Comparing Statistical Similarity Measures for Stylistic Multivariate Analysis

Marius Popescu
University of Bucharest
Department of Computer Science
Academiei 14, Bucharest, Romania
mpopescu@phobos.cs.unibuc.ro

Liviu P. Dinu
University of Bucharest
Department of Computer Science
Academiei 14, Bucharest, Romania
ldinu@funinf.cs.unibuc.ro

Abstract

The goal of this paper is to compare a set of distance/similarity measures, some motivated statistically, others motivated stylistically, regarding their ability to reflect stylistic similarity between texts. To assess the ability of these distance/similarity functions to capture stylistic similarity between texts, we have tested them in the two most frequently employed multivariate statistical analysis settings: cluster analysis and (kernel) principal components analysis.

Keywords

Stylistic Multivariate Analysis, Statistical Similarity Measures, Cluster Analysis, Kernel Principal Components Analysis

1 Introduction

Computational stylistics investigates texts from the standpoint of individual style (author identification) or functional style (genres, registers). Because in all computational stylistic studies/approaches, a process of comparison of two or more texts is involved, in a way or another, there was always a need for a distance/similarity function to measure similarity (or dissimilarity) of texts from the stylistic point of view.

Usually, the distance/similarity measures are implicitly or explicitly used by multivariate statistical analysis techniques typically applied in computational stylistic approaches. In [5], these approaches are characterized as: "[The]...technique essentially picks the N most common words in the corpus under investigation and computes the occurrence rate of these N words in each text or text-unit, thus converting each text into an N-dimensional array of numbers. Multivariate statistical techniques are then applied to the data to look for patterns. The two techniques most frequently employed are principal components analysis and cluster analysis."

The goal of this paper is to compare a set of distance/similarity measures, some motivated statistically, others motivated stylistically, regarding their ability to reflect stylistic similarity between texts.

As style markers we have used the function word frequencies. Function words are generally considered good indicators of style because their use is very unlikely to be under the conscious control of the author and because of their psychological and cognitive role [3]. Also function words prove to be very effective in many author attribution studies.

The distance/similarity between two texts will be measured as distance/similarity between the function words frequencies corresponding to the respective texts. For this study we selected some similarity/distance measures. We started with the most natural distance/similarity measures: euclidean distance and (taking into account the statistical nature of data) Pearson's correlation coefficient. Since function words frequencies can also be viewed as ordinal variables, we also considered for comparison some specific similarity measures: Spearman's rank-order coefficient, Spearman's footrule, Goodman and Kruskal's gamma, Kendall's tau. Finally, we have added a stylistically motivated similarity measure: Burrows's delta, that has interesting statistic interpretations.

To assess the ability of these distance/similarity functions to capture stylistic similarity between texts, we have tested them in the two most frequently employed multivariate statistical analysis settings: cluster analysis and principal components analysis.

Clustering is a very good test bed for a distance/similarity measure behavior. We plugged the distance/similarity measures selected for comparison into a standard hierarchical clustering algorithm and applied it to a collection of 21 nineteenth century English books [6]. The family trees thus obtained revealed a lot about the distance/similarity measures behavior.

If clustering explicitly uses a distance/similarity function as its base, principal components analysis implicitly uses the euclidean distance. Kernel principal components analysis [8] allows the replacement of the implicitly used euclidean distance with other similarity measures, *the kernels*. Not all the distance/similarity measures selected for comparison can be transformed into kernels because a kernel has to be a positive definite function. For those similarity measures that can be transformed into kernels (Spearman's rank-order coefficient, Kendall's tau) we have compared the results of kernel principal components analysis (using the respective kernels) with the result of standard principal components analysis (that implicitly uses the euclidean distance).

