Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Using restricted CCA for cross-language information retrieval

Emil Polajnar

COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2017, VOL. 46, NO. 6, 4618–4626
http://dx.doi.org/10.1080/03610918.2015.1122054

Taylor & Francis
Taylor & Francis Group

Using restricted CCA for cross-language information retrieval

Emil Polajnar

Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia

ABSTRACT
Canonical correlation analysis is a method of correlating linear relationship between two sets of variables. When not any linear combination of variables is allowed, restricted canonical correlation analysis is appropriate. The method was implemented with alternating least-squares and applied to the cross-language information retrieval on a dataset with officially translated and aligned documents in eight European languages.

ARTICLE HISTORY
Received 17 August 2015
Accepted 9 November 2015

KEYWORDS
Alternating least-squares; Cross-language information retrieval; Restricted canonical correlation

MATHEMATICS SUBJECT CLASSIFICATION
62H20; 93E24; 68P20

1. Introduction

Canonical correlation analysis (CCA), proposed by Hotelling (1936), is a method of correlating linear relationship between two sets of variables. In view of linguistic, CCA can also be seen as a feature selection method. It uses two views (two different languages) of the same semantic object (content of a document) to extract common semantic information (prevalent topics in a document). CCA is used across a wide range of scientific fields. A small sample of a more recent research includes papers in pattern recognition (Huang et al., 2010; Kim, 2012), social sciences (Heise and Lerner, 2006), psychology (Trank et al., 2002), and functional magnetic resonance imaging (Friman et al., 2001; Jin et al., 2012).

However, sometimes the nature of a problem does not allow any linear combination of variables and restricted CCA (RCCA) is a more appropriate method. RCCA was introduced by Das and Sen (1994), where one can find some motivating examples. More recently, RCCA was used to lower false positive detection of neural activity in functional magnetic resonance imaging (Ragnehed et al., 2009; Rydell et al., 2006).

A simple algorithm for solving the problem of RCCA that is based on a matrix generalized eigenvalue problem was proposed by Omladič and Omladič (2000). Although the algorithm solves the problem in a finite number of steps, total number of steps grows exponentially with the increasing number of variables in the two sets and quickly becomes impractical. Friman et al. (2003) limited the analysis to only a few variables due to the exponential time cost of computations. However, one can use a similarity between CCA and linear regression to obtain an iterative method known as alternating least-squares (ALS). It is fairly easy to implement and was put in use with a variety of multivariate methods (Branco et al., 2005; Golub and Zha, 1992; Young et al., 1976). There is a recent use of ALS with sparse CCA by Lykou and Whittaker (2010), but it is not exactly RCCA.

CONTACT Emil Polajnar ✉ emil.polajnar@fdv.uni-lj.si Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia.

Searching for information is a part of our everyday life and information retrieval techniques are in a growing need to help us use an ever faster growing amount of available information. Cross-language information retrieval enables us to retrieve information in a language we are unfamiliar with using a query in a language we are familiar with. One way to build cross-language information retrieval system is to automatically create a mapping between two languages. In Rehder et al. (1997), cross-language latent semantic indexing (LSI), introduced by Deerwester et al. (1990), was proposed and is based on a singular value decomposition. Vinokourov et al. (2002) and Li and Shawe-Taylor (2006) used kernel CCA (KCCA), which allows for nonlinear correlations with a mapping to a higher dimensional space and is therefore particularly suitable to cross-language information retrieval with a significantly better performance than LSI.

In this article, we present the results by using RCCA for cross-language information retrieval on a large set of parallel text corpora in eight European languages. There were two motivations to do this. One was to test, whether RCCA implemented with ALS algorithm would be fast enough that it could be used in problems with a large number of constraints. This could possibly open its use in other fields. The other was to see, if including natural nonnegativity constraints in the model improved the accuracy rates of mate retrieval. If RCCA performed on a par or better than CCA it would suggest further tests with restricted KCCA (RKCCA), that is already theoretically founded (Otopal, 2012).

