UCL Université catholique de Louvain

INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

ISBA

DISCUSSION PAPER

2018/05

Bayesian Inference For Bivariate Ranks

Guillotte, S., Perron, F. and J. Segers

BAYESIAN INFERENCE FOR BIVARIATE RANKS

SIMON GUILLOTTE, FRANÇOIS PERRON, AND JOHAN SEGERS

Abstract. A recommender system based on ranks is proposed, where an expert's ranking of a set of objects and a user's ranking of a subset of those objects are combined to make a prediction of the user's ranking of all objects. The rankings are assumed to be induced by latent continuous variables corresponding to the grades assigned by the expert and the user to the objects. The dependence between the expert and user grades is modelled by a copula in some parametric family. Given a prior distribution on the copula parameter, the user's complete ranking is predicted by the mode of the posterior predictive distribution of the user's complete ranking conditional on the expert's complete and the user's incomplete rankings. Various Markov chain Monte-Carlo algorithms are proposed to approximate the predictive distribution or only its mode. The predictive distribution can be obtained exactly for the Farlie–Gumbel–Morgenstern copula family, providing a benchmark for the approximation accuracy of the algorithms. The method is applied to the MovieLens 100k dataset with a Gaussian copula modelling dependence between the expert's and user's grades.

Key words. Bayes; Compatible ranking; Copula; Incomplete ranking; Markov chain Monte Carlo; Predictive distribution; Rank likelihood; Recommender systems; Simulated annealing.

1. Introduction

Recommender systems are part of many online businesses such as Amazon, e-Bay, Netflix, and others. They are tools to learn the interests of customers in order to make customer-specific recommendations of other products. These systems provide successful and valuable marketing strategies, especially due to the expansion of the world wide web and e-commerce. In 2006, Netflix organized the *Netflix-Prize*, awarding one million dollars for the best algorithm. The prize was won in 2009 by a team of researchers called *Bellkors Pragmatic Chaos* (AT&T Labs) after over three years of competition. The problem has attracted attention in the statistical community, with statisticians working on similar problems, see for instance Feuerverger, He & Khatri (2012), Fligner & Verducci (1986), Sun, Lebanon & Kidwell (2012), Zhu (2014), and the references therein.

We consider a version of recommender systems where an expert opinion ranking is available and is used, together with a partial ranking by a costumer, in order to predict that customer's complete ranking. Essentially, we want to predict an individual's ranking of a set of $n \geqslant 1$ different objects, given an expert opinion ranking of the same objects. More precisely, a set of objects indexed by $\mathscr{N} = \{1, \ldots, n\}$ is to be evaluated and ranked by both, an expert and an individual. The expert ranks all the objects, while the individual ranks only the subset of objects corresponding to the indices in the set $\mathscr{M} = \{i_1, \ldots, i_m\} \subset \mathscr{N}$. This can happen for instance if the individual does not have knowledge yet of the objects with indices in $\mathscr{N} \setminus \mathscr{M}$.

Assume ties are impossible. Let $\mathscr{S}_k$ be the permutation group on the set $\{1, \ldots, k\}$. The experiment provides a complete expert's ranking $r_x = (r_x(1), \ldots, r_x(n)) \in \mathscr{S}_n$ as well as an incomplete user's ranking $r_y^* = (r_y^*(1), \ldots, r_y^*(m)) \in \mathscr{S}_m$, where $r_y^*(j)$ is the user's rank of object i_j among the m objects $i_1, \ldots, i_m$. The choice for the subscripts x and y is clarified by the following. We think of the ranks as being induced by ratings or grades measured on a continuous

Date: February 12, 2018.

scale: if $x_1, \ldots, x_n$ denote the expert's grades and if $y_{i_1}, \ldots, y_{i_m}$ denote the individual's grades, then $r_x = \mathrm{rank}(x_1, \ldots, x_n)$ and $r_y^* = \mathrm{rank}(y_{i_1}, \ldots, y_{i_m})$.

If the user had been able to grade all objects, the user's grades would have been $y_1, \ldots, y_n$, with corresponding ranking $r_y = \mathrm{rank}(y_1, \ldots, y_n)$. In view of this, the model is constructed by assuming an underlying set of latent pairs of grades $(x_1, y_1), \ldots, (x_n, y_n)$ of all n objects attributed by both the expert and the individual. Concretely, we let (x_i, y_i), $i = 1, \ldots, n$, be realizations of independent random vectors (X_i, Y_i), $i = 1, \ldots, n$, each of which is distributed according to the same bivariate distribution with continuous margins. The continuity assumption makes the marginal distributions of the grades irrelevant to the rankings. We assume that the copula of the joint distribution belongs to some parametric family indexed by a parameter $\theta \in \Theta$ for which we select a prior. Let $R_X = \mathrm{rank}(X_1, \ldots, X_n)$ and $R_Y = \mathrm{rank}(Y_1, \ldots, Y_n)$ denote the random expert and individual rankings, respectively, of the objects in $\mathscr{N}$, and let $R_Y^* = \mathrm{rank}(Y_{i_1}, \ldots, Y_{i_m})$ denote the random user ranking of the objects in $\mathscr{M}$. Note that R_X and R_Y are random elements in $\mathscr{S}_n$, while R_Y^* is a random element in $\mathscr{S}_m$. The prediction of the user's complete ranking of all n objects is based on the mode of the posterior predictive distribution:

$$\hat{r}_y = \operatorname*{argmax}_{r_y} \mathrm{P}(R_Y = r_y \mid R_X = r_x, R_Y^* = r_y^*). \tag{1}$$

This predicted ranking is then used for instance to recommend new products to the customer.

One difficulty here is the evaluation of the joint probability mass function of the pair of rankings (R_X, R_Y): given a parameter value $\theta \in \Theta$, we need to compute

$$\mathrm{P}_\theta(R_X = r_x,\, R_Y = r_y), \qquad r_x, r_y \in \mathscr{S}_n. \tag{2}$$

In most cases, this probability is not analytically tractable. In the literature, it has been referred to as the rank likelihood; see for instance Hoff (2007), Hoff, Niu & Wellner (2014) and Segers, van den Akker & Werker (2014). The continuity assumption on the marginal cumulative distribution functions makes the margins irrelevant to the evaluation of the probability in (2). The assumption of uniform margins implies that the probability (2) can be evaluated by means of an integral over $[0, 1]^{2n}$. Within a Bayesian approach, it also means that we need only put a prior on the copula parameter.

An objective of this work is to find a family of copulas for which a closed-form expression of the posterior predictive distribution in (2) is available. We will show that the Farlie–Gumbel–Morgenstern (FGM) family (Nelsen, 2006, p. 77) satisfies this requirement.

Since the range of dependence that can be modelled by the FGM family is rather restricted, it is natural to ask how to proceed for other parametric copula families, when no explicit formulas for (2) exist. We develop a stochastic algorithm to compute (2) and we assess its accuracy by comparing its output to the results obtained from the exact formulas available for the FGM family.

Another problem for computing the prediction in (1) is that the cardinality of the set of rankings $r_y \in \mathscr{S}_n$ that are compatible with the observed ranking $r_y^* \in \mathscr{S}_m$ is equal to $n!/m!$. This number will usually be so high that it is infeasible to find the maximum in (1) by computing the probabilities on the right-hand side on (1) for all possible r_y. We will instead propose a solution based on an ergodic Monte Carlo Markov chain with the correct limiting distribution. The algorithm is applied to predict user rankings in the *MovieLens 100k* dataset with a Gaussian copula modelling dependence between expert and user grades.

2. Rank likelihood

Let $\mathscr{S}_n$ be the permutation group of the set $\mathscr{N} = \{1, \ldots, n\}$. A permutation $\sigma \in \mathscr{S}_n$ is a bijection from $\mathscr{N}$ to itself; notation $\sigma = (\sigma(1), \ldots, \sigma(n))$. The group operation, denoted by $\circ$, is the usual composition of functions, that is, $\sigma \circ \tau(i) = \sigma(\tau(i))$ for σ and τ in $\mathscr{S}_n$ and $i \in \mathscr{N}$.

The group's identity element is the identity map, $e = (1, \dots, n)$. The inverse of a permutation $\sigma \in \mathscr{S}_n$ is denoted by σ^{-1} and satisfies $\sigma \circ \sigma^{-1} = e = \sigma^{-1} \circ \sigma$.

Let $\mathbb{D}_n = \{x \in \mathbb{R}^n \colon x_{(1)} < \cdots < x_{(n)}\}$ be the set of vectors in $\mathbb{R}^n$ having no ties. The rank vector or ranking $\mathrm{rank}(x) = r_x = (r_x(1), \dots, r_x(n))$ associated to $x \in \mathbb{D}_n$ is defined by

$$r_x(i) = \sum_{j \in \mathscr{N}} \mathbb{1}(x_i \leqslant x_j), \qquad i \in \mathscr{N}.$$

We have $r_x \in \mathscr{S}_n$ for all $x \in \mathbb{D}_n$. We also define $r_x = \mathrm{rank}(x) = e$ if $x \in \mathbb{R}^n \setminus \mathbb{D}_n$, ensuring that the map $\mathrm{rank} : \mathbb{R}^n \to \mathscr{S}_n$ is well-defined. A simple but useful property is that the rank map behaves well under composition with permutations: for $x \in \mathbb{D}_n$ and $\sigma \in \mathscr{S}_n$, we have

$$\mathrm{rank}(x_{\sigma(1)}, \dots, x_{\sigma(n)}) = \mathrm{rank}(x_1, \dots, x_n) \circ \sigma. \tag{3}$$

Given two grading vectors $x, y \in \mathbb{D}_n$, we want to investigate the alignment, or the lack thereof, of the associated rankings $r_x = \mathrm{rank}(x)$ and $r_y = \mathrm{rank}(y)$. We would like to know the rank, under y, of the object that was attributed rank $j \in \mathscr{N}$ under the grading x. The original index of this object is equal to $i = r_x^{-1}(j)$, and its rank under y is equal to $r_y(i) = r_y(r_x^{-1}(j))$. This leads us to the study of the permutation

$$s = r_y \circ r_x^{-1} \in \mathscr{S}_n. \tag{4}$$

If $s = e$, for instance, the gradings x and y are perfectly aligned and induce the same ranking of the n objects.

Recall from the introduction that the random pairs (X_i, Y_i), $i = 1, \dots, n$, represent the expert's together with the individual's gradings of n objects. Assume that (X_i, Y_i), $i = 1, \dots, n$, are independent and identically distributed (iid) random pairs with continuous margins. With probability one, there are no ties among the gradings. Consider the random rank vectors

$$R_X = \mathrm{rank}(X_1, \dots, X_n), \qquad\qquad R_Y = \mathrm{rank}(Y_1, \dots, Y_n).$$

As in (4), we want to express the ranking induced by the random grading vector $Y = (Y_1, \dots, Y_n)$ in terms of the one induced by the random grading vector $X = (X_1, \dots, X_n)$. This motivates the definition of the rank statistic

$$S(X, Y) = R_Y \circ R_X^{-1}.$$

The joint distribution of (R_X, R_Y) is determined by the distribution of $S(X, Y)$.

