UNIVERSITY OF CALIFORNIA, SAN DIEGO

Applications of global seismic tomography and analysis of variational methods for the solution of the linearly attenuating frequency domain wave equation

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Earth Sciences

by

Stuart G. Johnson

Committee in charge:

Professor T. Guy Masters, Chairperson
Professor J. Freeman Gilbert
Professor Robert L. Parker
Professor Glenn R. Ierley
Professor Randolph E. Bank

1997

The dissertation of Stuart G. Johnson is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

1997

To my unwitting benefactor,
the Taxpayer of the United States of America

Table of Contents

List of Tables

List of Figures

ACKNOWLEDGMENTS

I would like to thank a number of individuals and groups for their friendship, good advice, quality work and life-saving talents. Vera Schulte-Pelkum, Keith Richards-Dinger, Harm van Avendonk, Harold Bolton and Paul Earle have been good friends and made the Barnyard computer room a great place to work. Megan Flanagan is too sweet for words. Mike Ravine has been an excellent source of carrots.

Glenn Ierley sold me on Chebyshev polynomials. Randy Bank suggested the H^0 FEM method. Bob Parker was always available for encouraging and energetic conversations on various thesis topics. Steve Shkoller has been the prime inspiration for the level of detail in chapter 4. After attempting to describe the problem to Steve, whose expertise in linear and non-linear PDEs is replete, I felt so stupid that I checked out enough books on the subject to produce the background material on elliptic PDEs in that chapter. Guy Masters suggested several avenues to explore, even if they led away from the direction of his research. His willingness even to pay me to do this is unfathomable. The content of chapter 4 is a direct outcome of a question he asked about someone else's research. Michael O'Brien was the first to *pooh-pooh* the eigenvalue method of solving the normal equations, leading me off on the parallel QR trail. Steve Constable invited me along on several research cruises, a great experience which no one at IGPP should go without. He also once gave me more allocation on the Cray by dial–up from home (over dinner, I think). Paul Henkart provided code and know-how with regard to numeric formats, since the odd supercomputer is often *other–endian.* The Netops crew of Ed Riesgo, Pat Russo and Chris Garrod have energetically kept the computers and network afloat. I also thank various nameless individuals at the San Diego Supercomputer Center for their help over the last few years. Since much of my computational work came down to linear algebra, I thank the teams which designed and engineered the LAPACK and SCALAPACK code for vector and parallel architectures.

I especially thank the San Diego Lifeguard Service for rescuing me and my sailboat from a nasty little storm off La Jolla. It would have been, at best, a very cold swim back. Thanks also to John B. Chandler of Austin, Texas and Eldon K. Shipp of Dripping Springs, Texas for good advice given and acted upon.

The text of chapter 2, in part, appeared in Masters, G., S. Johnson, G. Laske, and H. Bolton,1996, A shear velocity model of the mantle, *Phil. Trans. R. Soc. Lond.*, **354A**, 1385–1411. The dissertation author was a secondary researcher and Guy Masters directed and supervised the research which forms the basis for that chapter. The dissertation author was responsible for the process of constructing the shear velocity model S16B30 based on data collected and evaluated by the other authors.

VITA

March 19, 1963	Born, Austin, Texas
1991	B.S. Physics, University of Texas at Austin
1991	B.S. Geophysics, University of Texas at Austin
1992–1997	Research Assistant, Scripps Institution of Oceanography University of California, San Diego
1997	Doctor of Philosophy University of California, San Diego

PUBLICATIONS

Johnson, S. 1997 Convergence and computational efficiency of H^0 and H^1 spectral finite element solutions of the linearly attenuating frequency domain wave equation. *Geophysical Journal International*, in preparation.

Masters, G., Johnson, S., Laske, G., & Bolton, H. 1996 A shear-velocity model of the mantle. *Phil. Trans. R. Soc. Lond. A* **354**, 1385 – 1411.

Kuehne, J., Johnson, S. & Wilson, C. 1993 Atmospheric excitation of non-seasonal polar motion. *Journal of Geophysical Research* **98,11**, 19,973 – 19,978.

ABSTRACTS

Johnson, S., 1996. 1-D frequency domain Galerkin weak form solutions of the forced attenuating elastic wave equation. *EOS*, **77**.

Johnson, S., Masters, G., Shearer, P., Laske, G. and Bolton, H., 1994. A shear velocity model of the mantle. *EOS*, **75**.

Johnson, S., Masters, G., Tackley, P. & Glatzmaier, G., 1993. How well can we resolve a convecting Earth with seismic data?. *EOS*, **74**.

ABSTRACT OF THE DISSERTATION

Applications of global seismic tomography and analysis of variational methods for the solution of the linearly attenuating frequency domain wave equation

by

Stuart G. Johnson
Doctor of Philosophy in Earth Sciences
University of California, San Diego, 1997
Professor T. Guy Masters, Chair

Two distinct undertakings are made in this dissertation. First, the use of seismic tomography to investigate the internal structure of the Earth is examined. A tomographic image of seismic shear-wave velocity variations in the Earth's solid mantle is generated. Hypothetical seismic velocity variations in the Earth's mantle are obtained from the simulation of convection in a spherical shell, and the quality of tomographic imaging thereby assessed. It is found that the convection simulation does not, in detail, produce Earth-like structure within the mantle.

The second undertaking has as its motivation an improved understanding of the relationship of seismic observables to Earth structure. The variational solutions of the linearly attenuating frequency domain wave equation (LAFDWE) are studied. Some practical guidelines pertaining to the numerical approximations resulting from these variational methods are derived. A thorough discussion of elliptic partial differential equations is included to make concise the nature of the solutions of the LAFDWE in the presence of singular sources of typical use in seismology.

Chapter 1

Introduction

This dissertation investigates two aspects of the same endeavor. This endeavor is the attempt to constrain the internal structure of the Earth using recordings (seismograms) of vibrations at the surface. These vibrations are induced by earthquakes and have useful periods extending from one second, where they behave more or less as rays, to thousands of seconds, where they behave more or less as modes.

In trying to say something about the Earth given some set of observables, we must solve two problems: the forward problem and the inverse problem. Given a model of the Earth, that is, a specification of some of its physical parameters, the forward problem generates observables. These observables are seismograms or quantities derived from seismograms. The aim of the inverse problem is to estimate some properties of the Earth given the observables.

Chapters 2 and 3 assume the validity of fairly simple, "first-order" approximations of the forward problem. These assumptions are warranted if the Earth is dominated by low-amplitude, long-wavelength structure, and there is much evidence from long-period seismology and systematic patterns in traveltime data that this is the case. As well, the work in these two chapters formulates the inverse problem as a search for an Earth model which fits the data at some level of tolerance but also minimizes a measure of model roughness. This results in a practical means of generating models of the Earth's seismic velocity structure: tomography. At present, several authors are using tomography with somewhat different forward problems, data, and inverse problems to obtain generally similar results (Li & Romanowicz, 1996; Masters, et al., 1996; Su, et al., 1994; Woodward, et al., 1993).

Chapter 2 utilizes large data sets and tomographic methods to obtain an image of seismic velocity variations in the mantle of the Earth (see figure 2.13). The resulting model is dominated by large-scale structure and exhibits high variability only near the surface and the core-mantle boundary. The geodynamics community has become capable in the last few years of the simulation of convection in spherical shells with many physical properties believed to exist in the actual Earth (Tackley, et al., 1993; Tackley, 1993; Tackley, et al., 1994; Tackley, 1996a,b; Bunge, et al., 1996a,b; Bunge & Richards, 1996). These properties include abrupt radial changes in the crystallography of the Earth (referred to as phase changes), variations in viscosity and the presence of realistic tectonic plates at the surface. The tomographic image of the mantle obtained in chapter 2 is at odds with simulations of convection in homogeneous spherical shells, which exhibit spherical harmonic spectra with large amplitudes to very high degree. Understanding in detail what factors in the actual Earth are responsible for this difference is an important part of understanding the dynamics and evolution of the Earth.

The evaluation of tomographic imaging and the comparison of the *observable* Earth to convection results are the immediate tasks in combining tomography and 3-D geodynamic modelling. Chapter 3 utilizes the data sets and methods of chapter 2 to perform synthetic tomography experiments on two time snapshots of convecting mantle simulations. These simulations of flow in a convecting spherical shell have as their important features some combination of phase changes (at

410 and 670km depth) and a viscosity jump at 670km depth. Chapter 3 shows that while these convection simulations are somewhat Earth-like in that they are dominated by long-wavelength structure, the actual mantle of the Earth is fairly different: tomography of one simulated mantle predicts that we should see strong changes in velocity structure at 670km depth, and the other predicts relatively low amplitude structure associated with a simple pattern of upwelling throughout the entire mantle. Tomography of the Earth reveals no such features. This suggests that more research is required to understand the important controlling factors of mantle convection. Indeed, since the early convection model tested in this work was obtained, many improvements and additional features have been included in geodynamic simulations. These additional features include more realistic Rayleigh numbers and rigid surface plates (Mégnin, et al., 1997).

Since the answer to the inverse problem is known for the tomography of the synthetic convecting mantle, the tomographic process itself can be evaluated. The process of obtaining models in chapter 3 indicates that while models very much like the Earth probably result from tomography, it is not an easy task to choose amongst these models and declare, “this is the Earth”. Nevertheless, some characteristics of a large group of such models may be more robust. This suggests that we should attempt to determine these characteristics more directly from the observables, in a process more akin to hypothesis testing. It is not clear, however, how to actually go about this given the size and complexity of even the most simplified global inverse problems.

Chapter 4 changes gears considerably, and has as its motivation the study of the relationship of the details of seismograms to variations in elastic properties of an Earth model. Our ability to exploit the information in seismograms is limited by our ability to compute accurate wavefields in complicated media. To make progress, we often use approximations whose validity is hard to assess in the absence of complete calculations of the wavefield. Calculation of the full wavefield in Earth-like media, however, remains one of the great computational challenges. There are two general computational approaches to solving the appropriate equations: finite difference methods and variational methods. Finite difference methods (and pseudo-spectral methods) are the result of discretizing the differential equation, replacing differentiation with estimated derivatives. In domains with simple boundaries these estimated derivatives can be computed very efficiently. In domains with complicated boundaries and interfaces, less efficient estimates of the derivatives must be used. Alternatively, variational (or weighted-residual or Galerkin) methods are based on an integral formulation of the differential equation. This formulation can be used, via one or more integrations by parts, to implicitly include boundary conditions in a redefinition of the differential equation. This makes the task of obtaining solutions in complicated regions much easier. We can design highly accurate approximation schemes using higher order polynomials in subdomains (spectral elements) which can be used to pave any solution domain.

Chapter 4 analyzes the variational solutions of a form of the wave equation. This form is valid if the medium can be characterized as exhibiting linear attenuation mechanisms over some band of frequencies (Liu, et al., 1976). We will refer to this problem as the linearly attenuating frequency domain wave equation or LAFDWE. An examination of the nature of the solution of the LAFDWE is made from first principles. This involves casting the problem in the context of the theory of elliptic partial differential equations (PDE's). An elegant way of understanding the PDE as a mapping between vector spaces of distributions with clearly defined smoothness properties (Babuška & Aziz, 1972; Roitberg, 1996) is revealed. An examination of the convergence of two types of variational approximations to the solution follows. A surprising result about the nature of solutions to the LAFDWE in the presence of typical seismic sources is found as well. In particular, when the source is a derivative of the delta distribution (as earthquakes and explosive sources are typically modelled), neither the solution nor its first derivatives are square integrable in dimensions higher than 1.

Conclusions and prospects for further research are contained in the appropriate sections of each chapter.

1.1 References

Babuška, I. & Aziz, A.K. 1972 Survey lectures on the mathematical foundations of the finite element method. *in, The mathematical foundations of the finite element method with applications to partial differential equations. Academic Press, New York* , 1-359.

Bunge, H.-P., Richards, M. & Baumgardner, J. 1996 Effect of depth-dependent viscosity on the planform of mantle convection. *Nature* **379**, 436–438.

Bunge, H.-P., Richards, M. & Baumgardner, J. 1997 A sensitivity study of 3-D spherical mantle convection at 10^8 Rayleigh number: Effects of depth-dependent viscosity, heating mode and an endothermic phase change. *J. Geophys. Res.* **102**, 11991–12007.

Bunge, H.-P. & Richards, M. 1996 The origin of long wavelength structure in mantle convection: effects of plate motion and viscosity stratification. *Geophy. Res. Letters* **23**, 2987–2990.

Li, X. & Romanowicz, B. 1996 Global mantle shear-velocity model developed using nonlinear asymptotic coupling theory. *Geophys. J. R. astr. Soc.* **101**, 22245 – 22272.

Liu, H., Anderson, D.L. & Kanamori, H. 1976 Velocity dispersion due to anelasticity; implications for seismology and mantle composition. *Geophys. J. R. Astr. Soc.* **47**, 41 – 58.

Masters, G., S. Johnson, G. Laske, and H. Bolton 1996 A shear velocity model of the mantle. *Phil. Trans. R. Soc. Lond.* **354A**, 1385–1411.

Mégnin, C., Bunge, H.-P., Romanowicz, B. & Richards, M. 1997 Imaging 3-D spherical convection models: What can seismic tomography tell us about mantle dynamics?. *Geophysical Research Letters* **24**, 1299–1302.

Roitberg, Yakov 1996 *Elliptic Boundary Value Problems in the Spaces of Distributions.* Kluwer Academic Publishers, Dordrecht.

Su, W-J., Woodward, R.L. & Dziewonski, A.M. 1994 Degree 12 model of shear velocity heterogeneity in the mantle. *J. Geophys. Res.* **99**, 6945–6980.

Tackley, P., Stevenson, D., Glatzmaier, G. & Schubert, G. 1993 Effects of an endothermic phase transition at 670 km depth on a spherical model of convection in the Earth's mantle. *Nature* **361**, 699–704.

Tackley, P. 1993 Effects of strongly temperature-dependent viscosity on time-dependent, 3-dimensional models of mantle convection. *Geophys. Res. Letters* **20**, 2187 – 2190.

Tackley, P., Stevenson, D., Glatzmaier & Schubert, G. 1994 Effects of multiple phase transitions in a three-dimensional spherical model of convection in Earth's mantle. *J. Geophys. Res.* **99**, 15877 – 15901.

Tackley, P. 1996 On the ability of phase transitions and viscosity layering to induce long wavelength heterogeneity in the mantle. *Nature* **23**, 1985–1988.

Tackley, P. 1996 Effects of strongly variable viscosity on three-dimensional compressible convection in planetary mantles. *J. Geophys. Res.* **101**, 3311 – 3332.

Woodward, R.L., Forte, A.M., Su, W. & Dziewonski, A.M. 1993 Constraints on the large-scale structure of the mantle. In: *Geophys. Monograph 74, IUGG,* **v14***, Takahashi et al., eds. , AGU, Washington DC*, pp.89–109.

Chapter 2

A shear-velocity model of the mantle

2.1 Abstract

We present a new model of shear velocity structure in the mantle which is designed to fit data sets of absolute and differential body-wave traveltimes, surface-wave phase velocities over a broad range of frequencies, and structure coefficients of modes of free oscillation. The model is parameterized laterally by spherical harmonics (truncated at degree 16) and by 30 natural cubic B-splines in radius. Our best model features large-amplitude structure (up to ±6% anomalies in shear velocity) in the topmost 400 km of the mantle and in the lowermost 500 km (up to ±2.5%) with most of the power in the low harmonics ($\ell < 6$). The middle of the mantle is characterized by low-amplitude anomalies with a much whiter spectrum. Our models generally show no distinctive changes in structure (in either shape or amplitude) at the 660 km discontinuity supporting the idea that an endothermic phase transition is not a barrier to large-scale flow in the mantle.

2.2 Introduction

Many branches of geophysics would benefit from an accurate map of the three-dimensional structure of the mantle. Geochemists are interested in the correlation of isotopic anomalies with the seismic models, and geodynamicists are interested in using the models to constrain viscosity structure and characteristics of mantle flow. Seismology itself benefits by the improvement in realism in modeling wave propagation which then allows more detailed studies of the seismic source. While we are a long way from having an accurate global picture of mantle structure, it is now apparent that the spectrum of heterogeneity in the Earth is very red and it is reasonable to expect that accurate maps of the largest-scale structure are within our reach.

High-quality, long-period digital seismic data are now available for a timespan covering 15 years or more. Such data have been used in the past to construct models of mantle structure (Masters, et al., 1982; Woodhouse & Dziewonski, 1984; Nataf, et al., 1986; Dziewonski & Woodhouse, 1987; Tanimoto, 1988, 1990; Su & Dziewonski, 1991). Long-wavelength upper mantle structure has been quite well-determined since the early 80's (Woodhouse & Dziewonski, 1984) but it is only in the last two or three years that sufficiently large data sets have been constructed so that the largest-scale features are becoming well-resolved throughout the mantle (Woodward, et al., 1993; Masters, et al., 1992; Su, et al., 1994). Most of these studies have concentrated on the S-wave velocity structure (as we do here). P-wave structure has been mainly constrained using the short-period ISC traveltime data set (Dziewonski, 1984; Inoue, et al., 1990; Hager & Clayton, 1989) though, recently, long-period P-wave data sets have also been developed (Bolton & Masters, 1991) and reliable joint inversions for P and S structure will shortly be available.

The long-period instruments of the Global Digital Seismic Network (GDSN) pass signals with a dominant period of about 20 seconds. Such waves have horizontal wavelengths of 100 to 200 km and their traveltime residuals exhibit much greater spatial coherence than the short period (about 1 Hz) data that make up the bulk of the large catalogs of traveltimes. We have experimented with using later-arriving phase data from the ISC data set but generally find that such strong averaging is required to reveal any coherent large-scale signal that the effective geographical coverage of an ISC data set is usually poorer than that of the equivalent GDSN data set. Later arrivals in long-period data are usually much clearer than their short-period counterparts and the long-period stacks of Shearer (1990, 1991a,b) show many possible phases which have not yet been thoroughly analyzed on a global scale. In short, the existing long-period body-wave data set is an extremely rich one and provides many opportunities for significantly improving our knowledge of Earth structure. Inspection of all long-period recordings of earthquakes with body wave magnitude greater than 5.5 from 1976 to 1987 has resulted in a data set of over 45,000 usable transverse component seismograms. Since 1987, there has been a significant expansion of the global network with the addition of the GEOSCOPE network and the various components of the IRIS program adding to the global coverage. Data from other arrays such as POSEIDON and MEDNET are also becoming available. The new stations, particularly in the southern hemisphere, promise much more uniformity in the sampling of the Earth than has been available to date.

This paper describes the current status of our research in developing data sets of long-period body-wave differential and absolute traveltimes, surface-wave phase-velocity maps, and free-oscillation structure coefficients and using them to recover global-scale models of mantle structure. Given the rapid expansion of the global networks, it is unlikely that this will be the final word on the subject.

2.3 The Data sets

2.3.1 Absolute S times

The main feature which distinguishes the current model from others in the literature is that the inclusion of a new data set of long-period absolute S times with lower-mantle turning points significantly improves our resolution of lower-mantle structure. The measurement technique uses a cross-correlation of the data with a “synthetic seismogram”. The synthetic is taken to be the impulse response of the instrument convolved with a t^* operator (Doornbos, 1983). We concentrate on matching the first swing of the waveforms (Fig. 2.1) so that depth phases are unimportant for all but the shallowest earthquakes. Experiments with synthetic seismograms constructed using mode summation techniques indicate that the measurement algorithm is capable of a precision of 0.1 seconds. A subjective estimate of a typical measurement precision is about 0.5 seconds which is very competitive with other traveltime catalogs. The polarity and amplitude of the arrival is recorded as well as the arrival time. Comparisons of the observed polarities with the predictions of the Harvard CMT solutions shows that the polarity is correctly predicted 88% of the time and suggests that half-cycle errors (~ 12 s) are rare. About a third of the measurements with incorrectly predicted polarities are calculated to be near-nodal and small changes in focal mechanism can accommodate these. Yet another third are associated with source mechanisms which incorrectly predict the majority of observed polarities (about 1% of the CMT catalog) and the final third (about 4% of our total data set) appear to be mispicked. We have collected a total of 23,000 S arrivals though we have analyzed only lower-mantle turning rays in detail (about 18,000 S arrivals).

The data are first converted to ellipticity-corrected residuals relative to the PREM spherical Earth reference model (Dziewonski & Anderson, 1981) using NEIC (National Earthquake Information Center) event locations. The raw resid-

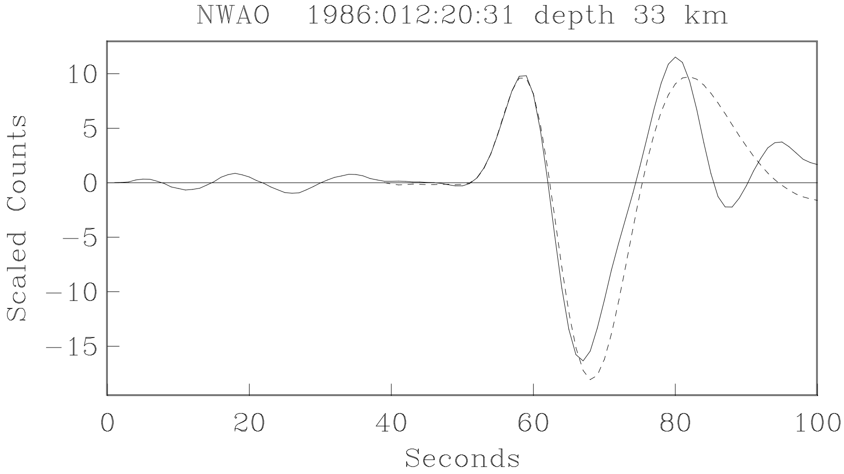


Figure 2.1: An example of the fit of a "synthetic" S arrival to a transverse component seismogram recorded at the SRO station NWAO.

uals show strong distance and event-depth dependent trends which we interpret as being due to inadequacies in the assumed t^* operator and the use of NEIC locations with the PREM model. With this much data, it is simple to empirically determine these trends and remove them from the data set. Histograms of residuals for the detrended S data are shown in Fig. 2.2 for a variety of turning-point depth ranges. Note the broadening of the S wave histograms for rays which turn in the lowermost mantle. If we plot the residuals of the deepest-turning rays at the turning point of the ray and lightly smooth the resulting map using a running mean in a spherical cap of radius 5°, we obtain the result shown in Fig. 2.3. The data show a remarkably coherent large-scale pattern which is in extremely good agreement with equivalent results for $ScS-S$ differential times shown in Fig. 2.4. (These $ScS-S$ differential times are an enlarged version of the data set described by Woodward & Masters (1991a).) Figures 2.3 and 2.4 provide direct evidence for large-scale structure in shear velocity near the base of the mantle. The pattern in the data is one that implies a distribution of fast velocity anomalies near the core-mantle boundary under the rim of the Pacific. This pattern is similar to that seen in an early P-wave tomography model (L02.56 of Dziewonski (1984)).