2. Methods

Suppose we are given n pairs of documents in two languages. Every document c_i in one language is a translation of document d_i in another language. One way to represent text documents is to use the vector space model. After some preprocessing the vector space for each language is formed by all the terms (words) we selected as relevant. A collection of terms for one language is actually a set of variables for that language. Suppose we selected p terms x_k for one language and q terms y_k for the other language. Then, we can obtain a numeric representation of documents. We form vector f_{xi} with p elements for every document c_i and vector f_{yi} with q elements for every document d_i. The simplest approach is to set the kth element of vector f_{xi} to 1, if term x_k appears in document c_i, and set it to 0, if the term is not present in the document. However, the most commonly used method is the term frequency-inverse document frequency (tf-idf) method. The kth element of vector f_{xi} is

$$f_{xik} = \mathrm{tf}_{ik} \cdot \log \frac{n}{\mathrm{df}_k}$$

where term frequency tf_{ik} counts the number of times that term x_k occurs in document c_i, n is the number of all documents in dataset, and document frequency df_k counts the number of documents containing the term. The more the term appears in a document (term frequency), the more it is important for that document. And the less the term is common among all the documents in the dataset (inverse document frequency), the more it is important for that document. We collect all vectors f_{xi} and f_{yi} as column vectors in matrices X and Y of size $p \times n$ and $q \times n$, where p and q are the number of selected terms (variables) in two languages. By using CCA, we can find directions in two spaces that would be maximally correlated. Because variables are terms (words) from documents, directions represent a collection of terms about the most prevalent topics in documents.

Formally, CCA finds a canonical correlation ρ and two directions u and v, such that

$$u = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p \qquad v = \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_q y_q$$

$$\rho = \max_{\alpha,\beta} \frac{\mathrm{cov}(u, v)}{\sqrt{\mathrm{var}(u)\,\mathrm{var}(v)}}$$

It is easy to see that the solution is not affected by rescaling α or β, therefore it is usually maximized under normalization conditions $\mathrm{var}(u) = \mathrm{var}(v) = 1$. After applying the Lagrange multiplier technique to the above optimization problem and some algebra, we are left with a generalized eigenvalue problem

$$\begin{bmatrix} 0 & S_{xy} \\ S_{yx} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \lambda \begin{bmatrix} S_{xx} & 0 \\ 0 & S_{yy} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

where $S_{xx} = \mathrm{var}(X)$, $S_{yy} = \mathrm{var}(Y)$, and $S_{xy} = S_{yx}^T = \mathrm{cov}(X, Y)$ are variance–covariance matrices and the largest eigenvalue λ is the canonical correlation ρ. The corresponding eigenvector has $p + q$ components and can be split into α and β, the first p components being α and the last q components β.

If elements in vectors α and β should not take any value, RCCA is to be used. And in the case of the vector space model with x_k and y_k representing terms (words) they should not. We may look at elements of vectors α and β as the weights of the corresponding terms. And weights should take only non-negative values. Only after we require constraints $\alpha \geq 0$ and $\beta \geq 0$ can we properly interpret directions u and v as a collection of terms about the most prevalent topics in documents.

There is a simple algorithm for solving the RCCA problem. It breaks down the main problem to many matrix generalized eigenvalue problems, where every submatrix of the original matrix eigenvalue problem has to be tested for a possible solution. RCCA solution is therefore also an eigenvector because it is a CCA solution of a submatrix problem. Details can be found in Omladič and Omladič (2000). We note that the algorithm finds the solution in a final number of steps but the number of steps grows exponentially with the number of variables in the two sets as $(2^p - 1)(2^q - 1)$. Even for modest values of p and q run times on a computer become long enough to make the solution unfeasible. Instead, we used an iterative approach known as alternating least-squares to find eigenvectors. The algorithm is described in a pseudo code as follows.

```
set elements of vector $\beta$ to some random non-negative values
repeat
    set $b = \beta^T Y$
    solve $\min_\alpha ||\alpha^T X - b||^2$ as least-squares with $\alpha \geq 0$
    normalize $\alpha$ so that $\mathrm{var}(u) = 1$
    set $b = \alpha^T X$
    solve $\min_\beta ||\beta^T Y - b||^2$ as least-squares with $\beta \geq 0$
    normalize $\beta$ so that $\mathrm{var}(v) = 1$
    compute $\rho = \mathrm{cov}(u, v)$
until convergence
```

The algorithm was implemented in Python. The most important part of the algorithm are two calls of least-squares regression with non-negativity constraints. For this we used nnls

Table 1. CCA and RCCA basis vectors corresponding to the largest eigenvalue. Ten terms respectively in English and Spanish ordered by the values.