Lemma 2.1. *If (X_i, Y_i), $i = 1, \dots, n$, are iid random pairs with continuous margins, then*

$$\mathrm{P}(R_X = r_x,\, R_Y = r_y) = \frac{1}{n!}\, \mathrm{P}\{S(X, Y) = r_y \circ r_x^{-1}\}, \qquad r_x, r_y \in \mathscr{S}_n.$$

The proof of Lemma 2.1 and of the other results in this paper are deferred to the Appendix. Let H and F, G be the joint and marginal cumulative distribution functions, respectively, of the random pairs (X_i, Y_i), i.e.,

$$H(x, y) = \mathrm{P}(X_1 \leqslant x, Y_1 \leqslant y), \qquad F(x) = H(x, \infty), \qquad G(y) = H(\infty, y),$$

for $x, y \in \mathbb{R}$. By assumption, F and G are continuous. Let $U_i = F(X_i)$ and $V_i = G(Y_i)$ for $i = 1, \dots, n$. The random variables U_i and V_i are uniformly distributed on $(0, 1)$ and their joint cumulative distribution function is a copula,

$$C(u, v) = \mathrm{P}(U_1 \leqslant u,\, V_1 \leqslant v), \qquad (u, v) \in [0, 1]^2.$$

Sklar's Theorem says that H admits the representation

$$H(x, y) = C(F(x), G(y)), \qquad (x, y) \in \mathbb{R}^2. \tag{5}$$

With probability one, the rankings induced by the random vectors $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ are the same as the ones induced by the random vectors $U = (U_1, \ldots, U_n)$ and $(V_1, \ldots, V_n)$, respectively. Since the pairs (U_i, V_i), $i = 1, \ldots, n$, are iid with cumulative distribution function given by the copula C, it follows that the joint distribution of the rank vectors (R_X, R_Y) is determined by C. This is formalized by the next theorem. In view of Lemma 2.1, it suffices to study the distribution of $S(X, Y)$.

Theorem 2.2. *Let (X_i, Y_i), $i = 1, \ldots, n$, be iid random vectors with continuous margins and copula C. If the copula C has density c, then we have*

$$\mathrm{P}\{S(X,Y) = s\} = \frac{1}{n!}\,\mathrm{E}\left\{\prod_{i=1}^{n} c\Big(\textstyle\sum_{j=1}^{i} W_{1,j}, \sum_{j=1}^{s(i)} W_{2,j}\Big)\right\}, \qquad s \in \mathscr{S}_n, \tag{6}$$

where $(W_{\ell,1}, \ldots, W_{\ell,n+1}) = (W_{\ell,1}, \ldots, W_{\ell,n}, 1 - \sum_{j=1}^{n} W_{\ell,j})$, $\ell = 1, 2$, are iid according to the Dirichlet$(1, \ldots, 1)$ *distribution.*

Among other things, Theorem 2.2 shows the intuitively obvious property that the marginal distributions of the gradings do not affect the joint distribution of the rank vectors. We shall therefore assume that the marginal distributions of the expert's and user's grades are both uniform on $(0, 1)$. Then we have $X_i = U_i$ and $Y_i = V_i$, and the random pairs (X_i, Y_i), $i = 1, \ldots, n$, are independent and identically distributed according to the copula C. This assumption also allows us to unambiguously write $S = S(X, Y) = S(U, V)$.

When the copula function belongs to a parametric family $(C_\theta : \theta \in \Theta)$, the distribution of S depends on θ. The probability mass function of S, seen as a function of θ, i.e., the map $\theta \mapsto \mathrm{P}_\theta(S = s)$, for $s \in \mathscr{S}_n$, is sometimes referred to as the rank likelihood. The expression (6) is particularly helpful as it suggests a certain Monte Carlo algorithm to compute this rank likelihood; see Algorithm 5.2.

3. Predictive distribution of compatible rankings

3.1. **Compatible rankings.** In the predictive distribution (1), any candidate complete ranking $r_y \in \mathscr{S}_n$ by the user should be compatible with the observed ranking $r_y^* \in \mathscr{S}_m$ of the m objects graded by the user. The notion of compatible rankings has already appeared in the literature, see for instance Alvo & Yu (2014). Recall that $\mathscr{S}_n$ is the group of permutations of the set $\mathscr{N} = \{1, \ldots, n\}$.

An incomplete ranking of size m, with $m \in \{1, \ldots, n-1\}$, is a couple $(r^*, \mathscr{M})$ consisting of a permutation $r^* \in \mathscr{S}_m$ and a subset $\mathscr{M} = \{i_1, \ldots, i_m\} \subset \mathscr{N}$, with $1 \leqslant i_1 < \cdots < i_m \leqslant n$. For example, for $n = 4$, an incomplete ranking of size $m = 3$ is given by $r^* = (3, 1, 2)$ and $\mathscr{M} = \{1, 3, 4\}$. This incomplete ranking corresponds to the partial ranking $r = (3, -, 1, 2)$, the second object not (yet) being ranked among the three other ones.

The set $\mathscr{C}(r^*, \mathscr{M})$ of compatible rankings associated to the incomplete ranking $(r^*, \mathscr{M})$ is defined as the set of rankings $r \in \mathscr{S}_n$ of all n objects such that the ranking of the m objects in $\mathscr{M}$ induced by r is equal to r^*. Formally, we have

$$\begin{aligned}\mathscr{C}(r^*, \mathscr{M}) &= \{r \in \mathscr{S}_n : \mathrm{rank}(r(i_1), \ldots, r(i_m)) = r^*\}. \\ &= \{r \in \mathscr{S}_n : r(i_{\sigma(1)}) < r(i_{\sigma(2)}) < \cdots < r(i_{\sigma(m)}),\ \sigma = (r^*)^{-1}\}.\end{aligned}$$

For the example above, we have

$$\mathscr{C}(r^*, \mathscr{M}) = \{(3,4,1,2), (4,3,1,2), (4,2,1,3), (4,1,2,3)\}.$$

To select an $(r^*, \mathscr{M})$-compatible ranking r, it suffices to choose the ranks, $r(i) \in \mathscr{N}$, of the $n - m$ objects $i \in \mathscr{N} \setminus \mathscr{M}$. The ranks of the remaining m objects in $\mathscr{M}$ are then determined by the compatibility constraint. This shows that the cardinality of $\mathscr{C}(r^*, \mathscr{M})$ is equal to $n!/m!$.

In the original formulation of the problem, the permutations $r_x \in \mathscr{S}_n$ and $r_y \in \mathscr{S}_n$ represent the expert and the individual's complete rankings respectively. The set $\mathscr{M} = \{i_1, \dots, i_m\} \subset \mathscr{N}$ represents the indices of the m objects ranked by the individual, with $1 \leqslant i_1 < \cdots < i_m \leqslant n$ and $1 \leqslant m < n$. One observes the expert's complete ranking r_x and the user's ranking $r_y^* = (r_y(i_1), \dots, r_y(i_m)) \in \mathscr{S}_m$ of the objects in $\mathscr{M}$. Notice that if the expert would have ranked only the m objects in $\mathscr{M}$, then the expert's ranks would have been

$$r_x^* = \mathrm{rank}(r_x(i_1), \dots, r_x(i_m)) \in \mathscr{S}_m.$$

Further, consider the permutation

$$s^* = r_y^* \circ (r_x^*)^{-1} \in \mathscr{S}_m.$$

In words, $s^*(j)$ is the user's rank of the object that was given rank $j = 1, \dots, m$ by the expert, among the m objects graded by the user. If $s^* = e$, the identity permutation in $\mathscr{S}_m$, then the rankings by the user and the expert are perfectly aligned.

By using Lemma 2.1, it will be shown in (9) below that for the calculation of the posterior predictive distribution (1), we can simply work with the transformation s^* as our observed data instead of with r_x, $\mathscr{M}$, and r_y^*. However, we need to translate the compatibility constraint to the transformed rankings. This is the purpose of Lemma 3.1 below, which says that r_y is compatible with r_y^* for the objects in $\mathscr{M}$ if and only if $r_y \circ r_x^{-1}$ is compatible with s^* for the objects in the set $\mathscr{M}^*$ defined in (7). This statement can be further interpreted as if the n objects were lined up in the order of the expert's preference (and so $r_x = e \in \mathscr{S}_n$) and the user was to rank the m objects presented to him in that order: the result would be s^* and $\mathscr{M}^*$.

The order statistics of the expert ranks $r_x(i_1), \dots, r_x(i_m) \in \mathscr{N}$ of the m objects graded by the user are denoted by $1 \leqslant i_1^* < \cdots < i_m^* \leqslant n$. For the reason explained above, it will be convenient to consider the incomplete ranking $(s^*, \mathscr{M}^*)$ with

$$\mathscr{M}^* = \{i_1^*, \dots, i_m^*\} = \{r_x(i_1), \dots, r_x(i_m)\} \subset \mathscr{N}. \tag{7}$$

To apply Lemma 2.1, we would like to switch from r_y to $r_y \circ r_x^{-1}$. The following lemma says how this transformation affects the compatibility constraint.

Lemma 3.1. *We have*

$$r_y \in \mathscr{C}(r_y^*, \mathscr{M}) \iff r_y \circ r_x^{-1} \in \mathscr{C}(s^*, \mathscr{M}^*).$$

3.2. **Predictive distribution.** Let $(C_\theta : \theta \in \Theta)$ be parametric family of bivariate copulas and let $\pi(\theta)$, $\theta \in \Theta$, be a prior density on θ. Conditionally on θ, the random pairs (X_i, Y_i), $i = 1, \dots, n$, are iid with common distribution given by C_θ. We observe the complete expert ranking $R_X = r_x \in \mathscr{S}_n$ as well as the partial user ranking $R_Y^* = r_y^* \in \mathscr{S}_m$ on $\mathscr{M} = \{i_1, \dots, i_m\} \subset \mathscr{N}$ as above. The posterior predictive distribution, or predictive distribution in short, of the complete user ranking R_Y given the data is

$$\mathrm{P}(R_Y = r_y \mid R_X = r_x,\, R_Y^* = r_y^*) = \frac{\int_\Theta \mathrm{P}_\theta(R_X = r_x,\, R_Y^* = r_y^*,\, R_Y = r_y)\, \pi(\theta)\, \mathrm{d}\theta}{\int_\Theta \mathrm{P}_\theta(R_X = r_x,\, R_Y^* = r_y^*)\, \pi(\theta)\, \mathrm{d}\theta},$$

for $r_y \in \mathscr{S}_n$. For the numerator, we use Lemma 2.1 to find that

$$\mathrm{P}_\theta(R_X = r_x,\, R_Y^* = r_y^*,\, R_Y = r_y) = \begin{cases} \mathrm{P}_\theta(R_X = r_x,\, R_Y = r_y) = \frac{1}{n!}\, \mathrm{P}_\theta(S = r_y \circ r_x^{-1}), & \text{if } r_y \in \mathscr{C}(r_y^*, \mathscr{M}), \\ 0 & \text{otherwise.} \end{cases}$$

Summing over all $r_y \in \mathscr{S}_n$, we find for the denominator that

$$\mathrm{P}_\theta(R_X = r_x,\, R_Y^* = r_y^*) = \sum_{r_y' \in \mathscr{C}(r_y^*, \mathscr{M})} \frac{1}{n!}\, \mathrm{P}_\theta(S = r_y' \circ r_x^{-1}).$$

The marginal distribution of S (marginal with respect to θ) is

$$\mathrm{P}(S = s) = \mathrm{E}_\pi\{\mathrm{P}_\theta(S = s)\} = \int \mathrm{P}_\theta(S = s)\, \pi(\theta)\, \mathrm{d}\theta, \qquad s \in \mathscr{S}_n. \tag{8}$$

As a consequence, the predictive distribution of R_Y given the data is

$$\mathrm{P}(R_Y = r_y \mid R_X = r_x,\, R_Y^* = r_y^*) = \mathbb{1}\{r_y \in \mathscr{C}(r_y^*, \mathscr{M})\}\, \frac{\mathrm{P}(S = r_y \circ r_x^{-1})}{\sum_{r_y' \in \mathscr{C}(r_y^*, \mathscr{M})} \mathrm{P}(S = r_y' \circ r_x^{-1})},$$

for $r_Y \in \mathscr{S}_n$. Write $r_y = s \circ r_x$ with $s = r_y \circ r_x^{-1}$ and note from Lemma 3.1 that $r_y \in \mathscr{C}(r_y^*, \mathscr{M})$ if and only if $s \in \mathscr{C}(s^*, \mathscr{M}^*)$. We find that the predictive distribution of R_Y given the data is

$$\begin{aligned} \mathrm{P}(R_Y = s \circ r_x \mid R_X = r_x,\, R_Y^* = r_y^*) &= \mathbb{1}\{s \in \mathscr{C}(s^*, \mathscr{M}^*)\}\, \frac{\mathrm{P}(S = s)}{\sum_{s' \in \mathscr{C}(s^*, \mathscr{M}^*)} \mathrm{P}(S = s')} \\ &= \mathrm{P}\{S = s \mid S \in \mathscr{C}(s^*, \mathscr{M}^*)\} =: p(s), \end{aligned} \tag{9}$$

for $s \in \mathscr{S}_n$. To ease the notation and terminology, we call $p(s)$, $s \in \mathscr{C}(s^*, \mathscr{M}^*)$, the predictive distribution of the compatible rankings. Since $p(s) = 0$ for $s \notin \mathscr{C}(s^*, \mathscr{M}^*)$, we do not need to consider such permutations.