The main finding of our comparison is that the similarity measures that treat function words frequencies as ordinal variables performed better than the others distance/similarity measures. Treating function words

International Conference RANLP 2009 - Borovets, Bulgaria, pages 349–354

frequencies as ordinal variables means that in the calculation of distance/similarity function the ranks of function words according to their frequencies in text will be used rather than the actual values of these frequencies. Usage of the ranking of function words in the calculation of the distance/similarity measure instead of the actual values of the frequencies may seem as a loss of information, but we consider that the process of ranking makes the distance/similarity measure more robust acting as a filter, eliminating the *noise* contained in the values of the frequencies. The fact that a specific function word has the rank 2 (is the second most frequent word) in one text and has the rank 4 (is the fourth most frequent word) in another text can be more relevant than the fact that the respective word appears 34% times in the first text and only 29% times in the second.

Also, the experiments shown that Burrows's Delta achieved good results. Burrows's Delta is a stylistically motivated distance function especially designed as a measure for authorship attribution and used until now only in classification experiments. As far as we know this is the first time when Burrows's Delta is used in a clustering setting.

In the next section we present the distance/similarity measures involved in the comparison study. Section 3 briefly describes the multivariate statistical analysis techniques used: cluster analysis and (kernel) principal component analysis. In section 4 are presented the experiments and the results obtained, and the last section contains discussion and suggestions for future work.

2 Similarity Measures

If we treat texts as random variables whose values are the frequencies of different words in the respective texts, then various statistical correlation measures can be used as similarity measures between that texts. For two texts X and Y and a fixed set of words $\{w_1, w_2, \ldots, w_n\}$ let us denote by x_1 the relative frequency of w_1 in X, by y_1 the relative frequency of w_1 in Y and so on by x_n the relative frequency of w_n in X, by y_n the relative frequency of w_n in Y.

2.1 Pearson's Correlation Coefficient

The Pearson's correlation coefficient [9] is:

$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)}{n-1}$$

where $\overline{x}$ is the mean of X, $\overline{y}$ the mean of Y and s_x is the standard deviation of X, s_y the standard deviation of Y.

The correlation coefficient measures the tendency of two variables to change in value together (i.e., to either increase or decrease). r is related with the Euclidean distance, the $\sqrt{2(1-r)}$ being the Euclidean distance between the standardized versions of X and Y.

2.2 Correlation Statistics for Ordinal Data

The random variables X, Y representing texts can also be treated as ordinal data, in which data is ordered but cannot be assumed to have equal distance between values. In this case the values of X (and Y respectively) will be the ranks of words $\{w_1, w_2, \ldots, w_n\}$ according to their frequencies in text X rather than of the actual values of these frequencies. The most common correlation statistic for ordinal data is *Spearman's rank-order coefficient* [9]:

$$r_{sc} = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (x_i - y_i)^2$$

To be noted that, this time, x_i, y_i are ranks and actually, the Spearman's rank-order coefficient is the Pearson's correlation coefficient applied to ranks.

The Spearman's footrule [9] is the l_1-version of Spearman's rank-order coefficient:

$$r_{sf} = 1 - \frac{3}{n^2-1} \sum_{i=1}^{n} |x_i - y_i|$$

Another set of correlation statistics for ordinal data are based on the number of concordant and discordant pairs among two variables. The number of concordant pairs among two variables X and Y is $P = |\{(i,j) : 1 \le i < j \le n, (x_i - x_j)(y_i - y_j) > 0\}|$. Similarly, the number of discordant pairs is $Q = |\{(i,j) : 1 \le i < j \le n, (x_i - x_j)(y_i - y_j) < 0\}|$.

Goodman and Kruskal's gamma [9] is defined as:

$$\gamma = \frac{P - Q}{P + Q}$$

Kendall developed several slightly different types of ordinal correlation as alternatives to gamma. *Kendall's tau-a* [9] is based on the number of concordant versus discordant pairs, divided by a measure based on the total number of pairs (n = the sample size):

$$\tau_a = \frac{P - Q}{\frac{n(n-1)}{2}}$$

Kendall's tau-b [9] is a similar measure of association based on concordant and discordant pairs, adjusted for the number of ties in ranks.It is calculated as $(P - Q)$ divided by the geometric mean of the number of pairs not tied on X (X_0) and the number of pairs not tied on Y (Y_0):

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + X_0)(P + Q + Y_0)}}$$

All the above three correlation statistics are very related, if n is fixed and X and Y have no ties, then P, X_0 and Y_0 are completely determined by n and Q.

