专著

Adams, M.J., *Chemometrics in Analytical Spectroscopy*, 2nd ed., The Royal Society of Chemistry: Cambridge. 2004.
Beebe, K.R., Pell, R.J., and Seasholtz, M.B. *Chemometrics: A Practical Guide.*, John Wiley & Sons: New York. 1998.
Box, G.E.P., Hunter, W.G., and Hunter, J.S. *Statistics for Experimenters*. John Wiley & Sons: New York. 1978.
Brereton, R.G. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. John Wiley & Sons: Chichester, U.K. 2002.
Draper, N.R. and Smith, H.S. *Applied Regression Analysis*, 2nd ed., John Wiley & Sons: New York. 1981.
Jackson, J.E. *A User's Guide to Principal Components*. John Wiley & Sons: New York. 1991.
Jollife, I.T. *Principal Component Analysis*. Springer-Verlag: New York. 1986.
Kowalski, B.R., Ed. *NATO ASI Series. Series C, Mathematical and Physical Sciences, Vol. 138: Chemometrics, Mathematics, and Statistics in Chemistry*. Dordrecht; Lancaster: Published in cooperation with NATO Scientific Affairs Division [by] Reidel, 1984.
Kowalski, B.R., Ed. *Chemometrics: Theory and Application*. ACS Symposium Series 52. American Chemical Society: Washington, DC. 1977.
Malinowski, E.R. *Factor Analysis of Chemistry*. 2nd ed., John Wiley & Sons: New York. 1991.
Martens, H. and Næs, T. *Multivariate Calibration*. John Wiley & Sons: Chichester, U.K. 1989.
Massart, D.L., Vandeginste, B.G.M., Buyden, L.M.C., De Jong, S., Lewi, P.J., and Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics*, Part A and B. Elsevier: Amsterdam. 1997.
Miller, J.C. and Miller, J.N. *Statistics and Chemometrics for Analytical Chemistry*, 4th ed., Prentice Hall: Upper Saddle River N.J. 2000.
Otto, M. *Chemometrics: Statistics and Computer Application in Analytical Chemistry*. John Wiley & Sons-VCH: New York. 1999.
Press, W.H.; Teukolsky, S.A., Flannery, B.P., and Vetterling, W.T. *Numerical Recipes in C. The Art of Scientific Computing*, 2nd ed., Cambridge University Press: New York. 1992.
Sharaf, M.A., Illman, D.L., and Kowalski, B.R. *Chemical Analysis, Vol. 82: Chemometrics*. John Wiley & Sons: New York. 1986.

参 考 文 献

[1] Bro, R., Multivariate calibration. What is in chemometrics for the analytical chemist? *Analytica Chimica Acta*, 2003. 500(1-2): 185–194.

2.3 正态分布的性质

正态分布曲线的实际形状和它在均值两边的对称性是标准偏差的函数。统计学显示出，68%的观测值落在 $\mu \pm \sigma$ 范围内，95%的观测值落在 $\mu \pm 2\sigma$ 范围内，99.7%的观测值落在 $\mu \pm 3\sigma$ 范围内(图 2.3)。我们可以用两个六面的骰子很容易地描述正态分布是怎样形成的。如果把两个骰子一起掷，就会产生小范围的可能结果，即 2、3、4、5、6、7、8、9、10、11 和 12。但是，有些结果会有较高的出现频率，它取决于每个骰子的可能的组合数。例如，一个可能事件结果为两个骰子的总数为 2 或 12，那么两个骰子必须都是 1 或都是 6。若掷一次骰子得到的总和是 7，这就会有很多可能的组合(1 和 6、2 和 5、3 和 4、4 和 3、5 和 2、6 和 1)。如果你掷骰子的次数很少，那么每个可能结果不一定都能得到，但是随着掷的次数增加，总体的各种结果就会慢慢填满，最终变为正态分布。对此你可以自己试试看。

(注：与遵循正态分布的测量结果有关的统计学分支称为参数统计。因为大多数的测量结果都遵循正态分布，所以它是最常用的统计学。统计学的另一个分支不遵循正态分布，称为非参数统计。)

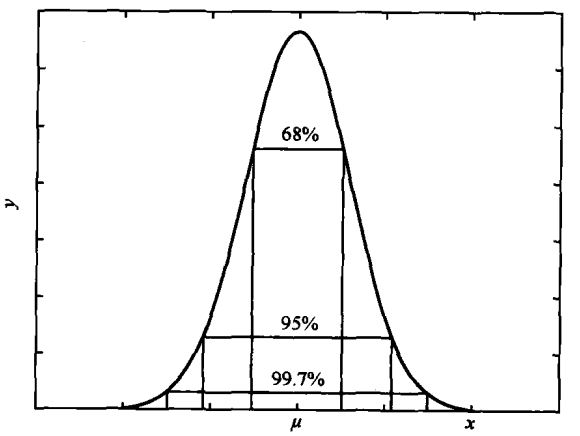


图 2.3 正态分布图。图中表明大约 68%的数值落在 $\mu \pm \sigma$ 范围内，95%的数值落在 $\mu \pm 2\sigma$ 范围内，99.7%的数值落在 $\mu \pm 3\sigma$ 范围内

置信区间是指我们可以合理地假设真值所在的一个范围，这个范围的极值称为置信限。术语“置信”暗示着我们在给定置信度，即一定概率下，可以得到一个结果。假设分布是正态分布，则样品均值有 95%的概率处在下面范围内：

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} < x < \mu + 1.96\frac{\sigma}{\sqrt{n}} \qquad (2.10)$$

但是，实际上我们通常获得了一个均值已知的对象，或者说一个取样的量测结果，然后，我们想获得 μ 的范围，所以，通过重新整理式(2.10)，可得

$$x-1.96\frac{\sigma}{\sqrt{n}}<\mu<x+1.96\frac{\sigma}{\sqrt{n}} \tag{2.11}$$

即

$$\mu=x\pm t\frac{\sigma}{\sqrt{n}} \tag{2.12}$$

t 的适当值(可以在统计表中找到)取决于自由度($n-1$)和所需的置信度(术语"自由度"是指用于计算 σ 的相互独立的偏差个数)。在无穷大的自由度和 95% 的置信限下，t 值为 1.96。

