各项误差的估计

科研菜鸟
http://www.sciencenet.cn/u/sanshiphy/

2008年5月19日-6月6日

一、用时间平均代替系综平均产生的误差

在湍流理论中，各种统计矩的定义都是用的系综平均。然而在实验或观测中，我们往往只能用有限时间长度的平均来代替系综平均，从而来评价理论的正确性。因此，有必要来计算用有限时间长度平均代替系综平均后产生的误差。LMK1[1] 将误差源分为两种（系统误差和统计误差），并给出了各种误差的计算方法。

1.1 系统误差

考虑一均值为0的平稳过程，其n阶中心矩为$\mu_n = \langle[w(t) - \langle w(t)\rangle]^n\rangle = \langle w^n(t)\rangle$，其中$\langle\cdot\rangle$表示系综平均。根据遍历性假设，

$$\langle w^n\rangle = \lim_{T\to\infty}\mu_n(T) = \lim_{T\to\infty}\frac{1}{T}\int_0^T [w(t) - w_T]^n\,\mathrm{d}t,$$

其中，

$$w_T = \frac{1}{T}\int_0^T w(t)\,\mathrm{d}t.$$

这就是说，当积分时间有限并且$n > 1$时，时间平均和系综平均直接会产生偏差，并且积分时间T越短，偏差越大。这种偏差造成的误差称为系统误差，定义为：

$$SE(n,T) = \frac{\mu_n - \langle\mu_n(T)\rangle}{\mu_n}.$$

为了计算系统误差，可构造一个非高斯的随机过程，并利用它来计算各阶矩所产生的误差。LMK 构造了如下随机过程：

$$w(t) = z(t) + a\frac{z^2(t) - \langle z^2(t)\rangle}{\sigma}, \tag{1}$$

其中，a是待定参数，$z(t)$是均值为0，方差为σ的高斯过程，并且自相关函数为

$$R_{zz}(t_2 - t_1) = \langle z(t_1)z(t_2)\rangle = \sigma^2\rho_0(t_2 - t_1) = \sigma^2\exp\left(-\frac{|t_2 - t_1|}{\mathcal{T}_0}\right). \tag{2}$$

[1] Journal of Atmospheric and Oceanic Technology, **11**, 661(1994)

等式（2）中的$\mathcal{T}_0$是$z(t)$的积分时间尺度，定义为

$$\mathcal{T}_0 = \int_0^\infty \rho_0(t)\,\mathrm{d}t.$$

同样地，假设$w(t)$的归一化自相关函数$\rho(t)$也是指数形式，

$$\rho(t_2 - t_1) = \frac{1}{\mu_2}\langle w(t_2)w(t_1)\rangle = \exp\left(-\frac{|t_2 - t_1|}{\mathcal{T}}\right), \tag{3}$$

其中积分时间尺度$\mathcal{T}$为

$$\mathcal{T} = \int_0^\infty \rho(t)\,\mathrm{d}t. \tag{4}$$

可以证明，1）$\mathcal{T}_0$ 和$\mathcal{T}$的关系是：

$$\mathcal{T} = \frac{1+a^2}{1+2a^2}\mathcal{T}_0;$$

2）$w(t)$的（双边）功率谱密度是：

$$S_{ww}(f) = \frac{2\mathcal{T}\mu_2}{1+(2\pi f)^2\mathcal{T}^2}; \tag{5}$$

3）$w(t)$的偏斜度(skewness)和陡峭度(kurtosis)是：

$$\begin{aligned} S &\equiv \frac{\mu_3}{\mu_2^{3/2}} = \frac{2a(3+4a^2)}{(1+2a^2)^{3/2}} \\ K &\equiv \frac{\mu_4}{\mu_2^2} = \frac{3(1+20a^2+20a^4)}{(1+2a^2)^2}. \end{aligned} \tag{6}$$

