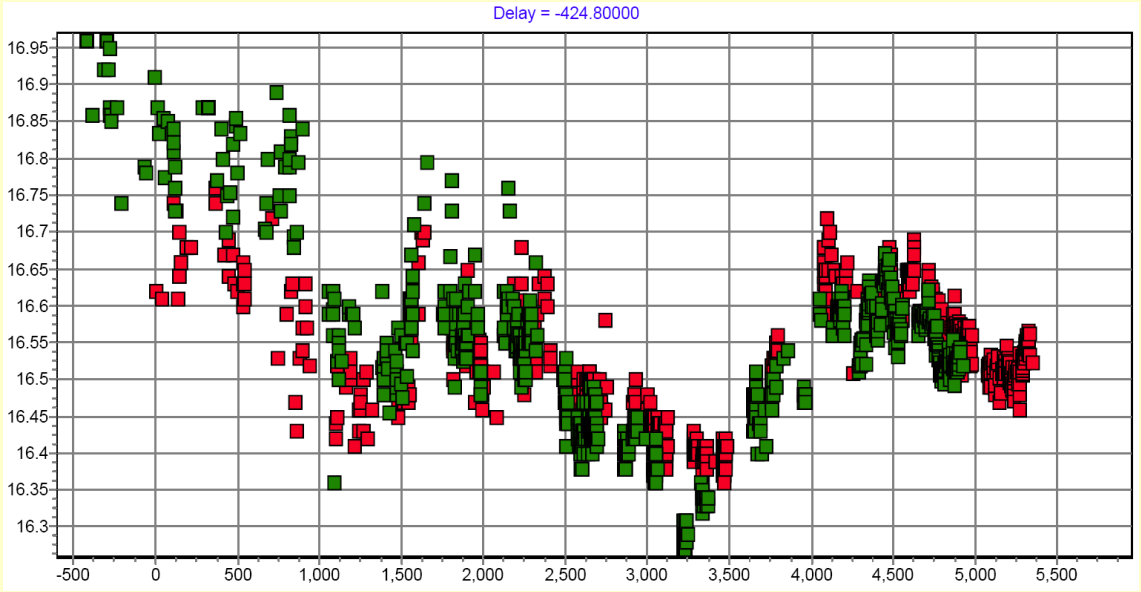
Nonparametric Dispersion Estimation

- Theory
- Applications in frequency analysis
- Time delay estimation



Estimation of dispersion through subsampling

For a set of observations $y_i, i = 1, 2, \ldots, N$ we can observe that

$$\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - y_j)^2 =$$

$$= \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\big((y_i - \mu) - (y_j - \mu)\big)^2 =$$

$$= \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 - \frac{1}{N^2}\Big(\sum_{i=1}^{N}(y_i - \mu)\Big)^2.$$

If we additionally set $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$

$$
\begin{aligned}
S^2 &= \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (y_i - y_j)^2 = \\
&= \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu)^2 .
\end{aligned}
$$

Let $K = \frac{N(N-1)}{2}$ be the total number of squares

$$S^2 = \frac{1}{K}\sum_{k=1}^{K}\frac{(y_{i(k)} - y_{j(k)})^2}{2} = \frac{1}{K}\sum_{k=1}^{K} d_k$$

The sequence of half squares can be looked upon as a general finite sample from which we can draw random subsamples.

$$d_k = \frac{(y_{i(k)} - y_{j(k)})^2}{2}$$

$$d_{k(l)}, l = 1, 2, ..., L$$

In this case the mean values:

$$\hat{S}^2 = \frac{1}{L}\sum_{l=1}^{L} d_{k(l)}$$

approximate the original estimate S^2

How good the estimates are?

$$\mathbf{D}(\hat{S}^2) = \mathbf{E}(\hat{S}^2 - \mathbf{E}\hat{S}^2) = \frac{(K-L)D^2}{KL}$$

D^2 is the dispersion for the full sample

$$D^2 = \frac{\sum_{k=1}^{K}(d_k - \bar{d})^2}{K-1}$$

We see that when L approaches K the approximation error vanishes

The important point is that even for small values of L the subsampling error is as small as D^2/L. The subsampling is used here as computing device to get approximates
for S^2 and the subsampling error arises as a result of using a subsample instead of the full set of squared differences.

Typical run of

D^2 – dependence on

number of pairs

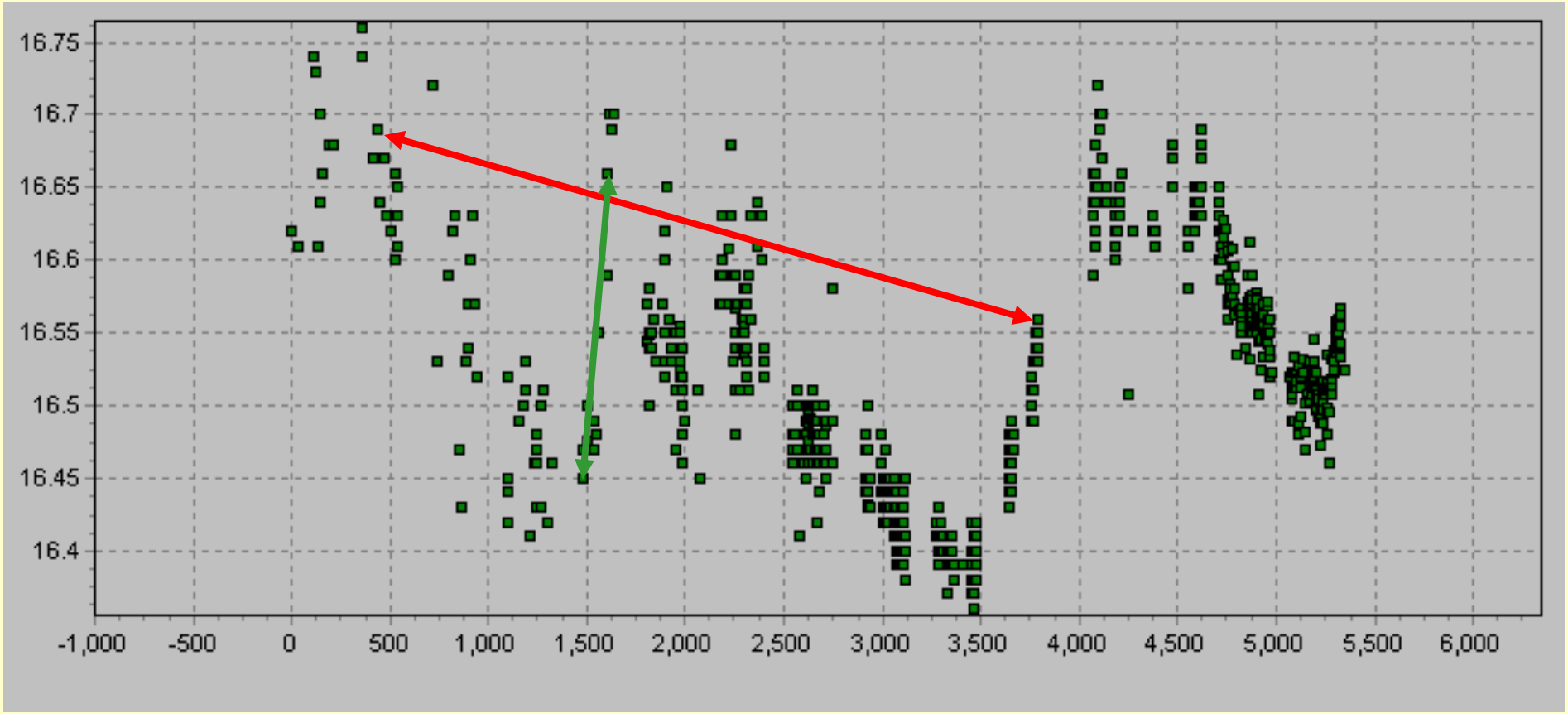
Trend

More interesting to us is a general regression model

$$y_i = g(t_i) + \epsilon_i = g_i + \epsilon_i, \quad i = 1, 2, ..., N$$

$$\begin{aligned}
S^2_{\text{obs}} &= \frac{1}{2K} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (y_i - y_j)^2 = \\
&= \frac{1}{2K} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} ((e_i - e_j) + (g_i - g_j))^2 = \\
&= S^2_{\text{err}} + S^2_{\text{trend}} + \\
&\quad + \frac{1}{K} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (e_i - e_j)(g_i - g_j),
\end{aligned}$$

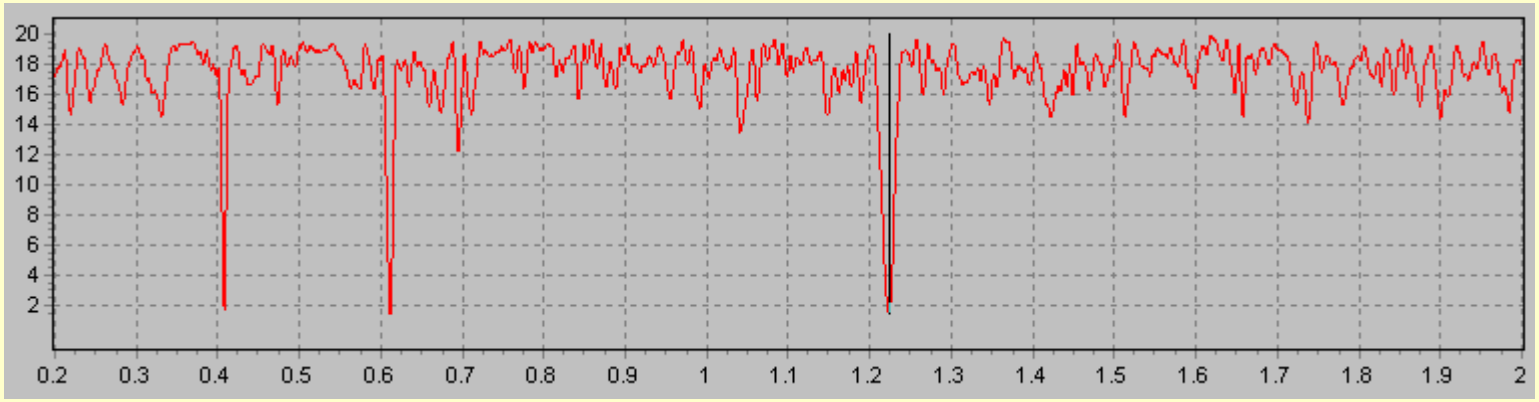
$$G(\delta, t_i, t_j) = \begin{cases} 1 & |t_i - t_j| \leq \delta \\ 0 & \text{otherwise,} \end{cases}$$

