**第七章 模型校验**

用最小二乘逼近观测的拟合数据后，实验者便得到所使用的模型函数的最佳估计参数。本章讨论（1）最优解是否充分；（2）假设是否正确，即所选择的模型是否真实反应了实验中的测量现象。

# 拟合优度、精度和准确度

## 统计模型和数据的一致性

首先概括一下数据拟合的目的。最小二乘逼近的思想是最小化如下函数



该实验中有个观测，每个观测值对应在某个条件下的测量。参数向量是待优化变量，而权值反映对应观测的可靠性。

通常情况下，无论模型是否正确，数据都不会完全符合所选模型函数。拟合优度表示统计模型描述数据的好坏程度。这个统计模型既包含选择的模型函数，也包含观测与该函数的假定偏差。

拟合优度由下式度量



其中，为用个模型参数拟合个数据点的自由度。表示统计模型与数据一致。这意味着权重被正确地选择为总体方差的倒数（），并且模型函数是适合的。然而，不排除存在其他可能更合适的模型。如果模型函数不适合描述观测，偏差将大于，因此。另一方面，小于1的拟合优度不一定表示更好的拟合; 相反，有可能因为被高估。

但是，许多应用中是未知的，这意味着权重必须用不同方式确定。理想地，可以估计正确的相对权重



则拟合优度仅仅是常数的估计值。

## 样本方差

拟合方差（样本方差）等于数据标准差平方的估计值，由下式表示



该公式为归一化的拟合优度，它不包括关于拟合质量的附加信息。

式中归一化的自由度和标准差的其它公式的使用理解如下：当与参数个数相比非常大时，差异可以忽略不计。考虑直线实例，并假设只有个观测。显然，经过这两点总是可以画出一条直线。但是，这是正确直线的概率接近零，且不确定性接近无穷大。因此分母变为，并且估计的偏差为无穷大。一般来说，这与过拟合的问题有关。使用具有足够参数的模型函数，可以拟合任何数据（甚至是噪音）。

标准差仅仅是精度的度量，它由观察的随机扰动引起。收集数据的实验者还想知道观测在真实范围内的准确度[1]。

图7. 1显示了实验的准确度和精度之间的差异。准确度表示观测与真值的接近程度; 它是结果正确性的度量（低系统误差）。精度是观测与模型函数拟合程度的度量，而不考虑真实值; 它受到随机误差的影响。

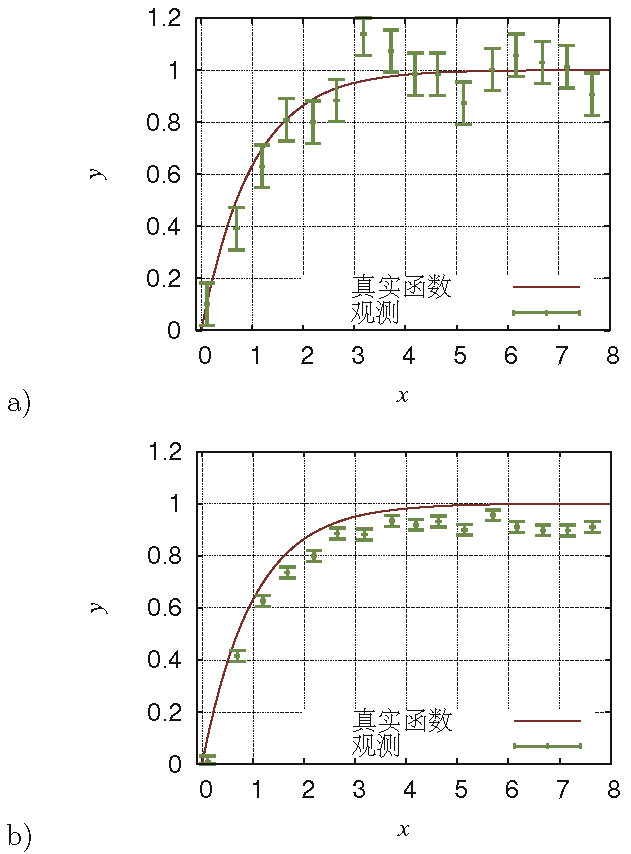


图7. 1 准确度与精度：a）低精度的准确观测，b）高精度但不准确的数据

遗憾的是，检测系统类型的误差非常困难，因此难以对结果的准确性作出说明。系统误差是由于设备校准不充分、依赖于未考虑的影响量测的参数（例如，温度等）或者观测系统偏差造成的。