当$T \gg \mathcal{T}$时，LMK2(p.48)[2]计算了二阶、三阶和四阶矩的统计误差：

$$\begin{aligned} SE(2,T) &= 2\frac{\mathcal{T}}{T} \\ SE(3,T) &= 3\left[2 - \frac{1}{(1+a^2)(3+4a^2)}\right]\frac{\mathcal{T}}{T} \\ SE(4,T) &= \frac{4(1+2a^2)(1+27a^2+18a^4)}{(1+a^2)(1+20a^2+20a^4)}\frac{\mathcal{T}}{T}. \end{aligned} \tag{7}$$

1.2 随机误差

由于样本有限，每一次针对不同样本的时间平均都会不同，并且与时间平均的系综平均$\langle\mu_n(T)\rangle$之间存在偏差，这种偏差产生的误差称为随机误差，并定义为：

$$RE(n,T) = \frac{\sigma_n(T)}{\mu_n} = \frac{\sqrt{\langle[\mu_n(T) - \langle\mu_n(T)\rangle]^2\rangle}}{\mu_n}.$$

[2]NCAR/TN-389+STR, 53pp (1993)

同样地，当$T \gg \mathcal{T}$时，利用上述非高斯过程的数学模型，LMK2(p.48)计算了二阶、三阶和四阶矩的随机误差：

$$\begin{aligned} RE(2,T) &= \sqrt{\frac{2(1+32a^2+22a^4)}{(1+a^2)(1+2a^2)}\frac{\mathcal{T}}{T}} \\ RE(3,T) &= \sqrt{\frac{(1+2a^2)(1+147a^2+1476a^4+780a^6)}{a^2(1+a^2)(3+4a^2)^2}\frac{\mathcal{T}}{T}} \\ RE(4,T) &= \sqrt{\frac{(1+2a^2)(7+1108a^2+24708a^4+109200a^6+46296a^8)}{3(1+20a^2+20a^4)^2(1+a^2)}\frac{\mathcal{T}}{T}}. \end{aligned} \tag{8}$$

比较等式（7）和（8），可得：

$$\frac{\sigma_n(T)}{\mu_n - \langle\mu_n(T)\rangle} = A(n,a)\sqrt{\frac{T}{\mathcal{T}}},$$

其中，

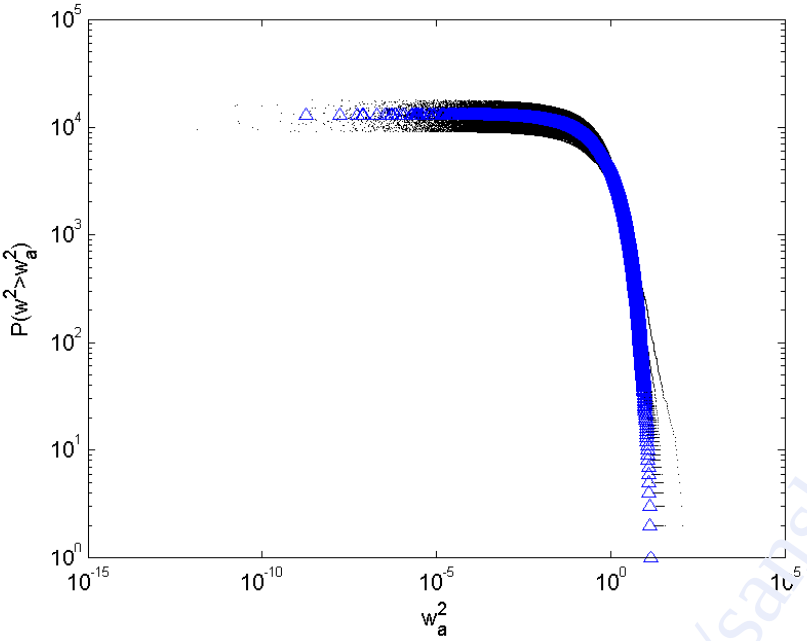
$$\begin{aligned} 0.7 \le A(2,a) < 2.345 \\ 1.06494 \le A(3, a > 0.1) < 1.646 \\ 0.3819 \le A(4,a) < 1.22. \end{aligned}$$

