LARS Information Note 040469
Purdue University

TESTING THE GAUSSIAN ASSUMPTION ON AIRCRAFT DATA

When feature selection and classification are done using LARSYSAA, the user makes the assumption that the data to be classified comes from a multivariate Gaussian distribution. This assumption is inherent in the algorithms used and if the data is far from Gaussian, results are unpredictable.

An experiment was conducted to test the assumption that aircraft data is Gaussianly distributed in each individual feature. Notice that the assumption that must be satisfied is that of multivariate Gaussian. If the data is Gaussian in each feature, this does not guarantee that it will be multivariate Gaussian. However, if it is not Gaussian in any particular feature, it is not multivariate Gaussian. Since no such test for multivariate data is known, it was felt that it would be useful to test the univariate assumption.

The method used is that of the Chi-Square Goodness of Fit test. The data is histogrammed and a vector of observed values (OBS(I)) is formed. A vector of expected values corresponding to the observed values is calculated. Given the mean and standard deviation of the data, the expected (EXP(I)) number of samples falling in each bin (given a Gaussian distribution) is calculated. If k bins are formed, the test values

$$\chi^2 = \sum_{i=1}^{k} (\mathrm{OBS}(I) - \mathrm{EXP}(I))^2/\mathrm{EXP}(I)$$

If χ^2 is less than $\chi^2_{1-\alpha}$ with k-3 degrees of freedom, the hypothesis that the data is Gaussianly distributed can be accepted with probability $1-\alpha$. Notice

that with a constant number of degrees of freedom, the smaller the χ^2 value, the better the data fits the hypothesized density. (This test can be used to hypothesize any density.)

A program was written (CHi2-Serial No. DA 0011) to analyze aircraft data in this manner. Training fields and classes and large areas (500 lines) were tested. (See program output in file.)

The following is a sample output.

Class stub
No. of samples = 960
Channel 12
Mean 166.92
St. Dev. 4.56

(bin)	EXP	OBS	CHI
155	4.3	5	0.1
157	9.9	11	0.2
159	25.4	35	3.9
161	53.6	55	3.9
163	93.9	109	6.3
165	136.1	138	6.4
167	163.2	172	6.8
169	162.0	164	6.9
171	133.1	127	7.2
173	90.6	79	8.6
175	51.0	41	10.6
177	23.8	15	13.8
179	9.2	5	15.7
225	3.9	4	15.7

XCHI = 15.7 with 11 degrees of freedom

If that χ^2 value with 13 degrees of freedom is looked up in a Chi-Square table, it can be seen that the hypothesis can be accepted with probability .10. After testing a random sample of fields and classes in this manner, it was

found that accepting the hypothesis with even that much probability was unusual. Most times, the χ^2 values obtained were much bigger and off the table.

A closer look at the observed and expected values for each bin shows that most discrepancies between the two occur at each end. The aircraft data does not have enough of a range to be really Gaussianly distributed.

A Laplace distribution was also fitted to the data. A Laplace density resembles a Gaussian distribution in general shape but has a much higher kurtosis (peak), and a smaller range. x is said to be distributed as a Laplace distribution if

$$p(x) = \int_{-\infty}^{\infty} \frac{1}{2\beta} \cdot e^{-\frac{|x-\alpha|}{\beta}} \cdot \quad -\infty < x < \infty$$

where α is the mean and $2\beta^2$ is the variance.

