DOI:10.1145/2076450.2076473

Technical Perspective Modeling High-Dimensional Data

By Santosh S. Vempala

DATA IN HIGH dimension is difficult to visualize and understand. This has always been the case and is even more apparent now with the availability of large high-dimensional datasets and the need to make sense of them.

The classical statistics approach to understanding data is to find a simple probabilistic model that could have generated it. The model is usually a probability distribution on a large domain and each data point is generated independently from the same distribution. While more complex models without the assumption of independent, identically distributed data points are also studied, the independent data model is the predominant one and is reasonable for many situations. Even this framework would not be very useful or interesting without further assumptions—the distribution could be considered uniform over the data encountered. The key is to model complex, large data with a *simple* distribution. Finding such a fit (if one exists) would likely give an insightful explanation and the parameters of the simple distribution might even have predictive powers.

What distribution to use? The central limit theorem suggests the most reasonable candidate would be a Gaussian distribution. This is what any aggregate distribution will eventually converge. Indeed, trying to find the single best-fit Gaussian to a given dataset is a commonly used and efficient approach. The Gaussian is estimated using the mean and covariance matrix of the data. Unfortunately, this works well only in rather special cases.

Thus we arrive at a widely used framework in statistics, called a *mixture model*. Here we assume that data is generated from a mixture of a small number of distributions of known type; the most common assumption is a mixture of Gaussians. The problem is to find the best-fit mixture of a small number of Gaussians to the given data. The number of component Gaussians, k, is much smaller than n, the ambient dimension. Unlike the case of a single Gaussian ($k = 1$), where it is straightforward to estimate the underlying Gaussian, the problem becomes much more difficult for general k. Even the case of a two-Gaussian mixture remained open for a long time.

Kalai, Moitra, and Valiant[3] show how to solve the problem for a mixture of two arbitrary n-dimensional Gaussians. Besides relying on a simple and ingenious reduction to the case of a mixture of two *1*-dimensional Gaussians, their analysis relies on the following fundamental fact about the identifiability of Gaussian mixtures: two distinct mixtures of Gaussians have different density functions; as the density of two mixtures gets closer, so must the mixtures (the means, variances, and mixing weights of one mixture must be approximated by those of the other). Such a property is not true for general mixtures, not even for mixtures of nice distributions such as those with logconcave densities, but it is true for Gaussian mixtures. Moreover, unlike the classical proof, they establish a polynomial bound on the number of samples needed to identify the components of a mixture to within a desired accuracy. Surprisingly, this is the first improvement on the sample complexity from the classical exponential bound, in spite of mixture models being studied for over a century.[5]

The key insight of their method is to show that a finite set of moments (six of them for the case of two 1-dimensional Gaussians) suffice to identify the components. With this tool in hand, they consider several random, 1-D projections of an n-dimensional mixture, identify the projections of the components in each, correctly cluster them according to component of origin, and thereby gather enough constraints on the original components to estimate their means, covariance matrices, and mixing weights.

In a follow-up paper, Moitra and Valiant[4] extended this approach to a mixture of k Gaussians, with complexity growing exponentially in k, but polynomially in all other parameters; thus a polynomial-time algorithm for any fixed k. A simple exponential dependence on k is unavoidable even in the sample complexity, at least if the goal is to identify the components of an arbitrary mixture of Gaussians with no separation condition. A similar bound was also proved independently using a different algorithm in a more abstract (and general) setting with possibly non-Gaussian components by Belkin and Sinha.[1]

The work presented in the following paper settles an important open problem, establishes fundamental facts and thereby advances classical statistics, and raises very interesting questions for computer science; among them: What can we hope to do for non-Gaussian mixtures (for which robust identifiability does not hold in general)? Can we handle Gaussian mixtures with some noise? In other words, is there an agnostic algorithm for learning Gaussian mixtures? Perhaps most interestingly, what reasonable assumptions lead to fully polynomial or even practical algorithms? (much work in the field assumes separation between components, which might be unavoidable for efficiency; for example, a polynomial-time algorithm is given assuming each component can be mostly separated by some hyperplane from the rest of the mixture;[2] one clean conjecture is that any probabilistically separable mixture is identifiable in polynomial-time).

References

1. Belkin, M. and Sinha, K. Polynomial learning of distribution families. *FOCS 2010*, 103–112.
2. Brubaker, S.C. and Vempala, S. Isotropic PCA and affine-invariant clustering. *Building Bridges. Bolyai Society Mathematical Studies* (special issue, M. Grötchel and G.O.H. Katona, Eds.). Also in *FOCS 2008*, 551–560.
3. Kalai, A., Moitra, A. and Valiant, G. Efficiently learning mixtures of two Gaussians. *STOC 2010*, 553–562.
4. Moitra, A. and Valiant, G. Settling the polynomial learnability of mixtures of Gaussians. *FOCS 2010*, 93–102.
5. Pearson, K. Contributions to the mathematical theory of evolution. *Philos. Trans. R. Soc. London 185* (1904), 71–110.

Santosh S. Vempala (vempala@cc.gatech.edu) is Distinguished Professor of Computer Science at Georgia Tech, Atlanta.