Overfitting and correlations in model fitting with separation ratios
(Research Note)

Ian W. Roxburgh

Astronomy Unit, Queen Mary University of London, Mile End Road, London E1 4NS, UK. e-mail: I.W.Roxburgh@qmul.ac.uk
Received / Accepted

ABSTRACT

The $r_{01}$ and $r_{02}$ separation ratios are not independent so combining them into a single series $r_{010}$ is overfitting the data, this can lead to almost singular covariance matrices with very large condition numbers, and hence to spurious results when comparing models and observations. Since the $r_{02}$ ratios are strongly correlated with $r_{01}$ and $r_{03}$ ratios, they should be combined into a single series $r_{012}$ (or $r_{021}$), which are not overfitted, and models and observation compared using the covariance matrix $\text{cov}_{012}$ (or $\text{cov}_{021}$) of the combined set. I illustrate these points by comparing the revised Legacy Project data with my results on the 10 Kepler stars in common.

Key words. stars: oscillations, - astroseismology - methods: data analysis - methods: analytical - methods: numerical

1. Introduction

Frequency separation ratios are widely used in astroseismic model fitting, i.e. finding models whose oscillation properties match an observed set, as these ratios are almost independent of the outer layers of a star (Roxburgh and Vorontsov 2003, 2013).

The ratios, constructed from frequencies $\nu_{nl}$ for angular degree $\ell = 0, 1, 2$, are customarily defined as

$\begin{align*}
r_{01}(n) &= \frac{(\nu_{n-1,0} - 4\nu_{n-1,1} + 6\nu_{n,0} - 4\nu_{n+1,0} + \nu_{n+1,1})}{8 (\nu_{n,1} - \nu_{n-1,1})} \quad \text{at } \nu_{n,0} \quad (1a) \\
r_{10}(n) &= \frac{(\nu_{n-1,0} + 4\nu_{n-1,1} + 6\nu_{n,0} - 4\nu_{n+1,0} + \nu_{n+1,1})}{8 (\nu_{n+1,0} - \nu_{n,0})} \quad \text{at } \nu_{n,1} \quad (1b) \\
r_{02}(n) &= \frac{\nu_{n,0} - \nu_{n-1,2}}{\nu_{n,1} - \nu_{n-1,1}} \quad \text{at } \nu_{n,0} \quad (1c)
\end{align*}$

Since the ratios for given $n$ have several frequencies in common (eg $\nu_{00}, \nu_{11}$), as do ratios of neighbouring $n$, they are strongly correlated. When comparing model and observed values this requires one to match models and observed values of the ratios using the covariance matrices of the ratios of the observed values. Care needs to be taken to ensure that one includes all the relevant correlations and that one does not overfit the data.

2. Combining $r_{01}$ and $r_{10}$ ratios into a single set $r_{010}$

Several authors combine the observed ratios $r_{01}$ and $r_{10}$ into a single sequence $r_{010}$ (cf Silva Aguirre et al 2013, 2017), which should then be compared with model values using the $r_{010}$ covariance matrix. To demonstrate this can lead to anomalies I show in Fig 1 the fits of the sequences $r_{01}, r_{10}, r_{010}$ for 16 Cyg A as given by the Legacy project in their revised MCMC analysis (Lund et al 2017a,b), to the very slightly different values (LegacyN) obtained directly from Eqns 1 using their frequencies. The fits for all 3 sequences using Legacy errors and for $r_{01}, r_{10}$ using the Legacy covariance matrices have $\chi^2 \sim 10^{-3}$, whereas the fit for $r_{010}$ sequence has $\chi^2 = 4$. As shown in Table 1 similar results are obtained for the 10 stars in common between the Legacy Project and those analysed by myself (Roxburgh 2017).

![Fig. 1. 16CygA: Fits of Legacy project ratios to LegacyN values derived from the definitions and the Legacy frequencies, $\chi^2_{\text{cov}}$, using Legacy errors and $\chi^2_{\text{err}}$ using the Legacy covariance matrices.](image)

Table 2 gives the fits between Legacy ratios and my values of the ratios for these 10 stars, again using the revised Legacy project covariance matrices (Lund 2017b). Several have substantial differences between $\chi^2_{\text{err}}$ for the $r_{010}$ sequence and $\chi^2_{\text{cov}}$ for the $r_{01}$ and $r_{10}$ sequences, while others show modest agreement. The result for 16 Cyg A is illustrated in Fig 2

