IFA - 93/40 Oct. 1993

AARHUS-ASTRO-1993-12 ISSN 0906-3870

10 p

GENERATION OF ARTIFICIAL HELIOSEISMIC TIME-SERIES

J. Schou and T.M. Brown

Inst. of Physics and Astronomy
Aarhus University
DK-8000 Aarhus C
Denmark

Submitted to: Astronomy and Astrophysics

A&A manuscript no.
(will be inserted by hand later)

Your thesaurus codes are:
03.13.2 06.15.1

ASTRONOMY
AND
ASTROPHYSICS
20.10.1993

Generation of artificial helioseismic time-series

J. Schou[1,2,*] and T. M. Brown[2]

[1] Institut for Fysik og Astronomi, Aarhus Universitet, DK-8000 Århus C, Denmark
[2] High Altitude Observatory, National Center for Atmospheric Research,** Box 3000, Boulder, CO 80307, U.S.A.

[the date of receipt and acceptance should be inserted later]

Abstract. We present an outline of an algorithm to generate artificial helioseismic time-series, taking into account as much as possible of the knowledge we have on solar oscillations. The hope is that it will be possible to find the causes of some of the systematic errors in analysis algorithms by testing them with such artificial time-series.

Key words: methods: data analysis – sun: oscillations

1. Introduction

As more helioseismic datasets have become available it has become apparent that at least some of the analysis algorithms used have systematic errors. This is perhaps most obvious when one compares frequency splittings from different data sets. These splitting show systematic differences, even when the observations were taken at essentially the same time, hence ruling out changes in the true solar rotation rate. Even analysing the same dataset with different methods has yielded small but statistically significant differences (Bachmann et al. 1993).

We hope that it will be possible to find the causes of at least some of the systematic errors by analysing artificial data sets for which the 'true' mode parameters are known. To do this it is necessary that the artificial data closely resemble the real observations. The danger in checking analysis programs this way is obviously that if one overlooks a crucial property of the real data, one may be led to believe that the analysis program performs well on the real data when, in reality, it is flawed. On the other hand, it might be argued that if one cannot successfully analyse artificial data, for which, in principle, one knows all the properties, there is not much chance that real data can be reduced without introducing systematic errors.

In the following we will start with a short summary of the relevant properties of solar oscillations and how the oscillations are observed. Thereafter we will go into more detail about how one can construct artificial time-series given this information. In a separate paper we will discuss some of the results obtained by analysing artificial data from this program using a number of different methods.

Send offprint requests to: J. Schou
* Present address: Center for Space Science and Astrophysics, Stanford University ERL 328, Stanford, CA 94305, USA
** The National Center for Atmospheric Research is sponsored by the National Science Foundation

2. Solar oscillations

The basic physical properties of solar oscillations and the techniques used for observing them have been described in a number of reviews (Christensen-Dalsgaard & Berthomieu 1991 and Hill et al. 1991). Here only the properties important for the construction of artificial data will be described.

Individual modes are generally described by their radial order n, degree l and azimuthal order m. For the purpose of observing the modes and hence for constructing artificial data, the radial order n is only important for determining the frequency ν_{nlm} of the mode. l and m on the other hand determine the appearance of the mode on the solar surface and are thus more important when reducing observations or generating artificial data. The radial component of the velocity (or the intensity) on the solar surface from a mode with a given (n, l, m) is given by

$$V_{n,l,m}(\phi,\theta,t) = \mathrm{Re}\left[a_{nlm}(t)Y_l^m(\phi,\theta)\right], \qquad (1)$$

where Re[ ] denotes the real part of a complex number (Im[ ] similarly denotes the imaginary part), and Y_l^m is a spherical harmonic given by

$$Y_l^m(\phi,\theta) = P_l^{|m|}(\theta)\mathrm{e}^{\mathrm{i}m\phi}\,. \qquad (2)$$

The coordinates ϕ and θ are longitude (For the purpose of analysing solar oscillations data, the zero point of longitude is usually placed at the Sun's sub-Earth meridian.) and colatitude respectively, and $a(t)$ is a time-series describing the (complex) mode amplitude. Notice that the definition of the Y_l^m's is not the most commonly used one, in that the sign is the opposite from the standard definition for odd negative m's. This has no effect on the time-series generated, as the phases of the oscillations on the Sun are random. The reason for this sign convention is historical.

The total surface velocity V is given by a sum over all modes of the individual mode velocities

$$V(\phi,\theta,t) = \sum_{n,l,m} V_{n,l,m}(\phi,\theta,t)\,. \qquad (3)$$

The time-series $a(t)$ is that of a stochastically excited damped oscillator and hence the real and imaginary parts of the Fourier transform $\tilde{a}(\nu)$ of the time-series $a(t)$ each have

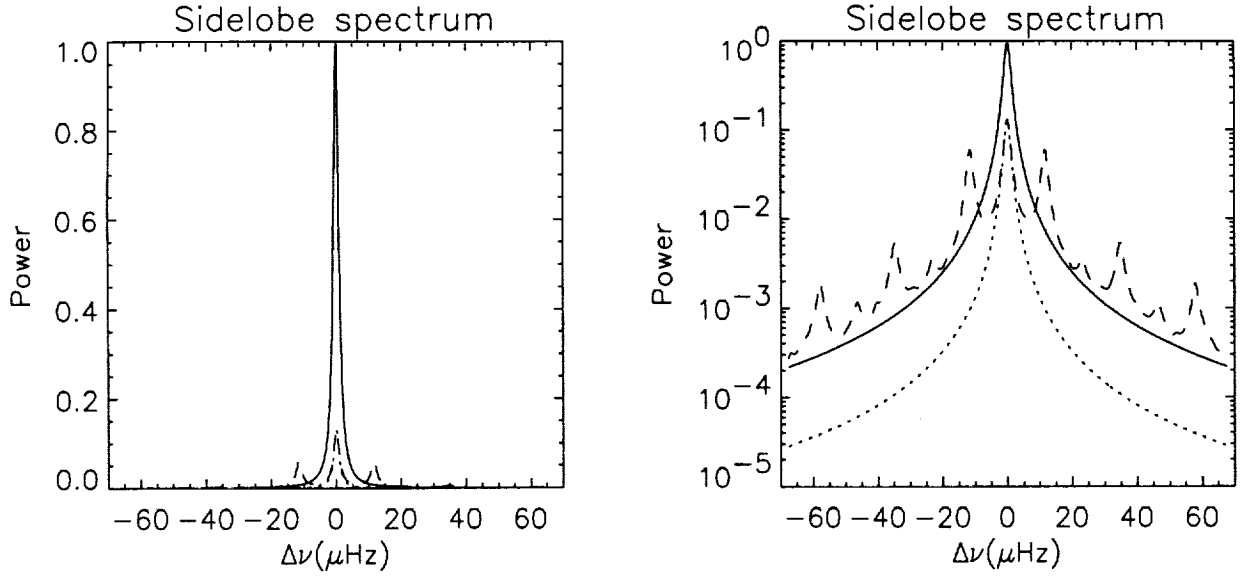

Fig. 1. The average theoretical power spectrum of a mode with a HWHM $w = 1\mu$Hz with (dashed lines) and without (solid lines) time-gaps. The left panel uses a linear scale, while the right panel uses a logarithmic scale. The dotted line shows the ungapped power scaled such that the peak power is the same as for the gapped power. The visibility function used was from a run with the Fourier Tachometer (which was run jointly by HAO and the National Solar Observatories) from the spring of 1989, and should thus represent real one site data well. The duty cycle was 34.94%