The predicted ranking, $\hat{r}_y$, for the user is equal to the mode of the predictive distribution. In view of the above identities, we have

$$\hat{r}_y = \operatorname*{argmax}_{r_y} \mathrm{P}(R_Y = r_y \mid R_X = r_x,\, R_Y^* = r_y^*) = \hat{s} \circ r_x$$

where

$$\hat{s} = \operatorname*{argmax}_{s \in \mathscr{C}(s^*, \mathscr{M}^*)} p(s) = \operatorname*{argmax}_{s \in \mathscr{C}(s^*, \mathscr{M}^*)} \mathrm{P}(S = s). \tag{10}$$

Two questions arise: How to compute the marginal and predictive distributions $\Pr(S = s)$ and $p(s)$, respectively? How to find the mode, $\hat{s}$, of the predictive distribution? For the family of Farlie–Gumbel–Morgenstern (FGM) copulas, we can find explicit formulas for the marginal probabilities $\Pr(S = s)$. For other copula families, we propose Monte Carlo algorithms, the performance of which we assess by using the explicit formulas for the FGM family as benchmark.

4. Farlie–Gumbel–Morgenstern copula family

4.1. **Rank likelihood.** The copulas in the FGM family have the form $C_\theta(u, v) = uv\{1 + \theta(1 - u)(1 - v)\}$, with densities $c_\theta(u, v) = 1 + \theta(1 - 2u)(1 - 2v)$, for $(u, v) \in [0, 1]^2$ and with parameter $\theta \in \Theta = [-1, 1]$. The fact that the density is polynomial allows us to evaluate the rank likelihood $\mathrm{P}_\theta(S = s)$ in (6) explicitly. The resulting expression is a polynomial of degree $n - 1$ in θ.

Theorem 4.1. *Let (X_i, Y_i), $i = 1, \ldots, n$, be iid random vectors with continuous margins and* FGM *copula C_θ, $\theta \in [-1, 1]$. Then*

$$\mathrm{P}_\theta(S = s) = \sum_{j=0}^{n-1} c_j(s)\, \theta^j, \qquad s \in \mathscr{S}_n, \tag{11}$$

with $c_0(s) = 1/n!$, and

$$c_j(s) = n! \sum_{1 \leqslant i_1 < i_2 < \cdots < i_j \leqslant n} d_j(i_1, \ldots, i_j)\, d_j(s(i_1), \ldots, s(i_j)), \quad j = 1, \ldots, n - 1, \tag{12}$$

where

$$d_j(i_1,\dots,i_j) = \frac{1}{(n+j)!}\sum_{k_1=1}^{n+1}\cdots\sum_{k_j=1}^{n+1}(-1)^{\sum_{\ell=1}^{j}\mathbb{1}(k_\ell>i_\ell)}\prod_{p=1}^{n+1}\left\{\sum_{\ell=1}^{j}\mathbb{1}(k_\ell=p)\right\}!. \tag{13}$$

The FGM model gives rise to some symmetries in the rank likelihood. Let $a=(n,\dots,1)\in\mathscr{S}_n$ be the anti-identity. Note that $a^{-1}=a$ and put $a^0=e$.

Lemma 4.2. *For $s\in\mathscr{S}_n$ and $\theta\in[-1,1]$, we have, for the* FGM *copula family,*

$$\mathrm{P}_\theta(S=s)=\mathrm{P}_{(-1)^{i+j}\theta}(S=a^i\circ s^k\circ a^j),\qquad i,j\in\{0,1\},\ k\in\{-1,1\}. \tag{14}$$

Inspecting the proof of Lemma 4.2, we see that the symmetry property (14) holds for any family of copula densities $(c_\theta:\theta\in\Theta)$ such that $c_\theta(u,v)=c_\theta(v,u)$ and $c_\theta(1-u,v)=c_{-\theta}(u,v)$ for all $(u,v)\in[0,1]^2$, where it is assumed that the parameter set $\Theta\subset\mathbb{R}$ is symmetric around the origin. Besides the FGM family, this includes, after reparametrization, the bivariate Frank, Plackett, and Gauss copula families.

4.2. **Marginal distribution of the rank statistic.** Let the FGM parameter θ have prior density π over $\Theta=[-1,1]$. It follows from Theorem 4.1 that the marginal distribution of S is

$$\mathrm{P}(S=s)=\mathrm{E}_\pi\{\mathrm{P}_\theta(S=s)\}=\int_{-1}^{1}\mathrm{P}_\theta(S=s)\,\pi(\theta)\,d\theta=\sum_{j=0}^{n-1}c_j(s)\int_{-1}^{1}\theta^j\,\pi(\theta)\,d\theta,$$

for $s\in\mathscr{S}_n$. The marginal probabilities $\mathrm{P}(S=s)$ are directly obtained via the calculation of the moments of order $1,\dots,n-1$ of the prior distribution.

The symmetries found in Lemma 4.2 for the rank likelihood carry through to the marginal distribution. If the prior on $[-1,1]$ is symmetric, which is the case for Jeffreys' prior discussed below, we obtain the equality

$$\mathrm{P}(S=s)=\mathrm{P}(S=a^i\circ s^k\circ a^j),\qquad i,j\in\{0,1\},k\in\{-1,1\}, \tag{15}$$

with $a=(n,n-1,\dots,1)\in\mathscr{S}_n$ the anti-identity. This symmetry property reduces the number of distinct values for the probabilities $\mathrm{P}(S=s)$, $s\in\mathscr{S}_n$, $n\geqslant 2$.

Next we investigate the mode of the marginal distribution of S. In many cases, the identity, e, or the anti-identity, a, are modes, and sometimes even both rankings are modal. The next result gives sufficient conditions for e or a to be modal rankings.

Theorem 4.3. *Let π be a (prior) density on the* FGM *parameter $\theta\in\Theta=[-1,1]$.*

(i) If the odd order moments of π are nonnegative, that is, if $\mathrm{E}_\pi(\theta^{2k+1})\geqslant 0$, for $k=0,\dots,\lfloor n/2\rfloor-1$, then the identity $s=e$ is a mode of the marginal distribution $\mathrm{P}(S=s)$, $s\in\mathscr{S}_n$.

(ii) If the odd order moments of π are nonpositive, then the anti-identity $s=a$ is a mode of the marginal distribution.

4.3. **Prior distributions.** We consider two priors on the FGM family, namely the Beta distribution and Jeffreys' prior.

For the Beta prior, let $\theta=2T-1$ with $T\sim\mathrm{Beta}(\alpha,\beta)$ and parameters $\alpha>0$ and $\beta>0$, and let $\pi_{\alpha,\beta}$ denote the resulting density. The moments of θ are easily computed:

$$\int_{-1}^{1}\theta^n\,\pi_{\alpha,\beta}(\theta)\,\mathrm{d}\theta=\frac{(-1)^n}{(n+1)B(\alpha,\beta)}\sum_{k=0}^{n}(-1)^k\frac{B(\alpha+k,\beta+n-k)}{B(1+k,1+n-k)},\qquad n=0,1,2,\dots,$$

where $B(\alpha,\beta)$, for $\alpha>0,\beta>0$, is the beta function. We have a corollary to Theorem 4.3 for this particular choice of prior.

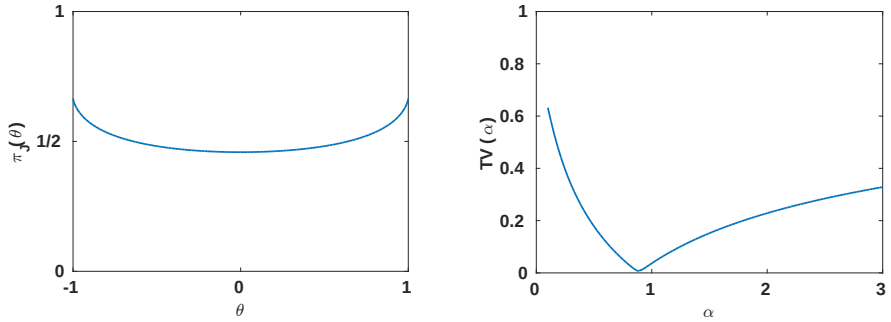


FIGURE 1. Jeffreys' prior (left) for the FGM family and its total variation distance (right) to the density of the random variable $\theta = 2T - 1$, where $T \sim \mathrm{Beta}(\alpha, \alpha)$ and $\alpha \in [0{\cdot}1, 3]$.

Corollary 4.4. *Let $\theta = 2T - 1$ with $T \sim \mathrm{Beta}(\alpha, \beta)$ for $\alpha > 0$ and $\beta > 0$.*
(i) If $0 < \beta \leqslant \alpha$, then the identity $s = e$ is a mode of the marginal distribution of S.
(ii) If $0 < \alpha \leqslant \beta$, then the anti-identity $s = a$ is a mode of the marginal distribution of S.

We now compute Jeffreys' prior $\pi_J(\theta) \propto \sqrt{I(\theta)}$, for $\theta \in [-1, 1]$, with $I(\theta)$ the Fisher information at θ. We have

$$I(\theta) = \int_{(0,1)^2} \left(\frac{\partial}{\partial \theta} \log c_\theta(u, v) \right)^2 \mathrm{d}(u, v) = \begin{cases} \frac{1}{9} & \text{if } \theta = 0, \\ \frac{1}{\theta^2} \left\{ \frac{\mathrm{LI}_2(\theta) - \mathrm{LI}_2(-\theta)}{2\theta} - 1 \right\} & \text{if } \theta \in [-1, 1] \setminus \{0\}, \end{cases}$$

where $\mathrm{LI}_2(x) = -\int_0^1 y^{-1} \log(1 - xy)\, \mathrm{d}y = \sum_{k=1}^\infty k^{-2} x^k$, for $x \leqslant 1$, is the dilogarithm function. It follows that the Fisher information is

$$I(\theta) = \sum_{k=0}^{\infty} \frac{\theta^{2k}}{(2k+3)^2}, \qquad \theta \in [-1, 1].$$

See Figure 1(a) for a graph of Jeffreys' prior, π_J. Its odd order moments vanish because of symmetry, and its even order moments can be computed by numerical quadrature.

