Spatial Processes for Recommender Systems

Fabian Bohnert, Daniel F. Schmidt, and Ingrid Zukerman
Faculty of Information Technology, Monash University
Clayton, VIC 3800, Australia
{fabianb,dschmidt,ingrid}@infotech.monash.edu.au

Abstract

Spatial processes are typically used to analyse and predict geographic data. This paper adapts such models to predicting a user's interests (i. e., implicit item ratings) within a recommender system in the museum domain. We present the theoretical framework for a model based on Gaussian spatial processes, and discuss efficient algorithms for parameter estimation. Our model was evaluated with a real-world dataset collected by tracking visitors in a museum, attaining a higher predictive accuracy than state-of-the-art collaborative filters.

1 Introduction

Recommender systems (RS) are designed to direct users to personally interesting items in situations where the amount of available information exceeds the users' processing capability [Resnick and Varian, 1997; Burke, 2002]. Typically, such systems (1) use information about a user to predict ratings of items that the user has not yet considered, and (2) recommend suitable items based on these predictions. *Collaborative* modelling techniques constitute one of the main model classes applied in RS [Albrecht and Zukerman, 2007]. They base their predictions upon the assumption that users who have agreed in their behaviour in the past will agree in the future.

The greatest strength of collaborative approaches is that they are independent of any representation of the items being recommended, and work well for complex objects, for which features are not readily apparent. The two main collaborative approaches are *memory-based* and *model-based*. Previous research has mainly focused on memory-based approaches, such as nearest-neighbour models (classic collaborative filtering), e. g., [Herlocker *et al.*, 1999], due to their intuitiveness. The main drawback of memory-based algorithms is that they operate over the entire user database to make predictions. In contrast, model-based approaches use techniques such as Bayesian networks, latent-factor models and artificial neural networks, e. g., [Breese *et al.*, 1998; Bell *et al.*, 2007], to first learn a statistical model in an offline fashion, and then use it to make predictions and generate recommendations. This decomposition can significantly speed up the recommendation generation process.

Spatial processes (*random fields*) are a subclass of *stochastic processes* which are applied to domains that have a geospatial interpretation, e. g., [Diggle *et al.*, 1998; Banerjee *et al.*, 2004]. Typical tasks of the field of spatial statistics include modelling spatial associations between a set of observations made at certain locations, and predicting values at locations where no observations have been made.

This paper applies theory from the area of spatial statistics to the prediction of a user's interests or item ratings in RS. We develop a *Spatial Process Model (SPM)* by adapting a Gaussian spatial process model to the RS scenario, and demonstrate our model's applicability to the task of predicting implicit ratings in the museum domain. The use of spatial processes requires a measure of distance between items in addition to user ratings. This measure, which is non-specific (e. g., it may be a physical or a conceptual distance), can be readily obtained in most cases.

Our application scenario is motivated by the need to automatically recommend exhibits to museum visitors, based on non-intrusive observations of their actions in the physical space. Employing RS in this scenario is challenging due to (1) the physicality of the domain, (2) having exhibit viewing times rather than explicit ratings, and (3) predictions differing from recommendations (we do not want to recommend exhibits that visitors are going to see anyway). We turn the first challenge into an advantage by exploiting the fact that physical distances between exhibits are meaningful, enabling the use of walking distance between exhibits to calculate (content) distance. This supports the direct, interpretable application of spatial processes by using a simple parametric Gaussian spatial process model (with the ensuing low variance in parameter estimates), compared to more complex non-parametric approaches, e. g., [Schwaighofer *et al.*, 2005]. The second challenge, which stems from the variable semantics of viewing times (time t for different exhibits could mean interest or boredom), is naturally addressed by *SPM*'s structure. The third challenge can be addressed by (a) using *SPM* to build a model of a visitor's interests in exhibits, (b) inferring a predictive model of a visitor's pathway through the museum [Bohnert *et al.*, 2008], and (c) combining these models to recommend exhibits of interest that may be overlooked if the predicted pathway is followed.

Our approach offers advantages over other model-based approaches in that, unlike neural networks (and memory-based techniques), it returns the confidence in a prediction, and its parameters have a clear interpretation; unlike Bayesian networks, our model does not require a domain-specific adaptation, such as designing the network topology. In addition, the distance measure endows our model with capabilities of hybrid RS [Burke, 2002; Albrecht and Zukerman, 2007] by seamlessly supporting the incorporation of other types of models (e. g., content-based). The distance measure also alleviates the *cold-start problem*. The *new-item problem* is addressed by utilising the (distance-based) correlation between this item and the other items. The *new-user problem* is similarly handled through the correlation between items rated by a user and the other items (our model can make useful personalised predictions after only one item has been rated).

SPM was evaluated with a real-world dataset of time spans spent by museum visitors at exhibits (viewed as implicit ratings). We compared our model's performance to that of (1) a baseline model which delivers a non-personalised prediction, and (2) a state-of-the-art nearest-neighbour collaborative filter incorporating performance-enhancing modifications, e. g., [James and Stein, 1961; Herlocker *et al.*, 1999]. Our results show that *SPM* significantly outperforms both models.