因此，当$T \gg \mathcal{T}$时，系统误差远小于随机误差，可以不考虑。

1.3 分析步骤

下面以结构函数$D_n(\Delta t) = \langle[u(t+\Delta t) - u(t)]^n\rangle$为例，介绍系统和随机误差的计算步骤。先对原始数据进行去趋势和去野点处理，然后在xy平面内做坐标变换，将x和y方向的风速投影到平均风速方向和垂直于平均风速的方向，从而得到处理后的风速分量$v(t)$。在此例中，$w(t,\Delta t) = u(t+\Delta t) - u(t)$，其中$\Delta t$的最大值取为15 min（总的资料长度取为30 min），并且在固定的Δt下，有$T(\Delta t) = t_{max} = (N - \frac{\Delta t}{f} - 1)f$. 我们根据以下步骤来分析系统误差和随机误差：

- 参数$a(\Delta t)$的估计。每隔10s计算一次方差的尾分布$P(\frac{w^2}{\mu_2^2} > \frac{w_a^2}{\mu_2^2})$，这样对于15分钟延滞时间而言，总共计算了90次尾分布。然后利用（1）式求出的尾分布与实测值比较，求出大致的a的大小，如图一所示$a \approx 0$。
- 求积分时间尺度$\mathcal{T}$，主要有三种方法：一是根据定义（4）式直接计算；二是根据（3）式拟合求出$\mathcal{T}(\Delta t)$；利用这两种方法，首先需要求出分段的自相关函数。计算每隔10s进行一次，如图二黑点所示。红线是实测平均值，在$[0, 50s]$ 内积分值是$\mathcal{T} = 2.9898s$，在$[0, 18.7s]$内积分是值是$\mathcal{T} = 3.3621s$（18.8s是自关联函数第一次为0的时间）。蓝线是在$[0, 50s]$内的拟合曲线，其表达式为$y = -0.006749exp(0.004651x) + 0.855exp(-0.2532x)$，因此可估计出$\mathcal{T} = 3.949s$。同样我们也可以在$[0, 18.7s]$ 内拟合，拟合曲线的表达式为$y = 0.8525exp(-0.2598x)$，因此估计出$\mathcal{T} = 3.849s$；三是求出信号的（双边）功率谱密度，并用（5）式拟合求出$\mathcal{T} \approx 3.32$，如图三所示；



图一：标准化方差尾分布。点是实测值，三角是标准正态分布。该图说明了$w(t)$可近似看做是高斯过程，因此$a \approx 0$.

- 如果$T(\Delta t) \gg \mathcal{T}$，根据（7）和（8）式便可求出积分时间为$T(\Delta t)$的系统和随机误差。本例中，$\mathcal{T} \approx 4s$，而$T_{min} = 15min \gg 4s$，因此根据上述两个式子估计出来的系统和随机误差如图四所示。

二、泰勒假设带来的系统误差

2.1 泰勒假设及其适用条件

许多湍流理论涉及速度场随空间的分布规律。然而，在实验中我们很难测量速度场的空间分布，而较容易得到的是它在某一点上随时间的变化。泰勒假设指出，在平均风速很大的情况下，可以设想上流一定范围内的脉动速度场还没有来得及演化，就在平均风速的携带下通过观测仪器，这时观测仪器记录的风速随时间的变化就反映了上流速度场的空间变化。更准确的泰勒假设的阐述如下，设湍流脉动速度场（去掉了平均速度）为$\mathbf{u}(\mathbf{x}, t)$，平均风速为$\mathbf{U}$，在某些情况下，泰勒假设指出：

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{u}(\mathbf{x} - \mathbf{U}t, 0).$$

