HodgeRank as a quantitative tool in social representations theory

Ana L Pereira\textsuperscript{a,b}, José T Lunardi\textsuperscript{a}, Marcos Calçada\textsuperscript{a} and Viviane A Bagio\textsuperscript{b}

\textsuperscript{a} Department of Mathematics & Statistics
\textsuperscript{b} Graduate Program in Education
State University of Ponta Grossa, Avenida Carlos Cavalcanti 4748, 84030-900 Ponta Grossa, PR, Brazil
E-mail: ana.pereira.173@gmail.com, jttlunardi@uepg.br, mcalcada@uepg.br, vivibagio@gmail.com

Abstract. HodgeRank is a technique proposed by Jiang \textit{et al} that provides a way for ranking data elements based on the relative importance that individuals associate to them. This technique has the advantage of working fine with incomplete and imbalanced data, besides providing a measure for the inconsistencies of the outcome. In this work we propose to use HodgeRank as a complementary quantitative tool for analysing the content and the structure of a Social Representation, in which data are spontaneous evocation of words or phrases by a group of individuals; such words or phrases are induced by a term which constitutes the object of the representation under study.

1. Introduction
Social representations (SR’s) were introduced in the social psychology by Moskovici in 1961, and consist in a framework of ideas, concepts, beliefs or feelings shared by individuals in a given group, regarding some social object \cite{1}. Since then, this theory evolved both in its conceptual aspects and in the development of methodological tools to analyse data. In the last decades a great effort was made in the development of quantitative tools, as complementary to the most usual qualitative analysis. In this regard, graphs of similarity, techniques of clustering and statistical analysis of frequency and evocation rank are often used to analyse the content and structure of a social representation (see \cite{2}, for instance, for a recent critical review of the methods in SR’s). In this work we propose to use the method of HodgeRank, a technique proposed by Jiang \textit{et al} \cite{3}, and that is based on the combinatorial Hodge theory, as an auxiliary quantitative tool to help the analysis of the content and the structure of an SR.

The rationale behind the HodgeRank technique is the following. Many modern datasets have one or more of the following features: (1) they contain \textit{cardinal scores} instead of ordinal orderings; (2) they are \textit{incomplete} in the sense that most entries lack information; (3) they are \textit{imbalanced} which means that the amount of available information varies widely from entry to entry and/or from criterion to criterion; (4) the data has the structure of a \textit{complex network} and this structure is important in the ranking process. The HodgeRank is a technique for ranking such datasets. It is formulated in terms of a graph where items to be ranked correspond to vertices and the edges have weights (edge flows) to quantify \textit{pairwise rankings}. The incompleteness and imbalance
of the data are reflected in the edge sparsity structure and vertex degree distribution of the underlying graph, respectively. So, naturally, the technique takes into account the underlying complex network structure. In fact, it exploits the Combinatorial or Discrete Hodge Theory, which is a discrete version of a theory originally coming from Geometry and Topology. By making use of the graph Helmholtzian (which is the graph theoretic analogue of the Helmolitz operator or vector Laplacian), the HodgeRank technique provides a way to extract ranking information from edge flows (pairwise rankings) by decomposing the space of edge flows into orthogonal components: a gradient flow that represents the global ranking and a divergence-free flow that measures the validity of the global ranking obtained - if this is large, then it indicates that the data does not have a good global ranking. This divergence-free flow can be further decomposed orthogonally into a curl flow (locally cyclic) and a harmonic flow (locally acyclic but globally cyclic); these provides information on whether inconsistency in the ranking data arises locally or globally.

In this work, the data elements to be ranked will be a set of words, expressions or phrases evoked by a group of individuals when each individual is asked to spontaneously write 5 words that come to their mind, ranked by the relative order of importance, when an inducing term is presented. The inducing term constitutes the object of the social representation being studied.

2. The “four quadrants” analysis
A popular methodology for analysing the content of an SR was proposed by Vergès in the 1990’s, and one of its versions consists of asking for a group of individuals to write, ranking in order of importance, a number \( n \) of words that come to their mind when the researcher says the word (inductor) that characterizes the object of representation under study [1, 2].