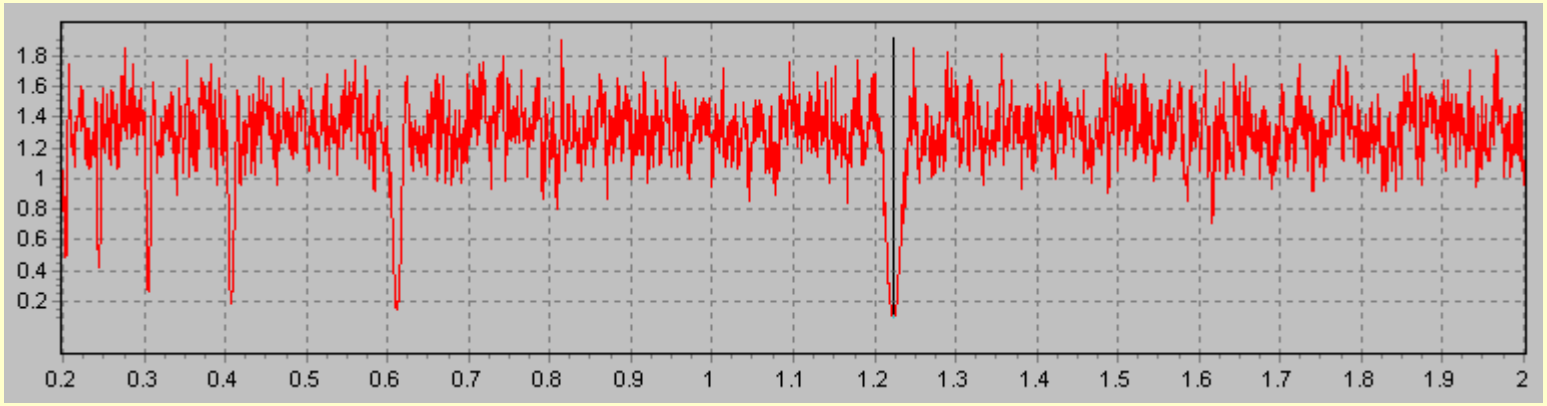


$$\hat{S}^2(\delta) = \frac{1}{2} \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} G(\delta, t_i, t_j)(y_i - y_j)^2}{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} G(\delta, t_i, t_j)}$$

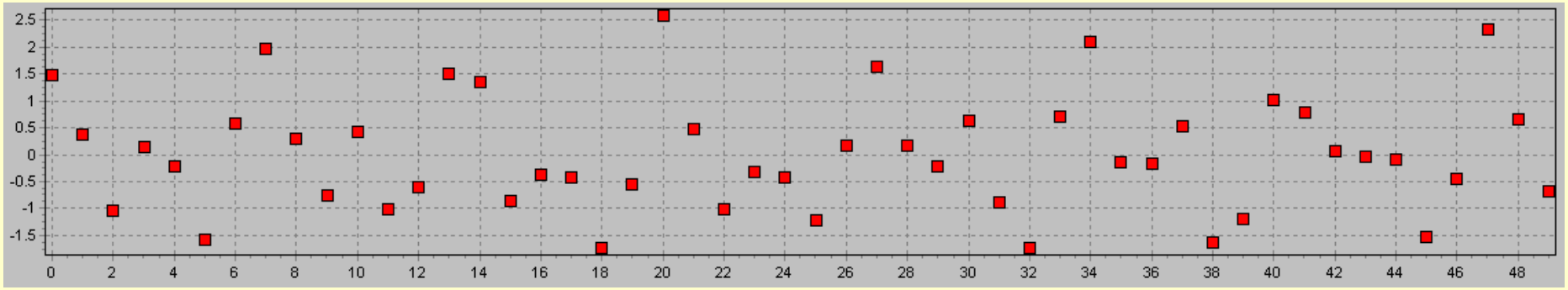
Simplest variant
(String Length?)

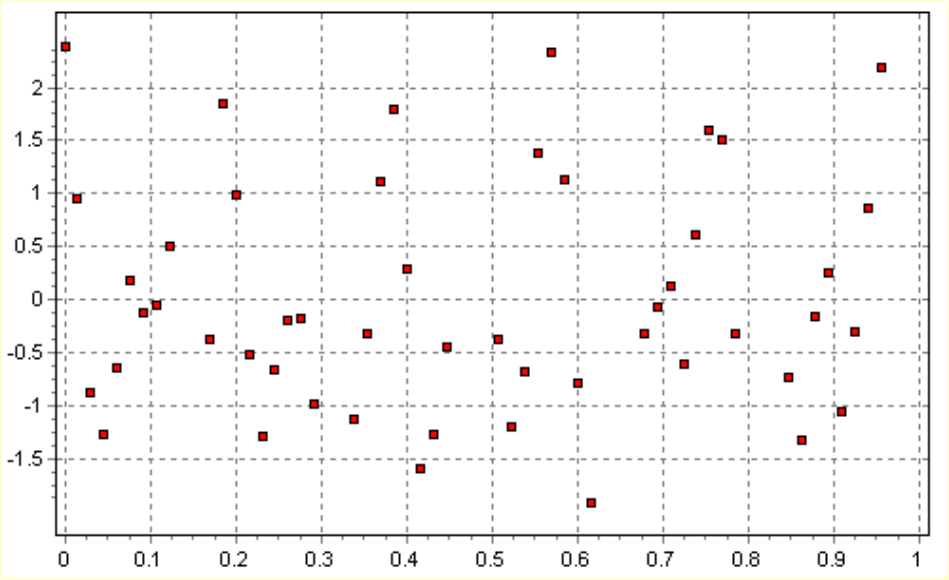
$$\hat{S}^2 = \frac{1}{2K} \sum_{k=1}^{K-1} (y_{k+1} - y_k)^2$$

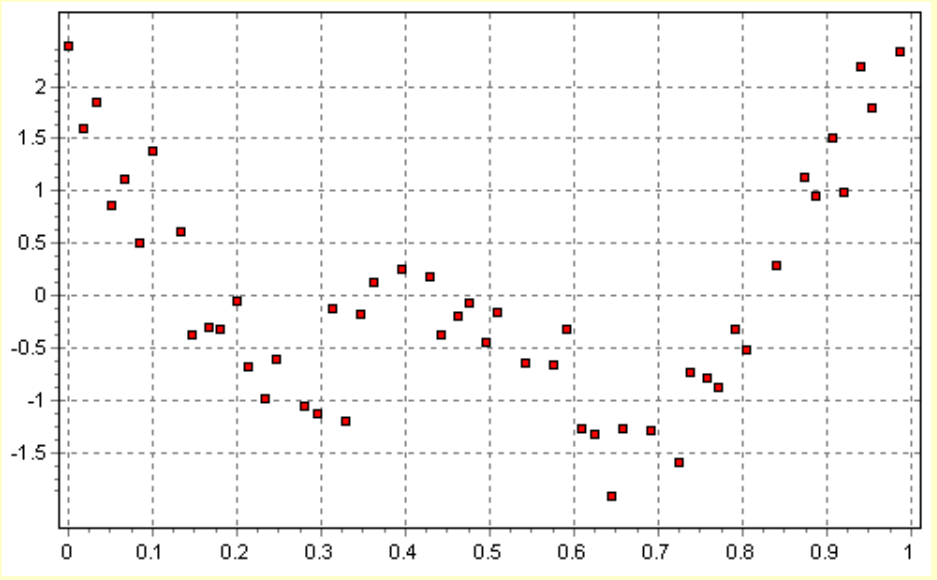




Phase-process diagram (folding)







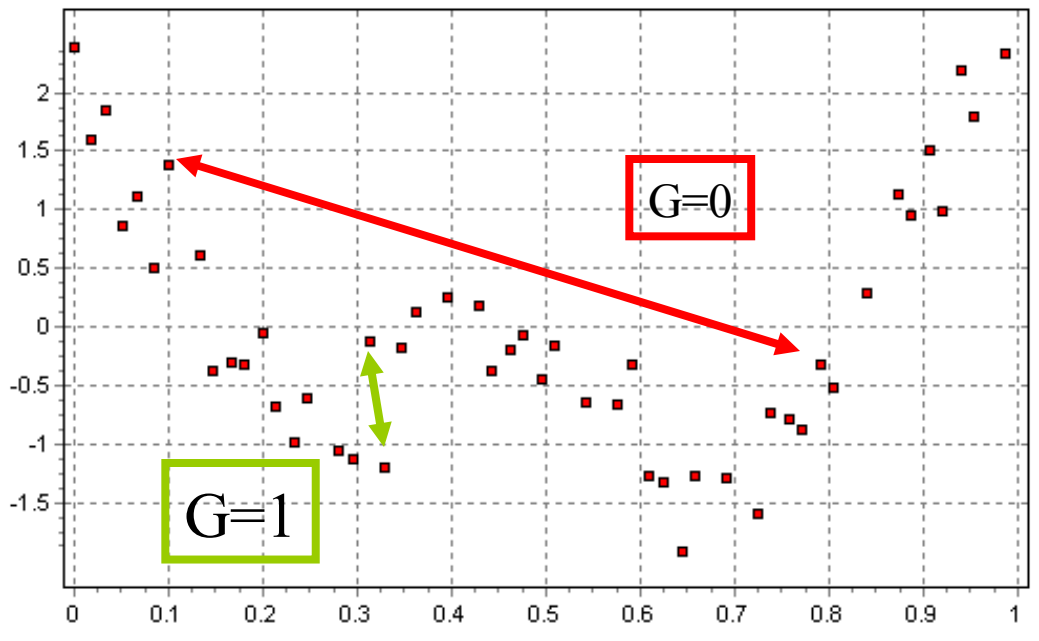
$$t_i, f(t_i), i = 1, 2, \ldots, N$$

$$\varphi_P(t_i) = \mathrm{Frac}(t_i P^{-1})$$

$$D^2(\nu) = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} G(t_i, t_j, \nu) \left[f(t_i) - f(t_j) \right]^2}{2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} G(t_i, t_j, \nu)}$$

Weights G are larger than zero when phases of two points in pair are similar, or:

$$t_i - t_j \approx \frac{k}{\nu}, k = 0, \pm 1, \pm 2, \ldots,$$



Smoother periodograms

$$D^2(\nu) = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} G^*(t_i, t_j, \nu) \left[f(t_i) - f(t_j) \right]^2}{2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} G^*(t_i, t_j, \nu)}$$

$$G^*(t_i, t_j, \nu) = G(t_i, t_j, \nu) W(t_i, t_j)$$

$$W(t_i, t_j) = \begin{cases} 1, & |t_i - t_j| \leq t_{\max} \\ 0, & \text{otherwise,} \end{cases}$$

How to compute?

$$t_i - t_j \approx k\delta t, \text{ for some } k$$

$$C_k = \sum_{t_i - t_j \approx k\delta t} \left[f(t_i) - f(t_j) \right]^2$$

$$D^2(\nu) = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} G(t_i, t_j, \nu) \left[f(t_i) - f(t_j) \right]^2}{2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} G(t_i, t_j, \nu)}$$

$$S_k = \sum_{t_i - t_j \approx k\delta t} 1$$

$$G(t_i, t_j, \nu) = d\big((t_i - t_j)\nu\big) \approx d(k\delta t\nu)$$

$$D^2(\nu) = \frac{\sum_{k=0}^{K} d(k\delta t\nu) C_k}{2 \sum_{k=0}^{K} d(k\delta t\nu) S_k}$$

