). "Comment: extending the latent position model for networks". In: *Journal of the American Statistical Association* 112.520, pp. 1531–1534.
- Raftery, A. E., X. Niu, P. D. Hoff, and K. Y. Yeung (2012). "Fast inference for the latent space network model using a case-control approximate likelihood". In: *Journal of Computational and Graphical Statistics* 21.4, pp. 901–919.
- Rastelli, R., N. Friel, and A. E. Raftery (2016). "Properties of latent variable network models". In: *Network Science* 4.4, pp. 407–432. DOI: 10.1017/nws.2016.23.
- Roberts, G. O. and J. S. Rosenthal (1998). "Two convergence properties of hybrid samplers". In: *The Annals of Applied Probability* 8.2, pp. 397–407.
- Roberts, G. O. and J. S. Rosenthal (2004). "General state space Markov chains and MCMC algorithms". In: *Probability surveys* 1, pp. 20–71.
- Rudolf, D. and N. Schweizer (2017). "Perturbation theory for Markov chains via Wasserstein distance". In: *Bernoulli* 24.4A, pp. 2610–2639.
- Ryan, C., J. Wyse, and N. Friel (2017). "Bayesian model selection for the latent position cluster model for Social Networks". In: *Network Science* 5.1, pp. 70–91.
- Salter-Townshend, M. and T. H. McCormick (2017). "Latent space models for multiview network data". In: *The Annals of Applied Statistics* 11.3, pp. 1217–1244.

- Salter-Townshend, M. and T. B. Murphy (2013). "Variational Bayesian inference for the latent position cluster model for network data". In: *Computational Statistics & Data Analysis* 57.1, pp. 661–671.
- Salter-Townshend, M., A. White, I. Gollini, and T. B. Murphy (2012). "Review of statistical network analysis: models, algorithms, and software". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 5.4, pp. 243–264.
- Sarkar, P. and A. W. Moore (2006). "Dynamic social network analysis using latent space models". In: Advances in Neural Information Processing Systems, pp. 1145–1152.
- Sewell, D. K. and Y. Chen (2015a). "Analysis of the formation of the structure of social networks by using latent space models for ranked dynamic networks". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64.4, pp. 611–633.
- Sewell, D. K. and Y. Chen (2015b). "Latent space models for dynamic networks". In: *Journal of the American Statistical Association* 110.512, pp. 1646–1657.
- Sewell, D. K. and Y. Chen (2016). "Latent space models for dynamic networks with weighted edges". In: *Social Networks* 44, pp. 105–116.
- Shortreed, S., M. S. Handcock, and P. Hoff (2006). "Positional estimation within a latent space model for networks." In: *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 2.1, p. 24.
- Wang, Y. J. and G. Y. Wong (1987). "Stochastic blockmodels for directed graphs". In: *Journal of the American Statistical Association* 82.397, pp. 8–19.

Appendix

Notation: throughout this appendix, $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and cumulative density function of a standard Gaussian random variable, respectively.

A On the LPM model assumptions

A.1 Proof related to Assumption 2

Proof. The truncated Gaussian prior on the latent positions satisfies Assumption 2 since:

$$\frac{\pi\left(\mathbf{z}_{i}^{\prime}\right)}{\pi\left(\mathbf{z}_{i}\right)} \leq \prod_{m=1}^{2} \frac{\pi\left(z_{im}^{\prime}\right)}{\pi\left(z_{im}\right)} \leq \prod_{m=1}^{2} \frac{\phi\left(\frac{z_{im}^{\prime}}{\gamma}\right)}{\phi\left(\frac{z_{im}}{\gamma}\right)} \leq \prod_{m=1}^{2} \frac{\phi\left(\frac{S}{\gamma}\right)}{\phi\left(0\right)} = e^{-S^{2}/\gamma^{2}} = \kappa_{\pi},\tag{43}$$

for all \mathbf{z}_i' and \mathbf{z}_i in $\mathcal{S}_{\mathcal{Z}}$.

A.2 Proof related to Assumption 4

Proof. A truncated Gaussian proposal satisfies Assumption 4 since, for any $\mathbf{z}_i, \mathbf{z}_i' \in \mathcal{S}_{\mathcal{Z}}$:

$$\frac{q_{\mathcal{Z}}\left(\mathbf{z}_{i}^{\prime} \to \mathbf{z}_{i}\right)}{q_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right)} \leq \prod_{m=1}^{2} \frac{\Phi\left(\frac{S-z_{im}}{v}\right) - \Phi\left(\frac{-S-z_{im}}{v}\right)}{\Phi\left(\frac{S-z_{im}^{\prime}}{v}\right) - \Phi\left(\frac{-S-z_{im}^{\prime}}{v}\right)} \leq \prod_{m=1}^{2} \frac{2S\phi\left(0\right)}{2S\phi\left(-\frac{2S}{v}\right)} = e^{4S^{2}/v^{2}}$$

$$(44)$$

where v > 0 is the proposal's standard deviation. The last inequality follows from the fact that, for a fixed v, the area under $\phi\left(\frac{z-c}{v}\right)$ in $\mathcal{S}_{\mathcal{Z}}$ is maximised when c=0 and minimised when c=S.