2.3 Burrows's Delta

In his 2001 Busa Award lecture, John F. Burrows proposed a new measure for authorship attribution which

he termed 'Delta', defined as: "the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text." [2]

Let $\mathcal{C} = \{X, Y, \ldots\}$ be a fixed set of texts, *a corpus*, and $\{w_1, w_2, \ldots, w_n\}$ a fixed set of words. Let σ_i be the standard deviation of the relative frequency of w_i in the corpus $\mathcal{C}$. For each $X, Y \in \mathcal{C}$, Delta is defined as [1]:

$$\Delta(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - y_i}{\sigma_i}\right|$$

As it is defined, $\Delta(X,Y)$ depends not only on X and Y, but also on the entire data set (corpus) from which X and Y are drawn. This will not be a problem for clustering, because the family tree obtained from a cluster analysis depends anyway on the entire data set (adding a new text to a data set can change the family tree completely).

3 Multivariate Analysis Techniques

3.1 Clustering Analysis

An agglomerative hierarchical clustering algorithm [4] arranges a set of objects in a family tree (dendogram) according to their similarity, similarity which in its turn is given by a distance function defined on the set of objects. The algorithm initially assigns each object to its own cluster and then repeatedly merges pairs of clusters until the whole tree is formed. At each step the pair of nearest clusters is selected for merging. Various agglomerative hierarchical clustering algorithms differ in the way in which they measure the distance between clusters. Note that although a distance function between objects exists, the distance measure between clusters (set of objects) remains to be defined. In our experiments we used the *complete linkage* distance between clusters, the maximum of the distances between all pairs of objects drawn from the two clusters (one object from the first cluster, the other from the second).

3.2 (Kernel) Principal Components Analysis

Principal components analysis (PCA) [4] is a method of dimensionality reduction. The motivation for performing PCA is often the assumption that directions of high variance will contain more information than directions of low variance. The PCA aims to transform the observed variables (function word frequencies, in our case) to a new set of variables which are uncorrelated and arranged in decreasing order of importance. These new variables, or components, are linear combinations of the original variables, the first few components accounting for most of the variation in the original data. Typically the data are plotted in the space of the first two components.

PCA works in the euclidean space and so implicitly use euclidean distance and standard inner product. Kernel principal components analysis [8] allows the replacement of the implicitly used euclidean distance with other similarity measures, *the kernels*.

Group	Author	Book
American Novelists	Hawthorne	Dr. Grimshawe's Secret
		House of Seven Gables
	Melville	Redburn
		Moby Dick
	Cooper	The Last of the Mohicans
		The Spy
		Water Witch
American Essayists	Thoreau	Walden
		A Week on Concord
	Emerson	Conduct Of Life
		English Traits
British Playwrights	Shaw	Pygmalion
		Misalliance
		Getting Married
	Wilde	An Ideal Husband
		Woman of No Importance
Bronte Sisters	Anne	Agnes Grey
		Tenant Of Wildfell Hall
	Charlotte	The Professor
		Jane Eyre
	Emily	Wuthering Heights

Table 1: *The list of books used in the experiments*

Kernel-based algorithms work by embedding the data into a feature space (a Hilbert space). The embedding is performed implicitly, that is by specifying the inner product between each pair of points rather than by giving their coordinates explicitly. The kernel function captures the intuitive notion of similarity between objects in a specific domain and can be any function defined on the respective domain that is symmetric and positive definite. Because of the positive definite restriction not all distance/similarity measures described in section 2 can be transformed into a kernel, but some can. For example, from the Spearman's rank-order coefficient the following kernel can be obtained: $k(X,Y) = e^{-\frac{r_{sc}(X,Y)}{2}}$ Also, P, the number of concordant pairs among two variables X and Y (see section 2.2) can be proved to be a kernel, but the prove of that is beyond the scope of this paper. For details of how a method like PCA ca be transformed into a kernel method and which distance/similarity functions can be a kernel see [8].

