M.-F. Ho

Report: Cross Validation for Bij Threshold

Composite SEDs
July 22, 2019

Grid Search : Bij Matrix

Calculating Bij matrix is a kind of grid-search method for searching maximum likelihood, which means we compute thoroughly every pairs of galaxies in the input data, $X = x_1, ..., x_N = f_{\lambda 1}, ..., f_{\lambda N}$, to get an optimal number of clusters. The computational expense of grid-search calculation is it usually grows as $O(N^2)$. While the size of the data grows, the computational expense will grow quadratically.

The formalism of Bij matrix is as simple as,

$$\begin{aligned} b_{ij} &= \sqrt{\frac{\Sigma(f_{\lambda i} - a_{ij} f_{\lambda j})^2}{\Sigma(f_{\lambda i})^2}} \\ a_{ij} &= \frac{\Sigma f_{\lambda i} \circ f_{\lambda j}}{f_{\lambda j}^2}, \end{aligned} \tag{1}$$

where b_{ij} is the similarity between a pair of galaxies and a_{ij} is the multiplication scaling factor between a pair of galaxies. The b_{ij} could also be considered as squared errors or the log Gaussian probabilities before normalization.

The grouping method of composite SEDs could be understood as an approximated version of Gaussian mixture model (GMM) with all indicators (π) and standard deviations (σ) of the Gaussians are equal to unity. The likelihood function for a Gaussian mixture with input data $X = x_i, ..., x_n$ is

$$p(X, Z \mid \mu, \sigma, \pi) = \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{I(Z_i = k)} N(x_i \mid \mu_k, \sigma_k)^{I(Z_i = k)}, \tag{2}$$

where μ_k is the mean of the kth group, σ_k is the standard deviation of the kth group, and π_k is the scale factor (for how high for the peak of kth Gaussian distribution) for kth group. The Eq 2 shows the likelihood of classifying input X into K groups with vectors of μ, σ, π, which parameterize the Gaussian mixture model.

We can rewrite the Bij calculation in the form of GMM, which allows us to get the likelihood function of our GridSearch method,

$$\begin{aligned} p(X, Z \mid \mu) &= \prod_{i=1}^{n} \prod_{k=1}^{K} N(x_i \mid \mu_k)^{I(Z_i = k)} \\ &\simeq \prod_{i=1}^{n} \prod_{k=1}^{K} \exp\left(-\frac{1}{2} \frac{\Sigma(f_{\lambda i} - a_{ik} f_{\lambda k})^2}{\Sigma(f_{\lambda k})^2}\right)^{I(Z_i = k)}. \end{aligned} \tag{3}$$

The $f_{\lambda k}$ can be understood as the rest-frame flux of primary galaxy in the kth group. The meaning of the likelihood function is the probability of input rest-frame fluxes $(f_{\lambda 1}, ..., f_{\lambda n})$

being explained by K Gaussian mixtures. Note that we approximate the means of Gaussian mixtures as the rest-frame fluxes of the primary galaxies.

Notice that the GMM of Bij have the same height and same width of Gaussians. I guess people made this choice because of the computational simplicity, since we only have to grid search for mean values in Eq 3. But note that it's still possible for introduce π and σ in our Bij calculation.

Cross Validation for Hyperparameters

One standard method in machine learning for preventing overfitting is using cross validation. The idea of cross validation is simple. We split the data into training set and validation set. We fit our model using only training set, and we use our trained model to predict the results of validation set. If we find our model performs well in the validation set, then it means our choice of hyperparameters is reasonable.

In the case of Bij calculation, the hyperparameter is "the upper bound of bij similarity (b_lim)". The b_lim is the largest value for a galaxy to be considered as a group member of the primary galaxy. In Kriek et al. (2011) paper, the choice of b_lim is $b < 0.05$. By using cross validation, it's possible to give a justifiable b_lim by maximizing or minimizing a chosen criterion.

Consider the case we randomly split the data $X = x_1, ..., x_N$ into training set $X^{(T)}$ and validation set $X^{(V)}$ (where $X^{(T)} \cup X^{(V)} = X$), and we perform the grid-search method with a given hyperparameter to find the composite SED groups,

$$\hat{\mu} = \text{GridSearch}(X^{(T)} \mid b < b_{lim}), \tag{4}$$

where $\hat{\mu} = f_{\lambda 1}, ..., f_{\lambda k}, ..., f_{\lambda M}$ is a vector of M mean values of M composite groups. Eq 4 means that we apply Bij calculation to training set ($X^{(T)}$) and set the upper bound of bij as b_{lim}, and then we find M mean values of composite groups.

To test how likely the composite groups we found in GridSearch is correct, we calculate the likelihood of the validation set belonging to $\hat{\mu} = f_{\lambda 1}, ..., f_{\lambda k}, ..., f_{\lambda M}$,

$$\begin{aligned} & p(X^{(V)} \mid \hat{\mu}, b < b_{lim}) \\ =& \hat{L} \sim \prod_{i=1}^{N^{(V)}} \prod_{k=1}^{M} \exp\left(-\frac{1}{2} \frac{\Sigma(f_{\lambda i} - a_{ik} f_{\lambda k})^2}{\Sigma(f_{\lambda k})^2} \right), \end{aligned} \tag{5}$$

where we simply just calculate how validation data deviated from the mixture Gaussians of composite SEDs. We use approximation sign here because we approximate the means of mixture Gaussians as rest-frame photometries of primary galaxies, according to the convention in the original paper.

If we change the choice of hyperparameter b_{lim}, the validation likelihood $\hat{L}$ will also change. The optimal choice of b_lim would be the one maximizing the validation likelihood $\hat{L}$.

Bayesian Information Criterion

In instead of maximizing validation likelihood, it's also sensible to choose other criterion to optimize our hyperparameter. One would expect if we have more composite groups, then the validation likelihood would also increase. This is the situation we want to avoid when we are running the clustering algorithm. Therefore, it would be more clever to choose a criterion which would penalize the number of clusters. Bayesian information criterion (BIC) could be a nice choice since it naturally penalizes the number of parameters used in the model.

The definition of BIC is,

$$\mathrm{BIC} = \ln(N) \times M - 2\ln\hat{L}, \tag{6}$$

where N is the number of observed data, M is the number of parameters, and $\hat{L}$ is the maximum likelihood estimation. The choice of b_lim with lowest BIC is preferred.