例如，考虑一组如下数据：

$$\bar{x}=100.5$$

$$s=3.27$$

$$n=6$$

使用 t=2.57(从表 2.1 查得)，置信度为 95%来计算置信区间：

$$\mu=100.5\pm2.57\frac{3.27}{\sqrt{6}}$$

$$\mu=100.5\pm3.4$$

表 2.1　t-分布中某置信区间下 t 的临界值

关于 P 值的 $\lvert t\rvert$ 的临界值 / P	90%	95%	98%	99%
自由度	0.10	0.05	0.02	0.01
1	6.31	12.71	31.82	63.66
2	2.92	4.30	6.96	9.92
3	2.35	3.18	4.54	5.84
4	2.13	2.78	3.75	4.60
5	2.02	2.57	3.36	4.03
6	1.94	2.45	3.14	3.71
7	1.89	2.36	3.00	3.50
8	1.86	2.31	2.90	3.36
9	1.83	2.26	2.82	3.25
10	1.81	2.23	2.76	3.17
12	1.78	2.18	2.68	3.05
14	1.76	2.14	2.62	2.98

这种检验可以通过两种方式使用：去检验两种样品的方差显著性差异或去检验两个数据组中哪个方差显著性高(或低)。因此，如表 2.2a 和表 2.2b 所示，一种是单尾检验，一种是双尾检验。

表 2.2a　单尾检验 F 的临界值(P=0.05)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	20	30	·
1	161.4	199.5	215.7	224.6	230.2	234	236.8	238.9	240.5	241.9	248	250.1	254.3
2	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4	19.45	19.46	19.5
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.66	8.62	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96	5.8	5.75	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.56	4.5	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	3.87	3.81	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.44	3.38	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	3.15	3.08	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	2.94	2.86	2.71
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.77	2.7	2.54
20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35	2.12	2.04	1.84
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	1.93	1.84	1.62
·	3.84	3	2.6	2.37	2.21	2.1	2.01	1.94	1.88	1.83	1.57	1.46	1

注：ν_1 为分子的自由度数；ν_2 为分母的自由度数。

表 2.2b　双尾检验 F 的临界值(P=0.05)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
1	647.7890	799.5000	864.1630	899.5833	921.8479	937.1111	948.2169	956.6562	963.2846	968.6274
2	38.5063	39.0000	39.1655	39.2484	39.2982	39.3315	39.3552	39.3730	39.3869	39.3980
3	17.4434	16.0441	15.4392	15.1010	14.8848	14.7347	14.6244	14.5399	14.4731	14.4189
4	12.2179	10.6491	9.9792	9.6045	9.3645	9.1973	9.0741	8.9796	8.9047	8.8439
5	10.0070	8.4336	7.7636	7.3879	7.1464	6.9777	6.8531	6.7572	6.6811	6.6192
6	8.8131	7.2599	6.5988	6.2272	5.9876	5.8198	5.6955	5.5996	5.5234	5.4613
7	8.0727	6.5415	5.8898	5.5226	5.2852	5.1186	4.9949	4.8993	4.8232	4.7611
8	7.5709	6.0595	5.4160	5.0526	4.8173	4.6517	4.5286	4.4333	4.3572	4.2951
9	7.2093	5.7147	5.0781	4.7181	4.4844	4.3197	4.1970	4.1020	4.0260	3.9639
10	6.9367	5.4564	4.8256	4.4683	4.2361	4.0721	3.9498	3.8549	3.7790	3.7168
20	5.8715	4.4613	3.8587	3.5147	3.2891	3.1283	3.0074	2.9128	2.8365	2.7737
30	5.5675	4.1821	3.5894	3.2499	3.0265	2.8667	2.7460	2.6513	2.5746	2.5112
·	5.0239	3.6889	3.1161	2.7858	2.5665	2.4082	2.2875	2.1918	2.1136	2.0483

注：ν_1 为分子的自由度数；ν_2 为分母的自由度数。

另一种可用于检验实验数据中异常值(或极限值)的方法是 Grubbs 检验法(Grubbs 1,Grubbs 2,Grubbs 3),如式(2.23)至式(2.25)所示。

$$G_1 = \frac{|\bar{x} - x_i|}{s} \tag{2.23}$$

$$G_2 = \frac{x_n - x_1}{s} \tag{2.24}$$

$$G_3 = 1 - \frac{(n-3) \times s_{n-2}^2}{(n-1) \times s^2} \tag{2.25}$$

式中,s 为标准偏差;x_i 为可疑极值;x_n 和 x_1 为最极值;s_{n-2} 为除去两个最极值后数据的标准偏差。

Grubbs 检验法的独特之处是在其应用之前,数据需按升序排列。像前面讨论的所有检验法一样,G_1、G_2 和 G_3 的检验值与从表 2.4 中得到的值相比较。如果检验值比表中的值大,那么我们拒绝原假设并把可疑值作为异常点拒绝,其中原假设是指假设它们来自于同一总体。同样,用于异常点拒绝的置信水平一般在 95% 和 99%。

表 2.4 Grubbs 检验法中 G 的临界值

n	95%置信水平			99%置信水平		
	G_1	G_2	G_3	G_1	G_2	G_3
3	1.153	2.00	—	1.155	2.00	—
4	1.463	2.43	0.9992	1.492	2.44	1.0000
5	1.672	2.75	0.9817	1.749	2.80	0.9965
6	1.822	3.01	0.9436	1.944	3.10	0.9814
7	1.938	3.22	0.8980	2.097	3.34	0.9560
8	2.032	3.40	0.8522	2.221	3.54	0.9250
9	2.110	3.55	0.8091	2.323	3.72	0.8918
10	2.176	3.68	0.7695	2.410	3.88	0.8586
12	2.285	3.91	0.7004	2.550	4.13	0.7957
13	2.331	4.00	0.6705	2.607	4.24	0.7667
15	2.409	4.17	0.6182	2.705	4.43	0.7141
20	2.557	4.49	0.5196	2.884	4.79	0.6091
25	2.663	4.73	0.4505	3.009	5.03	0.5320
30	2.745	4.89	0.3992	3.103	5.19	0.4732
35	2.811	5.026	0.3595	3.178	5.326	0.4270
40	2.866	5.150	0.3276	3.240	5.450	0.3896
50	2.956	5.350	0.2797	3.336	5.650	0.3328

