# Detecting higher order variable interactions: A spectral analysis approach

*LXAI Workshop at NIPS 2018 Submission*

David Uminsky[1], Rosa Garza[2], Lillian González Albino[3], Sylvia Akueze Nwakanma[4], Stephen Devlin[5] and Mario Banuelos[6]

[1]University of San Francisco
[2] California State University, Monterey Bay
[3]University of Puerto Rico, Río Piedras
[4]Pomona College
[5]University of San Francisco
[6]California State University, Fresno

September 2018

We propose the use of a powerful algebraic signal processing tool from non-abelian harmonic analysis (or spectral analysis, see [1]) to address the fundamental question of how to properly attribute the response to higher order co-variate interactions. Accurate attribution can create novel features for machine learning models and similar algebraic signal processing methods have garnered increasing attention from the machine learning community [2, 3, 4]. These techniques have also been used to extract insight in ranked or partially ranked data in contexts such as committee voting behavior ([5], economics and consumer choice [6], stability of biomarkers [7] and detecting supreme court voting blocks [8].

For this paper we specifically consider a data set which consists of $m$ rows of where a fixed set of $n$ input variables for a given row are measured to be either "on" or "off," and a response, typically a continuous value, is recorded. The core questions that arise are 1) How, if any, do subsets of our $n$ variables affect the response? 2) Do subsets of our $n$ variables interact with each other in a *higher order way* that affect the response and can we detect this? The first question is generally straightforward to address but question two is largely open and much more difficult to answer.

The above setting commonly manifests itself across many applications and to demonstrate the generalizability of this method we focus on two very unrelated applications: Genetics and Basketball. In basketball, 5 players (out of a 15 member team) are on the floor at any given time during a game and total points scored or plus-minus (total points scored minus total points scored by opposing team) is a natural continuously measured response. The key questions to investigate are of the form: Is player X good or is he good because he plays alongside Lebron James? More generally, are there natural synergies between 3 players that is more effective than the same 3 players playing individually? Can we detect this higher order interaction between players from the data?

The area of genomics is our second example of data in this form. Over the past several decades, thousands of Single Nucleotide Polymorphisms (SNPs) have been associated to diseases and other complex traits [9]. Single letter mutations of genes, SNPs, are measured ("on" or "off") and a phenotypic response (e.g. cancer/no cancer or increase/decrease in hemoglobin count) is recorded. Statistical analysis typically looks for association between a phenotype and a SNP taken individually via single-locus tests, though geneticists admit this is an oversimplified approach to tackle the complexity of underlying biological mechanisms.

Interaction between SNPs, known as epistasis [9], must be considered. Unfortunately, effective epistasis detection gives rise to significant analytical and computational challenges. Two main challenges include: 1) the computational complexity of exhaustive approaches to epistasis grows exponentially and 2) many of the more traditional statistical methods increase type I error associated with too many hypothesis testing and require Bonferroni-like corrections to partially address this.

In this paper, we show that the Fourier transform over the associated irreducible representations of the symmetric group for data in this specific structure yield precise insight into the higher order interaction and affect on response variables. The key factor to the Fourier transform success in both basketball data as well as genomics is the *orthogonal* decomposition into pure higher order interactions. In the case of basketball data, we look at the lineup level play-by-play data for the entire 2015-2016 NBA season. In Figure 1 one can plainly see one large spike in the spectrum in the pure 3rd order effects space for the Golden State Warriors. This spike has "detected" the well known synergistic trio of Draymond Green, Stephen Curry and Klay Thompson.