To identify the cause of this anomalous behaviour I note that the frequencies can be be expressed as (cf Roxburgh 2016)

$\nu_{nl} = \Delta[n + \ell/2 + \epsilon_{l}(\nu_{n\ell})]$, where $\epsilon_{l}(\nu) = \alpha(\nu) - \delta_{l}(\nu)$

$\alpha(\nu)$ is the (almost) $\ell$ independent outer phase shift and the $\delta_{l}(\nu)$ are the $\ell$ dependent inner phase shifts. The differences

$\epsilon_{l}(\nu) = \epsilon_{l}(\nu) - \delta_{l}(\nu)$

$\Delta(\nu_{n\ell}, l)$ which give approximations to the $\delta$ differences

$r_{01} \approx \delta_{1} - \delta_{0}$ at $\nu_{n,0}, \quad r_{10} \approx \delta_{1} - \delta_{0}$ at $\nu_{n,1}$

But these are not independent data.
Table 1. Fits of Legacy ratios and covariances to LegacyN ratios

| KIC no  | $r_{01}$ | $r_{02}$ | $r_{010}$ |
|---------|---------|---------|---------|
| 3427720 | $\chi_{cov}^2$ | 0.0006 | 0.0007 | 0.0424 |
| 6106415 | $\chi_{cov}^2$ | 0.0002 | 0.0003 | 4.1139 |
| 6116048 | $\chi_{cov}^2$ | 0.0006 | 0.0008 | 0.9903 |
| 6225718 | $\chi_{cov}^2$ | 0.0111 | 0.0033 | 2.0442 |
| 6603624 | $\chi_{cov}^2$ | 0.0004 | 0.0004 | 0.0524 |
| 8379927 | $\chi_{cov}^2$ | 0.0111 | 0.0016 | 5.6099 |
| 8760414 | $\chi_{cov}^2$ | 0.0009 | 0.0011 | 0.8123 |
| 9098294 | $\chi_{cov}^2$ | 0.0007 | 0.0007 | 0.1008 |
| 10963065 | $\chi_{cov}^2$ | 0.0004 | 0.0005 | 0.0266 |
| 12069424 | $\chi_{cov}^2$ | 0.0009 | 0.0012 | 4.0505 |
| 12069449 | $\chi_{cov}^2$ | 0.0014 | 0.0006 | 1.4373 |

Table 2. Fits of Legacy ratios and covariances to Roxburgh’s ratios

| KIC no  | $r_{01}$ | $r_{02}$ | $r_{010}$ |
|---------|---------|---------|---------|
| 3427720 | $\chi_{cov}^2$ | 0.345 | 0.411 | 0.545 |
| 6106415 | $\chi_{cov}^2$ | 1.200 | 1.247 | 2.622 |
| 6116048 | $\chi_{cov}^2$ | 0.697 | 0.648 | 1.048 |
| 6225718 | $\chi_{cov}^2$ | 0.905 | 0.806 | 2.786 |
| 6603624 | $\chi_{cov}^2$ | 0.217 | 0.130 | 0.302 |
| 8379927 | $\chi_{cov}^2$ | 0.391 | 0.398 | 14.786 |
| 8760414 | $\chi_{cov}^2$ | 0.751 | 0.660 | 3.836 |
| 9098294 | $\chi_{cov}^2$ | 0.394 | 0.467 | 0.575 |
| 12069424 | $\chi_{cov}^2$ | 1.043 | 1.057 | 8.813 |
| 12069449 | $\chi_{cov}^2$ | 1.467 | 1.576 | 27.518 |

Fig. 2. 16CygA: Fits Legacy–Roxburgh ratios (Leg errors/cov)

A simple illustration of this is to suppose one has $N$ independent values of $e_0(N)$ and $e_1(N)$ at the same frequencies $\nu_i$. Then since $e_0(N) = \alpha(N) - \delta_0(N)$ and $e_1(N) = \alpha(N) - \delta_1(N)$, subtraction eliminates the $N$ values of $\alpha$ leaving $N$ independent values of $\delta_1 - \delta_0$ at the $N$ frequencies $\nu_i$. One could determine additional values of $\delta_1 - \delta_0$ at any $\nu_i$ either by interpolating in the $N e(\nu_i)$’s and subtracting, or equivalently interpolating in the $N \delta_1 - \delta_0$ at $\nu_i$, but this does give any additional independent information.

