Technical Report
EHU-KZAA-TR-1-2011

Universidad del País Vasco
Euskal Herriko Unibertsitatea

UNIVERSITY OF THE BASQUE COUNTRY
Department of Computer Science and Artificial Intelligence

Learning Probability Distributions over Permutations by Means of Fourier Coefficients

Ekhine Irurozki, Borja Calvo, Jose A. Lozano

January 2011

San Sebastian, Spain
http://www.ccia-kzaa.ehu.es/

Learning Probability Distributions over Permutations by Means of Fourier Coefficients

Ekhine Irurozki, Borja Calvo, Jose A. Lozano
ekhine.irurozqui@ehu.es, borja.calvo@ehu.es, ja.lozano@ehu.es
Intelligent Systems Group, University of the Basque Country, Spain
http://www.sc.ehu.es/isg

Abstract

A large and increasing number of data mining domains consider data that can be represented as permutations. Therefore, it is important to devise new methods to learn predictive models over datasets of permutations. However, maintaining models, such as probability distributions, over the space of permutations is a hard task since there are $n!$ permutations of n elements. Recently the Fourier transform has been successfully generalized to functions over permutations and offers an attractive way to represent uncertainty over the space of permutations. One of its main advantages is that the Fourier transform compactly summarizes approximations to functions by discarding high order marginals information. Moreover, a lately proposed framework for making inference completely in the Fourier domain has opened new doors for efficiently reasoning over a space of permutations. In this paper, we present a method to learn a probability distribution that approximates the generating distribution of a given sample of permutations. Particularly, this method learns the Fourier domain information representing this probability distribution.

1 Introduction

Permutations and orders appear in a wide variety of real world combinatorial problems such as multi object tracking, structure learning of Bayesian networks, election, etc. Particularly, in the machine learning domain, the application of permutations which is receiving the most attention by the community is that of ranking [2], [4].

Exact probability representation over the space of permutations of n elements is intractable, in general, with the exception of very small n, since this space has size $n!$. However, different simplified models for representing or approximating probability distributions over a set of permutations can be found in the literature [3], [5], [7]. The most basic approach consists of storing the first order marginals [5]. However, the accuracy of the obtained approximation is limited to very smooth distributions. Two well-known approaches from the distance based exponential family models are the Mallows and Generalized Mallows Model [3], [14]. They compactly summarize distributions even when dealing with permutations of large n. While the Mallows Model defines a two parameter probability distribution, the central or consensus ranking and a spread

parameter, the Generalized Mallows Model considers n parameters, which are the consensus ranking and an $n-1$ spread parameters.

Another way to represent probability distributions over permutations is the Fourier-based approach. This is based on a generalization of the well-known Fourier transform in the real line for permutations. Permutations form an algebraic group under the composition operation, also known as the *symmetric group*, so we will use both expressions, permutations and symmetric group, interchangeably along this paper. Although the use of the Fourier transform for representing functions over permutations is not novel, recently this topic has once again come to the attention of the researchers. This is partly due to a framework recently provided by [7] and [13] which allows to carry out inference tasks entirely in the Fourier domain. Moreover, new concepts have been introduced, such as the probability independence over permutations in the Fourier domain [8], [6], [9]. Furthermore, the Fourier representation of functions has been also used in other data mining contexts. Particularly, in [12] it is shown that by using the Fourier analysis, some kernels can be efficiently computed.

The Fourier transform on the symmetric group decomposes a given function over the space of permutations of n elements into $n!$ complex numbers. These complex numbers that result from the transformation of the function into the Fourier domain are called Fourier coefficients. The Fourier coefficients can be inverse-transformed into the original function. The idea of bandlimiting functions (the use of a limited number of Fourier coefficients) over the real line to approximate functions has its equivalent in the Fourier transform over the symmetric group. In addition, this approximation in the context of probability distributions over permutations has a very interesting property: by bandlimiting a probability distribution, the higher order marginal probabilities are discarded. As a bandlimited approximation of a function over the real line smooths the original signal by discarding high frequency terms, its analogous in the symmetric group smooths the probability distribution, bringing it closer to the uniform distribution. Therefore, this approach approximates smooth distributions more accurately than sharp ones.