Jeffreys' prior π_J being symmetric, we compare it with the symmetric subfamily $\pi_{\alpha,\alpha}$, $\alpha > 0$, of the Beta prior. We consider the total variation distance between π_J and $\pi_{\alpha,\alpha}$, i.e.,

$$\mathrm{TV}(\alpha) = \frac{1}{2} \int_{-1}^{1} |\pi_J(\theta) - \pi_{\alpha,\alpha}(\theta)|\, \mathrm{d}\theta, \qquad \alpha > 0.$$

The minimal value of $\mathrm{TV}(\alpha)$, obtained numerically, is attained when $\alpha = 0{\cdot}88$, with $\mathrm{TV}(0{\cdot}88) = 0{\cdot}0082$, see Figure 1(b). The symmetric Beta prior with this value for α may be a numerically tractable alternative to Jeffreys' prior.

We illustrate the marginal distributions $\mathrm{P}(S = s)$, $s \in \mathscr{S}_n$, obtained with Jeffreys' prior and with an asymmetrical Beta prior. The lack of a universal total ordering on $\mathscr{S}_n$ makes graphing a bit difficult. We visualize the marginal distributions arising from both priors by plotting the marginal probabilities of $s \in \mathscr{S}_n$ against the Kendall distance, d_τ, of s from the modal rankings, the latter depending on the prior. The Kendall distance (Diaconis, 1988) on $\mathscr{S}_n$ is given essentially by the number of discordances between two permutations; more precisely,

$$d_\tau(s, s') = \sum_{1 \leqslant i < j \leqslant n} \mathbb{1}(s' \circ s^{-1}(i) > s' \circ s^{-1}(j)), \qquad s, s' \in \mathscr{S}_n. \tag{16}$$

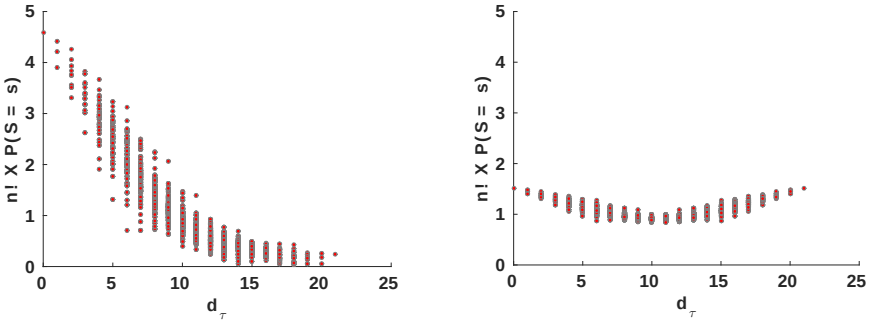


FIGURE 2. Rescaled marginal distribution $n!\,\mathrm{P}(S = s)$, $s \in S_7$, for the FGM copula family, plotted against the Kendall distance $d_\tau(s, s_\pi)$ of the rankings s from the modal rank s_π for the given prior π. Left: asymmetrical Beta prior $\theta = 2T - 1$ with $T \sim \mathrm{Beta}(1/10, 2)$. Right: Jeffreys' prior π_J.

In particular, we have the relation $0 \leqslant \binom{n}{2}^{-1} d_\tau(s, s') = \{1 - \tau(s, s')\}/2 \leqslant 1$, where τ is the sample version of Kendall's tau of the sample $(s(1), s'(1)), \ldots, (s(n), s'(n))$.

The marginal distributions are illustrated in Figure 2 for an asymmetric Beta prior on the left and for Jeffrey's prior on the right. The superposition of points is explained by Lemma 4.2 and equation (15). For the asymmetric Beta prior, Corollary 4.4 implies that the mode of the marginal distribution is the anti-identity $s = a$. Jeffrey's prior is symmetric, so that, by Theorem 4.3, both the identity, $s = e$, and the anti-identity, $s = a$, are modes of the marginal distribution. The symmetry that appears for Jeffrey's prior is also an artifact of the FGM model and will also appear for other exchangeable and radially symmetric copula families, as discussed after Lemma 4.2. Since $a \circ \sigma = (n + 1 - \sigma(1), \ldots, n + 1 - \sigma(n))$, with $a = (n, \ldots, 1)$ the anti-identity and $\sigma \in \mathscr{S}_n$, we get

$$d_\tau(s, a \circ s') = \binom{n}{2} - d_\tau(s, s'), \qquad s, s' \in \mathscr{S}_n. \tag{17}$$

Together with the equality (15), we obtain that the marginal probabilities are symmetrical with respect to the midrange distance $\binom{n}{2}/2$.

4.4. **Predictive distribution.** The posterior predictive distribution, $p(s)$, of the rank statistic S is equal to the marginal distribution conditioned on the event $\{S \in \mathscr{C}(s^*, \mathscr{M}^*)\}$; see (9). The polynomial form of the rank-likelihood (11) induced by the FGM family together with moment formulas for the prior distributions then allow us to compute predictive probabilities $p(s)$ exactly.

To provide an example, take as a toy problem the incomplete rankings $(-, 2, -, 1, 3, -, -)$ or in other words, $n = 7$, $\mathscr{M}^* = \{2, 4, 5\}$ and $s^* = (2, 1, 3)$, and consider the same two priors as in Figure 2. The predictive distribution $p(s)$, $s \in \mathscr{C}(s^*, \mathscr{M}^*)$, from equation (9) is illustrated in Figure 3. The predictive distribution associated to Jeffreys' prior has two modes, $s = (1, 4, 2, 3, 5, 6, 7)$ and $s^{-1} = (1, 3, 4, 2, 5, 6, 7)$. We will return to this toy example in Section 5.3.

In contrast to the marginal distribution, the posterior predictive distribution arising from Jeffrey's prior is no longer symmetric around the midrange distance $\binom{n}{2}/2$: compare the right-hand panels of Figures 2 and 3. Recall that the symmetry property of the marginal distribution is due to a combination of equations (15) and (17). For the predictive distribution, this explanation breaks down, because $s \in \mathscr{C}(s^*, \mathscr{M}^*)$ implies $a \circ s \notin \mathscr{C}(s^*, \mathscr{M}^*)$. Indeed, a compatible ranking

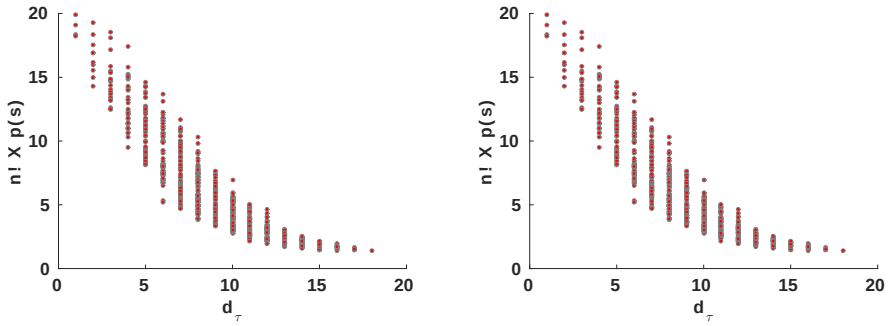


FIGURE 3. Rescaled predictive probabilities, $n!\,p(s)$, of all compatible rankings $s \in \mathscr{C}(s^*, \mathscr{M}^*)$, where $n = 7$, $\mathscr{M}^* = \{2, 4, 5\}$, and $s^* = (2, 1, 3)$, using the FGM copula family. The values are plotted against the Kendall distance of the rankings to the modal rank. Left: asymmetrical Beta prior $\theta = 2T - 1$ with $T \sim \text{Beta}(1/10, 2)$. Right: Jeffreys' prior π_J.

$s \in \mathscr{C}(s^*, \mathscr{M}^*)$ must satisfy

$$s(i^*_{\sigma(1)}) < s(i^*_{\sigma(2)}) < \cdots < s(i^*_{\sigma(m)}), \qquad \sigma = (s^*)^{-1},$$

but for $s' = a \circ s = (n+1-s(1), \ldots, n+1-s(n))$, we have $s'(i^*_{\sigma(1)}) > s'(i^*_{\sigma(2)}) > \cdots > s'(i^*_{\sigma(m)})$, and so $a \circ s \notin \mathscr{C}(s^*, \mathscr{M}^*)$. In passing, note that, in contrast to the ranking $a \circ s$, the ranking $s \circ a$ may or may not belong to the compatible rankings. Take for instance $n = 3$, $\mathscr{M}^* = \{1, 2\}$, and $s^* = (1, 2)$. On the one hand, we have $e = (1, 2, 3) \in \mathscr{C}(s^*, \mathscr{M}^*)$ but $a = (3, 2, 1) \notin \mathscr{C}(s^*, \mathscr{M}^*)$. On the other hand, if $s = (2, 3, 1)$, we have both $s \in \mathscr{C}(s^*, \mathscr{M}^*)$ and $s \circ a = (1, 3, 2) \in \mathscr{C}(s^*, \mathscr{M}^*)$.

5. ALGORITHMS

5.1. **Drawing compatible rankings.** The mode of the posterior predictive distribution is the ranking used in order to make a recommendation to the individual. The simulated annealing algorithm proposed in Section 5.2 below gives a way to approximate this mode. It is based on an algorithm to draw random compatible rankings. In practice, the cardinality, $n!/m!$, of $\mathscr{C}(s^*, \mathscr{M}^*)$ can be enormous, and a complete listing of all compatible rankings is elusive. One way to draw samples from the uniform distribution over $\mathscr{C}(s^*, \mathscr{M}^*)$ is to draw a permutation $\tilde{s}$ randomly from $\mathscr{S}_n$ and then turn it to a compatible ranking $s \in \mathscr{C}(s^*, \mathscr{M}^*)$ by rearranging $(\tilde{s}(i^*_1), \ldots, \tilde{s}(i^*_m))$ in such a way that $\text{rank}(s(i^*_1), \ldots, s(i^*_m)) = s^*$. This algorithm could be used for constructing Markov chain Monte Carlo algorithms with independent proposals. Here, we are interested in random walk type proposals, and so we construct an ergodic Markov chain on $\mathscr{C}(s^*, \mathscr{M}^*)$ which happens to have a uniform stationary distribution. It will be used as a proposal distribution in Algorithms 2 and 3 below.

——————————— ALGORITHM 1

Let $s \in \mathscr{C}(s^*, \mathscr{M}^*)$ be the current state of the chain. The next state $s' \in \mathscr{C}(s^*, \mathscr{M}^*)$ is obtained by selecting at random one move between moves M_1 and M_2 with equal probability.

M_1 – *Swap move.* Draw a pair $\{i, j\}$ where $i, j \in \mathscr{M}^c$, $i < j$ (with probability $1/\binom{n-m}{2}$), and take s' such that $s'(i) = s(j)$, $s'(j) = s(i)$, and $s'(t) = s(t)$, for every $t \in \{1, \ldots, n\} \setminus \{i, j\}$.

