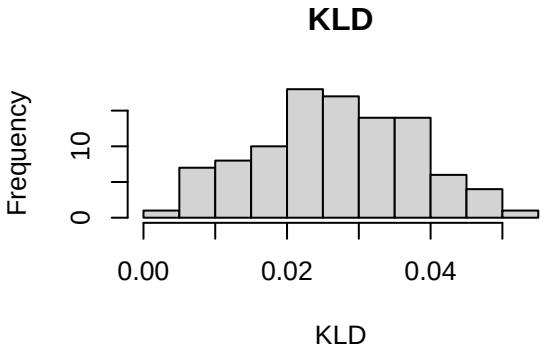
10/10/2024 Meeting

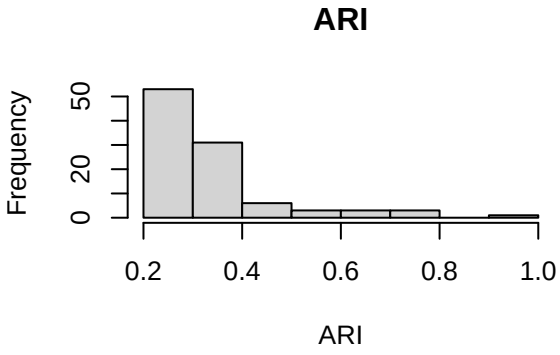
Jonathan Klus

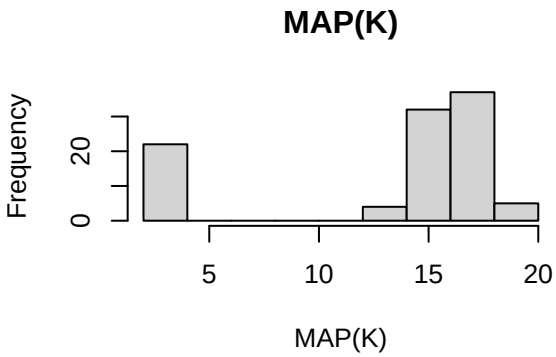
2024-10-10

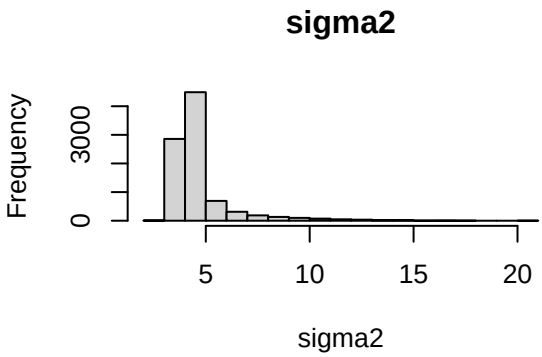
Close together scenario

n=300 with SM





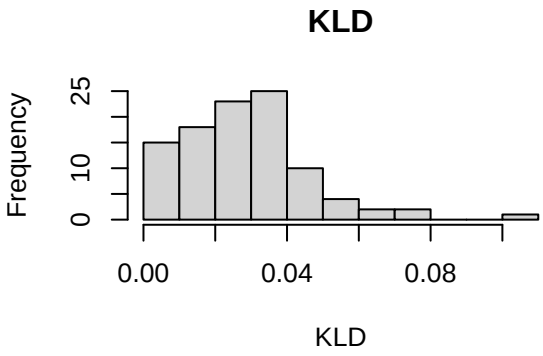


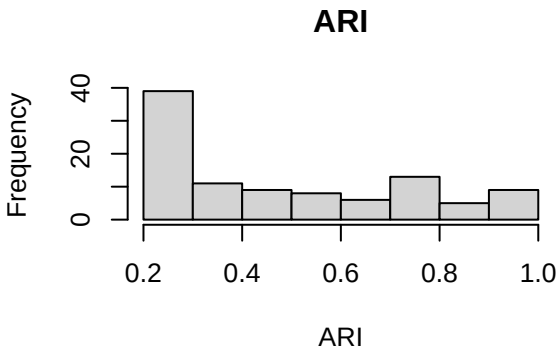


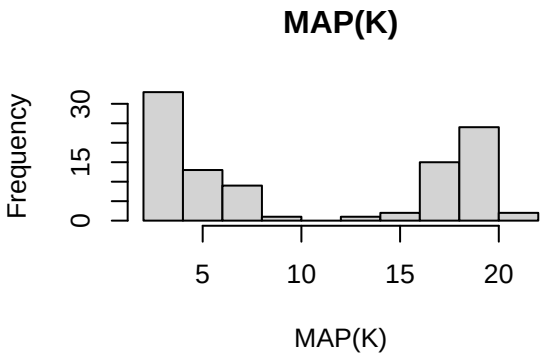

```
## [[1]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.55   4.32         1.26 3.75  7.86
##
## [[2]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.62   4.32         1.93 3.65  8.28
##
## [[3]]
```