It is also interesting to note that the amplitude of the traveltime anomalies in the S-wave data is larger than that in the $ScS-S$ data implying that shear velocity structure in the mid-mantle is mildly correlated with the structure at the base of the mantle.

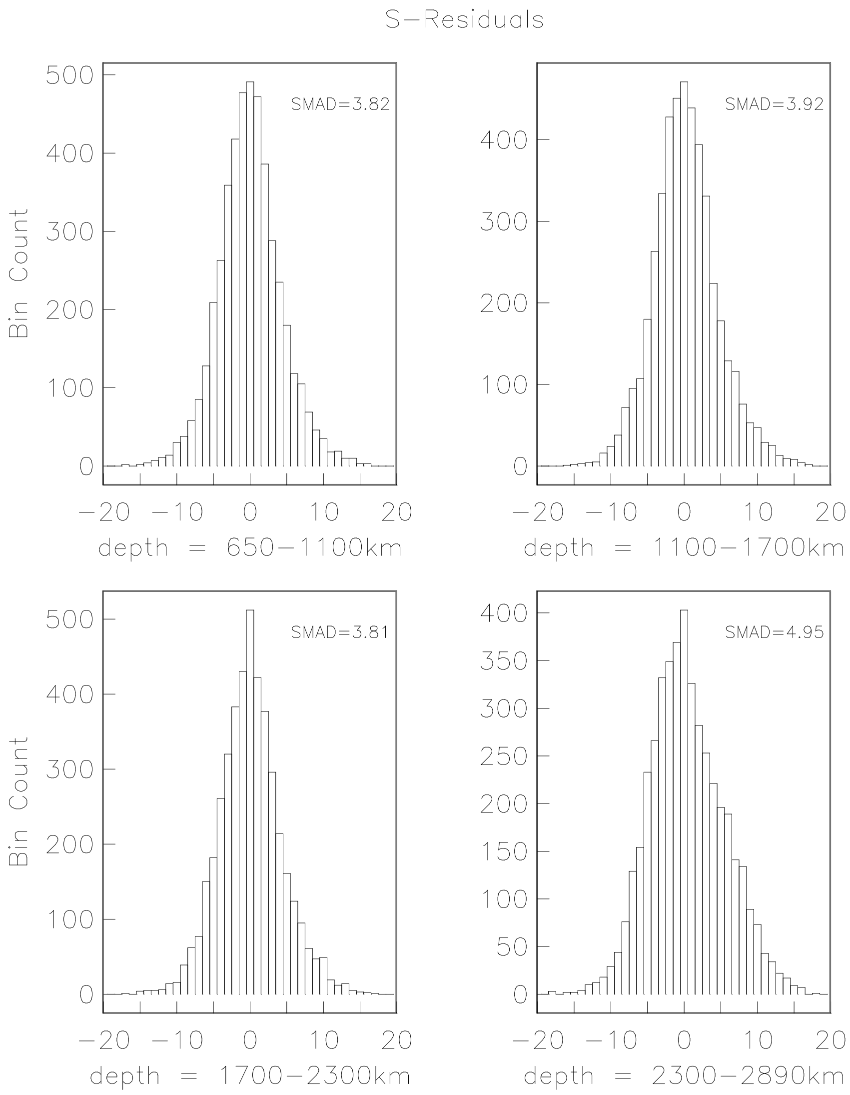


Figure 2.2: Histograms of the detrended S data in four ranges of turning depths in the lower mantle each about 550 km thick. Note the increased variance of the deepest-turning S rays.

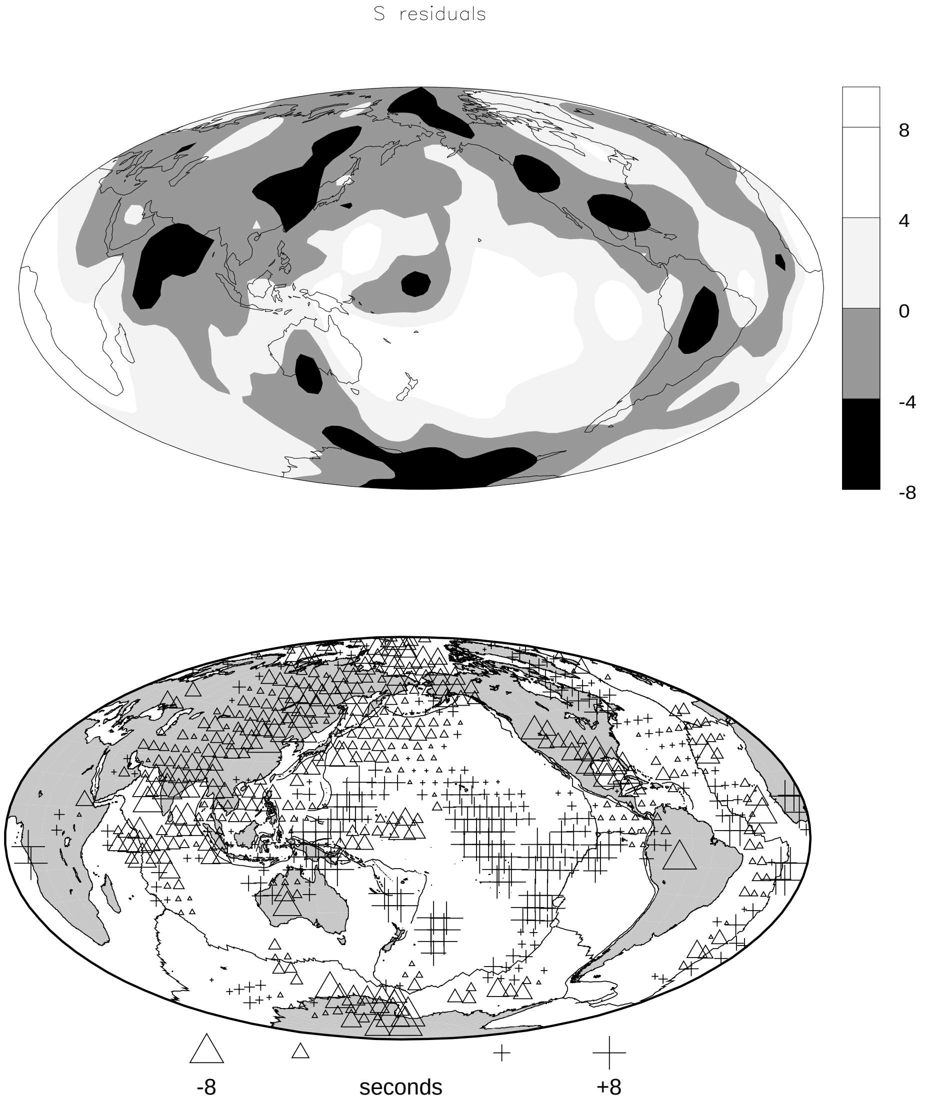


Figure 2.3: Lower panel: long-period S residuals plotted at the turning point of S for rays which bottom in the lowermost 300 km of the mantle. The raw data have been lightly smoothed by applying a running-mean smoothing filter which is a spherical cap of radius 5°. Note the ring of negative residuals (fast velocity) around the Pacific. The upper panel shows a map of the data constructed using spherical splines. The contour levels are in seconds.

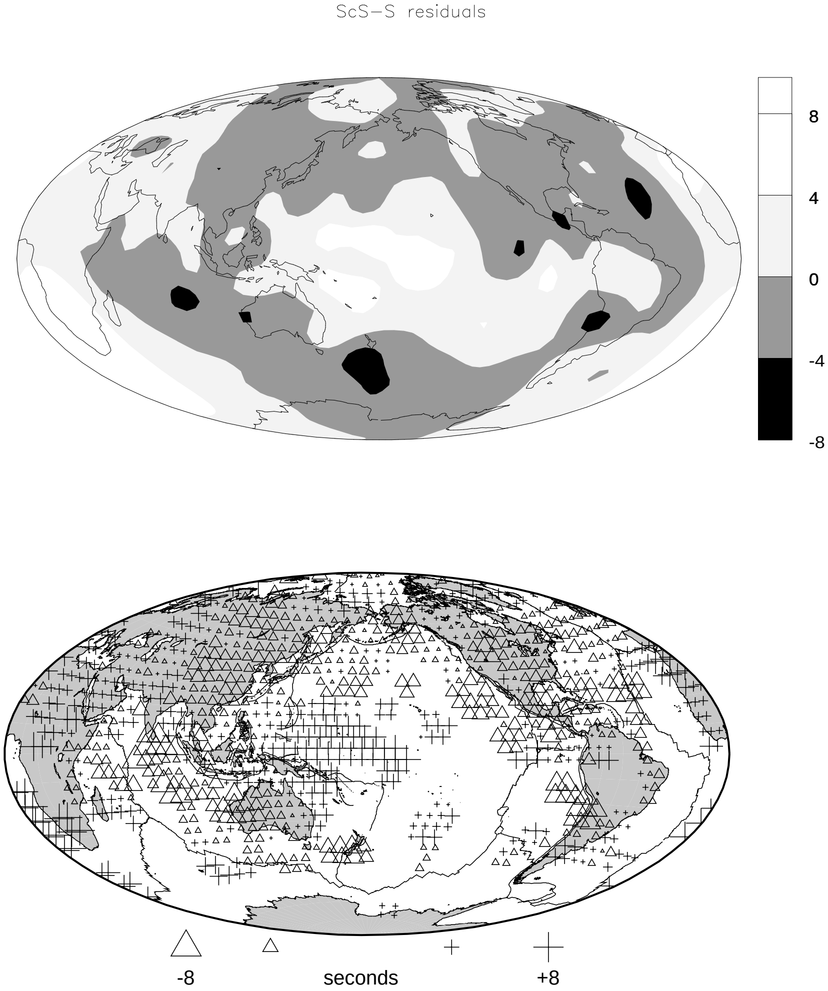


Figure 2.4: Long-period $ScS - S$ residuals plotted at the bounce point of ScS on the core-mantle boundary (see Fig. 2.3 for details). Again, note the ring of negative residuals (fast velocity) around the Pacific.

2.3.2 Differential times

An extensive study of $PP - P$ and $SS - S$ differential times was presented by Woodward & Masters (1991b) and these data sets have continued to be expanded. The data sets have now achieved almost complete geographical coverage of bounce point locations and provide good control of the vertically averaged structure in the upper mantle. Comparisons of the $PP - P$ and $SS - S$ patterns are also interesting, showing that $SS - S$ residuals are typically three times the equivalent $PP - P$ value. We have tried a similar experiment with $PcP - P$ (Pulver & Masters, 1990) for comparison with our extensive data set of $ScS - S$, but the small number of reliable long-period $PcP - P$ measurements makes the comparison rather uncertain. A preliminary result shows that $PcP - P$ times are also typically three times smaller than the equivalent $ScS - S$ times (in agreement with an earlier regional study of Jordan & Lynn (1974)). While this ratio is not as extreme as some that have been proposed, if taken at face value it is barely in agreement with a thermal interpretation for the source of the velocity anomalies in the lower mantle (Karato, 1993). It has therefore been suggested that lateral variations in composition are at least partly responsible for the velocity anomalies at the base of the mantle.

In the current inversion for S velocity, we use an expanded $SS - S$ data set of 10,000 measurements and an $ScS - S$ data set of about 5000 measurements. To give an idea of the coverage of the data sets, Fig. 2.5a shows the smoothed $SS - S$ residual distribution with residuals plotted at the SS bouncepoint at the surface. This map shows a clear correlation with surface tectonics though we have previously shown that lower-mantle structure can contribute significantly to $SS - S$ traveltimes (Masters, et al., 1993). In fact, if we use a shear-velocity model to calculate and remove the effects of lower mantle structure, we arrive at the map shown in Fig. 2.5b (this map uses model SH8/U4L8 of Forte, et al. (1993) but similar results are obtained with our models). This map now shows much clearer correlation with surface tectonics – in particular, the correlation with the age of oceanic lithosphere is now quite dramatic (Masters, et al., 1993) and suggests we should correct for this signal in the future.

2.3.3 Phase-velocity maps

Global surface wave inversions have been carried out by numerous workers either by measuring surface wave dispersion (Montagner & Tanimoto, 1991) or by modeling long-period waveforms (Woodhouse & Dziewonski, 1984; Tanimoto, 1988). We prefer the use of dispersion data since errors are more easily assigned and an independent evaluation of a proper parameterization for upper mantle structure is possible. The transfer function technique we use for phase estimation involves measuring phase and amplitude relative to a synthetic waveform which has been optimally aligned with the data. This robust technique is based on multi-taper spectral estimation (Thomson, 1982) and allows an estimate of measurement errors (Laske, 1993). To date, we have collected a total of about 10,000 phase measurements for the first two wave trains (R1, R2, G1, G2) and for great circle pairs (R3-R1, R4-R2, G3-G1, G4-G2) at each of various periods between 250 and 70 sec (Laske & Masters, 1996).

Inversions for surface-wave phase-velocity maps are dramatically improved by the addition of polarization data (i.e., the arrival angle of the incoming surface-wave packet). Such data are sensitive to the lateral derivative of phase velocity and so aid in the recovery of short wavelength structure. Our polarization measurement technique closely follows Park, et al. (1987) except in some minor respects (Laske, 1993) and gives a reliable measure of the frequency-dependent arrival angle of the surface-wave packet. Ray tracing calculations show that Woodhouse & Wong (1986) linear theory for the off-great circle arrival angle is sufficiently accurate for realistic long-wavelength phase velocity variations (up to $\ell = 16$) to explain the observed arrival angles. This is particularly true for low orbit wave trains which are of most interest since the polarization of the later orbits becomes more and more insensitive to structure of odd harmonic degree (Woodhouse & Wong, 1986). Polarization data have been collected for 150 selected events recorded by the global GEOSCOPE and IDA networks (a preliminary data set was presented in Laske, et al. (1994)).

A joint inversion of the phase and polarization data for a phase velocity map expanded up to $\ell = 16$ is shown in Fig. 2.6. The phase data do not require shorter wavelength structure (for example, no significant additional variance reduction is obtained by going from $\ell = 12$ to $\ell = 16$), but the polarization data are much better fit if shorter wavelength structure is added. Fig. 2.7 uses a checkerboard test ($\ell = 14$) to demonstrate the tremendous improvement in short-wavelength resolution when polarization data are included in the inversion. A more detailed description of the data and the final phase velocity maps is given in Laske & Masters (1996). The final Love wave phase velocity maps are presented in Fig. 2.8 and show a strong correlation with surface tectonics. It turns out that these data are critical for resolving upper-mantle shear velocity variations.

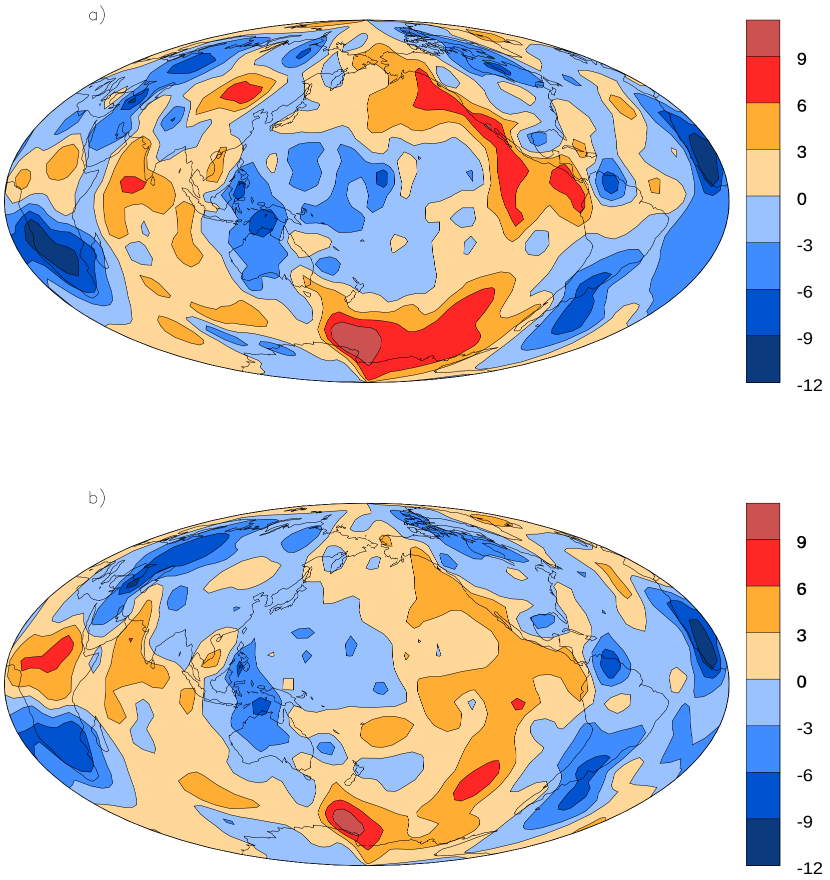


Figure 2.5: Smoothed $SS-S$ residuals relative to PREM, constructed by fitting spherical splines to $SS-S$ residuals placed at their surface bounce points. Contour levels are in seconds. Figure 2.5a is corrected for ellipticity and topography, and reflects both vertically averaged velocity structure in the upper mantle (from the SS leg) and low spherical harmonic degree structure in the lower mantle (from the S leg). Figure 2.5b is constructed from the same data but is corrected for the traversal of the S leg through the lower mantle structure of model SH8/U4L8. Figure 2.5b should represent vertically integrated upper mantle velocity structure, and fits a cooling halfspace velocity model of the oceanic crust better than Figure 2.5a.

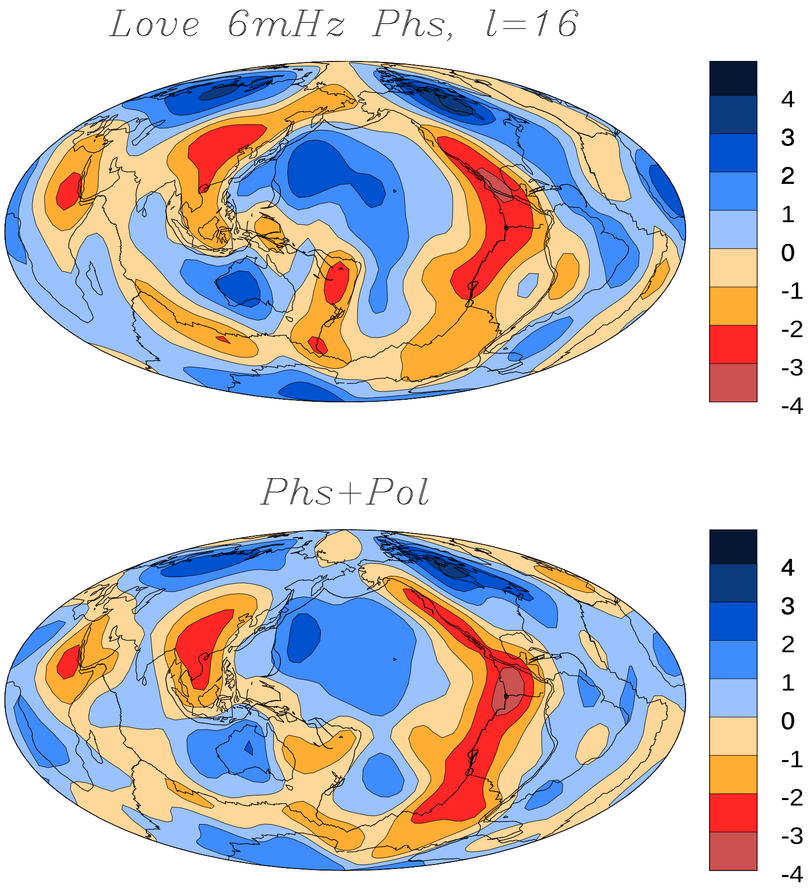


Figure 2.6: Comparison of the phase velocity maps obtained for Love waves at 6mHz. For the upper map, only phase data were used for the inversion, while phase and polarization data were used to obtain the lower map. The maps are expanded in spherical harmonics up to degree $\ell = 16$ and show phase velocity perturbation in %. The phase data do not require structure above $\ell = 12$ but the fit of the polarization data can be dramatically improved by low amplitude structure between $\ell = 12$ and 16.

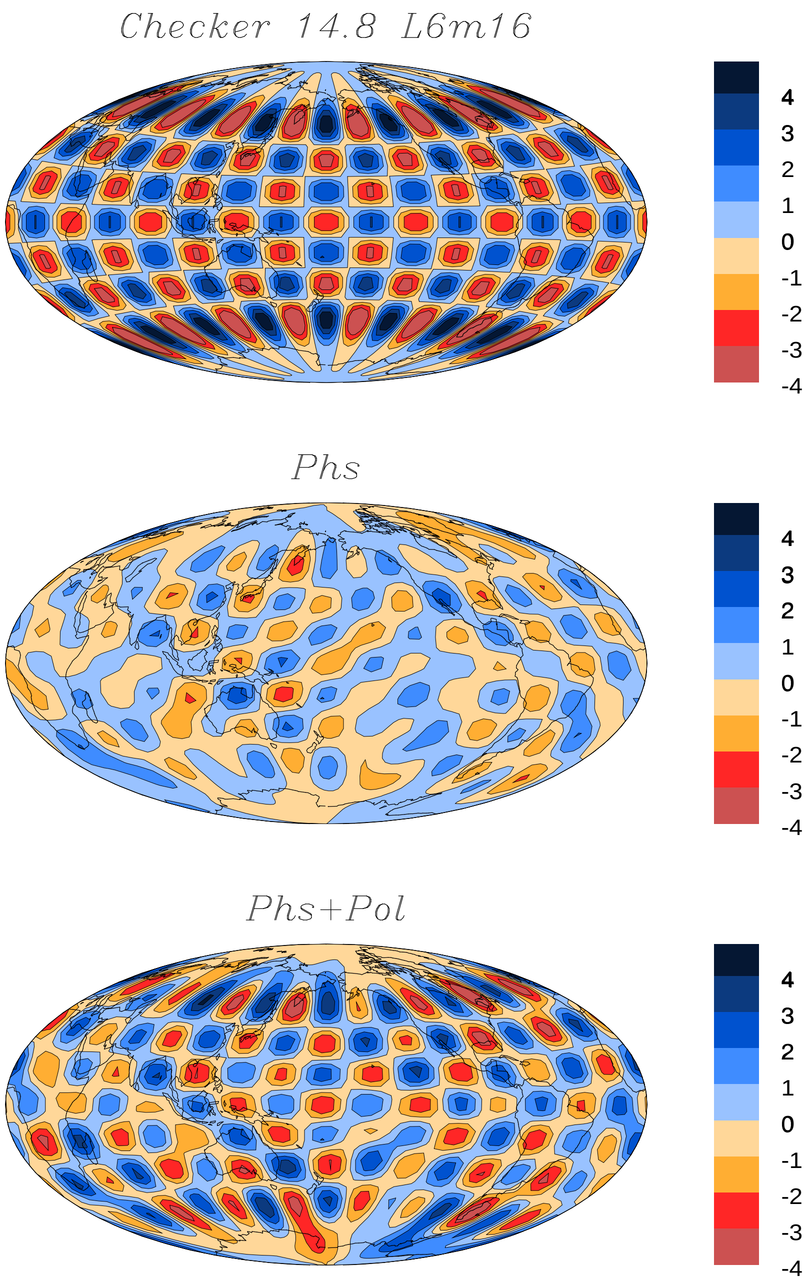


Figure 2.7: Resolution test using a checkerboard pattern (top panel). This test assumes that non-linear effects are unimportant and inverts a synthetic data set using the same operator that was used to invert the real data. The middle panel shows the recovery of the pattern when only phase data are included. The lower panel shows the recovery when phase and polarization data are used simultaneously in the inversion.

