ASTRONOMY & ASTROPHYSICS
SUPPLEMENT SERIES

Astron. Astrophys. Suppl. Ser. **117,** 197-201 (1996)

Parameter estimation in epoch folding analysis

S. Larsson

Stockholm Observatory, S-13336 Saltsjöbaden, Sweden
Internet: stefan@astro.su.se

Received January 16; accepted October 30, 1995

Abstract. — We describe a procedure that we have implemented to use epoch folding to estimate pulse period, shape, amplitude and phase (with uncertainties) for coherent oscillations in time series data. We improve on traditional techniques by fitting the χ^2 as a function of test period with a response function which takes into account both data sampling and the oscillation pulse shape. It is shown that the epoch folding procedure makes optimum usage of the full pulse shape information, or equivalently all its Fourier components, in the period determination. An analytic expression for the period error is also given for this general, non-sinusoidal, case. The error estimate, which is equivalent to that for a least-squares fit of a set of harmonically related Fourier components, is verified by Monte Carlo simulations.

Key words: methods: data analysis

1. Introduction

The problem of estimating periodicities in time series data is common in astronomy as well as in many other sciences. A large variety of methods, with different statistical properties, are in use. The choice of the method, such as Fourier transform, epoch folding, Rayleigh folding, etc., will depend on signal-to-noise ratio, data length, evenness of sampling and on the character of the signal to be analyzed. In this paper we will consider epoch folding, which is often used in cases where one wants to search for a coherent signal in large amounts of data with low or moderate signal-to-noise ratio. One advantage of epoch folding is that it is more easily applied to cases of non-evenly sampled data than e.g. Fourier methods. Investigations of the statistical properties of epoch folding have mainly concerned the estimate of detection significances (Schwarzenberg-Czerny 1989 and Davies 1990, 1991). In this paper we describe a procedure by which we use epoch folding to estimate pulse period, shape, amplitude and phase (with uncertainties) for a single coherent oscillation in time series data. Our two main improvements of the traditional epoch folding techniques are: 1. We estimate the oscillation period by fitting a χ^2 response function which takes into account both data sampling and oscillation pulse shape. 2. We give an analytic expression for the period uncertainty in the general, non-sinusoidal case, and show that our method gives errors consistent with basic statistical limits for least-squares fits to the time series. For non-sinusoidal oscillations the epoch folding automatically adds together the signal of all harmonic components, so that the period determination makes maximum use of the available information. Our parameter estimation procedure works well also for non-evenly sampled data, and as long as the sampling is reasonably even the error estimates will still provide quite reliable uncertainty limits for the oscillation parameters.

2. Method

In most applications of epoch folding, χ^2 (over the folded pulse) is calculated for a range of test periods and the oscillation period is taken as the period at the largest χ^2 value or at the χ^2 maximum estimated from a polynomial fit near the peak. A suggestion by Leahy (1987) to estimate period and amplitude by fitting an analytic χ^2 function to $\chi^2(P_{\rm test})$ was the starting point for the development of our estimation procedure.

In summary the analysis contain the following steps which will be further elaborated below.

1. Calculate χ^2 as a function of test folding period.
2. Calculate the χ^2 response function by sampling a synthetic pulse at the same times as the data. Then fit this χ^2 function to that of the data to estimate pulse period and amplitude.
3. Fold the data with the best fit period and make a Fourier decomposition of the pulse profile.
4. Using the Fourier pulse determination, go back to step 2. It is in most cases sufficient to rerun steps 2 & 3 only once or twice.

2.1. *Calculating* χ^2

In epoch folding the data is folded modulo a test period and coadded into N_b phase bins. We then calculate

$$\chi^2 = \sum_{i=1}^{N_\mathrm{b}} \frac{(x_i - \overline{x})^2}{\sigma_i^2}\,, \tag{1}$$

so that $\chi^2 \sim N_\mathrm{b} - 1$ if the scatter of the bin points over the pulse cycle is as expected for Gaussian noise with standard deviations σ_i. A high χ^2 value ($\gg N_\mathrm{b} - 1$) will signal the presence of a periodicity for which the significance can be estimated from the χ^2 distribution function for $N_\mathrm{b} - 1$ degrees of freedom. We want to make two comments on our definition and use of χ^2 from Eq. (1). First, since our aim is to determine the parameters for an oscillation we already know (or assume) is in the data, the σ_i values are calculated from either the variance of the data points in each phase bin or from initial error estimates associated to each data point. To estimate the significance of a χ^2 value the standard deviations should instead (under the null hypothesis) be defined as $\sigma_i = \sigma_\mathrm{tot}/\sqrt{n_i}$, where σ_tot is the standard deviation of the unfolded time series and n_i is the number of data points in bin i. Secondly, for short time series a more proper test statistic than χ^2 is to use the L-statistic for estimates of significances (Schwarzenberg-Czerny 1989 and Davies 1990, 1991).