CCA				RCCA			
English	α	Spanish	β	English	α	Spanish	β
president	0.872	president	0.842	president	0.874	president	0.850
european	0.162	union	0.148	european	0.153	union	0.142
union	0.150	europe	0.146	union	0.141	europe	0.139
countries	0.082	señor_president	0.068	parliament	0.089	parlament	0.084
mr_president	0.018	pais	0.062	countries	0.076	señor_president	0.068
parliament	0.016	parlament	0.020	mr	0.028	pais	0.055
mr	0.015	señor	0.003	mr_president	0.014	señor	0.009
european_union	0.012	inform	0.003	european_union	0.013	inform	0
members	0.006	la_comision	− 0.044	members	0.008	la_comision	0
commission	− 0.134	comision	− 0.097	commission	0	comision	0

function in scipy.optimize package from the SciPy library. This function is based on the Fortran code that was published in Lawson and Hanson (1987).

3. Results and discussion

Cross-language information retrieval. In the previous section, we have shown that CCA and RCCA produce a set of eigenvectors. The eigenvectors with the largest eigenvalues correspond to the maximally correlating directions. These eigenvectors form a base in each of two languages for documents in a training set. Every basis vector represents a topic or a mix of topics (see Table 1). And with the help of these basis vectors cross-language information retrieval is possible, as is discussed next.

We first select d eigenvectors with the largest eigenvalues to form a base in language 1 and language 2. These are d directions, each direction with its distinct topic or mix of topics. Then, we represent a query in language 1 as a linear combination of basis vectors in space of language 1. We also represent some documents in language 2 as a linear combination of basis vectors in space of language 2. Finally, we compare representations of the query and the documents to select (or order) the relevant documents in language 2 based on the query in language 1. Or, if we wish, we can also have a query in language 2 and some documents in language 1.

Formally, we take first d vectors α with the largest eigenvalues and form $p \times d$ matrix A with α as column vectors. We also take first d vectors β with the largest eigenvalues and form $q \times d$ matrix B with β as column vectors. We assume that vectors α correspond to language 1 and vectors β correspond to language 2. Next, we take a query as a column vector q in language 1 and project it onto the d canonical directions in space of language 1 as $q_{fs} = A^T q$. Similarly, we represent documents in language 2 as d dimensional vectors in space of language 2. Then, a document with the shortest Euclidean distance to the query is regarded as being relevant.

Table 1 shows one example of the basis vectors (α and β) obtained from a training set of 1000 documents for one pair of languages. For demonstrating purposes, we selected only 10 terms ($p = q = 10$) for the vector space in each language. This makes it possible to show all components of the basis vectors so that differences between CCA and RCCA can be spotted. CCA allows any value for coefficients and in the left of Table 1 a few negative values are present. Because components of vectors α and β are weights for the terms it is difficult to interpret such values. However, with RCCA we constrained all coefficients to non-negative values. In the right of Table 1, we see that some constraints were active and corresponding terms are excluded with the weight zero. Those terms are therefore not part of a topic covered by this direction. It is important to note that excluded terms by RCCA are not necessary those that have negative value of the weight by CCA. Also, the order of terms differs between CCA and RCCA. Therefore, RCCA is not the same as applying CCA and afterward setting all negative values to zero.