4 Experiments

For our experiments we used a collection of 21 nineteenth century English books written by 10 different authors and spanning a variety of genres (Table 1). The books were used by Koppel et al. [6] in their authorship verification experiments.

To perform the experiments, a set of words must be fixed. The most frequent function words may be selected or other criteria may be used for selection. In all our experiments we used the set of function words identified by Mosteller and Wallace [7] as good candidates for author-attribution studies.

In a first set of experiments we used the agglomerative hierarchical clustering algorithm coupled with the various distance similarity functions employed in the comparison to cluster the works the Table 1.

The resulted dendograms for euclidean distance and Pearson's correlation coefficient are very similar, which is no surprise taking into account the close relation between the two measures (see section 2.1). We present only the dendogram for Pearson's correlation coefficient in Figure 1. The problem of this family tree (and also of the family tree corresponding to euclidean distance) is that the works of Melville are not grouped together: one being clustered with the novels

of Cooper (Moby Dick) and the other with the novels of Hawthorne. Also, apart from authorship relation, the dendogram reflects no other stylistic relation between the works (like grouping the works according to genre or nationality of the authors: American / English).

The dendogram for Spearman's footrule (not shown because of lack of space) is a good one, accurately reflecting the stylistic relations between books. The books were grouped in three big clusters (the first three branches of the tree) corresponding to the three genre: dramas (lower branch), essays (middle branch) and novels (upper branch). Inside each branch the works were first clustered according to their author. The only exceptions are the two essays of Emerson which instead of being first clustered together and after that merged in the cluster of essays, were added one by one to this cluster.

Spearman's rank-order coefficient, Goodman and Kruskal's gamma and Kendall's tau produced the same dendogram (modulo the scale). Figure 2 shows the dendogram for Spearman's rank-order coefficient. The dendogram is perfect: all works are clustered according to theirs author. More over, the first two branches correspond to the nationality of the authors: British writers on lower branch, American writers on upper branch. Further more, inside each of these two branches, the works are clustered according to genre: drama and novels in the case of British writers, novels and essays in the case of American writers.

The family tree obtained when Burrows's Delta was used resembles the dendogram produced by Spearman's rank-order coefficient (Figure 2), but this time, in the case of American writers, the works are no longer grouped according to genre.

A second set of experiments aim to compare the standard principal components analysis (that implicitly uses the euclidean distance) with kernel principal components analysis, based on kernels derived from distance/similarity measures selected for this study. The works in the Table 1 are plotted in the space of the first two principal components, to see if the stylistic similarity is reflected in the spatial configuration.

The plot obtained using standard principal components analysis is shown in Figure 3. Generally, the works of the same author are plotted close together, but again (as in the case of the euclidean distance and Pearson's correlation coefficient clustering) the works of Melville ($\times$) are an exception. One is placed close to works of Emerson ($\square$) and the other alone in a different region. Also, the works of Emerson ($\square$) and the works of Cooper ($+$) are not clearly separated. An interesting fact is that the American writers and the British writers are separated in the plane by a vertical line ($x = 0$).

For comparison, in Figure 4 we present the plot obtained using kernel principal components analysis with the kernel derived from Spearman's rank-order coefficient[1] (see section 3.2). The works of the same author are plotted close together and different authors are clearly separated. Even more interesting than in the case of standard components analysis, a vertical line, $x = 0$, separates the British writers (left) from the American writers (right), and a horizontal line, $y = 0$ separates different genres: drama (above) from novels (below) in the case of British writers, essays (above) from novels (bellow) in the case of American writers.