2.2. *Estimating pulse period*

Just as in the case of power density spectra the peak in the $\chi^2(P_\mathrm{test})$ is a broadened function with sidelobes. The $\chi^2(P_\mathrm{test})$ function depends on the time series sampling and length as well as on the pulse shape. For an evenly sampled sinusoidal signal $\chi^2 \propto \sin^2(x)/x^2$ (Leahy 1987) as shown in Fig. 1. Leahy suggested that the pulsation period and amplitude be determined by fitting the analytic χ^2-function for a sinusoid to $\chi^2(P_\mathrm{test})$ calculated from the data. From Monte Carlo simulations Leahy also gave a relation for estimating the uncertainty in the period and amplitude determinations. To estimate the error in period we instead use an analytic expression derived for power density spectra and least-squares fitted sinusoids. Kovács (1980) estimated the frequency shift in power density spectra by nearby white noise frequency components and found the gaussian frequency error to be

$$\sigma_f = \frac{\sqrt{2}a\sigma_\mathrm{tot}}{\sqrt{N}AT}\,, \tag{2}$$

In this equation, N is the total number of data points, A is the sinusoidal amplitude and T is the total time length for the data. For the parameter a, Kovács had to rely on a Monte Carlo estimate which gave a value of $a \approx 0.45$. Gilliland & Fisher (1985), in the analysis of chromospheric activity data, also used Eq. (2) but with a value corresponding to $a = 0.53$. An expression which is essentially the same as Eq. (2) is given by Bloomfield (1976),

$$\sigma_\omega = \frac{\sqrt{24}\sigma_\mathrm{tot}}{N^{3/2}A} + \text{smaller terms}\,, \tag{3}$$

Since ω is given in radians per sampling time interval this corresponds to $a = 0.551$. Note that Bloomfield assumes that the time series is evenly sampled. Using the analytic value of a derived by Bloomfield, we can rewrite Eq. (2) in terms of period error as,

$$\sigma_P^2 = \frac{6\sigma_\mathrm{tot}^2}{\pi^2 N A^2 T^2} P^4\,, \tag{4}$$

Let us now consider a non-sinusoidal oscillation. In this case the time series can be described as a set of Fourier components with frequencies, $\nu_k = k\nu_1$. Since the frequencies are related by exact (integer) multiples, it is possible to increase the precision of the period determination by taking a weighted mean of the fundamental period associated with these harmonics. In epoch folding this is done automatically since the higher harmonics are folded with k cycles over one fundamental pulse cycle. We can then also generalize Eq. (4) to give

$$\sigma_P^2 = \frac{6\sigma_\mathrm{tot}^2}{\pi^2 N T^2} \frac{P^4}{\sum_{k=1}^{m} k^2 A_k^2}\,, \tag{5}$$

To apply this estimate in practice we need to know the harmonic content of the signal, which is determined in the next step of the analysis.

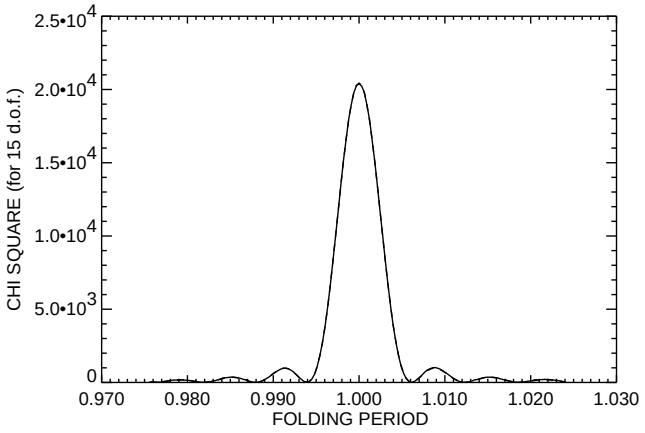

Fig. 1. The plot shows the expected χ^2 calculated over the pulse shape as a function of the folding period when the evenly sampled time series contains a pure sinusoidal oscillation