zero mean and a variance given by a Lorenzian profile in frequency

$$\mathrm{Var}(\nu) = \frac{P/w}{1 + \left(\frac{\nu-\nu_0}{w}\right)^2}, \tag{4}$$

where P is a measure of the average mode power, w is the half width at half maximum (HWHM) of the mode, and ν_0 is the mode frequency. If the mode is excited very often during its lifetime $\tau = (2\pi w)^{-1}$, the values of the discrete Fourier transform at different frequencies close to the central peak are independent. If the mode is only excited infrequently, the individual points in the Fourier transform are not independent over a frequency range of the order $t^{-1}_{\mathrm{excitation}}$, where $t_{\mathrm{excitation}}$ is the typical time between excitations.

The oscillations are observed by taking images of either the intensity or the Doppler velocity at the solar surface (see Hill et al. 1991) at regular intervals of typically 1 minute. The images are then usually interpolated to a uniform grid in ϕ and $x = \cos\theta$ and inner products with suitable masks are calculated to isolate the different target modes

$$o_{l,m}(t) = \int_{-1}^{1}\int_{-\pi/2}^{\pi/2} V_{\mathrm{obs}}(\phi, x, t) M_l^m(\phi, x)\mathrm{d}\phi\mathrm{d}x\,, \tag{5}$$

where $o_{l,m}$ is the observed time-series for a target (l,m), V_{obs} the observed surface velocity and M_l^m the mask used to isolate the (l,m). The reason for the interpolation to a fixed net in latitude and longitude is that it allows one to compensate for the varying B-angle (which is the angle between the solar rotation axis and the plane of the sky) and effective P-angle (which is the angle on the images between the solar rotation axis and a reference direction) without changing the masks as a function of time. A uniform net in longitude also allows one to use a Fast Fourier Transform to do the longitude integrals, substantially decreasing the computational burden. Notice that since modes with $+m$ and $-m$, but the same l, look identical at any given time, it is not possible to separate them at this point in the processing; thus only masks with $m \geq 0$ are used (see later). Also notice that all the n's at a given (l,m) appear in the corresponding time-series.

For reasons explained later we have chosen to use masks given by

$$M_l^m(\phi, x) = \frac{1}{\pi} Y_l^m(\phi, x)\mathrm{Ap}(r), \tag{6}$$

where $\mathrm{Ap}(r)$ is an apodization chosen to reduce the contribution from the noise close to the solar limb and $r = (\cos^2\theta + \sin^2\phi\sin^2\theta)^{\frac{1}{2}}$ is the fractional radius of the given point on the observed image of the Sun. The apodization function Ap, is typically 1 inside a certain radius and monotonically decreasing outside.

The time-series of each of the target (l,m)'s are then Fourier transformed and the complex conjugate of the negative frequency part of the spectrum is identified with $+m$, while the positive frequency part is identified with $-m$. Due to bad weather and other problems (such as the Sun being below the horizon) the Sun is normally not observed uninterrupted. It is thus necessary to zero-fill the time-series, which, unfortunately, leads to temporal sidelobes in the Fourier transforms. Recall that the Fourier transform of a product of two functions (the time-series without gaps and the gaps represented as a visibility function with, say, 1 when there is data and 0 when there is not) is the convolution of the Fourier transforms of the two functions. Since the visibility function tends to be highly periodic (at least at moderate latitudes) due to the presence of the day/night cycle, its Fourier transform (and thereby power spectrum) contains peaks at multiples of $1\mathrm{day}^{-1} \approx 11.57\mu$Hz. As the values of the Fourier transform of the uninterrupted time-series at the different frequency points are independent, the average power spectrum of the gapped time-series is the

convolution of the average power spectrum of the ungapped time-series (which is twice the variance in Eq. (4)) convolved with the power spectrum of the window function. The effects of the gaps are thus, as illustrated in Fig. 1, to introduce so-called temporal sidelobes in the power spectra and to introduce correlations among the previously independent points in the Fourier transform. Also the power level far from the peak is considerably higher relative to the peak when time-gaps are present.

Since Y_l^m's are not orthogonal on the part of the Sun we observe, individual (l,m)'s are not perfectly separated by the inner product operation described above. For Doppler velocity observations, only the velocity in the direction of the observer is detected and hence the effective area of the Sun observed is even less than a hemisphere. The observed velocity signal from a single mode is

$$\begin{aligned} V_{n,l,m,\mathrm{obs}}(\phi,x,t) &= \sqrt{1-r^2}V_{n,l,m}(\phi,\theta,t) \\ &= \sqrt{1-r^2}\mathrm{Re}\left[a_{nlm}(t)Y_l^m(\phi,\theta)\right] \\ &= \sqrt{1-r^2}P_l^m(x)\mathrm{Re}\left[a_{nlm}(t)\mathrm{e}^{\mathrm{i}m\phi}\right] , \end{aligned} \tag{7}$$

where $\sqrt{1-r^2}$ is the line of sight velocity projection factor. The observed time-series for a target (l,m) is now given by a sum over all modes on the Sun weighted by coefficients $c_{l,m,l',m'}$ describing the sensitivities to modes characterized by (l',m'), when the target mode is characterized by (l,m):