布被用来计算置信区间，以取代图 3.1 中所示的基于 z 值的正态概率分布。当 $n \geqslant 30$ 时，t-分布接近于标准正态概率分布。对样本容量小的样本，均值的置信区间变大，可用式(3.9)来估计：

$$\mu = \bar{x} \pm t_{\alpha/2} s_{\bar{x}} \tag{3.9}$$

式中，t 为在期望的置信水平下，自由度为 $n-1$ 时的临界值。自由度是指在计算标准偏差 s 时所用的独立偏差 $(x_i - \bar{x})$ 的个数。例如，如果想估计自由度为 $n-1=5$，置信水平为 $100\%(1-\alpha) = 95\%$ 时均值的置信区间，从标准 t-分布表中查得 $\alpha/2 = 0.025$ 时，临界值 $t_{\alpha/2} = 2.571$ 就可用于式(3.9)中。这里的 $\alpha/2 = 0.025$ 代表在 t-分布中右侧与左侧的百分率，在相关 t-分布中所选的值图示于图 3.2(b)。式(3.9)并不表示样本均值非正态分布；相反，它表明，除非当 n 很大时，否则 s 只是 σ 的一个不好的估计。

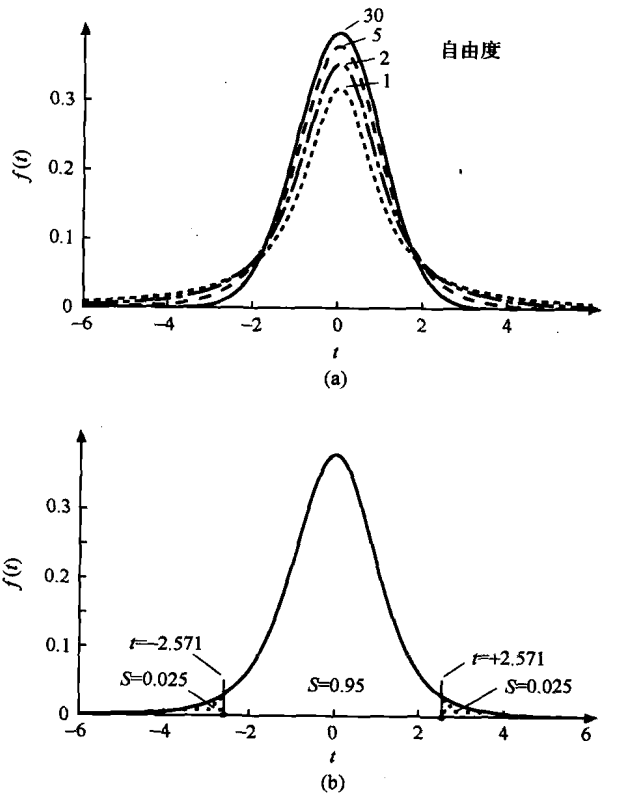


图 3.2 t-分布的图示

(a) 不同自由度；(b) 自由度为 5 时，曲线下面积为 0.95

检验量，并与统计表中的值进行了比较。测试结果有四种可能的情况：①如果 H_0 为真而接受 H_0，那么我们就已作出了正确的决定；②如果 H_a 为真却接受 H_0，我们就犯了第Ⅱ类错误，犯这类错误的概率记为 β；③如果 H_0 为假而拒绝 H_0，那么我们已作出了正确决定；④如果 H_0 为真而却拒绝 H_0，那么我们就犯了第Ⅰ类错误，犯这类错误的概率记为 α。

大部分统计假设检验的应用要求我们将允许犯第Ⅰ类错误的最大概率具体化，这就称为检验的显著性水平。我们常用的显著性水平为 0.01 或 0.05，这就意味着我们有较高的置信度来作拒绝 H_0 的判断。例如，当我们以 95%的置信水平来拒绝 H_0 时，5%的情况下我们会作出错误的决定。换句话说，这个差别是由于随机偶然引起的概率为 5%。所以，我们非常有把握认为我们作出了正确的决定。

为了对式(3.13)中所述的均值比较作测试，我们计算统计检验量(t_{calc})：

$$t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{s_p}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{3.14}$$

然后将其与自由度为 $n_1 + n_2 - 2$、显著性水平为 α 时从表中查得的 t 作比较，其中 s_p 为合并的标准偏差：

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{3.15}$$

如果 $t_{calc} > t_\alpha$，那么有 $100\%(1-\alpha)$ 的置信水平拒绝 H_0。对表 3.1 中所示的数据，我们有 $t_{calc} = 2.451$，$t_{\alpha=0.05,\nu=8} = 1.860$，所以我们有 95%的把握拒绝 H_0 而接受 H_1，有小于 5%的把握认为这种差别是由随机误差引起的。用来描述假设检验的显著性水平和作出决定的置信水平的语言表明了两者之间的关系，式(3.16)给出了用于计算两个均值之差的置信区间的公式：

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \tag{3.16}$$

式中，$t_{\alpha/2}$ 可从 t-分布表中查得，此时自由度为 $n_1 + n_2 - 2$；s_p 为合并的标准偏差。式(3.16)的简单变形可得到类似于式(3.14)的形式。用于比较两个均值的 t-检验等价于估计统计检验量的置信区间，然后检查该置信区间是否包含用于统计检验的假设值。

3.4.2　方差的统计推断和 F-分布

有时需要比较两个总体的方差。例如，表 3.1 所示的数据代表了两个不同的总体。在计算合并标准偏差前，检验方差是否相等可能是比较合适的。例如，我们可能会问："方差 s_1^2 与 s_2^2 的差别是否在统计上显著，或者这个差别能否仅用随机误差来解释？"F-分布是用来进行这种检验的，它描述的是样本方差比率的分布，