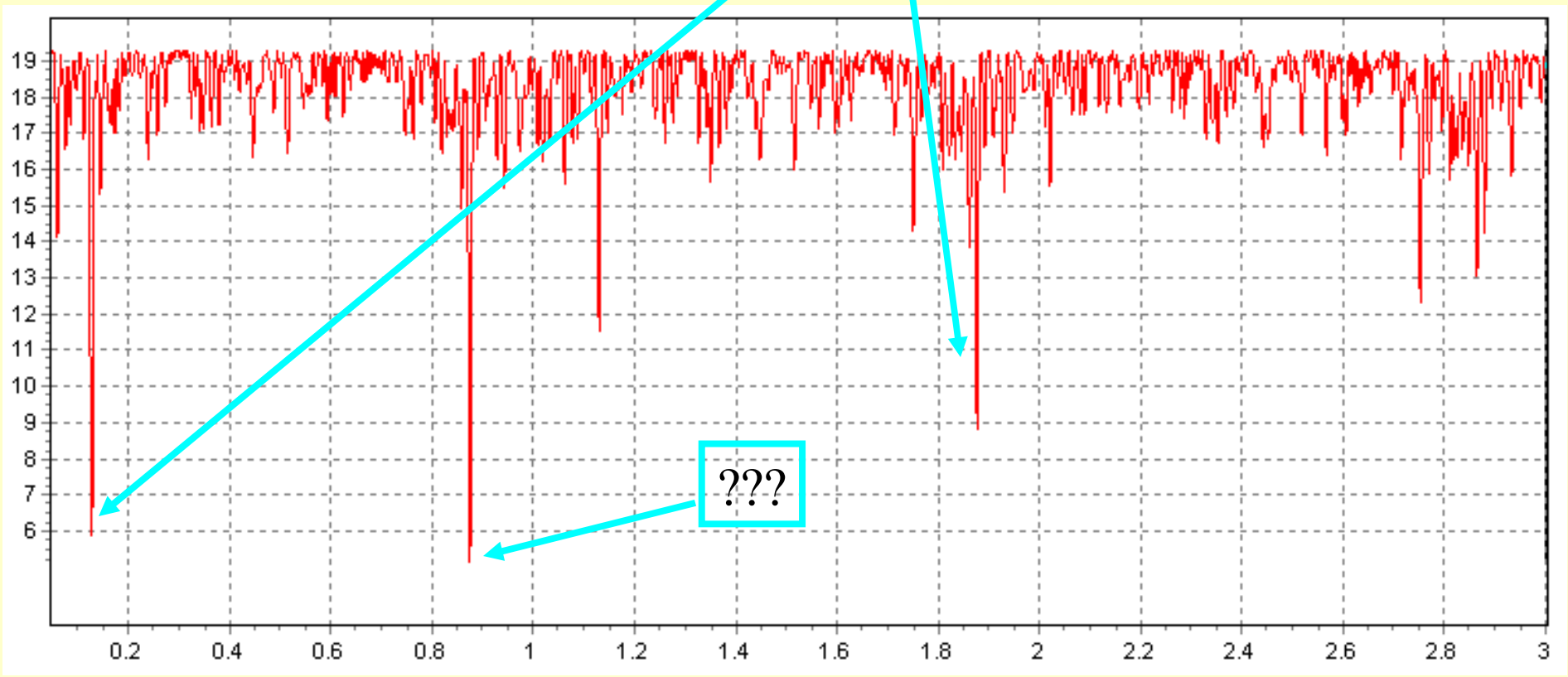
$$d(k\delta t\nu) = \sum_{r=0}^{\infty} d_r \cos\left(2\pi r k \delta t \nu\right)$$

Multiperiodic processes

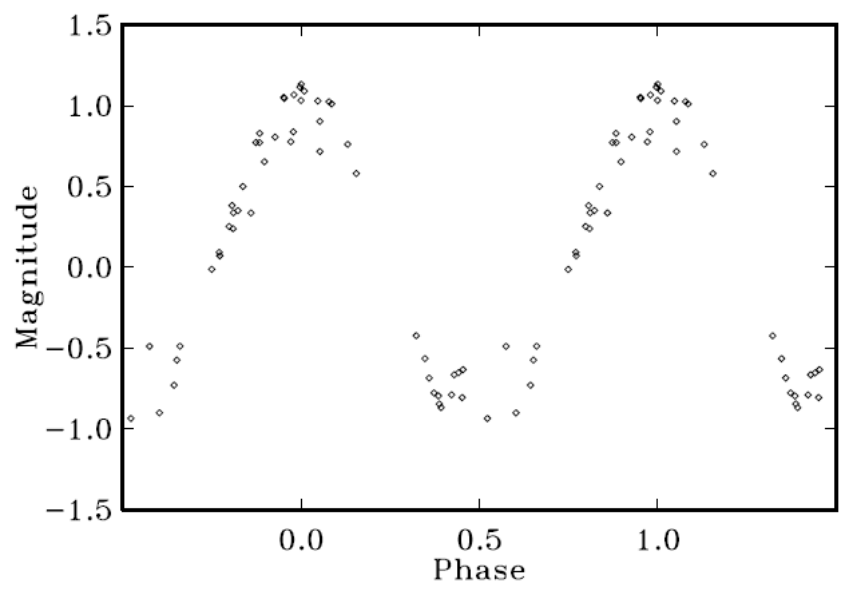
$$\frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} G(t_i,t_j,\nu_1,\nu_2)\left[f(t_i)-f(t_j)\right]^2}{2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} G(t_i,t_j,\nu_1,\nu_2)}$$

An example

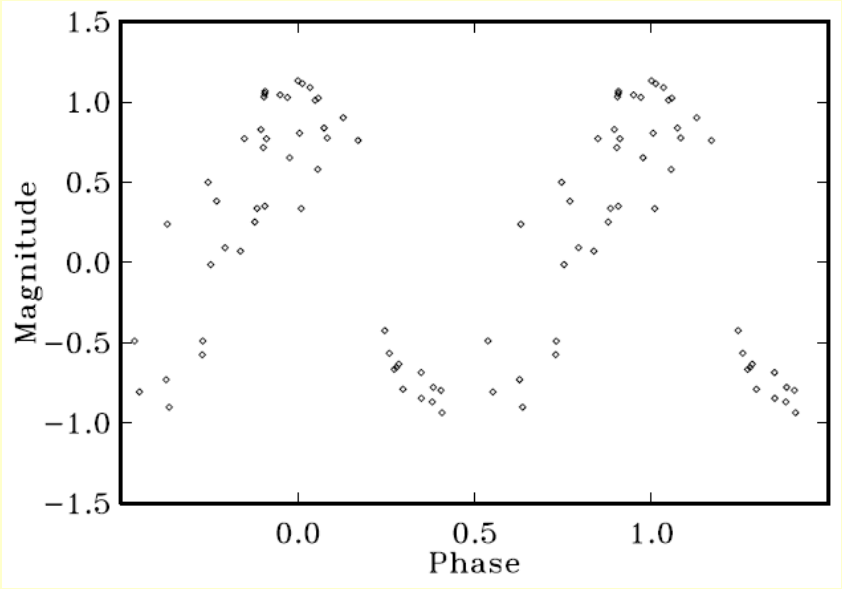
$$f(t) = 0.55073\cos(2\pi t/0.53289 + 0.09111) +$$
$$+0.58717\cos(2\pi t/7.83453 + 0.91564).$$



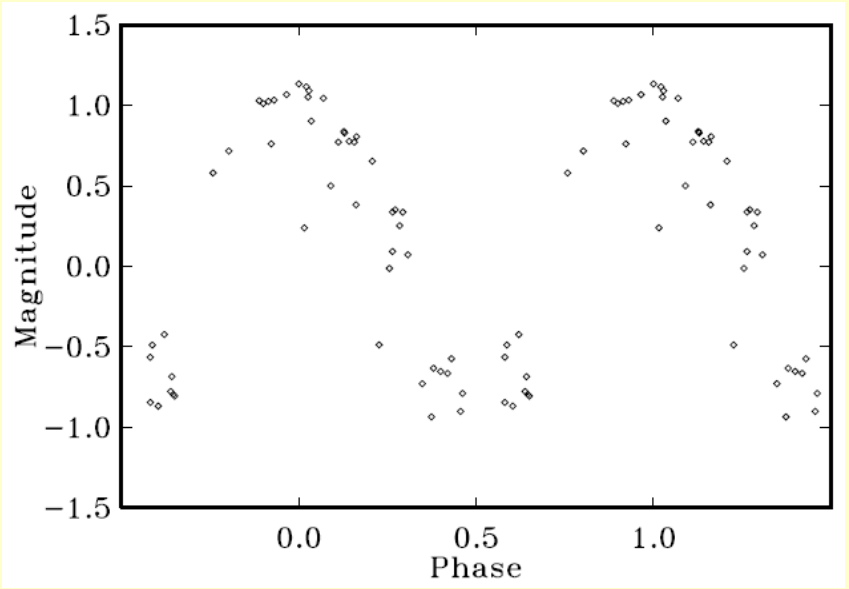
Why?