**例7.1** 在图像处理应用中，观测器可以是检测特殊目标的软件算法。这些目标位置的确定可能会受到照明条件的影响，引入系统偏差。例如，明亮的目标可以通过简单的灰度阈值与深色背景分离，该目标的中心可以从剩余的目标区域中加以确定。如果光源从右侧闪烁，则该目标的左侧显得较暗，这意味着由于所选的阈值相应的图像元素可能被视为背景，造成所得的目标中心自动偏向右侧。

简单地重复实验以获得更多观测通常会减少随机误差。如果随机误差是由仪器（或算法）不确定性引起的，那么使用更先进的仪器或算法可以减少随机误差。应该指出的是，在任何情况下将随机误差降低到远低于系统误差都是没有意义的。

# 估计参数的不确定度

除了确定模型函数的参数之外，还需要估计这些值的不确定度，通常用估计的标准差或标准不确定度表示。

参数的标准不确定度可以直接由下式得到



其中，为关于的Hessian矩阵，为残差向量。



为所有参数的方差—协方差矩阵[2]，它描述参数之间依赖选择的模型函相关性如何。对角元素包含参数的方差。显然不依赖于实际观测无关，但受所选条件和权重的影响。如果用表示观测的相对不确定度，那么通常乘以拟合优度[3][4]



以补偿常数因子（见7.1节）。不确定度表示模型参数的可靠性。它通常被指定为估计参数值的百分比。如果一个或多个参数显示出相对较高的不确定度，则可能指向不合适的模型函数。

的对角元素对应于模型参数的方差，而非对角线元素表示协方差，它可能为负值。

# 数据逼近的不确定度

数据拟合的目的不限于数据点的逼近。数据模型也可以用来在选择其他实验条件时，预测最有可能的测量值，这被称为模型预测。此外，我们感兴趣的是这种模型预测的不确定度。

根据参数的不确定度（式），并利用误差传播原理



得到模型预测的不确定度为



以直线拟合为例



得到偏差



因此



为建模函数的不确定度。由此可见，模型预测的不确定度是条件的一个函数，并随着的增加而增加。图7. 2显示了这种效应。越大，逼近越可能偏离真实的模型函数。

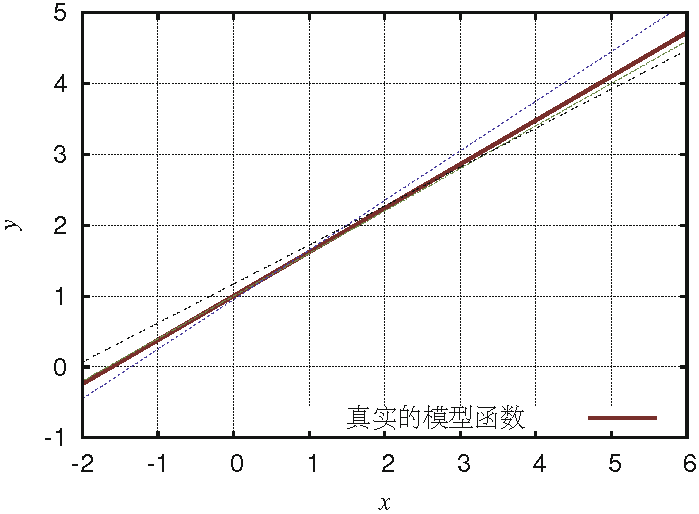


图7. 2 基于直线逼近的模型预测。如果估计的模型参数不准确，则随着的增加，对于给定的条件，预测正确的值变得更加不确定（绿色、蓝色和黑色线条示例）。

# 绘图检验

当数据拟合是自动处理时，数值评估方法至关重要。在一次实验的情况下，我们还建议至少仔细观察残差（）图，并且最好观察最终模型函数与观测的关系图。后者给出了拟合过程是否总体上成功，而前者可以揭示所选择的模型函数和条件与观测之间真实关系的失配程度。

