Sparse logistic regression analysis for 'gay', NYT headlines, Jan81 !! Jan07 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 20 40 60 80 100 120 140 160 180

Figure py Left panel: Sparse logistic regression allowing to visualize the image of the term "gay" in New York Times headlines from pxwp to qoovm Right panel: sparse graphical model of Senate voting datak obtained by recursive application of sparse logistic regressionm headlines containing the topic "gay" from those not containing that termm Thus for each month we obtained a short list of terms that are good predictors of the appearance of the query term in any headlinem Our proposed visualization of all the lists obtained in a sliding window fashion is shown in the left panel of figure pm The vertical axis refers to the terms in the resulting collection of listsk shown by order of appearancem Thusk the plot shows a staircase patternk where we have highlighted the terms with large weightsz these terms tell a consistent story of the evolution of the termk from "chorus" gwhen the term was not associated in the New York Times with homosexual behavior but only with New York's "gay men chorus" troupeh to "rights" to "marriage"m Some termsk such as "marriage"g occur over many yearsk as expected from the fact that they represent a current issue discussed in the mediam The plot confirms that it is possible to obtaink based on a purely statistical analysisk a historical summary of the evolution of the "image" of a topic in any text corporam

Sparse graphical models. We can build on sparse logistic regression models by regressing each feature against all the othersk in recursive fashionm The method gdescribed in 	 allows to compute a sparse graphical model,

where the absence of an edge denotes conditional independencem The right panel of figure p illustrates the results obtained on Senate voting data gqooslqoouhk and clearly shows that it allows to go beyond the simplistic results obtained via classical clustering methods gwhich typically only "discover" party lineshk and discover clustersk as well as important Senators that bridge these clustersm Applying the method to marketing data would allow to discover such clusters and bridges of features gbrandsk industriesk termsk etc

Sparse principal component analysis. Often it is desirable to obtain a global view of datak without any specific topic in mindm For examplek it can be interesting to plot di

different brands on a
twoldimensional plotk or to picture them in a graphm

Classical methods for doing sok such as principal component analysisk lack in a crucial aspecty the axes on which data is projected do not have any meaningful interpretationm We have recently prol sparse principal component analysis to provide such an interpretationy in that methodk

the axes on which data is projected are sparsek which means that they involve only a few featuresm