下，需要使用 Hotelling's T^2 统计，对在 $100\%(1-\alpha)$ 的置信水平下，由样本均值和分散度矩阵得到的置信区间做一个判定。

$$T^2 = (\bar{\boldsymbol{x}}-\boldsymbol{\mu})\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^{\mathrm{T}} > \frac{(n-1)}{(n-p)}F_{p,n-p,\alpha} \tag{3.35}$$

计算统计参量 T^2，并且与显著性水平为 α 时的 $[(n-1)/(n-p)]F$ 值进行比较，当 $T^2 > [(n-1)/(n-p)]F$ 时，我们就拒绝原假设 H_0。

检验统计参量：

$$F = \frac{s_1^2}{s_2^2}，\text{如果 } F > F_\alpha \text{ 则拒绝 } H_0^{①} \tag{3.36}$$

3.7　例子：多元距离

名为“smx. mat”的文件包含了一组数据。这个量测包含了 83 个不同含量磺胺甲噁唑的近红外反射光谱（图 3.8），磺胺甲噁唑是药物制备中所用的一种活性物质。这个数据集已被分为三部分：42 个光谱组成的训练集、13 个光谱组成的测试集和 28 个“不合格”光谱。不合格样本中有意掺入了磺胺甲噁唑的两个降解产物——对氨基苯磺酸和对氨基苯磺酰胺，掺入的质量分数为 0.5%～5%。在这些练习中，你将考察这些数据集，选择几个波长，并且用所选波长处的红外反射光谱测量值来计算样本的马氏（Mahalanobis）距离，以推断近红外量测是否可用于检测具有上述不纯样本的纯度。在每一节的后面都附了用于样本分析的 MATLAB 程序。

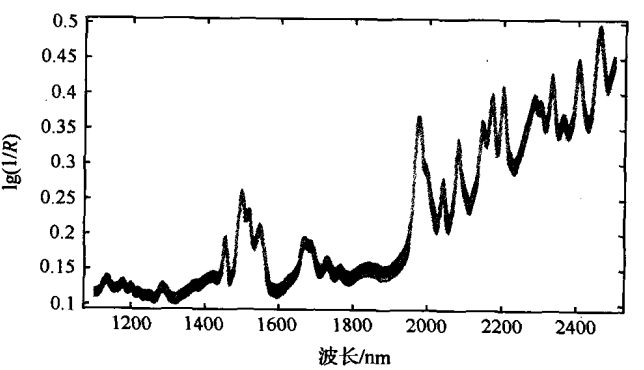


图 3.8　83 份磺胺甲噁唑粉末样本的近红外反射光谱

① 译者注：原书式(3.36) 中 F 的表达式排版有误，已依照原始文献更改。

向量将 $\boldsymbol{A}$ 的秩(或维度)从 50 降至 2,并且不会有信息的显著性丢失。也就是说,我们忽略了拟合误差的基向量。PCA 模型的数据压缩功能频繁地被使用,这是其最重要的特征之一。

4.3 主成分模型

绝大多数情况下,一开始我们并不知道如图 4.1 中所描述的数据中的纯色谱图或光谱图。否则,便无需使用 PCA。幸运的是,我们能在没有先验信息的情况下用 PCA 计算出张成矩阵 $\boldsymbol{A}$ 空间的基向量的唯一解集。

通过 PCA,我们能建立一个用式(4.4)表达的数学经验模型,$\boldsymbol{T}_k$ 是 $n\times k$ 阶主成分得分矩阵,$\boldsymbol{V}_k$ 是 $m\times k$ 阶特征向量矩阵。

$$\boldsymbol{A}=\boldsymbol{T}_k\boldsymbol{V}_k{}^{\mathrm{T}}+\boldsymbol{E} \tag{4.4}$$

$\boldsymbol{V}_k$ 中的特征向量构成了矩阵 $\boldsymbol{A}$ 的行正交基向量,特征向量称为载荷,有时也称为抽象因子或特征光谱,它们构成了 $\boldsymbol{A}$ 中行空间的基向量,并非总都能得到向量的物理意义解释(图 4.2)。$\boldsymbol{T}_k$ 中的每一列都称为得分,彼此正交但没有标准化,其可以构成 $\boldsymbol{A}$ 的一组列基向量。

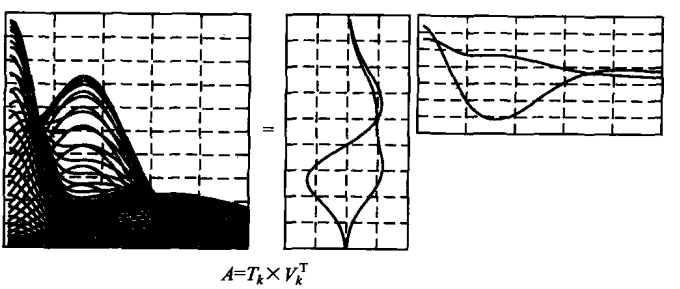


图 4.2 图 4.1 中色谱-光谱数据的主成分模型

数据中每一个独立的方差源对应于模型中一个独立的主成分。得分矩阵的第一列与第一个特征向量($\boldsymbol{V}_k$ 中的第一列)对应于第一因子。第一特征向量对应于最大的特征值。可以证明,第一因子拟合了原始数据中的最大方差(在最小二乘下最大)。第二因子由得分矩阵的第二列和特征向量矩阵的第二列构成,对应第二大特征值。它拟合了原始矩阵剩余的最大方差。简言之,数据阵可由 k 个秩为 1 的 $n\times m$ 阶矩阵加和表示:

$$\boldsymbol{A}=\boldsymbol{t}_1\boldsymbol{v}_1^{\mathrm{T}}+\boldsymbol{t}_2\boldsymbol{v}_2^{\mathrm{T}}+\cdots+\boldsymbol{t}_k\boldsymbol{v}_k^{\mathrm{T}}+\boldsymbol{E} \tag{4.5}$$