In this paper, we focus on the problem of learning the generating distribution of a given sample of permutations. We present a method for learning a limited number of Fourier coefficients that represent this probability distribution. Particularly, we propose a constrained formulation for finding the Fourier coefficients that maximize the likelihood of the sample.

The first attempt to learn a probability distribution by means of the Fourier coefficients was presented in [10]. The authors concentrated on getting a consensus ranking and a probability distribution under constrained sensing, when the available information is limited to the first order marginals. However, to the best of our knowledge, this work is the first attempt to do it in a general way.

The rest of the paper is organized as follows. The next section introduces the basis of the Fourier transform over permutations, including intuitive ideas about group representation theory. In Section 3 we detail how we formulate the maximum likelihood method in order to find the most likely Fourier coefficients for a given sample of permutations. Section 4 presents the experimental results of several tests over different kinds of probability distributions. In Section 5, we conclude the paper.

2 The Fourier Transform on the Symmetric Group

2.1 Preliminary Concepts

Formally, a permutation is defined as a bijection of the set {1,...,n} into itself which can be written as $\sigma = [\sigma(1), ..., \sigma(n)]$. S_n designates the set of all permutations of n elements.

The Fourier transform on the real line has been successfully generalized to several spaces of functions. We will focus our attention on the generalization to the symmetric group. Analogously to the operation on the real line, the Fourier transform over permutations decomposes a given function as a linear combination of a set of orthogonal basis functions. The elements playing the role of basis functions in the group theory generalization of the Fourier transform are the *irreducible representations*.

Since it is out of the scope of this paper to be a proper tutorial neither on the Fourier transform on the symmetric group nor in representation theory, we just give some ideas for intuition and refer the interested reader to [1] and [15] for further discussion. A representation $\rho : S_n \rightarrow \mathbb{C}^{d_\rho \mathrm{x}\ d_\rho}$ is a linear map from a group, such as S_n, that associates each element $s \in S_n$ with an invertible complex matrix $\rho(s)$ such that for every s, $t \in S_n$

$$\rho(st) = \rho(s)\rho(t). \tag{1}$$

The matrices in the image of ρ are called *representation matrices* and are said to be irreducible if they cannot be written as the direct sum of two representations. A set of basis for the Fourier transform in the symmetric group is given by l_n irreducible representation matrices. An order can be defined among those l_n irreducible representation matrices so they can be indexed as $\lambda = 1, ..., l_n$. The construction of irreducible representation matrices can be done in different ways by considering different basis vector spaces. Like other authors such as [7] we have chosen to construct these matrices with respect to the Gel'fand-Tsetlin (GZ) basis. An interesting property of the irreducible representation matrices constructed with respect to this basis is that they are real valued. The dimension of each irreducible representation matrix is denoted as d_ρ. Table 1 shows a set of irreducible representation matrices of S_3.

2.2 Fourier Transform and Inverse Fourier Transform

Now that the basic concepts have been introduced, we can define the Fourier transform on the symmetric group. Let $f : S_n \rightarrow \mathbb{R}$ be a function on the set of permutations of n elements, S_n, and ρ_λ the λ-th irreducible representation. The Fourier transform of f is the set of Fourier coefficients at the irreducible representations ρ_λ, $\hat{f} = \{\hat{f}_{\rho_\lambda}\}_{\lambda=1}^{l_n}$, where each coefficient $[\hat{f}_{\rho_\lambda}]_{ij}$ is computed as:

$$[\hat{f}_{\rho_\lambda}]_{ij} = \sum_{\sigma \in S_n} f(\sigma)[\rho_\lambda(\sigma)]_{ij} \tag{2}$$

where σ indexes the permutations in S_n. Note that, since the irreducible representation matrices are real valued, the obtained Fourier transforms are also real valued matrices.

