LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN INSTITUT FÜR STATISTIK

Surprises in topic model estimation and new Wasserstein document-distance calculations

Florentina Bunea
(Cornell University)

11.05.2022, 16.00 (c.t.)

Department of Statistics, Ludwigstr. 33, Room 144
and online via Zoom (Meeting-ID: 913-2473-4411; Password: StatsCol22)

Topic models have been and continue to be an important modeling tool for an ensemble of independent multinomial samples with shared commonality. Although applications of topic models span many disciplines, the jargon used to define them stems from text analysis. In keeping with the standard terminology, one has access to a corpus of n independent documents, each utilizing words from a given dictionary of size p. One draws N words from each document and records their respective count, thereby representing the corpus as a collection of n samples from independent, p-dimensional, multinomial distributions, each having a different, document specific, true word probability vector π. The topic model assumption is that each π is a mixture of K discrete distributions, that are common to the corpus, with document specific mixture weights. The corpus is assumed to cover K topics, that are not directly observable, and each of the K mixture components correspond to conditional probabilities of words, given a topic. The vector of the K mixture weights, per document, is viewed as a document specific topic distribution T, and is thus expected to be sparse, as most documents will only cover a few of the K topics of the corpus.

Despite the large body of work on learning topic models, the estimation of sparse topic distributions, of unknown sparsity, especially when the mixture components are not known, and are estimated from the same corpus, is not well understood and will be the focus of this talk. We provide estimators of T, with sharp theoretical guarantees, valid in many ractically relevant situations, including the scenario p >> N (short documents, sparse data) and unknown K. Moreover, the results are valid when dimensions p and K are allowed to grow with the sample sizes N and n. When the mixture components are known, we propose MLE estimation of the sparse vector T, the analysis of which has been open until now. The surprising result, and a remarkable property of the MLE in these models, is that, under appropriate conditions, and without further regularization, it can be exactly sparse, and contain the true zero pattern of the target.

When the mixture components are not known, we exhibit computationally fast and rate optimal estimators for them, and propose a quasi-MLE estimator of T, shown to retain the properties of the MLE. The practical implication of our sharp, finite-sample, rate analyses of the MLE and quasi-MLE reveal that having short documents can be compensated for, in terms of estimation precision, by having a large corpus.

Our main application is to the estimation of Wasserstein distances between document generating distributions. We propose, estimate and analyze Wasserstein distances between alternative probabilistic document representations, at the word and topic level, respectively. The effectiveness of the proposed Wasserstein distance estimates, and contrast with the more commonly used Word Mover Distance between empirical frequency estimates, is illustrated by an analysis of an IMDb movie reviews data set.

Biography:

Florentina Bunea obtained her Ph.D. in Statistics at the University of Washington, Seattle. She is now a Professor of Statistics in the Department Alrighof Statistics and Data Science, and she is affiliated with the Center for Applied Mathematics and the Department of Computer Science, at Cornell University. She is a fellow of the Institute of Mathematical Statistics, and she is or has been part of numerous editorial boards such as JRRS-B, JASA, Bernoulli, the Annals of Statistics. Her work has been continuously funded by the US National Science Foundation. Her most recent research interests include latent space models, topic models, and optimal transport in high dimensions.