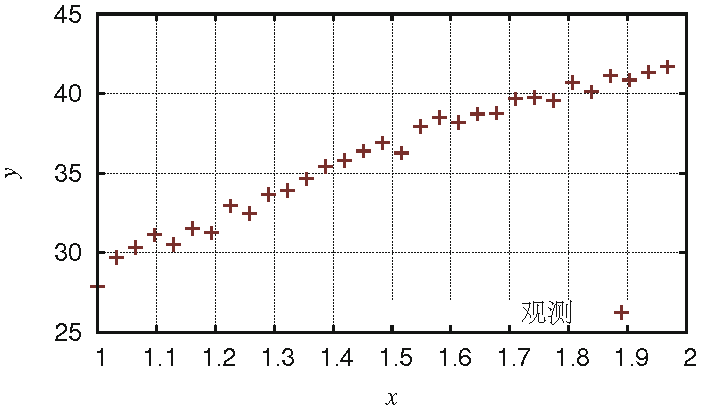


图7. 3某一特定实验的观测

图7. 3给出了31次量测的结果。和之间的关系似乎是线性的。因此，假定



可以正确地描述系统。拟合直线得到的图形如图7. 4所示。乍看之下，这种拟合似乎是合适的。然而，对偏差的检查表明它们在条件范围内并不是均等分布的，而是呈现出一种趋势，见图7. 5。这表明所选模型函数不足以描述该过程，故必须考虑高阶多项式或其他函数。