2.3. *Fourier decomposition of the pulse profile*

The Fourier decomposition is done as follows. The best fit period gives a folded pulse with value x_i and standard error σ_i (in bin i). After subtraction of the mean, the number of degrees of freedom is $N_\mathrm{b} - 1$. Each Fourier component has two parameters so fitting $(N_\mathrm{b}-1)/2$ harmonics will absorb all white noise in the folded pulse, or equivalently the

white noise power per fitted harmonic is $\sigma^2/[(N_\mathrm{b}-1)/2]$. After a linear regression fit of the desired number of Fourier components, this white noise contribution should be subtracted from each harmonic variance ($A_k^2/2$). Each harmonic is fitted as $a_k \sin(k\omega_1 t)+b_k \cos(k\omega_1 t)$ from which the phase for each harmonic can be calculated. The fitted amplitude values have to be corrected for binning, which is done by multiplying with the factor

$$g(N_\mathrm{b}) = \frac{1}{\sqrt{f(N_\mathrm{b})}} = \frac{1}{\sqrt{1-\frac{\pi^2}{3{N_\mathrm{b}}^2}+\frac{2\pi^4}{45{N_\mathrm{b}}^4}}}\,, \qquad (6)$$

(where $f(N_\mathrm{b})$ is the related factor used by Leahy 1987.)

To select significant Fourier components we use the uncorrected amplitudes and the noise variance to estimate the false alarm probability for each component. This can be done in the same way as for peaks in power density spectra. In our case we have error estimates for the points in the pulse profile, so we can test the significance of each Fourier amplitude.

For Gaussian white noise, the variance estimates, V_k, at the Fourier frequencies are independently distributed with a probability distribution proportional to a χ^2 with two degrees of freedom, i.e.

$$\frac{V_k}{\sigma^2_{N_\mathrm{b}}} = \chi^2_2\,, \qquad (7)$$

This distribution is exponential and gives a probability for $V_k/\sigma^2_{N_\mathrm{b}}$ to be larger than some value, z of (see e.g. Priestley 1981, p. 388),

$$P[(\frac{V_k}{\sigma^2_{N_\mathrm{b}}}) > z] = \int_0^z \frac{1}{2}\exp(-\frac{s}{2})\,\mathrm{d}s = \exp(-\frac{z}{2})\,, \qquad (8)$$

2.4. *Iteration*

For the Fourier decomposition we now select the components with amplitudes above some significance level, using the result in Eq. (8). This Fourier pulse is then sampled at the same time points as the data and a new χ^2 response function is calculated and fitted to the χ^2-values for the data. It is usually sufficient to rerun steps 2 & 3, only once or twice.

3. Example and verification

In order to verify our parameter and uncertainty estimation procedure we have made extensive Monte Carlo simulations injecting coherent oscillations with a range of different pulse shapes and signal-to-noise ratios. In each case the error estimates were compared with the standard deviation in the parameter estimates for some 200 to 1000 simulations. The results from these simulations were found to be consistent with the estimates in Sect. 2. In particular the deviation of the period estimate for a set of 2400