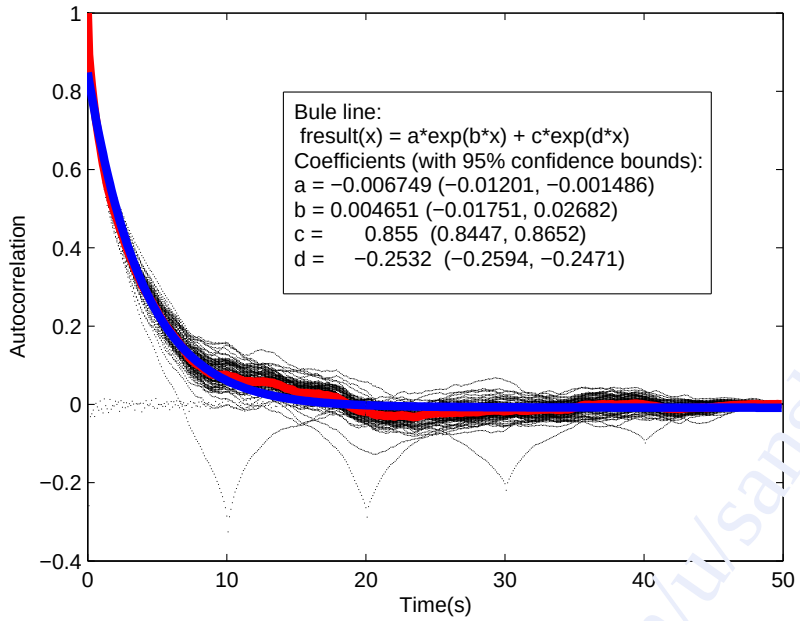
泰勒假设在以下条件下不成立：

- 相比于平均风速，脉动速度场的演化速度过快。根据K41理论，我们可以用下式来表示在特征范围r内的脉动速度场的特征演化速度：

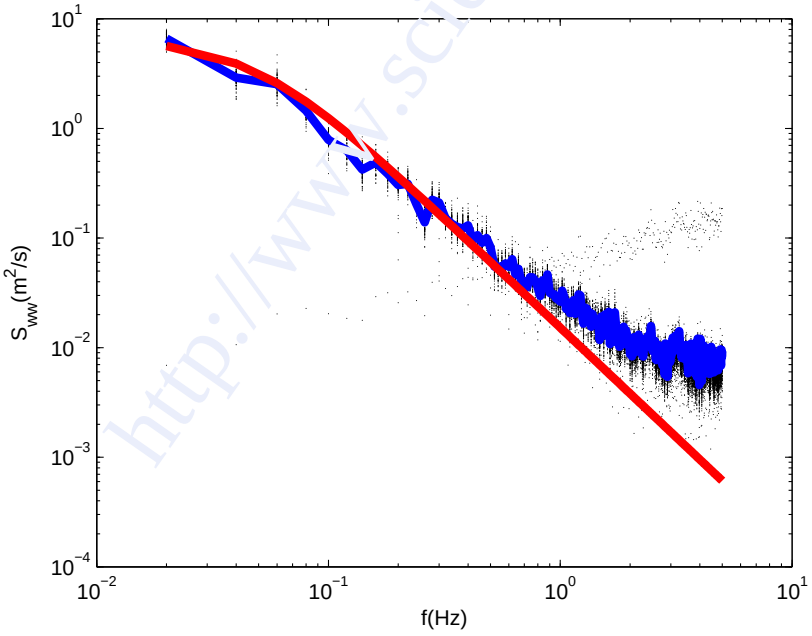
$$\sigma(r) \equiv \sqrt{\langle [u_1(\mathbf{x} + \mathbf{r}, t) - u_1(\mathbf{x})]^2 \rangle},$$