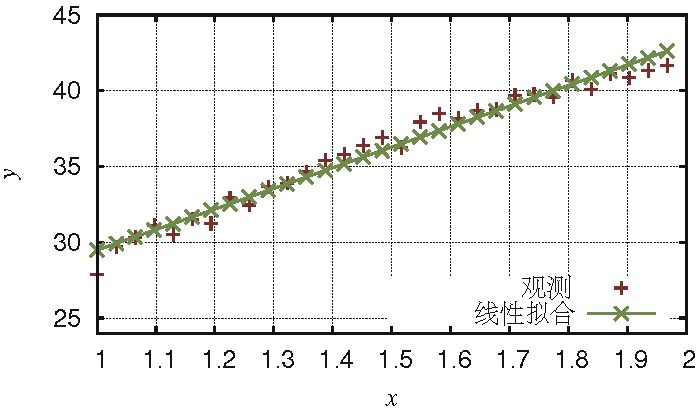


图7. 4 用直线模型拟合数据的结果

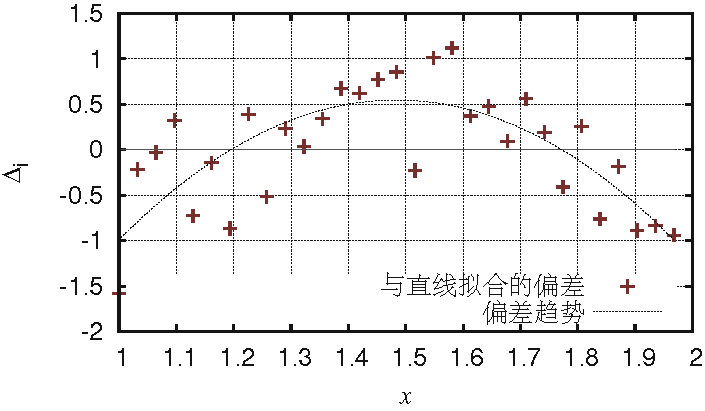


图7. 5 偏差（）图

# 计算示例

以下小节涉及前面章节中介绍的加权数据拟合结果。因为可以相应地推导出其它模型的方程，这里只详细讨论两个简单的模型函数。然后讨论模型失配的问题。

## 常值

常值的极小化问题化简为



使用式，得到协方差矩阵



在本示例中，只是一个标量。拟合优度等于

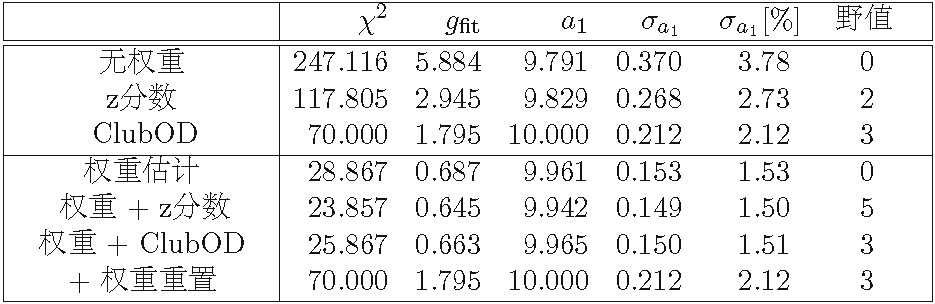


根据式和式，所得到的参数的方差为



因此，均值的不确定度等于。实验次数加倍，则不确定度（随机误差）减少倍。

下表显示了5.5.1节讨论的例子中使用不同权重和野值检测方法对数据拟合改进的比较。



表格的上半部分包含有无野值检测的普通最小二乘逼近（OLS）的结果。使用z分数法或者ClubOD法进行野值检测。表格的下半部分列出了基于加权最小二乘的结果。由此，得到以下结论：

* 因为是极小化的目标，所以它的减小至少表明去除野值满足优化准则。
* （无权重）OLS的应用意味着，即。这导致以下效果：
* 如果引入权重的平均值小于1，则只会减少。
* 假设模型函数被正确选择，表明观测的不确定度高于1。当然，去除干扰后不确定度降低，因此拟合优度也降低。
* 权重的应用使得拟合优度，表明高估了观测的不确定度，即权重平均太小。 这是由选定的加权方案引起的系统效应。
* 以参数值的百分比表示的估计参数的标准不确定度提供了最可靠的信息，即拟合过程是成功的。当然，如果与模型函数偏差较大的观测的权重被降低或甚至被移除，则不确定度降低。