σ	$\rho_1(\sigma)$	$\rho_2(\sigma)$	$\rho_3(\sigma)$
$(1,2,3)$	$[1]$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$[1]$
$(2,1,3)$	$[1]$	$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$	$[-1]$
$(1,3,2)$	$[1]$	$\begin{bmatrix} 1/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{bmatrix}$	$[-1]$
$(3,2,1)$	$[1]$	$\begin{bmatrix} 1/2 & -\sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \end{bmatrix}$	$[-1]$
$(3,1,2)$	$[1]$	$\begin{bmatrix} -1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{bmatrix}$	$[1]$
$(2,3,1)$	$[1]$	$\begin{bmatrix} -1/2 & \sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \end{bmatrix}$	$[1]$

Table 1: A set of irreducible representation matrices of S_3

Naturally, the function can be recovered from the Fourier domain by the Inversion Theorem, which can be expressed as:

$$f(\sigma) = \frac{1}{|S_n|} \sum_{\lambda} d_{\rho_\lambda} Tr[\hat{f}_{\rho_\lambda}^{\ T} \cdot \rho_\lambda(\sigma)] \tag{3}$$

where *"Tr"* refers to the trace, i.e. the sum of the diagonal elements of the matrix that results from the product of $[\hat{f}_{\rho_\lambda}^{\ T} \cdot \rho_\lambda(\sigma)]$.

As for the operation on the real line, several efficient algorithms have been designed and implemented for computing the fast Fourier transform on the symmetric group [11]. In this way, this operation that would naively run in $O(n!^2)$, can be computed in $O(n^2\ n!)$.

2.3 Representing Probability Distributions by Means of the Fourier Coefficients

In order to store the probability distribution of the set of permutations of n elements, the total number of required Fourier coefficients is $n!$ (which is, in fact, the total number of different permutations). However, one of the most attractive properties of transforming a probability distribution to the Fourier domain is the way in which probability distributions can be approximated. Since it is not our aim to go into detailed discussion, we will just drop some intuitive ideas. Further discussion can be found in [1].

As we have already stated, the irreducible representations can be indexed. There exists an order that corresponds to the one that can be defined over the marginal probabilities of a distribution in the sense that the Fourier transform at low index irreducible representations contains the low order marginal probabilities and the Fourier transform at high irreducible representations contains the high order marginal probabilities. Moreover, the $(k-1)$-th order marginal information is kept in the Fourier transform at the first k irreducible representations, $\{\hat{f}_{\rho_1}, ..., \hat{f}_{\rho_k}\}$. Therefore, to approximate a distribution by keeping its first $(k-1)$-th order marginal probabilities, it is enough to save the Fourier transform at the first k irreducible representations and discard the rest. While

this bandlimited model correctly approximates smooth distributions, it should be noted that its accuracy decreases as the probability distribution gets sharper.

Another interesting property of representing probability distributions by means of the Fourier coefficients is that storing the Fourier transform at the first k irreducible representations is more efficient than storing the first $k-1$ order marginals. Moreover, given the coefficients $\{\hat{f}_{\rho_1}, ..., \hat{f}_{\rho_k}\}$ and some constant real valued matrices, the first $(k-1)$-th order marginal probabilities can be computed in polynomial time. For a detailed description on the computation of such matrices see [7].

3 Learning Probability Distributions over the Fourier Domain

In this section we describe our proposed formulation for learning the Fourier coefficients from a given sample of permutations. The inverse Fourier transform in equation 3 defines the probability distribution in terms of the Fourier coefficients. Our proposal consists of finding the Fourier coefficients that maximize the likelihood of this function given a sample of permutations. However, due to the exponential nature of the symmetric group, we are interested in obtaining a bandlimited distribution which considers the $(k-1)$-th lowest marginal probabilities. In order to learn such an approximation, the Fourier coefficients in the formulation are restricted to the ones in the Fourier transform corresponding to the first k irreducible representations $\{\hat{f}_{\rho_1}, ..., \hat{f}_{\rho_k}\}$.