[1] Because of space limitation we have presented the kernel principal components analysis only in the case of the best performing kernel, the kernel derived from Spearman's rank-order coefficient.

5 Discussion

In this paper we have compared a set of distance/similarity measures, some motivated statistically, others motivated stylistically, regarding theirs ability to reflect stylistic similarity between texts. To assess the ability of these distance/similarity functions to capture stylistic similarity between texts, we tested them in the two most frequently employed multivariate statistical analysis settings: cluster analysis and (kernel) principal components analysis.

The experiments have shown that the similarity measures that treat function words frequencies as ordinal variables (Spearman's rank-order coefficient, Spearman's footrule, Goodman and Kruskal's gamma, Kendall's tau) performed better than the distance/similarity measures that use the actual values of function words frequencies (Euclidean distance, Pearson's correlation coefficient).

Also we have shown how Burrows's Delta, a distance function especially designed as a measure for authorship attribution, can be used in clustering analysis with good results.

In future work it would be useful to test these distance/similarity measures on other data sets. Also, it would be interesting to further investigate the ability of some of the similarity measures (Spearman's rank-order coefficient, Goodman and Kruskal's gamma, Kendall's tau) to distinguish between the different nationality of English language writers; for example, by adding to the data set works of Australian writers from the same period.

References

[1] S. Argamon. Interpreting burrows's delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147, 2008.

[2] J. Burrows. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

[3] C. K. Chung and J. W. Pennebaker. The psychological function of function words. In K. Fiedler, editor, *Social communication: Frontiers of social psychology*, pages 343–359. Psychology Press, New York, 2007.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd ed.)*. Wiley-Interscience Publication, 2001.

[5] D. I. Holmes, L. J. Gordon, and C. Wilson. A widow and her soldier: Stylometry and the american civil war. *Literary and Linguistic Computing*, 16(4):403–420, 2001.

[6] M. Koppel, J. Schler, and E. Bonchek-Dokow. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8:1261–1276, 2007.

[7] F. Mosteller and D. L. Wallace. *Inference and Disputed Authorship: The Federalist*. CSLI Publications, Stanford, 2007.

[8] J. S. Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

[9] G. Upton and I. Cook. *A Dictionary of Statistics*. Oxford University Press, Oxford, 2008.