其中，u_1是平均风速方向的脉动速度。因此，只有当$\dfrac{\sigma}{U} \ll 1$时，脉动速度场的演化才能被忽略。对于均匀湍流，

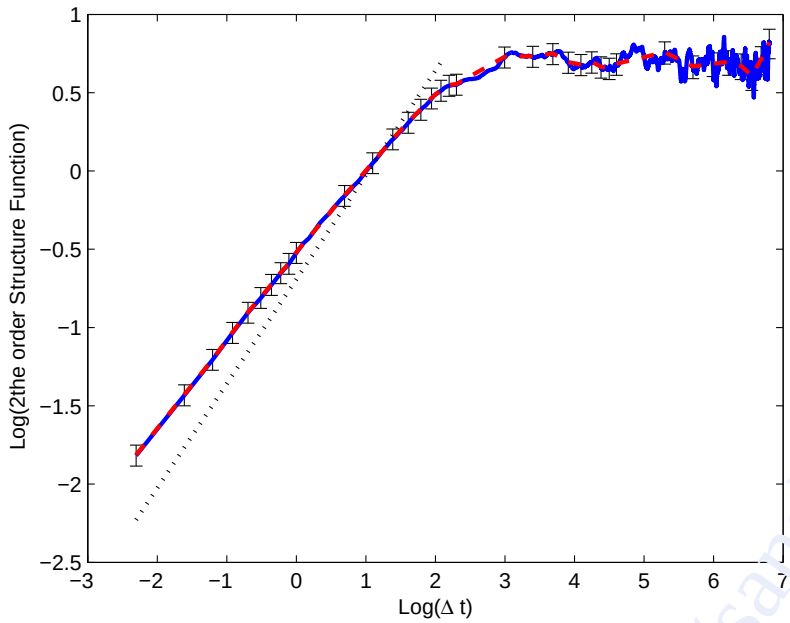
$$\frac{\sigma(r)}{U} = \frac{\sqrt{2\langle u_1^2(\mathbf{x}) \rangle (1 - f(r))}}{U} \ll 1. \tag{9}$$



图二：自关联函数。点是实测值，红线是实测的平均值. 蓝线是指数拟合曲线。



图三：功率谱密度。点是实测值，蓝线线是实测的平均值，红线是拟合曲线。



图四：结构函数。实线是实测值，红虚线是考虑系统误差后的值，点线是理论曲线$\sim \Delta t^{2/3}$。

- 携带脉动速度场大尺度风场的速度（平流速度，convection velocity）并不总是**U**，而是围绕这个平均值做随机变化，如果随机变化的值偏离平均值太远，泰勒假设不成立。这也就是说，如果大尺度速度场的特征脉动速度

$$\sigma(L) \approx \sqrt{2\langle u_1^2(\mathbf{x})\rangle} \ll U, \tag{10}$$

  那么大尺度风场的随机变化产生的影响可以被忽略掉。综合（9）和（10）式，对于均匀湍流场，泰勒假设成立的条件是：

$$I \equiv \frac{\sqrt{\langle u_1^2\rangle}}{U} \ll 1, \tag{11}$$

  其中I被称为湍流强度。

- 对于切变湍流场，平均速度的切变造成了上流湍流脉动场的形变，因此泰勒假设不成立。只有当脉动速度场的范围r 足够小时，平均速度的切变可以忽略。具体说来，如果假设速度场存在垂直方向的切变，那么

$$\frac{r}{U}\frac{dU(z)}{dz} \ll 1, \tag{12}$$

  才能保证范围为r的脉动速度场其平移不受切变的影响。对于切变湍流场，泰勒假设成立的条件是（11）和（12）均成立。

2.2 对什么样的湍流场进行泰勒假设修正

当湍流强度并不远小于1的时候，我们需要修正泰勒假设。在下面的推导中，我们假设湍流场存在切变，但是所考虑的脉动速度场范围比较小，以至于湍流场的切变效应对于它们可以忽略不计。准确地说，我们考虑满足（12）式的湍流速度场，也就是

$$r \ll \frac{U}{dU/dz}. \tag{13}$$

另外，我们希望湍流脉动场的演化速度不大，以至于脉动场可以被看作是冻结的，这要求（9）成立，也即是

$$\sqrt{[1-f(r)]} \ll \frac{\sqrt{\langle u_1^2(\mathbf{x})\rangle}}{U}.$$

若$A \leq aB$，我们就认为$A \ll B$，其中a是小于1的常数。参见图二，可设关联函数$f(r)$指数衰减，即

$$f(r) = \exp(-\frac{r}{L}),$$

其中，L是关联长度：

$$L \equiv \int_0^\infty f(r)dr.$$

当r很小时，可对$f(r)$进行Taylor展开：

$$f(r) = 1 - \frac{r}{L} + \mathcal{O}\left(\frac{r}{L}\right),$$

其中高阶项为；

$$\mathcal{O}(r/L) = \exp(-\frac{r}{L}) - 1 + \frac{r}{L}$$

保留一阶项，脉动场冻结的判据为

$$r \leq a^2 L \frac{\langle u_1^2\rangle}{U^2} \tag{14}$$

对于我们所关心的湍流，一般$I < 0.3$，如果取$a = 1$，那么用上式替代$1 - f \leq a^2\langle u_1^2\rangle/U^2$作为判据，其误差不大于:

