AUTOMATED p-MODE IDENTIFICATION USING BAYES' THEOREM

Timothy M. Brown
High Altitude Observatory/National Center for Atmospheric Research*
P.O. Box 3000
Boulder, CO 80303
U.S.A.

ABSTRACT. The task of interpreting p-mode spectra is complicated by the presence of a very large number of oscillation modes, each of which may appear (because of aliasing) in the power spectra corresponding to several values of l and m. Identifying peaks in a power spectrum with particular modes in an interactive fashion thus quickly becomes impractical. Here I describe an automated method for doing this identification. The method is based on an application of Bayes' theorem, which provides a simple way to use prior knowledge about the oscillation spectrum. The method takes as input the observed power spectra, and a model of the amplitudes and frequencies one expects to see.

1. BASIC IDEAS ABOUT PROBABILITY

This introduction is based on the excellent discussion by R.T. Cox (1961). Suppose that A, B, C are statements that may be either true or false. Then

$$P(A \mid C)$$

is defined as the probability that A is true, given that C is true. Then all the usual formulae of probability theory may be derived from two relations:

$$P(A \cdot B \mid C) = P(A \mid C)\, P(B \mid A \cdot C) \qquad (\textit{product rule})$$

$$P(A \wedge B \mid C) = P(A \mid C) + P(B \mid C) - P(A \cdot B \mid C) \qquad (\textit{sum rule})$$

where $\cdot$ denotes logical conjunction (AND), and $\wedge$ denotes the inclusive OR. In particular, Bayes' theorem comes from the product rule, and the fact that $A \cdot B = B \cdot A$:

$$P(A \cdot B \mid C) = P(B \cdot A \mid C)$$

$$P(A \mid C)\, P(B \mid A \cdot C) = P(B \mid C)\, P(A \mid B \cdot C)$$

Leading to the final result:

$$P(A \mid B \cdot C) = P(A \mid C) \frac{P(B \mid A \cdot C)}{P(B \mid C)}$$

*The National Center for Atmospheric Research is sponsored by the National Science Foundation

J. Christensen-Dalsgaard and S. Frandsen (eds.), Advances in Helio- and Asteroseismology, 491–494.

Here we may interpret A as some *inference* in which we are interested, C is our *initial information,* on which we base our initial estimate of the probability that A is true, and B is some *new information,* for example something we have observed.

A useful special case of Bayes' theorem occurs if one is concerned with a number of inferences A_i that are *exhaustive* and *mutually exclusive,* so that one and only one of the A_i can be true. Then one can show that

$$P(B \mid C) = \sum_i P(B \mid A_i \cdot C)\, P(A_i \mid C) \quad ,$$

and Bayes' theorem becomes

$$P(A_i \mid B \cdot C) = \frac{P(B \mid A_i \cdot C) P(A_i \mid C)}{\sum_i P(B \mid A_i \cdot C) P(A_i \mid C)} \quad .$$

2. PEAK IDENTIFICATION STRATEGY

All of the above suggests the following *Strategy for peak identification:*

For each observed peak in a power spectrum:

(1) Identify a (hopefully small) set of *possible causes* A_i for the peak in question.

(2) Assign *a priori* probabilities to each of the A_i, in whatever way seems reasonable (*e.g.*, based on relative frequencies of occurrence in the spectrum taken as a whole).

(3) For each *observed fact* B_j about the peak, and for each A_i, estimate the probability that the observation could have occurred given that A_i is true. *i.e.*, estimate

$$P(B_j \mid A_i \cdot C) \quad .$$

(4) Use Bayes' theorem to *update* the probabilities for each A_i.

(5) Repeat (3) and (4) until all the observed facts are exhausted.

(6) If, for some i, $P(A_i \mid B_1 \cdot B_2 \cdots B_n \cdot C)$ is large enough, identify A_i as the *cause* of the observed peak.