$$\begin{aligned} o_{l,m}(t) &= \int_{-1}^{1}\int_{-\pi/2}^{\pi/2} V_{\mathrm{obs}}(\phi,x,t)M_l^m(\phi,x)\mathrm{d}\phi\mathrm{d}x \\ &= \int_{-1}^{1}\int_{-\pi/2}^{\pi/2} \sum_{n',l',m'} V_{n',l',m'}(\phi,x,t)M_l^m(\phi,x)\mathrm{d}\phi\mathrm{d}x \\ &= \sum_{n',l',m'}\int_{-1}^{1}\int_{-\pi/2}^{\pi/2}\Big\{ P_{l'}^{m'}(x)\mathrm{Re}(a_{n'l'm'}(t)\mathrm{e}^{\mathrm{i}m'\phi}) \\ &\quad \sqrt{1-r^2}\mathrm{Ap}(r)\frac{1}{\pi}Y_l^m(\phi,x)\Big\}\mathrm{d}\phi\mathrm{d}x \\ &= \sum_{n',l',m'}\int_{-1}^{1}\int_{-\pi/2}^{\pi/2}\Big\{ \frac{1}{\pi}P_{l'}^{m'}(x)P_l^m(x) \\ &\quad \left[\mathrm{Re}(a_{n'l'm'}(t))\cos(m'\phi) - \mathrm{Im}(a_{n'l'm'}(t))\sin(m'\phi)\right] \\ &\quad \left[\cos(m\phi)+\mathrm{i}\sin(m\phi)\right]\sqrt{1-r^2}\mathrm{Ap}(r)\Big\}\mathrm{d}\phi\mathrm{d}x \\ &= \sum_{n',l',m'}\int_{-1}^{1}\int_{-\pi/2}^{\pi/2}\Big\{ P_l^m(x)P_{l'}^{m'}(x)\mathrm{Ap}(r)\sqrt{1-r^2} \\ &\quad \left[\mathrm{Re}(a_{n'l'm'}(t))\cos(m\phi)\cos(m'\phi)\right. \\ &\quad \left. - \mathrm{Im}(a_{n'l'm'}(t))\sin(m\phi)\sin(m'\phi)\right]\Big\}\frac{1}{\pi}\mathrm{d}\phi\mathrm{d}x \\ &= \sum_{n',l',m'}\Big\{ c_{l,m,l',m'}\mathrm{Re}(a_{n',l',m'}(t)) \\ &\quad - \mathrm{i}c'_{l,m,l',m'}\mathrm{Im}(a_{n',l',m'}(t))\Big\} . \end{aligned} \tag{8}$$

In the last equality, the sensitivity coefficients $c_{l,m,l',m'}$ and $c'_{l,m,l',m'}$ are defined as

$$\begin{aligned} c_{l,m,l',m'} &= \frac{1}{\pi}\int_{-1}^{1}\int_{-\pi/2}^{\pi/2}\Big\{ P_l^m(x)P_{l'}^{m'}(x) \\ &\quad \cos(m\phi)\cos(m'\phi)\mathrm{Ap}(r)\sqrt{1-r^2}\Big\}\mathrm{d}\phi\mathrm{d}x \\ &= \frac{4}{\pi}\delta_{l,m,l',m'}\int_0^{\pi/2}\int_0^{\pi/2}\Big\{ P_l^m(x)P_{l'}^{m'}(x) \\ &\quad \cos(m\phi)\cos(m'\phi)\mathrm{Ap}(r)\sin(\theta)\sqrt{1-r^2}\Big\}\mathrm{d}\phi\mathrm{d}\theta , \end{aligned} \tag{9}$$

$$\begin{aligned} c'_{l,m,l',m'} &= \frac{1}{\pi}\int_{-1}^{1}\int_{-\pi/2}^{\pi/2}\Big\{ P_l^m(x)P_{l'}^{m'}(x) \\ &\quad \sin(m\phi)\sin(m'\phi)\mathrm{Ap}(r)\sqrt{1-r^2}\Big\}\mathrm{d}\phi\mathrm{d}x \\ &= \frac{4}{\pi}\delta_{l,m,l',m'}\int_0^{\pi/2}\int_0^{\pi/2}\Big\{ P_l^m(x)P_{l'}^{m'}(x) \\ &\quad \sin(m\phi)\sin(m'\phi)\mathrm{Ap}(r)\sin(\theta)\sqrt{1-r^2}\Big\}\mathrm{d}\phi\mathrm{d}\theta \end{aligned} \tag{10}$$

and

$$\delta_{l,m,l',m'} = \begin{cases} 1 & \text{if } l+m+l'+m' \text{ is even.} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Notice that the tangential component of the surface velocity has been neglected in these calculations. For some modes (eg. g-modes) it may be necessary to include it if a high accuracy is needed.

The c's obviously satisfy some symmetry relations:

$$\begin{aligned} c_{l,m,l',m'} &= c_{l',m',l,m} \\ c'_{l,m,l',m'} &= c'_{l',m',l,m} \\ c_{l,-m,l',m'} &= c_{l,m,l',m'} \\ c'_{l,-m,l',m'} &= -c'_{l,m,l',m'} . \end{aligned} \tag{12}$$

For $m \gg 1$

$$c'_{l,m,l',m'} \approx \mathrm{sign}(mm')c_{l,m,l',m'} . \tag{13}$$

Note that these relations for the c's are only true if the geometry of the images is correctly understood. If there are scale errors, orientation errors, centering errors, or distortions they may not hold. The fact that the image is sampled on a fairly coarse grid may also introduce inaccuracies. On the other hand, they do hold for many other types of masks, as long as these have the same symmetric/antisymmetric properties around the equator and the central meridian as the spherical harmonic masks used here.

To find the corresponding crosstalks in the Fourier transforms, suppose a mode (l',m') has $a(t) = a\mathrm{e}^{\mathrm{i}\omega t}$ (in other words looking at a single frequency point in the Fourier transform). As previously mentioned, the part of the Fourier transform of the observed time-series o used for a given m is the positive frequency part $|m|$ if m is negative and the conjugate of the negative frequency part if m is non negative. The crosstalk to another mode (l,m) can now be found by noting that the contribution from the mode (l',m') is given by

$$o_{l,m}(t) = c_{l,m,l',m'}a\cos(\omega t) - \mathrm{i}c'_{l,m,l',m'}a\sin(\omega t) \tag{14}$$

and that

$$o_{l,m}(t) = b_-\mathrm{e}^{-\mathrm{i}\omega t} + b_+\mathrm{e}^{\mathrm{i}\omega t}, \tag{15}$$

where b_- is the amplitude in the negative frequency part of the Fourier transform and b_+ is the amplitude in the positive frequency part. This gives (dropping subscripts)

$$b_- = \frac{1}{2}(c + c')a \tag{16}$$

and

$$b_+ = \frac{1}{2}(c - c')a \, . \tag{17}$$

Thus a mode with given (l', m') will show up in the mode (l, m) with the amplitude multiplied by

$$c^+_{l,m,l',m'} = (c_{l,m,l',m'} + c'_{l,m,l',m'})/2 \, . \tag{18}$$

The reason for the choice of normalization of the masks is that it is convenient that

$$c^+_{l,m,l,m} = 1 \tag{19}$$

for all (l, m) if there is no velocity projection factor, no apodization and the integration is done over the whole Sun.

