Update 16/03/2020 Simulations

Stephen Coleman

16/03/2020

Cases

To test the idea of consensus inference we want to run simulations on a number of simple cases.

Note: I use $vec(\cdot)$ to denote a vector of a repeated value.

1. 2D Uniform: in this case $K = 1$ and $N = 100$ points are drawn from a uniform distribution for $p = 2$ features. The method should find a single clustering, but the 2D case is relatively challenging and it is possible that there will be misleading structure present.
2. 2D Gaussian: $p = 2$, $N = 100$, $K = 5$, $\pi = vec(1/K)$, $\mu = (-2, -2), (-2, 2), (0, 0), (2, -2), (2, 2)$, $\Sigma_k = \Sigma = (I)$. This is an easily visualised, trivial case included as a sense check.
3. Simple multivariate normal. Let each cluster have be defined by its mean which is constant across all features. $p = 10$, $N = 100$, $K = 5$, $\pi = vec(1/K)$, $\text{MVN}(\mu_k, \Sigma)$.
4. "Gene expression" case. Let each cluster be defined by up/down regulation in a number of genes. Let up regulation be represented by a univariate standard normal with mean 1.5, down regulation be represented by a univariate normal with mean -1.5 and expressing as normal be represented by a univariate Gaussian with mean 0. For example, let gene 1 be uniform across the range $(-2, 2)$; let gene 2 be up-regulated in cluster 1 and down regulated in clusters 2, 3 and 4; let gene 3 be up-regulated in clusters 2 and 3, down-regulated in cluster 4 and normally expressed in cluster 1; etc.
5. Repeat 4 with $\pi = (0.3, 0.1, 0.1, 0.5)$.
6. Extend idea of up and down regulation to a less "blocky" structure. Define K means (possibly by sampling from a univariate standard Gaussian distribution and/or a range $(0, \ldots, K)$. Ideally some means are relatively close to each other, others more separated. For each feature sample the mean associated with each cluster and make draws from this standard Gaussian. Consider something similar for drawing K values for σ from a Gamma dsitribution?
7. Consider 4 - 6 with an additional p_n noisy features. Should thep_s features containing signal be much less in quantity? What ratio? I was thinking $p_s << p_n$, e.g. $p_s = 100$, $p_n = 400$.

Guarantees on analysis

I have been trying to think about the problem of exploring a "good" space with the different chains. I think that there will be, similarly to random forest, some measure defined by ratio between the correlation between chains (or at least their initialisation) and chain-model strength (i.e. how well the individual chains describe the posterior) that we wish to minimise. I think showing that the first ten iterations (my expectation for the burn-in) is quite different across many of the chains would be reassuring. I think that we have no easy ability to judge chain model strength (maybe autocorrelation is sufficient) as clustering is unsupervised and there's no way of being sure we are sampling the posterior, so creating a statistic for this might not be feasible.