## 直线

在直线示例中，必须根据模型函数



确定两个参数的不确定度。协方差矩阵的计算变得稍微复杂一些。根据式



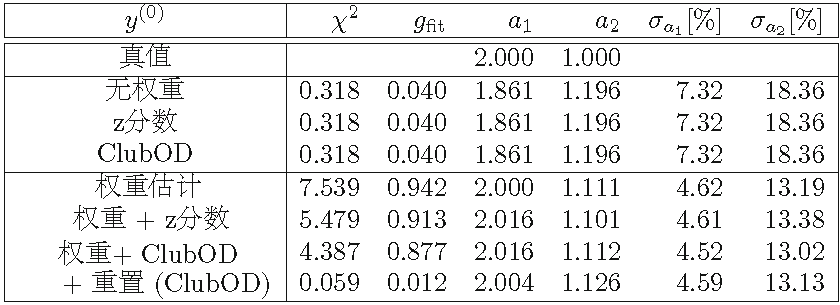
和矩阵求逆公式得到

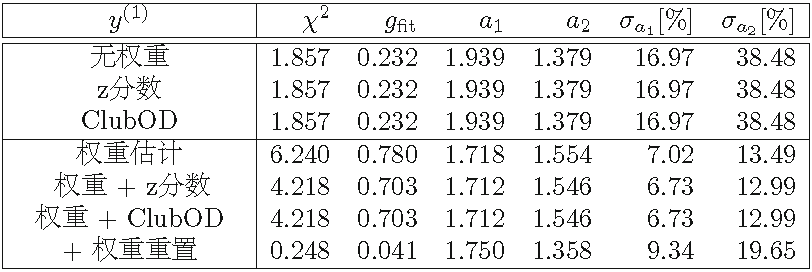


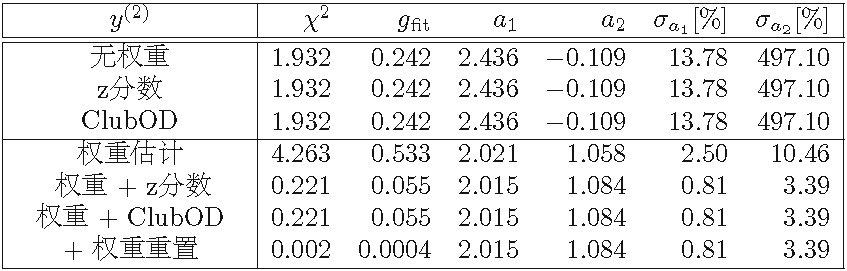
根据的矩阵结构，可以立即得出总共有四个方差或协方差，分别为

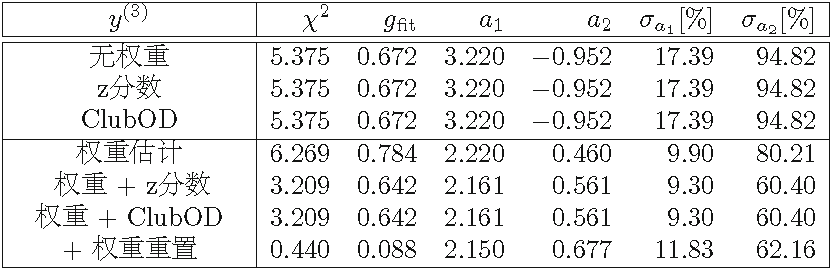


加权和野值检测的影响如下表所示，其中参考5.5.2节中讨论的例子。









根据前一小节的陈述，可以得出关于、和的结论。观测的不确定度小于1是和在所有四种情况下应用权重估计后都变得更高的原因。

这些直线回归的例子特别说明了估计参数的标准不确定度（以百分比的形式）与其正确性之间的高度相关性。超过50％的数值分别明确指向不可信的参数或不充分的模型函数。如果存在野值（且未被去除），那么所选模型函数通常不适合描述观测，并且可以观测到参数的高相对标准不确定度。低于10％的值表示参数具有合理值，然而，最小的不确定度不一定表示参数值最接近真值的，如的和以及的。

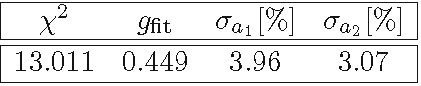
## 模型失配

**示例1**

回顾图7. 3-图7. 5所示的数据集。基于直线模型



和等权重拟合得到以下值



由于估计参数的不确定度较低，因此可以假设所选模型是合适的。然而，图7. 5中的残差图显示了一个更高阶的模型。

拟合优度小于1，即高估了观测的不确定度。可以将解释为式中的因子，由于使用了，所以得到的平方不确定度。实际上，的方差用于通过下式产生数据



估计的增大（与真实值相比）是模型失配的结果。

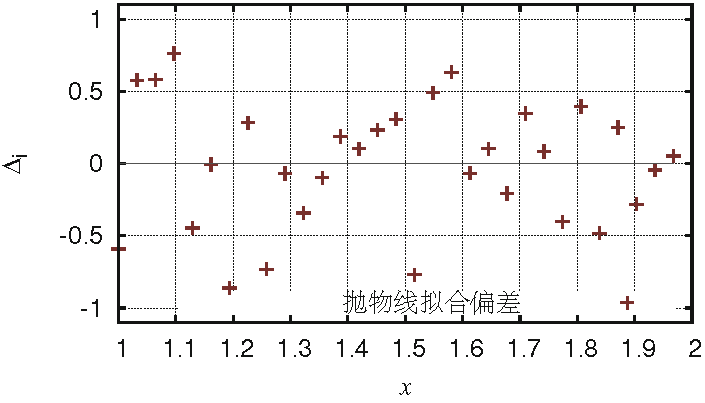
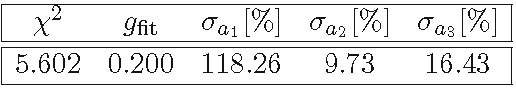


图7. 6 基于抛物线拟合的偏差（）图

现在我们使用抛物线代替直线时，即模型函数为



残差图看起来更合理（图7. 6）。数据拟合评估的结果为

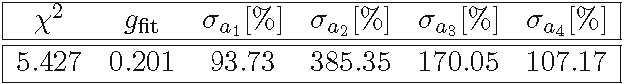


拟合优度变为，得到的观测的平方估计标准差，这已经接近于真值0.16。然而，参数的不确定度急剧增加。

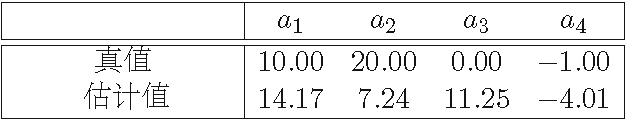
使用真实的三阶模型



结果变得更糟，即



估计的模型参数非常不确定。尽管有点奇怪，但一切都与理论一致。当比较估计参数与真值时，原因就变得清楚了。



估计的参数与真值相差甚远。与数据中的随机误差相比，要拟合的曲线的斜率变化（图7. 3）不够明显，使得无法以足够的准确度估计参数。另外，参数的不确定度通常随着参数个数的增加而增加，而观测的数量是固定的。