Coarse sampling, seeing and any interpolations performed on the images tend to smear the images, causing variations in response that depend on l and m. It is possible to simulate the effect of smearing by convolving the images of the oscillations with a point spread function (PSF), but since the smearing has to be performed in image coordinates it is fairly costly in terms of computing. The most important effect of smearing is to lower the sensitivity as a function of wavenumber. A second order effect (which is often important, as one normally attempts to push the l range) is to increase the leakages relative to the target mode as the effective area observed decreases due to the forshortening close to the solar limb.

The important properties of the noise are the temporal correlation (leading to frequency dependent noise) and the correlation between the noise in different time-series/Fourier transforms. The frequency dependent noise comes from the fact that most noise sources produce temporally correlated noise and not white noise. The resulting spectrum thus depends on the noise generation mechanism. Important contributions to the noise typically originate in the Sun, in the Earth's atmosphere, and in the instrument. Due to the different generation mechanisms, the temporal and spatial properties of these contributions generally are different from another. As an example, both the temporal and the spatial characteristics of various types of solar granulation, scintillation in the Earth's atmosphere and amplifier noise are very dissimilar.

The reason for the noise correlation between the different time-series is the same as the reason for the leakage into a target mode from other modes, namely that the masks used are not orthogonal on the observed part of the Sun. The covariance $e_{l,m,l',m'}$ between the real parts of the time-series for modes with (l, m) and (l', m') is given by

$$\begin{aligned} e_{l,m,l',m'} &= e_0 \int_{-1}^{1} \int_{-\pi/2}^{\pi/2} M_l^m(\phi, x) M_{l'}^{m'}(\phi, x) \mathrm{Var}(\phi, x) \mathrm{d}\phi \mathrm{d}x \\ &= e_0 \delta_{l,m,l',m'} \int_0^{\pi/2} \int_0^{\pi/2} \Big\{ P_l^m(x) P_{l'}^{m'}(x) \\ &\quad \cos(m\phi) \cos(m'\phi) \mathrm{Ap}(r)^2 \mathrm{Var}(r) \sin(\theta) \Big\} \mathrm{d}\phi \mathrm{d}\theta \, , \end{aligned} \tag{20}$$

where it has been assumed that the noise is uncorrelated between different points on the Sun and has a variance $\mathrm{Var}(\phi, x)$ that is symmetric around the equator and the central meridian. The constant e_0 has absorbed factors of 2 and π and the effects of the discreteness of the sampling. The discreteness of the sampling and the (lack of) independence between different points are clearly connected. If the original pixels are assumed independent then the integration elements in the previous equation cannot be, since different pixels are mapped into different sized areas in the ϕ-θ plane. In addition to this (in a sense trivial) problem, there are other more troublesome ones. One is that it is difficult to estimate the covariance matrix; another is that the computational burden goes up by several orders of magnitude if a full covariance matrix has to be used.

For the imaginary parts it similarly follows that the covariance of the noise time-series is given by

$$\begin{aligned} e'_{l,m,l',m'} &= c\delta_{l,m,l',m'} \int_0^{\pi/2} \int_0^{\pi/2} \Big\{ P_l^m(x) P_{l'}^{m'}(x) \\ &\quad \sin(m\phi) \sin(m'\phi) \mathrm{Ap}(r)^2 \mathrm{Var}(r) \sin(\theta) \Big\} \mathrm{d}\phi \mathrm{d}\theta \, . \end{aligned} \tag{21}$$

It also follows that the noise in the real parts of the time-series is uncorrelated with the noise in the imaginary parts if the noise is symmetric around the central meridian. Again it turns out that the covariance between the same frequency point in two different Fourier transforms is proportional to

$$e^+_{l,m,l',m'} = (e_{l,m,l',m'} + e'_{l,m,l',m'})/2 \, . \tag{22}$$

Under essentially the same assumptions as those made in the calculation of the crosstalks, it may be shown that the same symmetry relations (Eqs. (12) and (13)) hold for the e's.

3. Generation of artifical time-series

In the previous section we discussed the basic properties of the time-series. In this section we present a way to generate time-series with prescribed properties. First we will show how to generate a time-series for a single mode, then how to generate noise time-series, and finally how to combine the different time-series, taking into account the crosstalks and the noise correlations.

As noted in the previous section a single mode is well described by a stochastically excited damped oscillator, hence a straightforward way of generating time-series is to model such an oscillator. It is, however, worthwhile to note that for all relevant modes

$$\Delta t_{\mathrm{obs}} \ll \tau, \tag{23}$$

where Δt_{obs} is the time interval between samplings of the mode and τ is the mode lifetime. As previously discussed, this means that it is not necessary to 'kick' the mode each timestep, but only much more often than τ. Since the generation of the random numbers used to determine the kicks is fairly expensive computationally, a considerable reduction in running time can be acheived by applying kicks only at every nth observed time, with n fairly large.

Consider a mode with a frequency ν, lifetime τ and an rms velocity amplitude of v_{rms}, 'observed' with a time cadence Δt_{obs} and with kicks applied with a cadence $t_{\mathrm{kick}} =$

$N_{\rm kick}\Delta t_{\rm obs} \ll \tau$. In the simplest form of the algorithm we use, the time-series of the mode is started out by setting the initial value a_0 of the time-series to $v_{\rm rms}({\rm rand} + {\rm i} * {\rm rand})$, where rand is a normally distributed random number with unit variance. The first chunk of the time-series $a_k, k = 0, ..., N_{\rm kick} - 1$ is then set to

$$a_k = a_0 \exp(2\pi\nu {\rm i} k\Delta t_{\rm obs} - k\Delta t_{\rm obs}/\tau) \,, \tag{24}$$

for $k = 0, ..., N_{\rm kick} - 1$. The initial value a_0' of the next chunk is set to

$$\begin{aligned} a_0' &= a_0 \exp(2\pi\nu {\rm i} N_{\rm kick}\Delta t_{\rm obs} - N_{\rm kick}\Delta t_{\rm obs}/\tau) \\ &\quad + a_{\rm kick}({\rm rand} + {\rm i} * {\rm rand}) \\ &= a_0 \exp(2\pi\nu {\rm i} t_{\rm kick} - t_{\rm kick}/\tau) \\ &\quad + a_{\rm kick}({\rm rand} + {\rm i} * {\rm rand}) \,, \end{aligned} \tag{25}$$

where $a_{\rm kick} = v_{\rm rms}\sqrt{1 - \exp(-2t_{\rm kick}/\tau)} \approx v_{\rm rms}\sqrt{2t_{\rm kick}/\tau}$, such that the expectation value of the power stays constant. Finally the individual chunks are concatenated to yield the complete time-series. Notice that apart from a_0 all the chunks are identical and that it is therefore possible to compute the exponential once and for all. Also notice that the time-series generated will not have exactly the specified $v_{\rm rms}$, but rather $v_{\rm rms}\frac{\tau}{t_{\rm kick}}(1 - \exp(-t_{\rm kick}/\tau))$, as the decay of the mode during each chunk has not been taken into account in the calculation of the initial a_0 and $a_{\rm kick}$.