The paper is organised as follows. Our spatial processes approach for modelling and predicting item ratings is described in Section 2. In Section 3, we present the results of our evaluation, followed by our conclusions in Section 4.

2 Using Spatial Processes to Model and Predict Ratings in Recommender Systems

We briefly introduce stationary spatial process models (Section 2.1), before adapting a model based on Gaussian spatial processes to RS (Section 2.2). In Section 2.3, we describe an MCMC-based Bayesian approach for estimating the parameters of the model, and in Section 2.4 we outline the theory for predicting item ratings.

2.1 Stationary Spatial Process Models

Let $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))$ be a vector of observations at n sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$. Assuming stationarity both in the mean and variance, we can define the following basic model to capture spatial associations:

$$Y(\boldsymbol{s}_i) = \mu + \sigma w(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i) \text{ for all } i = 1, \ldots, n,$$

where $w(\boldsymbol{s}_i)$ are assumed to be realisations from a stationary Gaussian spatial process $W(\boldsymbol{s})$ with mean 0, variance 1, and isotropic[1] correlation function $\rho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|; \phi, \nu)$ capturing residual spatial association; and $\varepsilon(\boldsymbol{s}_i)$ are realisations from a white-noise process with mean 0 and variance τ^2, i. e., non-spatial uncorrelated error terms. That is, we assume that $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))$ are observations from a stationary Gaussian spatial process over $\boldsymbol{s}$ with mean μ, variance $\sigma^2 + \tau^2$, and correlation function $\rho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|; \phi, \nu)$. Generally, correlation is assumed to approach 0 with increasing distance $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$. Common choices for isotropic correlation functions are the powered exponential

$$\rho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|; \phi, \nu) = \exp\left(-\left(\phi\|\boldsymbol{s}_i - \boldsymbol{s}_j\|\right)^{\nu}\right),$$

where $\phi > 0$ and $0 < \nu < 2$, or correlation functions from the Matérn class [Banerjee *et al.*, 2004].[2]

The model provides a generic framework for modelling spatial associations between locations. Given estimates for μ, σ^2 and τ^2, it can make predictions for new locations. This model can be extended to observations $\boldsymbol{Y}$ that would not naturally be modelled using a Gaussian distribution, e. g., discrete binary variables indicating like/dislike [Diggle *et al.*, 1998].

2.2 Adaptation for Recommender Systems

RS help users find interesting information in a space of many options. Typically, RS identify items that suit the needs of a particular user given some evidence about his/her preferences. A collaborative filter, for example, utilises a set of ratings $\boldsymbol{Y}$ of users $U = \{u : u = 1, \ldots, m\}$ regarding a set of items $I = \{i : i = 1, \ldots, n\}$ to identify users who are similar to the current user. It then predicts this user's ratings on the basis of the ratings of the most similar users. Ratings of different users are usually considered to be independent, whereas ratings of related items tend to be correlated. Introducing a notion of spatial distance between items in order to functionally specify this correlation structure, we can use spatial process models (Section 2.1) for the prediction task, in a fashion similar to the Gaussian process model for preference prediction described by Schwaighofer *et al.* [2005]. The assumption made for spatial processes, that correlation between observations increases with decreasing site distance, fits well with RS, where ratings are usually more correlated the closer (i. e., more related) items are. As for the previous section, we use $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ to denote the locations of items $i, j = 1, \ldots, n$ in a space providing such a distance measure, i. e., $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$. For example, $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ could be computed from feature vectors representing the items (similarly to content-based RS), from item-to-item similarities (similarly to item-to-item collaborative filtering [Sarwar *et al.*, 2001]), or from physical distance, as done in this paper.

The following changes are necessary to adapt the model presented in the previous section to RS:

- **Multiple ratings.** Given a set of users U, we can have multiple (but independent) ratings for a given item i.
- **Non-stationarity.** Different items can have different rating means and variances. Hence, the underlying process cannot be assumed to be stationary in its mean and variance, as both μ and σ^2 depend on an item's location $\boldsymbol{s}$. We use the notation $\mu(\boldsymbol{s})$ and $\sigma^2(\boldsymbol{s})$ to indicate this.
- **Item set finiteness.** In contrast to traditional geospatial modelling, we require predictions only for a finite set of items, i. e., those at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$. Hence, it is sufficient (and necessary) to know $\mu(\boldsymbol{s})$ and $\sigma^2(\boldsymbol{s})$ only at these locations. That is, we do not require a special (functional) structure for $\mu(\boldsymbol{s})$ or $\sigma^2(\boldsymbol{s})$, which is usually the case for geospatial models.[3]

[1] A correlation function is isotropic if it depends on the separation vector $\boldsymbol{s}_i - \boldsymbol{s}_j$ only through its length $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$.

[2] In our experiments (Section 3), we use a powered exponential correlation function, as the Matérn class yielded inferior results.