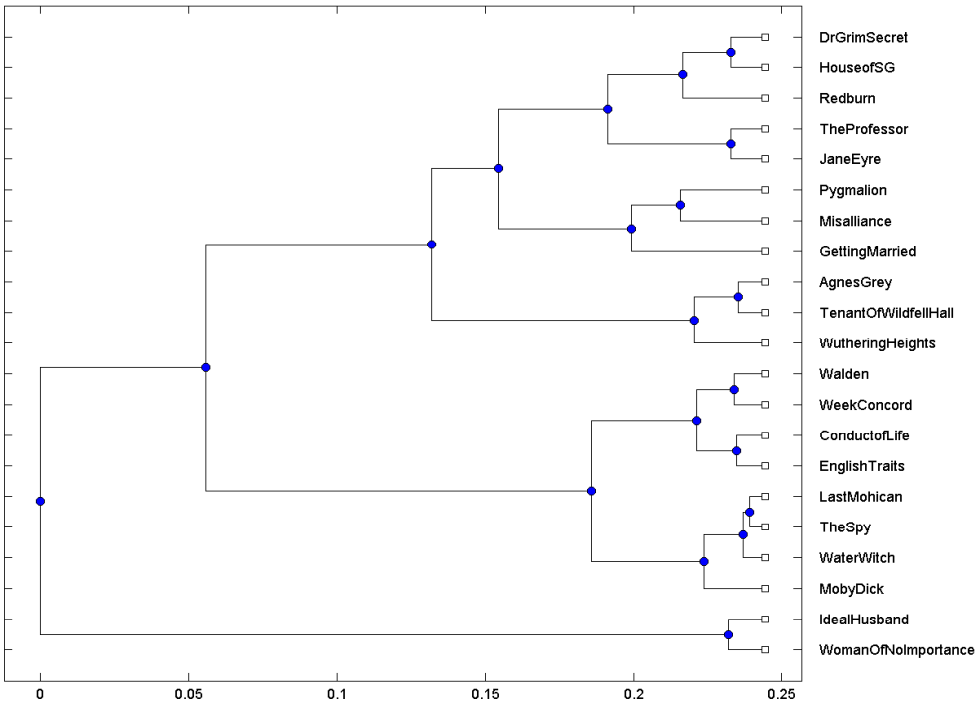

Fig. 1: *Dendogram of 21 nineteenth century English books (Pearson's correlation coefficient)*

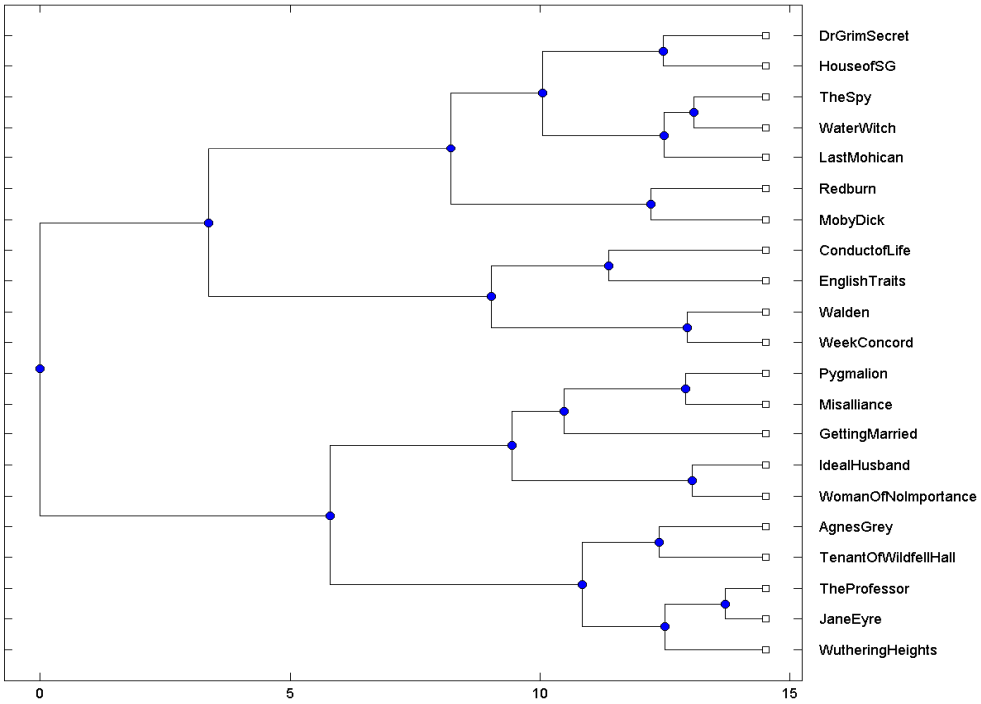

Fig. 2: *Dendogram of 21 nineteenth century English books (Spearman's rank-order coefficient)*