Maximizing the likelihood of a sample $\{\sigma_1, ..., \sigma_t\}$ given the model in equation 3 means solving the following nonlinear optimization problem:

$$\begin{aligned}(\hat{f}^{mle}_{\rho_{\lambda_1}}, ..., \hat{f}^{mle}_{\rho_{\lambda_k}}) &= \underset{\hat{f}_{\rho_1},...,\hat{f}_{\rho_k}}{\arg\max}\ \hat{\mathcal{L}}(\sigma_1, ..., \sigma_t | \hat{f}_{\rho_1}, ..., \hat{f}_{\rho_k}) \\ &= \underset{\hat{f}_{\rho_1},...,\hat{f}_{\rho_k}}{\arg\max} \prod_{i=1}^{t} \Big(\frac{1}{|S_n|} \sum_{\lambda=1}^{k} d_{\rho_\lambda} Tr[\hat{f}_{\rho_\lambda}^{\ T} \cdot \rho_\lambda(\sigma_i)] \Big)\end{aligned}$$

Unfortunately, not every set of Fourier coefficients leads to a valid probability distribution. Actually, maximizing the likelihood in this formulation will lead to a function whose values do not sum 1 and does not actually correspond to a probability distribution. Compactly describing the coefficients of a valid distribution is still an open problem [7]. We will restrict the search space by the addition of some constraints that forbid searching in regions of the space where no coefficient representing a valid distribution can be found. We have considered two kinds of constraints.

The first kinds of constraints ensure that the Fourier coefficients take values between the maximum and the minimum values of the irreducible representations that multiply it. It can be seen in equation 2 that each Fourier coefficient is the result of the product of a probability value and an irreducible representation term. Since the probability values range in the interval [0,1], the bounds of the Fourier coefficients for a valid probability distribution are constrained as follows:

$$min_\sigma([\rho_\lambda(\sigma)]_{ij}) \leq [\hat{f}_{\rho_\lambda}]_{ij} \leq max_\sigma([\rho_\lambda(\sigma)]_{ij})$$

In addition, by setting the constraint which makes the trivial coefficient (i.e. the Fourier transform at the first irreducible representation) equal 1, we ensure that the sum of the probability values at of the permutations is 1. However, this does not guarantee a positive distribution. The number of constraints is $2 * c + 1$, where c is the number of estimated Fourier coefficients, $c = \sum_{\lambda=2}^{k} d_\lambda^2$.

The second kinds of constraints ensure a positive probability for each permutation in the sample. The number of this second kinds of constraints is equal to the number of different elements in the sample.

The Fourier coefficients obtained by maximizing the likelihood restricted to these constraints correspond to a distribution whose sum is guaranteed to be 1. However, this does not ensure a valid probability distribution, so it is possible to have negative 'probabilities'. In that case we perform a normalization process. Let m be the minimum probability value associated to a permutation. This process consists of adding to every value of the probability distribution the absolute value of m and normalizing it. Note that if we add a second kind constraint for each $\sigma \in S_n$ the estimated distribution is valid. The experimental section describes several experiments in this scenario. However, it should be pointed out that this framework is intractable with the exception of very small values of n.

4 Experiments

In this section we will show the performance of the proposed formulation for learning probability distributions. Our aim is to demonstrate two points: (i) we will show that the learned distribution is significantly better than any random distribution and (ii) we will see how the algorithm behaves when the search space is restricted to the space of valid probability distributions. Particularly, we will show the accuracy of the estimated distributions as the sample size grows and higher order marginals are learned.