[3] In order to compute predictions for a new item i without ratings, $\mu(\boldsymbol{s}_i)$ and $\sigma^2(\boldsymbol{s}_i)$ must be externally estimated and supplied to our model, until they can be estimated from observed data.

For a single user with rating vector $\boldsymbol{Y}$ with $(\boldsymbol{Y})_i = Y(\boldsymbol{s}_i)$, we extend the basic model from Section 2.1 as follows. We set $Y(\boldsymbol{s}_i)$ to be observations from a **non-stationary** spatial process, i. e., $Y(\boldsymbol{s}_i) = \mu(\boldsymbol{s}_i) + \sigma(\boldsymbol{s}_i)w(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i)$. Exploiting **finiteness**, $\mu(\boldsymbol{s}_i)$ and $\sigma(\boldsymbol{s}_i)$ are respectively components of $\boldsymbol{\mu} = (\mu(\boldsymbol{s}_1), \ldots, \mu(\boldsymbol{s}_n))$, the vector of mean ratings for items $1, \ldots, n$, and $\boldsymbol{\sigma} = (\sigma(\boldsymbol{s}_1), \ldots, \sigma(\boldsymbol{s}_n))$, the vector of standard deviations. Let $\boldsymbol{\theta} = \left(\boldsymbol{\mu}, \boldsymbol{\sigma}, \tau^2, \phi, \nu\right)$ be a vector collecting all model parameters, and $\boldsymbol{W} = (w(\boldsymbol{s}_1), \ldots, w(\boldsymbol{s}_n))$. Then,

$$\boldsymbol{Y} \mid \boldsymbol{\theta}, \boldsymbol{W} \sim \mathcal{N}\left(\boldsymbol{\mu} + \boldsymbol{\sigma} \mathit{1}_n \boldsymbol{W}, \tau^2 \mathit{1}_n\right),$$

where $\mathit{1}_n$ is the identity matrix of dimension $n \times n$. Note that $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))$ are mutually independent given $\boldsymbol{\theta}$ and $\boldsymbol{W}$. Marginalising the model over W, we obtain[4]

$$\boldsymbol{Y} \mid \boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\sigma} \mathit{1}_n H(\phi, \nu) \boldsymbol{\sigma} \mathit{1}_n + \tau^2 \mathit{1}_n\right), \tag{1}$$

where $H(\phi, \nu)$ denotes a correlation matrix with components $(H(\phi, \nu))_{ij} = \rho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|; \phi, \nu)$. That is, $(H(\phi, \nu))_{ij}$ represents the correlation between the ratings for items i and j.

We are ready to generalise the model to the **multi-user** case. As above, we denote the set of users with U (cardinality m), and the set of items with I (cardinality n). Typically, for a user u in U, we have ratings for only a subset of I, say for n_u items in I. Denoting a rating by user u for item i with $Y_u(\boldsymbol{s}_i)$ and a user's rating vector with $\boldsymbol{Y}_u$, we collect all observed ratings into a vector $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m)$ of dimension $\sum_{u=1}^m n_u$. Similarly, we structure $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ such that $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m)$ and $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_m)$, where $\boldsymbol{\mu}_u$ and $\boldsymbol{\sigma}_u$ are the vectors of means and standard deviations for those items rated by a user u, respectively. For example, assume that $U = \{1, 2\}$ and $I = \{1, 2, 3\}$. If user 1 rated items 2 and 3, and user 2 rated items 1 and 2, then

$$\begin{aligned}
\boldsymbol{Y} &= (\boldsymbol{Y}_1, \boldsymbol{Y}_2) &&= (Y_1(\boldsymbol{s}_2), Y_1(\boldsymbol{s}_3); Y_2(\boldsymbol{s}_1), Y_2(\boldsymbol{s}_2)), \\
\boldsymbol{\mu} &= (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) &&= (\mu(\boldsymbol{s}_2), \mu(\boldsymbol{s}_3); \mu(\boldsymbol{s}_1), \mu(\boldsymbol{s}_2)), \\
\boldsymbol{\sigma} &= (\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) &&= (\sigma(\boldsymbol{s}_2), \sigma(\boldsymbol{s}_3); \sigma(\boldsymbol{s}_1), \sigma(\boldsymbol{s}_2)).
\end{aligned}$$

Similarly to Equation 1, $\boldsymbol{Y} \mid \boldsymbol{\theta}$ is multivariate normal of dimension $\sum_{u=1}^m n_u$, where $H(\phi, \nu)$ is block diagonal with diagonal elements $H_1(\phi, \nu), \ldots, H_m(\phi, \nu)$ (due to users $u = 1, \ldots, m$ being independent), and $H_u(\phi, \nu)$ denotes a user u's correlation matrix of dimension $n_u \times n_u$. That is, for all users $u = 1, \ldots, m$,

$$\boldsymbol{Y}_u \mid \boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\mu}_u, \boldsymbol{\sigma}_u \mathit{1}_{n_u} H_u(\phi, \nu) \boldsymbol{\sigma}_u \mathit{1}_{n_u} + \tau^2 \mathit{1}_{n_u}\right). \tag{2}$$

Thus, given the model parameters $\boldsymbol{\theta} = \left(\boldsymbol{\mu}, \boldsymbol{\sigma}, \tau^2, \phi, \nu\right)$, our model is fully specified, with $\boldsymbol{\mu} = (\mu(\boldsymbol{s}_1), \ldots, \mu(\boldsymbol{s}_n))$ and $\boldsymbol{\sigma} = (\sigma(\boldsymbol{s}_1), \ldots, \sigma(\boldsymbol{s}_n))$.