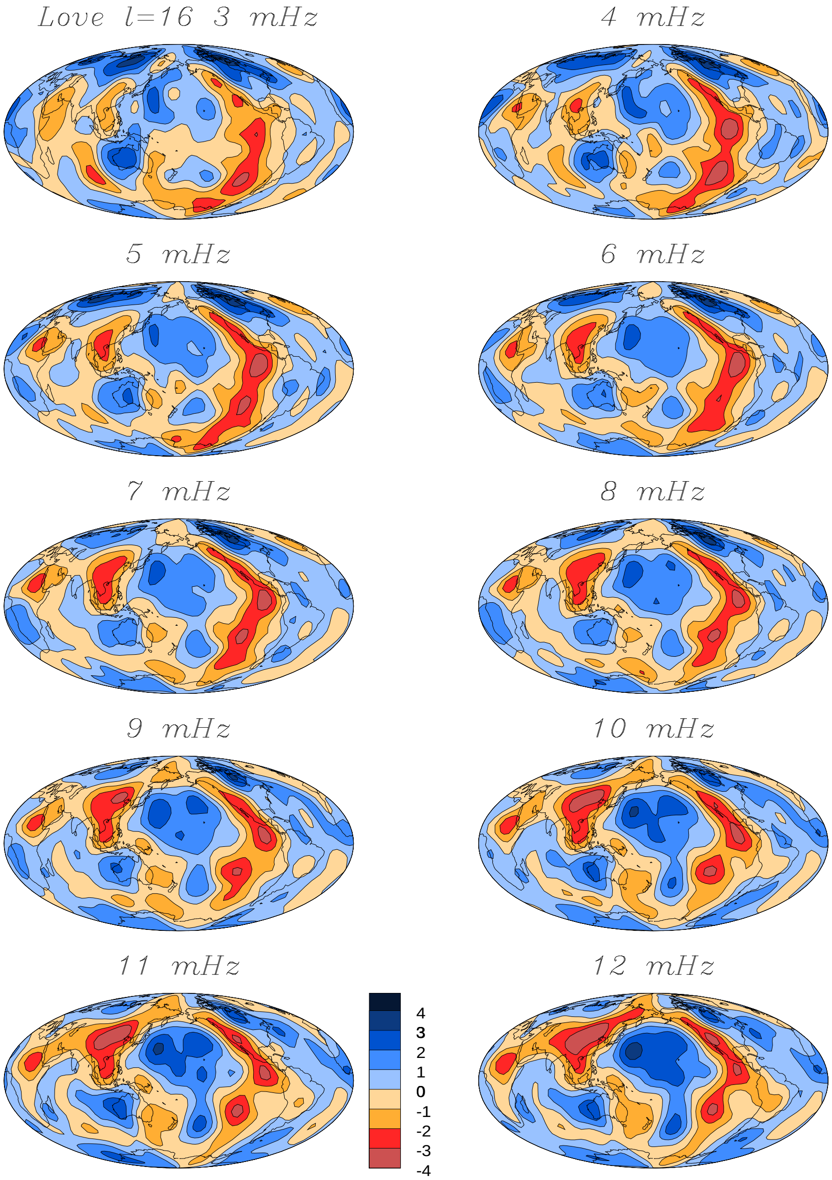

Love l=16 3 mHz
4 mHz
5 mHz
6 mHz
7 mHz
8 mHz
9 mHz
10 mHz
11 mHz
12 mHz

Figure 2.8: The ten Love wave phase velocity maps obtained from a smoothed inversion of phase and polarization data. The maps are expanded in spherical harmonics up to degree $\ell = 16$ and show phase velocity perturbation in %. The errors vary geographically and with frequency but do not exceed 0.3%.

2.3.4 Mode structure coefficients

The free oscillation data consist of 2034 structure coefficient estimates obtained from the studies of Smith & Masters (1989), Smith (1989), and Ritzwoller, et al. (1986, 1988). We use structure coefficients up to harmonic degree 8 for 69 spheroidal modes. The data set consists primarily of fundamental modes though there are 19 spheroidal overtones which are sensitive to lower-mantle structure. Details of the methods used for obtaining the structure coefficient estimates can be found in the original papers.

It should be noted that the structure coefficients for surface-wave equivalent modes and coefficients in the expansion of phase velocity maps are not independent data. Indeed, for large harmonic degree, we can convert the mode structure coefficients to an effective phase velocity map at the period of the mode though, because the mode coefficients are estimated using a standing wave analysis, there is no sensitivity to structure of odd harmonic degree (Jordan, 1978a). Fig. 2.9 shows the comparison of such a conversion with the even part of the phase-velocity map determined using surface-wave techniques by Laske & Masters (1996). It is remarkable that such different analysis techniques give such similar answers, which gives us confidence in both data sets.

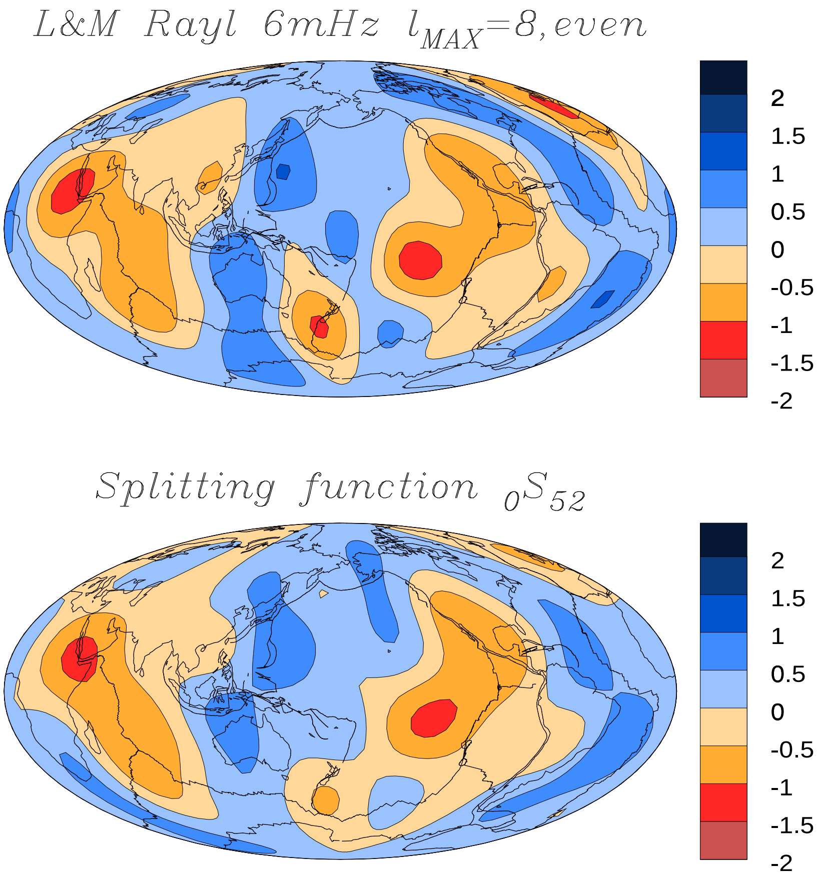


Figure 2.9: Comparison of the phase velocity map for Rayleigh waves at 6mHz and the splitting function of spheroidal fundamental mode $_0\mathrm{S}_{52}$. Since the splitting function is expanded only up to $\ell = 8$, the phase velocity map was truncated at $\ell = 8$ and only even harmonic degrees are shown.

2.3.5 Spectral content of the data sets

Inspection of Figs. 2.3, 2.4, 2.5 and 2.8 shows that the data sets are dominated by signals with long wavelength. This is made more obvious if we plot the amplitude spectra of the data (Fig. 2.10). The amplitude spectra for the traveltime data sets are estimated by performing a spherical spline fit to the cap values shown in Figs. 2.4 and 2.5 (Parker, 1994, chapter 2) and then computing the equivalent spherical harmonic expansion of the splined field. This procedure penalizes short wavelength structure but the amplitude spectra up to about degree 10 are insensitive to the degree of smoothing used in the spline fitting. In every case, we find that long-wavelength structure dominates the data sets and, in fact, most of the observed power is at harmonic degrees below 6. This observation is the primary motivation for parameterizing our shear-wave velocity structure in terms of spherical harmonics as they are likely to be an efficient basis for representing the dominant structure in the mantle.

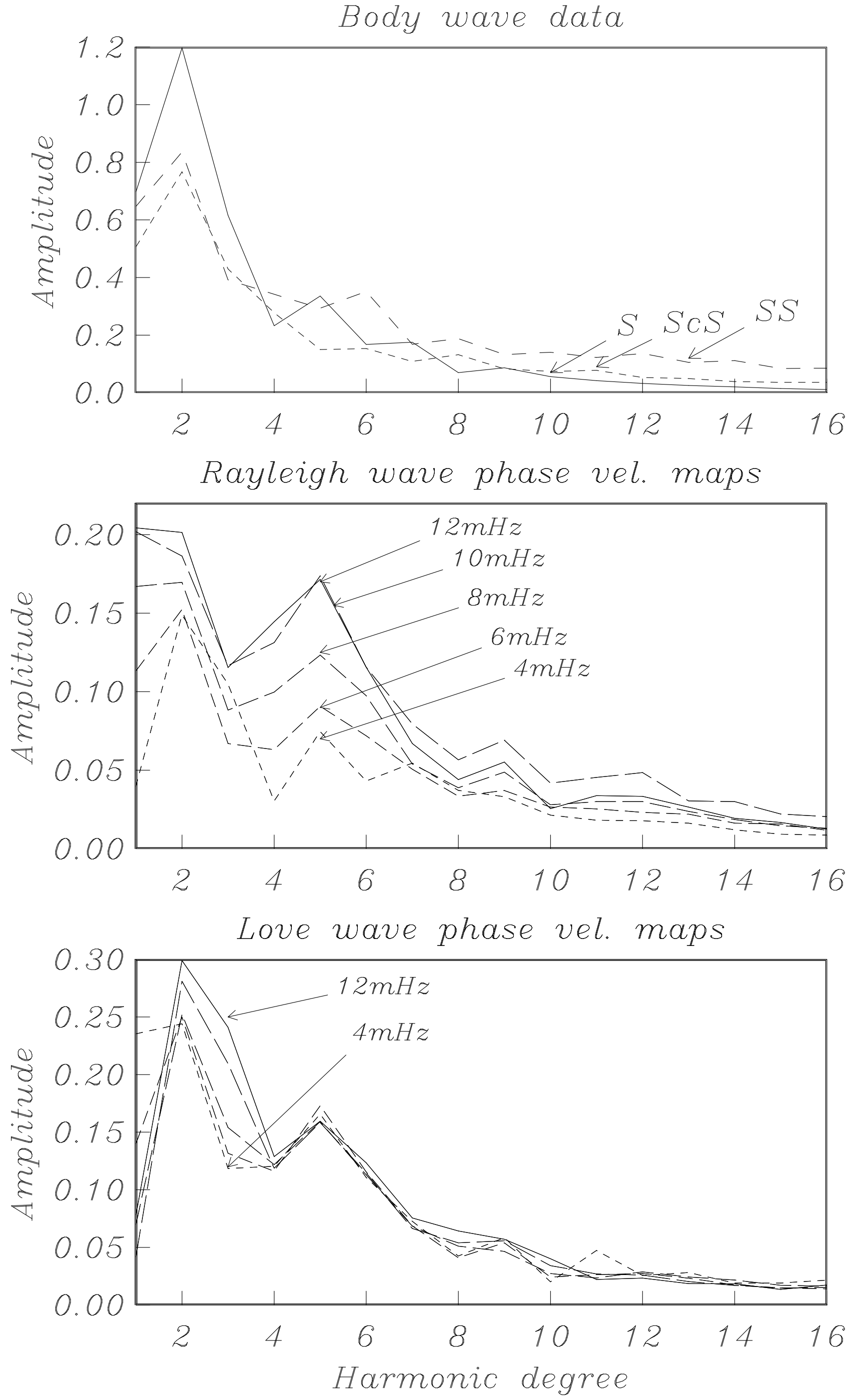


Figure 2.10: Amplitude spectra of the data sets used in this study (e.g. Figs. 2.3,2.4,2.5,2.8). The upper panel shows the spectra of the body wave traveltimes S, $SS-S$ and $ScS-S$ while the lower two panels are the spectra of the phase velocity maps for Rayleigh and Love waves at frequencies 4, 6, 8, 10 and 12 mHz. In all data sets, the long-wavelength component ($\ell = 6$ and below) is clearly dominant.

2.4 Inversion

2.4.1 Parameterization

One of our early inversions of the data used only $SS - S$ differential travel-time residuals in conjunction with a subset of the mode structure coefficients to invert for upper mantle structure (Woodward & Masters, 1992). This work demonstrated that simple models can be constructed that go a long way to simultaneously fitting both data sets but it also showed that there is a significant contribution to the $SS - S$ data from the lower mantle. Now that we also have extensive data sets of S and $ScS - S$ times, an inversion for whole mantle structure is both feasible and desirable. Several preliminary inversions have been done (Masters & Bolton, 1991) which initially followed Tanimoto (1990) and divided the mantle into eleven shells in each of which structure up to harmonic degree 10 was allowed. In our latest inversions, the model is parameterized in radius by natural cubic B-splines. We prefer these basis functions to the Legendre or Chebyshev polynomials used in other studies since these latter basis functions have global support and are capable of mapping structure in one part of the mantle to another. The cubic B-splines are the functions defined at a set of knots with continuous zeroth, first and second derivatives and with the minimum support (and with zero second derivatives at the end knots). The knots have been set at approximately 100 km spacing across the entire mantle, which clearly constitutes a radial overparameterization based on a posteriori evaluations of resolution (see below). This overparameterization and consistent knot spacing is desirable in mitigating the effects of a rather arbitrary choice of basis functions on the appearance of the resulting model. The shear velocity perturbation can now be written as

$$\delta v(r,\theta,\phi)=\sum_{l=1}^{16}\sum_{m=-l}^{+l}\sum_{k=1}^{30}{}_k\delta v_l^m b_k(r)Y_l^m(\theta,\phi)$$

where b_k is a B-spline and Y_l^m is a fully normalized surface spherical harmonic (Edmonds, 1960). The present experiments ignore the possibility of lateral variations in the topography of internal discontinuities. The traveltime data we use are rather insensitive to the location of discontinuities. On the other hand, it was shown by Masters, et al. (1982) that the long-period mode and surface-wave data are somewhat sensitive to the topography of mantle discontinuities giving a tradeoff with structure in the immediate vicinity of the boundary. The experiments of Woodward & Masters (1992) suggest that this tradeoff is not strong and that the boundary perturbations are only weakly constrained with the current data sets. The constraints from the stacking of mantle discontinuity phases (Shearer & Masters, 1992) can be included in the inversions and should help greatly in reducing these tradeoffs. This will be the subject of future work.

One final point is that the crust has a significant impact on most of the data sets but, at the same time, is too thin to be resolved by them. Most authors have handled this by applying a crustal correction to the data before inversion. We follow this procedure though it must be recognized that crustal structure is poorly known in many parts of the world and our correction may be quite uncertain. To compute the crustal corrections, we use the model of W. Mooney (personal communication 1994) which is based on the work of Christensen & Mooney (1995). This model is a seven layer model (including water and ice) for v_p and ρ, compiled for 52 crustal types. Except for water and ice, an average Poisson's ratio of 0.27 is taken to compute v_s. Topography and bathymetry data are taken from the ETOPO5 database (National Geophysical Data Center 1986). The crustal model used in earlier works included the compilation of crustal thickness by Soller, et al. (1982). Though our new crustal model is fairly detailed compared with the old one, it gives rather similar effects. As is well-known, the crustal correction actually increases the variance in most of the data sets leading to larger amplitude velocity perturbations in the uppermost mantle.

2.4.2 Model construction

The inverse solution is obtained following the philosophy espoused by Shaw & Orcutt (1985) and Constable, et al. (1987). The aim of this method is to obtain a solution which minimizes the misfit, to within a prescribed tolerance, while simultaneously minimizing a measure of roughness of the model produced.

In matrix notation, our forward problem is

$$d = Gm$$

where m is the model of spherical harmonic expansion coefficients. For the traveltime data, d is a vector of traveltime residuals and the corresponding rows of G are computed using the linearized theory given by Dziewonski (1984). The theoretical basis for this interpretation of the traveltime data is rather simplistic as it is based on ray theory and invokes Fermat's principle to justify the neglect of perturbations in ray paths from those in the spherically averaged Earth. Grand (1994) has shown that ray-path deviations can be quite severe for rays which bottom in the upper mantle, particularly in the presence of low-velocity zones. All of our traveltime data bottom in the lower mantle and 'a posteriori' ray-tracing experiments indicate that ray-path deviations are small and have almost no effect on the predicted traveltime (Michael Kendall – personal communication). For the mode and phase-velocity data, d is a vector of structure coefficients or phase-velocity expansion coefficients and the corresponding rows of the matrix G are computed using the results of Woodhouse & Dahlen (1978). In reality, the traveltime data (in particular, the absolute S times) are sensitive to the event location as well as to the velocity structure. We therefore use a projection algorithm to find linear combinations of all the traveltime data for a particular event which are, to first order, insensitive to the event location. Such techniques are quite common in tomographic analyses (Scott, et al., 1994) but it turns out that the velocity models that we obtain do not change much if we omit this step.

We desire a model such that the misfit is of a predicted size. That is, we want

$$\|d - Gm\|^2 = \chi^2$$

where χ^2 is an estimate of the expected misfit in the data and $\|.\|$ is the Euclidean norm. By normalizing the rows of d and G by the estimated errors of the individual data, χ^2 should be roughly the same as the number of data used. We also desire that the model be smooth, which is ensured by penalizing a measure of roughness

$$\|\partial m\|^2$$

where ∂ could, for example, be the first or second differencing operator. A solution which satisfies these constraints can be obtained by using Lagrange multipliers (Constable, et al., 1987). The role of the Lagrange multiplier, μ, is illustrated by the resulting normal equations

$$\left[G^T G + \mu \partial^T \partial\right] m = G^T d.$$

Solving this equation for $\mu = 0$ gives us the best fitting (least squares) solution, but there is no penalty for roughness and the data may well be overfit. As μ is increased the penalty for roughness increases, so the solution is smoother but the misfit is greater. In practice, we look at a family of solutions for a variety of Lagrange multipliers and choose a compromise in the tradeoff between roughness and misfit. Although the errors are difficult to establish in an *a priori* fashion, conservative error estimates from the analysis of the data are used to select models with a reasonable misfit to the data. In the interest of increased numerical stability over a wide range of Lagrange multipliers and smoothing matrix constructions a factored form of the normal equations is solved by the QR method (Parker, 1994). A parallel version of this algorithm is discussed in appendix A.

The construction of the smoothing operator is complicated by the fact that we usually wish to treat radial and lateral smoothing of the model differently (since the natural length scales and resolution of the data can be quite different in the radial and lateral directions). We follow Woodward & Masters (1992) and separate the smoothing of lateral and radial variations by writing

$$\partial^T \partial = D = D_L + \alpha D_R.$$

D_L controls lateral smoothing and D_R controls radial smoothing; α is an empirically determined weighting factor which is used to balance the radial and lateral smoothness. Lateral roughness (for each radial basis function) is expressed as

$$\int_{\Omega} |\nabla_1^2 \delta v_k(\theta, \phi)|^2 \, d\Omega.$$

where $\delta v_k = \sum_{l,m} {}_k\delta v_l^m Y_l^m(\theta, \phi)$, ∇_1^2 is the surface Laplacian and Ω is the surface of the sphere. It therefore follows that a total measure of lateral roughness of the model can be written:

$$m^T D_L m = \sum_{k,l,m} l^2(l+1)^2 |{}_k\delta v_l^m|^2$$

Thus D_L is a diagonal matrix with terms of the form $l^2(l+1)^2$ along the diagonal.

Radially, we again use a second derivative measure of roughness so that the total radial roughness of the model is written:

$$m^T D_R m = \sum_{l,m} \int_c^a |\sum_k {}_k\delta v_l^m \frac{\partial^2}{\partial r^2} b_k(r)|^2 \, dr$$

where c is the core-mantle boundary radius and a is the radius of the base of the crust.

