STT 843: Multivariate Analysis

7. The Multivariate Normal Distribution (Chapter 4.7-4.8)

Guanqun Cao

Department of Statistics and Probability
Michigan State University

Spring 2026

Outline

1. Evaluating Bivariate Normality

2. Detecting Outliers and Cleaning Data

Evaluating Bivariate Normality

- By Result 3.7, the set of bivariate outcomes $\mathbf{x}$ such that

$$(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq \chi_2^2(0.5)$$

  has probability 0.5.

- Thus we should expect roughly the same percentage, 50%, of sample observations lie in the ellipse given by

$$\left\{ \text{ all } \mathbf{x} \text{ such that } (\mathbf{x}-\hat{\mathbf{x}})^{\mathsf{T}}\mathbf{S}^{-1}(\mathbf{x}-\hat{\mathbf{x}}) \leq \chi_2^2(0.5) \right\}$$

  where $\boldsymbol{\mu}$ is replaced by $\hat{\mathbf{x}}$ and Σ^{-1} by its estimate $\mathbf{S}^{-1}$. If not, the normality assumption is suspect.

Example 3.12 (Checking bivariate normality)

Although not a random sample, data consisting of the pairs of observations ($x_1 =$ sales, $x_2 =$ profits) for the 10 largest companies in the world are listed in the following table. Check if (x_1, x_2) follows bivariate normal distribution.

The World's 10 Largest Companies [1]

Company	x_1 = sales (billions)	x_2 = profits (billions)	x_3 = assets (billions)
Citigroup	108.28	17.05	1,484.10
General Electric	152.36	16.59	750.33
American Intl Group	95.04	10.91	766.42
Bank of America	65.45	14.14	1,110.46
HSBC Group	62.97	9.52	1,031.29
ExxonMobil	263.99	25.33	195.26
Royal Dutch/Shell	265.19	18.54	193.83
BP	285.06	15.73	191.11
ING Group	92.01	8.10	1,175.16
Toyota Motor	165.68	11.13	211.15

[1] From www.Forbes.com partially based on Forbes The Forbes Global 2000 , April 18, 2005.

- A somewhat more formal method for judging normality of a data set is based on the squared generalized distances

$$d_j^2 = (\mathbf{x}_j - \overline{\mathbf{x}})^{\mathsf{T}} \mathbf{S}^{-1} (\mathbf{x}_j - \overline{\mathbf{x}})$$

- When the parent population is multivariate normal and both n and $n - p$ are greater than 25 or 30 , each of the squared distance $d_1^2, d_2^2, \ldots, d_n^2$ should behave like a chi-square random variable.

Although these distances are not independent or exactly chi-square distributed, it is helpful to plot them as if they were. The resulting plot is called a chi-square plot or gamma plot, because the chi-square distribution is a special case of the more general gamma distribution. To construct the chi-square plot

1. Order the square distance in the equation above from smallest to largest as $d^2_{(1)} \leq d^2_{(2)} \leq \cdots \leq d^2_{(n)}$.
2. Graph the pairs $\left(q_{c,p}\left(\left(j-\frac{1}{2}\right)/n\right), d^2_{(j)}\right)$, where $q_{c,p}\left(\left(j-\frac{1}{2}\right)/n\right)$ is the $100(j-\ \frac{1}{2})\ /n$ quantile of the chi-square distribution with p degrees of freedom.

Example 3.13 (Constructing a chi-square plot)

Let us construct a chi-square plot of the generalized distances given in Example 3.12. The order distance and the corresponding chi-square percentile for $p = 2$ and $n = 10$ are listed in the following table:

j	$d^2_{(j)}$	$q_{c,2}\left(\frac{j-\frac{1}{2}}{10}\right)$
1	.30	.10
2	.62	.33
3	1.16	.58
4	1.30	.86
5	1.61	1.20
6	1.64	1.60
7	1.71	2.10
8	1.79	2.77
9	3.53	3.79
10	4.38	5.99

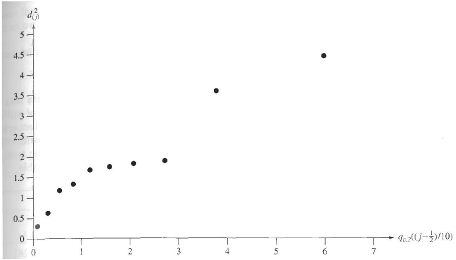

Figure 4.7 A chi-square plot of the ordered distances in Example 3.13.