The dataset. We used the European Parliament Proceedings Parallel Corpus dataset. With each enlargement of the European Union, there are more official languages and the newest dataset now includes aligned documents in 21 European languages. However, we used an older version of dataset, that has a total of 11,968 aligned documents in 11 languages, which were at the time the official languages of the European Union. More details about the dataset can be found in Koehn (2005). We used a subset of eight languages: Danish (da), German (de), English (en), Spanish (es), Italian (it), Dutch (nl), Portuguese (pt), and Swedish (sv). Preprocessing and forming vectors f_{xi} for each language was done using all the 11,968 documents in the dataset, although in later analyses only a fraction of documents was included in a training set. It is important to note that all documents are official translations and therefore even more valuable for such a large corpus. Besides, every document is translated in all the languages. This means that with eight languages we were able to get 28 language pairs with aligned documents.

Alternating least-squares. We first tested ALS algorithm in order to asses if RCCA is a feasible method. Code for CCA and ALS RCCA was written in Python and run on a personal computer. The experiment was set as follows. On a training set of 1000 randomly chosen documents we had always p equal to q and varied them from 5 to 100 in increments of 5. Each setting was run 100 times on every language pair and then the average run time was calculated. Run times for the matrix generalized eigenvalue RCCA were estimated as they were expected to be extremely long. We tried to estimate the lower bound. Suppose $p = q = 5$, then the matrix generalized eigenvalue method forms 961 submatrices of different sizes from $p + q = 2$ to $p + q = 10$ and tests each for a possible solution. In all estimated times, we took the average run time for the CCA method with $p = q = 5$ and multiplied it with the number of submatrices. This overestimates the $p = q = 5$ setting but with larger p and q it underestimates run times as the majority of submatrices are larger than $p + q = 10$. The results are presented in Fig. 1 and are encouraging.

As was anticipated, run times for ALS RCCA are longer than those for CCA. An iterative method makes more steps than the exact one step matrix method and the number of steps is also dependent on a convergence criterion. Nevertheless, the ALS RCCA run times stayed within the order of seconds and they approximately followed the CCA run times as a function of p and q. Therefore, we showed that it is possible to implement the RCCA method and carry out an analysis on a personal computer within a reasonable time frame. Also we showed, with the estimated matrix RCCA run times, that implementation of the RCCA method and its comparison to other methods in cross-language information retrieval would not be possible, if we did not use ALS approach.

Mate retrieval. In mate retrieval, a document in one language is selected as a query and only the mate document in another language is considered as relevant. The mate document is the same document as the query but in another language. Both, the query and a document in another language, are represented as vectors in a base of their language and distance between

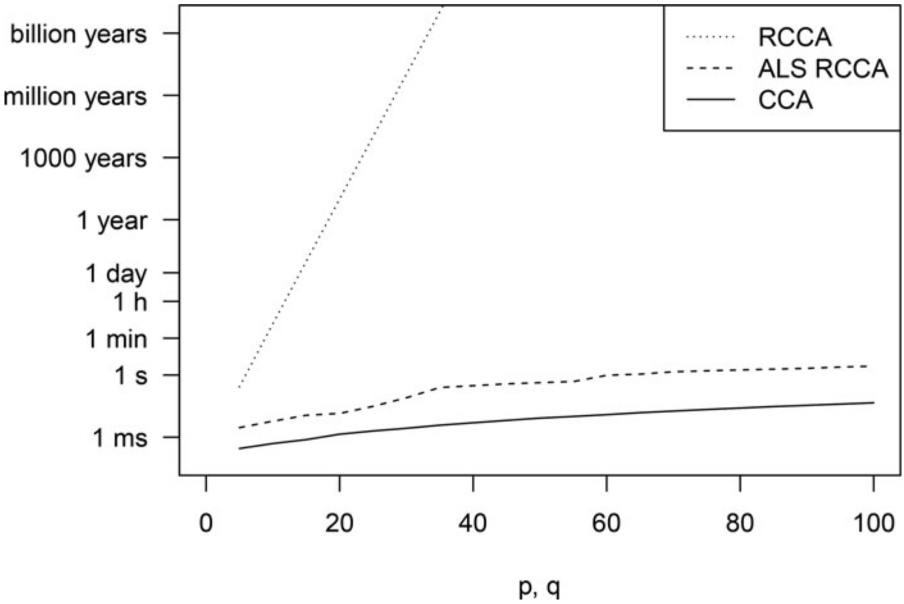

Figure 1. The run times of algorithms. Relationship to the dimensionality p and q of the eigenvectors, where $p = q$. The alternating least-squares RCCA algorithm was compared with the matrix generalized eigenvalue RCCA and CCA algorithms.

the vectors is calculated. The mate document is considered to be retrieved if it is most close to the query among all documents in a test set.

