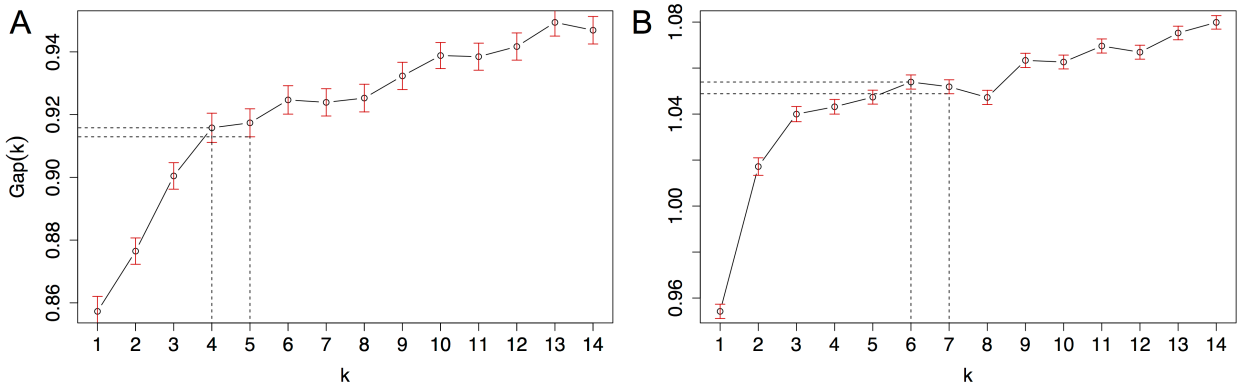

Figure e-1. **Gap statistics.** Plot of the gap statistic $\mathrm{Gap}(k)$ versus number of clusters with k-means on 500 bootstrapped samples of a) the domains clustering, and b) the symptoms clustering. Error bars represent ± 1 standard error (se). Per the method described in Tibshirani *et al*. (2001), the optimal number of clusters is the smallest k such that $\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - \mathrm{se}_{k+1}$. For the domains clustering, $k = 4$; for the symptoms clustering, $k = 6$. The gap statistic for the optimal k and the comparison to $k + 1$are marked with dotted lines.