Multivariate Spearman's rho for rank aggregation

Justin Bedő
IBM Research - Australia
and the University of Melbourne
bedoj@unimelb.edu.au

Cheng Soon Ong
NICTA Canberra, Australia
and the Australian National University
chengsoon.ong@anu.edu.au

Abstract

We study the problem of rank aggregation: given a set of ranked lists, we want to form a consensus ranking. Our main contribution is the derivation of a non-parametric estimator for rank aggregation based on multivariate extensions of Spearman's ρ, which measures correlation between a set of ranked lists. Multivariate Spearman's ρ is defined using copulas, and we show that the geometric mean of normalised ranks maximises multivariate correlation. Motivated by this, we propose a weighted geometric mean approach for learning to rank which has a closed form least squares solution. When only the best or worst elements of a ranked list are known, we impute the missing ranks by the average value, allowing us to apply Spearman's ρ. Finally, we demonstrate good performance on the rank aggregation benchmarks MQ2007 and MQ2008.

1 Introduction

This paper considers a novel formulation of rank aggregation based on multivariate extensions to Spearman's ρ. For a set of n objects from the domain Ω, we are given a set of d experts that rank these objects providing ranks $R_1, \ldots, R_d$. Each rank is a permutation of the n objects, and can be represented as a vector of unique integers from 1 to n. The problem of rank aggregation is to construct a new vector R that is most similar to the set of d ranks provided by the experts. In this paper we use Spearman's correlation ρ; instead of decomposing the association into a combination of pairwise similarities, $\rho(R,R_1), \rho(R,R_2), \ldots, \rho(R,R_d)$, we directly maximise the multivariate correlation

$$R^* = \arg\max_R \rho(R, R_1, R_2, \ldots, R_d).$$

Spearman's correlation is proportional to $Q(C,\pi)$, where Q is a concordance measure and C and π are copulas [1]. In section 2, we generalise the definition of concordance Q to the multivariate case, allowing us to define multivariate multivariate Spearman's ρ in an analogous way to the bivariate ρ as defined in [3]. We show that the geometric mean of a set of normalised ranks maximises multivariate Spearman's ρ in section 3. This motivates our method which finds a setting of weights that maximise multivariate Spearman's ρ for a specific target (supervised rank aggregation). In particular, we handle the case of extreme ranks where there are n distinct items, and d experts, but each expert only labels the top-k or bottom-k items, where k can be different for each expert.

2 Spearman's rho via multivariate concordance and copulas

Let R_1 and R_2 be ranking functions, which are bijections mapping n elements x in the domain Ω to $\{\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}\}$ (i.e., bijections to $(0,1)$). Similar to the approach of Pearson's correlation for the measure of dependence, Spearman's ρ is a measure of correlation between ranks:

$$\rho_n = \frac{\sum_x (R_1(x) - \bar{R}_1)(R_2(x) - \bar{R}_2)}{\sqrt{\sum_x (R_1(x) - \bar{R}_1)^2 \sum_x (R_2(x) - \bar{R}_2)^2}} = 1 - \frac{6 \sum_x (R_1(x) - R_2(x))^2}{n(n^2-1)}, \tag{1}$$

where $\bar{R}_j := \frac{1}{n}\sum_x R_j(x)$ for $j = 1,2$ are the empirical means of the respective random variables. This is equivalent to applying Pearson's correlation to the ranks.

To generalise ρ to the multivariate case, we use the copula form of Spearman's that measures the concordance between two multivariate distributions. A d copula is a function $C\colon [0,1]^d \longrightarrow [0,1]$ that is a joint distribution on uniformly distributed random variables, which directly captures the dependence of continuous random variables [2]. The theorem below directly generalises the concept of bivariate concordance in [3].