2.3 Parameter Estimation

This section describes efficient algorithms for estimating the $2n+3$ model parameters $\boldsymbol{\theta} = \left(\boldsymbol{\mu}, \boldsymbol{\sigma}, \tau^2, \phi, \nu\right)$. The most popular parameter estimation strategies are maximum-likelihood and Bayesian inference. We opt for a Bayesian solution, as it offers some attractive advantages over the classic frequentist approach. For instance, prior knowledge can be formally incorporated into parameter estimation via the prior distribution $p(\boldsymbol{\theta})$, and the uncertainty about the parameters $\boldsymbol{\theta}$ is captured by the posterior distribution. Parameter estimates for $\boldsymbol{\theta}$ can be obtained from the posterior distribution

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{3}$$

where $p(\boldsymbol{Y}|\boldsymbol{\theta})$ is the likelihood of $\boldsymbol{Y}$ given $\boldsymbol{\theta}$. Typically, independent priors are chosen for the different parameters, i. e., in our case, $p(\boldsymbol{\theta}) = p(\boldsymbol{\mu})p(\boldsymbol{\sigma})p(\tau^2)p(\phi)p(\nu)$.

Algorithm 1 Slice Gibbs sampling algorithm

Initialise $\boldsymbol{\theta}$, e. g., by drawing $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta})$.
repeat
  Updating of auxiliary variable $V \mid \boldsymbol{\theta}, \boldsymbol{Y}$.
  Draw $Z \sim \text{Exp}(1)$, and set $V = l(\boldsymbol{\theta}; \boldsymbol{Y}) + Z$.
  Component-wise updating of $\boldsymbol{\theta} \mid V, \boldsymbol{Y}$.
  for $k = 1, \ldots, |\boldsymbol{\theta}|$ **do**
    repeat
      Draw the k-th component θ_k of $\boldsymbol{\theta}$ from $p(\theta_k)$, using shrinkage sampling to truncate the domain of $p(\theta_k)$ after each iteration.
    until $l(\boldsymbol{\theta}; \boldsymbol{Y}) < V$.
  end for
  Keep acquired sample of $\boldsymbol{\theta}$.
until the number of MCMC samples of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\boldsymbol{Y})$ is sufficiently large.

The integrations required to calculate $p(\boldsymbol{\theta}|\boldsymbol{Y})$ in Equation 3 are generally not tractable in closed form. However, $p(\boldsymbol{\theta}|\boldsymbol{Y})$ can be approximated numerically using *Markov chain Monte Carlo (MCMC)* integration methods, such as the *Metropolis-Hastings algorithm* and the *Gibbs sampler*.[5] Following Banerjee *et al.* [2004], we use *slice sampling* [Neal, 2003], i. e., a *slice Gibbs sampler*, to sample from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{Y})$. This approach is favoured by Banerjee *et al.*, because it does not require tuning that is tailored to the application, and hence provides an automatic MCMC algorithm for fitting Gaussian spatial process models. Algorithm 1 summarises our sampling procedure. The algorithm consists of two iteratively applied steps: (1) slicing the likelihood, and (2) performing Gibbs updates using draws from the prior along with rejection sampling. When updating a component θ_k of $\boldsymbol{\theta}$, we use *shrinkage sampling* to reduce the number of draws required before a point in the slice is found [Neal, 2003].

For our model (Equation 2), the marginal negative log-likelihood $l(\boldsymbol{\theta}; \boldsymbol{Y}) = -\log p(\boldsymbol{Y}|\boldsymbol{\theta})$ associated with $\boldsymbol{Y}$ is

$$\begin{aligned}
l(\boldsymbol{\theta}; \boldsymbol{Y}) = {} & \tfrac{1}{2} \sum\nolimits_{u=1}^m \log |\Sigma_u| \\
& + \tfrac{1}{2} \sum\nolimits_{u=1}^m (\boldsymbol{Y}_u - \boldsymbol{\mu}_u)^T \Sigma_u^{-1} (\boldsymbol{Y}_u - \boldsymbol{\mu}_u) + C,
\end{aligned}$$

where $\Sigma_u = \Sigma_u(\boldsymbol{\sigma}_u, \tau^2, \phi, \nu) = \boldsymbol{\sigma}_u \mathit{1}_{n_u} H_u(\phi, \nu) \boldsymbol{\sigma}_u \mathit{1}_{n_u} + \tau^2 \mathit{1}_{n_u}$, and $C \equiv \mathit{const}.$ independent of $\boldsymbol{\theta}$. Given $\boldsymbol{\sigma}_u$, ϕ and

[4]Marginalisation over W is possible only in the Gaussian case, not in the more general case described in [Diggle *et al.*, 1998].