Kicking the mode only at selected timesteps does lead to a slightly distorted line profile, in particular small bumps appear around the peak with a spacing of $t_{\rm kick}^{-1}$. These bumps are, however, so small and so far away from the main peak that they drown in the noise for a realistic noise level. The infrequent driving also leads to correlations between different points in the Fourier transforms for a given mode (again at a distance of the order $t_{\rm kick}^{-1}$), but again these should only affect frequencies far from the main peak. In order to reduce the first of these problems, we have not used the method as just described. Rather than making the final time-series by concatenating chunks of length $t_{\rm kick}$, we make them by adding overlapping series of length $2t_{\rm kick}$, each multiplied by a triangular tapering function (see Fig. 2). This means that the kicks are applied more gradually, giving less distorted power spectra.

An example of mode spectra generated this way is shown in Fig. 3.

Frequency dependent noise can be created either by passing white (frequency independent) noise through a digital filter, or by generating frequency dependent noise in the Fourier transform and transforming it back to the time domain. We have chosen the latter approach. To avoid doing very long Fourier transforms and to avoid storing long noise time-series, we create small chunks of noise time-series and concatenate them. Since the noise in different chunks is not correlated, the noise spectrum will not be correct below frequencies of approximately $t_{\rm noise}^{-1}$, where $t_{\rm noise}$ is the length of the chunks. Since this is far below the frequency of the modes for even moderatly long series, this should not lead to any major problems, however.

The correlation of the noise between different time-series/Fourier transforms is slightly more difficult to handle. To generate time-series y_i with a prescribed covariance matrix E one starts with uncorrelated times series x_i with unit variance and sets $y = Gx$ at each time step, where $E = GG^T$. One way to find G is to perform a Cholesky decomposition of E (see eg. Golub & Van Loan 1989), where G a lower tridiagonal matrix. That this produces the desired covariance follows from

$$\begin{aligned} {\rm Cov}(y_i, y_j) &= \sum_k G_{ik}G_{jk}\sigma_k^2 = \sum_k G_{ik}G_{jk} \\ &= \sum_k G_{ik}(G^T)_{kj} = (GG^T)_{ij} = E_{ij} \,, \end{aligned} \tag{26}$$

and is the traditional way of generating vectors with a prescribed covariance matrix.

A problem with generating the noise time-series this way is that, in order to get the noise correlations between distantly spaced l's correct, one should generate all noise time-series for all l's and m's and use the covariance matrix for all l's and m's. This is clearly not realistic; the covariance matrix is far too big. Even for one l it is problematic to use the matrix for all m's when l is large, as the condition number for E becomes very large (meaning that the covariance matrix is close to singular).

A potential solution to this problem is to note that when the integrals in Eqs. (20) and (21) are discretized one obtains $E = AA^T$ for some matrix A (which generally has a very large number of columns). If the Singular Value Decomposition (SVD) of this matrix is $A = U\Sigma V^T$ then

$$E = AA^T = U\Sigma V^T V\Sigma U^T = U\Sigma\Sigma U^T = U'U'^T, \tag{27}$$

where $U' = U\Sigma$ can be used the same way that G from the Cholesky decomposition was used before. This U' can be truncated by only including the highest singular values. It can be shown that the error in E introduced by truncating U' goes like the sum of the squares of the neglected singular values, and is thus presumably insignificant. In any case the errors introduced this way are probably of the same magnitude as those in the Cholesky decomposition, if not smaller, if enough singular values are retained. In addition the truncated U' matrix has fewer coloumns than G and thus a smaller x vector can be used. If this method is used it should be possible to cover a substantial l-range consistently without running into numerical problems, but this method has not been tested.

Finally the finished time-series are produced by adding up the modes according to Eq. (8) and adding the noise as just described.

4. Discussion

It would clearly be useful to be able to compare the artificial data to real data to see if we have indeed been able to reproduce the essential properties, such as crosstalks, mode frequencies, amplitudes and linewidths. Apart from checking such obvious things as that the overall levels of modes and noise are correct and looking at the linewidths of the produced modes, it is, however, very difficult to check whether the properties of the artificial data are identical to those of the real data.

Unfortunately many of the possible comparisons depend critically on knowing parameters for the real data that are difficult to determine and are largely irrelevant for the purpose of testing analysis programs. Examples of such parameters are exact mode frequencies, linewidths, and splittings, which, although they may be the final goal of the analysis, are probably not critical to model exactly as long as they are aproximately correct. For instance, if all frequencies are in error by say 1μHz, this is unlikely to make any difference for the purpose of testing