We used a training set because the dataset was too big to work with, the number of documents was 11,968 and the number of terms varied from 51,116 to 171,821. The training set was also a test set. We randomly selected 1000 documents and then for each language chose the 100 most prevalent terms in those documents. Therefore, the dimension of our matrices X and Y was 100 × 1000. However, the number of variables p and q need not to be equal and can be both very large, because the number of terms in documents is large. When the number of variables is large, there are problems with collinearity and finding a solution in a traditional way is not possible. Interested reader is referred to ridge regression concept (Hoerl and Kennard, 1970) or sparse CCA approaches (Lykou and Whittaker, 2010; Wilms and Croux, 2015) that attempt to overcome the problem.

Next, we applied RCCA on documents in the training set for two languages at a time to learn the semantic. We chose either 5 or 10 eigenvectors with the largest eigenvalues to form a base in each of two languages. We did this for all 28 language pairs. Then, we selected queries. We limited our analysis to queries within the training set only and every document in the training set was a query. We first selected a document in the first language in a pair as a query and checked if the mate document in the second language in a pair was retrieved. Next, we reversed the roles of languages and a document in the second language was selected as a query. The last row in Table 2 shows that the correlation of the accuracy rates of mate retrieval between both language directions was very strong. Therefore, only the results for queries in the first language in a pair are discussed. For comparison, we also implemented LSI and CCA for cross-language information retrieval under the same experimental settings.

Table 2 contains various statistics for the accuracy rates of mate retrieval on the training set with $n = 1000$ documents and $p = q = 100$ terms (variables) for each language. These statistics were calculated over all 28 language pairs. If we take numbers for RCCA with five

Table 2. Various statistics for the accuracy rates of mate retrieval.

Eigenvectors	$d = 5$			$d = 10$		
Method	LSI	CCA	RCCA	LSI	CCA	RCCA
Min	0.017	0.229	0.095	0.070	0.415	0.153
Max	0.130	0.526	0.362	0.317	0.725	0.534
Average	0.060	0.381	0.226	0.159	0.603	0.365
StDev	0.036	0.079	0.073	0.075	0.078	0.109
Corr	0.848	0.989	0.987	0.835	0.992	0.994

eigenvectors we can see that for cross-language information retrieval the most difficult language pair (German–Italian and German–Portuguese) matched correctly only 95 documents out of 1000 and the easiest language pair (Spanish–Portuguese) matched correctly 362 documents out of 1000. The average number of correct matches for RCCA with 5 eigenvectors in the set of 28 language pairs was 226 documents out of 1000.

These results are consistent with those on the Japanese–English documents in Li and Shawe-Taylor (2006). That is, LSI performance was quite poor for low number of basis vectors. And we expected that CCA would not challenge KCCA performance as KCCA uses a mapping to a higher dimensional space for nonlinear correlations. Our main objective was to test, whether RCCA would perform on a par or better than CCA. Unfortunately, that was not the case. RCCA lagged quite a bit in performance behind CCA. CCA accuracy rates of mate retrieval were on average better for a factor of 1.69 in five-dimensional space and for a factor