Theorem 1 (Multivariate concordance) *Let $(X_1,\dots,X_d)$ and $(Y_1,\dots,Y_d)$ be two independent d-vectors with joint distributions $C_X(F(x))$ and $C_Y(F(y))$ where $F(x) = F_1(x_1),\dots,F_d(x_d)$ and $F(y) = F_1(y_1),\dots,F_d(y_d)$ are the marginal distributions, and C_X, C_Y are the respective d copulas. Then the concordance function Q is given by*

$$\begin{aligned} Q(C_X,C_Y) :=& P\left[\prod_{j=1}^{d}(X_j - Y_j) > 0\right] - P\left[\prod_{j=1}^{d}(X_j - Y_j) < 0\right] \\ =& 2^d \int_{[0,1]^d} C_X(v)\,\mathrm{d}C_Y(u) - 1 \end{aligned}$$

where $u = F(x)$ and $v = F(y)$.

In terms of the concordance function, Spearman's ρ is given by the concordance between the copula C and the independent copula π. There are several possible generalisations of Spearman's ρ to the multivariate case, all of which are equivalent in the bivariate case. By substituting the d copula C and the independent copula[1] π into theorem 1, we obtain the generalisation ρ_2 in [7].

$$\rho = h(d)Q(C,\pi) = h(d)\left[2^d \int_{[0,1]^d} \pi(u)\,\mathrm{d}C(u) - 1\right], \quad \text{where } h(d) = \frac{1}{Q(M,\pi)} = \frac{d+1}{2^d - (d+1)}. \tag{2}$$

The normalisation factor $h(d)$ is a scaling that ensures that the value of ρ is in the interval $[-1,+1]$. In contrast to [7], we obtain our expression using the principle of multivariate concordance.

3 Optimal rank aggregation with Spearman's rho

Since the copula is a cumulative distribution function (CDF), the empirical version is defined in the same way as an empirical CDF. The d dimensional empirical copula for n objects given by

$$C_n(u) = \frac{1}{n}\sum_x \prod_{j=1}^{d} \mathbf{1}\left(R_j(x) \leqslant u_j\right), \tag{3}$$

where $R_1(x),\dots,R_d(x)$ is the rankings of the d experts. Plugging the empirical copula (3) expression into Spearman's ρ (2), and observing that integrating the product over the copula is the product of the ranks [7], we obtain an empirical expression for multivariate Spearman's correlation:

$$\rho_n(R_1,\dots,R_d) = h(d)\left[\frac{2^d}{n}\sum_x \prod_{j=1}^{d} R_j(x) - 1\right]. \tag{4}$$

We are now in a position to derive the deceptively simple result: the ranking R that maximises correlation with a given set of ranks $\{R_1,\dots,R_d\}$ is given by the geometric mean of $R_1,\dots,R_d$.

Theorem 2 *Let $\{R_1,R_2,\dots,R_d\}$ be a set of ranks with common domain Ω and σ be a ranking function. Then*

$$\arg\max_{R\in \operatorname{codom}\sigma} \rho_n(R,R_1,R_2,\dots,R_d) = \sigma\left(\prod_{j=1}^{d} R_j\right).$$

Note that $\sigma\colon \Omega \to (0,1)$ is simply a ranking function that is invariant to scale and dependent only on the ordering of elements.

[1] The copulas π and M are the Fréchet–Hoeffding bounds corresponding to independence and full dependence respectively.

4 Supervised learning to rank with a weighted mixtured of experts

As a result of theorem 2 we have a way of finding the ranking (according to some rank generator) that is closest to a target set of ranks. Consider the learning problem where we have a ranking L which comprise our labels, and a set of d experts $\{R_j\}$. During training, we would like to find a weighting of the input rankings such that it gives the label. Given a target ranking L, we would like to optimise the weights ω,

$$\min_{\omega} \rho_n(L, R_1^{\omega_1}, R_2^{\omega_1}, \ldots, R_n^{\omega_d})$$

Here we have introduced weights ω over each rank to control the influence of each rank over the final consensus rank; the intuition here is that ranks with $\omega_i > 1$ are "replicated" with more influence, which is easy to see when ω_i are natural numbers. For example, a weight of 2 would mean the ranked list has appeared twice in the calculation of the consensus rank. While it is convenient to have integer weights for interpretability, the weights ω could be any real number in general. In the following, we consider $\omega \in \mathbb{R}^n$.