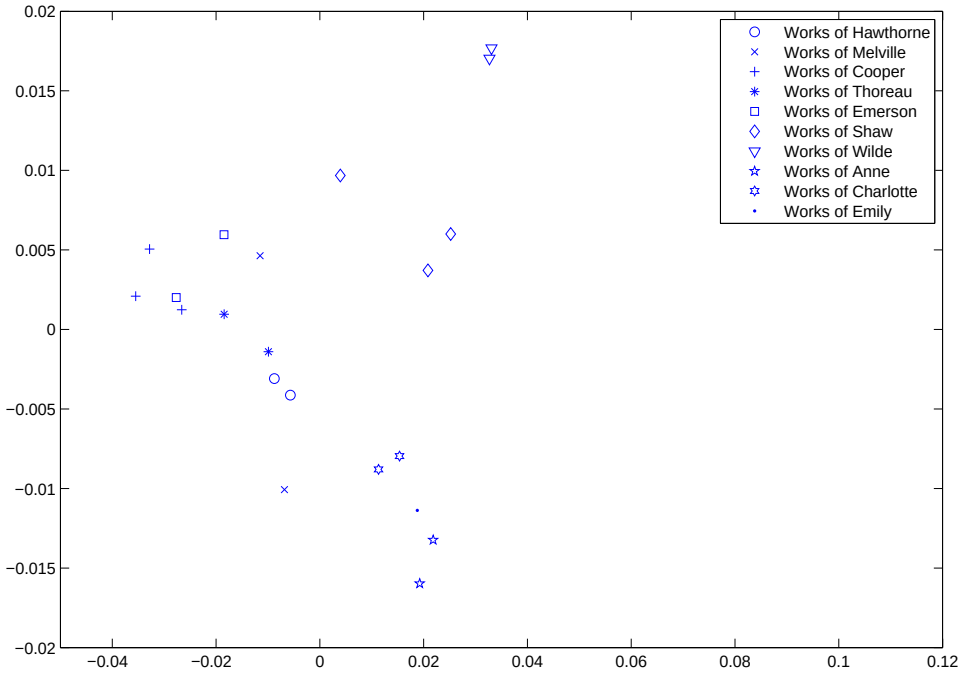

Fig. 3: *Standard principal components plot of 21 nineteenth century English books*

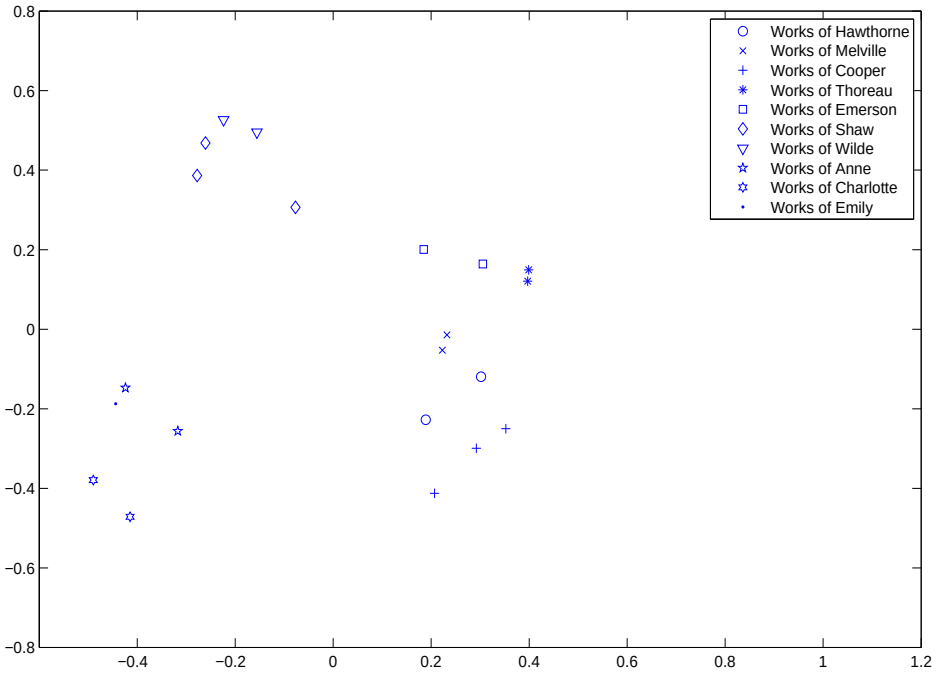

Fig. 4: *Kernel principal components plot of 21 nineteenth century English books (Spearman's rank-order coefficient kernel)*