**示例2**

让我们来看另一个模型失配的例子。根据下面的模型函数给出一个具有线性趋势的谐波振荡



数据点受随机误差的轻微干扰，服从的高斯分布。图7. 7显示了相应的图。模型参数为和（）。基于高斯分布（）生成加性噪声。该图可以表示一个三次多项式，即



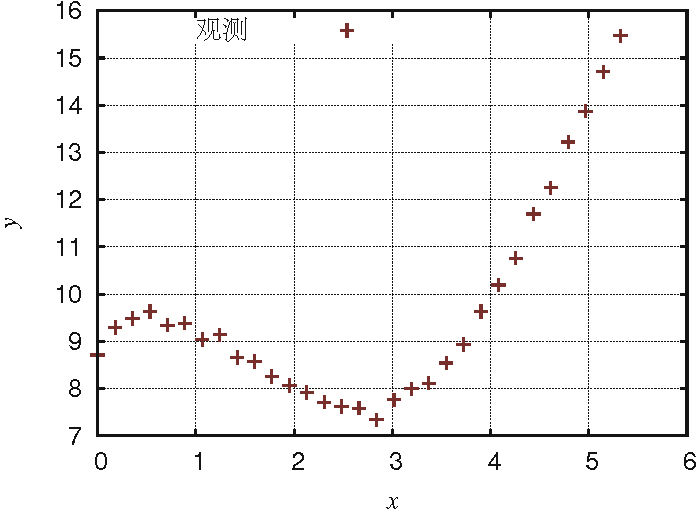
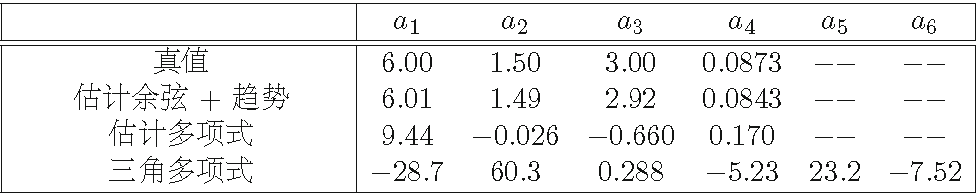