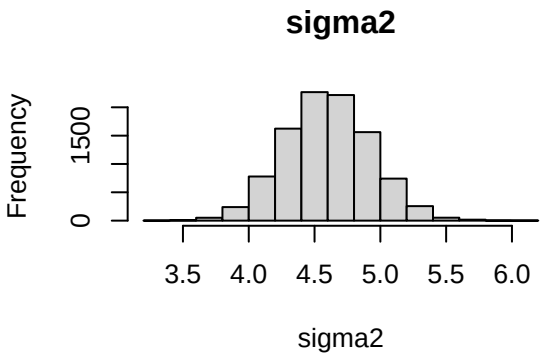
```
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.71   4.27         2.41  3.6 10.02
##
## [[4]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.53   4.19         2.17 3.54  9.09
##
## [[5]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.52   4.17         1.77 3.41  8.67
##
## [[6]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.59   4.21         2.17  3.4  9.27
##
## [[7]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.62   4.19         2.25 3.38  9.49
##
## [[8]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.68   4.15            3 3.32 10.14
##
## [[9]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.67    4.2         2.69  3.3  9.82
##
## [[10]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.75   4.25         2.65 3.34 10.28
##
## [[11]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.82   4.23         3.07 3.39  9.84
##
## [[12]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.89   4.34         2.97 3.38 10.51
##
## [[13]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1  5.1   4.42         4.13 3.44 11.01
```

n=300 without SM










```
## [[1]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.74   4.73         0.08  4.2  5.32
##
## [[2]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.66   4.65         0.09 4.09  5.26
##
## [[3]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.54   4.52          0.1 3.97  5.17
##
## [[4]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.49   4.48          0.1 3.91  5.16
##
## [[5]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1  4.4   4.39         0.11 3.78  5.05
##
## [[6]]
##           Mean Median Empirical SE 2.5% 97.5%
## sigma_1_1 4.37   4.35          0.1  3.7  5.06
```

KL divergence and overfitting

Let $p(x)$ and $q(x)$ represent the density of p and q evaluated at x, where $x \in \mathbb{R}^p$, and p is an integer greater than zero.

Then the Kullback–Leibler divergence (KLD) for a continuous density is defined as follows:

$$D_{KL}(p||q) = \int_{-\infty}^{\infty} p(x) log\left(\frac{p(x)}{q(x)}\right) dx$$

The KLD is defined on $\mathbb{R}^+$, the set of non-negative real numbers. When p and q are the same for all $x \in \mathcal{X}$, the the ratio in the log evaluates to 1, $log(1) = 0$, and KLD = 0.

When computing the KLD outside of a theoretical setting, we do not directly evaluate this integral since we are unable to observe all values of $x \in \mathbb{R}^p$. Instead, we compute a discrete approximation to the KLD by the true and estimated density at each of the observed data points $x \in \mathcal{X}$. Therefore our actual implementation of the KLD looks like its definition for a discrete density, which is:

$$D_{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) log\left(\frac{p(x)}{q(x)}\right)$$

There are a few possible issues associated with this implementation:

1. For problems with more data, our grid of points x will be larger and our estimate of the KLD will likely be better.
2. Let K denote the number of mixture components. In the realm of Bayesian nonparametric inference, we may approximate reference distribution p using a potentially infinite number of densities $q_1, \ldots, q_K$.

The second issue is of particular interest to our work. Even if the reality of the situation is that the reference distribution p is a mixture of three multivariate normal densities, we may closely estimate the true density p using a large number of multivariate normal densities. This constitutes overfitting, and is not penalized in any way by the KLD.

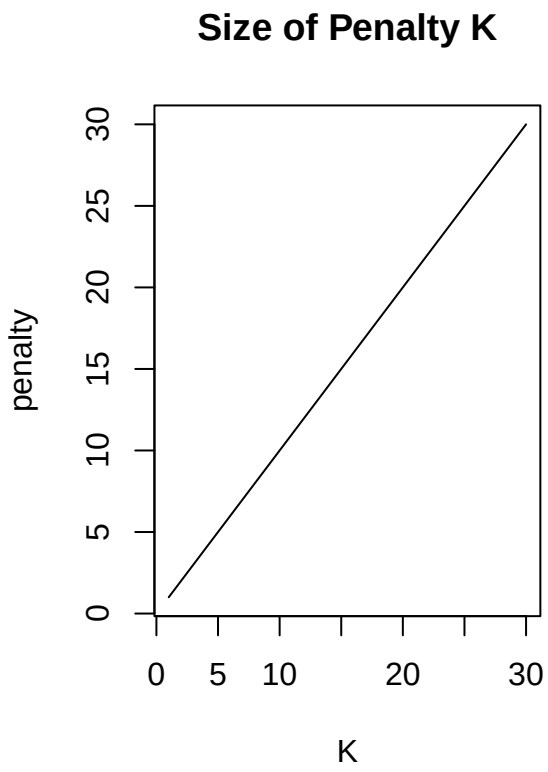
We wish to reward the simplest model that approximates the reference distribution accurately. This goal motivates the idea of penalizing the KL divergence in some way, but the question is how to do so.