向量的外积 $\boldsymbol{t}_1\boldsymbol{v}_1^{\mathrm{T}}$ 为第一因子拟合的方差。对于图 4.1 中的 HPLC-UV/Vis 数据,式(4.4)和式(4.5)只需两个抽象因子对其进行拟合,每个因子对应一个化学成分。此处强调的抽象因子是指两个因子,不一定就是两个化学成分。

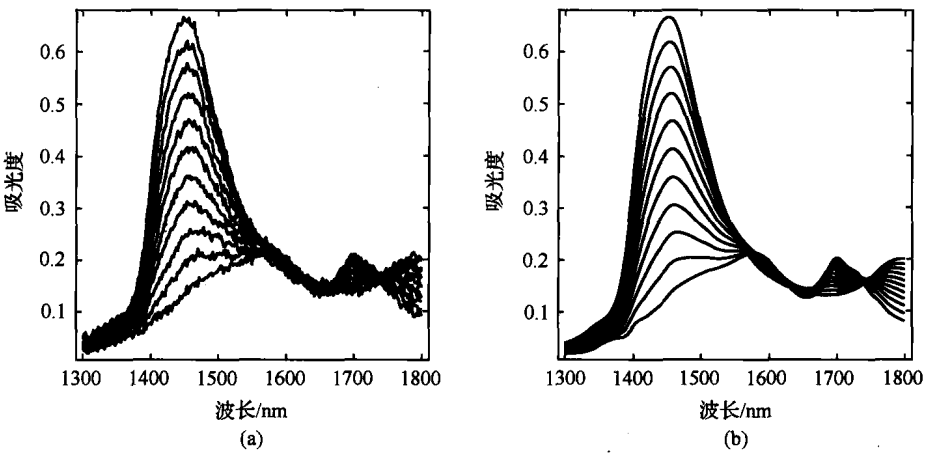


图 4.5　多项式平滑水-甲醇混合物的近红外光谱图示

(a) 原始数据；(b) 平滑后数据

一固定常数。平滑点是多项式拟合获得的中间点。在多项式对平滑点做出估算后，窗口摒弃旧数据，加入新数据向右移动。由另一个多项式拟合新的窗口并估算中间点。不断重复上述步骤，一次一个点，直至整个曲线得以平滑。通过改变窗口的宽度 w 和拟合多项式的级数，实现对平滑度的控制。增加窗口的宽度可以得到更强的平滑效果。增加多项式级数，如从二次幂到四次幂，将允许对数据进行更加复杂的曲线拟合。

多项式平滑并不是最理想的频率响应函数，并且很可能在平滑信号中引入扭曲和伪像[7]。其他平滑法并没有这种缺点。详细讨论见本书第 10 章。

.4.5　一阶导数与二阶导数

对连续函数求导可以用来扣除基线漂移，因为常数的导数为 0[8]。实际情况下，数字化曲线的一阶导数可通过数值方法来逼近，以有效移除基线漂移。导数变换是线性的，并且求导后产生的曲线保留了原始信号的定量特点。最常用的方法基于多项式平滑。在多项式平滑中应用了滑动窗口，但是，平滑操作的系数生成了拟合数据的多项式的导数。在多项式平滑中，这些滤波类型的频率响应函数都不理想，而且如果误用这种方法，将会引入扭曲和伪像。

图 4.6 展示了水-甲醇混合物近红外光谱一阶导数的数值逼近。除移除基线漂移外，求导也具有高通滤波的功能，即将光谱峰窄化和尖化。零点可用于鉴定峰在原始光谱中的位置。这一过程也移除了相当数量的信号，导致一阶导数曲线中的信噪比降低(注意图 4.6 中两幅图的纵坐标刻度不同)。

4.6.3 *F*-检验法确定因子数

可以用简化特征值设计一个简单的假设检验来检测因子 j 的显著性[17]。

$$F_{(1,n-j)} = \frac{\mathrm{REV}_j}{\sum_{i=j+1}^{n} \mathrm{REV}_i}(n-j) \tag{4.40}$$

F-检验通过从倒数第二个特征值开始，以决定数据阵的实际因子数。通过将对应于倒数第二个特征值的方差与剩余特征值的方差进行比较来检测该特征值的显著性。若计算出的 F 值小于在想要的显著性水平($\alpha=0.05$ 或 $\alpha=0.01$)下从表中查得的 F 值，我们就认为该特征值不显著。通过比较下一个最小特征值与非显著特征值组的方差之和来检验该特征值是否显著，重复将特征值添加到非显著因子组的过程，直到第 j 个特征值的方差比值超过了表中的 F 值，从而标示出真实矢量集与误差矢量集的区别。

例 4.6 中，表 4.1 计算出图 4.1 中模拟色谱数据集的统计量。该模拟数据加入了随机误差($\sigma=0.0005$，$\bar{x}=0$)。从列的底部到顶部的倒退工作方式在表 4.1 中以 F 比值(F-ratio)标出，我们看到第一个显著性主成分在 $j=2$ 处出现。REV_2 与剩余特征值之和的差异是由随机误差引起的这一事件的概率由表 4.1 中"Prob. Level"一列给出。$j=2$ 处的真实概率水平非常小(ca. 1×10^{-7})，故而它被取整为 0。选择两个主成分，我们发现估计的残余误差为 0.0005 个吸光度单位(AU)，这与加入到数据矩阵中的真实随机误差非常一致。例 4.7 中所示的 MATLAB 代码用于计算 REV 值和 F 比值。

表 4.1 模拟色谱数据的因子分析结果(Trace=47.890 421)

