Examine the Gaussian Mixture Assumption

R2018a

Discriminant analysis assumes that the data comes from a Gaussian mixture model (see Creating Discriminant Analysis Model). If the data appears to come from a Gaussian mixture model, you can expect discriminant analysis to be a good classifier. Furthermore, the default linear discriminant analysis assumes that all class covariance matrices are equal. This section shows methods to test these assumptions.

Bartlett Test of Equal Covariance Matrices for Linear Discriminant Analysis

The Bartlett test (see Box [1]) checks equality of the covariance matrices of the various classes. If the covariance matrices are equal, the test indicates that linear discriminant analysis is appropriate. If not, consider using quadratic discriminant analysis, setting the `DiscrimType` name-value pair to `'quadratic'` in `fitcdiscr`.

The Bartlett test assumes normal (Gaussian) samples, where neither the means nor covariance matrices are known. To determine whether the covariances are equal, compute the following quantities:

- Sample covariance matrices per class σ_i, $1 \le i \le k$, where k is the number of classes.
- Pooled-in covariance matrix σ.
- Test statistic V:

$$V = (n-k)\log(|\Sigma|) - \sum_{i=1}^{k}(n_i - 1)\log(|\Sigma_i|)$$

where n is the total number of observations, and n_i is the number of observations in class i, and $|\Sigma|$ means the determinant of the matrix Σ.

- Asymptotically, as the number of observations in each class n_i become large, V is distributed approximately χ^2 with $kd(d+1)/2$ degrees of freedom, where d is the number of predictors (number of dimensions in the data).

The Bartlett test is to check whether V exceeds a given percentile of the χ^2 distribution with $kd(d+1)/2$ degrees of freedom. If it does, then reject the hypothesis that the covariances are equal.

Example: Bartlett Test for Equal Covariance Matrices

Check whether the Fisher iris data is well modeled by a single Gaussian covariance, or whether it would be better to model it as a Gaussian mixture.

```
load fisheriris;
prednames = {'SepalLength','SepalWidth','PetalLength','PetalWidth'};
L = fitcdiscr(meas,species,'PredictorNames',prednames);
Q = fitcdiscr(meas,species,'PredictorNames',prednames,'DiscrimType','quadratic');
D = 4; % Number of dimensions of X
Nclass = [50 50 50];
N = L.NumObservations;
K = numel(L.ClassNames);
SigmaQ = Q.Sigma;
Sigmai = L.Sigma;
logV = (N-K)*log(det(Sigmai));
for k=1:K
    logV = logV - (Nclass(k)-1)*log(det(SigmaQ(:,:,k)));
end
nu = (K-1)*D*(D+1)/2;
pval = 1 - chi2cdf(logV,nu)
```

```
pval = 0
```

The Bartlett test emphatically rejects the hypothesis of equal covariance matrices. If `pval` had been greater than `0.05`, the test would not have rejected the hypothesis. The result indicates to use quadratic discriminant analysis, as opposed to linear discriminant analysis.

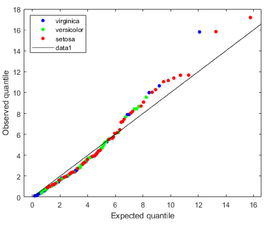
Q-Q Plot

A Q-Q plot graphically shows whether an empirical distribution is close to a theoretical distribution. If the two are equal, the Q-Q plot lies on a 45° line. If not, the Q-Q plot strays from the 45° line.

Check Q-Q Plots for Linear and Quadratic Discriminants

For linear discriminant analysis, use a single covariance matrix for all classes.

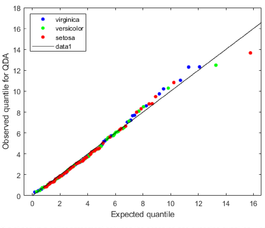
```
load fisheriris;
prednames = {'SepalLength','SepalWidth','PetalLength','PetalWidth'};
L = fitcdiscr(meas,species,'PredictorNames',prednames);
N = L.NumObservations;
K = numel(L.ClassNames);
mahL = mahal(L,L.X,'ClassLabels',L.Y);
D = 4;
expQ = chi2inv(((1:N)-0.5)/N,D); % expected quantiles
[mahL,sorted] = sort(mahL); % sorted observed quantiles
figure;
gscatter(expQ,mahL,L.Y(sorted),'bgr',[],[],'off');
legend('virginica','versicolor','setosa','Location','NW');
xlabel('Expected quantile');
ylabel('Observed quantile');
line([0 20],[0 20],'color','k');
```



Overall, the agreement between the expected and observed quantiles is good. Look at the right half of the plot. The deviation of the plot from the 45° line upward indicates that the data has tails heavier than a normal distribution. There are three possible outliers on the right: two observations from class `'setosa'` and one observation from class `'virginica'`.