The fact that with real data one needs to interpolate in either, or both the $e$’s does alter this. In principle the values $\delta_1 - \delta_0$ at $\nu_{n,1}$ can be derived by interpolation in the values at $\nu_{n,0}$, and likewise the values of $r_{01}$ by interpolation in the values of $r_{01}$. In this sense the combined sequence of $r_{010}$ which has $\sim N$ terms is overdetermined.

Since such interpolation is dominantly linear it follows that the ratio $r_{01}(n)$ is strongly correlated with the neighbouring values $r_{01}(n)$ and $r_{01}(n + 1)$, and likewise for all neighbouring triplets. For Legacy 16 Cyg A, $corr(r_0{(19)}, r_0{(19)}) = 0.84$ and $corr(r_0{(19)}, r_0{(20)}) = 0.80$. This in turn can lead to almost singular $r_{010}$ covariance matrices with large condition numbers (eg $\sim 1.6 \times 10^8$ for Legacy 16CygA, to be contrasted with values 883 and 587 for $r_{01}$, $r_{010}$ covariance matrices).

One could argue that the $r_{010}$ sequence is not overdetermined since one is simply comparing 2N frequencies; this is true, but one is comparing particular combinations of frequencies designed to eliminate the contributions of the outer layers and this introduces strong correlations between neighbouring terms which can lead to nearly singular covariance matrices.

As remarked by Lund et al (2017b) in their Erratum this leads to an inverse covariance matrix for $r_{010}$ with very large values oscillating in sign, which can lead to spurious values of $\chi^2$ when comparing 2 sets of ratios. For example for 16 Cyg A the elements on the leading diagonal of the inverse covariance matrix are all positive with values up to $1.6 \times 10^{11}$ whereas the elements on the neighbouring diagonals have similar values but are all negative.

To illustrate that this behaviour is not just due to the properties of the Legacy covariance matrices I show in Table 3 the comparisons for all 10 stars using my ratio covariance matrices. These show similar (but different) behaviour to those in Table 2.

Using the combined series $r_{010}$ therefore can, (but may not) give spurious results. One should avoid this possibility by comparing only one of the $r_{01}$ or $r_{10}$ sequences. As discussed below this should be combined with the $r_{02}$ ratios.

3. The combined sequence $r_{012}$ or $r_{102}$

As is clear from the definitions of $r_{02}$, $r_{10}$ (or $r_{01}$) in Eqns 1 these ratios are strongly correlated as they have several frequencies in common, one should therefore combine $r_{01}, r_{02}$ into a combined sequence $r_{102}$ and compare observed and model values using the covariance matrix of the combined set. Or equally combine the ratios $r_{01}, r_{02}$ into a combined set $r_{012}$, but not both. Such combined sequences are not overfitted as from 3N values of $\chi_{0,\ell}^2, \ell = 0, 1, 2$, one can determine $\sim N$ independent values of each of the differences $\delta_1 - \delta_0$, and $\delta_2 - \delta_0$. The resulting covariance matrices have small condition numbers, eg for 16CygA $1.7 \times 10^3$ for $r_{102}$ as compared to $1.6 \times 10^8$ for $r_{010}$. The $r_{02}$ fits for 16 Cyg A are shown in Fig 3.
Since the Legacy project does not give $r_{02}$ or $r_{012}$ covariance matrices I generated these from the frequencies $v_k$ and frequency covariance matrices $\text{cov}(v_k, v_m)$ as given in Lund (2017b), the same procedure as was used to generate my covariance matrices. The ratios and their derivatives with respect to the frequencies follow directly from the definitions in Eqns 1 and the covariance matrices. In this sense the Legacy project has $N$ components. In this sense they are approximated by the ratios $r_{01}$, $r_{02}$, and $r_{012}$. This is a reflection of the frequency differences and considerably smaller errors on the low frequencies from my MLE analysis as compared to those from the Legacy project’s MCMC analysis (see, for example, tables A1-A3 in Roxburgh 2017).

4. Conclusions

The ratios $r_{01}$, $r_{10}$ are not independent and in principle one set can be derived from the other by interpolation. From $N + 1$ frequencies one can only derive $\sim N$ independent values of the phase shift differences $\delta_1 - \delta_0$ (which are approximated by the ratios). But the $r_{010}$ sequence has $\sim 2N$ components. In this sense one is overfitting the data. Neighbouring elements of the covariance matrix can then be very strongly correlated leading to almost singular matrices with large condition numbers, and hence spurious results when comparing 2 sets of ratios. Only one of $r_{01}$, $r_{10}$ should be used in model fitting.