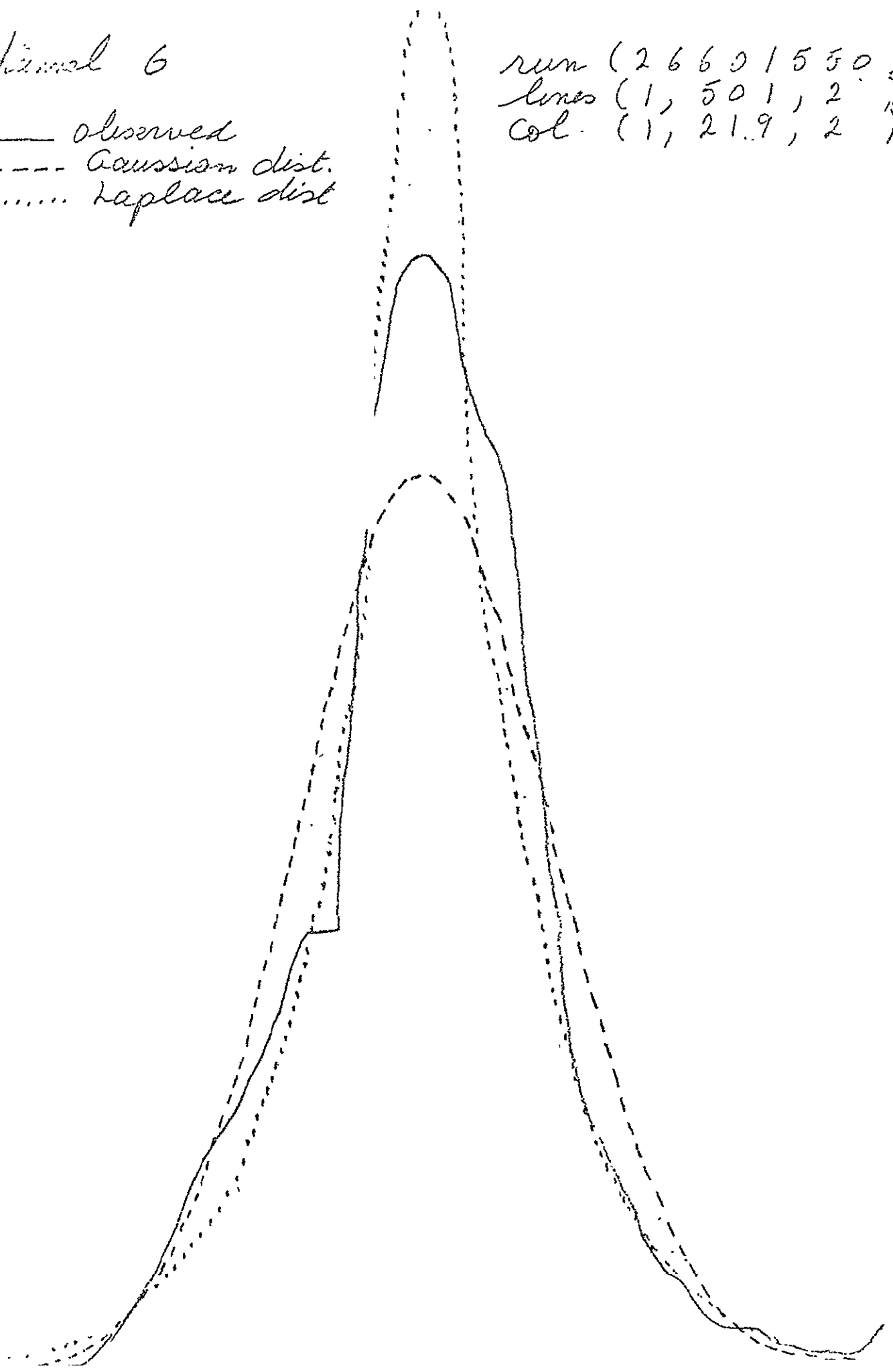
Results were about of the same magnitude. A Laplace distribution fit the data better in some channels, worse in others. However, the hypothesis that the data was distributed as a Laplace distribution could not be accepted with greater probability. The accompanying graphs show an area of 500 lines of run 26601550 (September 1966 - C3). The observed data lies in between the Gaussian and Laplace distribution. In certain places (channel 10) it is bimodal.