B Proofs of Section 7

B.1 Lemma

The following lemma contains a basic result on Lipschitz functions that is used in the main proofs of this paper.

Lemma 1. Let f be a function with domain C = [a, b], for some $0 < a \le b < 0$, such that:

- $\forall x \in \mathcal{C}: 0 < f^{\mathcal{L}} \le f(x) \le f^{\mathcal{U}} < 1;$
- f is Lipschitz in x, that is:

$$\forall x_1, x_2 \in \mathcal{C}: |f(x_2) - f(x_1)| \le \kappa_f |x_2 - x_1|$$

for some finite constant κ_f .

Then, the following inequalities hold:

$$\exp\left\{-\frac{\kappa_f}{f^{\mathcal{L}}}|x_2 - x_1|\right\} \le \frac{f(x_2)}{f(x_1)} \le \exp\left\{\frac{\kappa_f}{f^{\mathcal{L}}}|x_2 - x_1|\right\} \tag{45}$$

$$\frac{1 - f^{\mathcal{U}}}{1 - f^{\mathcal{U}} + \kappa_f |x_2 - x_1|} \le \frac{1 - f(x_2)}{1 - f(x_1)} \le \frac{1 - f^{\mathcal{U}} + \kappa_f |x_2 - x_1|}{1 - f^{\mathcal{U}}}$$
(46)

for any x_1 , x_2 in C.

Proof. Regarding (45):

$$\frac{f\left(x_{2}\right)}{f\left(x_{1}\right)} = \exp\left\{\log\left(\frac{f\left(x_{2}\right)}{f\left(x_{1}\right)}\right)\right\} \leq \exp\left\{\left|\log\left(f\left(x_{2}\right)\right) - \log\left(f\left(x_{1}\right)\right)\right|\right\}. \tag{47}$$

Since $[a,b] \subset \mathbb{R}^+$, the log function is Lipschitz with Lipschitz constant equal to $1/f^{\mathcal{L}}$, hence:

$$\frac{f(x_2)}{f(x_1)} \le \exp\left\{\frac{1}{f^{\mathcal{L}}} \left| f(x_2) - f(x_1) \right| \right\} \le \exp\left\{\frac{\kappa_f}{f^{\mathcal{L}}} \left| x_2 - x_1 \right| \right\}. \tag{48}$$

Now, we can flip the terms on both sides of the inequality and rename the variables to get:

$$\frac{f(x_2)}{f(x_1)} \ge \exp\left\{-\frac{\kappa_f}{f^{\mathcal{L}}} |x_2 - x_1|\right\};\tag{49}$$

hence proving (45). Regarding (46):

$$\frac{1-f(x_2)}{1-f(x_1)} = \frac{1-f(x_2)+f(x_1)-f(x_1)}{1-f(x_1)}$$

$$= 1 - \left[\frac{f(x_2)-f(x_1)}{1-f(x_1)}\right]$$

$$\leq \left|1 - \left[\frac{f(x_2)-f(x_1)}{1-f(x_1)}\right]\right|$$

$$\leq 1 + \frac{|f(x_2)-f(x_1)|}{1-f(x_1)}$$

$$\leq 1 + \frac{\kappa_f |x_2 - x_1|}{1-f^{\mathcal{U}}}$$

$$\leq \frac{1-f^{\mathcal{U}} + \kappa_f |x_2 - x_1|}{1-f^{\mathcal{U}}}$$

Again, we can flip the terms on both sides of the inequality and rename the variables to get:

$$\frac{1 - f(x_2)}{1 - f(x_1)} \ge \frac{1 - f^{\mathcal{U}}}{1 - f^{\mathcal{U}} + \kappa_f |x_2 - x_1|};$$
(51)

and, hence, conclude the proof.

B.2 Proof of Proposition 1

Proof.

$$||P_{r} - \tilde{P}_{r}|| = \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{r}(u, A) - \tilde{P}_{r}(u, A) \right|$$

$$= \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| \int_{A} \alpha \left(u \to u' \right) du' - \int_{A} \tilde{\alpha} \left(u \to u' \right) du' \right|$$

$$= \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| \int_{A} \left[\alpha \left(u \to u' \right) - \tilde{\alpha} \left(u \to u' \right) \right] du' \right|$$

$$\leq \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \int_{A} |\alpha \left(u \to u' \right) - \tilde{\alpha} \left(u \to u' \right) | du'$$

$$\leq ||S|| \kappa_{\alpha}$$

$$(52)$$

where, in the LPM context, the two quantities ||S|| and κ_{α} are constants determined by the parameter space size and Corollary 2, respectively.