[5]Refer to [Andrieu *et al.*, 2003] for an excellent introduction to MCMC for machine learning.

ν, computing the eigen decomposition of positive-definite $\boldsymbol{\sigma}_u \mathit{1}_{n_u} H_u(\phi, \nu) \boldsymbol{\sigma}_u \mathit{1}_{n_u}$ simplifies the calculation of $\log |\Sigma_u|$ and Σ_u^{-1}. It also speeds up the sampling procedure when updating τ^2 at a given iteration of the slice Gibbs sampling algorithm. To minimise the number of eigen decompositions, we update ϕ and ν together. We proceed similarly for the components of $\boldsymbol{\sigma}_u$, and hence for the components of $\boldsymbol{\sigma}$.

In the following section, we set $\sigma(\boldsymbol{s}_i) = \sqrt{\sigma^2_{\boldsymbol{Y}}(\boldsymbol{s}_i) - \tau^2}$, where $\sigma^2_{\boldsymbol{Y}}(\boldsymbol{s}_i)$ denotes the sample variance of the ratings for item i (at site $\boldsymbol{s}_i$), calculated from the ratings $Y_u(\boldsymbol{s}_i)$. This reduces the number of free model parameters from $2n + 3$ to $n + 3$, i. e., $\boldsymbol{\theta} = \left(\boldsymbol{\mu}, \tau^2, \phi, \nu\right)$, and significantly speeds up the slice Gibbs sampler.

2.4 Prediction

Given $\boldsymbol{\theta}$, the prediction of a user u's ratings of unseen items, say $\boldsymbol{Y}_{u,1}$, from a vector of observed ratings $\boldsymbol{Y}_{u,2}$ is straightforward. That is, we can use standard multivariate normal theory, because $\boldsymbol{Y}_u = (\boldsymbol{Y}_{u,1}, \boldsymbol{Y}_{u,2}) \mid \boldsymbol{\theta}$ is normally distributed (Section 2.2, similarly to Equation 1). If we use the following notation

$$\begin{bmatrix} \boldsymbol{Y}_{u,1} \\ \boldsymbol{Y}_{u,2} \end{bmatrix} \mid \boldsymbol{\theta} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{u,1} \\ \boldsymbol{\mu}_{u,2} \end{bmatrix}, \begin{bmatrix} \Sigma_{u,11} & \Sigma_{u,12} \\ \Sigma^T_{u,12} & \Sigma_{u,22} \end{bmatrix} \right),$$

then the conditional distribution $p\left(\boldsymbol{Y}_{u,1} | \boldsymbol{Y}_{u,2}, \boldsymbol{\theta}\right)$ is normal with mean vector and covariance matrix

$$\begin{aligned} \mathbb{E}\left(\boldsymbol{Y}_{u,1} | \boldsymbol{Y}_{u,2}, \boldsymbol{\theta}\right) &= \boldsymbol{\mu}_{u,1} + \Sigma_{u,12} \Sigma^{-1}_{u,22} \left(\boldsymbol{Y}_{u,2} - \boldsymbol{\mu}_{u,2}\right), \\ \mathrm{Cov}\left(\boldsymbol{Y}_{u,1} | \boldsymbol{Y}_{u,2}, \boldsymbol{\theta}\right) &= \Sigma_{u,11} - \Sigma_{u,12} \Sigma^{-1}_{u,22} \Sigma^T_{u,12}. \end{aligned}$$

The expectation $\mathbb{E}\left(\boldsymbol{Y}_{u,1} | \boldsymbol{Y}_{u,2}, \boldsymbol{\theta}\right)$ represents a personalised prediction of ratings $\boldsymbol{Y}_{u,1}$, and a measure of confidence can be easily derived from $\mathrm{Cov}\left(\boldsymbol{Y}_{u,1} | \boldsymbol{Y}_{u,2}, \boldsymbol{\theta}\right)$.

3 Evaluation with a Real-World Dataset

This section reports on the results of an evaluation performed with a real-world dataset from the museum domain, including comparison with a state-of-the-art collaborative filter.

In an information-seeking context, viewing time correlates positively with preference and interest. This observation was used in [Bohnert *et al.*, 2008] to propose a formulation of visitors' interests based on viewing times of exhibits. In this paper, we evaluate the predictive accuracy of our model on the basis of its predictions of viewing times.