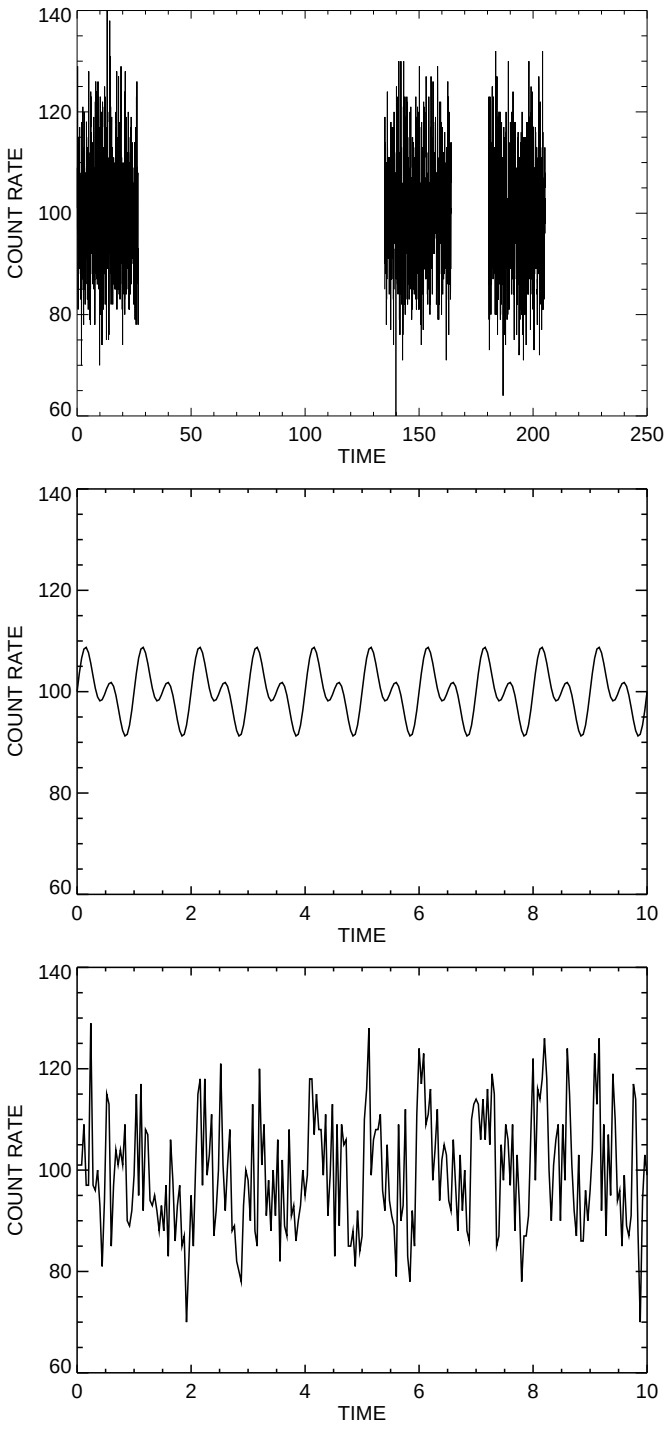

Fig. 2. a-c). An example of a data simulation used to test the analysis procedure. The data is evenly sampled (25 points per time unit) except for the two time gaps. **a)** The full time series, 2024 points. **b)** The first part of the time series without any noise added. The injected oscillation has a period equal to one and is composed of two equally strong ($A = 5$) Fourier components (the fundamental and the first overtone). **c)** The same time series segment as in **b)** but after adding Poisson noise