因子	特征值	方差/%	累积方差/%	RE	IND /(1×10^{-7})	REV	F 比值	Prob. Level
1	43.015 138	89.819 9	89.819 9	0.047 55	245.62	1.912×10^{-2}	371	0.000
2	4.874 733	10.178 9	99.998 9	0.000 52	2.79	2.261×10^{-3}	141 987	0.000
3	0.000 047	0.000 1	99.998 9	0.000 50	2.86	2.260×10^{-8}	1.43	0.238
4	0.000 039	0.000 1	99.999 0	0.000 50	2.95	1.955×10^{-8}	1.25	0.271
5	0.000 037	0.000 1	99.999 1	0.000 49	3.05	1.936×10^{-8}	1.24	0.272
6	0.000 033	0.000 1	99.999 2	0.000 48	3.15	1.831×10^{-8}	1.18	0.284
7	0.000 032	0.000 1	99.999 2	0.000 47	3.27	1.844×10^{-8}	1.19	0.281
8	0.000 029	0.000 1	99.999 3	0.000 46	3.39	1.779×10^{-8}	1.16	0.289
9	0.000 028	0.000 1	99.999 4	0.000 46	3.51	1.810×10^{-8}	1.18	0.284
10	0.000 026	0.000 1	99.999 4	0.000 45	3.65	1.735×10^{-8}	1.14	0.293
11	0.000 023	0.000 0	99.999 5	0.000 44	3.81	1.676×10^{-8}	1.10	0.301
12	0.000 021	0.000 0	99.999 5	0.000 43	3.98	1.608×10^{-8}	1.06	0.311

度。由式(4.46)给出，s_o^2 是类 q 的残余方差，n 是类 q 中的样本数。

$$s_o^2 = \frac{1}{(m-k)(n-k)} \sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij}^2 \tag{4.46}$$

式(4.46)中的求和为类 q 中全部样本在所有波长下残余光谱的方差之和。注意，该处 s_o^2 的定义与 Malinowski 的 RE 函数的定义一样。进行均值校准后，式(4.46)中分母所表示的自由度应该改为$(n-k-1)(m-k-1)$。

若假设原始数据服从正态分布并且主成分模型足以拟合该原始数据，则可认为残差也服从正态分布。这样，可以计算式(4.47)中的方差比值来检验原假设 $H_0: s_i^2 = s_o^2$，其与 $H_1: s_i^2 \neq s_o^2$ 互为对立假设。当计算的比值大于显著性水平 α 对应的 F 临界值时，拒绝原假设。在使用 NIR 光谱的这个实验中，我们发现用于式(4.47)所示的 F-检验中的自由度给出了满意的结果。在其他应用中，不同作者对 F-检验给出了不同的自由度。

$$\frac{s_i^2}{s_o^2} \geqslant F_{1,n-k}(\alpha) \tag{4.47}$$

项 1 和 $n-k$ 给出了用于比较计算出的单一未知光谱的 F 值与表中 F 值的自由度。没有均值校准时，自由度用 $n-k$ 表示，若已均值校准，式(4.47)就用项 $n-k-1$ 代替 $n-k$，然后将数据矢量根据 F-检验的概率水平进行分类。

例 4.9 中，用例 4.8 来计算残余方差和图 4.16 中数据集的 F 比值。10 个“未知”光谱的 F 比值见表 4.2。被微量杂质污染的未知光谱展示在第一行中。训练集中所有样本都仅包含小残余方差，并且其 F 比值比临界值 $F=4.105$ 低。“不能接受”的未知光谱有一个非常大的 F 比值，表明在高置信度下，它不属于训练集代表的样本集。

表 4.2 残余方差分析的例子

样本	F 比值
1	34.945 6
2	0.897 9
3	0.667 0
4	0.922 2
5	0.913 3
6	0.808 9
7	0.839 7
8	0.766 0
9	0.960 2
10	3.434 2

注：F 比值中自由度分别为 df=1 和 df=37。F 分布表中的 $F(\alpha=0.05,1,37)=4.105$。在概率水平 1.00−0.05=0.95 下，只要 F 比值>4.105，就拒绝原假设。

大多数日常分析任务中被认为是无用的，因为许多化合物在该波长区域有相互重叠的宽带吸收。现在近红外光谱分析法正在快速取代许多费时的常规分析方法，如 Karl-Fisher 湿度滴定法、用于测定总蛋白质含量的 Kjeldahl 定氮法、美国材料测试协会(ASTM)用于测定发动机汽油中辛烷值的量测方法等。正是由于有像多元校正这样的化学计量学方法，他们才能解析从信息含量丰富的 NIR 光谱区域观测到的相互重叠的宽吸收带的复杂模式，使上述应用成为可能。

5.1.2 振动的基本模式、倍频及和频

近红外光谱区(700～3000 nm)的吸收带是中红外范围(4000～600 cm^{-1})内基本振动模式的倍频或和频的结果。已有关联表格列出特定官能团在近红外光谱区可能给出的吸收带位置，可供查阅。现来看一个例子，OH 键的基本伸缩振动频率大概为 3600 cm^{-1}。这一基本振动模式的第一级、第二级和第三级倍频可以在近红外光谱区域观测到，波数分别为 7200 cm^{-1}、9800 cm^{-1}和 13800 cm^{-1}。吉他之类的弦乐器也提供了一个有用的模型，基调琴弦振动的第一级倍频会产生一个高八度的音，例如，其频率是基调的两倍。一个分子的基本振动模式对应着从基态能级水平 ($\upsilon=0, E_\upsilon=1/2h\nu$) 到第一激发态 [$\upsilon=1, E_\upsilon=(1+1/2)h\nu$] 的跃迁；第一级、第二级和第三级倍频对应于分子从基态分别跃迁到 $\nu=2$、$\nu=3$ 和 $\nu=4$ 能级的禁阻跃迁。如果分子键的振动是标准的简谐振动，各能态的间隔应该是均等的，实际上在分子振动偏离简谐振动模型时，能级间的间隔并不均等。正因为各能级间的能量有差异，我们才能观察到禁阻跃迁，尽管它的强度比基本跃迁弱 10～100 倍。每一级倍频吸收带都相继减弱，第三级倍频只有在基本振动吸收带非常强时才能被观测到，第四级倍频一般太弱以至于检测不到。

和频吸收带是指同时发生两种跃迁的组合，如一个分子同时含有位置靠近的羰基 ($\nu_1=1750$ cm^{-1})和羟基 ($\nu_2=3600$ cm^{-1})官能团，就会产生一个波数为 $\nu_1+\nu_2=5350$ cm^{-1}的组合吸收谱带。