4.1 Experimental setup

In order to evaluate our approach on the two above described statements we have designed two different experimental frameworks.

In the first one, the next procedure is followed. First of all, a probability distribution is randomly generated. Then, departing from this distribution, several permutation samples are generated. For these samples, the proposed algorithm learns the Fourier coefficients and the distributions corresponding to these coefficients are calculated. Finally, the Kullback-Leibler divergence between the reference and the resulting estimated distributions are calculated.

In order to prove our first point and see that the learned distributions are a significantly better approximation than a random distribution, we propose a comparison test based on Monte Carlo technics. The test consists in sampling a large number of random distributions and measuring the Kullback-Leibler divergence between the reference and each of the random distributions.

The reference and the random distributions are generated by sampling a Dirichlet distribution. In this way, the generation of each distribution requires $n!$ hyper-parameters $\alpha_1, ..., \alpha_{n!}$. We have set these hyper-parameters, for every

distribution, as $\alpha_1 = \alpha_2 = ... = \alpha_{n!} = \alpha$, where α is uniformly drawn from the interval [0.05, 0.25].

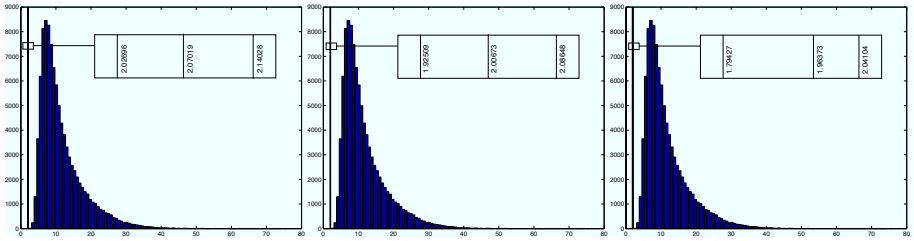
The hyper-parameter α in the Dirichlet distribution is related with the smoothness of the generated distribution. The smaller it is, the sharper the distribution. It seems reasonable thinking that, by learning higher order marginals, it will be possible to more accurately approximate the reference distribution. In order to prove this intuition, the Fourier coefficients corresponding to three different marginals have been learned for each test, that is, the ones at the second, third and fourth irreducible representations.

This first experimental framework has been set as follows. The tests have been made over the set of permutations of 5, 6 and 7 elements, S_5, S_6 and S_7 respectively. For each different S_n, three sample sizes are defined which are 5, 10 and 25 % of $n!$ for S_5 and S_6 and 1, 5 and 10% of $n!$ for S_7. Also, for each n and sample size, ten different samples are randomly generated and the average results are computed. The number of random distributions for the comparison test is 100,000, and their divergences with the reference distributions are used to draw a histogram.

The second experimental framework has been designed to show how this algorithm behaves when the search space is restricted to the space of valid probability distributions. In order to restrict the search space, some constraints have been added to the model described in the above section. Particularly, these constraints ensure that every value of the probability distribution is positive. Although this approach is not efficient for large values of n, we find these experiments particularly illustrative to see how the estimation algorithm behaves as the sample size and the number of learned Fourier coefficients increase. In the same way as in the first framework, a reference distribution is generated and several samples are obtained from it. Then, the Fourier coefficients for those samples are learned and the distributions corresponding to those coefficients are obtained. Finally, the Kullback-Leibler divergences between the estimated and the reference distributions are given.

The parameters of this second framework have been set as follows. The experiments have been performed for S_5. The Fourier transform of a function on S_5 consists on 120 Fourier coefficients which are grouped in seven matrices. For each sample seven different probability distributions are learned by using the coefficients corresponding to the first k irreducible representations, $k = 1..7$. We have considered two different reference distributions. The first one has also been generated by sampling a Dirichlet distribution. Its 5! parameters are equal to 0.05. The second one is a Mallows distribution with the spread parameter $\theta = 1.4$ and the central permutation $\pi_0 = [1, 2, 3, 4, 5]$. For each of them, one reference distribution is generated for which four different sample sizes are defined, 5%, 10%, 25% and 50 % of 5!. As in the previous case, for each reference distribution and sample size ten different samples are randomly generated and the given Kullback-Leibler divergence is the average of these ten runs.