图7. 7 具有趋势谐波振荡

使用二阶三角多项式作为模型函数



用真实的和推测的多项式模型函数拟合数据点得到以下数值结果



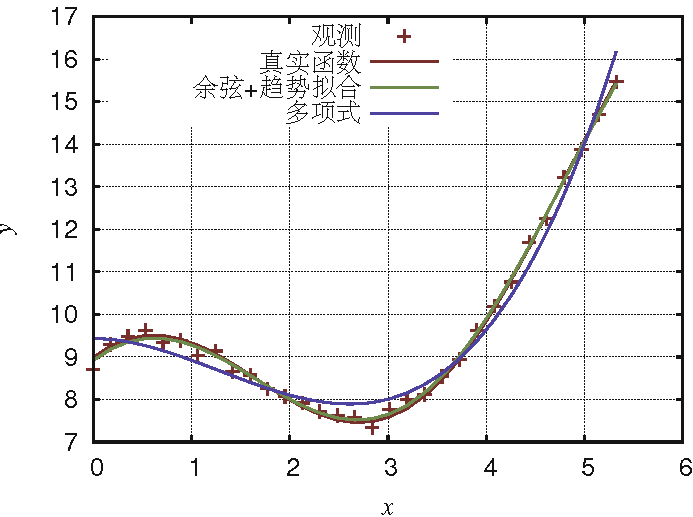


图7. 8 拟合有趋势的谐波振荡

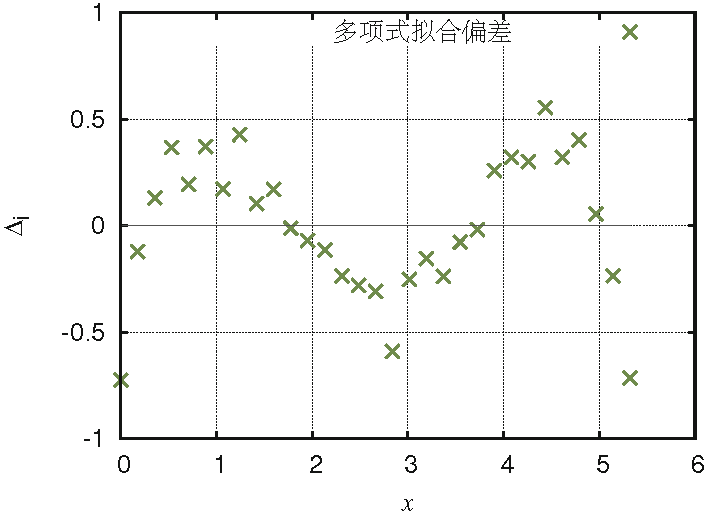
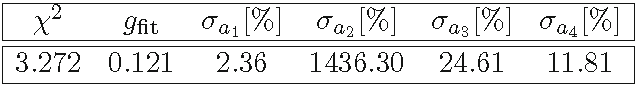


图7. 9 使用三阶多项式拟合有趋势的谐波振荡的偏差图

图7. 8中对应于多项式模型函数的拟合曲线已经揭示了一些问题，当检查残差图时，问题变得更加明显（图7. 9）。此外，参数估计的不确定度明确指出模型失配，如下表所示



参数完全不确定；此外，和的值似乎也不可信。

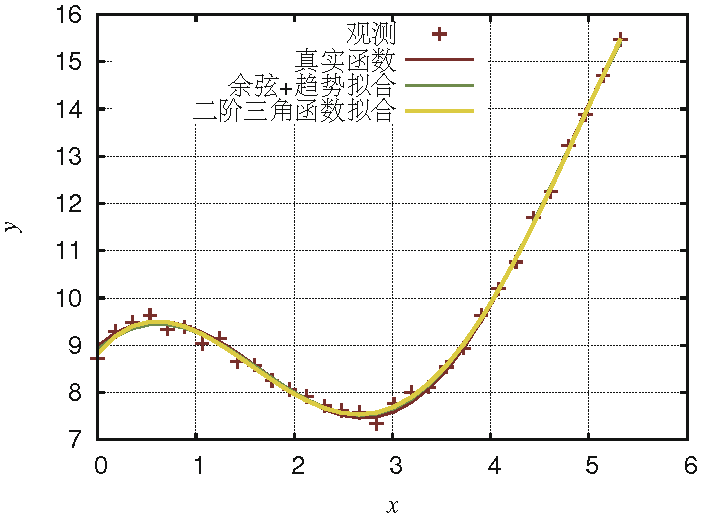
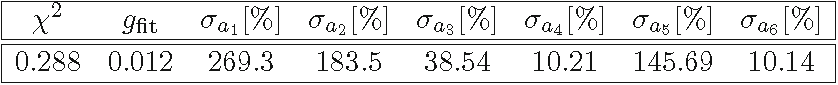