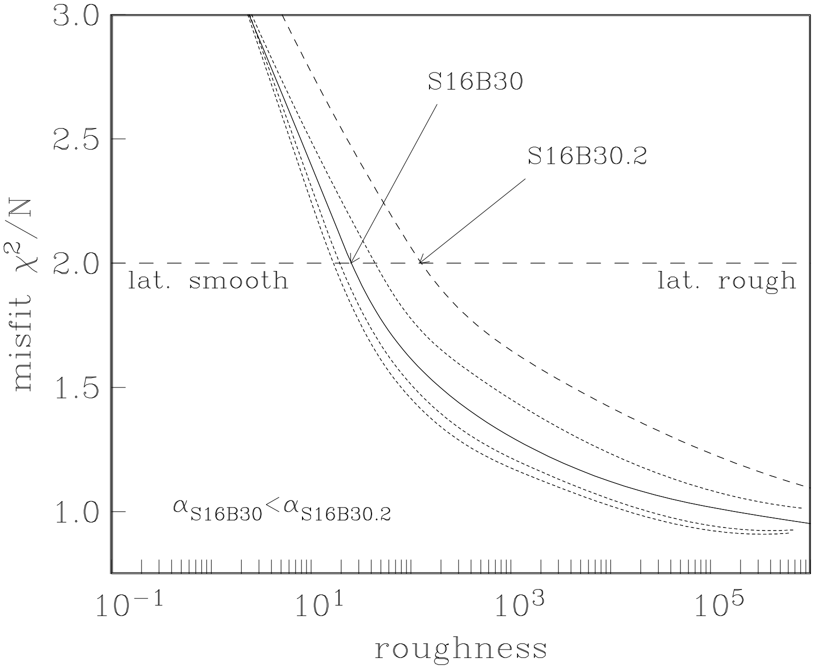


Figure 2.11: Tradeoff curves from the inversions for S16B30. A total data misfit level of $\chi^2/N = 2.0$ lies at the dashed line; models below this line are rougher, and models above smoother. The family of curves is generated with a range of values of the parameter α (see text). At a given misfit, models on the solid curve (e.g. S16B30) are radially rougher and laterally smoother than models on the dashed curves to the right (e.g. S16B30.2).

The inversion is now done for a variety of values of α and the resulting tradeoff curves are shown in Fig. 2.11. We find that there is a range of models which give acceptable fits to the data but vary in their smoothness characteristics. Choosing a particular value for the combined misfit of all data sets yields a range of misfits for each of the data sets. In the case of choosing a combined misfit χ^2/N of 1.0 the Rayleigh and Love wave data sets are overfit at χ^2/N of 0.45 and 0.85, respectively. In the interest of choosing a conservative model (i.e., not unnecessarily rough), models with a combined misfit of 2.0 are deemed preferable. The range of misfits for each of the data subsets is shown in Fig. 2.12. We present two models in Figs. 2.13 and 2.14, each of which fit the data at the combined misfit level of 2.0. The model in Fig. 2.13 is laterally very smooth but radially rough while the other one is laterally rougher but radially smoother (Fig. 2.14). The gross features of the models are similar with large-amplitude structure near the top and bottom of the mantle and with lower-amplitude structure in the mid mantle. Small-scale features are, however, quite different. Our prejudice for the model in Fig. 2.13 is founded on considerations of Fig. 2.12 and supported by the comparison of small scale structure in the model with small scale structure in the primary data sets. Models less laterally rough (smaller α) are a poorer fit to the Love wave data set, and models radially smoother (larger α) are a poorer fit to the Rayleigh wave and spheroidal mode data. A reasonable compromise is to choose the model which satisfies all the data sets fairly well. The spectrum of lateral variations in the ScS-S data set is also more consistent with the model in Fig. 2.13 suggesting that this model is reasonable throughout the mantle.

Fig. 2.13, therefore, shows our “best” model (S16B30) and, for comparison, model SH12WM13 of Su, et al. (1994) is plotted in Fig. 2.15. In Fig. 2.16, we show the rms amplitude as a function of depth of both models along with the correlation between them. S16B30 is larger in amplitude near both boundaries of the mantle but both models correlate well in shape in these regions. The correlation is poorest in the mid-mantle, where structure is of low amplitude, but the correlation is significant at the 95% level (which is about 0.3 for this number of parameters). We believe that our model exhibits higher amplitudes near the core-mantle boundary because of the inclusion of the data set of absolute S times which sample this region. Such data are absent from SH12WM13 though recent models by the Harvard group which include $SKS-S$ differential times have amplitudes similar to S16B30 (Adam Dziewonski, personal communication). The reason that S16B30 has larger amplitude structure near the surface is less clear though we believe it to be necessary to explain the short-period Love wave

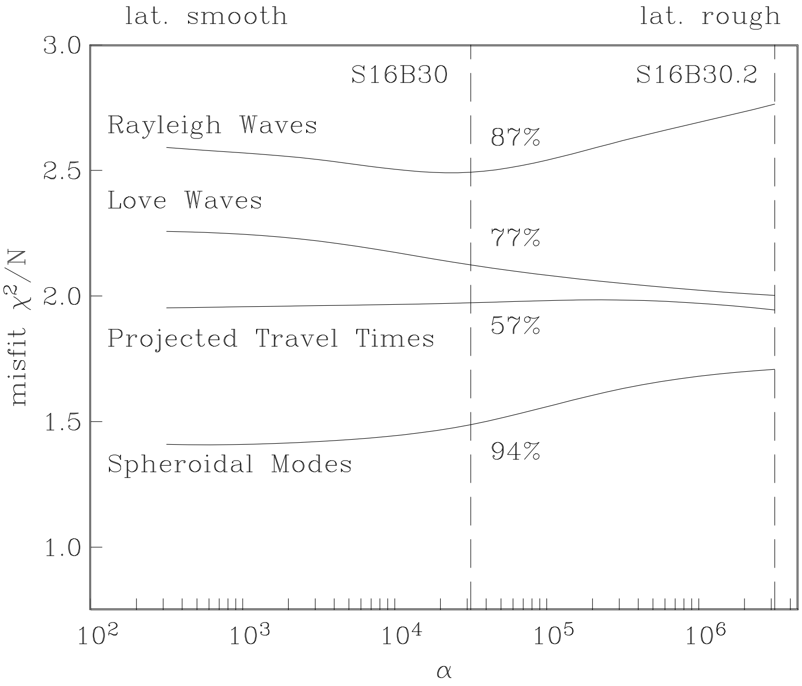


Figure 2.12: Misfit to various data sets for models with a total misfit of $\chi^2/N = 2.0$. Sampling along the tradeoff curves in Fig. 2.11 with this misfit yields a set of models with varying smoothness characteristics. These models fit the subsets of the data differently, presumably according to how inconsistent the smoothing constraint is with the structure preferred by the data subset. The vertical line corresponds to the value of α of S16B30, and is chosen to satisfy the surface wave and mode data sets generally the best. Variance reductions using model S16B30 are 57% for traveltimes and 77-94% for the surface waves and spheroidal modes.

data of Fig. 2.8.

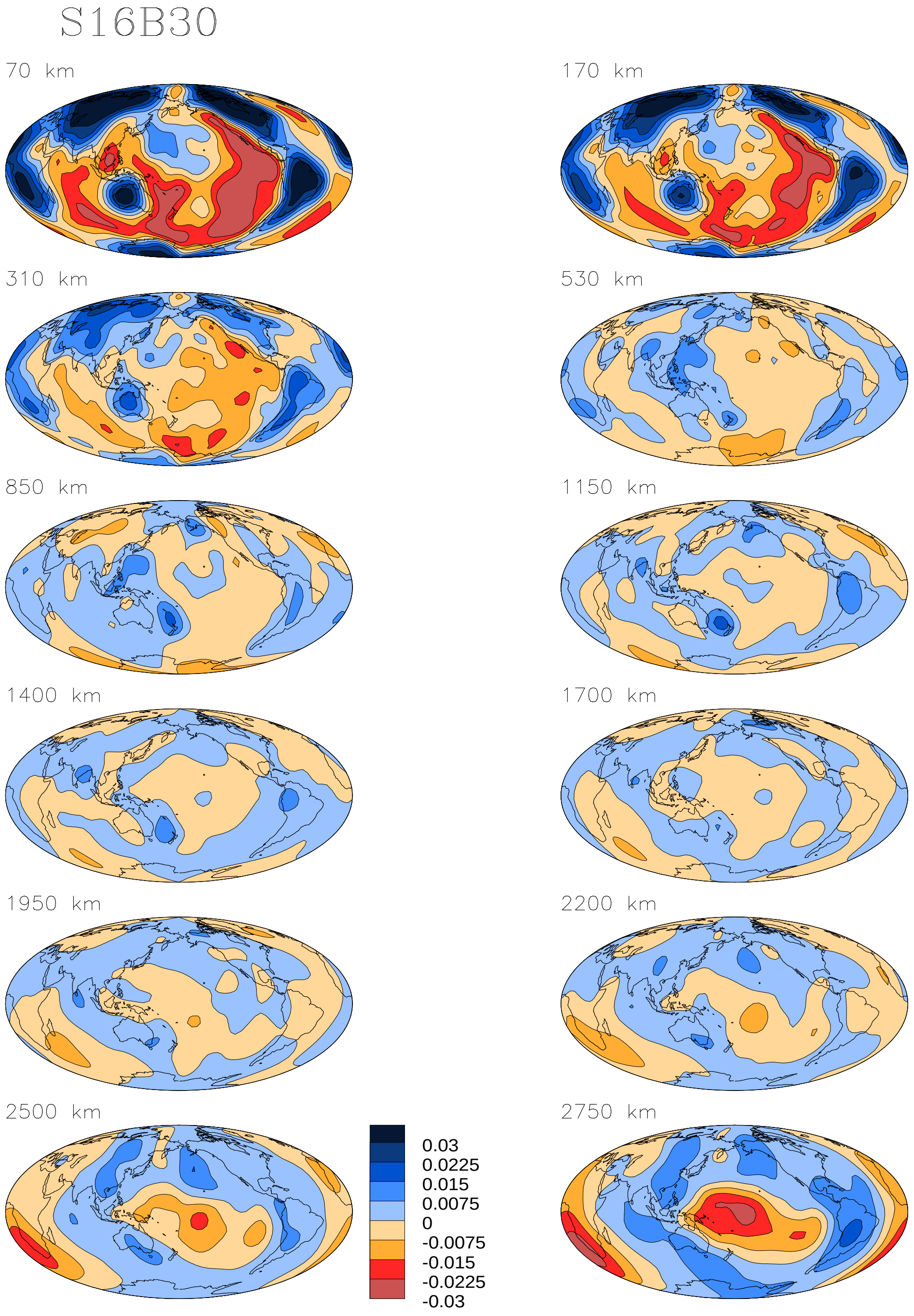


Figure 2.13: Twelve depth slices of model S16B30. Contoured values are shear velocity perturbation relative to the global model average at a given depth.

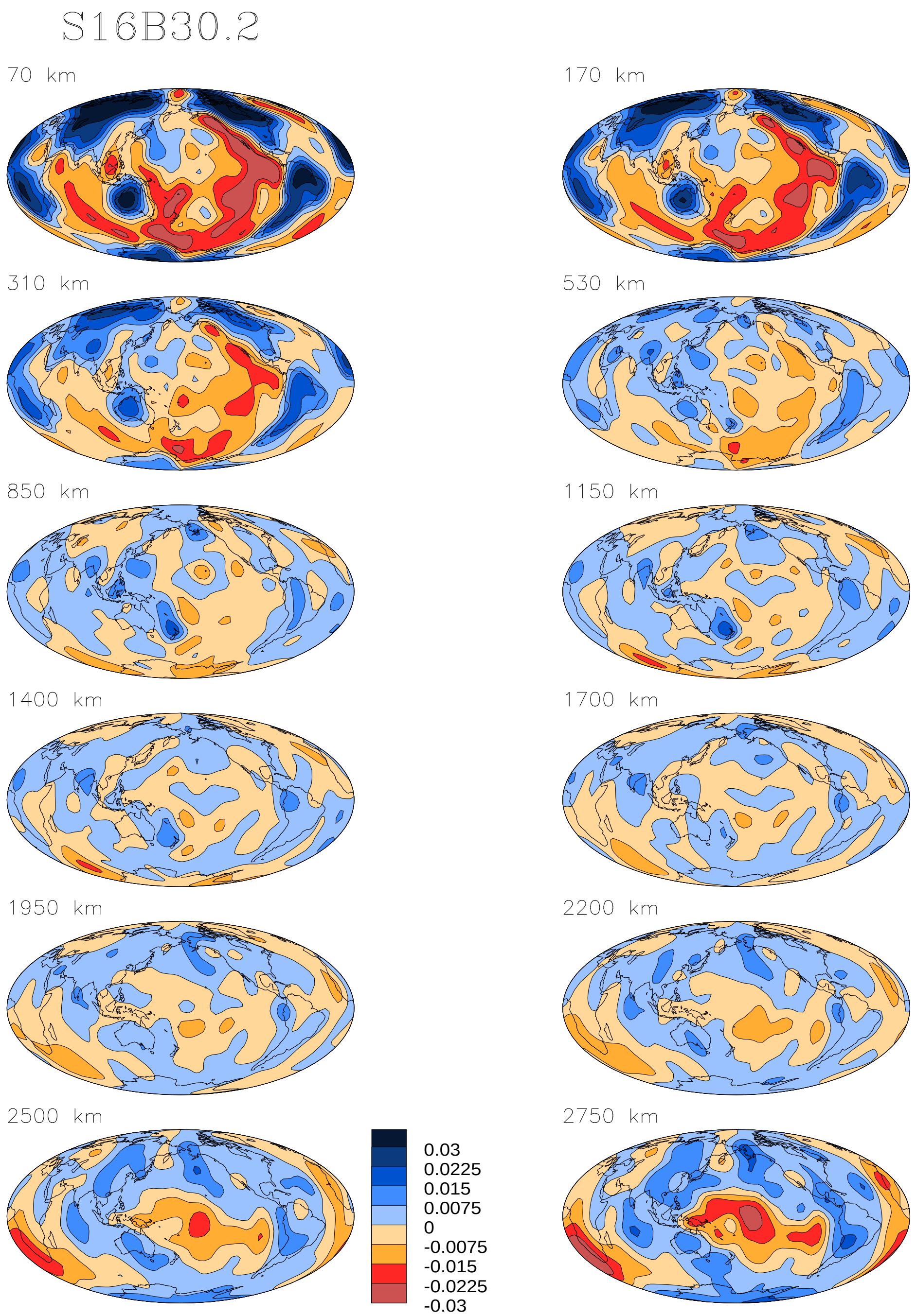


Figure 2.14: Model S16B30.2, plotted as in Figure 2.13. While radially smoother than S16B30 (Fig. 2.13), this model is laterally rougher, as is especially evident in the comparison of the lower mantle slices of Figs. 2.13 and 2.14. Even though the total data misfit expresses no preference for model S16B30 vs. S16B30.2, it is argued that S16B30 is a better compromise amongst the misfits shown in Fig. 2.14, and that S16B30.2 possesses spurious structure when compared to data sets such as $ScS - S$.

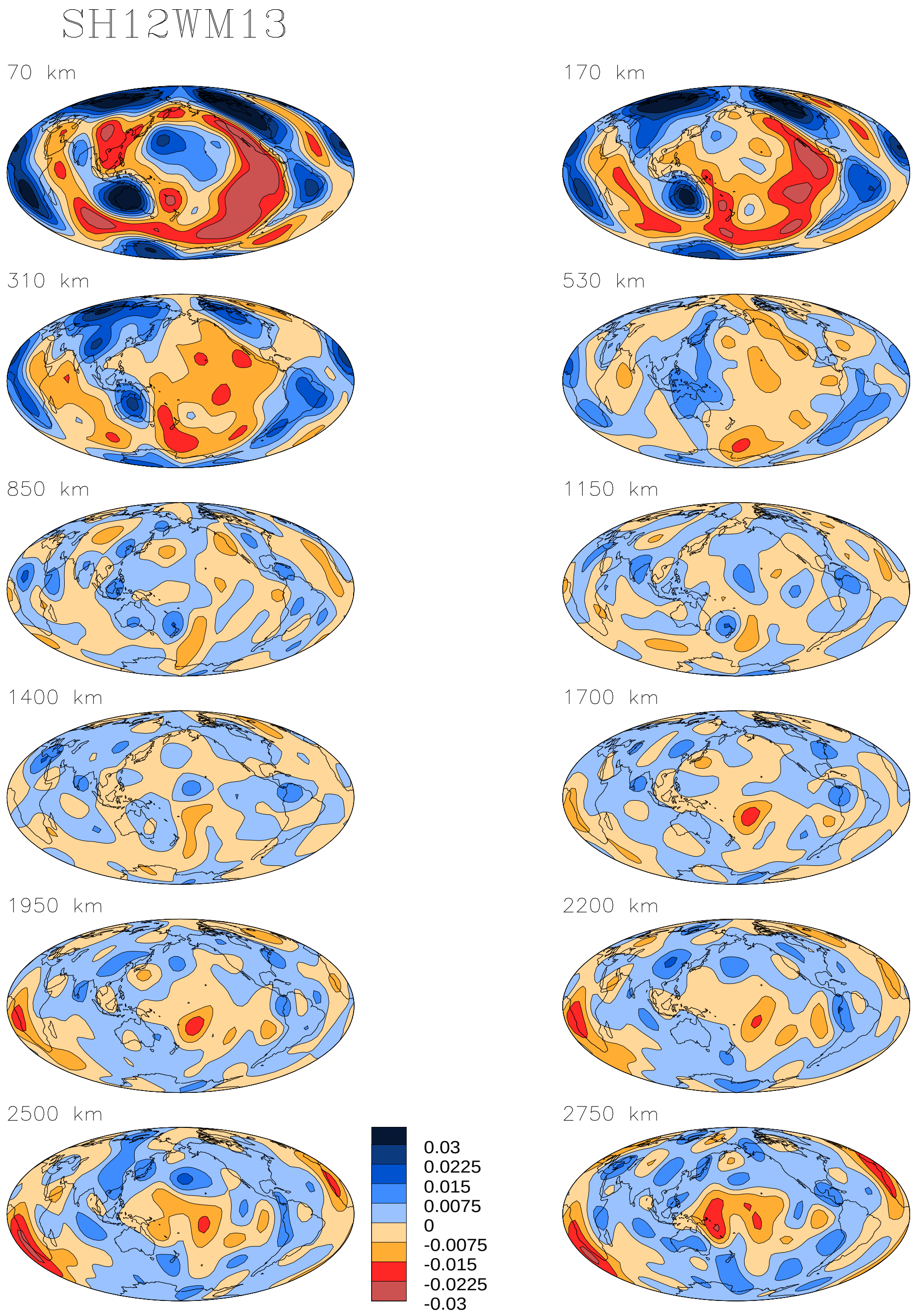


Figure 2.15: Model SH12WM13 plotted as in Fig. 2.13. All the models presented here have a relatively small fraction of data in common (part of the $SS-S$ and $ScS-S$ data sets), but were obtained by different methods and employ different radial basis functions. The main differences between S16B30 (Fig. 2.12) and SH12WM13 are in the patterns of mid-mantle structure and in the size of anomalies in the lowermost mantle.

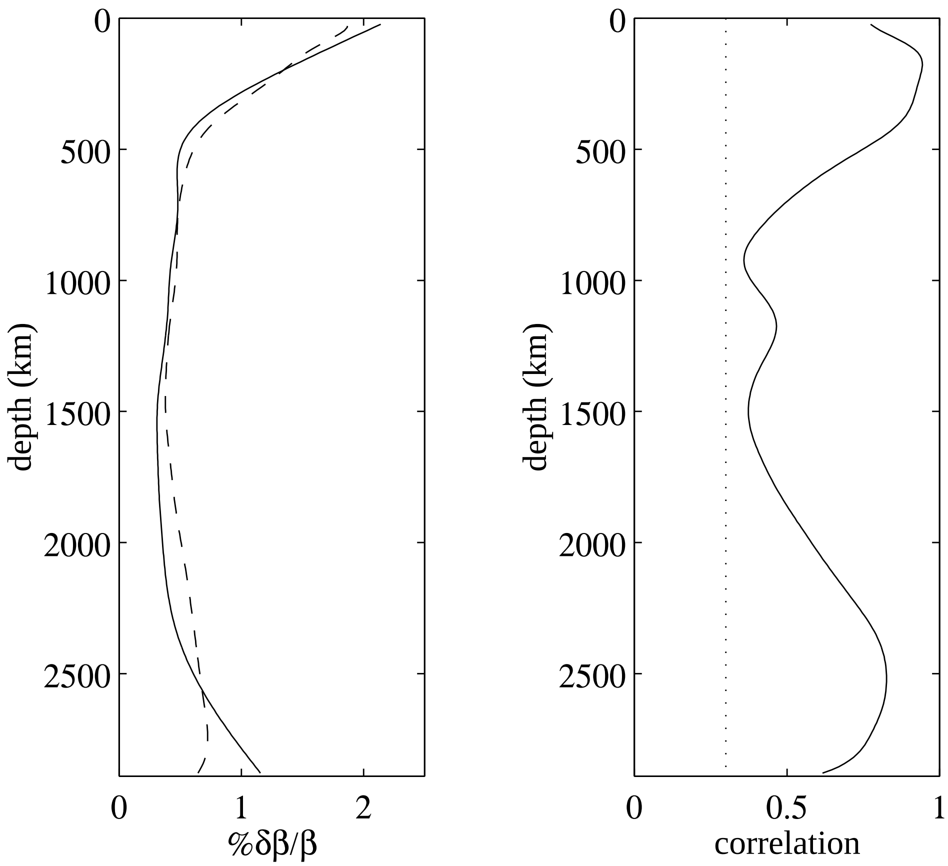


Figure 2.16: Comparison of S16B30 and SH12WM13 as a function of depth. The left panel shows the RMS amplitude as a function of depth (S16B30 is the solid line). The main difference between the two models occurs near the base of the mantle and is probably due to the inclusion of absolute S traveltimes in the inversion for S16B30. The right panel shows the cross-correlation between the models as a function of depth (computed for spherical harmonics up to degree 12). The dotted line indicates the 95% significance level (about 0.3).

2.4.3 Resolution and error analysis

We evaluate the resolution of the modeling algorithm by performing several different inversion experiments using the operator which produced the model from the real data. This operator was applied to synthetic data sets constructed from models consisting of checkerboard patterns of approximately 250km thickness placed at various levels in the mantle. This technique allows depth resolution as well as lateral resolution to be evaluated. The results for three layers are shown in Fig. 2.17 for a pattern that is a Y_7^4 spherical harmonic. Since most of the power in the model in Fig. 2.13 is of lower harmonic degree, the results of this test may be somewhat pessimistic. When the pattern is placed in the upper mantle (Fig. 2.17c), the inversion smears the answer in depth a little (upper panel) but recovers the geographic pattern very well (lower panel). Our data are actually capable of accurately recovering harmonic patterns up to degree 16 in this region. In the mid and lower mantle (Figs. 2.17a and 2.17b) the pattern is recovered well in the northern hemisphere but reveals the effects of sampling inadequacies in the southern hemisphere. The recovered patterns tend to have low amplitudes where resolution is poor so the total rms amplitude of the map is underestimated. On the other hand, where resolution is good, the peak values are quite close to the input values.