Instead of performing this high-dimensional optimisation, we decompose it into a pairwise (bivariate) comparison between the label L and the weighted geometric mean, where we now explicitly show the fact that the ranks are a function of the n objects x

$$\min_{\omega} \sum_x \rho_n(L(x), \sigma(R_1^{\omega_1} \otimes R_2^{\omega_2} \otimes \cdots \otimes R_d^{\omega_n})(x))$$

where the notation $\otimes$ indicates the product operator. Observe that we have used theorem 2 to convert the d dimensional problem into the product of ranks R_j and the Spearman's correlation above is only two dimensional. For bivariate Spearman's ρ, this can be expressed in terms of the squared difference (1). We further assume that σ is the identity mapping to simplify the problem, giving us:

$$\min_{\omega} \sum_x \left(L(x) - R_1^{\omega_1}(x) R_2^{\omega_2}(x) \ldots R_d^{\omega_n}(x)\right)^2. \tag{5}$$

The objective (5) minimises the distance between the label ranks and the weighted expert ranks.

4.1 Least squares method on logarithm of ranks

Recall that we consider normalised ranks (divided by $n+1$). By using the logarithm identity, we convert the power scaling in (5) into a multiplicative scaling. Our algorithm is:

1. Extend incomplete ranks $\{R_i\}$ to $\{R_i'\}$ by imputing the average missing value such that $\mathrm{Dom}\, R_i' = \mathrm{Dom}\, L$;
2. Convert to log-ranks $r_i' = \log \circ R_i'$ and $l = \log \circ L$;
3. Learn weights ω by minimising

$$\sum_x \left(l(x) - \sum_{j=1}^{d} \omega_j r_j'(x) \right)^2, \qquad \text{where the outer sum is over the } n \text{ examples } x.$$

A log transformation of the ranks is used as it naturally encodes the weights as a power scaling in the framework of theorem 2, i.e., the weighted consensus rank is given by $\prod_i r_i'(x)^{\omega_i}$. Note that this is still solving (5) as we are optimising Spearman's ρ, which is sensitive only to ordering, and therefore though the final weights are different ρ is maximised via (1).

4.2 Benchmarking on LETOR 4.0

We tested our method on the MQ2007-agg and MQ2008-agg list aggregation benchmarks [6]. The goal in these challenges is to aggregate 21 and 25 different rankers respectively over a set of query-document pairs. Each dataset has 5 pre-defined cross-validation folds with each fold providing a training, testing and validation dataset (60%/20%/20%). We trained our model on the training set and validation set and tested on the testing set.

In the following we consider two types of experts: either experts $\{R_j\}$ are top-k experts, that is they only rank the best k samples from Ω, or experts are bottom-k experts, that is they identify the worst