B.3 Proof of Theorem 1

Proof. If R = 1 then $||P_{11} - \tilde{P}_{11}|| = ||P_1 - \tilde{P}_1|| \le \Re$. If R = 2, then:

$$||P_{[2]} - \tilde{P}_{[2]}|| = ||P_{1}P_{2} - \tilde{P}_{1}\tilde{P}_{2}||$$

$$= \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{1}P_{2}(u, A) - \tilde{P}_{1}\tilde{P}_{2}(u, A) \right|$$

$$= \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{1}P_{2}(u, A) + P_{1}\tilde{P}_{2}(u, A) - P_{1}\tilde{P}_{2}(u, A) - \tilde{P}_{1}\tilde{P}_{2}(u, A) \right|$$

$$\leq \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{1}P_{2}(u, A) - P_{1}\tilde{P}_{2}(u, A) \right| + \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{1}\tilde{P}_{2}(u, A) - \tilde{P}_{1}\tilde{P}_{2}(u, A) \right|.$$
(53)

The first term of the last line satisfies the following:

$$\sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{1} P_{2}(u, A) - P_{1} \tilde{P}_{2}(u, A) \right| =$$

$$= \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| \int P_{1}(u, dv) P_{2}(v, A) - \int P_{1}(u, dv) \tilde{P}_{2}(v, A) \right|$$

$$= \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| \int P_{1}(u, dv) \left[P_{2}(v, A) - \tilde{P}_{2}(v, A) \right] \right|$$

$$\leq \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \int P_{1}(u, dv) \left| P_{2}(v, A) - \tilde{P}_{2}(v, A) \right|$$

$$\leq \|P_{2} - \tilde{P}_{2}\| \sup_{u \in \mathcal{S}} \int P_{1}(u, dv)$$

$$\leq \|P_{2} - \tilde{P}_{2}\| \leq \mathfrak{K}.$$
(54)

Similarly, the second term satisfies the following:

$$\sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{1} \tilde{P}_{2}(u, A) - \tilde{P}_{1} \tilde{P}_{2}(u, A) \right| =
= \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| \int P_{1}(u, dv) \tilde{P}_{2}(v, A) - \int \tilde{P}_{1}(u, dv) \tilde{P}_{2}(v, A) \right|
= \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| \int \left[P_{1}(u, dv) - \tilde{P}_{1}(u, dv) \right] \tilde{P}_{2}(v, A) \right|
\leq \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \int \left| P_{1}(u, dv) - \tilde{P}_{1}(u, dv) \right| \tilde{P}_{2}(v, A)
\leq \|P_{1} - \tilde{P}_{1}\| \sup_{A \subset \mathcal{S}} \int \tilde{P}_{2}(dv, A)
\leq \|P_{1} - \tilde{P}_{1}\| < \Re$$
(55)

As a consequence:

$$||P_{[2]} - \tilde{P}_{[2]}|| \le 2\mathfrak{K}. \tag{56}$$

Now, we assume that (34) is valid for every $r \leq R - 1$, and prove the statement for r = R.

$$\begin{split} \|P_{[R]} - \tilde{P}_{[R]}\| &= \|P_{[R-1]}P_R - \tilde{P}_{[R-1]}\tilde{P}_R\| \\ &= \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{[R-1]}P_R(u, A) - \tilde{P}_{[R-1]}\tilde{P}_R(u, A) \right| \\ &= \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{[R-1]}P_R(u, A) + P_{[R-1]}\tilde{P}_R(u, A) - P_{[R-1]}\tilde{P}_R(u, A) - \tilde{P}_{[R-1]}\tilde{P}_R(u, A) \right| \\ &\leq \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{[R-1]}P_R(u, A) - P_{[R-1]}\tilde{P}_R(u, A) \right| + \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{[R-1]}\tilde{P}_R(u, A) - \tilde{P}_{[R-1]}\tilde{P}_R(u, A) \right|. \end{split}$$

$$(57)$$

The first term of the last line satisfies the following:

$$\sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{[R-1]} P_R(u, A) - P_{[R-1]} \tilde{P}_R(u, A) \right| =
= \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| \int P_{[R-1]}(u, dv) P_R(v, A) - \int P_{[R-1]}(u, dv) \tilde{P}_R(v, A) \right|
\leq \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \int P_{[R-1]}(u, dv) \left| P_R(v, A) - \tilde{P}_R(v, A) \right|
\leq \|P_R - \tilde{P}_R\| \sup_{u \in \mathcal{S}} \int P_{[R-1]}(u, dv)
\leq \|P_R - \tilde{P}_R\| \leq \mathfrak{K}.$$
(58)

In (58) we have used that:

$$\sup_{u \in \mathcal{S}} \int P_{[R-1]}(u, dv) = \sup_{u \in \mathcal{S}} \int P_{[R-2]} P_{R-1}(u, dv)
= \sup_{u \in \mathcal{S}} \int \int P_{[R-2]}(u, dw) P_{R-1}(w, dv)
= \sup_{u \in \mathcal{S}} \int P_{[R-2]}(u, dw) \int P_{R-1}(w, dv)
= \sup_{u \in \mathcal{S}} \int P_{[R-2]}(u, dw)
= \cdots
= \sup_{u \in \mathcal{S}} \int P_{1}(u, dw) = 1$$
(59)