In all cases, sharp radial features are smeared 150-250km in depth with some low amplitude contamination spreading across the lower mantle. This may, in part, be due to the second derivative smoothing which strongly penalizes sharp changes with radius and we are currently experimenting with different radial smoothing criteria to see if we can reduce radial leakage.

Maps of model variance are computed using a Monte-Carlo technique by applying the matrix inversion algorithm to several hundred randomly chosen realizations of the data errors. Model variances can also be obtained by forming the covariance matrix of the model parameters and propagating errors from the data to errors in the model, but it turns out to be computationally easier to use the Monte-Carlo technique given the size of the matrices we have to deal with. Errors are largest near the boundaries of the mantle but peak at about 0.2% (more than a factor of ten smaller than the peak signal). In the mid-mantle, formal errors are about 0.1%, which, again, is much smaller than the model there. Scaling up the errors by a factor of $\sqrt{2.0}$ to account for the fact that our data are not perfectly fit by the models doesn't change the conclusion that most of the detail in the models is formally above the noise.

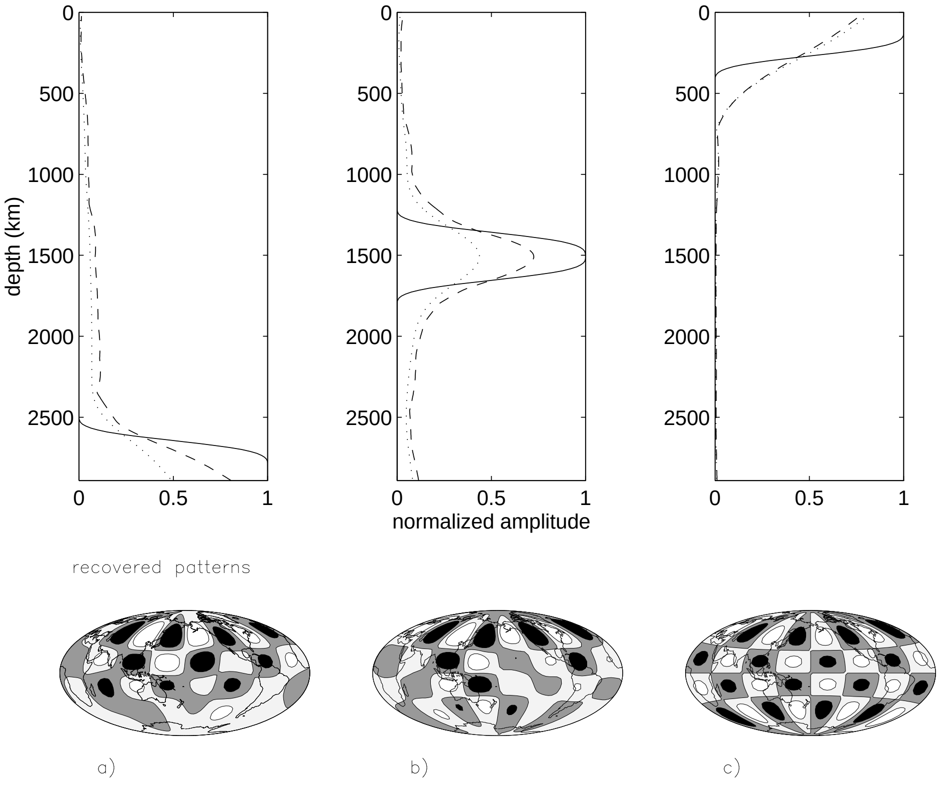


Figure 2.17: Attempts to recover checkerboard patterns at three levels in the mantle. The input pattern is the linear combination of three natural B-splines in radius and is a Y_7^4 spherical harmonic laterally. The solid curve in the upper plots is the amplitude vs. depth of the input pattern, normalized to have a peak value of 1. The dashed line is the peak recovered value at each depth as a fraction of the input pattern peak at that depth. The dotted line is the fraction of the RMS amplitude recovered. The lower plots show the recovered pattern at the depth of the peak recovered RMS. Although not shown, the input pattern is essentially identical to the recovered pattern in c). a) An attempt to recover a checkerboard pattern situated at the bottom of the mantle (the bottom 3 B-splines). Although some of the input pattern is smeared into low amplitudes across the lower mantle, the radial resolution is taken to be the increase in radial extent of the main peak of the recovered pattern which is roughly 200km. The difference in the fraction of recovered RMS and peak amplitudes is due to the geographic variation of the resolution. b) An attempt to recover a checkerboard pattern situated in the mid-mantle. As in a), the input pattern is the linear combination of three B-splines in radius. The edges of the input peak are smeared 200-300km radially. c) Attempt to recover a checkerboard pattern situated in the upper mantle. The model construction algorithm tends to smear the anomaly about 200km radially although the shape is very well-recovered.

2.5 Discussion

Our final model, S16B30, is shown in Fig. 2.13 and four cross-sections are shown in Fig. 2.18. The model is characterized by large-amplitude, large-scale structure, peaking at about ±6% in the upper mantle and at about ±2.5% at the base of the mantle. The RMS amplitude of velocity perturbation as a function of depth changes smoothly from high amplitudes near the surface to low amplitudes from 500-2200 km depth to higher amplitudes near the CMB. Structure shallower than 300-400 km has peak amplitudes in spherical harmonic degrees 1,4 and 5 while degree 2 is predominant in the transition zone. The spectra in the mid mantle are whiter though, below about 2000 km depth, the lower mantle is again dominated by degree 2 structure.

Structure in the uppermost and lowermost mantle appears to be the best established, presumably because of its high amplitude and red spectrum. The cross sections reveal low velocity anomalies along mid-ocean ridges extending in some cases to at least 300km depth. The effects of the cooling oceanic lithosphere are evident in the shallow model structure in oceanic regions. Fast regions extend 300-400 km beneath continents, supporting the hypothesis of a continental tectosphere (Jordan, 1978b). Backarc regions are characterized by shallow low velocities. In the lower mantle, a ring of high velocity surrounds the Pacific and large slow regions are present in the Central Pacific and beneath Africa. Although SH12WM13 and S16B30 are less similar in the transition zone, both show fast regions beneath the subduction zones in the western Pacific and beneath South America.

The resolution tests (Fig. 2.17) and comparison of SH12WM13 and S16B30 show that interpretation of features in the mid-mantle is more dubious. Resolution is strongly a function of geographic location, with the highest resolution in the circum-Pacific and Asian regions. We should also bear in mind that radial smearing of short wavelength structure is more severe than for long wavelengths and so is likely to be more of a problem in this region. Fig. 2.13 and the cross sections show several more or less radially continuous fast features. Examples are beneath the Tasman Sea, India, the North Central Pacific, and South America, all regions which may well have been supplied with downgoing material in the past (Ricard, et al., 1993). All of these features are in regions of fairly good lateral resolution (Fig. 2.17). Synthetic tests of convection simulations show that similarly sized cylindrical features which are oriented vertically in the mid-mantle can be resolved by tomography (see chapter 3 of this dissertation) implying that we are on the verge of imaging some short-wavelength mid-mantle anomalies.

There appear to be several slow features extending through the mid mantle. The large slow feature beneath the Azores (Fig. 2.18) is the most notable and is in a moderately well-resolved region. Slow material also extends upward from the CMB beneath South Africa towards the mid-ocean ridge south of Africa. A generally slow, low amplitude region in the Central and Western Pacific connects the high amplitude slow anomalies at the CMB to the slow regions in the upper mantle. As stated above, resolution suffers greatly in the mid mantle in both the Central Pacific and beneath South Africa, so that these connections may be less (or more) substantial than they appear.

It is tempting to interpret the general features of the tomographic models in terms of the results of recent 3D simulations of mantle convection. Such an interpretation is dangerous given the relatively simple rheologies used in most convection simulations (see Tackley, et al. (1994) for a thorough discussion). For example, we might attribute the dominance of low spherical harmonic degree structure in the Earth (or, more precisely, the radial variation of power spectra) to the effects of the phase transition at 660 km depth (Tackley, et al., 1993) though the presence of plates is likely to be at least as important. As discussed in Jordan, et al. (1993), global diagnostics of mantle structure reveal major differences between the predictions of some 3D convection simulations and mantle tomography. RMS amplitude of shear velocity and the radial correlation function (RCF) introduced in Jordan, et al. (1993) are two simple diagnostics of interest. While 3D numerical convection simulations modeling the effects of the 400 and 660 phase changes predict high RMS amplitudes across the upper mantle and a sharp decrease at the 660 discontinuity (Tackley, et al., 1994), such a pattern is not evident in S16B30 (Fig 2.16a). The cross-sections and Fig. 2.16a show that large-amplitude structure is confined to about the top 450 km in S16B30 and apparently does not extend significantly into the transition zone. This is also true of model SH12WM13 where structure extends a little deeper but is essentially confined above 500 km. Chapter 3 shows that our tomographic methods should faithfully reproduce the change in RMS amplitude in the 3D mantle convection simulation of Tackley, et al. (1993) near the 660 discontinuity, suggesting that such a change in structure across the 660 does not exist in the Earth. The radial decorrelation in structure in the vicinity of the 660 in the RCF of this convection simulation is also shown to be well-recovered by the tomography, so the absence of such a feature in the RCF of S16B30 (fig. 2.19) also supports the conclusion that phase transition at 660 km may not be the dominant controlling factor in the convection of the actual mantle. The RCF of S16B30 shows only a slight decorrelation at shallower depths, which may be related to the disappearance of continental roots.

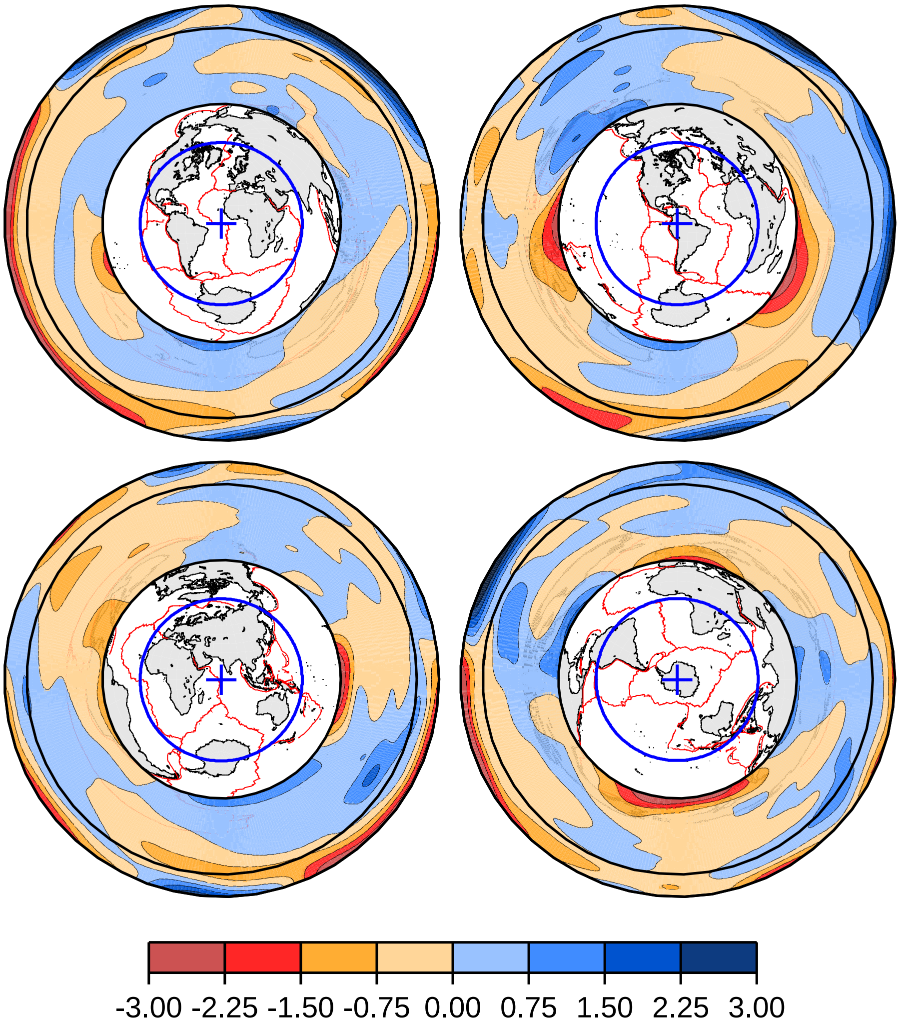


Figure 2.18: Cross-sections through model S16B30. The shear velocity perturbation relative to the model average at depth is contoured. The slices intersect the surface along the great circles plotted on azimuthal equidistant projections in the centers of the cross sections.

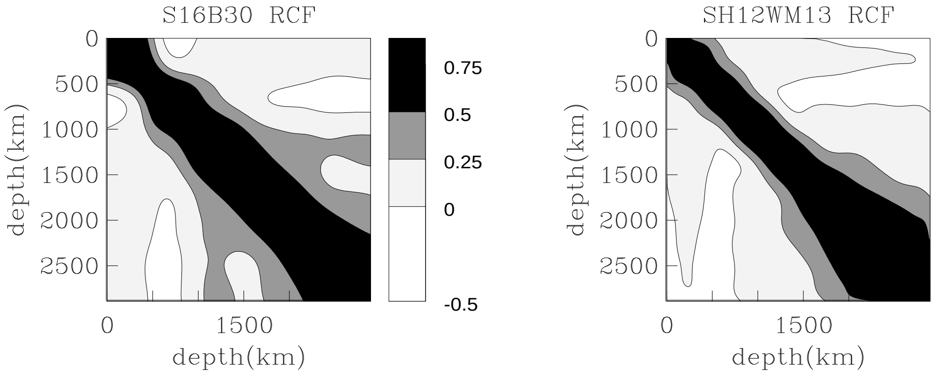


Figure 2.19: Radial Correlation Functions for models S16B30 and SH12WM13. These are contours of the cross-correlation between the model patterns on two spherical surfaces specified by the depths on the axes. The diagonal of the plot represents the correlation of model patterns at a given depth with themselves and therefore must have the value 1.0. If the model has similar structure in radius, the diagonal ridge will broaden, and vice versa. Recent numerical modeling of convection in a spherical shell suggests that the width of the diagonal correlation ridge with depth may provide some insight into the important factors in mantle convection. For example, neither of the RCF's for models S16B30 nor SH12WM13 have the narrowing of the correlation ridge predicted by numerical simulations of mantle convection with a phase change at the 670 km discontinuity.

2.6 Future directions

A comparison of models S16B30 and SH12WM13 suggests that mantle tomography has matured to the point where we can reliably recover the large-scale variation of shear velocity in the mantle. Clearly, neither model S16B30 nor model SH12WM13 is the final word in mantle tomography. Some improvement can be achieved simply by expanding the current data sets to take advantage of the rapidly growing digital global seismic network. Many of the problems associated with imaging structure in the southern hemisphere of the lower mantle can be expected to be addressed in this way. We should also be including boundary perturbations in the inversions though these tend to be poorly resolved with the existing data (Woodward & Masters, 1992). The addition of data sets of conversions and reflections from mantle discontinuities (Shearer, 1991b) should do much to help in this regard.

As the data sets expand and extend to higher frequencies, we will be tempted to try and resolve finer and finer scale structure. This will force an evaluation of the rather simple theoretical techniques currently employed to interpret the data. In particular, we still lack fast and/or accurate methods to calculate synthetic body wave and short-period surface wave seismograms for realistic 3D models. We also need to properly incorporate attenuation into the inversions and we need to evaluate the importance of anisotropy on a global scale. While rapid progress is being made on many of these fronts, global seismology still has a long way to go.

Acknowledgments

The text of this chapter, in part, appeared in Masters, G., S. Johnson, G. Laske, and H. Bolton,1996, A shear velocity model of the mantle, *Phil. Trans. R. Soc. Lond.*, **354A**, 1385–1411. The dissertation author was a secondary researcher and

Guy Masters directed and supervised the research which forms the basis for this chapter. The dissertation author was responsible for the process of constructing the shear velocity model S16B30 based on data collected and evaluated by the other authors. This research has been funded by grants from the National Science Foundation. G. Laske has been partially supported by the Green Foundation of La Jolla and S. Johnson has been partially supported by an IGPP/Los Alamos minigrant. The inversion computations were performed on an Intel Paragon parallel supercomputer at the San Diego Supercomputer Center using funds from the SIO block grant. We would also like to thank the network operators (USGS, IDA, GEOSCOPE) for producing such high-quality data and the IRIS DMC for making the data so easy to access.

A Appendix: A parallel QR algorithm

Global tomography seeks to determine N model parameters from M data. In the case of a linear (or linearized) relationship between the model and data, this can result in extensive computations to solve the resulting linear algebra problem. In translating the constrained least squares tomography problem to numerical computation, we have several choices of algorithm. These choices are tradeoffs of floating point operations, computer memory requirements and numerical stability.

Recent advances in parallel computing technology and parallel algorithms have combined to allow us to do very fast linear algebra on very large matrices and vectors. For example, recent experience with the tomography problems above has resulted in the manipulation of matrices several billion bytes (Gbytes) in size at computational rates of up to ten billion floating point operations per second (Gflops). These computational speeds are attained on clusters of fast workstation-level computing nodes, and so amplify the capabilities of fast single workstations roughly by the number of nodes, since many linear algebra tasks are very efficiently parallelized.

Repeating section 2.4.2, the equations we wish to solve are the normal equations:

$$\left[G^T G + \lambda \partial^T \partial\right] m = G^T d, \tag{A.0.1}$$

for a number of Lagrange multipliers λ. The matrix G is of dimension $M \times N$ and ∂ is of dimension $N \times N$. In the global seismology problems solved in chapters 2 and 3, $G^T G$ is positive definite and $\partial^T \partial$ is positive semidefinite. As well, the normal equations constitute a numerically stable problem given double precision arithmetic and the parameterizations used to specify the model m. In general, the numerical stability of the problem may change drastically with the Lagrange multiplier, and a numerically stable approach is important.

There are (at least) two direct method approaches to solving equation A.0.1. We do not discuss iterative methods (e.g., the conjugate gradient method). The method of joint diagonalization (JD) solves the normal equations in a computationally efficient way for many Lagrange multipliers, at least on serial machines, but with an unnecessarily large condition number γ^2. Of course, we note that there might be other reasons for desiring an SVD or eigenvalue-eigenvector decomposition in the process of solving an inverse problem (Gilbert, 1971). The QR method (Lawson & Hanson, 1974) solves a factored form of the normal equations, thereby reducing the effective condition number to γ.

The JD method (Gilbert, 1971) utilizes a matrix eigenvalue/eigenvector routine to decompose two matrices. Although the computational efficiency (either for serial or parallel platforms) of this algorithm is not discussed in detail, the algorithm is briefly described. The matrix $G^T G$ can be diagonalized so that $G^T G = S^T \Lambda S$, where Λ is the diagonal eigenvalue matrix and S contains the eigenvectors. We can now rewrite equation A.0.1 as

$$S^T \Lambda^{1/2} \left[I + \lambda \Lambda^{-1/2} \partial^T \partial \Lambda^{-1/2}\right] \Lambda^{1/2} S m = G^T d. \tag{A.0.2}$$

Diagonalizing the symmetric matrix $\Lambda^{-1/2}\partial^T\partial\Lambda^{-1/2}$ as $V^T\Xi V$,

$$S^T\Lambda^{1/2}\left[I+\lambda V^T\Xi V\right]\Lambda^{1/2}Sm=G^Td, \tag{A.0.3}$$

whereby

$$S^T\Lambda^{1/2}V^T\left[I+\lambda\Xi\right]V\Lambda^{1/2}Sm=G^Td. \tag{A.0.4}$$

Clearly, we can rewrite this as

$$\left[I+\lambda\Xi\right]V\Lambda^{1/2}Sm=V\Lambda^{-1/2}SG^Td. \tag{A.0.5}$$

Solving this equation is equivalent to two simple steps. First, perform the vector operation of obtaining the solution to

$$D_\lambda m'=d', \tag{A.0.6}$$

where D_λ is a diagonal matrix which is a simply obtained function of the Lagrange multiplier λ. Second, transform the vector m' to m using $m=S^T\Lambda^{-1/2}V^Tm'$. Since this procedure involves only a vector operation $(O(N))$ to obtain each model, and a matrix vector product $(O(N^2))$ to transform each model, each additional model only costs $(O(N^2))$. The diagonalization processes themselves are $(O(N^3))$ processes.

The problem with the JD process is the excessive condition number of the squared matrices in the normal equations in equation A.0.1. In order to avoid unnecessary numerical instability, we instead use a factored form of the normal equations. It is apparent that the normal equations are equivalent to $C^TCm=C^Td'$, where

$$C=\begin{bmatrix} G \\ \lambda^{1/2}\partial \end{bmatrix}, \tag{A.0.7}$$

and

$$d'=\begin{bmatrix} d \\ 0 \end{bmatrix}, \tag{A.0.8}$$

Constructing the QR decomposition of C, so that $C = QR$, where R is upper triangular and Q is orthogonal, we obtain $R^T Rm = R^T Q^T d'$. Assuming R^T is invertible, this is equivalent to $Rm = Q^T d'$. We then obtain m by back-substitution. While it can be shown that the effective condition number of this procedure is the square-root of that for JD (Parker, 1994), the number of operations per model is $O(N^3)$.

In practice, QR proceeds across the column space of the matrix C, converting C to R in the process. Q can be written as a sequence of *Householder* transformations:

$$Q^T = (I - \boldsymbol{w}_n \boldsymbol{w}_n^T)(\cdots)(I - \boldsymbol{w}_2 \boldsymbol{w}_2^T)(I - \boldsymbol{w}_1 \boldsymbol{w}_1^T), \tag{A.0.9}$$

where the transformation vectors are

$$\boldsymbol{w}_j = \frac{1}{\sqrt{s_j^2 - s_j A_{j-1}(j,j)}} \begin{bmatrix} 0 \\ \vdots \\ A_{j-1}(j,j) - s_j \\ A_{j-1}(j+1,j) \\ \vdots \\ A_{j-1}(m,j) \end{bmatrix}. \tag{A.0.10}$$

Given that we set

$$s_j = -sign(A_{j-1}(j,j))||A_{j-1}(j:m,j)||, \tag{A.0.11}$$

where

$$A_{j-1} = (I - \boldsymbol{w}_{j-1} \boldsymbol{w}_{j-1}^T)(\cdots)(I - \boldsymbol{w}_2 \boldsymbol{w}_2^T)(I - \boldsymbol{w}_1 \boldsymbol{w}_1^T)C. \tag{A.0.12}$$

It is easy to verify that we have established a recursive definition of Q with the desired properties; Q is orthogonal and upper-triangularizes the matrix A_0 in n steps. Note that $A_{j-1}(k,l)$ refers to the element in the kth row and lth column of the matrix A_{j-1}. This recursive definition of QR reduction is illustrated in figure 2.20.