3.1 Dataset and Model Justification

We obtained the dataset by manually tracking visitors to Melbourne Museum (Melbourne, Australia) from April to June 2008. In general, visitors do not require recommendations to travel between individual, logically related exhibits in close physical proximity. Hence, with the help of museum staff, we grouped the collection exhibited at Melbourne Museum, which comprises a few thousand exhibits, into 126 coherent exhibit areas. We tracked first-time adult visitors travelling on their own, to ensure that neither prior knowledge about the museum nor other visitors' interests influenced a visitor's decisions about what to look at. The resulting dataset comprises 158 complete visitor pathways in the form of time-annotated sequences of visited exhibit areas, with a total visit length of 291:22:37 hours, and a total viewing time of 240:00:28 hours.[6] A total of 8327 exhibit areas were viewed, yielding 52.7 areas per visitor on average (41.8% of the exhibit areas). Hence, 58.2% of the entries are missing from the viewing time (rating) matrix, indicating a potential for pointing a visitor to personally relevant but unvisited exhibit areas. Table 1 summarises further statistics of the dataset.

Table 1: Museum dataset statistics

	Mean	**Stddev**	**Min**	**Max**
Visit length (hrs)	1:50:39	0:47:54	0:28:23	4:42:12
Viewing time (hrs)	1:31:09	0:42:05	0:14:09	4:08:27
Exhibit areas / visitor	52.70	20.69	16	103
Visitors / exhibit area	66.09	25.36	6	117

Clearly, the deployment of an automated RS in a museum requires suitable positioning technologies to non-intrusively track visitors, and models to infer which exhibits are being viewed. Although our dataset was obtained manually, it provides information that is of the same type as information inferable from sensing data. Additionally, the results obtained from experiments with this dataset are essential for model development, as they provide an upper bound for the predictive performance of our model.

The museum space is carefully themed by curatorial staff, such that closely-related exhibits are in physical proximity. Based on this observation, we hypothesise that physical walking distance between exhibits is inversely proportional to their (content) similarity. Thus, in our experiments, we use physical walking distance for measuring (content) distance between exhibits. To calculate walking distances, we employed an SVG file-based representation of Melbourne Museum's site map, mapped onto a graph structure which preserves the physical layout of the museum (i. e., preventing paths from passing through walls or ceilings), and normalised the distances to the interval $[0, 1]$.

We used the *Bayesian Information Criterion (BIC)* to select the most appropriate family of probability distributions for approximating the distribution of viewing times at each exhibit area. We tested exponential, gamma, normal, $\log$-normal and Weibull distributions. The $\log$-normal family fitted best, with respect to both number of best fits and average BIC score (averaged over all exhibit areas). Hence, we transformed all viewing times to their $\log$-equivalent to obtain normally distributed data.

3.2 Parameter Estimation

We performed slice Gibbs sampling (Algorithm 1) to obtain estimates for the parameters $\boldsymbol{\theta}$ of our *Spatial Process Model (SPM)* (Section 2). For each of the 129 free model parameters,[7] we used (uninformative) independent uniform

[6] For our experiments, we ignore travel time between exhibit areas, and collapse multiple viewing events of one area into one event.

[7] We set $\sigma(\boldsymbol{s}_i) = \sqrt{\sigma^2_{\boldsymbol{Y}}(\boldsymbol{s}_i) - \tau^2}$ to speed up the sampling process (Section 2.3), which reduces the number of free parameters from 255 to 129.

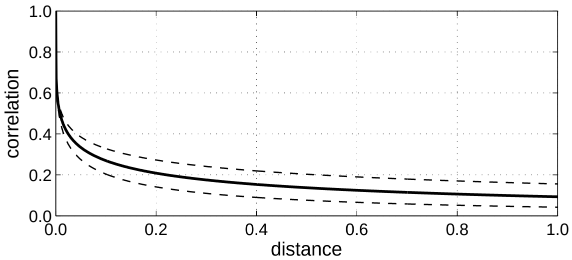

Figure 1: $\rho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|; \phi = 28.6334, \nu = 0.2583)$

prior distributions. We used every 20-th sample after a burn-in phase of 1000 iterations as a sample of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\boldsymbol{Y})$, and stopped the sampling procedure after 8000 iterations. Thus, in total, this provides 350 samples of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\boldsymbol{Y})$.

For visualisation purposes, we ran a slice Gibbs sampler on the complete dataset of 8327 $\log$ viewing times, and used the output of this run to obtain posterior mean point estimates of ϕ (28.6334), ν (0.2583) and τ^2 (0.1578); we omit the estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ due to space limitations. Figure 1 depicts a plot of $\rho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|; \phi, \nu)$ for this parameterisation of ϕ and μ, showing the shape of the fitted powered exponential correlation function. The dashed lines indicate the correlation functions obtained with ϕ and ν set to the lower and upper bounds of their 95% credible intervals respectively. The correlation function rapidly drops to values around 0.4, and then more slowly approaches 0.1 as $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ approaches 1.0. The shape of this function confirms the existence of a relatively high correlation between viewing durations at exhibit areas in close physical proximity.