The resulting constrained nonlinear optimization problems from both frameworks have been solved using MATLAB. Particularly, the *fmincon* function in the *optimization toolbox*.



(a) Sample size of 5% (b) Sample size of 10% (c) Sample size of 25%

Figure 1: Kullback-Leibler divergence between the reference and estimated distributions and the reference random distributions S_5

4.2 Results

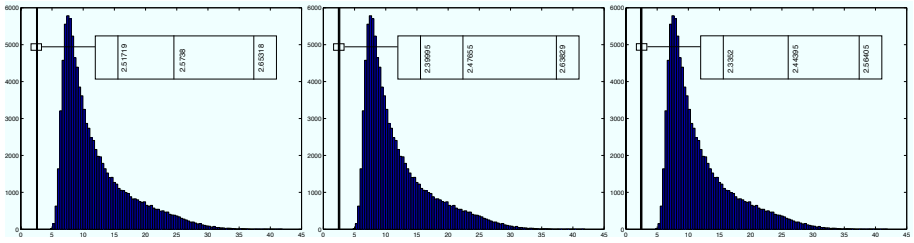
In this section we give a detailed description of the results of the experiments in the two described frameworks. The result of the experiments considering the first framework for S_5, S_6 and S_7 are shown in Figures 1, 2 and 3 respectively.

Figures 1a, 1b and 1c show the results of estimating a sample of 5%, 10% and 25% of $n!$ respectively. Particularly, these figures show the Kullback-Leibler divergence between the reference and the estimated distributions and the reference and random distributions. The first point to consider in each figure is that the divergence between the reference and the random distributions are spanned in a wide interval, being the higher concentration in the first half of the range. However, none of the random distributions is closer to the reference distribution than any of the ones obtained by learning the Fourier coefficients, which are plot with a vertical line. Note that each line represents the average divergence of ten distributions obtained from ten different samples of the same size. The estimated distributions are significantly better than any random distribution. The three lines correspond to the distributions obtained by estimating the Fourier coefficients at the first 2, 3 and 4 irreducible representations. Since the differences cannot be clearly appreciated in the plots, a zoom over them is done in the top right side of each figure. In every figure the line in the right corresponds to the estimation of the lowest order marginals ($k = 2$) and the line in the left to the estimation of highest order marginals considered ($k = 4$). This means that as the number of learned Fourier coefficients grows, the resulting distribution gets closer to the reference distribution.

Figures 2a, 2b and 2c show the results of the tests over the group S_6 using sample sizes of 5% 10% and 25% of $n!$. Similar conclusions can be drawn, since none of the random distributions is closer to the reference one than any of the distributions recovered from the learned Fourier coefficients. Here again, the larger the number of estimated Fourier coefficients, the better the approximation.

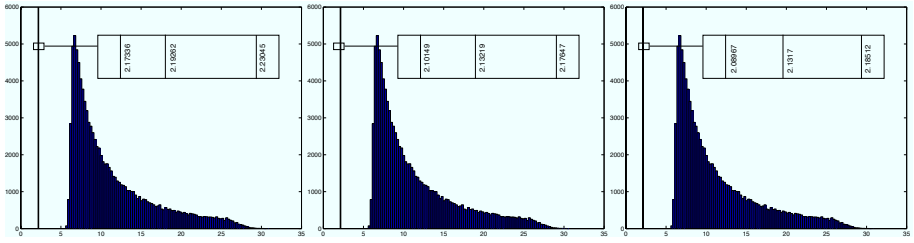
Figures 3a, 3b and 3c show quite a similar performance on the group S_7. Moreover, one can see that, as the number of elements in the set of permutations grows, the divergences between the reference and the random distributions quickly increase, while the divergence of the learned distributions are quite stable.