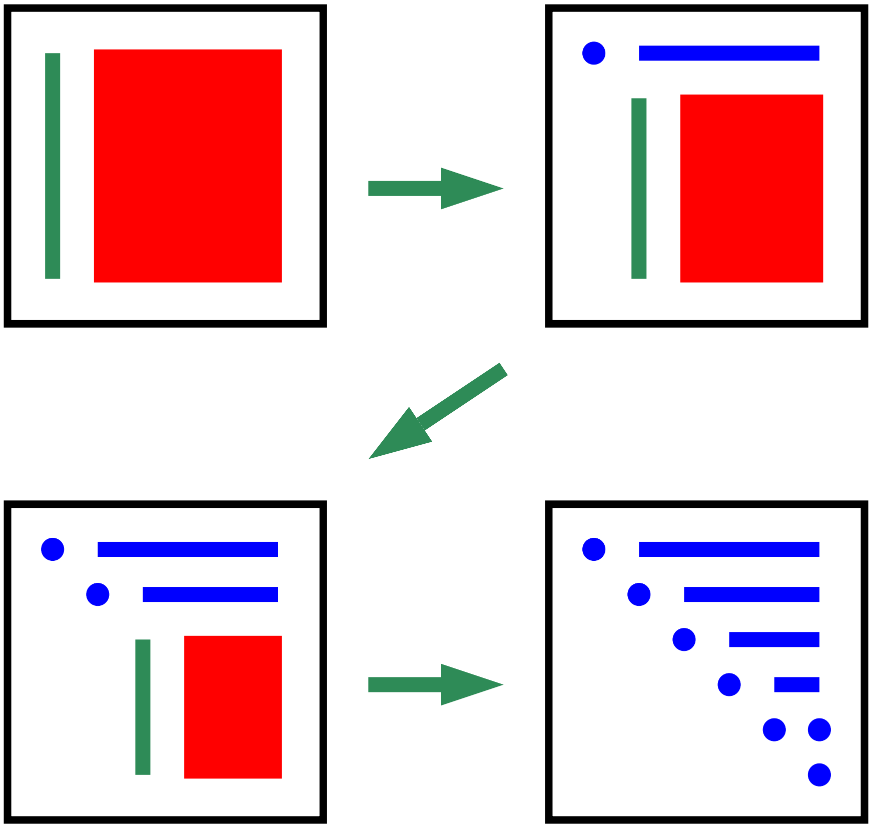


Figure 2.20: QR reduction of a square dense matrix. Shown are the matrices A_{j-1} in the $j = 1$, $j = 2$, $j = 3$, and $j = n + 1$ steps in the recursive algorithm described in equations A.0.9 - A.0.12. The green vectors are the basis for the formation of the $\boldsymbol{w}_j$ and the quantities in blue are completed parts of the matrix $R = A_n$. The red and green sections of the matrix are those components operated on by the corresponding Householder transformation. The blue dots on the diagonal end up being the s_j defined in equation A.0.11.

QR factorization can also operate on a subset of C and d' at a time. This process, known as sequential QR, subdivides the rows of the problem into several sections. Schematically, the application of Q^T would appear as the sequence of operations

$$\begin{bmatrix} R_a \\ 0 \\ C_3 \\ C_4 \end{bmatrix} = \begin{bmatrix} Q_{a1} & Q_{a2} & 0 & 0 \\ Q_{a3} & Q_{a4} & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}^T \begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{bmatrix}, \tag{A.0.13}$$

$$\begin{bmatrix} R_b \\ 0 \\ 0 \\ C_4 \end{bmatrix} = \begin{bmatrix} Q_{b1} & 0 & Q_{b2} & 0 \\ 0 & I & 0 & 0 \\ Q_{b3} & 0 & Q_{b4} & 0 \\ 0 & 0 & 0 & I \end{bmatrix}^T \begin{bmatrix} R_a \\ 0 \\ C_3 \\ C_4 \end{bmatrix}, \tag{A.0.14}$$

and finally,

$$\begin{bmatrix} R \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} Q_{c1} & 0 & 0 & Q_{c2} \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ Q_{c3} & 0 & 0 & Q_{c4} \end{bmatrix}^T \begin{bmatrix} R_b \\ 0 \\ 0 \\ C_4 \end{bmatrix}. \tag{A.0.15}$$

These expressions represent the QR decomposition of matrices with two distinct structures. Expression A.0.13 is the decomposition of the dense matrix composed of C_1 and C_2. This QR algorithm will be referred to as QR1. Expressions A.0.14 and A.0.15 are operations on the matrices composed of upper triangular sections R_a and R_b and the dense submatrices C_3 and C_4. We will refer to this algorithm as QR2.

The sequential QR process is therefore composed of a single application of QR1 and repeated application of QR2. Once sequential QR has completed with the data matrix, the final QR algorithm is needed. This algorithm applies QR to the composition of two upper triangular matrices, the R matrix from the data, and the ∂ matrix from equation A.0.7. This algorithm is referred to as QR3.

QR2 and QR3 are distinct from QR1 in that they avoid operations on quantities known to be zero. QR3 is additionally tailored to make the most of "extra"

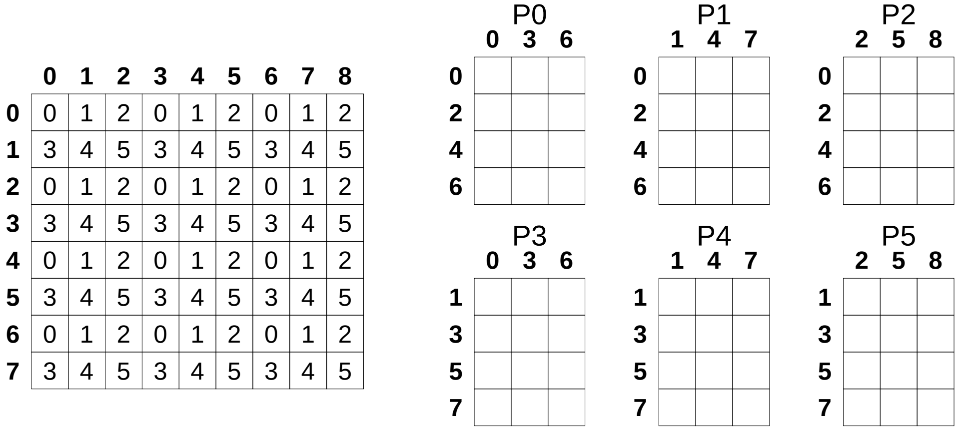
Table 2.1: $QR1$, $QR2$ and $QR3$ operations and operation counts . The submatrices listed are: $D = n \times n$ dense, $R = n \times n$ upper triangular, $\partial = n \times n$ upper triangular and $d = n \times nd$ dense.

<table>
<tr><td>$QR1$</td><td>$\left[\begin{array}{c|c} D & d \end{array}\right] \rightarrow \left[\begin{array}{c|c} R & d' \end{array}\right]$</td><td>$\frac{2n^3}{3}(+nd\;n^2)$</td></tr>
<tr><td>$QR2$</td><td>$\left[\begin{array}{c|c} R_1 & d_1 \\ D & d_2 \end{array}\right] \rightarrow \left[\begin{array}{c|c} R & d' \\ 0 & \delta \end{array}\right]$</td><td>$n^3(+nd\;n^2)$</td></tr>
<tr><td>$QR3$</td><td>$\left[\begin{array}{c|c} R_1 & d \\ \lambda^{1/2}\partial & 0 \end{array}\right] \rightarrow \left[\begin{array}{c|c} R & d' \\ 0 & \delta \end{array}\right]$</td><td>$\frac{n^3}{3}(+nd\;n^2)$</td></tr>
</table>

space in the lower triangle. This is fairly important in avoiding unnecessary disk IO, which can be a significant bottleneck on parallel supercomputers. QR2 and QR3 utilize copy operations which assure that the application of Householder transformations occurs on vectors which are, as much as possible, adjacent in memory. This maximizes the advantages of vectorization on each node. The functionality and asymptotic (large n) operation counts of the three QR algorithms are summarized in table 2.1.

A.1 Dense linear algebra in parallel

It is fairly straightforward to understand the advantages of parallel computation for dense linear algebra. It is somewhat less obvious how to keep all processors busy in a given task on global matrices distributed across all processors. A very fruitful approach is to distribute the necessary matrices and vectors across the processor grid in some type of interleaved fashion. This is typically referred to in the literature as *square-block-scattered* (SBS). An example of an SBS matrix is shown in figure 2.21. The details of parallel computation with SBS matrices are discussed in Blackford, et al. (1997), and more exhaustively in papers referenced therein.



	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0	0	1	2	0	1	2	0	1	2
1	3	4	5	3	4	5	3	4	5
2	0	1	2	0	1	2	0	1	2
3	3	4	5	3	4	5	3	4	5
4	0	1	2	0	1	2	0	1	2
5	3	4	5	3	4	5	3	4	5
6	0	1	2	0	1	2	0	1	2
7	3	4	5	3	4	5	3	4	5

Figure 2.21: An 8 row × 9 column matrix *square-block-scattered* across a 2 row × 3 column processor grid with a blocking factor of $nb = 1$. A more typical value is $nb > 10$. The matrix row and column is indicated in bold numbers. The matrix elements are labeled with the number of the processor on which the element is actually stored. This scheme assures that operations on a submatrix of any size are always distributed across all processors. The right diagram is the same matrix as stored on each processor of the grid (P0 – P5). (After Blackford, et al. (1997))

The SBS data distribution lends itself to an easy *single program, multiple data* (SPMD) programming style. All nodes follow distinct execution paths through the same code. The location of any part of the distributed matrix is quickly calculated on any node, and appropriate action taken by both the senders and receivers of intermediate results.

$QR1$, $QR2$, $QR3$ and miscellaneous operations of importance in the solution of our constrained least squares problem were developed from an early version (1.0 beta) of the SCALAPACK linear algebra code described in Blackford, et al. (1997). A code map describing the important components of the resulting code is included in figure 2.22. Modifications and extensions of the original SCALAPACK code were made to implement the $QR1$,$QR2$ and $QR3$ algorithms and to make use of the unused lower triangular space to minimize IO.

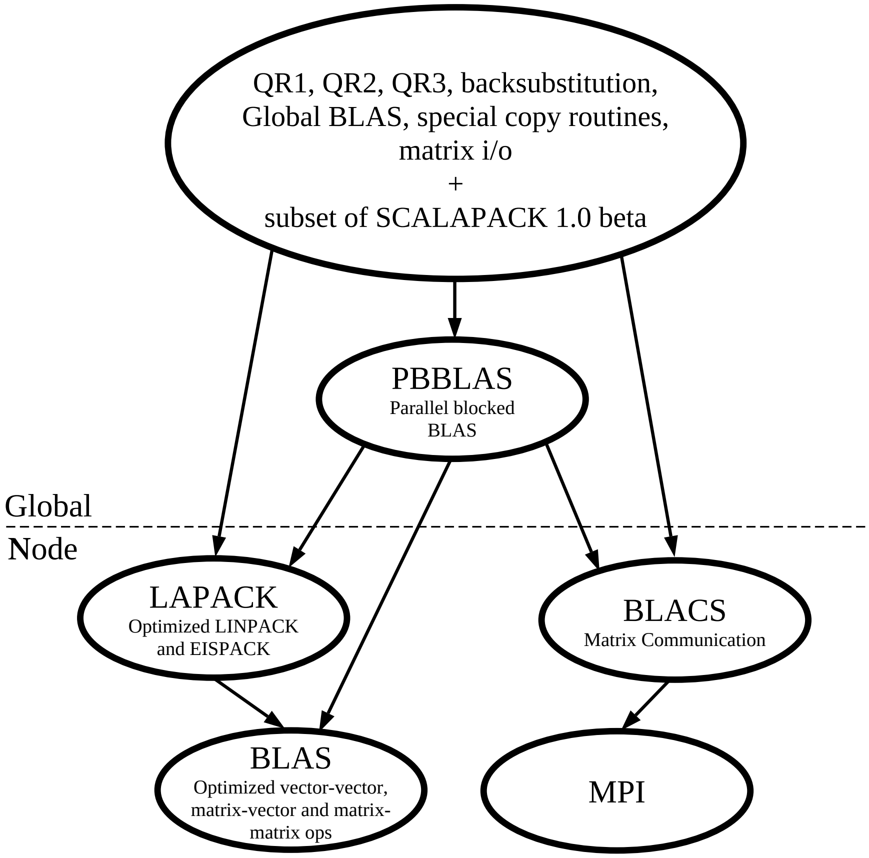


Figure 2.22: Map of parallel linear algebra code used in global tomography of chapters 2 and 3. The performance of the global or parallel algorithms above the dotted line is due to calls to the optimized BLAS routines on the individual nodes of the processor grid. (Modified from Blackford, et al. (1997))

Parallel performance

The parallel efficiencies of the $QR1$,$QR2$ and $QR3$ algorithms are evaluated by first determining the performance of a single node on the same problem, and then timing the algorithms on multiple nodes. Parallel efficiency is computed from

$$E = \frac{t_s}{Pt_p}, \tag{A.1.16}$$

where t_s is the single node computation time, t_p the parallel time and P the number of nodes. Single node performance is shown in table 2.2, where 4 QR algorithms and a matrix-matrix multiply are timed, and performance determined using the operation counts of table 2.1.

Table 2.2: T3E single node performance for 30-50Mb/node.

Algorithm	QR1	QR2	QR3	SGEQRF	SGEMM
Mflop/s	180	180	160	150	280

Since the parallel algorithms will be run on problems which will consume 30-50 Mbytes per node, the single node tests are adjusted to consume the same amount. The performance only varies a few percent over the 30-50 Mbyte range. For the parallel algorithms, the algorithms are run on 1 node with the same blocking factor *nb* (Blackford, et al., 1997) as the parallel runs. *nb* is set to 32 for all runs presented. The best performance is obtained with the BLAS level 3 routine SGEMM (Anderson, et al., 1992), which executes a matrix-matrix multiply. SGEMM is optimized to make the maximum use of adjacent memory, and vectorizes very well due to the fact that all vector operations are as large as possible for a matrix of the given size. All of the QR routines suffer from the diminishing vector length and lack of adjacent memory use which are intrinsic to the QR process. The units in table 2.2 are Mflop/s, or million floating point operations per second, where a floating point operation is either a multiply or an add.

Parallel efficiency is shown in table 2.3. The two runs described are both used in research in this dissertation. Run 1 is used to used to find model S16B30 (this chapter), and Run 2 is used in Chapter 3. Run 2 is more efficient due to higher compute to communicate ratios. These efficiencies are certainly acceptable.

Efficient IO on parallel platforms

The matrices G and ∂ constructed for the above problems result from the execution of a number of programs. These programs perform a variety of tasks, generating a number of subproblems which eventually lead to the construction

Table 2.3: Parallel efficiency E for two runs. Run 1 is a problem with parameters $n = 8670$,$nd = 441$ consuming about 27 Mbytes per node on 64 nodes. Run 2 has parameters $n = 3150$,$nd = 199$ and consumes about 52 Mbytes per node on 4 nodes.

Algorithm	QR1	QR2	QR3
E(Run 1)	89%	85%	82%
E(Run 2)	95%	92%	95%

of the final G matrix. Since these codes are stable and quick to execute on local workstations, the actual G matrices are not constructed in memory on the supercomputer. These matrices are constructed elsewhere and sent to the supercomputer in manageable sections via ftp. Since the problems of interest frequently involve several Gbytes of G matrix, a fast way of loading these disk files into distributed memory on the parallel platform is needed.

Generally speaking, performance IO is obtained in two ways. First, and most importantly, large IO requests must be made. Table 2.4 shows an example of this for the T3E. In the case of SBS matrix IO, this requires redistribution of data from one or more IO nodes to the rest of the processor grid. Second, the disk file can be *striped* over several physical disk devices and IO can occur simultaneously to several large blocks of the file at once. To make use of striping, IO requests from multiple nodes of the processor grid must be sequenced, which adds additional complexity to the IO code. The ability to stripe files on the T3E is fairly straightforward if IO is performed in C (as it is). However, striped IO is not essential for the problems solved in this dissertation, which can be seen as follows.

Assume for the moment that we can achieve single node peak IO rates. Most of the disk IO is associated with the $QR2$ algorithm. For a problem running on P nodes, the IO and CPU times can be predicted from tables 2.4 and 2.2. Assuming double precision arithmetic (one has no other choice on the T3E),

Table 2.4: IO rate versus IO request size for a single node of the T3E. Tests are for loading approximately 42 Mbytes in sections of the given request size.

IO request(Kbytes)	16	65	256	1024
IO rate(Mbytes/sec)	5	8	17	38

and approximating parallel efficiency as 100%, IO times should be about $8n^2/38$, whereas CPU is roughly $2n^3/180P$. This results in a predicted CPU/IO ratio of $n/19P$. CPU/IO is therefore predicted to be approximately 7 for Run 1, and about 40 for Run 2. Due to technicalities in the way the large-request IO is actually implemented, the observed rates differ for Run 2, yielding a CPU/IO ratio of no more than 20. In any case, this indicates that IO is not a large part of the actual run time for problems of interest. In general, this will not be the case, and further work on scalable IO is indicated.

2.7 References

Anderson, E., Bai, Z., Bischof, C., Demmel, J., Dongarra, J., DuCroz, J., Greenbaum, A., Hammarling, S., McKenney, A., Ostrouchov, S., Sorenson, D. 1992 *LAPACK User's Guide.* SIAM, Philadelphia.

Blackford, L., Choi, J., Cleary, A., D'Azevedo, E., Demmel, J., Dhillon, J., Dongarra, J., Hammarling, S., Henry, G., Petitet, A., Stanley, K., Walker, D. & Whaley, R.C. 1997 *ScaLAPACK Users' Guide.* SIAM, Philadephia.

Bolton, H. & Masters, G. 1991 Long Period Absolute P Times and Lower Mantle Structure. *EOS Trans AGU* **72**, 339.

Christensen, N.I. & Mooney, W.D. 1995 Seismic velocity structure and the composition of the continental crust: a global view. *J. Geophys. Res.* **100**, 9761–9788.

Constable, S.C., Parker, R.L. & Constable, C.G. 1987 Occam's Inversion: A practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics* **52**, 289–300.

Doornbos, D.J. 1983 Observable effects of the seismic absorption band in the Earth. *Geophys. J. R. Astron. Soc.* **75**, 693–711.

Dziewonski, A.M. 1984 Mapping the lower mantle: determination of lateral heterogeneity in P velocity up to degree and order 6. *J. Geophys. Res.* **89**, 5929–5952.

Dziewonski, A.M. & Anderson, D.L. 1981 Preliminary reference Earth model. *Phys. Earth Planet. Inter.* **25**, 297–356.

Dziewonski, A.M. & Woodhouse, J.H. 1987 Global images of the Earth's interior. *Science* **236**, 37–48.

Edmonds, A.R. 1960 *Angular Momentum and Quantum Mechanics.* Princeton University Press, Princeton, N.J..

Forte, A.M., Dziewonski, A.M. & Woodward, R.L. 1993 Aspherical structure of the mantle, tectonic plate motions, nonhydrostatic geoid, and topography of the core-mantle boundary. In: *Dynamics of the Earth's Deep Interior and Earth Rotation, Geophysical Monography 72, IUGG vol 12, eds LeMouel, Smylie, and Herring.*, pp.135–166.

Gilbert, F. 1971 Ranking and winnowing gross Earth data for inversion and resolution. *Geophys. J. R. Astr. Soc.* **23**, 125 – 128.

Grand, S.P. 1994 Mantle shear structure beneath the Americas and surrounding oceans. *J. Geophys. Res.* **99**, 11.591–11.621.

Hager, B.H. & Clayton, R.W. 1989 Constraints on the structure of mantle convection using seismic observations, flow models, and the geoid. In: *Mantle Convection, ed W.R. Peltier, Gordon and Breach, New York*, pp.657–763.

Inoue, H., Fukao, Y., Tanabe, K. & Ogata, Y. 1990 Whole mantle P-wave travel time tomography. *Phys. Earth Planet. Inter.* **59**, 294–328.

Jordan, T.H. 1978a A procedure for estimating lateral variations from low-frequency eigenspectra data. *Geophys. J. R. Astron. Soc.* **52**, 441–455.

Jordan, T.H. 1978b Composition and development of the continental tectosphere. *Nature* **274**, 544–548.

Jordan, T.H. & Lynn, W.S. 1974 A velocity anomaly in the lower mantle. *J. Geophys. Res.* **79**, 2679–2685.

Jordan, T.H., Puster, P., Glatzmaier, G.A. & Tackley, P.J. 1993 Comparisons of seismic earth structures and mantle flow models using radial correlation functions. *Science* **261**, 1427–1431.

Karato, S. 1993 Importance of anelasticity in the interpretation of seismic tomography. *Geophys. Res. Lett.* **20**, 1623–1626.

Laske, G. 1993 *The frequency-dependent polarization of long-period surface waves and its implication for global phase velocity maps.* Doctoral Thesis, Universität Karlsruhe , Germany.

Laske, G., Masters, G. & Zürn, W. 1994 Frequency dependent polarization measurements of long period surface waves and their implications for global phase velocity maps. *Phys. Earth Planet. Int.* **84**, 111 – 137.

Laske, G. & Masters, G. 1996 Constraints on global phase velocity maps by long-period polarization data. *J. Geophys. Res.* **101**, 16059 – 16075.

Lawson, L. & Hanson, R. 1974 *Solving least squares problems.* Prentice Hall, New Jersey.

Masters, G. & Bolton, H. 1991 Large-scale shear velocity structure of the mantle. *EOS, Trans AGU* **72**, 316.

Masters, G., Bolton, H. & Shearer, P. 1992 Large-scale 3-dimensional structure of the mantle. *EOS, Trans. AGU* **73**, 201.

Masters, G., Johnson, S. & Sheehan, A.F. 1993 $SS - S$ and the depth of shear velocity anomalies under ridges. *EOS Trans. AGU* **74**, 76.

Masters, G., Jordan, T.H., Silver, P.G. & Gilbert, F. 1982 Aspherical earth structure from fundamental spheroidal-mode data. *Nature* **298**, 609–613.

Montagner, J.-P. & Tanimoto, T. 1991 Global upper mantle tomography of seismic velocities and anisotropies. *J. Geophys. Res.* **96**, 20,337–20,351.