3.3 Predictive Model Performance

Experimental Setup

To evaluate *SPM*'s predictive performance, we implemented two additional models: *Mean Model (MM)* and *Collaborative Filter (CF)*. *MM*, which we use as a baseline, predicts the $\log$ viewing time of an exhibit area i to be its (non-personalised) mean $\log$ viewing time $\mu(\boldsymbol{s}_i)$. For *CF*, we implemented a nearest-neighbour collaborative filtering algorithm, and added modifications from the literature that improve its performance, such as shrinkage to the mean [James and Stein, 1961] and significance weighting [Herlocker *et al.*, 1999]. Additionally, to ensure that varying exhibit area complexity does not affect the similarity computation for selecting the nearest neighbours (viewing time increases with exhibit complexity), we transformed the $\log$ viewing times into z-scores by normalising the values for each of the exhibit areas separately. Visitor-to-visitor differences with respect to their mean viewing durations were neutralised by transforming predictions to the current visitor's viewing-time scale [Herlocker *et al.*, 1999]. We tested several thousand different parameterisations, but in this paper, we report only on the performance of the best one.

Due to the relatively small dataset, we used leave-one-out cross validation to evaluate the performance of the different models. That is, for each visitor, we trained the models with a reduced dataset containing the data of 157 of the 158 visit trajectories, and used the withheld visitor pathway for testing. For *SPM*, we first obtained the posterior mean of $p(\boldsymbol{\theta}|\boldsymbol{Y})$ by performing slice Gibbs sampling on the training data (Section 3.2), and then used this posterior mean to compute predictions by conditioning a multivariate normal distribution (Section 2.4). For *CF*, predictions were computed from the ratings of the nearest neighbours; and for *MM*, we used $\mu(\boldsymbol{s}_i)$, estimated from the appropriate reduced dataset, as a prediction.

We performed two types of experiments: *Individual Exhibit (IE)* and *Progressive Visit (PV)*.

- *IE* evaluates predictive performance for a single exhibit. For each observed visitor-exhibit area pair (u, i), we removed the observation $Y_u(\boldsymbol{s}_i)$ from the vector of visitor u's $\log$ viewing durations, and computed a prediction $\hat{Y}_u(\boldsymbol{s}_i)$ from the other observations. This experiment is lenient in the sense that all available observations except the observation for exhibit area i are kept in a visitor's viewing duration vector.
- *PV* evaluates performance as a museum visit progresses, i. e., as the number of viewed exhibit areas increases. For each visitor, we started with an empty visit, and iteratively added each viewed exhibit area to the visit history, together with its $\log$ viewing time. We then predicted the $\log$ viewing times of all yet unvisited exhibit areas.

For both experiments, we used the *mean absolute error (*MAE*)* to measure predictive accuracy as follows:

$$\mathrm{MAE} = \frac{1}{\sum_{u \in U} |I_u|} \sum_{u \in U} \sum_{i \in I_u} |Y_u(\boldsymbol{s}_i) - \hat{Y}_u(\boldsymbol{s}_i)|,$$

where I_u denotes a visitor u's set of exhibit areas for which predictions were computed. For *IE*, we calculated the total MAE for all valid visitor-exhibit area pairs; and for *PV*, we computed the MAE across the yet unvisited exhibit areas and all visitors for each time fraction of a visit (to account for different visit lengths, we normalised all visits to a length of 1).

Results

Table 2 shows the results for the *IE* experiment, where *SPM* achieves an MAE of 0.7548 (stderr 0.0066), outperforming both *MM* and *CF*. The performance difference between *SPM* and the other models is statistically significant with $p \ll 0.01$.

The performance of *SPM*, *CF* and the baseline *MM* for the *PV* experiment is depicted in Figure 2. *CF* outperforms *MM* slightly (statistically significantly for visit fractions 0.191 to 0.374 and for several shorter intervals later on, $p < 0.05$). More importantly, *SPM* performs significantly better than both *MM* and *CF* (statistically significantly for visit fractions 0.019 to 0.922, $p < 0.05$). Drawing attention to the initial portion of the visits, *SPM*'s MAE decreases rapidly, whereas the MAE for *MM* and *CF* remains at a higher level. Generally, the faster a model adapts to a visitor's interests, the more likely it is to quickly deliver (personally) useful recommendations. Such behaviour in the early stages of a museum visit is essential in order to build trust in the RS, and to guide a visitor in a phase of his/her visit where such guidance is most likely needed. As expected, *MM* performs at a relatively constant MAE level. For *CF* and *SPM*, we expected to see an improvement in performance (relative to *MM*) as the

Table 2: Model performance for the *IE* experiment (MAE)

	MAE	Stderr
Mean Model (MM)	0.8618	0.0071
Collaborative Filter (CF)	0.7868	0.0068
Spatial Process Model (SPM)	0.7548	0.0066

