High-Dimensional Data and Dimension Reduction

Joshua Mitts*

The papers in this volume highlight the promise and potential of textual data in facilitating the computational analysis of law. Law is largely text, and text is uniquely high-dimensional: words might be conceptually related, but every word is distinct from another. Dimensionality is a double-edged sword: on the one hand, a large vocabulary allows for separating individual words from each other, but it is impossible to estimate correlations between an outcome and an entire vocabulary when the size of the vocabulary exceeds the number of observations. In the statistics literature, this is known as the "curse of dimensionality," and gives rise to techniques like regularization (e.g., the LASSO), topic modeling, and so forth (Hastie, Tibshirani, and Friedman 2009).

Dimension reduction is essential if one is to make sense of data where the number of parameters (i.e., size of a vocabulary) exceeds the number of observations (i.e., number of documents). But dimension reduction often entails making difficult choices which can shape the conclusions that are drawn from high-dimensional data. In this essay, I illustrate the promise and challenges of dimension reduction by discussing a recent working paper, (Bubb and Catan 2018), which addresses a question of

*Associate Professor of Law, Columbia Law School. I would like to thank Michael Livermore, Daniel Rockmore and participants in the Computational Study of the Law, held at the Sante Fe Institute on December 12-13, 2017.

first-order importance in the corporate-governance literature: the voting behavior of mutual funds.[1]

(Bubb and Catan 2018) identify latent structure in patterns of mutual fund voting on public-company shareholder proposals. Because any given fund only holds a relatively small number of portfolio companies, there is a tremendous dimensionality problem: 3,619 voters × 33,189 proposals = 120,089,277 potential fund-votes in their estimation sample. While their project does not employ a bag-of-words representation or model other textual data directly, the structure is very similar to textual data, as these proposals occur sparsely just like text: there are only 5,315,876 actual votes in their analysis sample, so 95.6% of their data cells are empty.

(Bubb and Catan 2018) estimate a two-dimensional latent factor model using Principal Components Analysis (PCA). After identifying two factors that explain the vast majority of the voting variance and are substantively interpretable, they apply cluster analysis to divide funds into three groups, which they call the Traditional Governance Party, the Shareholder Veto Party and the Shareholder Intervention Party. (Bubb and Catan 2018) show these clusters reflect a kind of "party structure": funds within these two groups vote consistently on a range of different kinds of shareholder proposals, especially on interventionist policies like proxy contents where dissident shareholders nominate a competing slate of directors. They argue that these parties "represent distinctive philosophies of shareholders' role in corporate governance."

This analysis reflects a tremendous step forward in understanding mutual-fund voting patterns. (Bubb and Catan 2018) is a highly innovative computational project that makes an outstanding contribution to the corporate-governance literature. That said, it does illustrate some of

1. Contemporaneous work by (Bolton et al. 2018) uses a slightly different method, W-NOMINATE, a which is just another form of unsupervised dimension reduction.

the inherent limitations of dimension reduction. Consider the use of PCA to represent latent factors. PCA projects a high-dimensional space onto vectors which capture the greatest variance in the underlying data. This can be intuitively visualized as drawing lines through high-dimensional space in the direction of the bulk of the data. This makes a great deal of intuitive sense, but are these vectors the "best" way to describe the data? To answer that question, we need to dive more deeply into what "best" might mean.

Suppose that true model (i.e., the "data generating process") underlying the setting in (Bubb and Catan 2018) is that fund j's expected vote on a proposal i at time t is given by a linear combination of proposal-specific and fund-specific characteristics, e.g.:

$$y_{i,j,t} = \alpha + \beta^{'}\mathbf{Z_i} + \gamma^{'}\mathbf{X_j} + \mathbf{\Delta}^{'}\left(\mathbf{X_j} \times \mathbf{Z_i}\right) + \epsilon_{i,j,t}$$

where $y_{i,j,t}$ is 1 if the fund voted for the proposal and 0 otherwise; $\mathbf{Z_i}$ is a vector of proposal-specific characteristics (e.g., the topic of the proposal); $\mathbf{X_j}$ is a vector of fund-specific characteristics (e.g., whether a fund is likely to vote for interventionist proposals); and $\epsilon_{i,j,t}$ is a random error term. Crucially, the length of $\mathbf{Z_i}$ + $\mathbf{X_j}$ exceeds the number of data points, so it is impossible to estimate the full parameter vectors β and γ directly, much less the interaction coefficients Δ. For this reason, it is necessary to engage in some kind of dimension reduction. There are several different ways to go about this. Three general possibilities are as follows:

Ex ante theory. The classical method underlying traditional inference in the social sciences is to choose a parsimonious subset of $\mathbf{Z_i}$ and $\mathbf{X_j}$ which are theoretically meaningful. Typically, this is obtained by specifying a model, either qualitatively or formally, which yields specific predictions. For example, one might theorize that mutual funds seek to maximize certain objectives, like increasing future fund flows or fee revenue, and voting is a means to achieving those ends. This hypothetical theory would yield a prediction as to a subset of $\mathbf{Z_i}$ and $\mathbf{X_j}$ which explains fund voting behavior. The relevant subset of $\mathbf{Z_i}$ might include whether the proposal involves a social or environmental issue that could matter to a

fund's investors, and the relevant subset of $\mathbf{X_j}$ might include the average age of the fund's investor base.