M_2 – *Swap and rearrange move.* Draw $\ell \in \{1,\dots,m\}$ and $j \in \mathscr{M}^c$ (with probability $1/[m(n-m)]$) and then take s' such that $s'(j) = s(i^*_\ell)$, $s'(t) = s(t)$, for every $t \in \mathscr{M}^c \setminus \{j\}$, and such that $\{s'(i^*_k)\colon k = 1,\dots,m\} = \{s(i^*_k)\colon k = 1,\dots,m, k \neq \ell\} \cup \{s(j)\}$, with

$$s'(i^*_{\sigma(1)}) < s'(i^*_{\sigma(2)}) < \cdots < s'(i^*_{\sigma(m)}), \quad \sigma = (s^*)^{-1}.$$

□

Lemma 5.1. *If $1 < m < n$, then Algorithm 1 generates an irreducible and aperiodic Markov chain with uniform stationary distribution on $\mathscr{C}(s^*, \mathscr{M}^*)$.*

5.2. **Finding the modal ranking of the predictive distribution.** A simple way to compute the predictive probabilities $p(s)$ in (9) is by standard Monte-Carlo approximations, exploiting Theorem 2.2 for the rank likelihood. Combining this with draws from the compatible rankings in $\mathscr{C}(s^*, \mathscr{M}^*)$ according to Algorithm 1, we propose the following simulated annealing algorithm for approximating (10).

——————————— Algorithm 2

At each iteration t, let $S_t \in \mathscr{C}(s^*, \mathscr{M}^*)$ be the current state of the rankings.

SA_1. Draw S' according to Algorithm 1.

SA_2. For $k = 1,\dots,K$, repeat move MC_1 and when completed, do move MC_2.

MC_1. Draw $(W'_{\ell,1},\dots,W'_{\ell,n+1}) = (W'_{\ell,1},\dots,W'_{\ell,n}, 1 - \sum_{j=1}^n W'_{\ell,j})$, for $\ell = 1, 2$, from Dirichlet$(1,\dots,1)$, and independently draw $\theta' \sim \pi$, and evaluate

$$p_k(S') = \frac{1}{n!} \prod_{i=1}^n c_{\theta'}\Big(\sum_{j=1}^i W'_{1,j}, \sum_{j=1}^{S'(i)} W'_{2,j} \Big).$$

MC_2. Compute $\hat{p}(S') = \frac{1}{K} \sum_{k=1}^K p_k(S')$.

SA_3. Set $S_{t+1} = S'$ with probability

$$1 \wedge \exp\left\{ \frac{\hat{p}(S') - \hat{p}(S_t)}{T_t} \right\},$$

where $T_t = 1/\log t$ is the temperature. Otherwise, $S_{t+1} = S_t$.

□

As an illustration, we consider the toy example of Figure 3, with incomplete permutation $(-, 2, -, 1, 3, -, -)$, the FGM model and the Jeffreys' prior on the copula parameter. In Figure 4, the excursion eventually oscillates between the two modes $s = (1, 4, 2, 3, 5, 6, 7)$ and $s^{-1} = (1, 3, 4, 2, 5, 6, 7)$, found at the beginning of this section, and $d_\tau(s, s^{-1}) = 2$.

5.3. **Approximating the predictive distribution.** It can be interesting for a recommendation system developper to explore more than a single ranking in order to make the recommendation to the individual. One reason may be to prevent local maxima situations inherent to simulated annealing algorithms. We therefore propose a second, more involved algorithm to obtain the few most likely rankings according to the predictive distribution $p(s)$, $s \in \mathscr{C}(s^*, \mathscr{M}^*)$. The idea is to construct a Markov chain with limiting (stationary) distribution equal to the conditional distribution of (S, W_1, W_2, θ) given $\{S \in \mathscr{C}(s^*, \mathscr{M}^*)\}$. In line with Theorem 2.2, the random vector (S, W_1, W_2, θ) has joint density given by

$$f(s, w_1, w_2, \theta) = n! \prod_{i=1}^n c_\theta\Big(\sum_{j=1}^i w_{1,j}, \sum_{j=1}^{s(i)} w_{2,j} \Big) \pi(\theta), \qquad s \in \mathscr{S}_n, (w_1, w_2) \in \Delta^2, \theta \in \Theta, \tag{18}$$

with respect to $\nu \times \lambda_1 \times \lambda_2$, where ν is the counting measure, and λ_1 and λ_2 are the Lebesgue measures on Δ^2 and Θ, respectively, with $\Delta = \{w \in (0,1)^n : w_1 + \cdots + w_n < 1\}$.

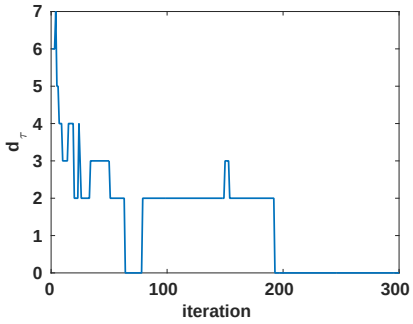


FIGURE 4. Excursion $S_1, S_2, \ldots$ from the simulated annealing Alorithm 2 excursion. Here, as in Figure 3, the incomplete ranking is $n = 7$, $\mathscr{M}^* = \{2, 4, 5\}$ and $s^* = (2, 1, 3)$, that is, $(-, 2, -, 1, 3, -, -)$. The illustration shows the Kendall distance, d_τ, of the visited permutations from the mode plotted against the iteration number. The FGM copula parameter θ is distributed according to Jeffreys' prior π_J.

Essentially, $S_1, S_2, \ldots$ travels through the space $\mathscr{C}(s^*, \mathscr{M}^*)$, with limiting relative frequencies of occupancy converging to the predictive probabilities $p(s)$ in (9). The move M_2 below concerns the proposal of the variables (W_1, W_2) given S and θ. We have looked at two ways to do so, resulting in two variations of the algorithm: the proposal is drawn either independently of the current value of (W_1, W_2) or from the instrumental density in equation (20) below. The two variations will be called (MHI) and (MHRW) in Move M_2 in Algorithm 3. Move M_3 of the algorithm requires a draw of a $\theta' \in \Theta$ according to some instrumental density $q(\theta' \mid \theta)$, given the previous value $\theta \in \Theta$; for the FGM family, this is the density of the uniform distribution on the interval $[\theta - \varepsilon, \theta + \varepsilon] \cap [-1, 1]$, for some tuning parameter $\varepsilon > 0$.

——————————— ALGORITHM 3

Let (S, W_1, W_2, θ) be the initial state of the chain, with $S \in \mathscr{C}(s^*, \mathscr{M}^*)$, $(W_1, W_2) \in \Delta^2$, and $\theta \in \Theta$. Let f be the density given in (18).

(1) At each iteration $t = 1, \ldots, N$, select a move at random (equiprobably) between moves M_1, M_2, and M_3.

M_1 – Draw S' according to Algorithm 1 and replace S by S' with probability

$$1 \wedge \frac{f(S', W_1, W_2, \theta)}{f(S, W_1, W_2, \theta)}.$$

M_2 – (MHI) Draw $W'_\ell = (W'_{\ell,1}, \ldots, W'_{\ell,n+1}) \sim \mathrm{Dirichlet}(1, \ldots, 1)$, for $\ell = 1, 2$, and and replace (W_1, W_2) by (W'_1, W'_2) with probability

$$1 \wedge \frac{f(S, W'_1, W'_2, \theta)}{f(S, W_1, W_2, \theta)}. \tag{19}$$

(MHRW) Choose $\ell \in \{1, 2\}$ at random and draw $W'_\ell = (W'_{\ell,1}, \ldots, W'_{\ell,n})$ according to the distribution with density given by (20) with $w_0 = W_\ell$. Put $W'_j = (W'_{j,1}, \ldots, W'_{j,n}) = (W_{j,1}, \ldots, W_{j,n})$ for $j \in \{1, 2\} \setminus \{\ell\}$, and replace (W_1, W_2) by (W'_1, W'_2) with probability

$$1 \wedge \frac{f(S, W'_1, W'_2, \theta)\, q_{W_\ell}(W_\ell \mid W'_\ell)}{f(S, W_1, W_2, \theta)\, q_{W'_\ell}(W'_\ell \mid W_\ell)}.$$

M_3 – Draw θ' according to some density $q(\theta' \mid \theta)$; we use $\theta' \mid \theta \sim U(-1 \vee \{\theta - \varepsilon\}, 1 \wedge \{\theta + \varepsilon\})$. Replace θ by θ' with probability

$$1 \wedge \frac{f(S, W_1, W_2, \theta')\, q(\theta \mid \theta')}{f(S, W_1, W_2, \theta)\, q(\theta' \mid \theta)}.$$

The current state of the chain then becomes the (possibly unchanged) state, denoted for simplicity by (S, W_1, W_2, θ); set $S_t = S$.

(2) Let $\mathscr{O}(s^*, \mathscr{M}^*) = \{S_1, \ldots, S_N\}$ be the set of all the (distinct) values taken by $S_1, \ldots, S_N$. For each $s \in \mathscr{O}(s^*, \mathscr{M}^*)$, compute the relative frequency of s:

$$\hat{p}(s) = \frac{1}{N} \sum_{t=1}^{N} \mathbb{1}(S_t = s).$$

□

Lemma 5.2. *Let $w_0 = (w_{01}, \ldots, w_{0n}) \in \Delta$. If $W' \sim \text{Dirichlet}(1, \ldots, 1)$ and if Λ is an independent random variable on $(0, 1)$ with density g, then $W = (1 - \Lambda)w_0 + \Lambda W'$ has density*

$$q_W(w \mid w_0) = n!\, \mathbb{1}_\Delta(w) \int_{1-\delta(w;w_0)}^{1} \lambda^{-n}\, g(\lambda)\, \mathrm{d}\lambda, \tag{20}$$

where $\delta(w; w_0) = \min\{\frac{w_i}{w_{0i}} : 1 \leqslant i \leqslant n\} \wedge \{(1 - \sum_{i=1}^n w_i)/(1 - \sum_{i=1}^n w_{0i})\}$.

Let us compare the two variations, (MHI) and (MHRW), of Algorithm 3 and use the FGM family as a validating benchmark to assess their accuracy in approximating the entire predictive distribution. We look at the total variation distance from the true predictive distribution, p, to the approximation $\hat{p}_t$, as a function of the iteration t. Here, the total variation distance is defined as

$$\text{TV}(\hat{p}_t, p) = \frac{1}{2} \sum_{s \in \mathscr{C}(s^*, \mathscr{M}^*)} |\hat{p}_t(s) - p(s)|.$$

We have again taken the same situation as Figure 3 with Jeffreys' prior and we have considered the six incomplete rankings with $n = 7$ and $\mathscr{M}^* = \{2, 4, 5\}$, one such ranking for every $s^* \in \mathscr{S}_3$. Figure 5 shows the results for $s^* = (1, 2, 3)$ and $s^* = (3, 2, 1)$. The results for the other four permutations in $\mathscr{S}_2$ are similar.