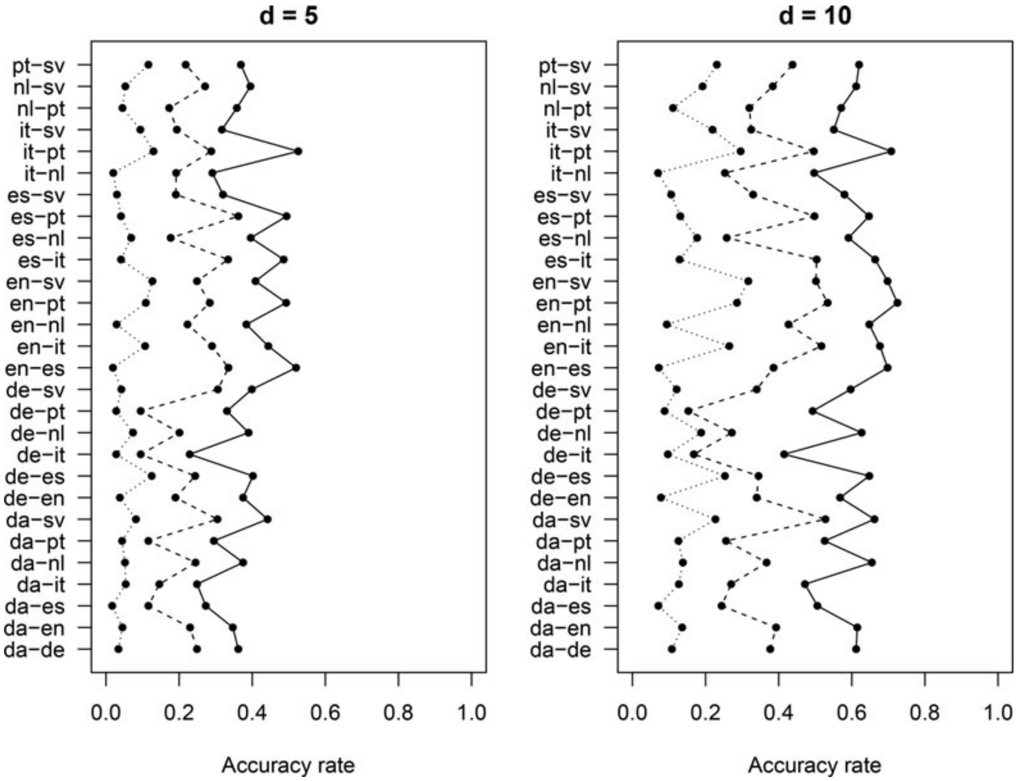

Figure 2. The accuracy rates of mate retrieval for 28 language pairs on the training set of 1000 documents. Learning was done with $p = q = 100$ terms and 5 or 10 eigenvectors with the largest eigenvalues was chosen as a basis. Three methods LSI (dotted), RCCA (dashed), and CCA (solid) were compared.

of 1.65 in ten-dimensional space. If there was an advantage we wanted it at low-dimensional space. Because this was not the case, we did not pursue analysis to higher dimensions as there were no benefits in doing so. It may be that RCCA directions have better and proper explanation as a collection of terms about the most prevalent topics in documents. But since accuracy rate of mate retrieval is measurement of performance CCA is better suitable for cross-language information retrieval than RCCA. We also did not pursue things further with RKCCA, as KCCA and CCA have so many similarities that we cannot expect to outperform KCCA.

In Fig. 2, we present detailed picture of accuracy rates of mate retrieval for 28 language pairs to show variability between pairs. The highest accuracy rate language pairs were English–Portuguese and English–Spanish, while the lowest accuracy rate language pairs were German–Italian and German–Portuguese. It is also clear that RCCA performance was always better than LSI performance and, in turn, CCA performance was always better than that of RCCA.

4. Conclusions

We have presented a novel approach for cross-language information retrieval. In the vector space model representation, each element in the document or the query vector is the term frequency-inverse document frequency of the corresponding term in the document or in the query and as such is a non-negative value. We incorporated this non-negativity in a model through constraints on coefficients and used RCCA to extract semantic information. We demonstrated that alternating least-squares approach offers an elegant and fast solution to RCCA problems even with a large number of constraints in which case the solution with a standard matrix generalized eigenvalue approach would not be feasible.

We tested the proposed method for cross-language information retrieval on a large set of parallel text corpora in eight European languages, that is on 28 language pairs. We compared the accuracy rates of mate retrieval of RCCA to those of LSI and CCA. Although RCCA performed better than LSI it was a little bit disappointing to see it lag behind the accuracy rates of CCA. Nevertheless, it could still be beneficial to use RCCA in some cases. Since eigenvectors have only non-negative values they offer a more natural interpretation of a solution and can be truly viewed as a set of distinct topics in documents.