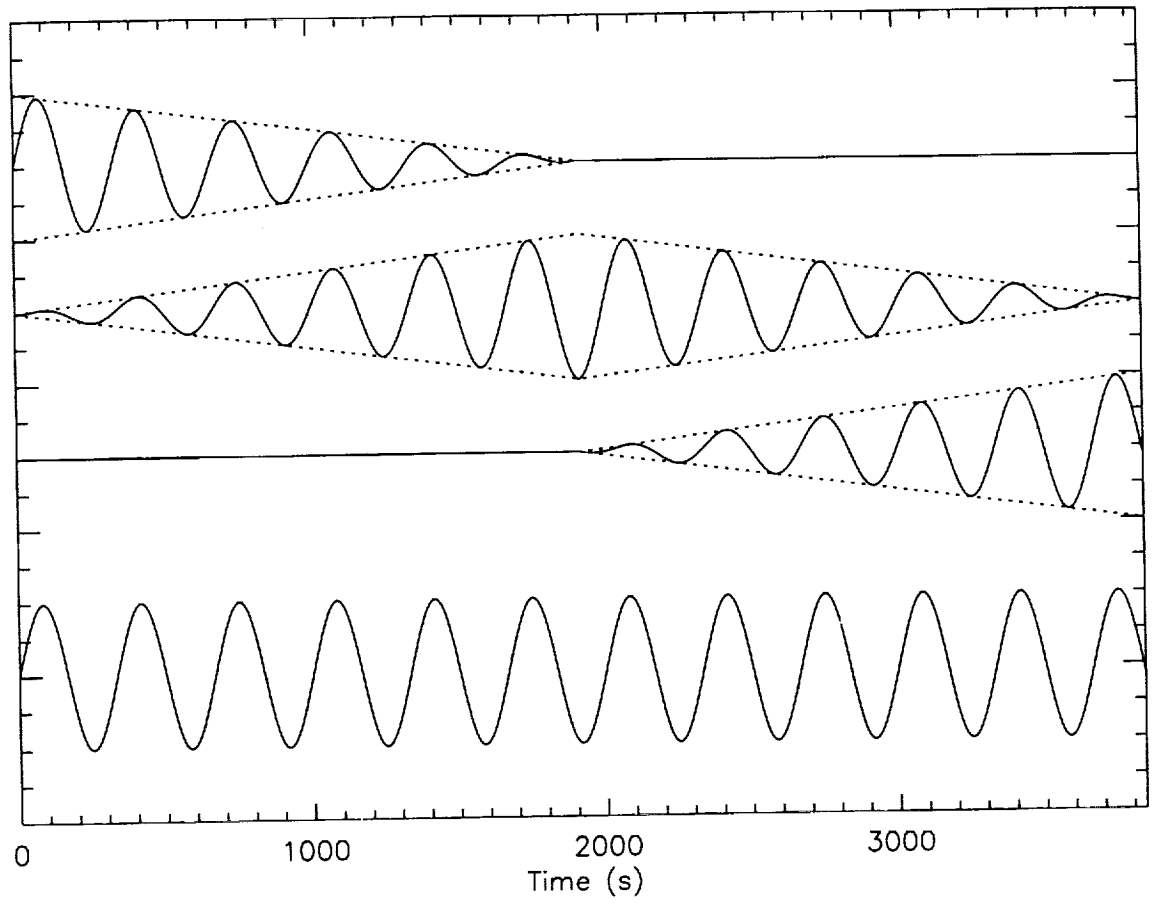

Fig. 2. The principle behind the triangular functions. Three consecutive chunks and their corresponding envelopes have been shown together with their sum. A mode with a frequency of 3000μHz and a chunk length of 2 minutes has been used. For clarity no decays or excitations were used. Also the coarse sampling of the modes generally used was ignored

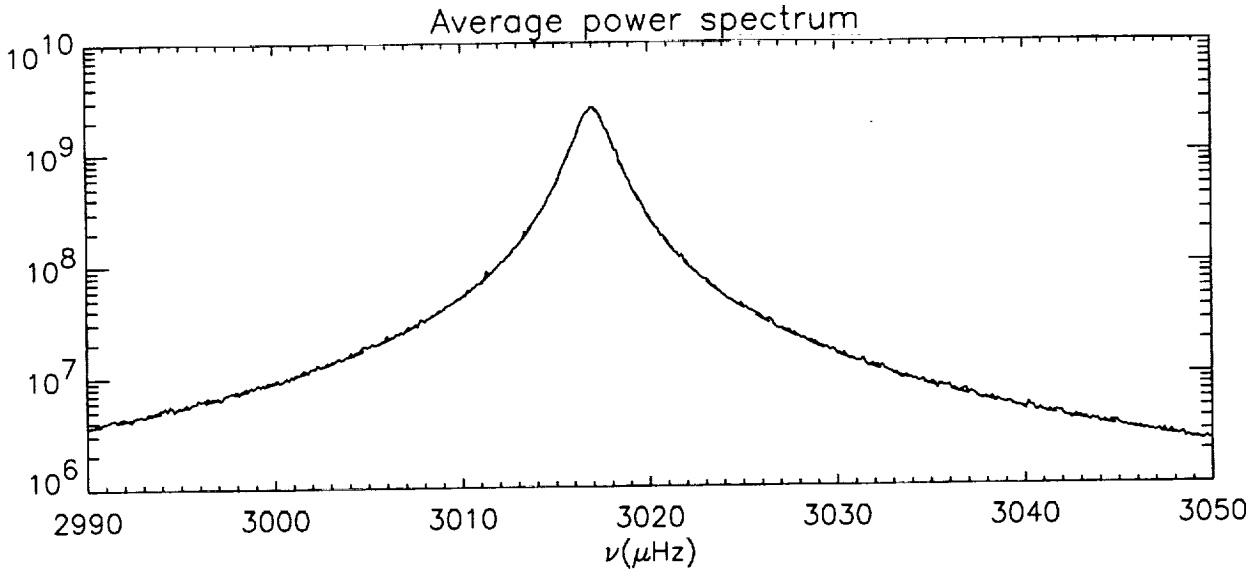

Fig. 3. The average of 1000 power spectra for a mode with a frequency $\nu = 3017\mu$Hz and a HWHM $w = 1\mu$Hz. The theoretical limit (scaled to match the power level) has been indicated by a dashed line (which is not visible)

analysis programs, as long as the frequency spacings are close to their observed values.

A section of the p-mode spectrum for artificial data produced using this code and the corresponding spectrum from a run with the Fourier Tachometer are shown in Fig. 4. For illustrative purposes, the power spectra were collapsed in m, shifting the spectra for the individual m's to remove an approximation the the frequency shifts caused by the solar rotation. The different plots of artificial data in Fig. 4 show where each of the peaks in the spectrum of the real data comes from, and illustrate the complexity introduced by crosstalks and time gaps. The increased complexity introduced by time-gaps is, indeed, the strongest argument for projects such as GONG and SOHO, which aim to reduce the time-gaps to negligible proportions.

As can be seen from the third and fourth plots in the right hand column of Fig. 4, the real and the artificial data look very much alike. Note that it was not attempted to match the noise level perfectly between the real and artificial data. Although the general appearance of the real and artifical data are very similar, there are minor differences, such as different ratios of the power between peaks. Among the reasons for these differences are the different signal and noise levels, slightly different dependencies of the mode amplitudes and linewidths with frequencies, slightly different crosstalks due to (for instance) problems in estimating the PSF, and slightly different a-coefficients. Also the fact that modes with $\Delta l \geq 4$ were neglected in the artificial data may be the cause of some of the differences.

Among the more subtle things that can be checked is that the correlations in both the time-series and the Fourier-transforms are as in the real data. A scatter plot of the values in a real and artificial time-series for $(l, m) = (30, 0)$ and $(l, m) = (30, 2)$ is shown in Fig. 5. Notice that the time-series of modes with $\Delta l = 0$ and $\Delta m = 2$ are generally anti-correlated, as the corresponding P_l^m's have opposite sign around the equator ($x = 0$), where the weighting by the masks is the highest. In order not to have the plot dominated by the low frequency noise (which is different), the time-series were high pass filtered. Again it can be seen that the real and artificial data behave similarly. However the match is not perfect, in particular it appears that the overall power level is somewhat higher for the real data.