Although roughly noticeable, there is and should be a certain symmetry in the results between Figure 5 (a) and (b). This is explained by the equality of the events

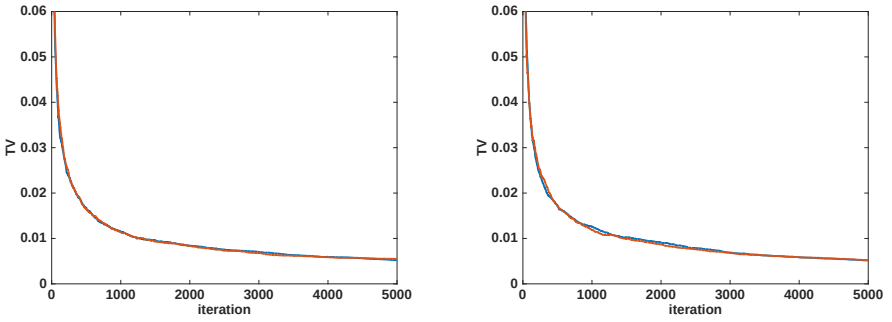
$$\{S \in \mathscr{C}(a \circ s^*, \mathscr{M}^*)\} = \{a \circ S \in \mathscr{C}(s^*, \mathscr{M}^*)\},$$

where, by a little abuse of notation, a stands for the anti-identity in $\mathscr{S}_m$ on the left-hand side and for the anti-identity in $\mathscr{S}_n$ on the right-hand side. It follows that for the FGM copula family and a symmetrical prior like Jeffreys' prior, since $\mathrm{P}(a \circ S = s) = \mathrm{P}(S = s)$ for all $s \in \mathscr{S}_n$, we obtain

$$\begin{aligned} \mathrm{P}\{S = s \mid S \in \mathscr{C}(a \circ s^*, \mathscr{M}^*)\} &= \mathrm{P}\{S = s \mid a \circ S \in \mathscr{C}(s^*, \mathscr{M}^*)\} \\ &= \mathrm{P}\{S = a \circ s \mid S \in \mathscr{C}(s^*, \mathscr{M}^*)\}. \end{aligned}$$

6. Comparisons with other recommender systems

We compare our method in Algorithm 2 with those available from the *Personalized Recommendation Algorithms* (PREA) java software, see Lee, Sun & Lebanon (2012), which contains some state-of-the-art techniques. Our competitors are the algorithms *SlopeOne*, introduced by Lemire & Maclachlan (2007), *Non-negative Matrix Factorization* (NMF), see for instance Lee & Seung (1999), *Probabilistic Matrix Factorization* (PMF), see Salakhutdinov & Mnih (2007), *Bayesian*



(a) $s^* = (1, 2, 3)$ (f) $s^* = (3, 2, 1)$

FIGURE 5. Total variation distance of the true predicted distribution to the approximation using Algorithm 3 (MHI) in blue and (MHRW) in orange, using Jeffreys' prior on the FGM copula parameter θ. The incomplete rankings are given by $n = 7$, $\mathscr{M}^* = \{2, 4, 5\}$ and $s^* = (1, 2, 3)$ (left) and $s^* = (3, 2, 1)$ (right). The values are plotted against the iteration number (in thousands).

Probabilistic Matrix Factorization (BPMF), see Salakhutdinov & Mnih (2008), *Regularized Singular Value Decomposition* (RegSVD), see Paterek (2007), and *Fast Non-negative Principal Component Analysis* (NPCA), see Yu, Zhu, Lafferty & Gong (2009). We refer to our method as *Bayesian Bivariate Ranks* (BBR).

We consider the well known *MovieLens 100k* (https://grouplens.org/datasets/movielens/100k/) dataset, comprising 1664 users, 943 movies, with a total of 99392 ratings, from which we select a subset of users and movies with many ratings to act as all the data. More precisely, we consider a matrix of 100 users and 35 movies, with 2687 available ratings, from one to 5 stars. In the selected matrix, the movies are ranked in order from the highest rated movie (by averaging the user ratings of each movie in the entire *MovieLens 100k* dataset) to the lowest rated one. Note that this overall ranking is important only for our method and acts as the expert opinion. In view of the notation of Section 2, this ordering implies that $r_x = e$ and $r_y = s$ for every user, although the movies ranked for one user may differ from that of another user.

The user ratings have ties. While this does not cause problems for the overall ordering of the movies using the entire *MovieLens 100k* dataset, it does require a choice for obtaining an individual user's permutation $r_y = s$. We break the ties using the expert opinion, in that if a user has given identical ratings to two or movies, the overall ordering determines their mutual ranks.

From the ratings data, we keep (at random) a certain proportion p of the data, and this we do for each user. We have considered the proportions 5%, 10%, 15%, 20%, 25%, 50% and 75%, see the graphs below. For a given proportion p of ratings kept, the matrix thus obtained becomes the data which we shall use to predict the rankings of the users. This is done, in turn, for each of the 100 users. Apart from the prior specification, see below, our methodology uses only the current user's retained rankings, obtained as discussed above. In contrast, the other methodologies use the entire matrix of retained ratings (not rankings), for each user. We then repeat all of this 30 times. There are many metrics used to evaluate recommender systems, see Gunawardana & Shani (2009) and Lee et al. (2012), among which the Kendall distance d_τ in (16). Since we are interested on the predicted rankings and not ratings, we shall focus on this distance for evaluating the methodologies. Note that the methodologies considered here do not give ties very often, especially when p is large, and so the predictions, either of ratings (the competitors) or

rankings (us) are all (or can be transformed into) genuine permutations. In the few encountered events where there were ties in the predicted ratings (for small p), again we broke the ties using the expert opinion, to obtain permutations.

We have used the Gaussian copula family with parameter $-1 < \rho < 1$ as model for the dependence between the expert and user ratings. For the prior on ρ, we consider the one-to-one relation $\rho = \sin(\tau\pi/2)$, where τ is Kendall's τ. We have obtained the empirical distribution of τ using the complete MovieLens data set accross all users and have applied the above transformation to obtain an estimate of the distribution of ρ. We finally fit the density of $2T - 1$, where $T \sim \text{Beta}(\alpha, \beta)$, to this estimate which gave the approximate values $\alpha = 6$ and $\beta = 2$. This is the prior for ρ.

For a fixed proportion p of available data in the matrix, and for each user u, let s_u be his or her (true) movie rankings and let $\hat{s}_{u,k,p}$ be the prediction based on the kept data, and this for repetition $k = 1, \ldots, 30$. The plots in Figure 6 show boxplots accross the repetitions k of

$$d(p, k) = \frac{1}{100} \sum_{u=1}^{100} d_\tau(\hat{s}_{u,k,p}, s_u) \tag{21}$$

for all the methods considered and for various choices of p, the proportion of data kept. To give a global idea of the performance of each method, Figure 7 shows the values of

$$\bar{d}(p) = \frac{1}{30} \sum_{k=1}^{30} d(p, k) \tag{22}$$

as a function of p.

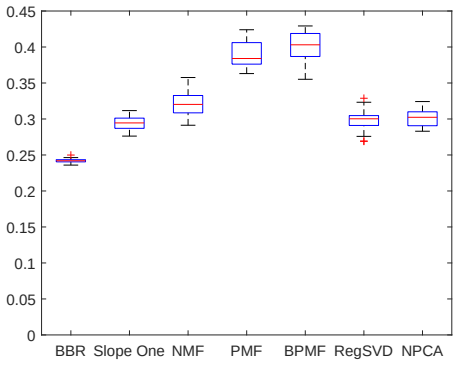
While the method that we propose seems to do better than the other methods when the information provided by the users is limited, it is also less variable than most of the other methods. In fact, the other methods that we have looked at depend only on the available users' data, whereas our method incorporates expert information. Our method could be of interest to a start-up company having little available data at first, until maybe switching too another method when the amount of data it has increases.

Besides the methods mentioned, we have also tried collaborative filtering methods and local-low ranks matrix factorization methods provided by the PREA toolkit. These are useful for large-scale data but did not perform as well as the other methods considered in our experiments.

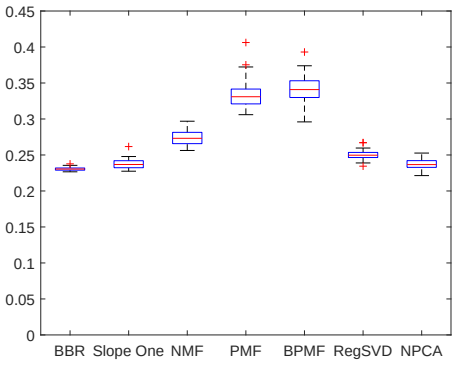
Our method could also be used to predict the ranks of a user's top n' movies more quickly. One way to do this is to run the chain on the entire set of n movies with the m partial (relative) rankings, therefore using all the information at hand, and then stop the chain. Consider only those n' movies that have appeared most often during this first run. For those n' movies, consider their relative rankings on $\mathscr{S}_{n'}$ obtained by the initial general (or expert) rankings. Consider also the relative m' partial rankings on $\mathscr{S}_{m'}$ of the $m' \in \{0, 1, \ldots, m\}$ movies of the user's top n' that belong to the list of the m initial partial rankings. Finally, apply the algorithm to the resulting permutation, with n replaced by n' and m by m', to obtain the predicted ranking of the user's top n' movies.

Acknowledgement

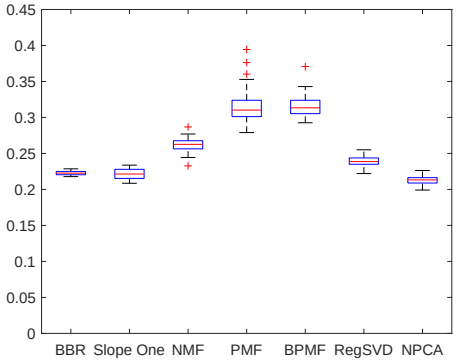
J. Segers gratefully acknowledges financial support from the Projet d'Actions de Recherche Concertées programme of the Communauté française de Belgique and from a Interuniversity Attraction Pole research network grant of the Belgian federal government.



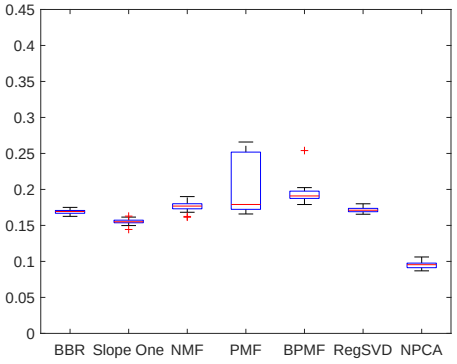
(a) $p = 0.05$



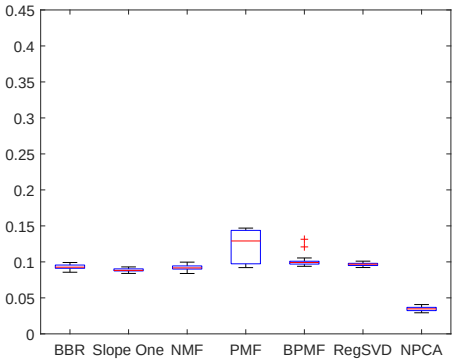
(b) $p = 0.15$



(c) $p = 0.2$



(d) $p = 0.25$



(e) $p = 0.75$

FIGURE 6. Boxplots of the $d(p, k)$ in (21) over $k = 1, \ldots, 30$, for various values of p and for various recommender systems. Our method is BBR.

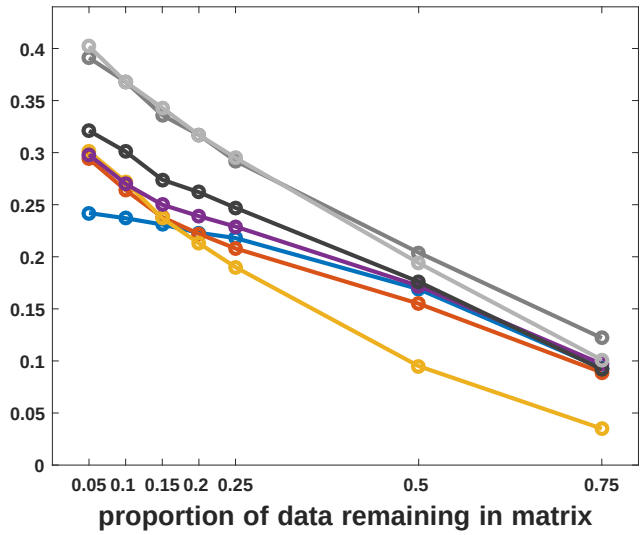


FIGURE 7. Values of $d(p)$, given by (22), for $p = 5\%, 10\%, 15\%, 20\%, 25\%, 50\%$ and 75%. Our method (BBR) is in blue, yellow is NPCA, orange is Slope One, purple is RegSVD, black is NMF, dark grey is PMF, and light grey is BPMF.