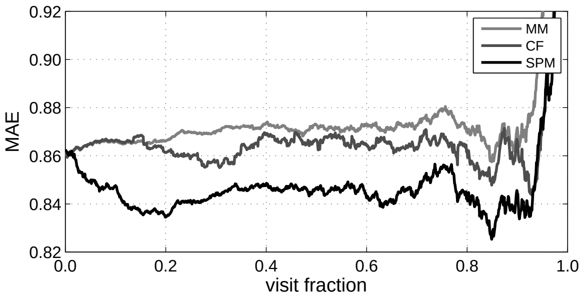

Figure 2: Model performance for the *PV* experiment (MAE)

number of visited exhibit areas increases. However, this trend is rather subtle (it can be observed when plotting the models' performance relative to *MM*). Additionally, for all three models, there is a performance drop towards the end of a visit. We postulate that these phenomena may be explained, at least partially, by the increased influence of outliers on the MAE as the number of exhibit areas remaining to be viewed is reduced with the progression of a visit. This influence in turn offsets potential gains in performance obtained from additional observations. Our hypothesis is supported by a widening in the standard error bands for all models as a visit progresses, in particular towards the end (not shown in Figure 2 for clarity of presentation). However, this behaviour requires further, more rigorous investigation.

4 Conclusions and Future Work

In this paper, we utilised the theory of spatial processes to develop a model-based approach for predicting users' interests or item ratings in RS. We applied our model to a real-world dataset from the museum domain, where our model attains a higher predictive accuracy than state-of-the-art nearest-neighbour collaborative filters. In addition, under the realistic *Progressive Visit* setting, our model rapidly adapts to a user's ratings (starting from as little as one rating), thus alleviating the *new-user problem* common to collaborative filtering.

Our dataset is relatively small compared to other real-world RS applications. Although a high number of ratings per user slows down the slice Gibbs sampler due to repeated inversion of matrices of high dimension, employing our model with larger datasets should not represent a problem in practice. This is because the number of ratings per user is usually small compared to the number of users and items, and the computational complexity of evaluating the likelihood function depends only linearly on the number of users in the database.

In the future, we intend to hybridise our model by incorporating content-based item features into our distance measure. We also plan to extend our model to fit non-Gaussian item ratings, e. g., [Diggle *et al.*, 1998; Yu *et al.*, 2006], and consider negative correlations between items.

Acknowledgments

This research was supported in part by grant DP0770931 from the Australian Research Council. The authors thank Carolyn Meehan and her team from Museum Victoria for their assistance; and David Abramson, Jeff Tan and Blair Bethwaite for their help with the computer cluster.

References

[Albrecht and Zukerman, 2007] D.W. Albrecht and I. Zukerman. Introduction to the special issue on statistical and probabilistic methods for user modeling. *UMUAI*, 17(1-2):1–4, 2007.

[Andrieu *et al.*, 2003] C. Andrieu, N. de Freitas, A. Doucet, and M.I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1):5–43, 2003.

[Banerjee *et al.*, 2004] S. Banerjee, B.P. Carlin, and A.E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, 2004.

[Bell *et al.*, 2007] R. Bell, Y. Koren, and C. Volinsky. Chasing $1,000,000: How we won the Netflix progress prize. *ASA Statistical and Computing Graphics Newsletter*, 18(2):4–12, 2007.

[Bohnert *et al.*, 2008] F. Bohnert, I. Zukerman, S. Berkovsky, T. Baldwin, and L. Sonenberg. Using interest and transition models to predict visitor locations in museums. *AI Communications*, 21(2-3):195–202, 2008.

[Breese *et al.*, 1998] J.S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *UAI-98 Proc.*, pages 42–52, 1998.

[Burke, 2002] R. Burke. Hybrid recommender systems: Survey and experiments. *UMUAI*, 12(4):331–370, 2002.

[Diggle *et al.*, 1998] P.J. Diggle, J.A. Tawn, and R.A. Moyeed. Model-based geostatistics. *Applied Statistics*, 47(3):299–350, 1998.

[Herlocker *et al.*, 1999] J.L. Herlocker, J.A. Konstan, A. Borchers, and J.T. Riedl. An algorithmic framework for performing collaborative filtering. In *SIGIR-99 Proc.*, pages 230–237, 1999.

[James and Stein, 1961] W. James and C.M. Stein. Estimation with quadratic loss. In *Proc. of the Fourth Berkeley Symp. on Mathematical Statistics and Probability, Vol. 1*, pages 361–379, 1961.

[Neal, 2003] R.M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

[Resnick and Varian, 1997] P. Resnick and H.R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[Sarwar *et al.*, 2001] B. Sarwar, G. Karypis, J.A. Konstan, and J.T. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW-01 Proc.*, pages 285–295, 2001.

[Schwaighofer *et al.*, 2005] A. Schwaighofer, V. Tresp, and K. Yu. Learning Gaussian process kernels via hierarchical Bayes. In *NIPS-04 Proc.*, pages 1209–1216. 2005.

[Yu *et al.*, 2006] S. Yu, K. Yu, V. Tresp, and H.-P. Kriegel. Collaborative ordinal regression. In *ICML-06 Proc.*, pages 1089–1096, 2006.