Similarly, the second term satisfies the following:

$$\sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| P_{[R-1]} \tilde{P}_{R}(u, A) - \tilde{P}_{[R-1]} \tilde{P}_{R}(u, A) \right| =
= \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \left| \int \left[P_{[R-1]}(u, dv) - \tilde{P}_{[R-1]}(u, dv) \right] \tilde{P}_{R}(v, A) \right|
\leq \sup_{u \in \mathcal{S}} \sup_{A \subset \mathcal{S}} \int \left| P_{[R-1]}(u, dv) - \tilde{P}_{[R-1]}(u, dv) \right| \tilde{P}_{R}(v, A)
\leq \|P_{[R-1]} - \tilde{P}_{[R-1]}\| \sup_{A \subset \mathcal{S}} \int \tilde{P}_{R}(dv, A)
\leq \|P_{[R-1]} - \tilde{P}_{[R-1]}\| \leq (R-1) \Re$$
(60)

Finally, using the inequalities (58) and (60) in (57), we obtain the following:

$$||P_{[R]} - \tilde{P}_{[R]}|| \le R\mathfrak{K},$$
 (61)

proving the theorem by mathematical induction.

B.4 Proof of Theorem 2

Proof. We first show (36). The likelihood ratios for a latent position update are defined in (13), whereas the noisy counterparts are defined in (26). Since the edge probabilities are always greater than zero (Assumption 3), the likelihood ratios must be strictly positive, thus the following holds true:

$$\left| \mathcal{LR}_{\mathcal{Z}} \left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime} \right) - \widetilde{\mathcal{LR}}_{\mathcal{Z}} \left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime} \right) \right| \leq \left| \mathcal{LR}_{\mathcal{Z}} \left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime} \right) \left[1 - \frac{\widetilde{\mathcal{LR}}_{\mathcal{Z}} \left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime} \right)}{\mathcal{LR}_{\mathcal{Z}} \left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime} \right)} \right] \right|,$$

$$= \mathcal{LR}_{\mathcal{Z}} \left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime} \right) \cdot \left| 1 - \frac{\widetilde{\mathcal{LR}}_{\mathcal{Z}} \left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime} \right)}{\mathcal{LR}_{\mathcal{Z}} \left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime} \right)} \right|.$$
(62)

Now, we address the two terms on the right hand side of the above inequality separately.

First, we point out that the first term, $\mathcal{LR}_{\mathcal{Z}}(\mathbf{z}_i \to \mathbf{z}_i')$, is a finite constant independent of the box sidelength b, due to:

$$\mathcal{LR}_{\mathcal{Z}}(\mathbf{z}_{i} \to \mathbf{z}_{i}') = \prod_{j \in Y_{i}^{1}} \frac{p\left(\mathbf{z}_{i}', \mathbf{z}_{j}; \boldsymbol{\psi}\right)}{p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}\right)} \prod_{j \in Y_{i}^{0}} \frac{1 - p\left(\mathbf{z}_{i}', \mathbf{z}_{j}; \boldsymbol{\psi}\right)}{1 - p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}\right)}$$

$$\leq \prod_{j \in Y_{i}^{1}} \frac{p^{\mathcal{U}}}{p^{\mathcal{L}}} \prod_{j \in Y_{i}^{0}} \frac{1 - p^{\mathcal{L}}}{1 - p^{\mathcal{U}}}$$

$$\leq \left[\frac{p^{\mathcal{U}}}{p^{\mathcal{L}}}\right]^{D_{i}} \left[\frac{1 - p^{\mathcal{L}}}{1 - p^{\mathcal{U}}}\right]^{N - 1 - D_{i}};$$
(63)

and, similarly:

$$\mathcal{LR}_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right) \geq \left[\frac{p^{\mathcal{L}}}{p^{\mathcal{U}}}\right]^{D_{i}} \left[\frac{1-p^{\mathcal{U}}}{1-p^{\mathcal{L}}}\right]^{N-1-D_{i}}.$$
 (64)

Now we focus instead on the second term at the RHS of (62) to show that it goes to zero as b decreases. Here, we use the fact that the edge probability function satisfies the assumptions of Lemma 1. First we focus on the following ratio:

$$\frac{\widetilde{\mathcal{LR}}_{\mathcal{Z}}(\mathbf{z}_{i} \to \mathbf{z}_{i}')}{\mathcal{LR}_{\mathcal{Z}}(\mathbf{z}_{i} \to \mathbf{z}_{i}')} = \prod_{j \in Y_{i}^{1}} \frac{p(\mathbf{z}_{i}', \mathbf{c}_{j}; \boldsymbol{\psi})}{p(\mathbf{z}_{i}', \mathbf{z}_{j}; \boldsymbol{\psi})} \cdot \frac{p(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi})}{p(\mathbf{z}_{i}, \mathbf{c}_{j}; \boldsymbol{\psi})} \prod_{j \in Y_{i}^{0}} \frac{1 - p(\mathbf{z}_{i}', \mathbf{c}_{j}; \boldsymbol{\psi})}{1 - p(\mathbf{z}_{i}', \mathbf{z}_{j}; \boldsymbol{\psi})} \cdot \frac{1 - p(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi})}{1 - p(\mathbf{z}_{i}, \mathbf{c}_{j}; \boldsymbol{\psi})},$$