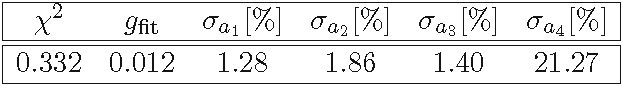
图7. 10 用二阶三角多项式拟合非谐振荡趋势

应用三角多项式得到更好的拟合效果，主要是因为它包含更多的模型参数，见图7. 10。 评估结果为



所有参数的不确定性都比较高。尽管不确定度通常随着（参数个数/观测个数）的减小而增加，但仍然可以得出结论：所选三角函数模型函数是失配的。

使用正确的模型拟合得到



只有相移（）似乎有些不确定。然而，这在拟合振荡时相当典型，特别是由于对偶性。

**示例3**

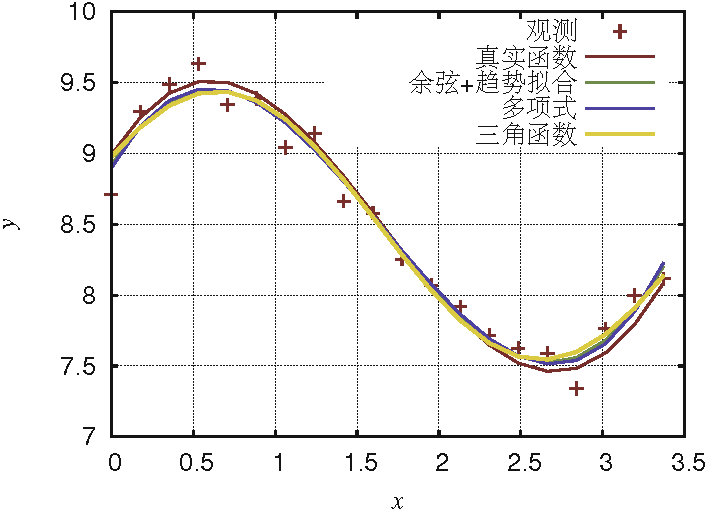
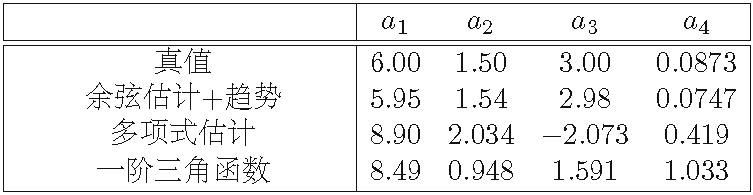
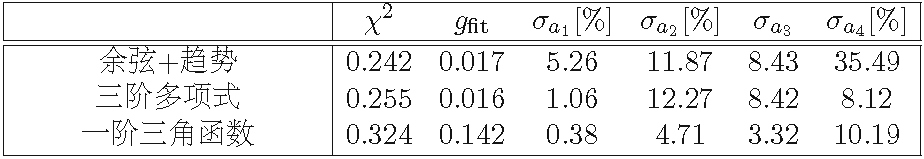
****

图7. 11 拟合有限条件范围内具有趋势的谐波振荡

遗憾的是，模型失配的迹象并不总是如示例2那样明显。从前面的示例中获取相同的数据点，但使用更窄范围的条件，见图7. 11。现在这个范围大约在时被截断，从而减少了趋势的影响（式中的参数a2）。估计参数变为



它们的不确定度变为



尽管正确模型的参数估计几乎保持不变，但它相对多项式方法的优势已经消失。这也反映在拟合曲线图7. 11上，残差图是可信的，见图7. 12。

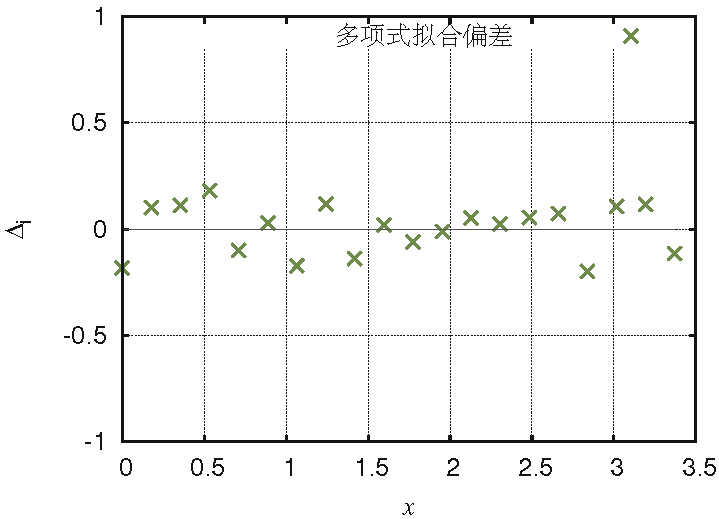


图7. 12 使用三阶多项式拟合带有趋势的谐波振荡的偏差图（有限的条件范围）

总结如下：

* 假定条件范围与模型阶数相比足够窄，可以通过多项式或三角多项式来准确拟合源自未知模型的数据点。
* 观测的条件范围越宽，模型失配的测试就越可靠。

# 参考文献

1. Bevington, P.R.; Robinson, D.K.: *Data reduction and error analysis for the physical science*., 2nd edition, WCB/McGraw-Hill, 1992, ISBN 0-07-911243-9
2. Ryan, T.R.: *Modern Regression Methods*. John Wiley & Sons, Inc., 1997, ISBN 0-471-52912-5
3. Prince, E.; Boggs, P.T.: Least squares. in *International Tables for Crystallography*. Vol.C, Mathematical, physical and chemical tables, 1992, ISBN 0-7923-1638-X, 672–682
4. Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P.: *Numerical Recipes in* C. 2nd edition, Cambridge University Press, 1992, ISBN 0-521-43108-5