From looking at observations, it turns out that the distribution of noise on the Sun is not a smooth function of radius only, but often has additional contributions around active regions, in addition to various bizarre instrumental effects. This makes it very difficult to reproduce the correlations in the observed noise. It is of course possible to use the observed correlations from real data to generate the artificial noise, but this approach has other problems. Also, it is far from obvious that the noise properties are the same at the frequencies where they are easy to measure (typically low frequencies) as they are where one cares about them (in the p-mode band). Different noise sources are likely to have very different spatial and temporal characteristics, leading to variations in the correlations with frequency. Also, a large contribution to the noise in the p-mode band seems to be unresolved modes and temporal sidelobes of those. As can be seen from Fig. 4, the addition of a realistic noise level makes a very small difference to the apparent noise level at the peak of the p-mode power distribution, especially if time gaps are present. When looking at the data without time gaps (which would presumably be similar to that obtainable with the GONG network), it is important to note that the contributions from modes with $\Delta l \geq 4$ have been neglected, and that, given the level of the modes with $\Delta l = 3$, they are likely to contribute significantly to the apparent background noise level.

In a sense one of the best checks of whether the essential properties have been modelled properly is to check that various analysis programs give mode parameters close to the input values. Unfortunately this does not prove (or disprove) that the underlying assumptions are correct, only that they have been consistently implemented between this program and the analysis programs. This is of course not totally useless, but it is not entirely satisfactory either.

All in all it thus appears that the best one can do is use the program assuming that things have been properly implemented. If it later turns out that the analysis program behaves differently on real and artificial data by, for instance, indicating that the statistical properties are different, it will be necessary to find out which property of the real data was neglected or incorrectly implemented.

Despite these problems with verifying the program, it has been an extremely useful tool for testing our different analysis procedures. One of the very useful features is the ability to turn various properties on and off. In particular it is useful to be able to turn off the noise, the time gaps and/or various parts of the crosstalks, as those can easily lead to problems if not properly taken care of. Some of the effects of turning various properties of the time series on and off have been shown in Fig. 4. As can be seen the spectrum is considerably simpler when crosstalks and/or time gaps are neglected.

Using this program we have been able to identify problems in some of the earlier versions of our analysis codes in the determination of the mode linewidths and the a-coefficients. These parameters are very sensitive to certain errors in modelling the crosstalks, as the individual m's are generally not resolved in the power spectra.

Also, when we have been concerned that some particular neglected effect has been causing problems, or when modifications to the programs have been made, it has been extremely useful to know what the correct results were. If one had had to rely on real data for these tests, it would only have been possible to see that some parameter changed, and not whether the change improved the results or made them worse.

A description of one analysis method and some of the results obtained by analysing artificial data generated by the program described here can be found in Schou (1993). A more systematic comparison of results from analysing the output of this program using a number of different analysis methods is in preparation.

It is possible to address some problems using this program that it is not possible to treat using programs taking fewer of the physical properties of the modes into account (eg. Anderson et. al. 1991). These include the effects of crosstalks, both when it comes to introducing interfering peaks from neighbouring modes and correlations among the time-series, and the effects of correlated noise. These are problems that are likely to affect the mode linewidths and the so-called a-coefficients (describing the effects of asphericities on the mode frequencies).

Also note that the dependency of the points in the Fourier transform caused by the gaps in the time-series are treated properly here, which is not the case if the Fourier transforms are generated by multiplying white noise by the limit spectrum (from Eq. (4) convolved by the sidelobe spectrum of the