The results of the second experimental framework can be seen in Figure 4.



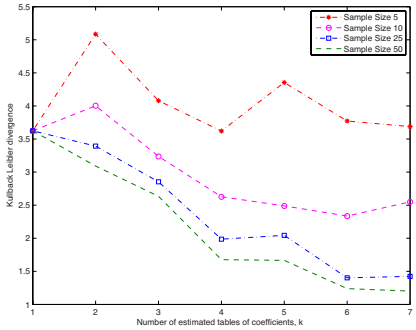
(a) Sample size of 5% (b) Sample size of 10% (c) Sample size of: 25%

Figure 2: Kullback-Leibler divergence between the reference and estimated distributions and the reference random distributions S_6

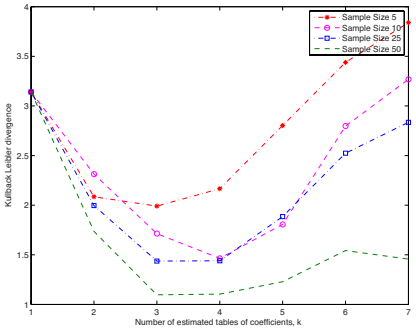


(a) Sample size of 1% (b) Sample size of 5% (c) Sample size of 10%

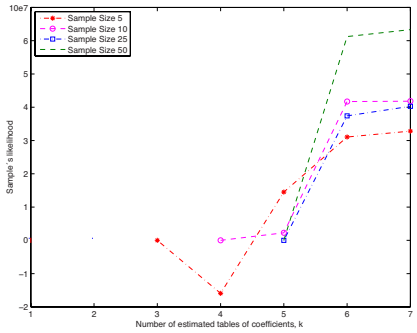
Figure 3: Kullback-Leibler divergence between the reference and estimated distributions and the reference random distributions S_7



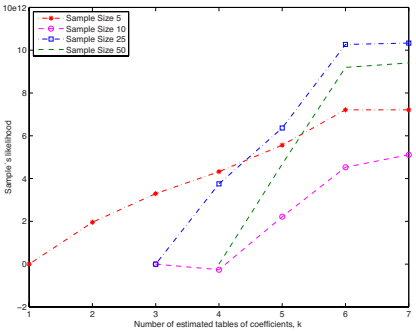
(a) Kullback Leibler divergence, Dirichlet generated distribution over S_5



(b) Kullback Leibler divergence, Mallows distribution over S_5



(c) Likelihood of the sample, Dirichlet generated distribution over S_5



(d) Likelihood of the sample, Mallows distribution over S_5

Figure 4: Kullback-Leibler divergence between the reference and the estimated distributions for different sample sizes and different number of coefficients considered in 4a and 4b. Figures 4c and 4d plot the likelihood of the samples for different sample sizes and different number of coefficients considered.

Figures 4a and 4b show the Kullback-Leibler divergence between the estimated and the reference distributions, which have been generated by sampling a Dirichlet and a Mallows distribution respectively. In both figures one can see how, as the sample size grows, the estimated distribution tends to get closer to the reference distribution. Moreover, in Figure 4a we can also see how as the number of estimated Fourier coefficients grow, the divergence between the distribution represented by these coefficients and the reference Dirichlet generated distribution tends to decrease. However, this is not the case for the samples coming from the Mallows distribution. Initially, the Kullback Leibler divergences decrease, but after $k = 3$ the divergences increase. In order to better understand the behavior of the estimated distributions it is illustrative to compare the likelihood of the samples for both estimated, e, and reference distributions, r, which is shown in Figures 4c and 4d. Particularly, these figures plot $log((r - e)/r)$ for positive values of $(r - e)$. As it can be seen in Figure 4d the likelihood of the estimated distribution from the Mallows samples quickly increases from $k = 3$. This is not the case of the Dirichlet generated distribution, Figure 4c. The main difference between both distributions is the number of parameters. While the Dirichlet

generated distribution needs $n!$ parameters in its definition, for the Mallows distribution only two parameters are required. Therefore, we can conclude that when learning the Mallows distribution an overfitting phenomenon is happening due to the fact that the number of parameters we are learning is much bigger than the number of parameters that describe the distribution.