At the moment, KCCA is the state-of-the-art method in cross-language information retrieval and outperforms CCA in terms of accuracy rates. Since KCCA is very similar to CCA, we see little possibility that restricted KCCA would outperform KCCA.

Acknowledgment

The author would like to thank the reviewer for the helpful comments, which substantially improved this article.

References

Branco, J. A., Croux, C., Filzmoser, P., Oliveira, M. R. (2005). Robust canonical correlations: A comparative study. *Computational Statistics* 20:203–229.

Das, S., Sen, P. K. (1994). Restricted Canonical Correlations. *Linear Algebra and its Applications* 210:29–47.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41:391–407.

Friman, O., Borga, M., Lundberg, P., Knutsson, H. (2003). Adaptive analysis of fMRI data. *NeuroImage* 19:837–845.

Friman, O., Cedefamn, J., Lundberg, P., Borga, M., Knutsson, H. (2001). Detection of neural activity in functional MRI using canonical correlation analysis. *Magnetic Resonance in Medicine* 45:323–330.

Golub, G., Zha, H. (1992). The canonical correlations of matrix pairs and their numerical computation. Stanford, CA: Stanford University.

Heise, D. R., Lerner, S. J. (2006). Affect control in international interactions. *Social Forces* 85:993–1010.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12:55–67.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* 28:321–377.

Huang, H., He, H., Fan, X., Zhang, J. (2010). Super-resolution of human face image using canonical correlation analysis. *Pattern Recognition* 43:2532–2543.

Jin, M., Nandy, R., Curran, T., Cordes, D. (2012). Extending local canonical correlation analysis to handle general linear contrasts for fMRI data. *International Journal of Biomedical Imaging* 2012:1–14.

Kim, M. (2012). Correlation-based incremental visual tracking. *Pattern Recognition* 45:1050–1060.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *In MT summit* 5:79–86.

Lawson, C. L., Hanson, R. J. (1987). Solving least-squares problems. Philadelphia, PA: SIAM.

Li, Y., Shawe-Taylor, J. (2006). Using KCCA for Japanese–English cross-language information retrieval and document classification. *Journal of Intelligent Information Systems* 27:117–133.

Lykou, A., Whittaker, J. (2010). Sparse CCA using a Lasso with positivity constraints. *Computational Statistics and Data Analysis* 54:3144–3157.

Omladič, M., Omladič, V. (2000). More on restricted canonical correlations. *Linear Algebra and its Applications* 321:285–293.

Otopal, N. (2012). Restricted kernel canonical correlation analysis. *Linear Algebra and its Applications* 437:1–13.

Ragnehed, M., Engstrom, M., Knutsson, H., Soderfeldt, B., Lundberg, P. (2009). Restricted canonical correlation analysis in functional MRI—Validation and a novel thresholding technique. *Journal of Magnetic Resonance Imaging* 29:146–154.

Rehder, B., Littman, M. L., Dumais, S. T., Landauer, T. K. (1997). Automatic 3-language cross-language information retrieval with latent semantic indexing. *In TREC* pp. 233–239.

Rydell, J., Knutsson, H., Borga, M. (2006). On rotational invariance in adaptive spatial filtering of fMRI data. *NeuroImage* 30:144–150.

Trank, C. Q., Rynes, S. L., Bretz, R. D. (2002). Attracting applicants in the war for talent: Differences in work preferences among high achievers. *Journal of Business and Psychology* 16:331–345.

Vinokourov, A., Shawe-Taylor, J., Cristianini, N. (2002). Inferring a semantic representation of text via cross-language correlation analysis. In *Advances of Neural Information Processing Systems*. pp. 1473–1480.

Wilms, I. and Croux, C. (2015). Sparse canonical correlation analysis from a predictive point of view. *Biometrical Journal* 57:834–851.

Young, F. W., De Leeuw, J., Takane, Y. (1976). Regression with qualitative and quantitative variables: An alternating least-squares method with optimal scaling features. *Psychometrika* 41:505–529.