To implement this strategy, one must enumerate the possible causes for peaks, which requires a model of the expected frequencies and amplitudes of oscillation. One must also allow for peaks that do not correspond to oscillation modes, *e.g.* 1/day sidelobes and noise peaks. the model I use provides for the following sources:

(1) Oscillation modes with l and m near the nominal values for that spectrum (within parity constraints). Frequencies are determined from a Duvall-law fit (with frequency-dependent α) to previous observations (Duvall 1982). Amplitudes depend on ν, and on the expected response to the given l and m.

(2) Sidelobes of all oscillation modes, with frequency separations that are multiples of 1/day. Relative amplitudes of the sidelobes are determined from the observation window function.

(3) Noise peaks, with typical amplitudes assumed independent of ν.

https://doi.org/10.1017/S0074180900158590 Published online by Cambridge University Press

(4) Sidelobes with typical frequency spacings of $1/T$, where T is the total observing run length.

Currently, the observed facts used to update probabilities are the following:

(1) $\frac{Power_{peak}}{Power_{model}}$: One can estimate probablility distributions for this ratio for each of the possible causes, giving the probability that the peak height is equal to or greater than the observed power for each A_i. For power spectra, these are appropriately modeled as exponential distributions.

(2) $\nu_{peak} - \nu_{model}$: The probability distribution for this difference depends on errors both in the observations and in the frequency model. Currently, model errors dominate. I use a Gaussian distribution, with a width based on the scatter around the Duvall-law fit.

(3) *Neighboring Power:* Define a "sidelobe index"

$$s \equiv \int_{-\infty}^{-\epsilon} \frac{\Pi(\delta\nu)}{(T\,\delta\nu)^2} d\,\delta\nu + \int_{\epsilon}^{\infty} \frac{\Pi(\delta\nu)}{(T\,\delta\nu)^2} d\,\delta\nu \quad ,$$

i.e., a weighted mean of the power Π in the neighborhood of the observed peak, but not including it. Then one may estimate the probability of observing a sidelobe with power S as a function of the ratio S/s.

After the $P(A_i)$ have been updated to reflect all of the available information, one has final estimates of the probabilities that the observed facts could have been produced by each of the assumed causes. If one inference has a large enough probability of being correct (I typically choose a threshold $P > 0.95$), then that identification of the power spectrum peak is adopted.

Because the spatial filter responses overlap in l, and because 1/day sidelobes are always present, I then insist that each mode be confidently identified several times on several different spectra. The final estimates of mode frequency and amplitude are obtained by averaging the results for the various identifications.

3. RESULTS AND COMMENTS

This technique (embodied in a program named IDMODE) appears to work reasonably well. Like most techniques, it has advantages and drawbacks.

Among its advantages are:

(1) When the data are good and the probability distributions have been chosen appropriately, it identifies modes correctly, at least as far as one can judge by eye.

(2) It refuses to commit itself to a mode identification unless the data warrant commitment. This behavior is particularly apparent for large degrees, where the overlap between sidelobes of successive l values is particularly severe.

(3) It behaves quite poorly if the probability distributions are poorly chosen. This calls direct attention to the importance of correctly estimating errors, and forces one to think hard about definitions, expectations, and uncertainties. All of these are good things.

(4) The form of the output data (estimated frequencies and amplitudes as functions of n, l, m) encourage statistical analysis, novel display methods, and comparison with theory.

The method principal drawback is that it tends to see what it expects to see, in the sense that the mode identifications cannot differ by very much from the parameters allowed by the frequency and amplitude model. A subtle effect of this problem is that regular behavior not parameterized in the model tends to be underestimated. The only guaranteed solution to this problem is to correct the model to reflect the quantities one wished to measure. However, it is worth noting that the Bayes' Theorem method closely mimics the thought processes of a person trying to identify power spectrum peaks by hand. I therefore speculate that similar bias exists in other methods of mode identification; an automated technique simply makes it easier to examine many cases, and thereby test for such problems.

REFERENCES

Cox, R.T. 1961, *The Algebra of Probable Inference*, The Johns Hopkins Press, Baltimore.
Duvall, T.L. 1982, *Nature*, **300**, 242.

https://doi.org/10.1017/S0074180900158590 Published online by Cambridge University Press