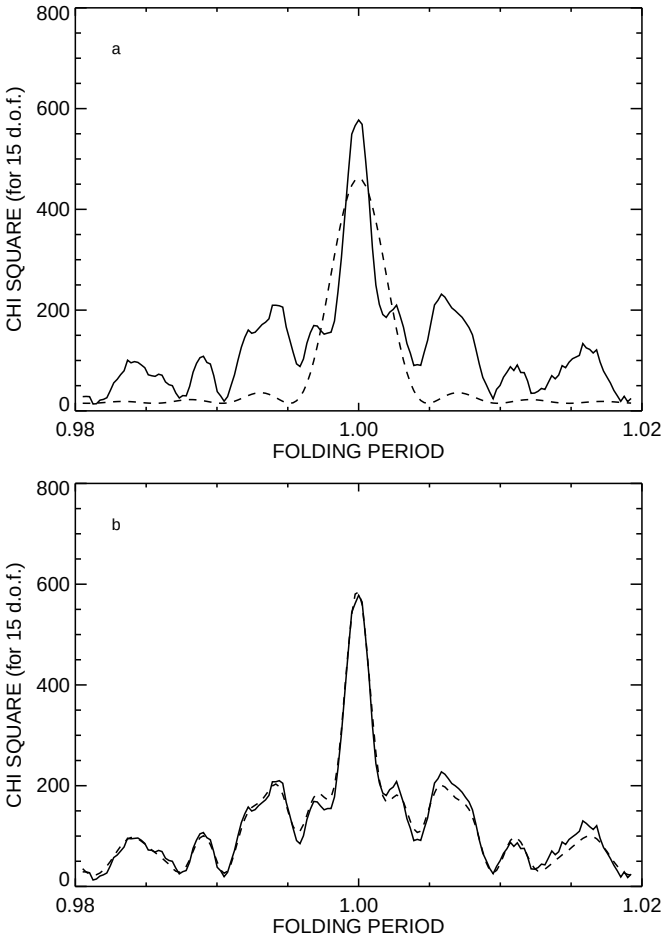

Fig. 3. a-b). χ^2 calculated over the pulse shape for a range of different folding periods near the true fundamental period. The solid line is for the simulated data plotted in Fig. 2a. **a)** Best fit (dashed curve) theoretical χ^2-function for the case of a sinusoidal oscillation. **b)** Best fit (dashed curve) numerically calculated χ^2-function for the particular data sampling and pulse shape (determined from the folded pulse) of this simulation

simulations containing a single sinusoidal oscillation was compared to the predicted value from our Eq. (4). The ratio of the observed to predicted standard deviation was 1.026 ± 0.018. The corresponding value for the parameter a is 0.565 ± 0.010, to be compared with the theoretical value of 0.551 and the estimates of 0.45 and 0.53 given by Kovács (1980) and Gilliland & Fisher (1985) respectively. In our simulation the data was folded in 16 phase bins. The binning effectively reduces the amplitude of the folded data by the factor $\sqrt{f(N_\mathrm{b})}$ (see Eq. (6)). The expected standard deviation does therefore, in this case, correspond to $a = 0.555$ so our simulation estimate of a is consistent with the analytic expression to within the statistical uncertainty (1σ). Note that these results were obtained for evenly sampled data. As already mentioned the error estimates will be quite reasonable also for mildly non-even