Nataf, H.-C., Nakanishi, I. & Anderson, D.L. 1986 Measurements of mantle wave velocities and inversion for lateral heterogeneities and anisotropy, 3. Inversion. *J. Geophys. Res.* **91**, 7261–7307.

Parker, R.L. 1994 *Geophysical Inverse Theory.* Princeton Univ. Press, Princeton.

Park, J., Vernon III, F.L. & Lindberg, C.R. 1987 Frequency dependent polarization analysis of high-frequency seismograms. *J. Geophys. Res.* **92**, 12,664–12,674.

Pulver, S. & Masters, G. 1990 $PcP - P$ travel times and the ratio of S to P velocity variations in the lower mantle. *EOS Trans. AGU* **71**, 1464.

Ricard, Y., Richards, M., Lithgow-Bertollini, C. & LeStunff, Y. 1993 A geodynamic model of mantle density heterogeneity. *J. Geophys. Res.* **98**, 21,895–21,909.

Ritzwoller, M., Masters, G. & Gilbert, F. 1986 Observations of anomalous splitting and their interpretation in terms of aspherical structure. *J. Geophys. Res.* **91**, 10,203–10,228.

Ritzwoller, M., Masters, G. & Gilbert, F. 1988 Constraining aspherical structure with low frequency interaction coefficients: Application to uncoupled multiplets. *J. Geophys. Res.* **93**, 6369–6396.

Scott, J.S., Masters, T.G. & Vernon, F.L. 1994 3-D velocity structure of the San Jacinto fault zone near Anza, California – I. P waves. *Geophys. Journ. Int.* **119**, 611 – 626.

Shaw, P.R. & Orcutt, J.A. 1985 Waveform inversion of seismic refraction data and applications to young Pacific crust. *Geophys. J. R. Astron. Soc.* **82**, 375–414.

Shearer, P.M. 1990 Seismic imaging of upper-mantle structure with new evidence for a 520-km discontinuity. *Nature* **344**, 121–126.

Shearer, P.M. 1991a Imaging global body wave phases by stacking long-period seismograms. *J. Geophys. Res.* **96**, 20,353–20,364.

Shearer, P.M. 1991b Constraints on upper mantle discontinuities from observations of long-period reflected and converted phases. *J. Geophys. Res.* **96**, 18,147–18,182.

Shearer, P.M. & Masters, G. 1992 Global mapping of topography on the 660 km discontinuity. *Nature* **355**, 791–796.

Smith, M.F. 1989 *Imaging the Earth's aspherical structure with free oscillation frequency and attenuation measurements.* Ph.D Thesis, Univ. Cal. San Diego, La Jolla, CA.

Smith, M.F. & Masters, G. 1989 Aspherical structure constraints from free oscillation frequency and attenuation measurements. *J. Geophys. Res.* **94**, 1953–1976.

Soller, D.R., Ray, R.D. & Brown, R.D. 1982 A new global crustal thickness model. *Tectonics* **1**, 125–149.

Su, W-J. & Dziewonski, A.M. 1991 Predominance of long-wavelength heterogeneity in the mantle. *Nature* **352**, 121–126.

Su, W-J., Woodward, R.L. & Dziewonski, A.M. 1994 Degree 12 model of shear velocity heterogeneity in the mantle. *J. Geophys. Res.* **99**, 6945–6980.

Tackley, P.J., Stevenson, D.J., Glatzmaier, G.A. & Schubert, G. 1993 Effects of an endothermic phase transition at 670 km depth on a spherical model of convection in the Earth's mantle. *Nature* **361**, 699–704.

Tackley, P.J., Stevenson, D.J., Glatzmaier, G.A. & Schubert, G. 1994 Effects of multiple phase transitions in a three dimensional spherical model of convection in Earth's mantle. *J. Geophys. Res.* **99**, 15,877–15,901.

Tanimoto, T. 1988 The 3-D shear wave structure in the mantle by overtone waveform inversion. II, Inversion of X-waves, R-waves and G-waves. *Geophys. J.* **93**, 321–333.

Tanimoto, T. 1990 Long-wavelength S-velocity structure throughout the mantle. *Geophys. J. Int.* **100**, 327–336.

Thomson, D.J. 1982 Spectrum estimation and harmonic analysis. *IEEE Proc.* **70**, 1055–1096.

Woodhouse, J.H. & Dahlen, F.A. 1978 The effect of a general aspherical perturbation on the free oscillations of the earth. *Geophys. J. R. Astron. Soc.* **53**, 335–354.

Woodhouse, J.H. & Dziewonski, A.M. 1984 Mapping of the upper mantle: three-dimensional modeling of earth structure by inversion of seismic waveforms. *J. Geophys. Res.* **89**, 5953–5986.

Woodhouse, J.H. & Wong, Y.K. 1986 Amplitude, phase and path anomalies of mantle waves. *Geophys. J. R. Astron. Soc.* **87**, 753–773.

Woodward, R.L., Forte, A.M., Su, W. & Dziewonski, A.M. 1993 Constraints on the large-scale structure of the mantle. In: *Geophys. Monograph 74, IUGG, v14, Takahashi et al., eds. , AGU, Washington DC*, pp.89–109.

Woodward, R.L. & Masters, G. 1991a Lower mantle structure from $ScS - S$ differential travel times. *Nature* **352**, 231–233.

Woodward, R.L. & Masters, G. 1991b Global upper mantle structure from long-period differential travel-times. *J. Geophys. Res.* **96**, 6351–6377.

Woodward, R.L. & Masters, G. 1992 Upper mantle structure from long-period differential travel times and free-oscillation data. *Geophys. J. Int.* **109**, 275–293.

Chapter 3

Imaging a convecting mantle with seismic data

3.1 Abstract

Numerical simulations of fully three-dimensional flow in a spherical shell near the regime of convection believed to characterize the Earth's mantle have produced synthetic snapshots of mantle internal structure. As these synthetic mantle structures become increasingly realistic, they can be used in conjunction with seismic data for two purposes: 1) evaluation of tomographic methods, and 2) constraining the important controlling factors and parameters of mantle convection. Two simulations of mantle convection which include "earth-like" features have been tested with these two purposes in mind. This test consists of constructing two sets of tomographic images: first, images based on synthetic seismic data from a snapshot of a numerical convection experiment and, second, images based on actual seismic data. The first images are compared to the original snapshot to evaluate the effect of tomographic imaging, and the first and second sets of images are compared to see if the Earth exhibits similar features. Based on this test, three general conclusions are made. First, the convection simulations have generated mantle structure which is similar to what appears to exist in the actual

Earth in terms of RMS variability and distribution of spatial scales. This is very significant in that convection in a homogeneous spherical shell exhibits a very different range of spatial scales. Second, seismic tomography is capable of imaging the simulated mantles. For example, tomography is able to see the strong "layering" effects of the upper-lower mantle transition of one convection simulation on the resulting mantle structure. Third, the Earth does not exhibit the same type of structure seen in the simulated convection experiments, implying that further exploration into the controlling factors of convection in the Earth is necessary.

3.2 Introduction

Over the last decade, seismologists have been developing global data sets to constrain the large scale velocity structure of the Earth's mantle. Such data sets include differential and absolute traveltime data sets, phase velocity maps and free oscillation mode peak shifts. These data constrain large-scale structure throughout the mantle, and must therefore tell us something about the nature of mantle convection.

Comparing tomographic images of structure to synthetic convection structure suffers from the drawback that the reliability of the tomographic method has not been taken into account. In order to address this problem, we test the synthetic structures in a more direct fashion. Our testing entails three steps: 1) generate synthetic data from the synthetic structure and compare qualitatively with actual data, 2) add noise to the synthetic data and construct tomographic images of the synthetic structure, 3) compare the tomographic images of the synthetic structure to tomographic images constructed from the actual data. Clearly, this process holds the promise of both informing us about the nature of tomographic imaging and the state of convection in the Earth's mantle.

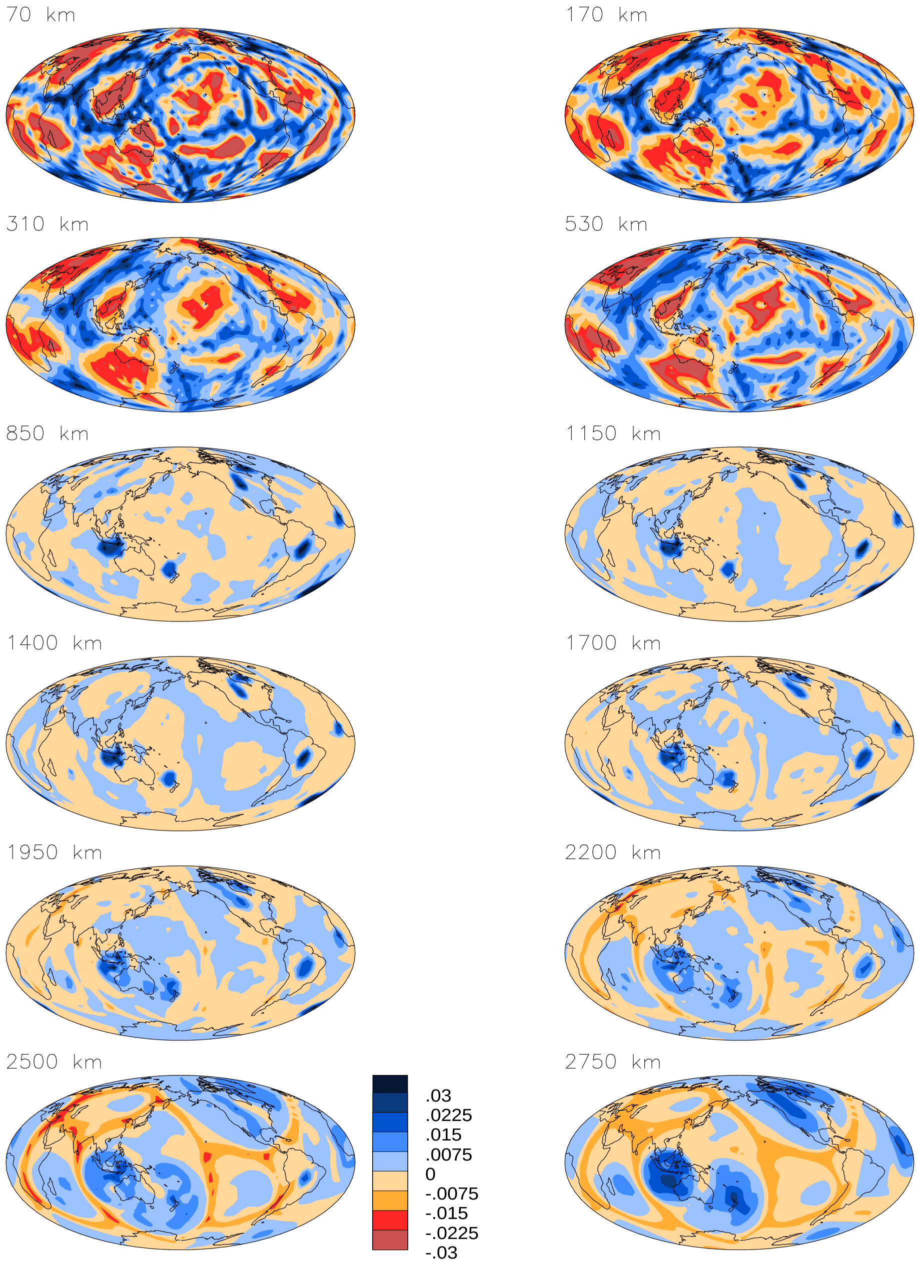


Figure 3.1: Twelve depth slices of the final frame (CVCT1) of a convection simulation. Contoured values are shear velocity perturbation relative to the global model average at a given depth. Continents are for scale only.

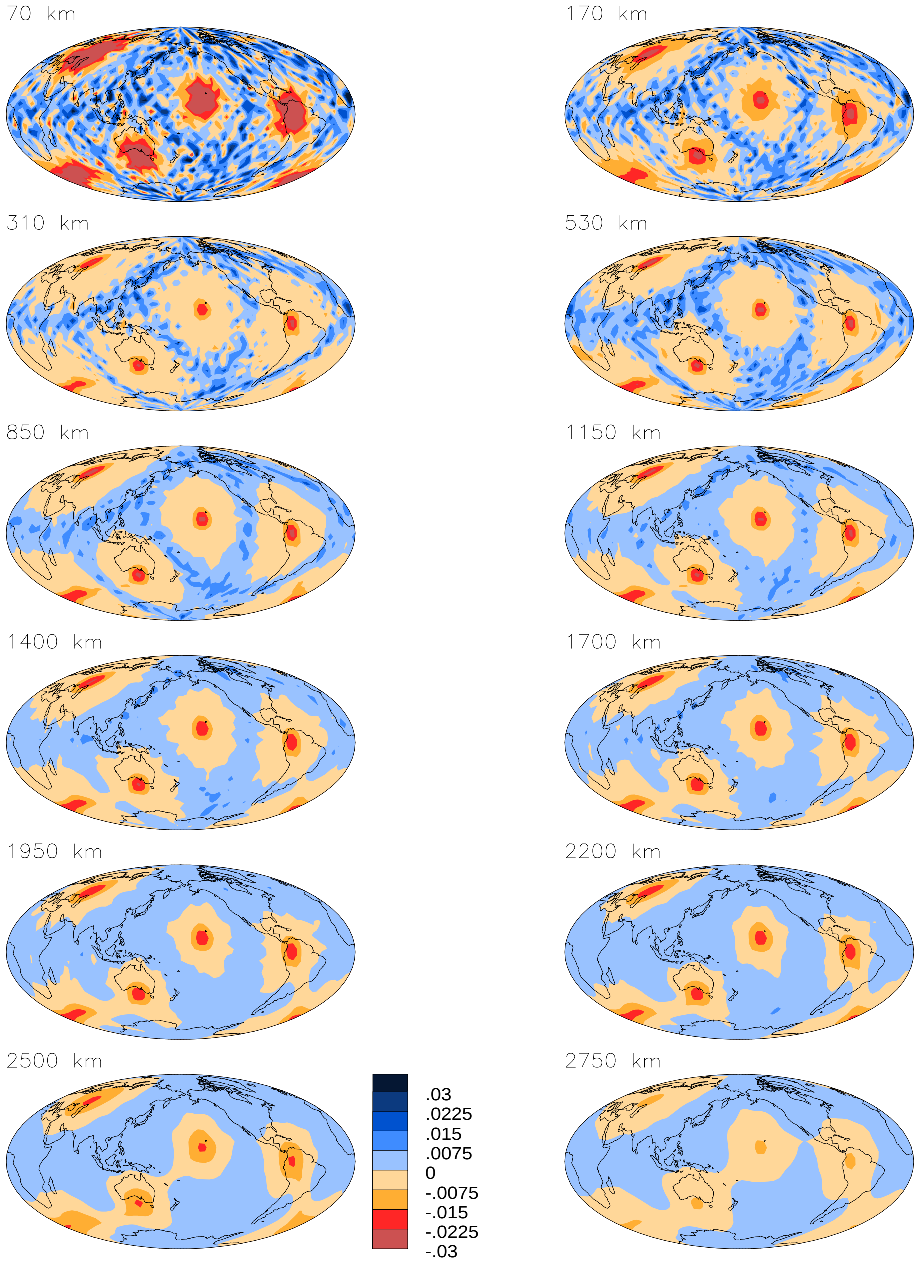


Figure 3.2: Twelve depth slices of the final frame (CVCT2) of a convection simulation.

3.3 The convection simulations

The above procedure is carried out on the final frames of two convection simulations. The first is described in Tackley, et al. (1993). This frame (denoted CVCT1) is a solution of compressible, self-gravitating flow, with radially varying viscosity and isothermal boundary conditions. An endothermic phase transition is present at 670km depth with a Clapeyron slope of -4 MPaK^{-1}, which is very likely to have a stronger effect on the flow than the phase transition in the actual Earth (Tackley, et al., 1993). This phase transition impedes the flow of material from the upper to lower mantle. The flow change at 670km depth results in a time dependent flow pattern consisting of the intermittent flushing of cold material into the lower mantle. The simulation is thereby given a distinct three-layer structure: a high variability upper mantle, a low variability mid mantle whose characteristic features are narrow stems of cylindrical downwellings, and a lowermost mantle of downwellings spreading out on the core-mantle boundary. In CVCT1, 40 % of the surface heat flow comes from the core. The convection frame scaled to shear velocity perturbations is shown in figure 3.1.

The second simulation (CVCT2) is an unpublished result by the same author, and possesses two phase changes, one at 410km depth, and the second at 670km depth. These phase changes have Clapeyron slopes of +3 MPaK^{-1} and -3 MPaK^{-1}, respectively. These Clapeyron slopes are considered to be more representative of the Earth than the -4 MPaK^{-1} of CVCT1 (personal communication, P. Tackley). Although the 670km phase change has a similar effect on slowing the flow (as in CVCT1), the effect of the 410km phase change has the opposite effect. This model also features a viscosity increase of a factor of 30 at 670km depth, also believed to characterize the actual Earth. The heat flow from the core is 40 % of the surface heat flow in CVCT2. The structure of CVCT2 is very different (see figure 3.2) than that of CVCT1, being dominated by several strong upwelling plumes with linear, time dependent downwellings (personal communication, P.

Tackley).

Conversion of the convection simulation state variables to seismic observables (density and compressional- or shear-wave velocity) is accomplished through the use of experimentally determined scalings. CVCT1 and CVCT2 are specified as entropy and pressure fields, and the entropy field can be converted directly to density using scalings which are an intrinsic part of the convection simulation. The effect of pressure perturbations on the the density is on the order of 1% of the entropy effect (personal communication, P. Tackley) and is ignored. The density perturbations are then scaled to shear velocity perturbations using $d\log\beta/d\log\rho = 2.5$ (Anderson, 1989). The density profile of the spherical reference state of the convection frame is within several percent of PREM at all radii. We will therefore make the approximation that PREM is the spherical reference model of CVCT1 and CVCT2. This is convenient for the evaluation of our seismic forward problem.

Two diagnostics will be used to characterize the nature of the velocity structures of CVCT1, CVCT2 and the tomographic results. These are the radial correlation function (RCF) (Jordan, et al., 1993) and the RMS variability vs. depth (RMSD). The RMSD is straightforward, and the RCF is simply the correlation of the velocity heterogeneity at any two given depths. Explicitly,

$$\mathrm{RCF}(d_1, d_2) = \frac{\iint\, d\Omega\; v(d_1)v(d_2)}{\left(\iint\, d\Omega\; v(d_1)v(d_1)\right)^{1/2}\left(\iint\, d\Omega\; v(d_2)v(d_2)\right)^{1/2}}, \tag{3.3.0.1}$$

where $v(d_1)$ is the velocity heterogeneity in a spherical shell relative to the spherical average at d_1 and $d\Omega$ is the area element on the sphere. $\mathrm{RCF}(d_1, d_1) = 1$, and rapid changes in V with depth will result in constrictions in the width of the RCF maximum near the diagonal. We note that the 95% confidence level for a spherical harmonic degree 14 RCF is .36.

The RMSD and RCF of CVCT1 are shown in figures 3.3 and 3.5, and clearly show the three layer nature of the flow pattern and the large jump in amplitude

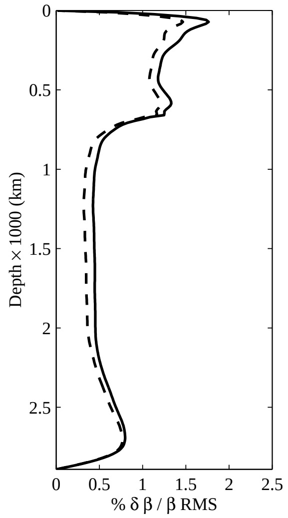


Figure 3.3: RMS amplitude of velocity heterogeneity in CVCT1 relative to the spherical average velocity at each depth. The dashed line corresponds to the truncation of CVCT1 to spherical harmonic degree 14. The prominent feature is the rapid decrease in amplitude below 670km depth.

across 670km depth. The same plots for CVCT2 are shown in figures 3.4 and 3.6. Clearly, the RCF is not such an interesting diagnostic of this simulation. The broadening of the RCF with depth is a characteristic of convection simulations in a mantle featuring an increase of viscosity with depth (personal communication, P. Tackley). For both CVCT1 and CVCT2, it is clear from these figures that most of the structure lies below spherical harmonic degree 14. This is important since our tomographic parameterization extends to degree 14.

We will also use the cross correlation as a function of depth between two functions u and v defined on the sphere

$$\sigma_{uv}(d) = \frac{\iint \, d\Omega \; u(d)v(d)}{\left(\iint \, d\Omega \; u(d)u(d)\right)^{1/2} \left(\iint \, d\Omega \; v(d)v(d)\right)^{1/2}}. \tag{3.3.0.2}$$

The entropy field of CVCT1 is parameterized in the convection simulations as spherical harmonics to degree 127, and in separate Chebyshev expansions in the upper and lower mantle. These upper and lower mantle polynomials are of order 13 and 37, respectively. CVCT2 is specified as 2 upper and 1 lower mantle layer, parameterized by Chebyshev polynomials of order 13, 13 and 33, respectively. For practical reasons, only the coefficients to spherical harmonic degree 64 are used in this work.

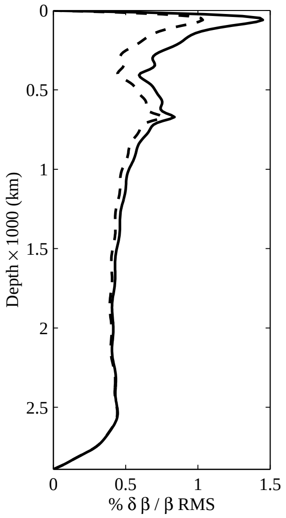


Figure 3.4: RMS amplitude of velocity heterogeneity in CVCT2 relative to the spherical average velocity at each depth. The dashed line corresponds to the truncation of CVCT2 to spherical harmonic degree 14.

3.4 Data

The seismic data sets used in our inversion for Earth structure are discussed in chapter 2. The data sets are composed of three distinct groups, each with its own methods of measurement and interpretation. Traveltimes of both direct S and differential phases $ScS - S$ and $SS - S$ are included. These times in combination constrain structure throughout the mantle. Phase velocity maps (Laske & Masters, 1996) are included to improve imaging of the upper mantle. Finally, normal mode structure coefficients (Smith & Masters, 1989) constrain the structure of the whole mantle at large spatial scales. The data sets used in the inversions are specified in table 3.1, and figure 3.7 schematically describes the sensitivity of the data.

