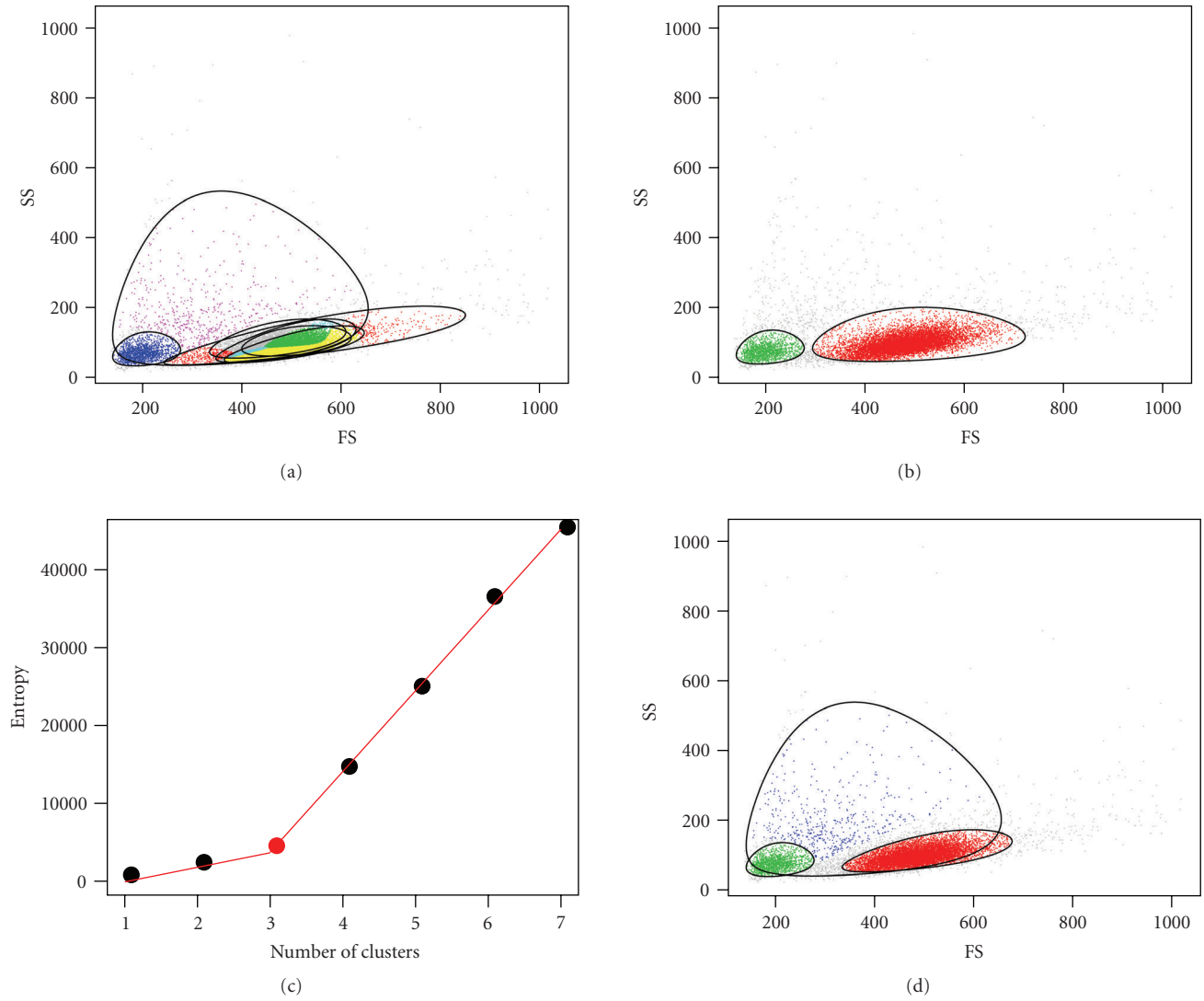

FIGURE 2: Examples of the flowClust$_{BIC}$, flowClust$_{ICL}$, flowMerge cluster solutions for forward versus side scatter in a sample of CLL flow cytometry data. (a) The flowClust$_{BIC}$ solution with seven clusters. (b) The flowClust$_{ICL}$ solution with two clusters. (c) The entropy versus number of clusters plot, fit to a two-component piecewise linear regression model. The best fitting model has a changepoint at three clusters. (d) The flowMerge solution corresponding to $K = 3$ clusters provides a better fit to the lymphocyte population than either the flowClust$_{BIC}$ or flowClust$_{ICL}$ solutions and provides a good estimate of the true number of cell populations.

these samples contain between two and three predominant cell populations that correspond to lymphocytes, debris, and outliers. The number of clusters identified by the flowClust$_{BIC}$ solution shows large variability across all samples. This solution generally required more mixture components than the true number of cell populations (median 6 clusters, range 3–15). Importantly, multiple components were often required to model the lymphocyte population (Figure 2(a)), which is the cell population of interest.

In contrast, the flowClust$_{ICL}$ fit is better but tends to underestimate the true number of cell populations. Across the 137 CLL samples, ICL identified a median of two populations per sample (range from 1 to 3). The ICL also provides a poor fit to the data, inadequately modeling the lymphocyte population (Figure 2(b)).

The flowMerge solution derived from the flowClust$_{BIC}$ solution provides both a good fit to the underlying data, including the lymphocyte cell population, as well as an improved estimate of the true number of cell populations (Figures 2(c) and 2(d)). The number of clusters estimated through merging is generally between the flowClust$_{BIC}$ and flowClust$_{ICL}$ solutions (median of 4 populations, range 2 to 8 clusters).

We performed automated gating in the fluorescence channels on the lymphocyte subpopulation derived from the previous autogating step. In 60/137 cases (43%), the GMM$_{BIC}$ solution returned more clusters than the flowClust$_{BIC}$ solution. In 95% of those cases the GMM$_{BIC}$ fit was within 5 components of the flowClust$_{BIC}$ fit. These two models returned an equal number of clusters in 29/137