The evocation rank (or score) \( r_{i}^{\alpha} \) of a word \( i \), considering the answer of the individual \( \alpha \), is defined as the order of importance this individual assign to it. For example, if \( n = 5 \) words, \( r_{i}^{\alpha} \) can vary from 1 (higher importance) to 5 (lower importance). The average evocation rank (AER) \( \bar{r}_{i} \) of the word \( i \) is the average of its scores, where the average is taken over all the individuals that evoked that word.

The words evoked by the individuals of the group are then organized in a double entry table, which defines 4 “quadrants”: the upper two contain the words evoked with higher frequency in the group; the two left contain the words with lower AER’s (higher average importance). The words in the upper left quadrant are the candidates for the so-called “central core” of the SR.

Table 1 shows the 4 quadrants for a SR of the term Didactics in a group of students from 10 initial teacher education courses, before they attended the discipline of “didactics”, from a public university in the State of Parana, Brazil (we did not translate the words/expressions from Portuguese). They were asked to answer the question “Which are the 5 words or expressions that best define the concept of didactics? Write down the words according to the importance you assign to them”. The words/expressions spontaneously evoked by the group were aggregated according with their meanings and the resulting words/expressions were used into the analysis.

The thresholds in the frequency and in the AER, which define the boundaries between the quadrants, are ad hoc parameters, and this feature constitutes a weakness of the 4 quadrants methodology [2]. Below we illustrate the application of the HodgeRank technique to provide a ranking of these words, built from the structure of the graph of pairwise comparisons among them, defined from the scores each individual assigned to the 5 chosen words.

3. The HodgeRank technique
Our aim is to rank the words of the SR in the previous Section, based on the relative importance the whole group of individuals assign to pairs of words. We associate the measure of the relative importance the individual \( \alpha \) assigns to a pair of words \( i \) and \( j \) by the function \( Y_{ij}^{\alpha} = r_{j}^{\alpha} - r_{i}^{\alpha} \).
Table 1. Four quadrants for the SR of Didactics in a group of students of 10 courses of initial teacher education in a public university in the State of Parana, Brazil, before they attended the discipline of “didactics”. The upper cells contains words/expressions evoked with an overall frequency (the second column in each cell) ≥ 15%; the left cells contains words having AER ≤ 3 (the last column in each cell). Only words/expressions with overall frequency ≥ 5% were considered.

| Campo de conhecimento | Aplicação docente | Estratégias específicas |
|-----------------------|-------------------|-------------------------|
| 37.94 2.6             | 27.66 3.0         | 13.43 3.4               |
| Metodologia de ensino | Encomenda docente | Instrumental             |
| 30.03 2.5             | 12.64 3.2         | 12.64 3.2               |
| Atuacao docente       | Escola            | 12.25 3.4               |
| 28.06 2.9             | Gestão da matéria| 11.06 3.4               |
| Vida comum            | Habilidades sociais| 10.27 3.1              |
| 24.11 2.3             | Saberes docentes  | 9.09 3.1                |
| Desenvolvimento profissional docente | Gestão da classe | 8.69 3.0                |
| Planejamento          | Ação didática     | 8.69 3.7                |
| 22.92 2.8             | Clima da aula     | 7.9 3.5                 |
| Aprendizagem          | Processo didático | 6.71 3.9                |
| 20.55 2.7             | Ensino como mediado | 5.92 3.1            |
|                       | Conteúdo-forma    | 5.13 3.3                |

if the individual α evokes both the words i and j, and zero otherwise. If \( r_α^i < r_α^j \), then the individual assigned a higher importance to i than to j. Obviously, \( Y_{ij}^α = -Y_{ji}^α \).