5.1.3 水-甲醇混合物

在这里制备了 9 个甲醇体积分数分别为 10%，20%，…，90%的甲醇-水混合样本。9 个样本再加上纯水和纯甲醇样本，使用 NIRSystems 公司的 Model 6500 NIR 光谱仪，依次在 0.5 mm 的流动样品池中测量。扫描范围为 1100 nm 到 2500 nm，间隔为 2 nm，每个光谱共记录 700 个数据点。在 1 h 的测量过程中并未刻意保持样品池恒温。11 个光谱绘制于图 5.1。

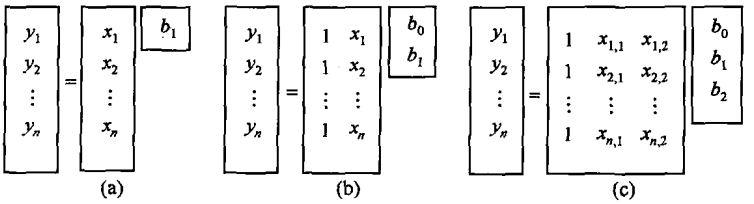


图 5.2 含有 n 个标准样的三种不同类的线性模型图示

左图表达最简单仅有斜率项而不含截距项的模型。中间的模型添加了非零的截距项。右边的模型在文献中一般记为多元线性回归(MLR)模型，因为它使用多个响应变量。另外，$n \geqslant (m+1)$ 时有截距项，$n \geqslant m$ 时没有截距项。右边的这个模型有一个非零的截距项

通过最小二乘处理，用 $\boldsymbol{y}$ 和 $\boldsymbol{x}$ 的值来估计模型参数 b_1，这个最小二乘估计值 $\hat{b}_1$，计算如下：

$$\hat{b}_1 = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} \tag{5.2}$$

式中，$\hat{b}_1$ 读作"b 帽"，用于强调它作为 b_1 的估计值这一角色。最终产生的校正模型被用来预测未知样本中待分析物的浓度 $\hat{y}_{\mathrm{unk}}$：

$$\hat{y}_{\mathrm{unk}} = x_{\mathrm{unk}}\hat{b}_1 \tag{5.3}$$

式中，x_{unk} 为未知样本在校正量测波长点的响应值。由于仅使用一个响应变量，所以这类校正称为一元校正。

5.2.2 非零截距

式(5.2)和式(5.3)假设为当分析物浓度为零时，仪器响应值为零。也就是说，上述校正模型使得校正曲线通过原点，即当仪器响应为零时，估计的浓度必须同样为零。针对此，常常通过从校正样读数中减去空白样本响应来将仪器响应设置为零。空白仪器响应就被归为零，对所有的校正量测亦都如此。空白的重复量测将会给出在零值两侧的小范围、正态分布的、随机的波动。然而，对很多样本来说，如果不能获得基质匹配并且不含待分析物的空白样，那么就很难这样做。在使用式(5.1)至式(5.3)之前可以通过对 $\boldsymbol{y}$ 和 $\boldsymbol{x}$ 均值中心化来消除截距项，但是由此在式(5.3)中得到的浓度估计值需要去均值中心化。虽然 $\boldsymbol{y}$ 和 $\boldsymbol{x}$ 均值中心化后校正曲线截距为零，但本质上，一般还会存在一个非零的截距项。因此，对 $\boldsymbol{y}$ 和 $\boldsymbol{x}$ 均值中心化以产生为零的截距项，与使用原始数据并限制模型以使截距为零是不同的。

当没使用均值中心化时，在校正模型中可以包含一个非零截距项，模型表达为

$$y_i = b_0 + x_i b_1 + e_i \tag{5.4}$$

矩阵表示中，可以通过增加一列全为 1 的列来扩展仪器响应矢量 $\boldsymbol{x}$ 从而产生响应矩阵，如图 5.2(b)所示 ($\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$)。模型参数 b_0 和 b_1 的最小二乘估计计算如下：

Carlo)交互验证(MCCV)[7-9],也称为一次移出多个的交互验证(leave-multiple-out CV, LMOCV)[8],就可以克服上述不足。使用 MCCV,校正集被分组,使得验证样本的数目比校正样本的数目大。通过多次的随机分割可以得到 MCCV 的平均值。该方法的一个变形是使用重复的 v-fold 交互验证,其中 v-fold 交互验证进行 B 个循环,每次都不同地随机分割成 v 个分开的组[10]。总的来说,当使用小数目的校正样本时,许多研究人员偏向于采用 LOOCV 方法,RMSECV 一般也倾向于给出对校正模型预测能力过分乐观的估计。

5.3 校正应用实例

5.3.1 水-甲醇混合体系近红外数据的图形化研究

在建立任何模型之前,都应该先绘制出光谱,因为这是用于构建模型的数据,而光谱的图形化研究可评估光谱的质量,评估内容包括全波长段的信噪比、共线性、背景漂移和任何观察到的非正常现象(例如,某一个光谱明显与其他不同,这表明这个光谱是一个异常者)。图 5.1 是 11 个水-甲醇混合样本的近红外光谱图。图 5.3 和图 5.4(a)(b)分别是纯水、纯甲醇光谱和上述 11 个样本的一阶导数和二阶导数光谱。图 5.1 中并没有出现任何观测的异常现象。但是,随波长增加,基线有明显上升趋势。一阶导数光谱在短波段对基线漂移有校正效果,而二阶导数光谱在整个波长段的校正效果都很明显。因此,在校正模型中使用二阶导数光谱或许可以获得最好的结果。然而,随着连续取导,信噪比会逐渐降低。