5 Conclusions and Future work

In this paper we propose a novel method for learning probability distributions from a set of permutations. The model for representing such distributions is the Fourier-based approach. We have described a formulation that, by maximizing the likelihood function, learns the Fourier coefficients that best represent the probability distribution of a given sample.

We have tested this formulation on two different models of probability distributions over permutations. The experiments showed that the original distribution can be better approximated when the sample sizes grow and higher order marginals are learned.

Acknowledgments

This work has been partially supported by the Saiotek and Research Groups 2007-2012 (IT-242-07) programs (Basque Government), TIN2008-06815-C02-01 and projects (Spanish Ministry of Science and Innovation) and COMBIOMED network in computational biomedicine (Carlos III Health Institute). Ekhine Irurozki holds the grant BES-2009-029143 from the Spanish Ministry of Science and Innovation.

References

[1] P. Diaconis. *Group representations in probability and statistics*. Institute of Matematical Statistics, 1988.

[2] John C. Duchi, Lester W. Mackey, and Michael I. Jordan. On the consistency of ranking algorithms. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 327–334, Haifa, Israel, June 2010. Omnipress.

[3] M. A. Fligner and J. S. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society*, 48(3):359–369, 1986.

[4] Robert Gwadera and Fabio Crestani. Ranking sequential patterns with respect to significance. In Mohammed Javeed Zaki, Jeffrey Xu Yu, B. Ravindran, and Vikram Pudi, editors, *PAKDD (1)*, volume 6118 of *Lecture Notes in Computer Science*, pages 286–299. Springer, 2010.

[5] D. P. Helmbold and M. K. Warmuth. Learning permutations with exponential weights. *Journal of Machine Learning Research (JMLR)*, 10:1705–1736, July 2009.

[6] J. Huang and C. Guestrin. Riffled independence for ranked data. In *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2009.

[7] J. Huang, C. Guestrin, and L. Guibas. Fourier theoretic probabilistic inference over permutations. *Journal of Machine Learning Research (JMLR)*, 10:997–1070, May 2009.

[8] J. Huang, C. Guestrin, X. Jiang, and L. Guibas. Exploiting probabilistic independence for permutations. In *Artificial Intelligence and Statistics (AISTATS)*, April 2009.

[9] Jonathan Huang and Carlos Guestrin. Learning hierarchical riffle independent groupings from rankings. In *International Conference on Machine Learning (ICML 2010)*, Haifa, Israel, June 2010.

[10] S. Jagabathula and D. Shah. Inferring rankings under constrained sensing. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada*, pages 753–760, 2008.

[11] R. Kondor. $\mathbb{S}_n$`ob`: a C++ library for fast Fourier transforms on the symmetric group, 2006. Available at `http://www.cs.columbia.edu/~risi/Snob/`.

[12] R. Kondor and M. Barbosa. Ranking with kernels in fourier space. In *Conference on Learning Theory*, Haifa/Israel, June 2010.

[13] R. Kondor, A. Howard, and T. Jebara. Multi-object tracking with representations of the symmetric group. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, San Juan, Puerto Rico*, March 2007.

[14] M. Meila, K. Phadnis, A. Patterson, and J. Bilmes. Consensus ranking under the exponential model. In *Proceedings of the Proceedings of the Twenty-Third Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*, pages 285–294, Corvallis, Oregon, 2007.

[15] J.P. Serre. *Linear Representations in Finite Groups*. Springer, 1977.