$$\frac{\mathcal{O}(r/L)}{a^2\langle u_1^2\rangle/U^2} = 5\%.$$

因此，当通过实验确定在某一误差范围内的a的值后，我们修正满足以下条件的泰勒假设：1）湍流强度并不远小于1；2）（12）和（13）式同时成立，当然在同一确定的误差范围内，两式中的系数a可能并不相同。这也就是说，小尺度湍流脉动场被大尺度均匀风场携带，近似冻结地经过观测点，但是由于湍流强度并不远小于1，或者说大尺度风场存在较大的随机起伏，使得原始的泰勒假设需要修正。

2.3 泰勒假设的修正方法

Hill在1996年提出了一种系统地修正泰勒假设的方法[3]，该方法可以修正结构函数，功率谱以及关联函数。这种方法实际上是Lumely两项展开模型[4]的推广，而Wyngaard和Cllfford通过平流速度的高斯分布模型[5]说明了Lumely的模型在湍流强度小于0.3时，能较为准确地替代全展开。例如：对于速度场的一阶导数两者没有差别，对于速度场的二阶导数，两者的差别不大于15%. 因此，Hill提出的修正方法，不仅2.1中列出的两个条件要满足，并且$I < 0.3$，下面我们就来介绍这种修正方法。

[3]Atmospheric Research, **40**, 153(1996)
[4]The Physics of Fluids, **8**, 1056(1965)
[5]Journal of the Atmospheric Science, **34**, 922 (1977)

设观测点的位置为原点，观测区间是[0,T]，观测到的湍流场脉动风速为$\mathbf{u}(0,t)$. 由于平流速度的起伏主要与大尺度风场有关，因此我们将原始风场分为高通成分和低通成分，低通的这部分就近似代表了大尺度风场，或平流速度，用$\mathbf{U}(0,t)$表示。我们将整个观测时间分成许多段，每一段内的平流速度可以近似看作是常数。正因为如此，每一段内的风场满足2.1中泰勒假设适用的条件，因此有

$$z(0,t+\tau) \approx z(\mathbf{U}\tau,t),$$

其中z可以表示结构函数，关联函数或者混合矩等统计量。令每一段内的平流速度与整个区间的平均速度差为$\Delta\mathbf{U}$, 即

$$\Delta\mathbf{U} = \bar{\mathbf{U}} - \mathbf{U}$$

其中用上横线表示整个区间的时间平均。因此，如果令$\rho = -\Delta\mathbf{U}\tau$，$\mathbf{r} = \bar{\mathbf{U}}\tau$，则：

$$z(\mathbf{U}\tau,t) = z(\rho+\mathbf{r},t)$$

其中ρ表示由于平流速度的偏差引起的小尺度速度脉动场的位移变化。对ρ进行Taylor展开，

$$\begin{aligned} z(\rho+\mathbf{r},t) &= z(\mathbf{r},t) + \rho_n z(\mathbf{r},t)|_n + \frac{1}{2}\rho_n\rho_p z(\mathbf{r},t)|_{np} + \cdots, \\ &= z(\mathbf{r},t) - \tau\Delta U_n z(\mathbf{r},t)|_n + \frac{\tau^2}{2}\Delta U_n\Delta U_p z(\mathbf{r},t)|_{np} + \cdots \end{aligned}$$

其中"$\cdots|_n$"表示对r_n的导数。对上式在整个时间段进行平均，考虑到$\Delta\bar{\mathbf{U}} = 0$并假设$\Delta U$与$z$统计独立，因此可以得到：

$$Z(\rho+\mathbf{r},0) = Z(\mathbf{r},0) + r^2\sigma_{np}Z(\mathbf{r},0)|_{np} + \cdots,$$

其中，

$$\begin{aligned} Z(\mathbf{r},0) &= \bar{z}(\mathbf{r},t) \\ \mathbf{r} &= \bar{\mathbf{U}}t \\ \sigma_{np} &= \frac{\overline{\Delta U_n\Delta U_p}}{2\bar{U}^2}. \end{aligned}$$