$$\leq \prod_{j \in Y_{i}^{1}} \exp \left\{ \exp \left\{ \frac{\varkappa_{p}}{p^{\mathcal{L}}} |d(\mathbf{z}_{i}', \mathbf{z}_{j}) - d(\mathbf{z}_{i}', \mathbf{c}_{j})| + \frac{\varkappa_{p}}{p^{\mathcal{L}}} |d(\mathbf{z}_{i}, \mathbf{z}_{j}) - d(\mathbf{z}_{i}, \mathbf{c}_{j})| \right\} \right\}$$

$$\times \prod_{j \in Y_{i}^{0}} \exp \left\{ \log \left(\frac{1 - p^{\mathcal{U}} + \varkappa_{p} |d(\mathbf{z}_{i}', \mathbf{z}_{j}) - d(\mathbf{z}_{i}', \mathbf{c}_{j})|}{1 - p^{\mathcal{U}}} \right) + \log \left(\frac{1 - p^{\mathcal{U}} + \varkappa_{p} |d(\mathbf{z}_{i}, \mathbf{z}_{j}) - d(\mathbf{z}_{i}', \mathbf{c}_{j})|}{1 - p^{\mathcal{U}}} \right) \right\}.$$
(65)

Remark 8. By construction, the points \mathbf{z}_i and \mathbf{c}_i belong to the same box, hence the following holds:

$$d\left(\mathbf{z}_{j},\mathbf{c}_{j}\right)\leq b\sqrt{2}.$$

The triangular inequality guarantees that, for any \mathbf{z}_i :

$$d(\mathbf{z}_i, \mathbf{z}_j) + d(\mathbf{z}_j, \mathbf{c}_j) \ge d(\mathbf{z}_i, \mathbf{c}_j)$$

which implies

$$d(\mathbf{z}_i, \mathbf{c}_j) - d(\mathbf{z}_i, \mathbf{z}_j) \le d(\mathbf{z}_j, \mathbf{c}_j)$$

Also:

$$d(\mathbf{z}_i, \mathbf{c}_j) + d(\mathbf{z}_j, \mathbf{c}_j) \ge d(\mathbf{z}_i, \mathbf{z}_j)$$

which implies

$$d(\mathbf{z}_i, \mathbf{c}_i) - d(\mathbf{z}_i, \mathbf{z}_i) \ge -d(\mathbf{z}_i, \mathbf{c}_i)$$
.

Hence, the two results combined imply:

$$|d(\mathbf{z}_i, \mathbf{z}_j) - d(\mathbf{z}_i, \mathbf{c}_j)| \le d(\mathbf{z}_j, \mathbf{c}_j) \le b\sqrt{2}.$$

Using the results in the above remark, we obtain:

$$\frac{\widetilde{\mathcal{LR}}_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right)}{\mathcal{LR}_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right)} \leq \prod_{j \in Y_{i}^{1}} \exp \left\{2\sqrt{2} \frac{\varkappa_{p}}{p^{\mathcal{L}}} b\right\} \prod_{j \in Y_{i}^{0}} \exp \left\{2\log \left(\frac{1 - p^{\mathcal{U}} + \varkappa_{p} b\sqrt{2}}{1 - p^{\mathcal{U}}}\right)\right\},$$

$$\leq \exp \left\{2\sqrt{2} \frac{\varkappa_{p}}{p^{\mathcal{L}}} D_{i} b + 2\left(N - 1 - D_{i}\right) \log \left(\frac{1 - p^{\mathcal{U}} + \varkappa_{p} b\sqrt{2}}{1 - p^{\mathcal{U}}}\right)\right\}.$$
(66)

In other words, the above inequality can be summarised by:

$$\frac{\widetilde{\mathcal{LR}}_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right)}{\mathcal{LR}_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right)} \leq \exp\{\eta\left(b\right)\},\tag{67}$$

where

$$\eta(b) := \chi_1 b + \chi_2 \log(1 + \chi_3 b) , \qquad (68)$$

with positive constants $\chi_1 := 2\sqrt{2}\varkappa_p(N-1)/p^{\mathcal{L}}$, $\chi_2 := 2(N-1)$ and $\chi_3 := \varkappa_p\sqrt{2}/1 - p^{\mathcal{U}}$ independent of b. Using the majorisation parts of the lemma (49) and (51), the following can be obtained analogously:

$$\frac{\widetilde{\mathcal{LR}}_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right)}{\mathcal{LR}_{\mathcal{Z}}\left(\mathbf{z}_{i} \to \mathbf{z}_{i}^{\prime}\right)} \ge \exp\{-\eta\left(b\right)\}. \tag{69}$$

Using those inequalities yields

$$e^{-\eta(b)} \le \frac{\widetilde{\mathcal{LR}}_{\mathcal{Z}}(\mathbf{z}_i \to \mathbf{z}_i')}{\mathcal{LR}_{\mathcal{Z}}(\mathbf{z}_i \to \mathbf{z}_i')} \le e^{\eta(b)} \quad \iff \quad 1 - e^{\eta(b)} \le 1 - \frac{\widetilde{\mathcal{LR}}_{\mathcal{Z}}(\mathbf{z}_i \to \mathbf{z}_i')}{\mathcal{LR}_{\mathcal{Z}}(\mathbf{z}_i \to \mathbf{z}_i')} \le 1 - e^{-\eta(b)}, \quad (70)$$

which implies

$$\left| 1 - \frac{\widetilde{\mathcal{LR}}_{\mathcal{Z}} \left(\mathbf{z}_i \to \mathbf{z}_i' \right)}{\mathcal{LR}_{\mathcal{Z}} \left(\mathbf{z}_i \to \mathbf{z}_i' \right)} \right| \le \max \left\{ 1 - e^{-\eta(b)}, 1 - e^{\eta(b)} \right\} = 1 - e^{-\eta(b)}$$

$$(71)$$

hence proving (36) and the first part of the proof relating to the updates of the latent positions.