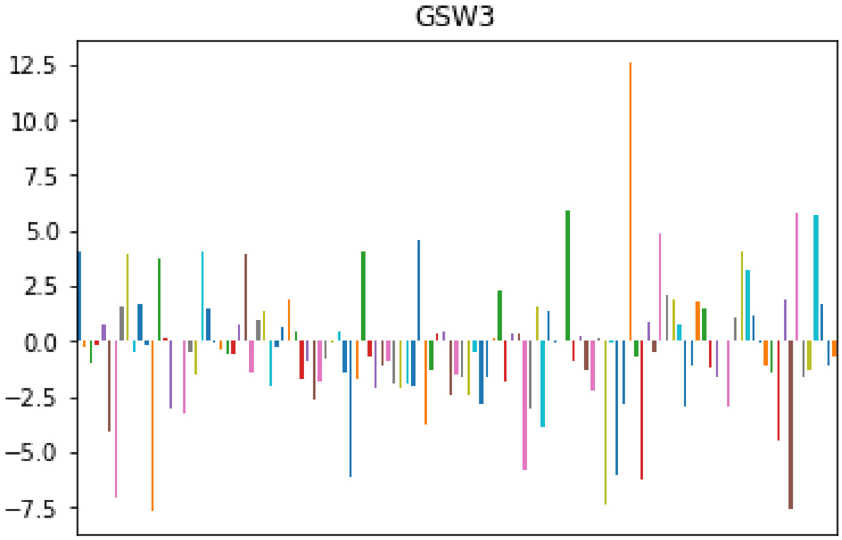

Figure 1: Pure third order effects for the 2015-2016 Golden State Warriors.

In our genomics application, we consider a subset $n$ genetic markers of 1000 indigenous ethnically Tibetan women from Nepal, adapted to high altitude which was recently considered [10]. To statistically test all possible genetic marker interactions, we would need to conduct on order of $2^n$ tests. Our Fourier transform approach on the genetic mutation data yields a few high frequency spikes that allow us to only conduct a handful of statistical tests. This short list of statistical tests of which mutations affect phenotype such as hemoglobin count and infant mortality rates and thus minimizing the need for Bonferroni corrections.

To support the result of these two applications we also use simulated data to show that these Fourier transform are more accurate than traditional Lasso and Ridge regression techniques for detecting higher order interactions in data with high signal to noise ratio. These results on the simulated data provides further robustness of the detection methods found in our two applications.

# References

[1] P. Diaconis. “Group Representations in Probability and Statistics”. English. In: *Lecture Notes-Monograph Series* 11 (1988), pp. i-vi+1-192. ISSN: 07492170. URL: http://www.jstor.org/stable/4355560.

[2] R. Kakarala. “A Signal Processing Approach to Fourier Analysis of Ranking Data: The Importance of Phase”. In: *Signal Processing, IEEE Transactions on* 59.4 (Apr. 2011), pp. 1518–1527. ISSN: 1053-587X. DOI: 10.1109/TSP.2010.2104145.

[3] R. Kondor, A. Howard, and T. Jebar. “Multi-object tracking with representations of the symmetric group”. In: 2 (2007), pp. 211–218.

[4] R. Kondor and W. Dempsey. “Multiresolution analysis on the symmetric group”. In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1637–1645. URL: http://papers.nips.cc/paper/4720-multiresolution-analysis-on-the-symmetric-group.pdf.

[5] P. Diaconis. “A Generalization of Spectral Analysis with Application to Ranked Data”. English. In: *The Annals of Statistics* 17.3 (1989), pp. 949-979. ISSN: 00905364. URL: http://www.jstor.org/stable/2241705.

[6] K. P. Paudel, M. Pandit, and M. A. Dunn. “Using spectral analysis and multinomial logit regression to explain householdsâ choice patterns”. English. In: *Empirical Economics* 44.2 (2013), pp. 739–760. ISSN: 0377-7332. DOI: 10.1007/s00181-012-0558-4. URL: http://dx.doi.org/10.1007/s00181-012-0558-4.

[7] G. Jurman et al. “Algebraic stability indicators for ranked lists in molecular profiling”. In: *Bioinformatics* 24.2 (2008), pp. 258–264. DOI: 10.1093/bioinformatics/btm550. eprint: http://bioinformatics.oxfordjournals.org/content/24/2/258.full.pdf+html. URL: http://bioinformatics.oxfordjournals.org/content/24/2/258.abstract.

[8] B. L. Lawson, M. E. Orrison, and D. T. Uminsky. “Spectral Analysis of the Supreme Court”. English. In: *Mathematics Magazine* 79.5 (2006), pp. 340-346. ISSN: 0025570X. URL: http://www.jstor.org/stable/27642969.

[9] C. Niel et al. “A survey about methods dedicated to epistasis detection”. In: *Frontiers in Genetics* 6 (2015), p. 285. ISSN: 1664-8021. DOI: 10.3389/fgene.2015.00285. URL: https://www.frontiersin.org/article/10.3389/fgene.2015.00285.

[10] C. Jeong et al. “Detecting past and ongoing natural selection among ethnically Tibetan women at high altitude in Nepal”. In: *PLOS Genetics* 14.9 (Sept. 2018), pp. 1–30. DOI: 10.1371/journal.pgen.1007650. URL: https://doi.org/10.1371/journal.pgen.1007650.