Abstract

Among the nonparametric methods of estimating the number of communities (K) in a community detection problem, methods based on the spectrum of the Bethe Hessian matrices ($\mathbf{H}_\zeta$ with the scalar parameter ζ) have garnered much popularity for their simplicity, computational efficiency, and robustness to the sparsity of data. For certain heuristic choices of ζ , such methods have been shown to be consistent for networks with N nodes with a common expected degree of $\omega(\log N)$. In this article, we obtain several finite sample results to show that if the input network is generated from either stochastic block models or degree-corrected block models, and if ζ is chosen from a certain interval, then the associated spectral methods based on $\mathbf{H}_\zeta$ is consistent for estimating K for the sub-logarithmic sparse regime, when the expected maximum degree is both $o(\log N)$ and $\omega(1)$, under some mild conditions even in the situation when K increases with N. We also propose a method to estimate the aforementioned interval empirically, which enables us to develop a consistent K estimation procedure in the sparse regime. We evaluate the performance of the resulting estimation procedure theoretically, also empirically through extensive simulation studies and application to a comprehensive collection of real-world network data.