We follow the same ideas to show that a similar bound holds for the update of the static parameters, as in (37). Thanks to Assumption 3 we can write:

$$\left| \mathcal{L}\mathcal{R}_{\psi} \left(\psi_{a} \to \psi_{a}^{\prime} \right) - \widetilde{\mathcal{L}}\mathcal{R}_{\psi} \left(\psi_{a} \to \psi_{a}^{\prime} \right) \right| \leq \mathcal{L}\mathcal{R}_{\psi} \left(\psi_{a} \to \psi_{a}^{\prime} \right) \cdot \left| 1 - \frac{\widetilde{\mathcal{L}}\mathcal{R}_{\psi} \left(\psi_{a} \to \psi_{a}^{\prime} \right)}{\mathcal{L}\mathcal{R}_{\psi} \left(\psi_{a} \to \psi_{a}^{\prime} \right)} \right|. \tag{72}$$

Similarly as before, the exact likelihood ratios do not depend on b and are never zero or infinity:

$$\mathcal{LR}_{\boldsymbol{\psi}}\left(\psi_{a} \to \psi_{a}^{\prime}\right) = \left[\prod_{i=1}^{N} \prod_{j \in Y_{i}^{1}} \frac{p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}^{\prime}\right)}{p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}\right)} \prod_{j \in Y_{i}^{0}} \frac{1 - p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}^{\prime}\right)}{1 - p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}\right)}\right]^{1/2},$$

$$\leq \left[\frac{p^{\mathcal{U}}}{p^{\mathcal{L}}}\right]^{\sum_{i=1}^{N} D_{i}/2} \left[\frac{1 - p^{\mathcal{L}}}{1 - p^{\mathcal{U}}}\right]^{(N^{2} - N - \sum_{i=1}^{N} D_{i})/2} \tag{73}$$

and

$$\mathcal{LR}_{\psi}\left(\psi_{a} \to \psi_{a}^{\prime}\right) \ge \left[\frac{p^{\mathcal{L}}}{p^{\mathcal{U}}}\right]^{\sum_{i=1}^{N} D_{i}/2} \left[\frac{1-p^{\mathcal{U}}}{1-p^{\mathcal{L}}}\right]^{(N^{2}-N-\sum_{i=1}^{N} D_{i})/2}.$$

$$(74)$$

Now we focus instead on the second and last term of (72) to show that it goes to zero as b decreases.

$$\frac{\widetilde{\mathcal{LR}}_{\boldsymbol{\psi}}(\psi_{a} \to \psi_{a}')}{\mathcal{LR}_{\boldsymbol{\psi}}(\psi_{a} \to \psi_{a}')} = \left[\prod_{i \in \mathcal{V}} \prod_{j \in Y_{i}^{1}} \frac{p\left(\mathbf{z}_{i}, \mathbf{c}_{j}; \boldsymbol{\psi}'\right)}{p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}'\right)} \frac{p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}'\right)}{p\left(\mathbf{z}_{i}, \mathbf{c}_{j}; \boldsymbol{\psi}'\right)} \prod_{j \in Y_{i}^{0}} \frac{1 - p\left(\mathbf{z}_{i}, \mathbf{c}_{j}; \boldsymbol{\psi}'\right)}{1 - p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}'\right)} \cdot \frac{1 - p\left(\mathbf{z}_{i}, \mathbf{z}_{j}; \boldsymbol{\psi}\right)}{1 - p\left(\mathbf{z}_{i}, \mathbf{c}_{j}; \boldsymbol{\psi}\right)} \right]^{1/2},$$

$$\leq \left[\prod_{i \in \mathcal{V}} \prod_{j \in Y_{i}^{1}} \exp\left\{ 2 \frac{\varkappa_{p}}{p^{\mathcal{L}}} \left| d\left(\mathbf{z}_{i}, \mathbf{z}_{j}\right) - d\left(\mathbf{z}_{i}, \mathbf{c}_{j}\right) \right| \right\} \right]^{1/2} \times \prod_{j \in Y_{i}^{0}} \exp\left\{ 2 \log\left(\frac{1 - p^{\mathcal{U}} + \varkappa_{p} \left| d\left(\mathbf{z}_{i}, \mathbf{z}_{j}\right) - d\left(\mathbf{z}_{i}, \mathbf{c}_{j}\right) \right|}{1 - p^{\mathcal{U}}} \right) \right\} \right]^{1/2},$$