In the classical paradigm, empirical researchers would set out to test the hypothesis whether these theoretically motivated subsets of $\mathbf{Z_i}$ and $\mathbf{X_j}$ are correlated with differences in $y_{i,j,t}$, at a given level of statistical significance. However, any regression model like this one is effectively a predictive model as well. It is natural to ask how well do these subsets of $\mathbf{Z_i}$ and $\mathbf{X_j}$ predict fund voting behavior, $y_{i,j,t}$, as measured by accuracy, mean squared error, etc. In general, a high degree of statistical significance for a coefficient on $z_i \in \mathbf{Z_i}$ or $x_j \in \mathbf{X_j}$ does not imply that this coefficient does a good job predicting the outcome $y_{i,j,t}$.

Supervised learning. The limitations of traditional inference in the high-dimensional setting can be overcome by embracing predictive accuracy as the normative goal. The approach is to estimate a model which yields a prediction for $y_{i,j,t}$ given the data, which we can denote generically by $\hat{y}_{i,j,t} = f(\mathbf{Z_i}, \mathbf{X_j})$. The goal is to minimize some function of the deviations between $y_{i,j,t} - \hat{y}_{i,j,t}$ like overall accuracy, sensitivity (true positive rate) or specificity (true negative rate), which involves choosing some subset of $\mathbf{Z_i}$ and $\mathbf{X_j}$ which maximizes this objective function. This is precisely what regularization (e.g., the LASSO) does.

Much of the machine learning literature focuses on the problem of over-fitting, i.e., making sure that the chosen subset of $\mathbf{Z_i}$ and $\mathbf{X_j}$ is not tailored too tightly to the sample used to train the model (and thus will perform poorly on another sample). But the more fundamental challenge is that predictors which maximize this objective function in the data might not have a meaningful theoretical interpretation. Even a correctly trained model could yield predictors of fund voting behavior that do not map clearly onto a priori theory.

Unsupervised learning. Yet a third way to reduce dimensionality is to find a lower dimension representation of $\mathbf{Z_i}$ and $\mathbf{X_j}$ that is unrelated to $y_{i,j,t}$, at least at the training stage. This is the approach taken in (Bubb and Catan 2018), and the techniques here range from PCA to cluster analysis

to topic modeling (more common in the analysis of textual data). The challenge with unsupervised learning methods is that they lack an objective function. Without performance criteria, the degrees of freedom seem unbounded: how many dimensions of PCA should be employed? How many clusters (or topics) should the data be grouped into?

The answers to these questions are often ad hoc, leading to a concern that they may be subject to bias and lack the sort of objective falsification characteristic of hypothesis testing. For example, in (Bubb and Catan 2018), how can one reject the hypothesis that three clusters ("parties") are appropriate for as opposed to two or four? While their story makes a great deal of intuitive sense, it is difficult to identify a clear hypothesis test in this regard.

The key point here is not to bemoan the limitations of dimension reduction but rather to encourage researchers to employ all these methods in complementary fashion when engaging with high-dimensional data. There are at least two ways this can be done. One is to utilize these methods alongside each other in the same project. In previous work (Macey and Mitts 2014), we set out three theoretical rationales for corporate veil piercing (method #1), identified phrases predictive of veil-piercing decisions using multinomial inverse regression (method #2), and employed topic modeling to examine whether the co-occurrence of phrases corresponds to these theoretical rationales (method #3). Exploiting multiple methods gives readers confidence that the analysis is not driven by the idiosyncratic strengths and limitations of a single approach.

A second way to employ these methods in a complementary fashion is to test one with the other. For example, future work on mutual fund voting might compare a predictions using the two-dimensional PCA decomposition (see Figure 6 in (Bubb and Catan 2018)) to a classifier employing regularization among high-dimensional covariates. These covariates could include the full text of the proposal and fund strategy, e.g., $\mathbf{Z_i}$ is a $k \times 1$ vector of word and phrase frequencies in proposal itself, where k is the size of the fund proposal vocabulary, and $\mathbf{X_j}$ is a $l \times 1$ vector of word

and phrase frequencies in the fund's stated strategy, where l is the size of the fund strategy vocabulary. This approach would yield a subset of words and phrases for proposals and fund strategies, respectively, that are most predictive of voting outcomes.

These could be compared to the two-dimensional PCA decomposition with respect to (1) predictive accuracy — which predicts voting better; (2) theoretical fit – which seems to be the most consistent with existing theory; and (3) stability, i.e., does the party structure of mutual fund voting remain consistent when using free-form words and phrases as predictors? A powerful critique of (Macey and Mitts 2014) is that it fails to perform this kind of comparative analysis between different kinds of dimension reduction, raising the question whether the results are specific to the sequence of methods utilized there. At the dawn of a new era of computational analysis of the law, there is an exciting opportunity to guide the next generation of scholars in the "best practices" of working with high-dimensional data.

References

Bolton, Patrick, Tao Li, Enrichetta Ravina, and Howard Rosenthal. 2018. "Investor ideology."

Bubb, Ryan, and Emiliano Catan. 2018. "The party structure of mutual funds."

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. "Unsupervised learning." In *The elements of statistical learning*, 485–585. Springer.

Macey, Jonathan, and Joshua Mitts. 2014. "Finding order in the morass: The three real justifications for piercing the corporate veil." *Cornell L. Rev.* 100:99.