因为高通成分的幅度相对于低通成分很小（参见Frisch书中23页的示意图[6]），因此我们可以用观测到的全湍流场代替高通统计量：

$$\sigma_{np} = \frac{\overline{u_n u_p}}{2\bar{U}^2},$$

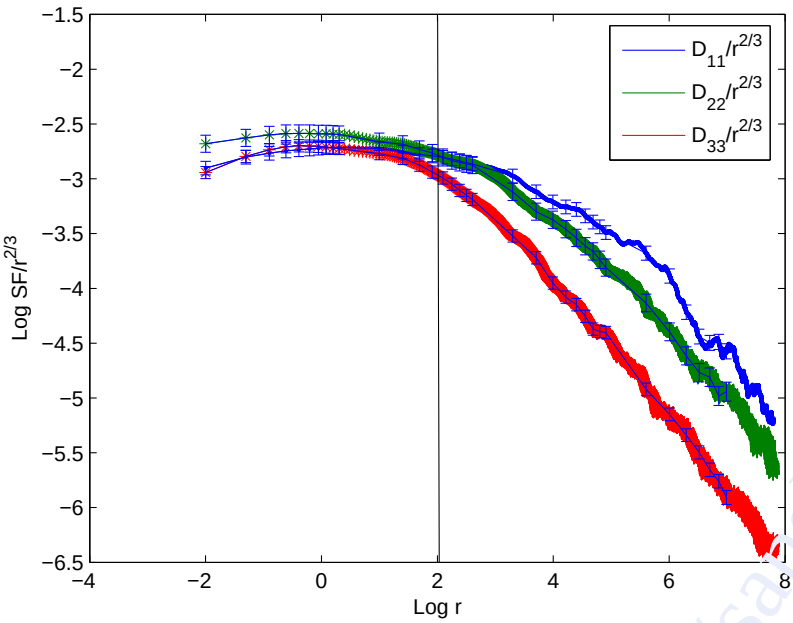
因此，保留到二阶项的泰勒假设修正为：

$$Z(0,\tau) \approx Z(\mathbf{r},0) + r^2\sigma_{np}Z(\mathbf{r},0)|_{np} \equiv Z(\mathbf{r},0) + P.$$

Hill计算了二阶结构函数D_{ij}的修正项P在局地各向同性假设下的表达式：

$$\begin{aligned} P_{11} &= D''_{11}\sigma_{11} + D'_{11}(\sigma_{22}+\sigma_{33}) + 2(D_{22}-D_{11})(\sigma_{22}+\sigma_{33}) \\ P_{22} &= 2(D_{22}-D_{11})\sigma_{22} + D''_{22}\sigma_{11} + D'_{22}(\sigma_{22}+\sigma_{33}), \end{aligned}$$

[6]U. Frisch, *Turbulence*,Cambridge University Press, 1995



图五：补偿结构函数函数。点是Taylor修正值，线是实测值，误差棒是随机误差，湍流强度0.5

其中

$$\begin{aligned} D' &= r\frac{dD}{dr} \approx \frac{dD}{d\tau} \\ D'' &= r^2\frac{d^2D}{dr^2} \approx \tau^2\frac{d^2D}{d\tau^2}. \end{aligned}$$

利用上述误差计算式我们对风速观测资料做了分析，结果如图五所示，可以发现即使湍流强度达到0.5，Taylor修正的影响也不是很大。

Wyngaard 和Clifford 根据Lumely两项展开模型计算了导数二阶矩的修正项P在局地各向同性假设下的表达式：

$$\begin{aligned} P_{1,1} &= \overline{(u_{1,1})^2}(\sigma_{11} + 2\sigma_{22} + 2\sigma_{33}) \\ P_{2,1} &= \overline{(u_{2,1})^2}(\sigma_{11} + 0.5\sigma_{22} + \sigma_{33}). \end{aligned}$$

因此单位质量能量耗散率的修正项为：

$$P_\varepsilon = 15\nu P_{1,1}.$$