Table 4. Fits of LegacyN ratios and covariances to Roxburgh’s ratios

| KIC no  | $r_{01}$ | $r_{02}$ | $r_{012}$ | $r_{10}$ | $r_{102}$ | $r_{012}$ |
|---------|---------|---------|---------|---------|---------|---------|
| 3427720 | $\chi^2_{\text{cov}}$ 0.427 | 0.772 | 0.582 | 0.580 |
| 6106415 | $\chi^2_{\text{cov}}$ 1.198 | 2.507 | 2.206 | 2.188 |
| 6116048 | $\chi^2_{\text{cov}}$ 0.657 | 0.597 | 0.562 | 0.567 |
| 6225718 | $\chi^2_{\text{cov}}$ 0.830 | 1.397 | 1.060 | 1.042 |
| 6603624 | $\chi^2_{\text{cov}}$ 0.138 | 1.926 | 1.162 | 1.226 |
| 8379927 | $\chi^2_{\text{cov}}$ 0.368 | 2.958 | 1.576 | 1.544 |
| 8760414 | $\chi^2_{\text{cov}}$ 0.661 | 2.831 | 1.742 | 1.550 |
| 9098294 | $\chi^2_{\text{cov}}$ 0.481 | 0.244 | 0.430 | 0.402 |
| 12069424 | $\chi^2_{\text{cov}}$ 1.059 | 0.857 | 0.919 | 0.947 |
| 12069449 | $\chi^2_{\text{cov}}$ 1.577 | 1.022 | 1.360 | 1.225 |

Table 5. Fits of Roxburgh’s ratios and covariances to LegacyN ratios

| KIC no  | $r_{01}$ | $r_{02}$ | $r_{012}$ | $r_{10}$ | $r_{102}$ | $r_{012}$ |
|---------|---------|---------|---------|---------|---------|---------|
| 3427720 | $\chi^2_{\text{cov}}$ 0.489 | 0.683 | 0.544 | 0.536 |
| 6106415 | $\chi^2_{\text{cov}}$ 1.270 | 2.762 | 2.259 | 2.361 |
| 6116048 | $\chi^2_{\text{cov}}$ 0.950 | 0.735 | 0.783 | 0.793 |
| 6225718 | $\chi^2_{\text{cov}}$ 0.663 | 1.045 | 0.859 | 0.924 |
| 6603624 | $\chi^2_{\text{cov}}$ 0.162 | 4.139 | 2.591 | 2.798 |
| 8379927 | $\chi^2_{\text{cov}}$ 0.441 | 1.069 | 0.701 | 0.646 |
| 8760414 | $\chi^2_{\text{cov}}$ 1.938 | 10.322 | 5.557 | 5.523 |
| 9098294 | $\chi^2_{\text{cov}}$ 0.564 | 0.268 | 0.479 | 0.425 |
| 10963065 | $\chi^2_{\text{cov}}$ 1.524 | 1.821 | 1.535 | 1.496 |
| 12069424 | $\chi^2_{\text{cov}}$ 2.607 | 6.608 | 4.697 | 4.790 |
| 12069449 | $\chi^2_{\text{cov}}$ 1.935 | 6.090 | 4.883 | 3.965 |

Since the $r_{02}$, $r_{10}$ and $r_{01}$ ratios are correlated they should be combined into single sequence $r_{010}$ or $r_{102}$ when comparing 2 sets of ratios. These sequences are not overfitted since from $N + 1$ frequencies subtraction gives $\sim N$ values of both $\delta_1 - \delta_0$ and $\delta_2 - \delta_0$, which are approximated by the ratios $r_{01}$, $r_{02}$. The $r_{012}$ covariance matrices have reasonable condition numbers.

References

Lund M N, Silva-Aguirre V , Davies G R, et al, 2017a, ApJ, 835, 172
Lund M N, Silva-Aguirre V , Davies G R, et al, 2017b, ApJ, 850,110
Roxburgh I W, 2016, A&A , 585, A63
Roxburgh I W, 2017, A&A , 604, A42
Roxburgh I W, Vorontsov S V, 2003, A&A , 411, 215
Roxburgh I W, Vorontsov S V, 2013, A&A , 560, A2
Silva Aguirre V , Basu S, Brandao I M, et al, 2013, ApJ, 769, 141
Silva Aguirre V , Lund M N, Antia H M, et al 2017 ApJ , 835, 173