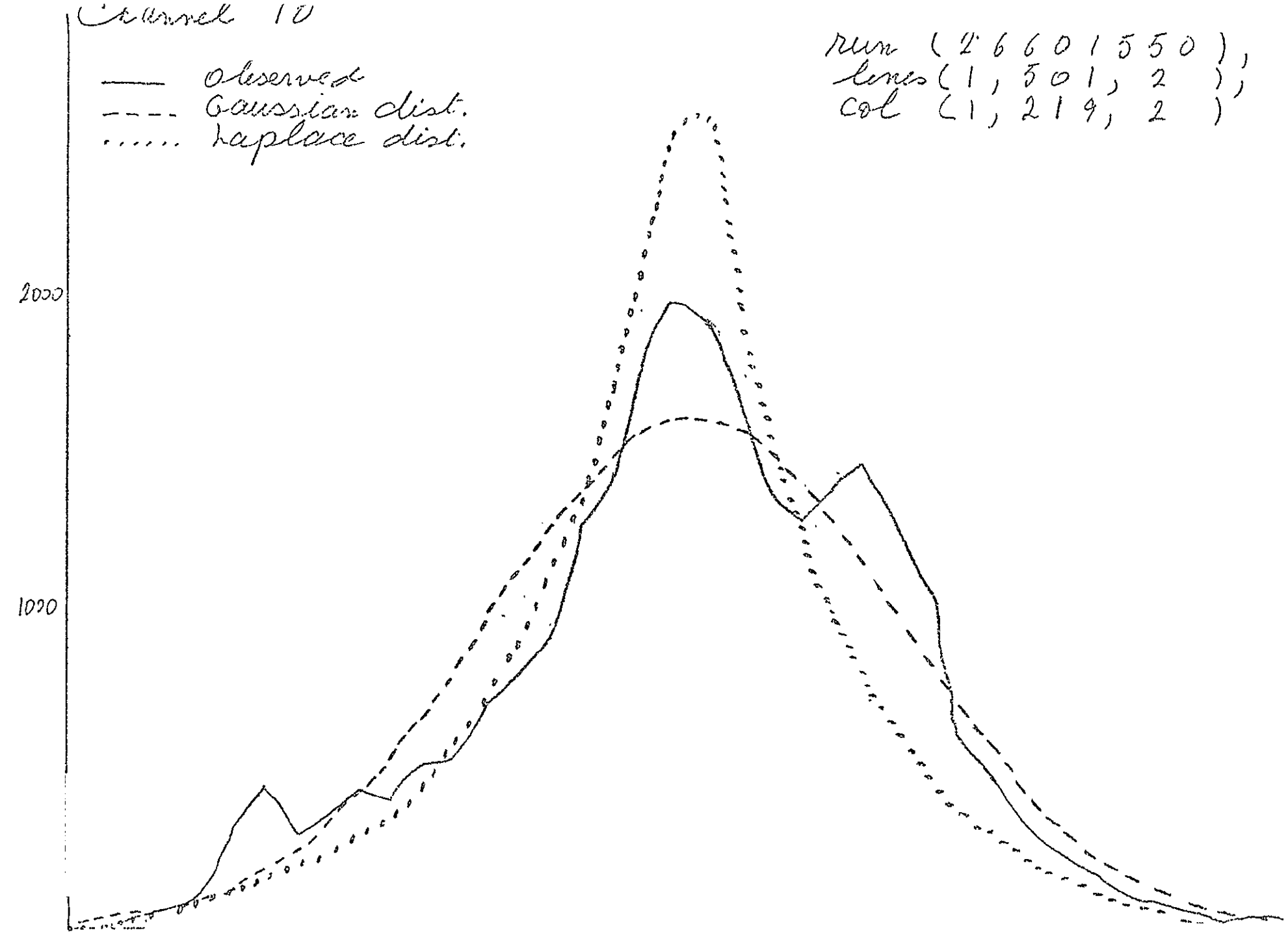
The aircraft data analyzed was found not to be Gaussianly distributed. Nothing can really be said about the population to which it belongs. However, because of the small range of the data, it is doubtful whether the population is Gaussian. It is therefore possible that errors occurring between classes that are close together are due to non-Gaussian data.

Chanel 6

run (26621550),
lines (1, 501, 2),
col. (1, 21.9, 2)

—— observed
- - - - Gaussian dist.
..... Laplace dist




channel 10
observed
Gaussian dist.
Laplace dist.
run (2660150),
lines(1, 501, 2),
col (1, 219, 2)
2000
1000

Channel 12

—— observed
---- Gaussian dist.
..... laplace dist.

run (2 6 6 0 1 5 5 0),
lines (1, 5 0 1, 2),
col (1, 2 1 9, 2)