Stability, not goodness-of-fit, selects the number of cell types: from a T4 benchmark to transcriptomic atlases

Abstract

How many cell types are there? Goodness-of-fit criteria like BIC and ICL cannot distinguish discrete types from continuous within-type variation—retinotopic gradients, developmental position, activity-dependent wiring—and keep buying components indefinitely. We demonstrate this on a ground-truth benchmark: the four direction-selective T4 subtypes (T4a–d) in the *Drosophila* optic lobe. For 3,103 right-hemisphere T4 neurons from the FlyWire reconstruction, with features defined as log-synapse-counts onto post-synaptic cell types, both a factor model and a tied-covariance Gaussian mixture have monotone-decreasing BIC and ICL across $k = 1, \ldots, 30$, placing their minima at $k \geq 29$ even though the correct answer is 4. We propose a two-part criterion built on resampling stability: the partition must be reproducible under 80 % subsampling, measured by the mean pairwise adjusted-Rand index across bootstraps; and each further split $k \to k{+}1$ must be a rearrangement across clusters, not a nested sub-split of one cluster. On T4 with 85 fine-grained partner types, stability peaks at $k = 4$ with ARI $= 1.00$ across 435 bootstrap pairs; the partitions at $k = 5, 6, 7$ have stability > 0.97 but are perfect nested refinements of $k = 4$, each halving one subtype along the retinotopic axis. The criterion recovers the curated T4a–d labels at ARI $= 0.998$ (2 of 3,103 misassigned), where likelihood criteria fail by a factor of 7. To rule out circularity—partner labels come from the same annotation pipeline as the T4 labels—we repeat the analysis after coarsening every partner label to its class prefix (Mi1/Mi4/Mi9 $\to$ Mi; T1/T2/T4a/T4b/T5c $\to$ T); subtypes remain recoverable at ARI $= 0.942$, and with only 28 coarse features the criterion gracefully degrades to $k = 2$ rather than over-claiming. The method is modality-agnostic: we are extending it to the remaining optic-lobe columnar types (T5a–d, Tm1–Tm20, Mi1–Mi15) and to transcriptomic atlases, where ground-truth labels are unavailable and the selection problem is most acute.

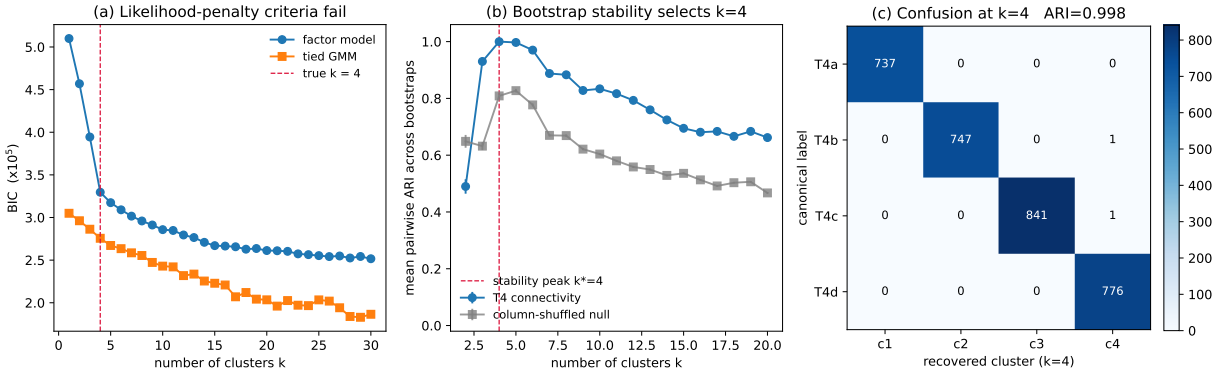

Figure 1: **A likelihood-penalty criterion fails to select the correct number of T4 subtypes; bootstrap stability succeeds. (a)** BIC as a function of the number of clusters k for a factor model (blue) and a tied-covariance Gaussian mixture (orange) fit to the 3103×85 log-connectivity matrix of right-hemisphere T4 neurons. Both criteria decline monotonically; their minima lie at $k \geq 29$, far beyond the four canonical T4 subtypes (dashed line). **(b)** Mean pairwise adjusted-Rand index between clusterings of 80 % subsamples ($B = 30$ replicates, 435 pairs per point) peaks sharply at $k^* = 4$ (ARI= 1.00) for T4 connectivity (blue) and clears a column-shuffled null control (grey) by ~ 0.19. **(c)** At $k = 4$, the recovered partition matches the curated T4a/b/c/d labels with adjusted Rand index 0.998; only 2 of 3,103 neurons are misassigned.