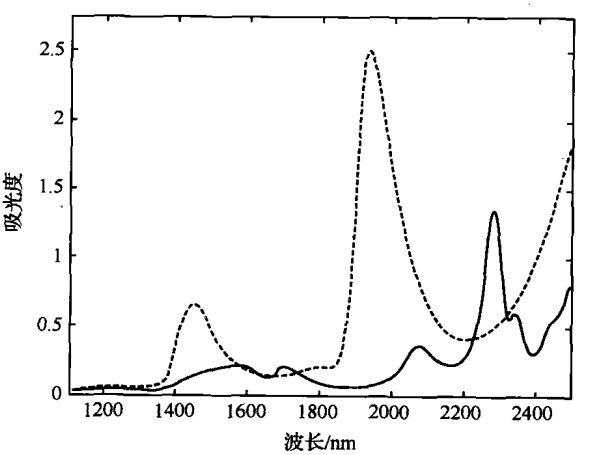
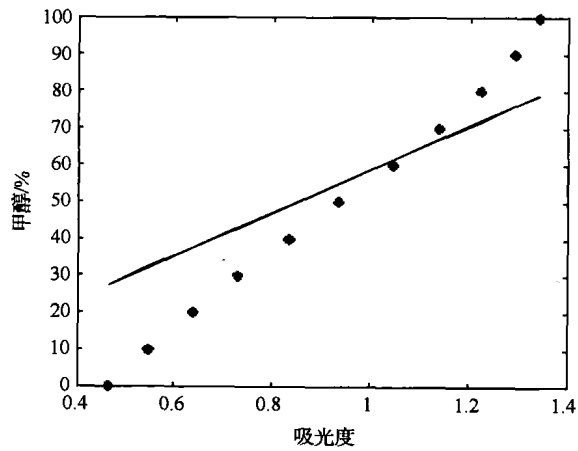


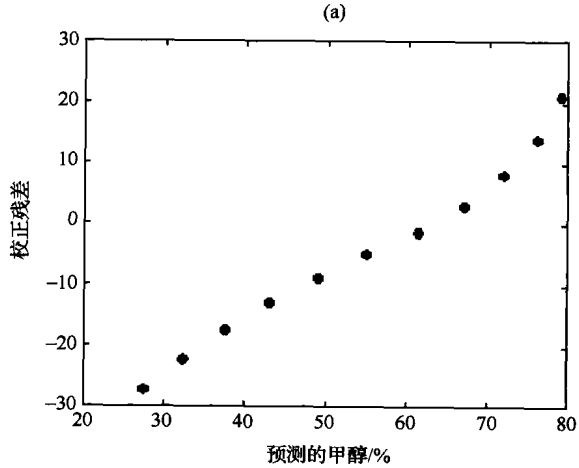
图 5.3 水(----)和甲醇(—)的纯组分 NIR 光谱

表 5.2 水-甲醇混合物进行一元和多元校正得到的甲醇的结果

波长(截距模式)/nm	RMSEC(甲醇)/%*	RMSEV(甲醇)/%*
2072(无截距项)	49.35	37.85
2072(有截距项)	2.82	1.73
2274(无截距项)	18.16	13.46
2274(有截距项)	3.03	1.86
1452、2274(无截距项)	2.07	1.29
1452、2274(有截距项)	0.48	0.36
1452、1932、2072、2274(无截距项)	0.45	0.37
1452、1932、2072、2274(有截距项)	0.24	0.18

* 这些值对应于 6 个校正样本和 5 个验证样本。





式(5.15)右侧的两个平方和项与各自的自由度一起，单独列于表 5.3。“均方”通过对平方和项除以其对应的自由度获得。在量测值 y 中，误差的估计值 s_y(y 的标准误差)按照下式估计：

$$s_y = \left[\frac{1}{n-m-1}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right]^{1/2}$$

F-检验中的原假设一般假定真实系数全为零(注意不包含 b_0)，F-分布的 F 比值中，分子和分母的自由度分别为 $df_{regr} = m$ 和 $df_{resid} = n-m-1$。

$$H_0 : b_1 = b_2 = \cdots = b_m = 0$$

$$H_1 : \text{有一个或多个 } b_i \text{ 不为零}$$

$$\text{如果} \quad \frac{SS_{regr}/df_{regr}}{SS_{resid}/df_{resid}} \geqslant F_{m,n-m-1}(\alpha)\text{，那么拒绝 } H_0$$

式中，SS_{regr}为回归模型所解释的方差和；SS_{resid}为残差平方和[见式(5.15)]。对于足够大的 F 值，我们在显著性水平为 α 时拒绝原假设，这意味着回归方差相对于其偶然出现的情况太大了。

5.4.4 回归系数的置信区间和假设检验

计算完校正系数后，很值得考察 $\hat{b}$ 中存在的误差并建立置信区间。每个回归系数的标准偏差按照如下公式计算：

$$s_{b_i} = \sqrt{\mathrm{Var}(\hat{b}_i)}$$

式中，$\mathrm{Var}(\hat{b}_i)$ 为最小二乘回归系数 $\hat{b}_i$ 的方差估计，也即 $s_y^2(\mathbf{X}^T\mathbf{X})^{-1}$ 的第 i 个对角元。回归系数中标准误差的分析可以通过对回归系数计算 t 比值和置信区间而简单地实现，其中，

$$t_{b_i} = \frac{b_i}{s_{b_i}}$$

$$b_i = \hat{b}_i \pm \sqrt{\mathrm{Var}(\hat{b}_i)}\ \sqrt{(m+1)F_{m+1,n-m-1}(\alpha)} \tag{5.16}$$

然而，许多计算机程序和使用者往往忽视了式(5.16)中计算置信区间的“同步性”，而替代地按照下面的公式用“一次一个”(one-at-a-tine)的 t 值来计算 F。

$$b_i = \hat{b}_i \pm t_{n-m-1}\left(\frac{\alpha}{2}\right)\sqrt{\mathrm{Var}(\hat{b}_i)}$$

一般当回归系数的标准偏差远远小于其数值时，该系数被认为是重要的。t 值可以用来帮助判断单个回归系数的显著性。一般地，如果 t 值在 2.0 至 2.5 之间或更小，则认为该回归系数对于预测 y 值不是特别有用。特别是当回归系数的 t 值小于临界值 t_{n-m-1} 时，我们接受原假设 $H_0 : b_i = 0$。这种条件表明系数的置信区间包含零值。t 值也可以看作是信噪比。