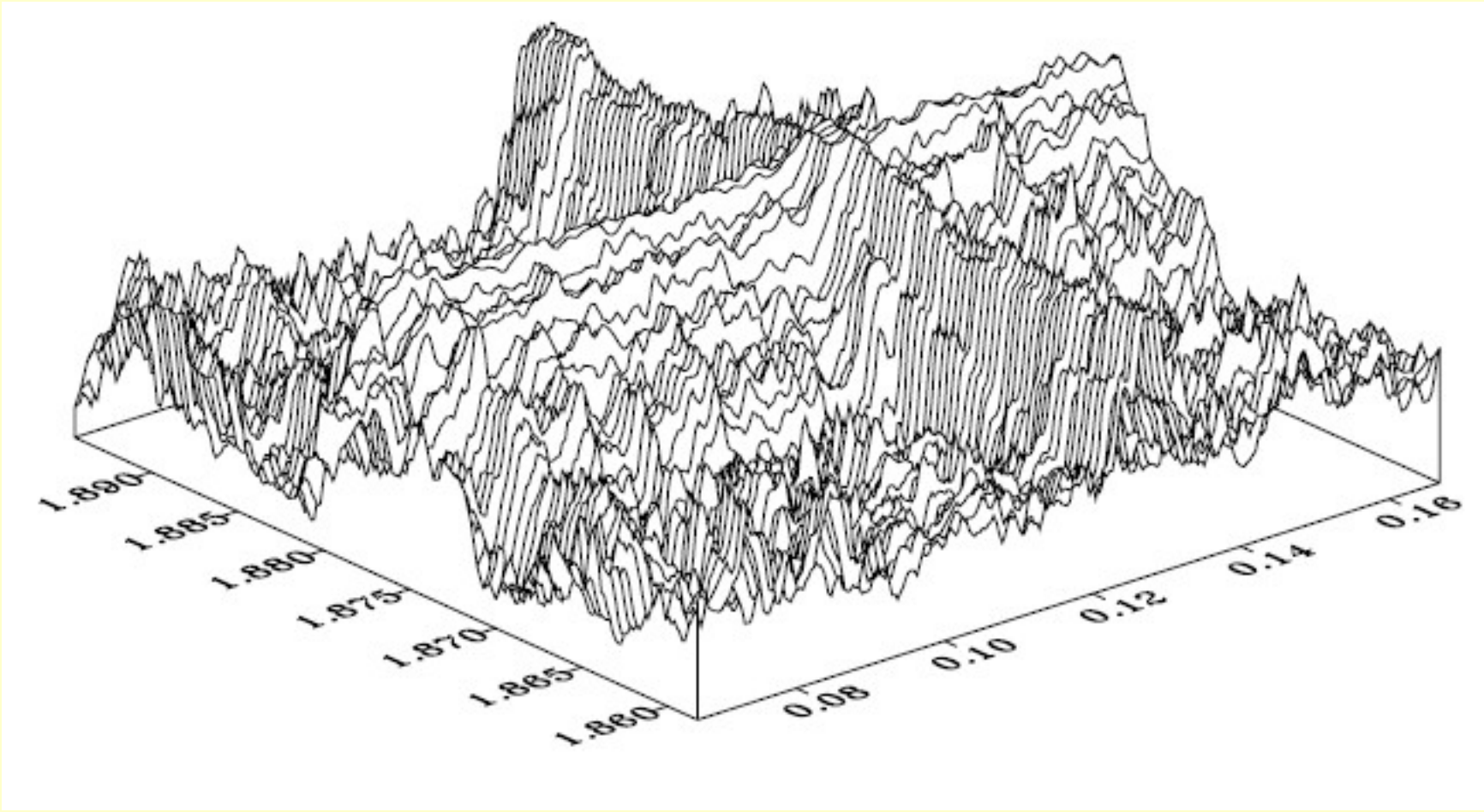
deepest peak in spectrum $P = 1.143455$



first real period $P = 0.5328956$

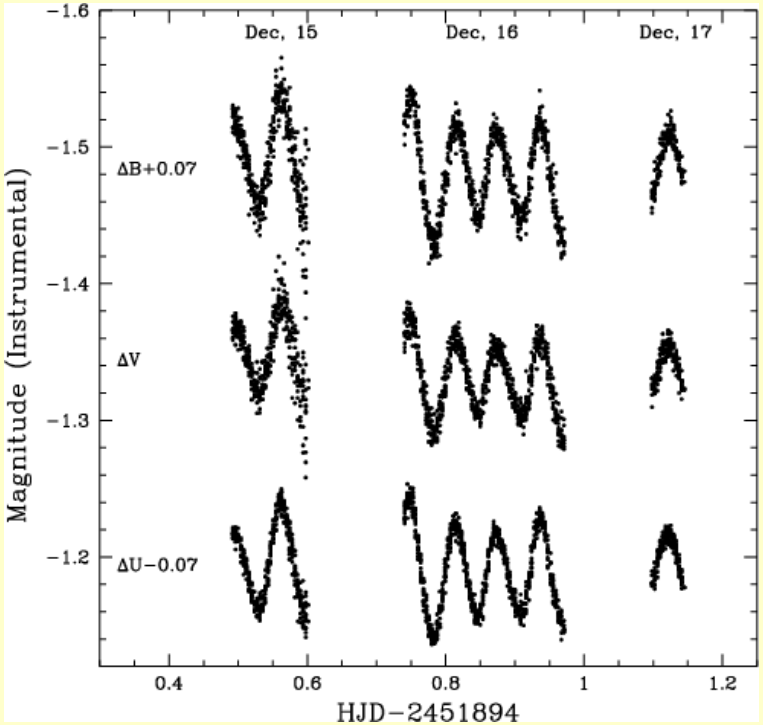


second real period $P = 7.8345346$



Two-dimensional "power spectrum" $1 - D^(\nu_1, \nu_2)$

Multichannel search



$$t_{ij} = |t_i - t_j|,$$

$$w^c_{ij} = \frac{1}{\sigma^c(t_i)^2 + \sigma^c(t_j)^2} = \frac{w^c_i \cdot w^c_j}{w^c_i + w^c_j},$$

$$y^c_{ij} = w^c_{ij} \cdot |y^c_i - y^c_j|^2.$$

Total dispersion

$$D(P) = \frac{\sum_{c=1}^{C} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} g(t_{ij}, P) L(t_{ij}) y_{ij}^{c}}{\sum_{c=1}^{C} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} g(t_{ij}, P) L(t_{ij}) w_{ij}^{c}}.$$

Phases $$\phi_{ij}(P) = \mathrm{Frac}\frac{t_i - t_j}{P}.$$

must be near by

$$g(t_{ij}, P) = \begin{cases} 1, & \phi_{ij}(P) \le \tau \quad \text{or} \\ & \phi_{ij}(P) > 1 - \tau \\ 0, & \text{otherwise.} \end{cases}$$

Resolution can be controlled

$$L(t_{ij}) = \begin{cases} 1, & D_{\min} \le |t_i - t_j| < D_{\max} \\ 0, & \text{otherwise.} \end{cases}$$