The function \( Y_{ij} = \sum_{\omega} \frac{Y_{ij}^\omega}{\omega_{ij}} \), if \( \omega_{ij} \neq 0 \), and zero otherwise, where \( \omega_{ij} \) is the number of individuals that compare words i and j, will be associated to the average relative importance the group assigns to the words i and j. We have also \( Y_{ij} = -Y_{ji} \) (the solution for our ranking (grad \( s^* \)) can be viewed as the elements of a skew symmetric matrix). If there were no inconsistencies in the set of pairwise comparisons \( Y_{ij} \) (the edge flows), one may seek for a “potential” function \( s_i \) which will be the “global rank” of the word i, considering the whole group of individuals) such that \( Y_{ij} = s_j - s_i \equiv (\text{grad} s)_{ij} \); solving such a problem is analogous to solve an electric circuit in which between nodes i and j, with electric potentials \( s_i \) and \( s_j \), respectively, flows an electric current \( Y_{ij} \), from i to j (assuming the convention that the current flows from lower to higher potentials). However, inconsistencies in the set of edge flows \( Y_{ij} \) will often be present, and a solution \( s_i \) may not exist. However, one can seek for the “best” solution \( s^* \), such that \( \text{grad} s^* \) is as close as possible to \( Y \) (we are using matrix notation).

In Combinatorial Hodge theory one introduces inner products on the spaces of functions \( C^0 \cong s_i^* \), \( C^1 \cong Y_{ij} \) and \( C^2 \cong \Phi_{ijk} \), where \( \{i, j, k\} \) is a triangle and \( \Phi_{ijk} \) is a hypermatrix (symmetric w.r.t cyclic permutations of \( \{i, j, k\} \), and skew symmetric otherwise). Then, after introducing the function \( (\text{curl} X)_{ijk} = X_{ij} + X_{jk} + X_{ki} \) and calculating its adjoint \( \text{curl}^* \), one can prove the Hodge decomposition:

\[
C^1 = G \oplus \mathcal{H} \oplus I,
\]

where \( G = \text{Im} (\text{grad}) \), \( I = \text{Im} (\text{curl}) \) and \( \mathcal{H} = G^\perp \cap I^\perp \). Now, \( Y_{ij} \) can be decomposed as

\[
Y = \text{grad} s^* \oplus h \oplus \text{curl}^* \varphi,
\]

where \( s^* \) is the best solution we were seeking for the potential (the solution for our ranking problem). This best solution is that which minimizes the squared norm \( ||Y - \text{grad} s^*||^2 \), and this quantity measures the overall quality of the optimal solution obtained (the squared norms of \( h \) and \( \text{curl}^* \varphi \) give, respectively, the magnitude of the global and local inconsistencies).

In Table 2 we show the result of the HodgeRank technique applied to the same data used in Section 2.
Table 2. Ranking of the words/expressions of Table 1, obtained by the HodgeRank technique. The overall quality of the ranking may be measured by the ratio \( R = \frac{||Y - \text{grad} s^*||^2}{||Y||^2} \approx 0.41 \), which may be interpreted as a relative measure of inconsistency in the matrix of pairwise comparisons \( Y \).

4. Discussion
Comparing the ranking in Table 2 with the 4 quadrants scheme of Table 1, we observe that the words/expressions in the two upper and the two left cells of the 4 quadrants tend to have higher global ranks, as expected. In some sense, the ranking procedure allows one to overcome the issues associated to the arbitrariness related to the thresholds in frequency and AER in the 4 quadrants scheme; besides that, the ranking reveals more structure in the data, coming from the complex network structure of pairwise comparisons.

It is important to notice that the outcomes of the HodgeRank depend of the definitions of the pairwise flux \( Y_{ij}^\alpha \), the aggregate flux \( Y_{ij} \) and of the definitions of the inner products in the spaces \( C^0, C^1 \) and \( C^2 \). Different definitions of these quantities may lead to different ranking outcomes. Establishing criteria for suitable choices of these functions are crucial for obtaining meaningful results.

Acknowledgements
ALP and JTL thank Fundação Araucária (State of Parana, Brazil) for partial support.

References
[1] W Wagner, G Duveen, R Farr, S Jovchelovitch, F Lorenzi-Cioletti, I Marková and D Rose 1999. Theory and method of social representations. Asian Journal of Social Psychology 2 95
[2] G Lo Monaco, A Piermattéo, P Rateau and JL Tavani 2016. Methods for studying the structure of social representations: a critical review and agenda for future research. Journal for the Theory of Social Behaviour 47 3. 4
[3] X Jiang, L-H Lim, Y Yao and Y Ye 2011. Statistical ranking and combinatorial Hodge theory. Mathematical Programming, Series B, 127 203