As shown in Bartlett Test of Equal Covariance Matrices for Linear Discriminant Analysis, the data does not match a single covariance matrix. Redo the calculations for a quadratic discriminant.

```
load fisheriris;
prednames = {'SepalLength','SepalWidth','PetalLength','PetalWidth'};
Q = fitcdiscr(meas,species,'PredictorNames',prednames,'DiscrimType','quadratic');
Nclass = [50 50 50];
N = L.NumObservations;
K = numel(L.ClassNames);
mahQ = mahal(Q,Q.X,'ClassLabels',Q.Y);
expQ = chi2inv(((1:N)-0.5)/N,D);
[mahQ,sorted] = sort(mahQ);
figure;
gscatter(expQ,mahQ,Q.Y(sorted),'bgr',[],[],'off');
legend('virginica','versicolor','setosa','Location','NW');
xlabel('Expected quantile');
ylabel('Observed quantile for QDA');
line([0 20],[0 20],'color','k');
```



The Q-Q plot shows a better agreement between the observed and expected quantiles. There is only one outlier candidate, from class `'setosa'`.

Mardia Kurtosis Test of Multivariate Normality

The Mardia kurtosis test (see Mardia [2]) is an alternative to examining a Q-Q plot. It gives a numeric approach to deciding if data matches a Gaussian mixture model.

In the Mardia kurtosis test you compute M, the mean of the fourth power of the Mahalanobis distance of the data from the class means. If the data is normally distributed with constant covariance matrix (and is thus suitable for linear discriminant analysis), M is asymptotically distributed as normal with mean $d(d+2)$ and variance $8d(d+2)/n$, where

- d is the number of predictors (number of dimensions in the data).
- n is the total number of observations.

The Mardia test is two sided: check whether M is close enough to $d(d+2)$ with respect to a normal distribution of variance $8d(d+2)/n$.

Example: Mardia Kurtosis Test for Linear and Quadratic Discriminants

Check whether the Fisher iris data is approximately normally distributed for both linear and quadratic discriminant analysis. According to Bartlett Test of Equal Covariance Matrices for Linear Discriminant Analysis, the data is not normal for linear discriminant analysis (the covariance matrices are different). Check Q-Q Plots for Linear and Quadratic Discriminants indicates that the data is well modeled by a Gaussian mixture model with different covariances per class. Check these conclusions with the Mardia kurtosis test:

```
load fisheriris;
prednames = {'SepalLength','SepalWidth','PetalLength','PetalWidth'};
L = fitcdiscr(meas,species,'PredictorNames',prednames);
mahL = mahal(L,L.X,'ClassLabels',L.Y);
D = 4;
N = L.NumObservations;
obsKurt = mean(mahL.^2);
expKurt = D*(D+2);
varKurt = 8*D*(D+2)/N;
[~,pval] = ztest(obsKurt,expKurt,sqrt(varKurt))
```

```
pval = 0.0208
```

The Mardia test indicates to reject the hypothesis that the data is normally distributed.

Continuing the example with quadratic discriminant analysis:

```
Q = fitcdiscr(meas,species,'PredictorNames',prednames,'DiscrimType','quadratic');
mahQ = mahal(Q,Q.X,'ClassLabels',Q.Y);
obsKurt = mean(mahQ.^2);
[~,pval] = ztest(obsKurt,expKurt,sqrt(varKurt))
```

```
pval = 0.7230
```

Because `pval` is high, you conclude the data are consistent with the multivariate normal distribution.

References

[1] Box, G. E. P. A General Distribution Theory for a Class of Likelihood Criteria. Biometrika 36(3), pp. 317–346, 1949.

[2] Mardia, K. V. Measures of multivariate skewness and kurtosis with applications. Biometrika 57 (3), pp. 519–530, 1970.

See Also

Functions

`cvshrink` | `fitcdiscr`

Objects

`ClassificationDiscriminant` | `gmdistribution`

Related Topics

- Discriminant Analysis Classification
- Bartlett Test of Equal Covariance Matrices for Linear Discriminant Analysis
- Check Q-Q Plots for Linear and Quadratic Discriminants