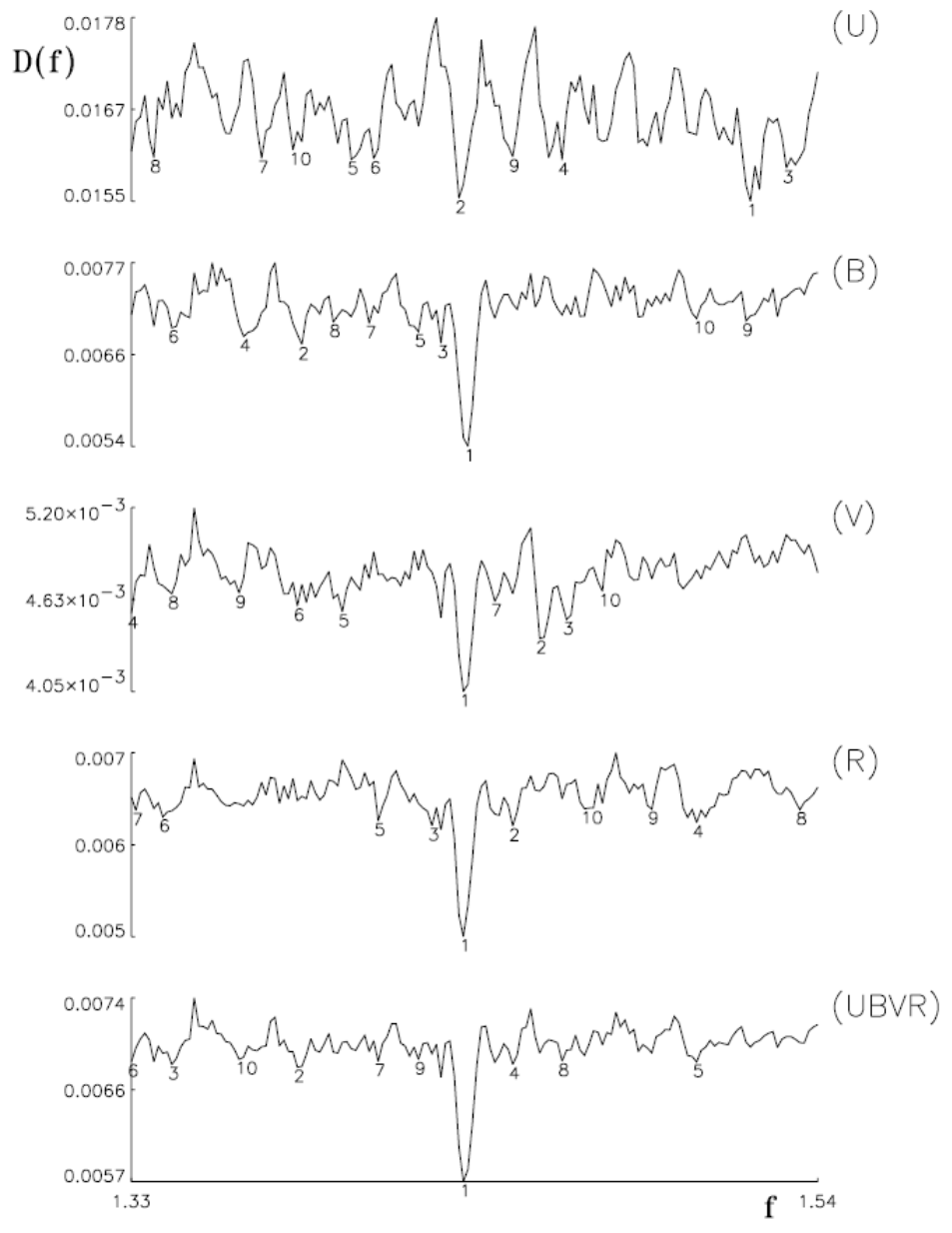
Many channels in LSQ format

Dispersions in separate channels

$$WRSS^c(P) = \sum_{i=1}^{N} w_i^c [y_i^c - M(t_i, \beta^c(P))]^2$$

Total dispersion

$$WRSS(P) = \sum_{c=1}^{C} WRSS^c(P)$$



Time delays

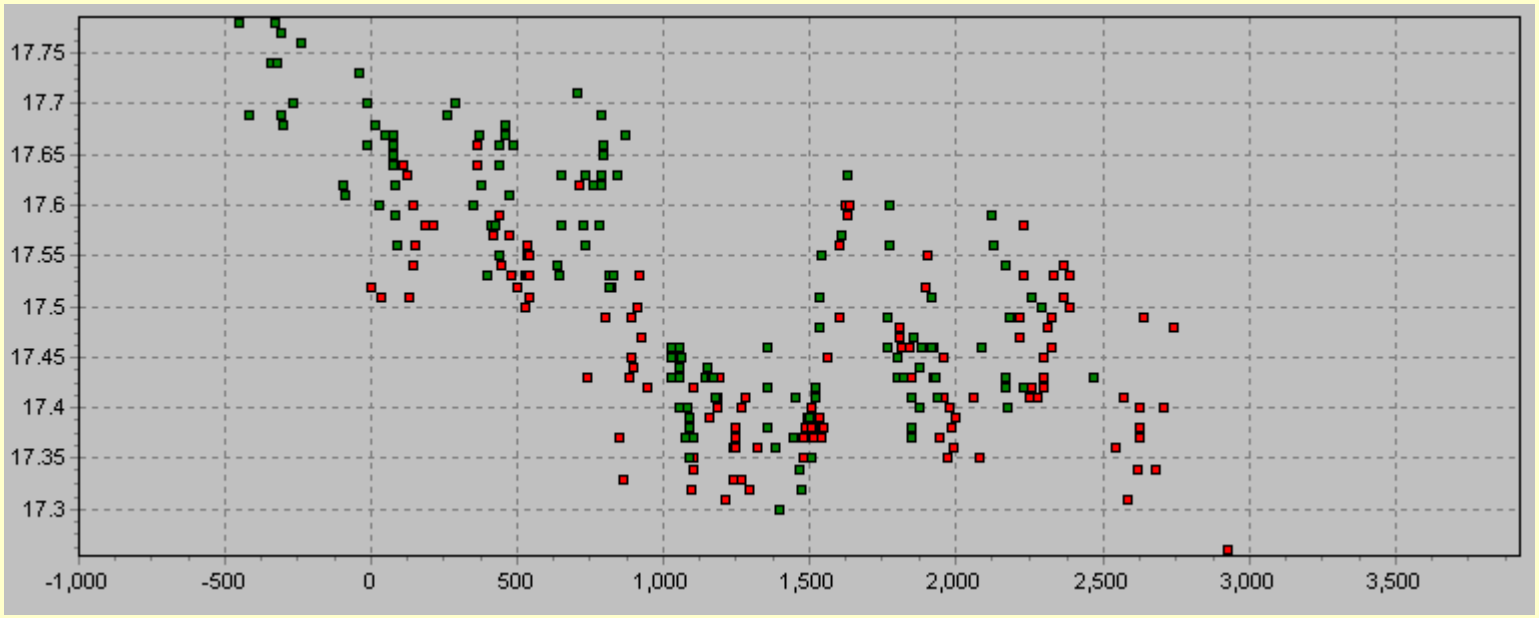
Model for two photometric curves

Time delay

$$
\begin{aligned}
A_i &= g(t_i) + \epsilon_{\mathrm{A}}(t_i), \\
&\quad i = 1, 2, \ldots, N_A, \\
B_j &= \big(g(t_j + \tau) - b\big)/a + \epsilon_{\mathrm{B}}(t_j) \\
&\quad j = 1, 2, \ldots, N_B.
\end{aligned}
$$

Dispersion of the combined curve

$$C_k(t_k) = \begin{cases} A_i, & \text{when } t_k = t_i, \\ aB_j + b, & \text{when } t_k = t_j + \tau \end{cases}$$

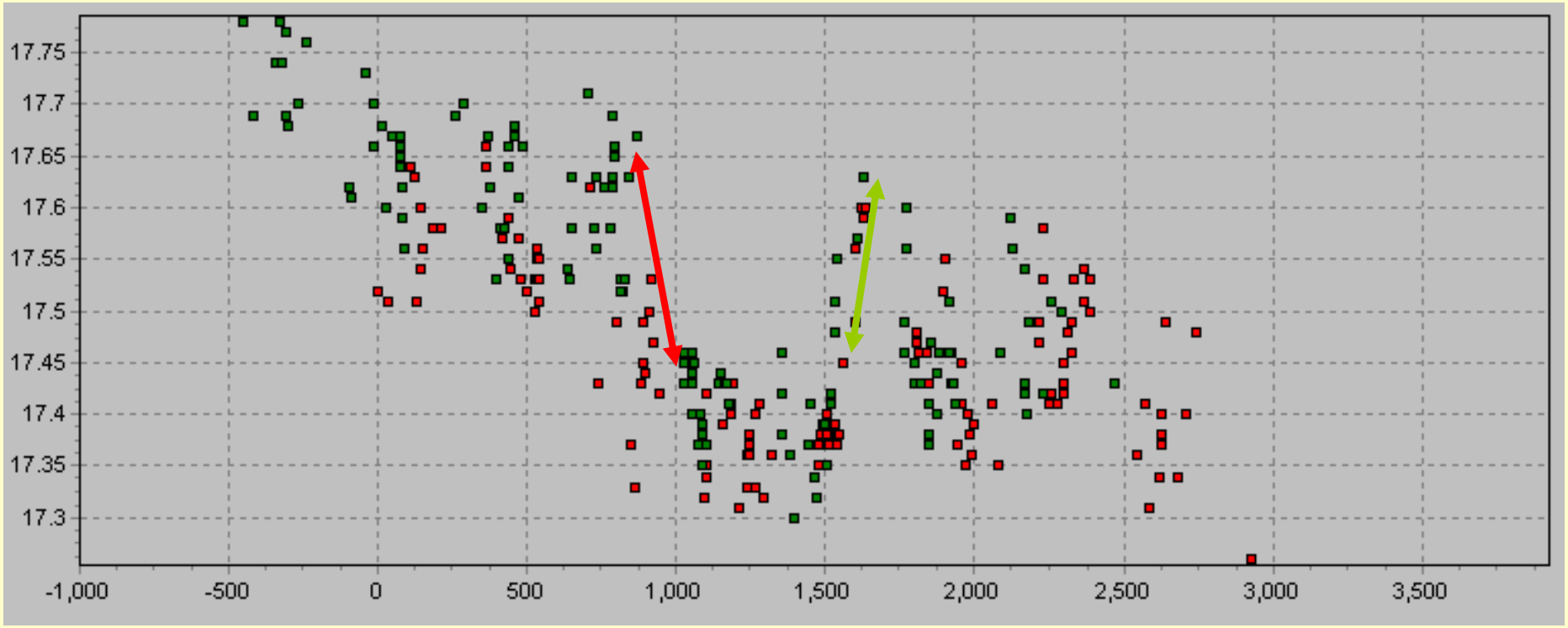


$$D^2(\tau) = \min_{a,b} D^2(\tau, a, b)$$

$$D^2_{\text{all}} = \min_b \frac{\sum_{k=1}^{K-1} W_k (C_{k+1} - C_k)^2}{2 \sum_{k=1}^{K-1} W_k}$$

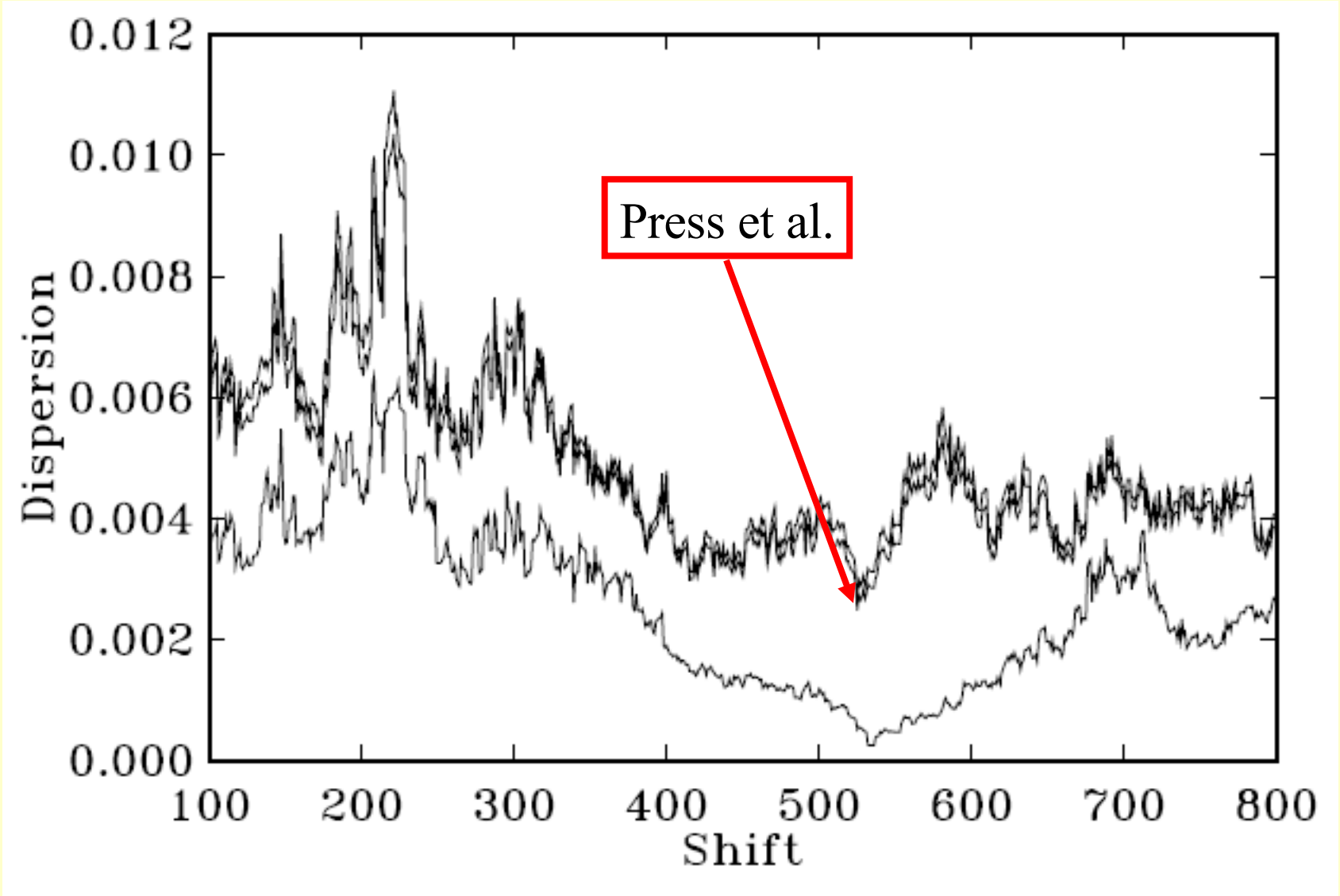
$$W_{i,j} = \frac{W_i W_j}{a^2 W_i + W_j}$$

