

Table 1: Kendall τ_b ranking performance measure, for Algorithm 1 and the 20 Newsgroups data set. The mean Kendall τ_b correlation coefficient is reported alongside the standard error (numerical value shown is the standard error $\times 10^3$). This numerical results are plotted in figure 1 below.

Data	Input	$p = 500$	$p = 1000$	$p = 5000$	$p = 7500$	$p = 12818$
Newsgroups	$\mathbf{Y}_{1:n}$	0.026 (0.55)	0.016 (1.0)	0.13 (2.2)	0.17 (2.5)	0.26 (2.9)
	$\zeta_{1:n}$	0.017 (0.72)	0.047 (1.2)	0.12 (1.9)	0.15 (2.5)	0.24 (2.6)

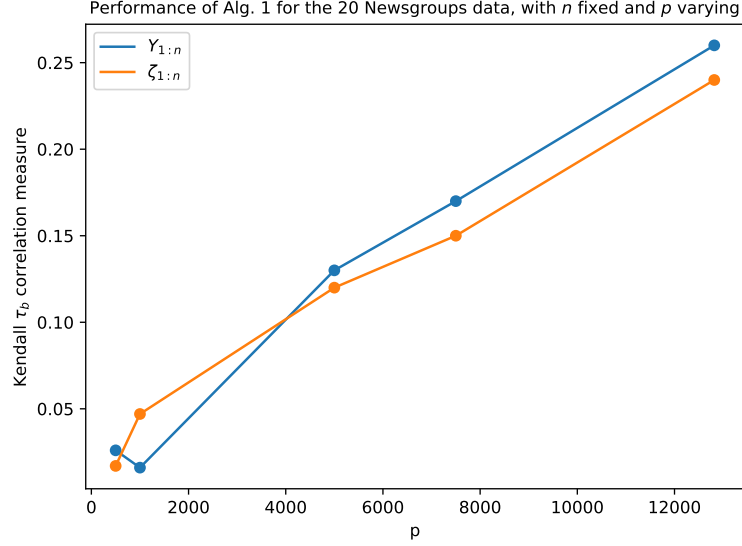


Figure 1: Performance of Algorithm 1 for the 20 Newsgroups data set as a function of number of TF-IDF features, p , with n fixed. See table 2 for numerical values and standard errors.

Table 2: Kendall τ_b ranking performance measure. The mean Kendall τ_b correlation coefficient is reported alongside the standard error (numerical value shown is the standard error $\times 10^3$).

Data	Input	UPGMA with dot product *dissimilarity*	UPGMA with Manhattan distance
Newsgroups	$\mathbf{Y}_{1:n}$	-0.0053 (0.24)	-0.0099 (1.3)
	$\zeta_{1:n}$	0.0029 (0.33)	0.052 (1.6)
Zebrafish	$\mathbf{Y}_{1:n}$	0.0012 (0.13)	0.16 (2.4)
	$\zeta_{1:n}$	0.00046 (0.12)	0.050 (2.8)
Reviews	$\mathbf{Y}_{1:n}$	-0.0005 (0.29)	0.0018 (0.44)
	$\zeta_{1:n}$	-0.0015 (0.41)	0.061 (1.3)
S&P 500	$\mathbf{Y}_{1:n}$	0.0026 (7.7)	0.37 (9.4)
	$\zeta_{1:n}$	0.0028 (7.5)	0.39 (11)
Simulated	$\mathbf{Y}_{1:n}$	-0.0026 (1.6)	0.55 (8.7)
	$\zeta_{1:n}$	-0.0023 (1.8)	0.84 (2)