APPENDIX A. PROOFS

Proof of Lemma 2.1. For $\sigma \in \mathscr{S}_n$, consider the random vectors $X \circ \sigma = (X_{\sigma(1)}, \ldots, X_{\sigma(n)})$ and $Y \circ \sigma = (Y_{\sigma(1)}, \ldots, Y_{\sigma(n)})$. The joint distribution of $(X \circ \sigma, Y \circ \sigma)$ is the same as the one of (X, Y). By (3), we have $R_{X\circ\sigma} = R_X \circ \sigma$ and $R_{Y\circ\sigma} = R_Y \circ \sigma$ with probability one, i.e., in the absence of ties. Setting $\sigma = \tau^{-1} \circ r_x$ with $\tau \in \mathscr{S}_n$, we obtain

$$\begin{aligned} \mathrm{P}(R_X = r_x,\, R_Y = r_y) &= \mathrm{P}(R_X \circ \tau^{-1} \circ r_x = r_x,\, R_Y \circ \tau^{-1} \circ r_x = r_y) \\ &= \mathrm{P}(R_X = \tau,\, R_Y \circ \tau^{-1} = r_y \circ r_x^{-1}) \\ &= \mathrm{P}\{R_X = \tau,\, S(X,Y) = r_y \circ r_x^{-1}\}. \end{aligned}$$

Summing over $\tau \in \mathscr{S}_n$, we find that

$$\mathrm{P}\{S(X,Y) = r_y \circ r_x^{-1}\} = \sum_{\tau \in \mathscr{S}_n} \mathrm{P}\{R_X = \tau,\, S(X,Y) = r_y \circ r_x^{-1}\} = n!\, \mathrm{P}(R_X = r_x,\, R_Y = r_y),$$

since the cardinality of $\mathscr{S}_n$ is $n!$. □

Proof of Theorem 2.2. As explained before the theorem, we have $R_X = R_U$ and $R_Y = R_V$ almost surely, and thus $S(X,Y) = S(U,V)$ almost surely. Let $s \in \mathscr{S}_n$ and let

$$E_s = \{(u,v) \in (0,1)^{2n} : u_1 < \ldots < u_n,\, v_{s^{-1}(1)} < \ldots < v_{s^{-1}(n)}\}.$$

Since $(0,1)^{2n} \setminus \mathbb{D}_n^2$ has Lebesgue measure zero, we find, by Lemma 2.1 with $r_x = e$,

$$\mathrm{P}\{S(X,Y) = s\} = n!\, \mathrm{P}(R_U = e,\, R_V = s) = n! \int_{E_s} \prod_{i=1}^{n} c(u_i, v_i)\, \mathrm{d}u_i\, \mathrm{d}v_i. \tag{23}$$

For $(u,v) \in (0,1)^{2n} \cap \mathbb{D}_n^2$, the vector $(u_{(1)}, \ldots, u_{(n)}; v_{(s(1))}, \ldots, v_{(s(n))})$ belongs to E_s; here, $(z_{(1)}, \ldots, z_{(n)})$ denotes the vector of ascending order statistics of the vector $(z_1, \ldots, z_n) \in \mathbb{D}_n$.

We find

$$\mathrm{P}\{S(X,Y)=s\} = \frac{1}{n!}\int_{(0,1)^{2n}} \prod_{i=1}^{n} c(u_{(i)}, v_{(s(i))})\,\mathrm{d}u_i\,\mathrm{d}v_i = \frac{1}{n!}\,\mathrm{E}\left\{\prod_{i=1}^{n} c\big(\tilde{U}_{(i)}, \tilde{V}_{(s(i))}\big)\right\}, \tag{24}$$

where $\tilde{U}_1,\ldots,\tilde{U}_n,\tilde{V}_1,\ldots,\tilde{V}_n$ are iid random variables, uniformly distributed on $(0,1)$. The result then follows from the fact that $\tilde{U}_{(i)} = \sum_{j=1}^{i} W_{1,j}$ and $\tilde{V}_{(s(i))} = \sum_{j=1}^{s(i)} W_{2,j}$, for $i=1,\ldots,n$, where $(W_{\ell,1},\ldots,W_{\ell,n+1})$, for $\ell=1,2$, is the vector of $n+1$ spacings on $(0,1)$ based on $\tilde{U}_1,\ldots,\tilde{U}_n$ and $\tilde{V}_1,\ldots,\tilde{V}_n$ for $\ell=1$ and $\ell=2$, respectively. □

Proof of Lemma 3.1. Applying (3) with $\sigma = (r_x^*)^{-1}$, we find

$$\begin{aligned} r_y \in \mathscr{C}(r_y^*, \mathscr{M}) &\iff \mathrm{rank}(r_y(i_1),\ldots,r_y(i_m)) = r_y^* \\ &\iff \mathrm{rank}(r_y(i_1),\ldots,r_y(i_m)) \circ (r_x^*)^{-1} = r_y^* \circ (r_x^*)^{-1} \\ &\iff \mathrm{rank}(r_y(i_{(r_x^*)^{-1}(1)}),\ldots,r_y(i_{(r_x^*)^{-1}(m)}) = s^*. \end{aligned}$$

By definition, $\mathscr{M}^* = \{i_1^*,\ldots,i_m^*\}$, where $i_1^* < \ldots < i_m^*$ are the order statistics of the vector $(r_x(i_1),\ldots,r_x(i_m))$. Since $\mathrm{rank}(r_x(i_1),\ldots,r_x(i_m)) = r_x^*$, we find

$$i_{(r_x^*)^{-1}(j)} = r_x^{-1}(i_j^*), \qquad j=1,\ldots,m.$$

We obtain that

$$\begin{aligned} r_y \in \mathscr{C}(r_y^*, \mathscr{M}) &\iff \mathrm{rank}(r_y \circ r_x^{-1}(i_1^*),\ldots,r_y \circ r_x^{-1}(i_m^*)) = s^* \\ &\iff r_y \circ r_x^{-1} \in \mathscr{C}(s^*, \mathscr{M}^*). \end{aligned}$$

□

Proof of Theorem 4.1. By (23), we have

$$\mathrm{P}_\theta(S=s) = n!\int_{E_s}\prod_{i=1}^{n} c_\theta(u_i,v_i)\,\mathrm{d}u_i\,\mathrm{d}v_i = n!\int_{E_s}\prod_{i=1}^{n}\{1+\theta(2u_i-1)(2v_i-1)\}\,\mathrm{d}u_i\,\mathrm{d}v_i,$$

where $E_s = \{(u,v)\in(0,1)^{2n} : u_1<\cdots<u_n, v_{s^{-1}(1)}<\cdots<v_{s^{-1}(n)}\}$. Let $\Delta = \{w\in(0,1)^n : w_1+\cdots+w_n<1\}$ be the standard n-dimensional simplex. Consider the following change of variables from E_s to Δ^2:

$$w_{1i} = \begin{cases} u_1 & \text{for } i=1, \\ u_i - u_{i-1} & \text{for } 2\leqslant i\leqslant n, \end{cases} \quad\text{and}\quad w_{2i} = \begin{cases} v_{s^{-1}(1)} & \text{for } i=1, \\ v_{s^{-1}(i)} - v_{s^{-1}(i-1)} & \text{for } 2\leqslant i\leqslant n. \end{cases}$$

If we put $w_{\ell,n+1} = 1 - \sum_{j=1}^{n} w_{\ell j}$, for $\ell=1,2$, then we can write

$$\begin{aligned} 2u_i - 1 = u_i - (1-u_i) &= \sum_{k=1}^{i} w_{1k} - \sum_{k=i+1}^{n+1} w_{1k} = \sum_{k=1}^{n+1}(-1)^{\mathbb{1}(k>i)} w_{1k}, \\ 2v_i - 1 = v_i - (1-v_i) &= \sum_{k=1}^{s(i)} w_{2k} - \sum_{k=s(i)+1}^{n+1} w_{2k} = \sum_{k=1}^{n+1}(-1)^{\mathbb{1}(k>s(i))} w_{2k}. \end{aligned}$$

We obtain

$$\begin{aligned} \prod_{i=1}^{n} c_\theta(u_i(w_1), v_i(w_2)) &= \prod_{i=1}^{n}\left\{1+\theta\left(\sum_{k=1}^{n+1}(-1)^{\mathbb{1}(k>i)}w_{1k}\right)\left(\sum_{k=1}^{n+1}(-1)^{\mathbb{1}(k>s(i))}w_{2k}\right)\right\} \\ &= 1 + \sum_{j=1}^{n}\theta^j \sum_{1\leqslant i_1<i_2<\cdots<i_j\leqslant n} d_j(i_1,\ldots,i_j;w_1)\,d_j(s(i_1),\ldots,s(i_j);w_2), \end{aligned}$$

where, for all $w \in \Delta$,

$$\begin{aligned} d_j(i_1,\dots,i_j;w) &= \sum_{k_1=1}^{n+1}\cdots\sum_{k_j=1}^{n+1}(-1)^{\sum_{\ell=1}^{j}\mathbb{1}(k_\ell>i_\ell)}w_{k_1}\cdots w_{k_j} \\ &= = \sum_{k_1=1}^{n+1}\cdots\sum_{k_j=1}^{n+1}(-1)^{\sum_{\ell=1}^{j}\mathbb{1}(k_\ell>i_\ell)}w_1^{\sum_{\ell=1}^{j}\mathbb{1}(k_\ell=1)}\cdots w_{n+1}^{\sum_{\ell=1}^{j}\mathbb{1}(k_\ell=n+1)}. \end{aligned}$$

The equality

$$\mathrm{P}_\theta(S=s) = n!\int_\Delta\int_\Delta\left\{\prod_{i=1}^{n}c_\theta(u_i(w_1),v_i(w_2))\right\}\mathrm{d}w_1\,\mathrm{d}w_2,$$

shows that $c_0(s) = 1/n!$; indeed, $\int_\Delta \mathrm{d}w = 1/n!$. For $j = 1,\dots,n$, the expression for the coefficient $c_j(s)$ is obtained via

$$d_j(i_1,\dots,i_j) = \int_\Delta d_j(i_1,\dots,i_j;w)\,\mathrm{d}w = \sum_{k_1=1}^{n+1}\cdots\sum_{k_j=1}^{n+1}(-1)^{\sum_{\ell=1}^{j}\mathbb{1}(k_\ell>i_\ell)}\frac{\prod_{p=1}^{n+1}\left\{\sum_{\ell=1}^{j}\mathbb{1}(k_\ell=p)\right\}!}{(n+j)!},$$

for $\{i_1,\dots,i_j\}\subset\{1,\dots,n\}$, an expression which is seen by recognizing the Dirichlet density normalizing constants. This gives (13), and (12) follows.