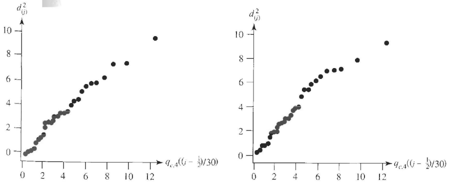

Figure 4.8 Chi-square plots for two simulated four-variate normal data sets with $n = 30$.

Example 3.14 (Evaluating multivariate normality for a four-variable data set)

The data in Table 4.3 were obtained by taking four different measures of stiffness, x_1, x_2, x_3, and x_4, of each of $n = 30$ boards. the first measurement involving sending a shock wave down the board, the second measurement is determined while vibrating the board, and the last two measurements are obtained from static tests. The squared distances $d_j = (\mathbf{x}_j - \overline{\mathbf{x}})^{\mathsf{T}} \mathbf{S}^{-1} (\mathbf{x}_j - \overline{\mathbf{x}})$ are also presented in the table

Table 4.3 Four Measurements of Stiffness

	x_1	x_2	x_3	x_4	d^2		x_1	x_2	x_3	x_4	d^2
1	1889	1651	1561	1778	.60	16	1954	2149	1180	1281	16.85
2	2403	2048	2087	2197	5.48	17	1325	1170	1002	1176	3.50
3	2119	1700	1815	2222	7.62	18	1419	1371	1252	1308	3.99
4	1645	1627	1110	1533	5.21	19	1828	1634	1602	1755	1.36
5	1976	1916	1614	1883	1.40	20	1725	1594	1313	1646	1.46
6	1712	1712	1439	1546	2.22	21	2276	2189	1547	2111	9.90
7	1943	1685	1271	1671	4.99	22	1899	1614	1422	1477	5.06
8	2104	1820	1717	1874	1.49	23	1633	1513	1290	1516	.80
9	2983	2794	2412	2581	12.26	24	2061	1867	1646	2037	2.54
10	1745	1600	1384	1508	.77	25	1856	1493	1356	1533	4.58
11	1710	1591	1518	1667	1.93	26	1727	1412	1238	1469	3.40
12	2046	1907	1627	1898	.46	27	2168	1896	1701	1834	2.38
13	1840	1841	1595	1741	2.70	28	1655	1675	1414	1597	3.00
14	1867	1685	1493	1678	.13	29	2326	2301	2065	2234	6.28
15	1859	1649	1389	1714	1.08	30	1490	1382	1214	1284	2.58

Source: Data courtesy of William Galligan.

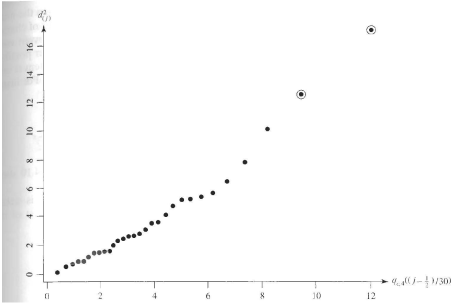

Figure 4.9 A chi-square plot for the data in Example 4.14.

Detecting Outliers and Cleaning Data

- Outliers are best detected visually whenever this is possible
- For a single random variable, the problem is one dimensional, and we look for observations that are far from the others.
- In the bivariate case, the situation is more complicated. Figure 4.10 shows a situation with two unusual observations.
- In higher dimensions, there can be outliers that cannot be detected from the univariate plots or even the bivariate scatter plots. Here a large value of $(\mathbf{x}_j - \bar{x})^{\mathsf{T}} \mathbf{S}^{-1} (\mathbf{x}_j - \bar{x})$ will suggest an unusual observation. even though it cannot be seen visually.

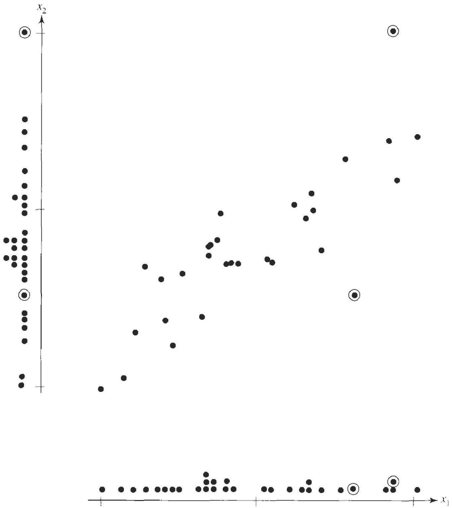

Figure 4.10 Two outliers; one univariate and one bivariate.

