Simon Freyaldenhoven (Federal Reserve Bank of Philadelphia)

Title:

On the Testability of the Anchor Words Assumption in Topic Models

Abstract:

Topic models are a simple and popular tool for the statistical analysis of textual data. Their identification and estimation is typically enabled by assuming the existence of anchor words; that is, words that are exclusive to specific topics. In this paper we show that the existence of anchor words is statistically testable: there exists a test with correct size that has nontrivial power. This means that, in general, the anchor word assumption cannot be viewed simply as a convenient normalization. At the core of our result lies a simple characterization of when a column-stochastic matrix with known nonnegative rank admits a separable factorization. We test for the existence of anchor words in two different datasets derived from the transcripts of the Federal Open Market Committee (FOMC). We use a simulation study to analyze the power of a bootstrapped version of our suggested procedure and to discuss its computational limitations.