Weights



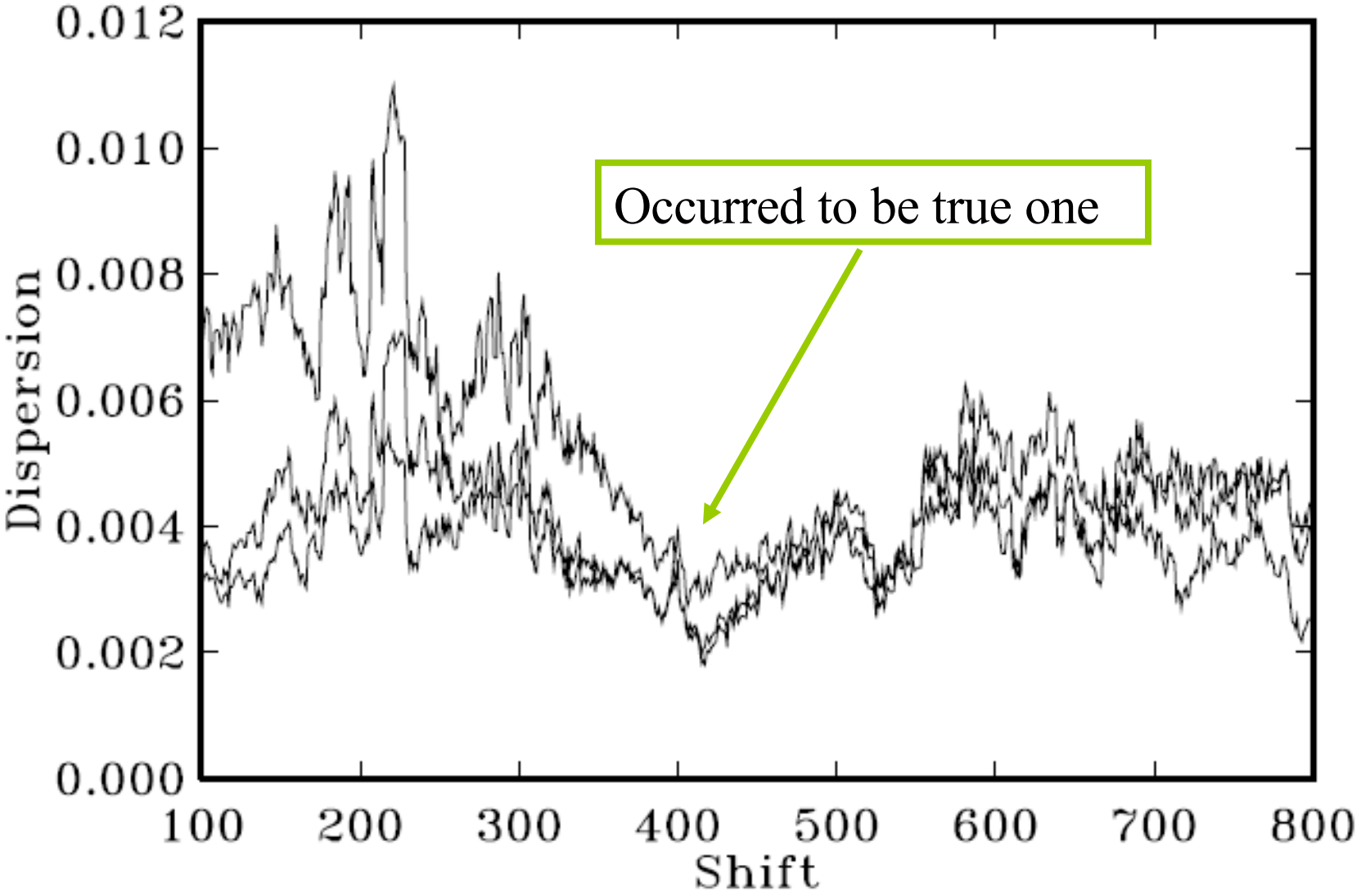
Not all pairs are equal!

$$D^2_{\text{A,B}} = \min_b \frac{\sum_{k=1}^{K-1} W_k G_k (C_{k+1} - C_k)^2}{2 \sum_{k=1}^{K-1} W_k G_k}$$



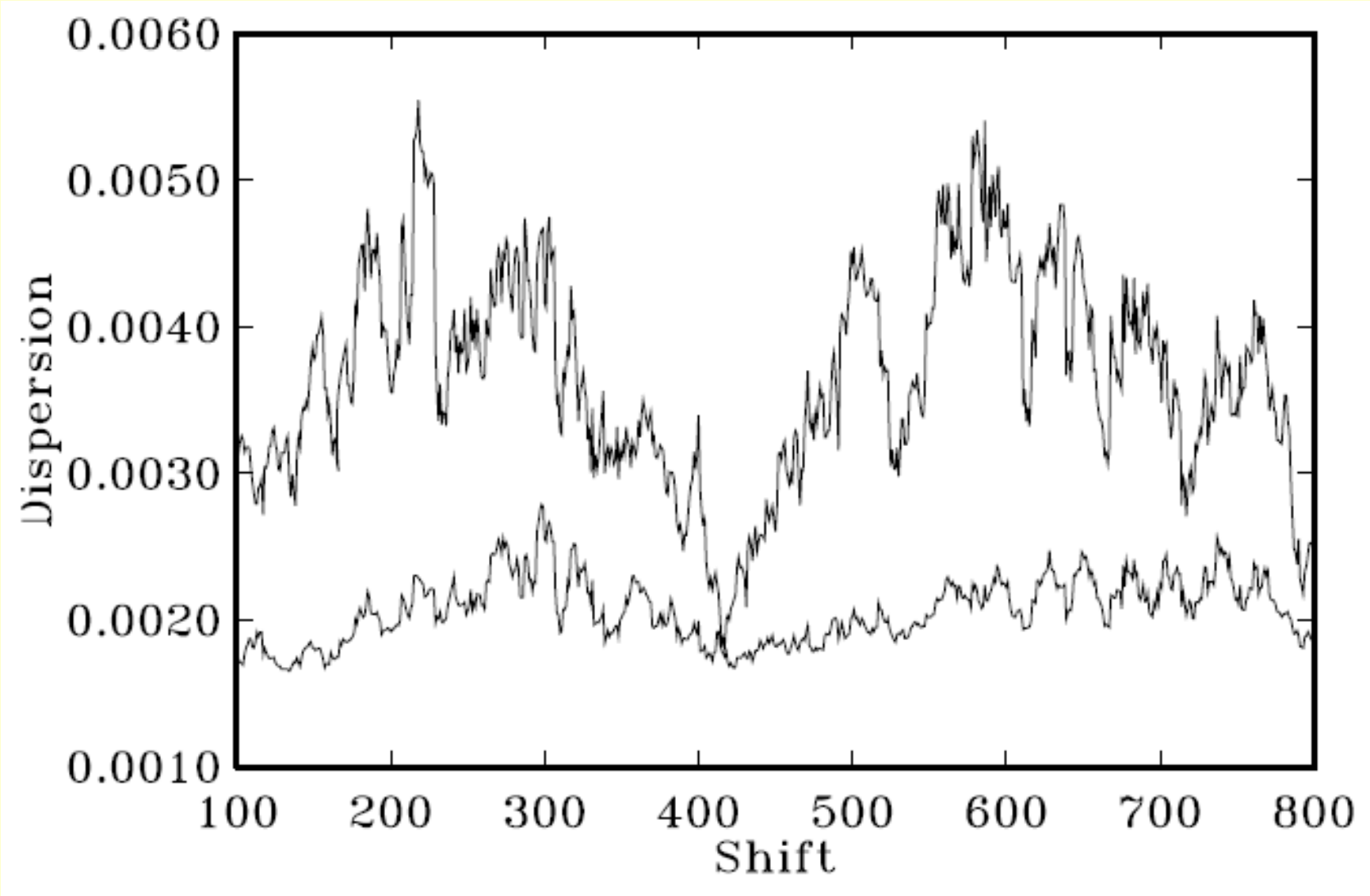
Dispersion spectra for original and model data sets

The lower curve is $D^2_{\mathrm{A,B}}$ for model data set



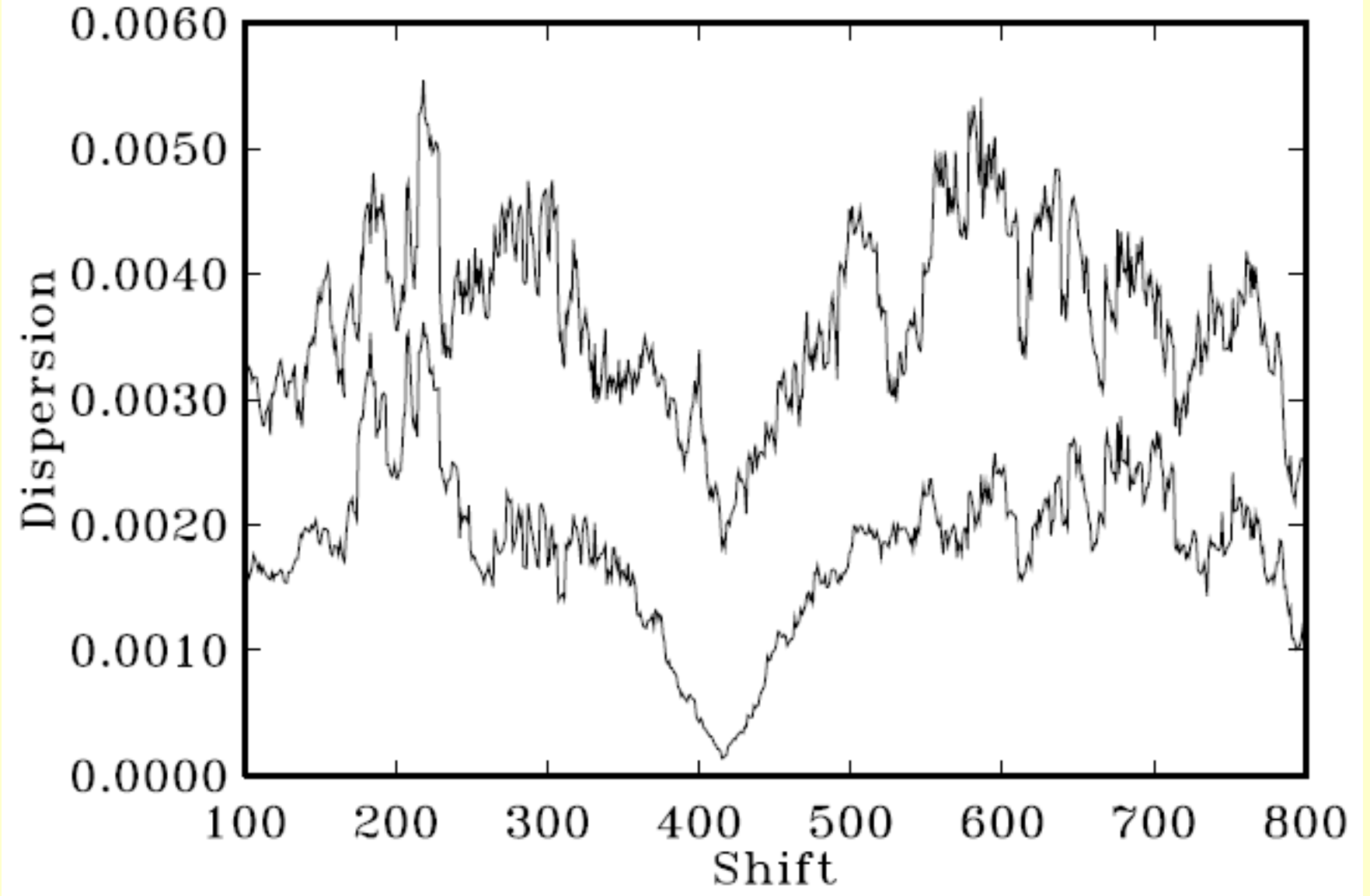
Dispersion spectra for the detrended data sets