Steps for Detecting Outliers

1. Math a dot plot for each variable.
2. Make a scatter plot for each pair of variables.
3. Calculate the standardize variable $z_{jk} = (x_{jk} - \bar{x}_k) / \sqrt{s_{kk}}$ for $j = 1, 2, \ldots, n$ and each column $k = 1, 2, \ldots, p$. Examine these standardized values for large or small values.
4. Calculate the generalized squared distance $(\mathbf{x}_j - \overline{\mathbf{x}})^{\mathsf{T}} \mathbf{S}^{-1} (\mathbf{x}_j - \overline{\mathbf{x}})$. Examine these distances for unusually values. In a chi-square plot, these would be the points farthest from the origin.

x_1	x_2	x_3	x_4	x_5	z_1	z_2	z_3	z_4	z_5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
1631	1528	1452	1559	1602	.06	-.15	.05	.28	-.12
1770	1677	1707	1738	1785	.64	.43	1.07	.94	.60
1376	1190	723	1285	2791	-1.01	-1.47	-2.87	-.73	4.57
1705	1577	1332	1703	1664	.37	.04	-.43	.81	.13
1643	1535	1510	1494	1582	.11	-.12	.28	.04	-.20
1567	1510	1301	1405	1553	-.21	-.22	-.56	-.28	-.31
1528	1591	1714	1685	1698	-.38	.10	1.10	.75	.26
1803	1826	1748	2746	1764	.78	1.01	1.23	4.65	.52
1587	1554	1352	1554	1551	-.13	-.05	-.35	.26	-.32
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Example 3.15 (Detecting outliers in the data on lumber)

Table 4.4 contains the data in Table 4.3, along with the standardized observations. These data consist of four different measurements of stiffness x_1, x_2, x_3 and x_4, on each $n = 30$ boards. Detect outliers in these data.

x_1	x_2	x_3	x_4	no.	z_1	z_2	z_3	z_4	d^2
1889	1651	1561	1778	1	-.1	-.3	.2	.2	.60
2403	2048	2087	2197	2	1.5	.9	1.9	1.5	5.48
2119	1700	1815	2222	3	.7	-.2	1.0	1.5	7.62
1645	1627	1110	1533	4	-.8	-.4	-1.3	-.6	5.21
1976	1916	1614	1883	5	.2	.5	.3	.5	1.40
1712	1712	1439	1546	6	-.6	-.1	-.2	-.6	2.22
1943	1685	1271	1671	7	.1	-.2	-.8	-.2	4.99
2104	1820	1717	1874	8	.6	.2	.7	.5	1.49
2983	2794	2412	2581	9	3.3	3.3	3.0	2.7	12.26
1745	1600	1384	1508	10	-.5	-.5	-.4	-.7	.77
1710	1591	1518	1667	11	-.6	-.5	.0	-.2	1.93
2046	1907	1627	1898	12	.4	.5	.4	.5	.46
1840	1841	1595	1741	13	-.2	.3	.3	.0	2.70
1867	1685	1493	1678	14	-.1	-.2	-.1	-.1	.13
1859	1649	1389	1714	15	-.1	-.3	-.4	-.0	1.08
1954	2149	1180	1281	16	.1	1.3	-1.1	-1.4	16.85
1325	1170	1002	1176	17	-1.8	-1.8	-1.7	-1.7	3.50
1419	1371	1252	1308	18	-1.5	-1.2	-.8	-1.3	3.99
1828	1634	1602	1755	19	-.2	-.4	.3	.1	1.36
1725	1594	1313	1646	20	-.6	-.5	-.6	-.2	1.46
2276	2189	1547	2111	21	1.1	1.4	.1	1.2	9.90
1899	1614	1422	1477	22	-.0	-.4	-.3	-.8	5.06
1633	1513	1290	1516	23	-.8	-.7	-.7	-.6	.80
2061	1867	1646	2037	24	.5	.4	.5	1.0	2.54
1856	1493	1356	1533	25	-.2	-.8	-.5	-.6	4.58
1727	1412	1238	1469	26	-.6	-1.1	-.9	-.8	3.40
2168	1896	1701	1834	27	.8	.5	.6	.3	2.38
1655	1675	1414	1597	28	-.8	-.2	-.3	-.4	3.00
2326	2301	2065	2234	29	1.3	1.7	1.8	1.6	6.28
1490	1382	1214	1284	30	-1.3	-1.2	-1.0	-1.4	2.58

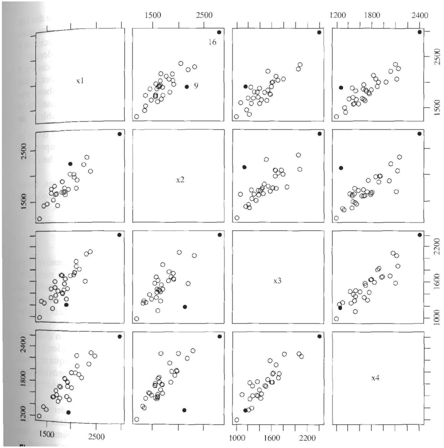

Figure 4.1l Scatter plots for the lumber stiffness data with specimens 9 and 16 plotted as solid dots.