Finally, to see that $c_n(s) = 0$ for all $s \in \mathscr{S}_n$, note that

$$1 = \sum_{s\in\mathscr{S}_n}\mathrm{P}_\theta(S(U,V)=s) = \sum_{j=0}^{n}\left\{\sum_{s\in\mathscr{S}_n}c_j(s)\right\}\theta^j = 1+\sum_{j=1}^{n}\left\{\sum_{s\in\mathscr{S}_n}c_j(s)\right\}\theta^j.$$

It follows that $\sum_{s\in\mathscr{S}_n}c_j(s) = 0$ for $j = 1,\dots,n$. Since $c_n(s) = n!\{d_n(1,\dots,n)\}^2$ does not depend on s, we conclude that $c_n(s) = 0$, as required. □

Proof of Lemma 4.2. First, the FGM copula density c_θ is symmetric in its arguments and so

$$\mathrm{E}\left\{\prod_{i=1}^{n}c_\theta\Big(U_{(i)},V_{(s^{-1}(i))}\Big)\right\} = \mathrm{E}\left\{\prod_{i=1}^{n}c_\theta\Big(U_{(s(i))},V_{(i)}\Big)\right\} = \mathrm{E}\left\{\prod_{i=1}^{n}c_\theta\Big(U_{(i)},V_{(s(i))}\Big)\right\}.$$

By using equation (24), we find $\mathrm{P}_\theta(S=s^{-1}) = \mathrm{P}_\theta(S=s)$.

Second, the FGM family also has the property that $c_\theta(u,v) = c_{-\theta}(1-u,v)$ for all $(u,v)\in[0,1]^2$. We find

$$\mathrm{E}\left\{\prod_{i=1}^{n}c_\theta\Big(U_{(i)},V_{(s(i))}\Big)\right\} = \mathrm{E}\left\{\prod_{i=1}^{n}c_{-\theta}\Big(U_{(i)},1-V_{(s(i))}\Big)\right\} = \mathrm{E}\left\{\prod_{i=1}^{n}c_{-\theta}\Big(U_{(i)},V_{(a\circ s(i))}\Big)\right\},$$

and so, again by (24), $\mathrm{P}_\theta(S=s) = \mathrm{P}_{-\theta}(S=a\circ s)$.

These two results imply $\mathrm{P}_{-\theta}(S=s\circ a) = \mathrm{P}_{-\theta}(S=a\circ s^{-1}) = \mathrm{P}_\theta(S=s^{-1}) = \mathrm{P}_\theta(S=s)$. Equation (14) follows. □

Proof of Theorem 4.3. Recall the expression for $\mathrm{P}_\theta(S=s) = \sum_{j=0}^{n-1}c_j(s)\,\theta^j$ for $s\in\mathscr{S}_n$ in Theorem 4.1. Since $\mathrm{P}_\theta(S=s) = \mathrm{P}_{-\theta}(S=a\circ s)$, for $\theta\in[-1,1]$, we have the relation

$$c_j(a\circ s) = \begin{cases} c_j(s), & \text{if } j \text{ is even,} \\ -c_j(s), & \text{if } j \text{ is odd.} \end{cases} \tag{25}$$

For every permutation $s\in\mathscr{S}_n$ and every $j = 1,\dots,n$, we have the identity $\{\{i_1,\dots,i_j\} : 1\leqslant i_1<i_2<\ldots<i_j\leqslant n\} = \{\{s(i_1),\dots,s(i_j)\} : 1\leqslant i_1<i_2<\ldots<i_j\leqslant n\}$. Moreover, the

expression d_j in (13) is symmetric in its arguments, that is, $d_j(i_1, \ldots, i_j) = d_j(i_{\sigma(1)}, \ldots, i_{\sigma(j)})$ for $\sigma \in \mathscr{S}_j$. It follows that

$$\sum_{1 \leqslant i_1 < i_2 < \cdots < i_j \leqslant n} \{d_j(i_1, \ldots, i_j)\}^2 = \sum_{1 \leqslant i_1 < i_2 < \cdots < i_j \leqslant n} \{d_j(s(i_1), \ldots, s(i_j))\}^2.$$

By the Cauchy–Schwarz inequality, $|c_j(s)| \leqslant c_j(e)$ for $j = 0, \ldots, n-1$. Finally, by the triangle inequality,

$$\begin{aligned} \mathrm{P}(S = s) = \mathrm{E}_\pi\{\mathrm{P}_\theta(S = s)\} &\leqslant \sum_{j=0}^{n-1} |c_j(s)\, \mathrm{E}_\pi(\theta^j)| \leqslant \sum_{k=0}^{\lfloor n/2 \rfloor - 1} c_{2k+1}(e)\, |\mathrm{E}_\pi(\theta^{2k+1})| + \sum_{k=0}^{\lfloor (n-1)/2 \rfloor} c_{2k}(e)\, \mathrm{E}_\pi(\theta^{2k}) \\ &= \begin{cases} \mathrm{P}(S = e), & \text{if all odd order moments are nonnegative,} \\ \mathrm{P}(S = a), & \text{if all odd order moments are nonpositive,} \end{cases} \end{aligned}$$

where the last equality follows from (25), with $s = e$. □

Proof of Corollary 4.4. By Theorem 4.3, we need to look at the signs of the odd order moments. Let $1 \leqslant k \leqslant n-1$ be an odd integer. We have

$$\mathrm{E}_{\pi_{\alpha,\beta}}(\theta^k) = \frac{B(\beta, \beta)}{B(\alpha, \beta)}\, \mathrm{E}[(2X-1)^k \{X^{\alpha-\beta} - (1/2)^{\alpha-\beta}\}], \qquad X \sim \mathrm{Beta}(\beta, \beta).$$

This is nonnegative if $\alpha \geqslant \beta$ and nonpositive if $\alpha \leqslant \beta$. □

Proof of Lemma 5.1. Let $S_0, S_1, S_2, \ldots$ be the Markov chain constructed by the algorithm.

First, two states $s, r \in \mathscr{C}(s^*, \mathscr{M}^*)$ are equal if and only if $s(t) = r(t)$, for every $t \in \mathscr{N} \setminus \mathscr{M}^*$. If $s = r$, then $\mathrm{P}(S_2 = r \mid S_0 = s) > 0$. Otherwise, let $\mathscr{N} \setminus \mathscr{M}^* = \{t_1, \ldots, t_{n-m}\}$, with $1 \leqslant t_1 < t_2 < \cdots < t_{n-m} \leqslant n$. Now if $s(t_1) \neq r(t_1)$, then there are two possible cases; either $r(t_1) = s(t_k)$ for some $k \in \{2, \ldots, n-m\}$, and a call to move M_1 can generate S_1 with $S_1(t_1) = r(t_1)$. Or, in the second case, $r(t_1) = s(i^*_\ell)$ for some $\ell \in \{1, \ldots, m\}$, and then a call to move M_2 can generate S_1 with $S_1(t_1) = r(t_1)$. Continuing this way for $t_2, \ldots, t_{n-m}$, we see that $\mathrm{P}(S_k = r \mid S_0 = s) > 0$ for some $k = 1, \ldots, n-m$, for every s and r in $\mathscr{C}(s^*, \mathscr{M}^*)$.

To show aperiodicity in the case where $1 < m < n$, take $s \in \mathscr{C}(s^*, \mathscr{M}^*)$, and let $i^*_{j_1}, i^*_{j_2} \in \mathscr{M}^*$ and $t_1 \in \mathscr{N} \setminus \mathscr{M}^*$. Three successive calls to move M_2 can generate S_1, S_2, and S_3 with $S_1(t_1) = s(i^*_{j_1})$, $S_2(t_1) = s(i^*_{j_2})$, and $S_3(t_1) = s(t_1)$. Therefore $\mathrm{P}(S_2 = s \mid S_0 = s) \wedge \mathrm{P}(S_3 = s \mid S_0 = s) > 0$.

By irreducibility and aperiodicity, the above Markov chain has a unique stationary distribution on $\mathscr{C}(s^*, \mathscr{M}^*)$. By symmetry of the transition kernel, i.e., $\mathrm{P}(S_1 = s \mid S_0 = r) = \mathrm{P}(S_1 = r \mid S_0 = s)$ for all $s, r \in \mathscr{C}(s^*, \mathscr{M}^*)$, it follows that this stationary distribution must be the uniform one. □

Proof of Lemma 5.2. We have $q_W(w \mid w_0) = \int_0^1 q_{W,\Lambda}(w, \lambda \mid w_0)\, \mathrm{d}\lambda$, with

$$q_{W,\Lambda}(w, \lambda \mid w_0) = n!\, \mathbb{1}_\Delta\left(\frac{w - (1-\lambda)w_0}{\lambda}\right) \lambda^{-n}\, g(\lambda) = n!\, \mathbb{1}_\Delta(w)\, \lambda^{-n}\, g(\lambda)\, \mathbb{1}_{(1-\delta(w;w_0),1)}(\lambda),$$

for $\lambda \in (0, 1)$ and $w \in \Delta$. □

References

Alvo, M. & Yu, P. L. H. (2014). *Statistical methods for ranking data.* Frontiers in Probability and the Statistical Sciences. Springer, New York.

Diaconis, P. (1988). *Group representations in probability and statistics.* Institute of Mathematical Statistics Lecture Notes—Monograph Series, 11. Hayward, CA: Institute of Mathematical Statistics.

Feuerverger, A., He, Y. & Khatri, S. (2012). Statistical significance of the netflix challenge. *Statist. Sci.* **27**, 202–231.

Fligner, M. A. & Verducci, J. S. (1986). Distance based ranking models. *J. Roy. Statist. Soc. Ser. B* **48**, 359–369.

Gunawardana, A. & Shani, G. (2009). A survey of accuracy evaluation metrics of recommendation tasks. *J. Mach. Learn. Res.* **10**, 2935–2962.

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics* **1**, 265–283.

Hoff, P. D., Niu, X. & Wellner, J. A. (2014). Information bounds for gaussian copulas. *Bernoulli* **20**, 604–622.

Lee, D. D. & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791.

Lee, J., Sun, M. & Lebanon, G. (2012). PREA: personalized recommendation algorithms toolkit. *J. Mach. Learn. Res.* **13**, 2699–2703.

Lemire, D. & Maclachlan, A. (2007). Slope one predictors for online rating-based collaborative filtering **5**.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer Series in Statistics. New York: Springer, 2nd ed.

Paterek, A. (2007). Improving regularized singular value decomposition for collaborative filtering .

Salakhutdinov, R. & Mnih, A. (2007). Probabilistic matrix factorization. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07. USA: Curran Associates Inc.

Salakhutdinov, R. & Mnih, A. (2008). Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08. New York, NY, USA: ACM.

Segers, J., van den Akker, R. & Werker, B. J. M. (2014). Semiparametric gaussian copula models: Geometry and efficient rank-based estimation. *Ann. Statist.* **42**, 1911–1940.

Sun, M., Lebanon, G. & Kidwell, P. (2012). Estimating probabilities in recommendation systems. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61**, 471–492.

Yu, K., Zhu, S., Lafferty, J. & Gong, Y. (2009). Fast nonparametric matrix factorization for large-scale collaborative filtering. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09. New York, NY, USA: ACM.

Zhu, M. (2014). Making personalized recommendations in e-commerce. In *Statistics in action*. CRC Press, Boca Raton, FL, pp. 259–268.

Département de mathématiques, Université du Québec à Montréal, 201, avenue Président-Kennedy, Montréal (Québec) H3X 2Y7, Canada
E-mail address: guillotte.simon@uqam.ca

Département de mathémathiques et de statistique, Université de Montréal, Pavillon André-Aisenstadt 2920, chemin de la Tour, Montréal (Québec) H3T 1J4, Canada
E-mail address: perronf@dms.umontreal.ca

Institut de statistique, biostatistique et sciences actuarielles (ISBA), Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium
E-mail address: johan.segers@uclouvain.be