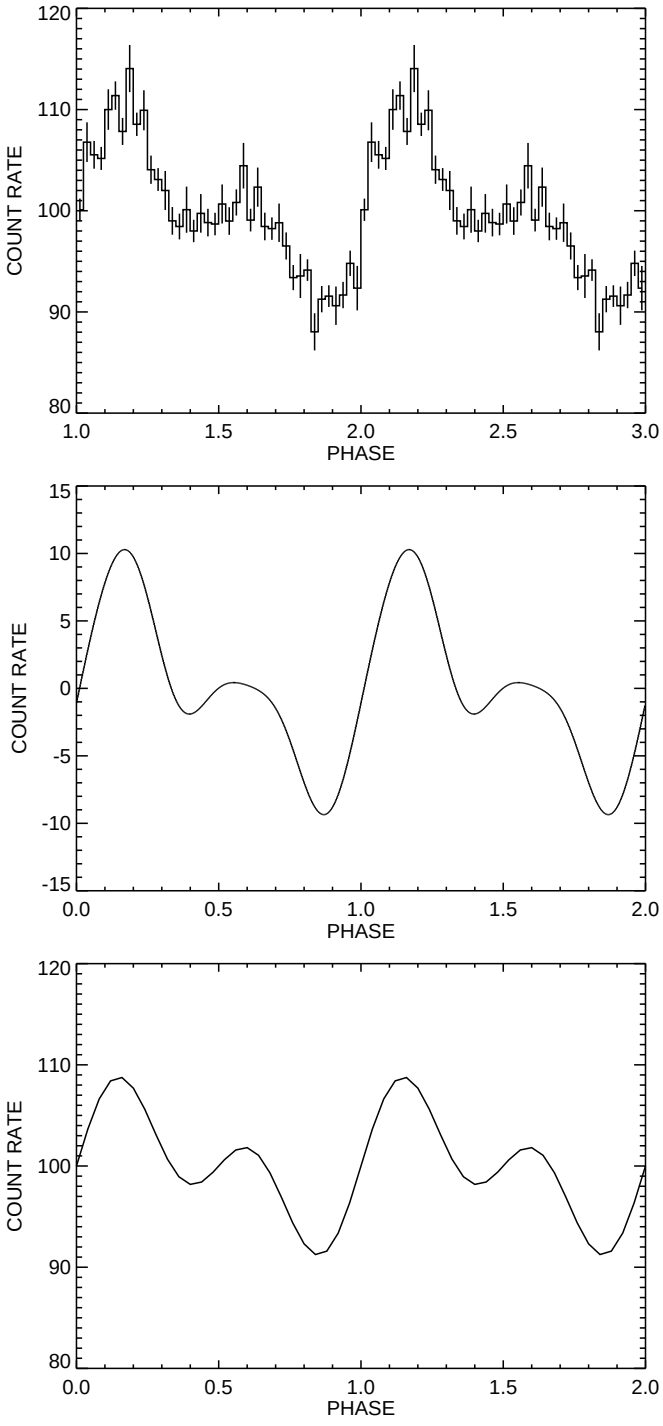

Fig. 4. a-c). Oscillation pulse shapes (two cycles). **a)** Pulse shape obtained by folding the simulated data at the best fit period. **b)** Fourier pulse obtained by fitting Fourier components to the pulse in **f**. **c)** The shape of the original pulse injected into the simulation

sampling. One example of a simulations with two substantial data gaps is shown together with analysis results in Figs. 2-4. The simulated data is shown in Figs. 2 a-c with a blow up of the first 10% of the data to show the injected oscillation (with two equally strong Fourier components) before and after the Poisson noise was added. The calculated χ^2 as a function of test period is plotted in Figs. 3a and 3b. In the first of these figures it is shown together with the best fit χ^2 function for an evenly sampled sinusoid. The narrow width of the χ^2 peak compared to that of the analytic function is because it contains a Fourier component also at the first harmonic. A more precise period determination is of course possible when the peak is more narrow. This is just a graphical representation of how the period error in Eq. (5) depends on the harmonic content of the pulse shape. In Fig. 3b we show the resulting fit after we have applied our full parameter estimation procedure. It is clear from this fit that most of the χ^2 modulation pattern is an effect of windowing and pulse shape. Finally in Fig. 4 we compare the pulse shape of the inject oscillation with the folded pulse shape for the best fit period and the pulse shape resulting from the Fourier decomposition of the folded pulse.

For unevenly sampled data the uncertainty in the period estimate depends on the actual sampling distribution. The more asymmetric and clumped the data distribution is, the larger is the deviation of errors from the estimate of Eq. (5). For data that consist of evenly sampled segments separated by gaps (such as in Fig. 2) Eq. (5) will remain a good error estimate as long as the data is not strongly concentrated to one part of the total time range. This is what one would expect intuitively, but it is also what we found from simulations with $1-12$ data gaps covering up to 99% of the length of the time series. The two most extensive sets of simulations that we ran gave period deviations corresponding to $a = 0.469 \pm 0.022$ and 0.532 ± 0.037. The first set consisted of 380 simulations with a sampling distribution almost identical to that in Fig. 2, i.e. 2024 points that were evenly sampled except for two gaps covering a fraction 0.74 of the total time series length. The second set of 160 simulations had the same number of points, total length and gap fraction but with six gaps instead of two. These gaped data simulations were made with the same pulsation parameters as in the previous example (Figs. 2-4).

In the limit of a time series with few data points our folding procedure provides no advantage to directly fitting Fourier components to the data. The method can still be applied in such cases however, as long as there is good, but not necessarily complete, coverage of the oscillation cycle. Also, for time series with gaps the length of data subsegments with respect to the oscillation period will only have an affect on the aliasing problem (the sidelobe pattern of the χ^2 function) and not on the applicability of the method. This last statement is true if the time series contains no variability component in addition to the coherent oscillation and the white noise. This is really the most important assumption on which our analysis procedure is based. It is also the main limitation that has to be considered when applying the method to real observational data.

4. Summary

We have presented a procedure by which epoch folding is used to estimate pulse period, shape, amplitude and phase for coherent oscillations in time series data. We also give a generalized analytic expression for the uncertainty in the period estimate. The error estimate is verified by Monte Carlo simulations for evenly sampled data with and without data gaps.

Acknowledgements. This work was supported by the Swedish Natural Science Research Council and the Swedish National Space Board.

References

Bloomfield P., 1976, Fourier analysis of time series: An introduction. John Wiley & Sons, New York
Davies S.R., 1990, MNRAS 244, 93
Davies S.R., 1991, MNRAS 251, 64
Gilliland R.L., Fisher R., 1985, PASP 97, 285
Kovács G., 1981, Ap&SS 78, 175
Leahy D.A., 1987, A&A 180, 275
Priestley M.B., 1981, Spect. Anal. of Time Ser. Academic Press, London
Schwarzenberg-Czerny A., 1989, MNRAS 241, 153