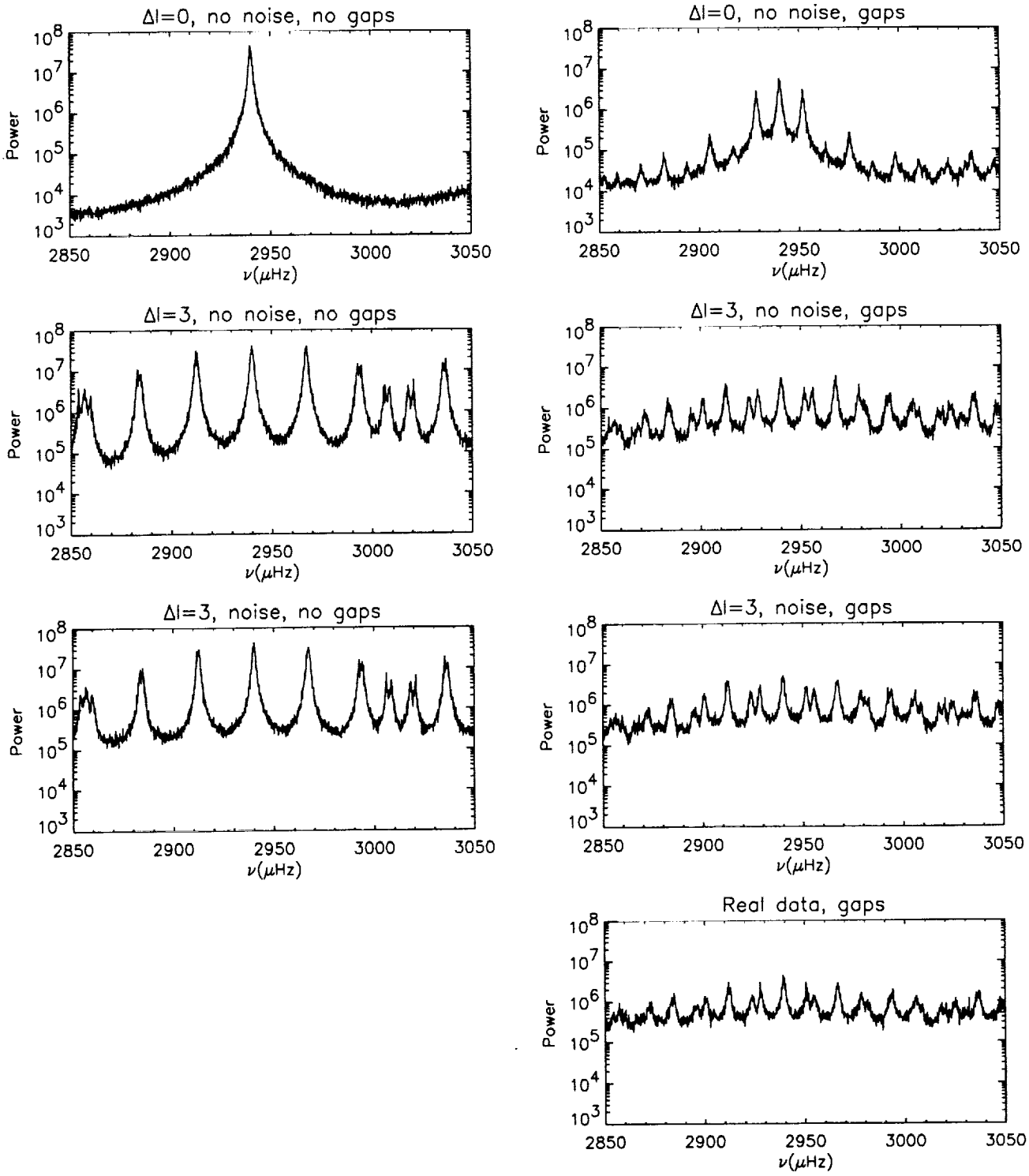

Fig. 4. Examples of power spectra for $l = 30$. The power spectra for the individual m's have been shifted according to the a-coefficients used to generate the artificial data and summed over m. Only a small fraction of the spectrum around $n = 12$, which is close to the peak of the power in the p-mode band, has been shown. The left column of plots show data without time-gaps, the right hand column shows plots with the time-gaps used for Fig. 1. The top row shows the spectrum with no crosstalk from neighbouring l's and no noise. The second row shows a spectrum in which crosstalk out to Δl of 3 has been included, but still without noise. The third row is similar to the second, except that a noise level similar to that from observations with the Fourier Tachometer has been added. The bottom row shows a spectrum using real data from the run with the Fourier Tachometer used for Fig. 1. Unfortunately, it was not possible to make the lower left hand plot

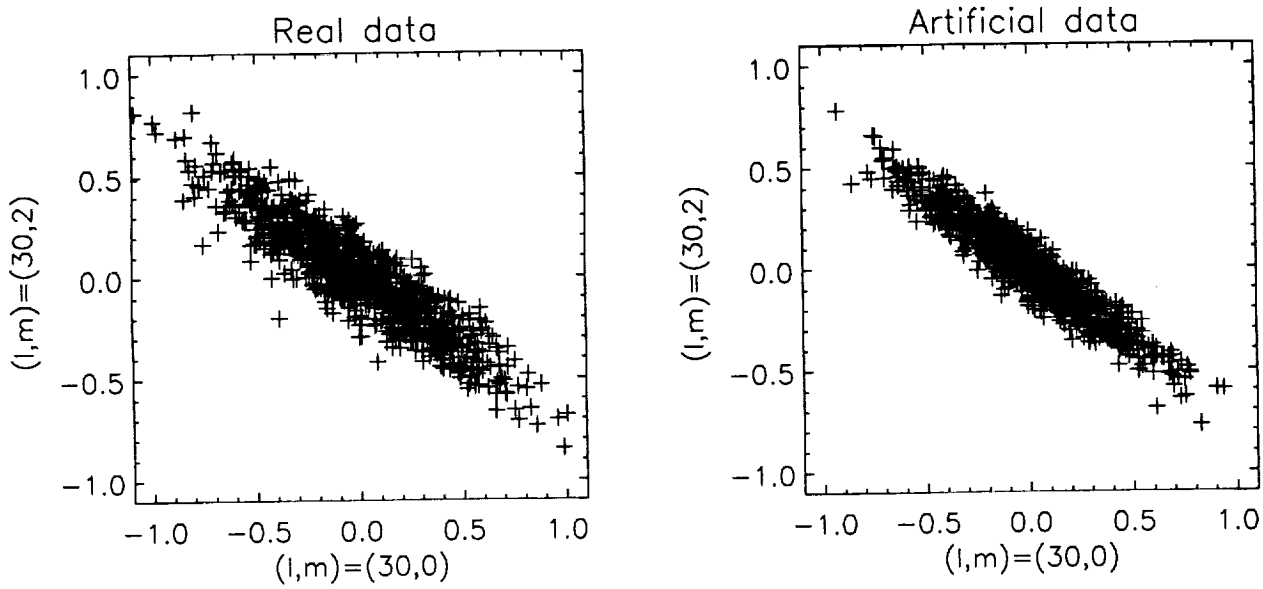

Fig. 5. Scatter plots showing the values of points in a time-series for $(l, m) = (30, 2)$ versus those for a time series with $(l, m) = (30, 0)$. The left hand plot shows results from real data, while the right hand plot shows artificial data. In order to reduce the effect of the low frequency noise, the time-series were high pass filtered. Notice that the overall power level is not matched perfectly

time gaps). One time when these correlations are important is when estimating the errors on the fitted parameters. If the correlations are not taken into account, the values of the Fourier transform in the side lobes and the main lobe are independent, allowing lower standard errors on the fitted parameters than if the correlations are properly modelled.

Unfortunately, we have not been able to eliminate all problems from our analysis procedures by using artificial data. In particular it appears that we have a systematic problem with our determination of the a-coefficients when analysing observations taken with the Fourier Tachometer (see Bachmann et al. 1993). Despite extensive tests using artificial data as described here, we have not been able to find a problem in our analysis programs. On the other hand if it had not been for these test we would probably not have been able to convince ourselves that the problem is not in the time-series analysis, given that this is by far the most complicated part of the analysis. We are therefore inclined to believe that the problem is in our understanding of either the physics or the instrument, rather than in the time-series analysis programs as such.

Acknowledgements. This research was supported in part by the National Aeronautics and Space Administration, through grant # W-17,678, by Statens Naturvidenskabelige Forskningsråd (Danish Natural Science Research Council) and by Forskerakademiet (Danish Research Academy). Also JS wishes to thank Stanford University at which the last revisions were made under NASA contract NAS 5-30386.

References

Anderson, E. R., Duvall, T. L. Jr., Jefferies, S. M., 1991, ApJ 364, 699.

Bachmann, K. T., Schou, J., Brown, T. M., 1993, ApJ (submitted).

Bachmann, K. T., Brown, T. M., Schou, J., 1993 (in preparation).

Christensen-Dalsgaard, J., Berthomieu, G., 1991, Theory of Solar Oscillations. In: Cox, A. N., Livingston, W. C. and Matthews, M. (eds.) Space Science Series, Solar Interior and Atmosphere. University of Arizona Press, Tucson, p. 401.

Golub, G. H., Van Loan, C. F., 1989, Matrix Computations, 2nd. edition, The Johns Hopkins University Press, Baltimore.

Hill, F., Deubner, F.-L., Isaak, G. R., 1991, Oscilation Observations. In: Cox, A. N., Livingston, W. C. and Matthews, M. (eds.) Space Science Series, Solar Interior and Atmosphere. University of Arizona Press, Tucson, p. 329.

Schou, J., 1993 (in preparation).

This article was processed by the author using Springer-Verlag TEX A&A macro package 1991.