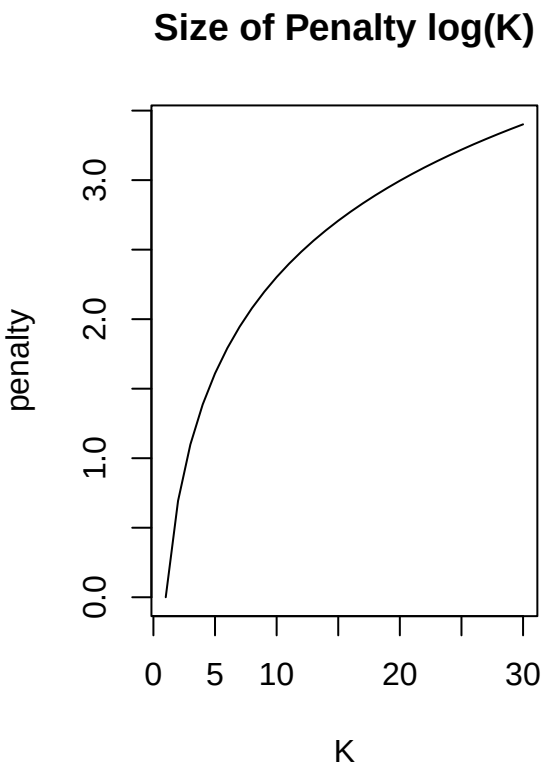
We could penalize:

1. The number of mixture components K that make up q.
2. The total number of parameters in q, which would address not just the number of mixture components but the complexity of the covariance matrix, for example.

What should this penalty term look like? Since larger values of the KLD indicate poor agreement between the estimated and reference densities, we want a penalty that will add positive value to the calculated KLD to adjust for model complexity.

1. K
2. $log(K)$

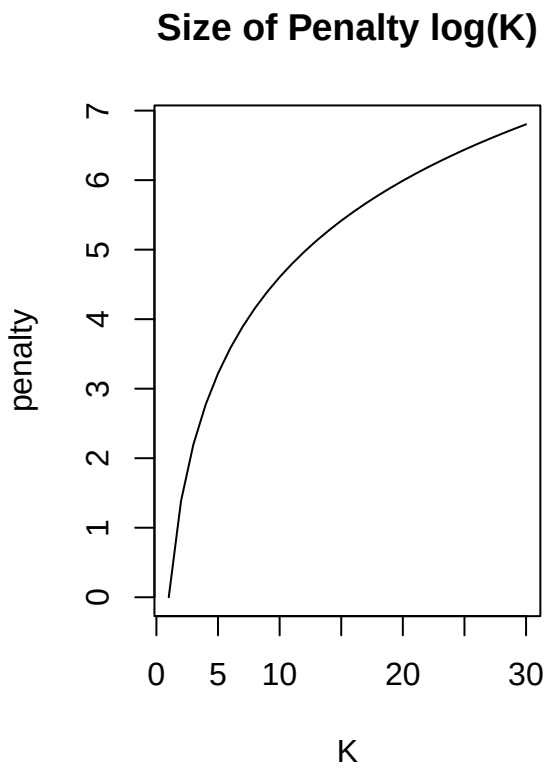


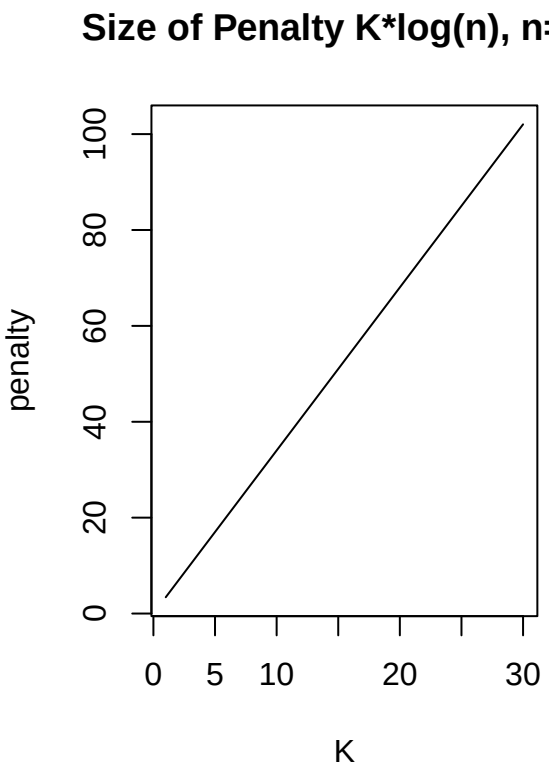


How do we choose a penalty that is of the appropriate magnitude?

How do we avoid biasing the divergence against mixtures that truly have a large number of components?

Both of the examples plotted above seem a bit too simplistic to be effective across a wide range of modeling scenarios. It may be beneficial to use a penalty structure more similar to the AIC $2k - 2log(L)$, where k is the total number of parameters estimated and L is the estimated likelihood. Or the BIC, $klog(n) - 2log(L)$.





Should we think about this as starting from the unadjusted KLD and adjust up to account for model complexity, or start from the penalty and see if the KLD can overcome it to get closer to 0?