Diagnosis from dimensionality reduction based on density estimation

David G. Aragonés, D. Cerezo-Wallis, A. Rubio-Ponce, I. Ballesteros, A. Hidalgo, G.F. Calvo

Instituto de Matemática Aplicada a la Ciencia y la Ingeniería, Universidad de Castilla-La Mancha

david.aragones@uclm.es

Resumen. Understanding the organization of immune cell states from high-dimensional transcriptomic data requires the development of rigorous mathematical frameworks. In this work, we introduce a quantitative methodology to characterize and compare neutrophil states using probability density estimation and information-theoretic overlap measures. Our approach begins with a representation $\{x_j, y_j\}$ of single-cell transcriptional data from murine neutrophils across diverse tissues and conditions, obtained using dimensionality reduction. To extract continuous density fields over the transcriptomic landscape, we employ kernel density estimation (KDE) with the Epanechnikov kernel. This procedure ensures accurate estimation even for multimodal or non-Gaussian distributions observed in neutrophil states. To quantify similarity between distributions associated with different physiological or pathological conditions, we compute the Bhattacharyya coefficient, which yields a normalized score in $[0, 1]$ measuring the degree of overlap between two probability density functions. Numerical evaluation of this integral is performed via adaptive Monte Carlo methods combining importance and stratified sampling, which reduce estimator variance by concentrating samples around density peaks. This mathematical framework allows us to represent cellular state similarity as a collection of overlap barcodes, providing a diagnostic tool for inferring host conditions from blood neutrophil distributions. Beyond its biological applications, our study highlights the utility of combining nonparametric density estimation, information-theoretic similarity measures, and advanced Monte Carlo integration in analyzing high-dimensional single-cell data.

Palabras clave: Kernel density estimation; Bhattacharyya coefficient; Monte Carlo integration; Machine learning; Data analysis.

Indicar la preferencia (subrayar la opción elegida): póster o charla.

Indicar la preferencia (subrayar la opción elegida): Lunes/Martes o Jueves/Viernes.