Only results with polynomial degrees $1, 5$ and 9

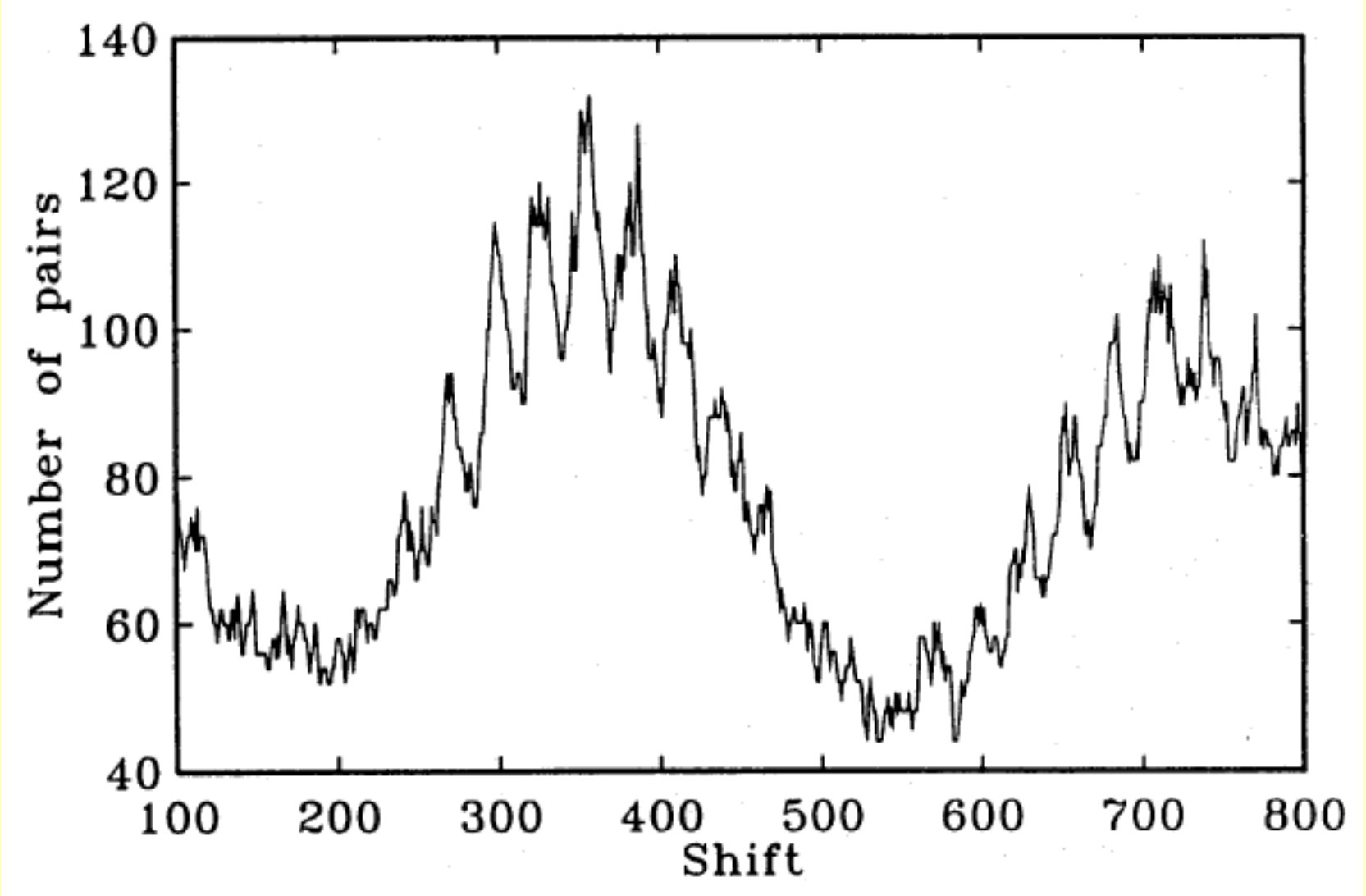


Dispersion spectra for detrended data sets

The upper curve is $D^2_{\mathrm{A,B}}(\tau)$ and the lower curve is $D^2_{\mathrm{all}}(\tau)$



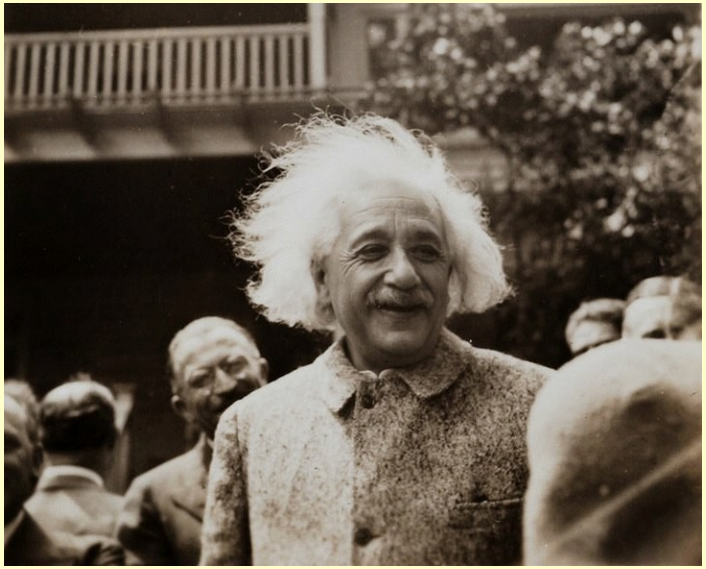
Dispersion spectra for detrended and artificial data sets



Window function for optical data

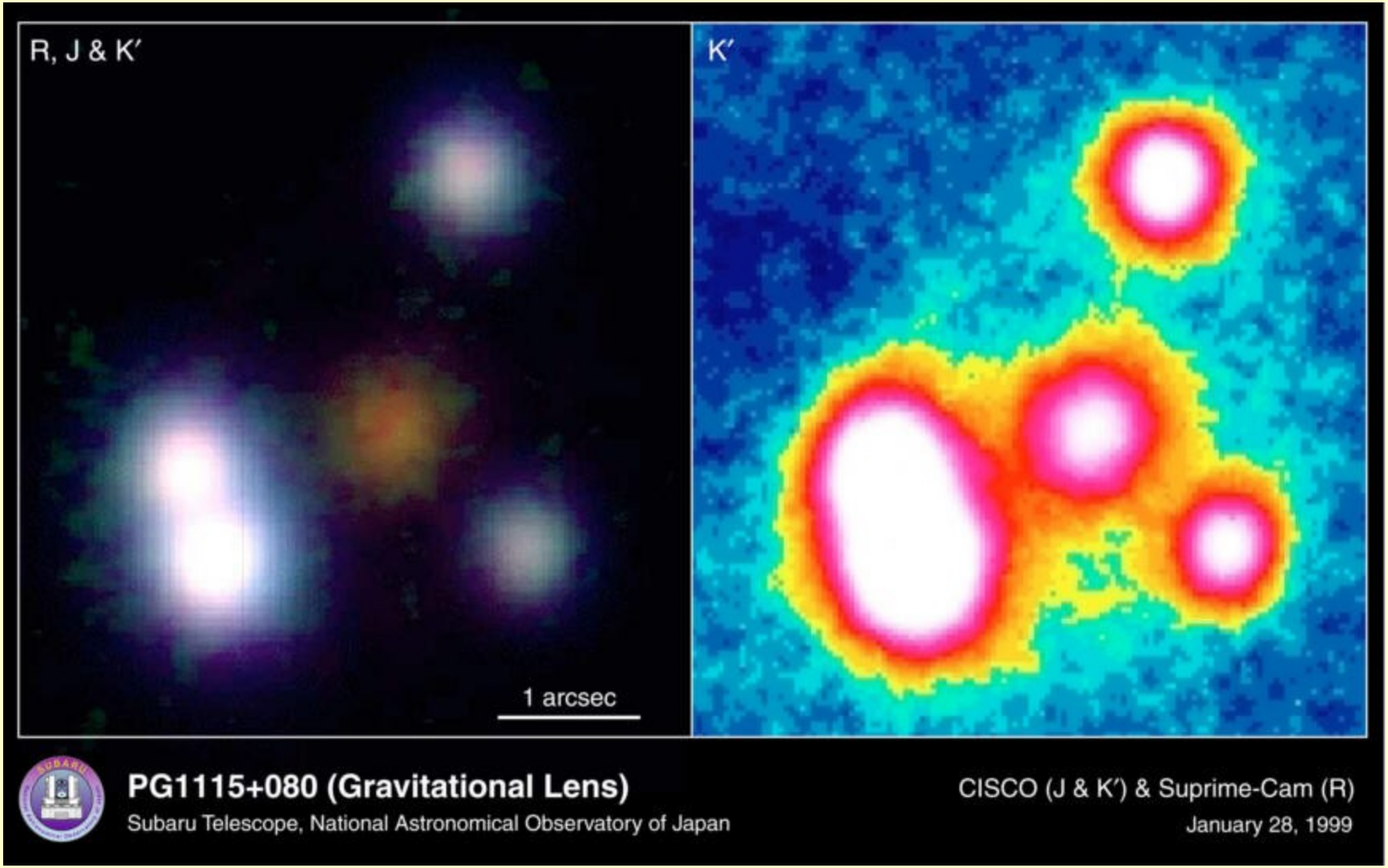
What we learned?

- Look first at data window
- Use only information truly needed
- Check methods on model data
- Do not hope that nature is kind(?)



Subtle is the Lord, but malicious He is not.

Multiple delays



More curves

$$A_i = q(t_i) + \epsilon_A(t_i),\ i = 1, \ldots, N^{(A)}$$

$$\begin{aligned} B_j^{(r)} &= q(t_j - \tau^{(r)}) + l^{(r)}(t_j) + \epsilon^{(r)}(t_j), \\ & \qquad j = 1, \ldots, N^{(r)}, \\ & \qquad r = 1, \ldots, R, \end{aligned}$$

Dispersion of the combined curves

$$C_k(t_k) = \begin{cases} A_i, & \text{if } t_k = t_i, \\ B_j^{(r)} - l_0^{(r)}, & \text{if } t_k = t_j^{(r)} + \tau^{(r)} \end{cases}$$

$$D^2(\tau^{(1)}, \ldots, \tau^{(R)}) = \min_{l_0^{(1)}, \ldots, l_0^{(R)}} D^2(\tau^{(1)}, \ldots, \tau^{(R)}, l_0^{(1)}, \ldots, l_0^{(R)})$$

$$D_1^2(\tau^{(1)},\ldots,\tau^{(R)}) =$$

$$\min_{l_0^{(1)},\ldots,l_0^{(R)}} \frac{\sum_{k=1}^{N-1} W_{k,k+1} G_{k,k+1} (C_{k+1} - C_k)^2}{2 \sum_{k=1}^{N-1} W_{k,k+1} G_{k,k+1}}$$

$$D_2^2(\tau^{(1)},\ldots,\tau^{(R)}) =$$

$$\min_{l_0^{(1)},\ldots,l_0^{(R)}} \frac{\sum_{n=1}^{N-1} \sum_{m=n+1}^{N} W_{n,m} S_{n,m} G_{n,m} (C_n - C_m)^2}{2 \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} W_{n,m} S_{n,m} G_{n,m}}$$

$$S_{n,m} = \begin{cases} 1 - \frac{|t_n - t_m|}{\Delta t}, & \text{if } |t_n - t_m| \le \Delta t, \\ 0, & \text{if } |t_n - t_m| > \Delta t \end{cases}$$

