Referees Report: A Mathematical Programming Approach to Sparse Canonical Correlation Analysis

by L. Amorosi, T. Padellini, J. Puerto and C. Valverde

This paper investigates the effectiveness of integer programming formulations, solvers and heuristics on finding (good) feasible solutions for the sparse canonical correlation analysis (CCA). Several algorithms are discussed and compared in an experimental context. This paper contributes to a growing literature on the use of operations research techniques in statistics, data science and machine learning.

The methods explored are traditional operation research techniques and it is mainly the application and the heuristic methods, as applied to the given problem in question, that forms the main contribution of this paper. The author demonstrate that modern optimization solvers have progressed to a point where mixed integer formulations provide better avenues to solutions that more traditional continuous optimization formulation and solvers. This is presumably due to the fact that these tools now solve much larger problems more quickly.

There are places in the paper where a little bit more detail would be informative to the reader. The introduction contains an extensive summary of the literature relating to the use of operations research techniques in data sciences and in addition a "state of the art" section which extends this discussion further. This exceeds 4 pages in length and I wonder if it is clear the reader in the end how all these papers relate to the work explored later in this paper. Moreover the English is wanting in some sections so I recommend that the authors look towards a revision of the first two sections to focus the discussion more tightly on the areas of interest of this work and also to improve the English (I will give an example in the comments below). There is also a heavy use of abbreviations and I wonder at times if this is always necessary i.e. do we need DS for data science? On the other had EEG (page 5 line 32) is used without a definition.

Detailed Comments:

- NB: I will use line numbers given in the left margin of the review copy.
- I am presuming that the formulation referred to on page 2 line 36 is the formulation (8)-(12) on page 7?
- The sentence concluding at the top of page 3. The reason is not clear to this referee why sparsity was chosen as one of the first properties to be studied. Could the authors elaborate further.
- Page 3 line 41: "...the maximum number of non-zero components of each PC *(is less than?)* (s)."
- The sentences starting on page 4 line 17 to line 22. Could I suggest the follow rewording, as an example of the need to refine some English expression: "A cutting plane algorithm *is use* to solve *this* on instances *where*

the number of samples and regressors *present are* in the *hundreds of thousands.* Moreover, the authors observed that the sparse regression problem has the property that, as the sample size (n) increases, the problem*s* become *more tractable in that one may* perfectly recover the support of the true signal, faster than LASSO. *On the other hand* for small n values, their approach takes *a* long*er* time to solve the problem. "

- Page 6 line 47; what is L_0? this has not been specifically defined.

- I am making a presumption that the formulation (8)-(12) is what the authors later (first on page 12) refer to as being $CCA(K_1, K_2)$ where $j \in K_1$ are in indices of the (possibly) nonzero variables $z_{j,1}$ and $j \in K_2$ are in indices of the (possibly) nonzero variables $z_{j,2}$? Please clarify.

- Page 8 line 40: The author claim from a theoretical point of view, it can produce local optimal solutions?? Gauss-Sidel can only produce partial optimal solutions (as rightly discussed shortly after). If the problem is partly smooth then partial optimal solutions are stationary. As problem at hand cannot be formulated as partial smooth problem due to he presence of constraints on both the w_1 and w_2 variables. Thus feasibility is all that is guaranteed. Please correct this.

- In discussion of the Algorithms 2 and 3 could the authors be more specific as to how one constructs the initial kernels K_1 and K_2?

- Page 14 line 37. The multi-start strategy (BR_Ms) is referred to. Could me detail be provided as to how this done?

- In the discussion of the algorithms on page 14 it is not always absolutely clear which acronym refers to which algorithm. I am presuming BRKS refers to algorithm 3 and BR to that in section 3.2 etc. Could the authors map the acronyms to the relevant sections or better still number all algorithms.

- From table 2 it is clear that the LASSO algorithm 1l-rgl executes in a very fast time. Despite it having inferior results, its output could be used to initialize another method much like the author have used the output of the best response method to initialize BD and KS?