Transformations to Near Normality

If normality is not a viable assumption, what is the next step ?

- Ignore the findings of a normality check and proceed as if the data were normality distributed. (Not recommend)
- Make non-normal data more "normal looking" by considering transformations of data. Normal-theory analyses can then be carried out with the suitably transformed data.

Appropriate transformations are suggested by

- theoretical consideration
- the data themselves.

Helpful transformations to near normality

Original Scale	Transformed Scale
1. Counts, y	$\sqrt{y}$
2. Proportions, $\hat{p}$	$\text{logit} = \frac{1}{2}\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$
3. Correlations, r	Fisher's $z(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$

Box and Cox transformation

$$x^{(\lambda)} = \begin{cases} \frac{x^{\lambda}-1}{\lambda,} & \lambda \neq 0 \\ \ln x, & \lambda = 0 \end{cases}$$

or

$$y_j^{(\lambda)} = \frac{x_j^{\lambda} - 1}{\lambda \left[\left(\prod_{i=1}^{n} x_i \right)^{1/n} \right]^{\lambda - 1}}, \quad j = 1, \ldots, n$$

Given the observations $x_1, x_2, \ldots, x_n$, the Box-Cox transformation for the choice of an appropriate power λ is the solution that maximizes the express

$$\ell(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^{n} \left(x_j^{(\lambda)} - \overline{x^{(\lambda)}} \right)^2 \right] + (\lambda - 1) \sum_{j=1}^{n} \ln x_j$$

where $\overline{x^{(\lambda)}} = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{x_j^{\lambda} - 1}{\lambda} \right)$

Example 3.16 (Determining a power transformation for univariate data)

We gave readings of microwave radiation emitted through the closed doors of $n = 42$ ovens in Example 3.10.

- The Q-Q plot of these data in Figure 4.6 indicates that the observations deviate from what would be expected if they were normally distributed.
- Since all the positive observations are positive, let us perform a power transformation of the data which, we hope, will produce results that are more nearly normal.
- We must find that value of λ maximize the function $\ell(\lambda)$.

λ	$\ell(\lambda)$	λ	$\ell(\lambda)$
-1.00	70.52		
-.90	75.65	.40	106.20
-.80	80.46	.50	105.50
-.70	84.94	.60	104.43
-.60	89.06	.70	103.03
-.50	92.79	.80	101.33
-.40	96.10	.90	99.34
-.30	98.97	1.00	97.10
-.20	101.39	1.10	94.64
-.10	103.35	1.20	91.96
.00	104.83	1.30	89.10
.10	105.84	1.40	86.07
.20	106.39	1.50	82.88
.30	106.51		

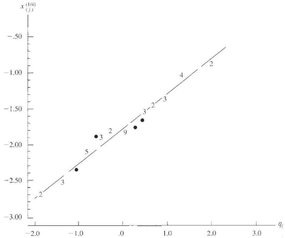

Figure 4.13 A Q–Q plot of the transformed radiation data (door closed). (The integers in the plot indicate the number of points occupying the same location.)

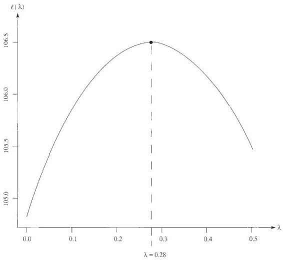

Figure 4.12 Plot of $\ell(\lambda)$ versus λ for radiation data (door closed).