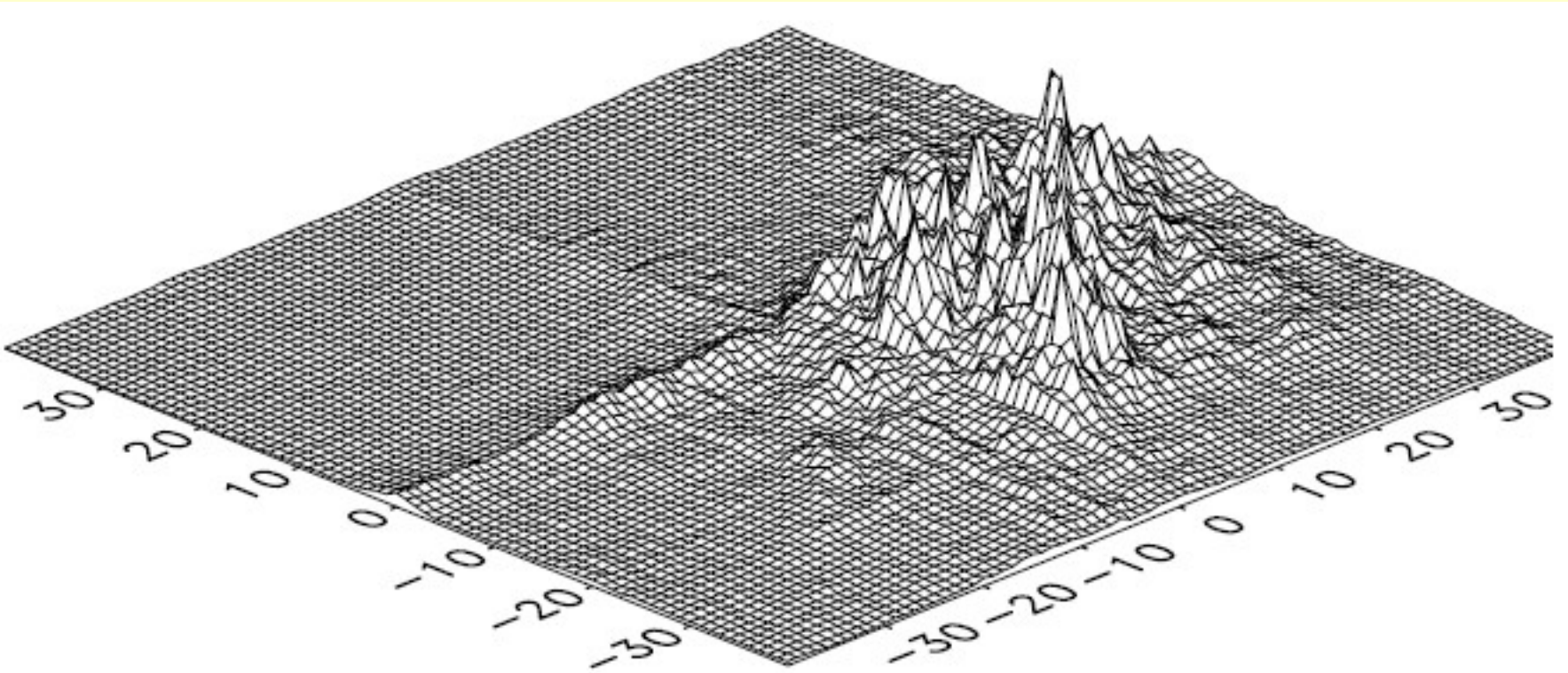


Fig. 1. Contrast enhanced plot $\rho(t_{BA}, t_{CA})$ of the dispersion spectrum D_1^2

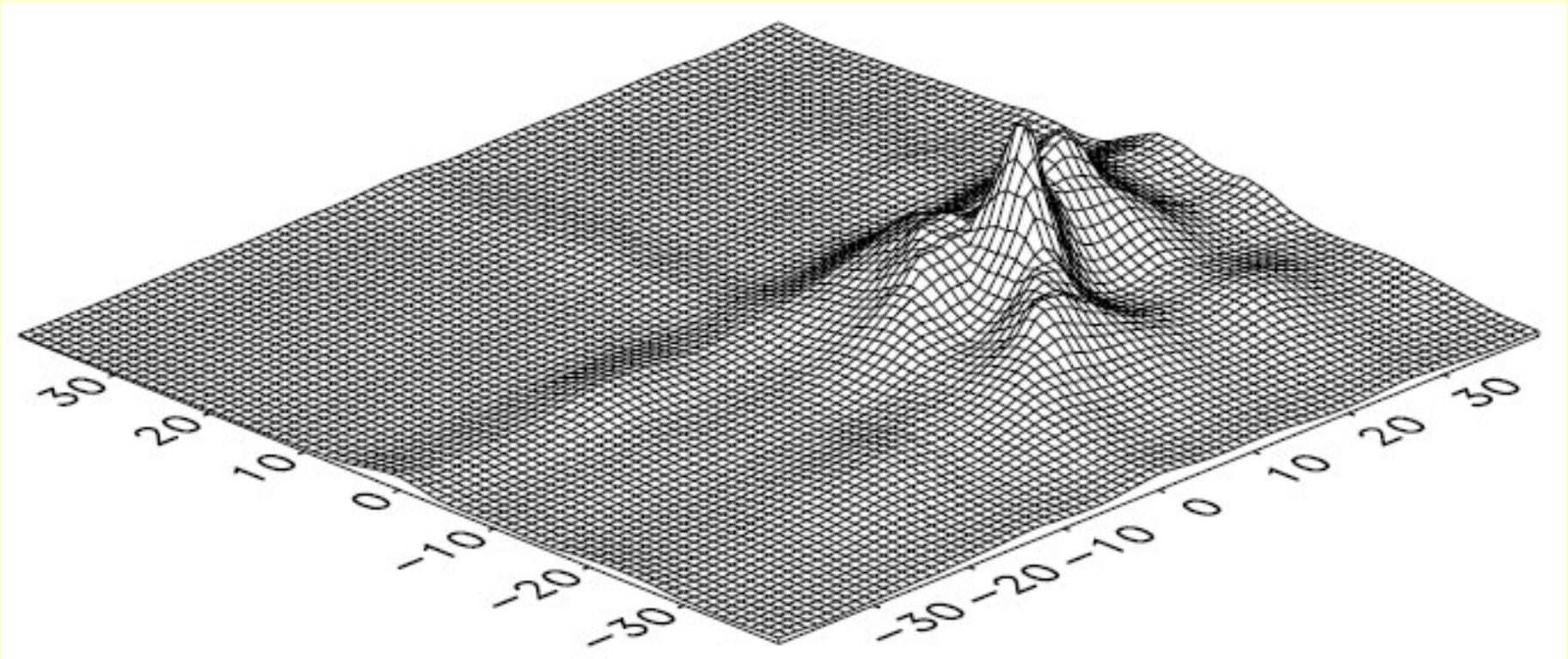


Fig. 2. Contrast enhanced plot $\rho(t_{BA}, t_{CA})$ of the dispersion spectrum D_2^2, $\Delta t = 5.5$

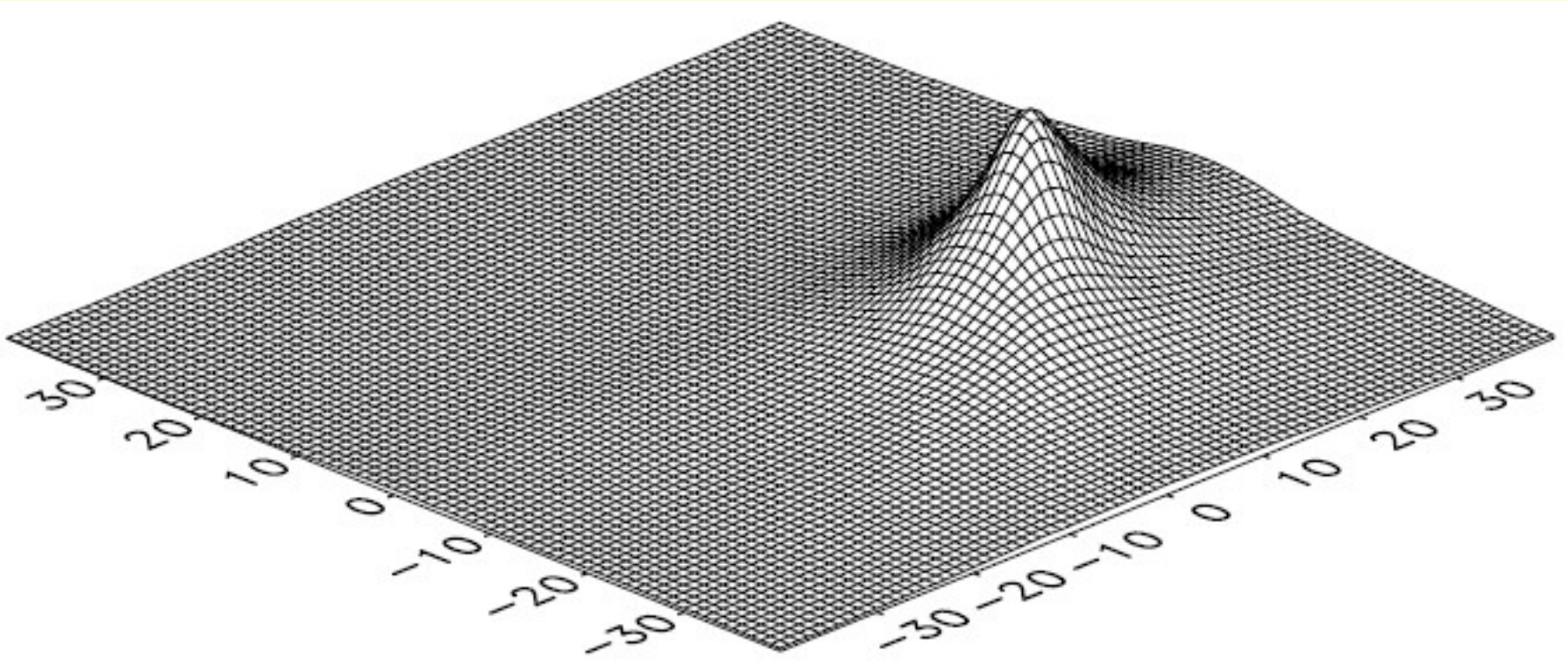


Fig. 3. Contrast enhanced plot $\rho(t_{BA}, t_{CA})$ of the dispersion spectrum D_2^2, $\Delta t = 15.5$