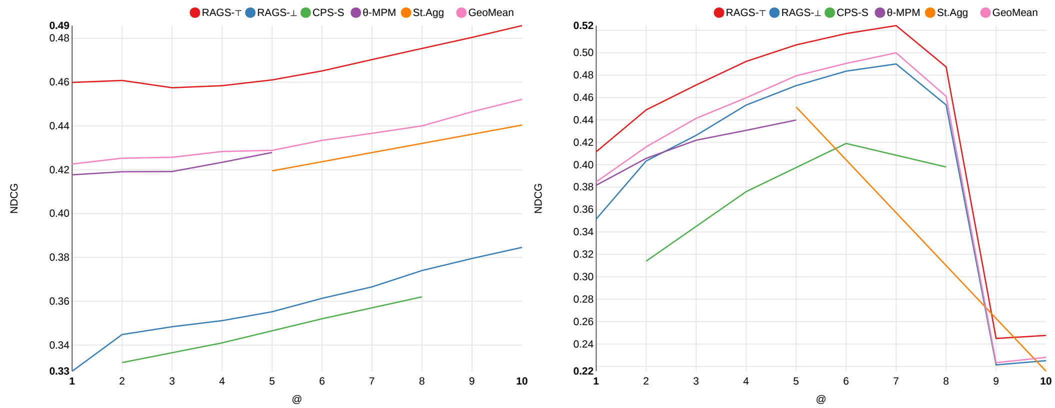

Figure 1: Results on MQ2007-agg (left) and MQ2008-agg (right): NDCG@k. Our method is labelled RAGS-$\top$ and RAGS-$\perp$ corresponding to top and bottom imputation respectively. The results for CPS-S was the best reported in [5]. The results of θ-MPM was the best among the reported results in [8] from CPS, SVP, Bradley-Terry model, and Plackett-Luce model. The results of St.Agg was the best among the reported results in [4] and was the best among MCLK, SVP, Plackett-Luce model, θ-MPM and RRF.

k samples from Ω. We call our proposed method RAGS-$\top$ and RAGS-$\perp$ respectively. We assume that the ranked documents in the benchmark datasets are either top-k or bottom-k respectively, with potentially different numbers of documents k labelled by each expert. Ties are given the average rank of tied documents.

To evaluate the agreement, we use the standard evaluation tool from the LETOR website[2], which implements the Normalised Discounted Cumulative Gain (NDCG). In fig. 1 (left), we see that our approach RAGS-$\top$ performs better than all other methods at any selection size on the MQ2007-agg dataset. Indeed, we also perform better than [5] where the best result uses a coset-permutation distance based stagewise (CPS) model with Spearman's ρ in a probabilistic model. Recall that our approach considers the multivariate Spearman's ρ whereas [5] uses bivariate Spearman's ρ in a pairwise fashion. For MQ2008-agg (fig. 1 (right)) both our approaches perform better than all other methods.

To tease apart the effect of imputing missing ranks and the effect of weighting the experts, we compared our proposed method with and without training (uniform weights, denoted GeoMean). Our proposed approach outperforms the geometric mean, which is a good sanity check, and also Borda count.

5 Conclusion

We propose an approach for learning weights between experts for the task of rank aggregation. By generalising the derivation of concordance functions, we obtain an expression for multivariate Spearman's ρ. Furthermore we show that the geometric mean of the expert ranks is the optimal aggregator under Spearman's correlation. Motivated by this, our method solves a least squares estimation problem for logarithmic normalised ranks to find optimal weights. We propose an imputation method for completing top-k ranked lists that allows us to apply Spearman's ρ to aggregate ranks from partial lists. Surprisingly, our weighted geometric mean shows state of the art results on benchmark datasets, without the need for tuning hyperparameters or expensive computation. The simplicity of our model makes it easier to interpret, and the weights give a direct estimate of the influence of each expert. This problem has wide applications to ensemble learning, voting, text mining, recommender systems and bioinformatics.

[2] `http://research.microsoft.com/letor`

References

[1] Christian Genest and Anne-Catherine Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368, 2007.

[2] Harry Joe. *Dependence Modeling with Copulas*. CRC Press, 2014.

[3] Roger B. Nelsen. *An Introduction to Copulas*. Springer, second edition, 2006.

[4] Shuzi Niu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. Stochastic rank aggregation. In *UAI*, 2013.

[5] Tao Qin, Xiubo Geng, and Tie-Yan Liu. A new probabilistic model for rank aggregation. In *NIPS*, 2010.

[6] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. LETOR: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval Journal*, 2010.

[7] Friedrich Schmid and Rafael Schmidt. Multivariate extensions of spearman's rho and related statistics. *Statistics & Probability Letters*, 77:407–416, 2007.

[8] Maksims N. Volkovs and Richard S. Zemel. A flexible generative model for preference aggregation. In *WWW*, 2012.