$$\leq \exp\left\{ \sqrt{2} \left(\sum_{i \in \mathcal{V}} D_{i} \right) \frac{\varkappa_{p}}{p^{\mathcal{L}}} b + \left(N^{2} - N - \sum_{i \in \mathcal{V}} D_{i} \right) \log\left(\frac{1 - p^{\mathcal{U}} + \varkappa_{p} b\sqrt{2}}{1 - p^{\mathcal{U}}} \right) \right\}.$$

$$(75)$$

In other words, the above inequality can be summarised through

$$\frac{\widetilde{\mathcal{LR}_{\psi}}(\psi_a \to \psi_a')}{\mathcal{LR}_{\psi}(\psi_a \to \psi_a')} \le \exp\{(N/2)\eta(b)\}$$
(76)

where η is the function defined at Eq. (68). This essentially concludes the proof of the theorem, since all the following steps are exactly the same as in (69), (70) and (71).

B.5 Proof of Corollary 2

Proof. First note that the function $g(x) = 1 \land x \ (x \in \mathbb{R})$ is Lipschitz in any compact $\mathcal{C} \subset (0,1)$ since it satisfies:

$$|1 \wedge x_1 - 1 \wedge x_2| \le |x_1 - x_2|, \qquad \forall (x_1, x_2) \in \mathcal{C}^2.$$
 (77)

Then, the following holds:

$$|\alpha_{\mathcal{Z}}(\mathbf{z}_{i} \to \mathbf{z}_{i}') - \widetilde{\alpha}_{\mathcal{Z}}(\mathbf{z}_{i} \to \mathbf{z}_{i}')| =$$

$$= \left| \frac{q_{\mathcal{Z}}(\mathbf{z}_{i}' \to \mathbf{z}_{i})}{q_{\mathcal{Z}}(\mathbf{z}_{i} \to \mathbf{z}_{i}')} \cdot \frac{\pi(\mathbf{z}_{i}')}{\pi(\mathbf{z}_{i})} \cdot \mathcal{L}\mathcal{R}_{\mathcal{Z}}(\mathbf{z}_{i} \to \mathbf{z}_{i}') - \frac{q_{\mathcal{Z}}(\mathbf{z}_{i}' \to \mathbf{z}_{i})}{q_{\mathcal{Z}}(\mathbf{z}_{i} \to \mathbf{z}_{i}')} \cdot \frac{\pi(\mathbf{z}_{i}')}{\pi(\mathbf{z}_{i})} \cdot \widetilde{\mathcal{L}}\mathcal{R}_{\mathcal{Z}}(\mathbf{z}_{i} \to \mathbf{z}_{i}') \right|$$

$$\leq \frac{q_{\mathcal{Z}}(\mathbf{z}_{i}' \to \mathbf{z}_{i})}{q_{\mathcal{Z}}(\mathbf{z}_{i} \to \mathbf{z}_{i}')} \frac{\pi(\mathbf{z}_{i}')}{\pi(\mathbf{z}_{i})} \left| \mathcal{L}\mathcal{R}_{\mathcal{Z}}(\mathbf{z}_{i} \to \mathbf{z}_{i}') - \widetilde{\mathcal{L}}\mathcal{R}_{\mathcal{Z}}(\mathbf{z}_{i} \to \mathbf{z}_{i}') \right|$$

$$\leq \varkappa_{q} \varkappa_{\pi} \left\{ \frac{1 - 1/p^{\mathcal{L}}}{1 - 1/p^{\mathcal{U}}} \right\}^{N(N-1)/2} \left(1 - e^{-(N/2)\eta(b)} \right)$$

$$(78)$$

hence proving (40). The proof of (39) is analogous.

B.6 A note on the remark in Section 8.1

The likelihood function of a LPM is made of a number of terms, say:

$$\mathcal{L}_a = \prod_{i=1}^M a_i. \tag{79}$$

In our framework, we construct estimators $\{b_i\}_i$ for each of the likelihood terms $\{a_i\}_i$. The noisy likelihood may be written as:

$$\mathcal{L}_b = \prod_{i=1}^M b_i; \tag{80}$$

whereas the log-likelihoods are defined as:

$$\ell_a = \sum_{i=1}^{M} \log(a_i), \qquad \qquad \ell_b = \sum_{i=1}^{M} \log(b_i).$$
 (81)

In our paper we do not show whether the estimators b_i are biased or not, so there is no way to know if the noisy likelihood is unbiased. As a consequence, we cannot say much on the bias of the noisy log-likelihood, either. Now, even if we assume that the estimators are unbiased, i.e.

$$\mathbb{E}[b_i] = a_i, \qquad \forall i \in \{1, \dots, N\}. \tag{82}$$

then, by Jensen's inequality, we can only achieve the following result:

$$\mathbb{E}\left[\ell_{b}\right] = \mathbb{E}\left[\sum_{i=1}^{M} \log\left(b_{i}\right)\right] = \sum_{i=1}^{M} \mathbb{E}\left[\log\left(b_{i}\right)\right] \le \sum_{i=1}^{M} \log\mathbb{E}\left[\left(b_{i}\right)\right] = \sum_{i=1}^{M} \log\left(a_{i}\right) = \ell_{a}; \tag{83}$$

which is in agreement with the results shown in Section 8.1.