Transforming Multivariate Observations

- With multivariate observations, a power transformation must be selected for each of the variables.
- Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the power transformations for the p measured characteristics. Each λ_k can be selected by maximizing

$$\ell(\lambda_k) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^{n} \left(x_{jk}^{(\lambda_k)} - \overline{x_k^{(\lambda_k)}} \right)^2 \right] + (\lambda_k - 1) \sum_{j=1}^{n} \ln x_{jk}$$

where $x_{1k}, x_{2k}, \ldots, x_{nk}$ are n observations on the k th variable, $k = 1, 2, \ldots, p$. Here

$$\overline{x_k^{(\lambda_k)}} = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{x_{jk}^{\lambda_k} - 1}{\lambda_k} \right)$$

- Let $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p$ be the values that individually maximize the equation above. Then the j th transformed multivariate observation is

$$\mathbf{x}_j^{(\hat{\lambda})} = \left[\frac{x_{j1}^{\hat{\lambda}_1} - 1}{\hat{\lambda}_1}, \frac{x_{j2}^{\hat{\lambda}_2} - 1}{\hat{\lambda}_2}, \cdots, \frac{x_{jp}^{\hat{\lambda}_p} - 1}{\hat{\lambda}_p} \right]^{\mathsf{T}}$$

- The procedure just described is equivalent to making each marginal distribution approximately normal. Although normal marginals are not sufficient to ensure that the joint distribution is normal, in practical applications this may be good enough.
- If not, the value $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p$ can be obtained from the preceding transformations and iterate toward the set of values $\boldsymbol{\lambda}^{\mathsf{T}} = [\lambda_1, \lambda_2, \ldots, \lambda_p]$, which collectively maximizes

$$\ell\left(\lambda_1, \lambda_2, \ldots, \lambda_p\right) = -\frac{n}{2} \ln \left|\mathbf{S}(\boldsymbol{\lambda})\right| + \left(\lambda_1 - 1\right) \sum_{j=1}^{n} \ln x_{j1} + \left(\lambda_2 - 1\right) \sum_{j=1}^{n} \ln x_{j2} + \cdots + \left(\lambda_p - 1\right) \sum_{j=1}^{n} \ln x_{jp}$$

where $\mathbf{S}(\boldsymbol{\lambda})$ is the sample covariance matrix computed from

$$\mathbf{x}_j^{(\lambda)} = \left[\frac{x_{j1}^{\lambda_1} - 1}{\lambda_1}, \frac{x_{j2}^{\lambda_2} - 1}{\lambda_2}, \ldots, \frac{x_{jp}^{\lambda_p} - 1}{\lambda_p}\right]^{\mathrm{T}}, \quad j = 1, 2, \ldots, n$$

Example 3.17 (Determining power transformations for bivariate data)

Radiation measurements were also recorded though the open doors of the $n = 42$ microwave ovens introduced in Example 3.10. The amount of radiation emitted through the open doors of these ovens is list Table 4.5.
Denote the door-close data $x_{11}, x_{21}, \ldots, x_{42,1}$ and the door-open data $x_{12}, x_{22}, \ldots, x_{42,2}$. Consider the joint distribution of x_1 and x_2, Choosing a power transformation for (x_1, x_2) to make the joint distribution of (x_1, x_2) approximately bivariate normal.

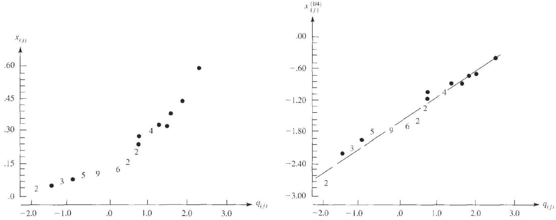

Figure 4.14 $Q-Q$ plots of (a) the original and (b) the transformed radiation data (with door open). (The integers in the plot indicate the number of points occupying the same location.)

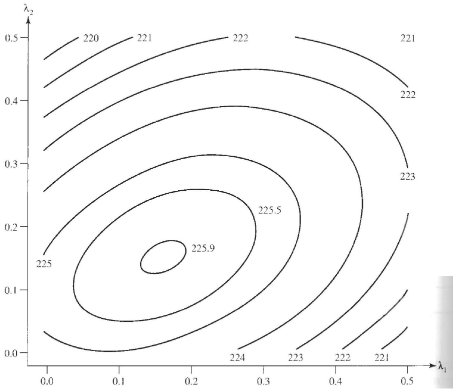

Figure 4.15 Contour plot of $\ell(\lambda_1, \lambda_2)$ for the radiation data.

If the data includes some large negative values and have a single long tail, a more general transformation should be applied.

$$x^{(\lambda)} = \begin{cases} \left\{(x+1)^{\lambda} - 1\right\}/\lambda & x \geq 0, \lambda \neq 0 \\ \ln(x+1) & x \geq 0, \lambda = 0 \\ -\left\{(-x+1)^{2-\lambda} - 1\right\}/(2-\lambda) & x < 0, \lambda \neq 2 \\ -\ln(-x+1) & x < 0, \lambda = 2 \end{cases}$$