C Uniform convergence of Metropolis-within-Gibbs kernels operating on a compact state space

Theorem 4. Let S be a bounded state space with $S \subset \mathbb{R}^d$ (for some d > 0) and A be a sigma-algebra on S. Let P be a Gibbs kernel operating on $S \times A$ with invariant distribution π defined on (S, A). Then the function $u \mapsto \|P(u, \cdot)^t - \pi\|$ converges uniformly to 0 as $t \to \infty$, at a geometric rate.

Proof. For simplicity, we take the case d=3, but generalizing the following reasoning for all d>0 is straightforward. Denoting with P_i the MwG kernel that keeps $x_{-i}:=(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d)$ fixed, we have for all $x\in\mathcal{S}$

$$P_i(x, dx') = \{Q_i(x, dx'_i)\alpha_i(x, x') + \delta_{x_i}(dx'_i)\rho_i(x)\} \delta_{x_{-i}}(dx'_{-i}),$$
(84)

where Q_i is the proposal kernel of the *i*-th dimension, $\alpha_i(x, x') = 1 \wedge \pi(x')Q_i(x'x)/\pi(x)Q_i(x, x')$ and $\rho_i(x) = 1 - \int Q(x, dx')\alpha_i(x, x')$. With regulatory conditions on the proposal kernels Q_1, Q_2, \ldots and since the state space is compact, we have for all $i \in \{1, \ldots, d\}$:

$$\overline{Q}_i := \sup_{(x,y) \in S^2} Q_i(x,y) < \infty, \qquad \underline{Q}_i := \inf_{(x,y) \in S^2} Q_i(x,y) > 0.$$
(85)

Moreover, since the pdf of π is a continuous function and S is bounded, we have:

$$0 < \underline{\pi} \le \pi(x) \le \overline{\pi} < \infty. \tag{86}$$

Assuming that, for all i, Q_i is absolutely dominated by a common dominating measure, we have that $Q_i(x, dx'_i) = Q(x, x'_i)dx'_i$ which combined with Eqs. (84), (85) and (86) yields

$$P_i(x, \mathrm{d}x') \ge Q_i(x, x_i') \alpha_i(x, x') \delta_{x_{-i}}(\mathrm{d}x_{-i}') \mathrm{d}x_i' \ge Q_i \underline{\alpha}_i \delta_{x_{-i}}(\mathrm{d}x_{-i}') \mathrm{d}x_i', \tag{87}$$

where $\underline{\alpha}_i := \underline{\pi}Q_i/\overline{\pi}\overline{Q}_i$. Now, the (systematic-scan) Metropolis-within-Gibbs transition kernel writes

$$\begin{split} P(x,\mathrm{d} x') &:= P_1 P_2 P_3(x,\mathrm{d} x') = \int P_1(x,\mathrm{d} y) P_2 P_3(y,\mathrm{d} x') \,, \\ &\geq \int \underline{Q}_1 \underline{\alpha}_1 \mathrm{d} y_1 P_2 P_3(y_1,x_2,x_3,\mathrm{d} x') \,, \\ &\geq \int \underline{Q}_1 \underline{\alpha}_1 \mathrm{d} y_1 \int \underline{Q}_2 \underline{\alpha}_2 \mathrm{d} z_2 P_3(y_1,z_2,x_3,\mathrm{d} x') \,, \end{split}$$

$$\geq \int \underline{Q}_{1}\underline{\alpha}_{1} dy_{1} \int \underline{Q}_{2}\underline{\alpha}_{2} dz_{2} \underline{Q}_{3} \left\{ 1 \wedge \frac{\pi(x')\overline{Q}_{3}}{\underline{\pi}\underline{Q}_{3}} \right\} dx'_{3} \delta_{y_{1}}(dx'_{1}) \delta_{z_{2}}(dx'_{2}),$$

$$\geq \left\{ \prod_{i=1}^{2} \underline{Q}_{i}\underline{\alpha}_{i} \right\} \underline{Q}_{3} \left\{ 1 \wedge \frac{\pi(x')\underline{Q}_{3}}{\overline{\pi}\overline{Q}_{3}} \right\} dx',$$

since $\iint \mathrm{d}y_1 \delta_{y_1}(\mathrm{d}x_1') \mathrm{d}z_2 \delta_{z_2}(\mathrm{d}x_2') = \mathrm{d}x_1' \mathrm{d}x_2'$. Hence, defining ν as the absolutely continuous probability measure with pdf $\nu(x) \propto 1 \wedge \pi(x') \underline{Q}_3/\overline{\pi} \overline{Q}_3$, we have

$$P(x, dx') \ge \beta \nu(dx'),$$
 (88)

with $\beta := \left\{\prod_{i=1}^2 \underline{Q}_i \underline{\alpha}_i\right\} \underline{Q}_3 \int 1 \wedge \pi(x') \underline{Q}_3 / \overline{\pi} \overline{Q}_3 dx'$. We conclude from Eq. (88) that the whole state space $\mathcal S$ is small for P and that therefore P is uniformly ergodic (with geometric rate $1-\beta$), see e.g. Theorem 8 in Roberts and Rosenthal 2004.