3.4.1 Synthetic Data

Computing broadband full waveform synthetics from the convection frame and a realistic distribution of stations and sources is beyond the reach of modern computing capabilities. Such a computation would be a valuable study, as it would allow the evaluation of approximations made in formulating the forward

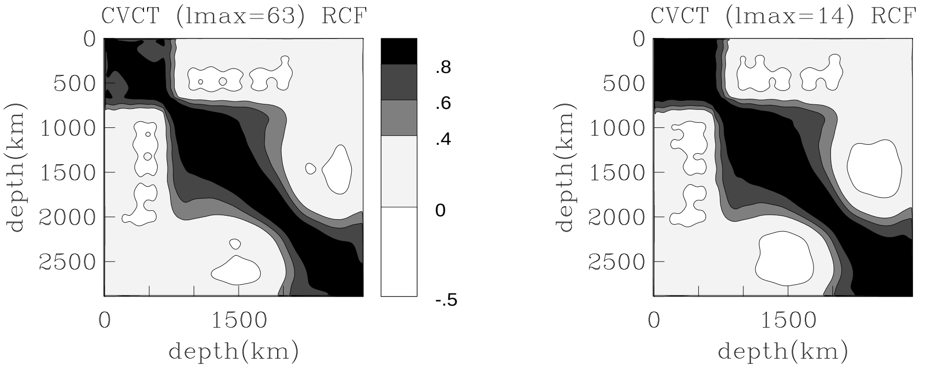


Figure 3.5: Radial correlation functions (RCF) of CVCT1 and its truncation to spherical harmonic degree 14. The three-layer structure of the model is evident as the RCF narrows near 670 km depth and as the avalanches pool at the CMB.

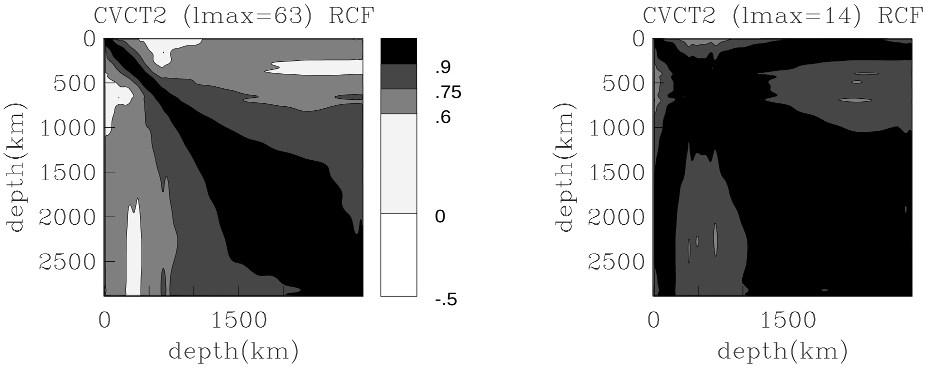


Figure 3.6: Radial correlation functions (RCF) of CVCT2 and its truncation to spherical harmonic degree 14.

problem and biases in measurement methods. In our case simplifications are in order. For the purpose of formulating the forward problems for body and surface waves, we assume two things. First, that the perturbation in shear velocity is small: $\delta\beta \ll \beta$. Second, we assume that the characteristic wavelengths of velocity heterogeneity are much larger than the wavelengths of the seismic waves we utilize. In this limit, the traveltimes are sensitive only to structure along a ray. As well, surface waves are only sensitive to integrals along great circle paths. The mode forward problem assumes only that $\delta\beta \ll \beta$.

Table 3.1: Data used in global tomography

Data set		comments
Travel	$SS - S$	18275 picks
Times	S	41098 picks
	$ScS - S$	8014 picks
Surface	Rayleigh	4 mHz - 15 mHz degrees 1 – 16
Waves	Love	4 mHz - 15 mHz degrees 1 – 16
Modes		even order sensitivity to various degrees
	Fundamental	0 S 3 – 0 S 10, 0 S 12 – 0 S 17
		0 S 21 – 0 S 31, 0 S 34 – 0 S 52
	Overtones	1 S 3 – 1 S 9, 2 S 4 – 2 S 6, 2 S 8
		2 S 10, 2 S 12, 4 S 3, 5 S 4 – 5 S 8

The extent and exact nature of the failure of these approximations in the real earth or even in the convection frame is not well understood. For some discussion of these issues see Wang & Dahlen (1995), Li & Tanimoto (1995), and Stark & Nikolayev (1993). The objective of Chapter 4 of this dissertation is to furnish additional tools by which the quality of our tomographic forward problems can be assessed.

Synthetic data are calculated from the convection frame using the linearized forward problems of typical use in global tomography. The synthetic traveltimes are computed using the method of Dziewonski (1984). The forward problems for modes and surface waves are computed as described in Woodhouse & Dahlen (1978). We directly calculate the perturbation of surface wave equivalent mode frequency from the velocity heterogeneities of CVCT1 and CVCT2. These frequency perturbations are then converted to phase velocity perturbations. The resulting spherical harmonic coefficients are perturbed by Gaussian errors with the magnitude of the estimated errors of the phase velocity coefficients estimated

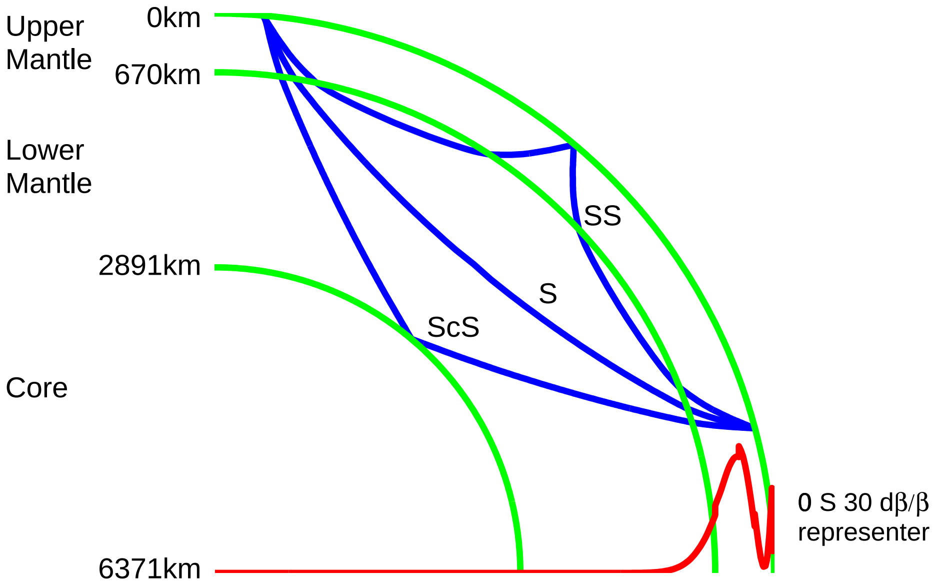


Figure 3.7: Raypaths of typical S, ScS, and SS phases in the mantle of the Earth. Differential phases $ScS-S$ and $SS-S$ allow for the reduction of the effect of source and receiver unknowns on the traveltimes. The sensitivity of a typical spheroidal mode is also shown. The surface waves have shallower sensitivity than the mode shown.

from actual data.

Generation of synthetic noise for the traveltime data is accomplished using error estimates obtained in Bolton (1996), section 2.4.3. We characterize the variance σ^2 in our traveltime measurements as due to three sources: earthquake source mislocation error, 3D structure, and the combined contributions of picking error, instrument response errors, etc. The standard deviations of the associated Gaussian random variables will be referred to as σ_X, σ_{3D}, and σ_N, so that $\sigma^2 = \sigma_X^2 + \sigma_{3D}^2 + \sigma_N^2$. The σ_N are obtained as described in Bolton (1996) for each of the traveltime data sets. The σ_X are assumed to be the result of a simple Gaussian model of the three components of spatial source mislocation, adjusted so that 50% of the resulting χ^2 distributed (with 3 degrees of freedom) mislocations lie within 15 km (Kennett & Engdahl, 1991; Smith & Ekström, 1996) of the location of the source in the PDE catalog. The origin time is perturbed using the empirical relation $\delta t(seconds) = \delta z(km)/9$ (Bolton, 1996, p.35).

The construction of synthetic noise in the traveltimes is summarized in table 3.2. The values in parentheses refer to estimated standard deviations in the actual data sets, assuming that these data sets have σ_N and σ_X as characterized by the synthetic errors. For CVCT1, our synthetic data and errors have a signal-to-noise ratio (SNR) comparable to the actual data, although the SNR for the synthetic S and $SS - S$ is somewhat larger than for the actual data. For CVCT2, the SNR is very similar to the actual data except for the $ScS - S$ data, where the SNR is very low.

Table 3.2: Traveltime synthetic noise

			σ_{3D}			σ		
Data set	σ_N	σ_X	CVCT1	CVCT2	Earth	CVCT1	CVCT2	Earth
$SS - S$	2.26s	0.20s	5.08s	3.49s	3.92s	5.63s	4.10s	4.53s
S	1.64s	1.86s	3.86s	4.03s	3.30s	4.63s	4.66s	4.13s
$ScS - S$	1.18s	0.20s	2.60s	0.76s	2.47s	2.75s	1.48s	2.75s

We note that although σ_X is quite large for the S traveltimes, the projection algorithm (see section 2.4.2) takes linear combinations of data to reduce the effect of this error.

3.4.2 Actual vs. Synthetic Data

Before performing some tomographic experiments to determine what we can image using all of the seismic data sets, we compare actual data and synthetic data generated from CVCT1 and CVCT2. There are a number of immediately obvious differences and some interesting similarities.

Traveltimes

Many of the data sets used have fairly simple relationships to velocity structure in particular regions of the mantle. The $SS - S$ and $ScS - S$ differential times have no sensitivity in regions where their raypaths coincide (see figure 3.7). The S leg of these differential phases averages over large regions along the the raypath in the mid-mantle. The smaller scale variations in a map of these differential phases plotted at their bounce points is therefore due to variations in structure in the vicinity of the bounce point. Also, if the mid-mantle structure is of low amplitude, variations are dominated by the bounce-point structure at all spatial scales. Figures 3.8 - 3.11 display synthetic and actual $SS - S$ and $ScS - S$ differential times, plotted at their bounce points and also smoothed after being fit by spherical splines (Parker, 1994, Chapter 2) and expanded in spherical harmonics.

The CVCT1 synthetic $ScS - S$ clearly show a smoothed lower mantle, based on a comparison of figure 3.10 and the deeper levels of figure 3.1. Figure 3.12 reveals that the spectra of observed and CVCT1 synthetic $ScS - S$ both exhibit a prominent peak at degree 2. $SS - S$ exhibits a large difference in total amplitude between the actual and CVCT1 synthetic data, due to the radial similarity of structure in the upper mantle of CVCT1. The shapes of the spectra are different, although both CVCT1 and the Earth exhibit broad low degree spectra. Presumably, the difference in degree 1 amplitude is due to the lack of continents in CVCT1.

The CVCT2 synthetic $ScS - S$ are strongly corrupted by noise (figure 3.10), and the resulting spectra have little to do with the structure of CVCT2 near the CMB, other than the fact that they are of low amplitude (see figure 3.12). CVCT2 $SS - S$ spectra (figure 3.12) differ enormously from the actual data. The CVCT2 spectra are very low in amplitude and are also dominated by degrees 3,4 and 5, in contrast with both the Earth and CVCT1.

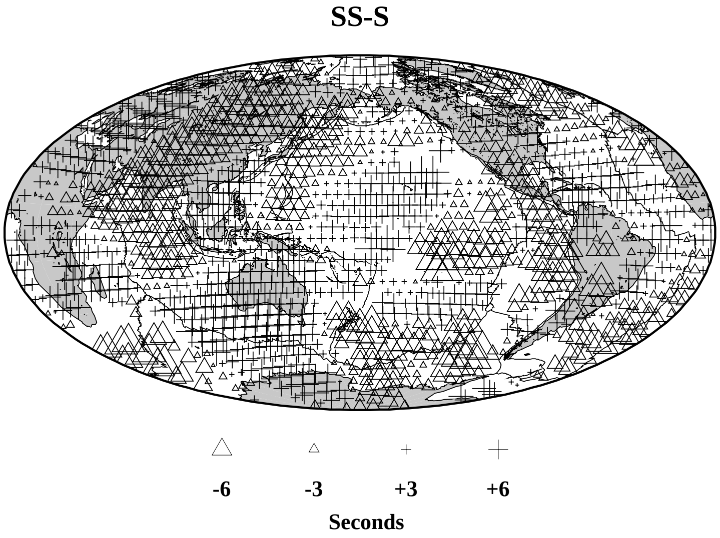


Figure 3.8: Synthetic $SS - S$ residual at the surface bounce point averaged over 5° radius spherical caps centered on the symbols. Caps are shown with 2 or more hits, and caps with values of zero will still plot as a point. Coverage is more or less global, and it is clear that these data more or less sample the structure at the bounce point.

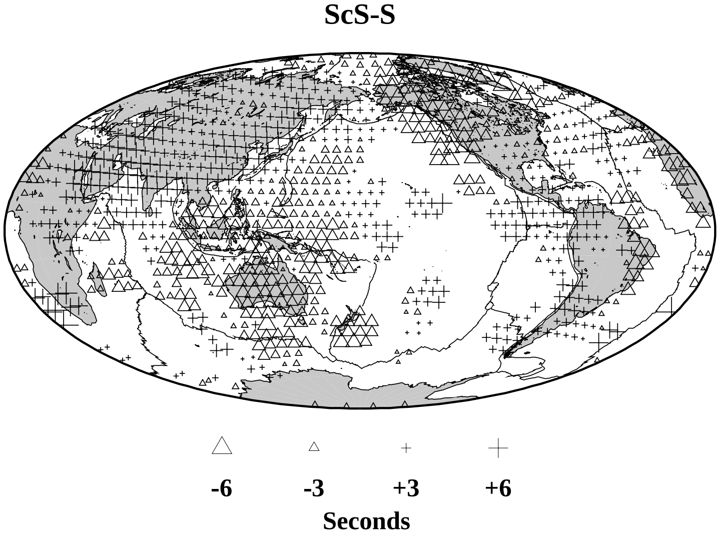


Figure 3.9: Synthetic $ScS - S$ residual at the surface bounce point averaged over 5° radius spherical caps centered on the symbols. Caps are shown with 2 or more hits, and caps with values of zero will still plot as a point. Coverage is very patchy, although these data clearly sample structure at the CMB bounce point. Continents for scale only.

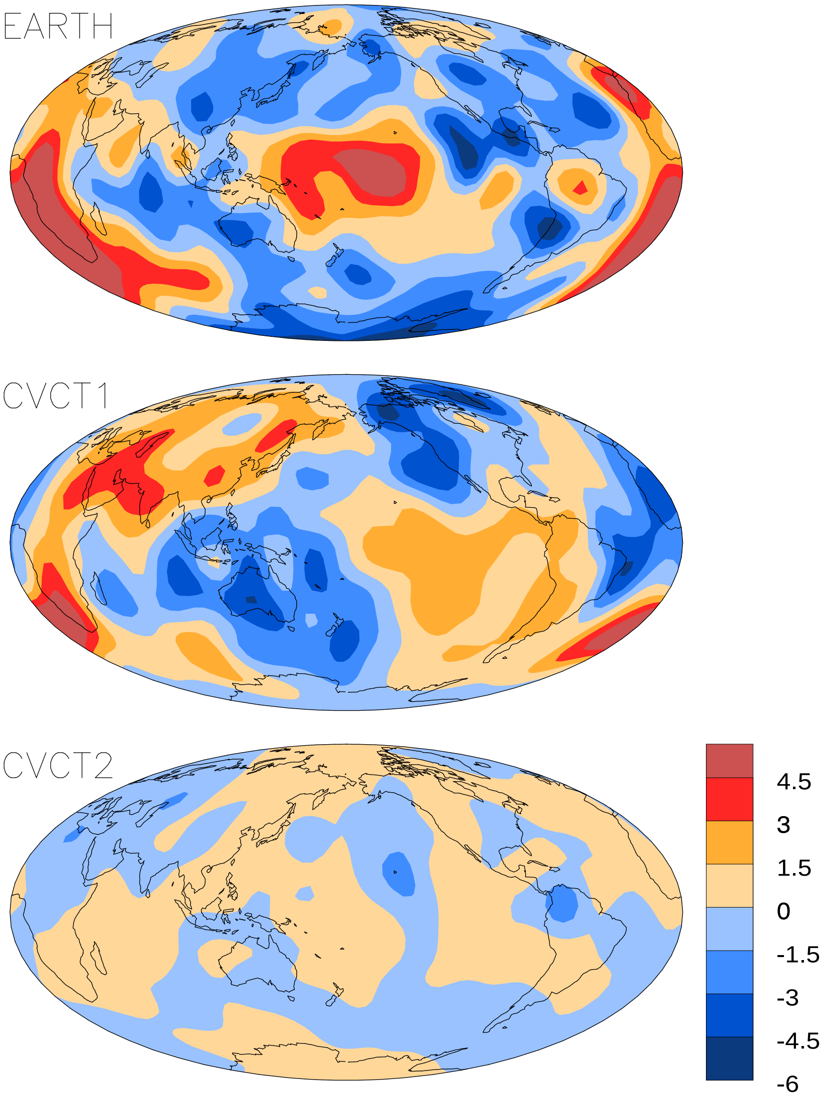


Figure 3.10: Synthetic and observed $ScS - S$ residuals at their surface bounce points fit by spherical splines and expanded in spherical harmonics to degree 16. The synthetic fits should be compared to the structure of CVCT1 (figure 3.1) and CVCT2 (figure 3.1) near the CMB. Note that the apparent CMB structure of CVCT2 is completely obscured by the addition of synthetic noise. Continents for scale only.

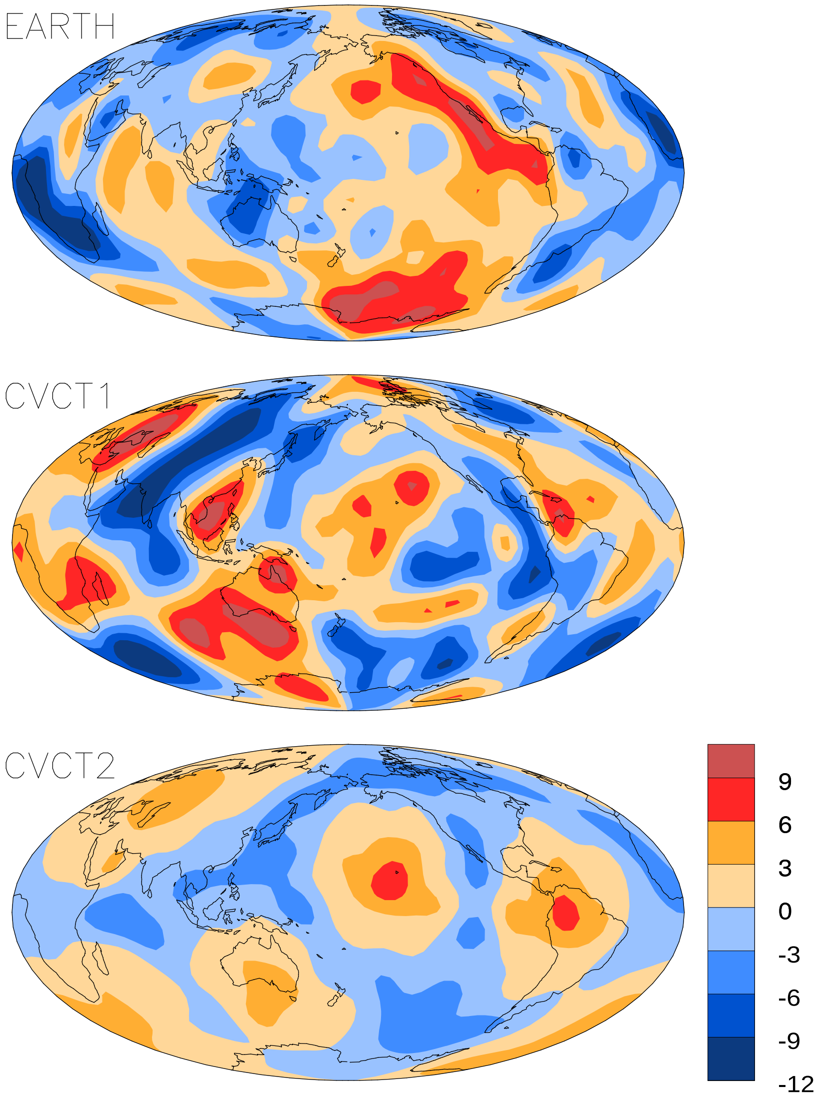
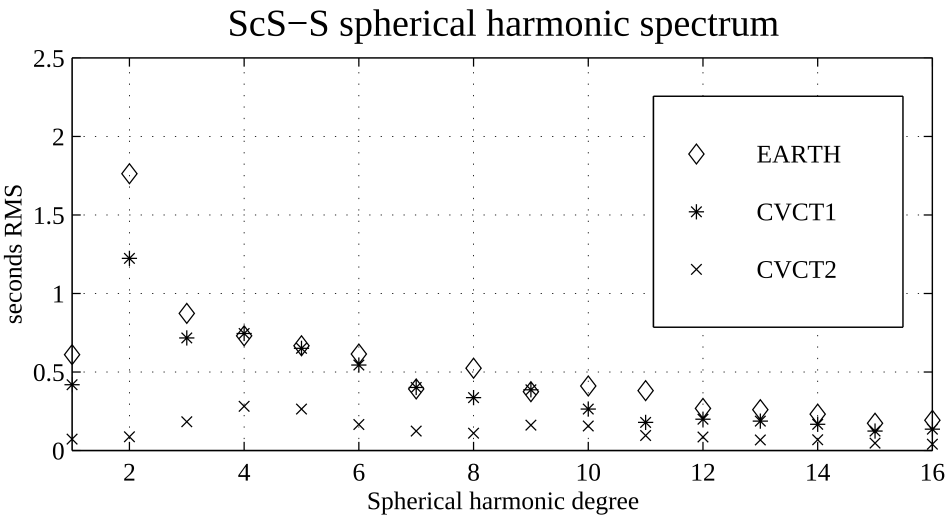


Figure 3.11: Synthetic and observed $SS - S$ residuals at their surface bounce points fit by spherical splines and expanded in spherical harmonics to degree 16. The fits of synthetic data should be compared to the structure of CVCT1 (figure 3.1) and CVCT2 (figure 3.2) in the upper mantle.



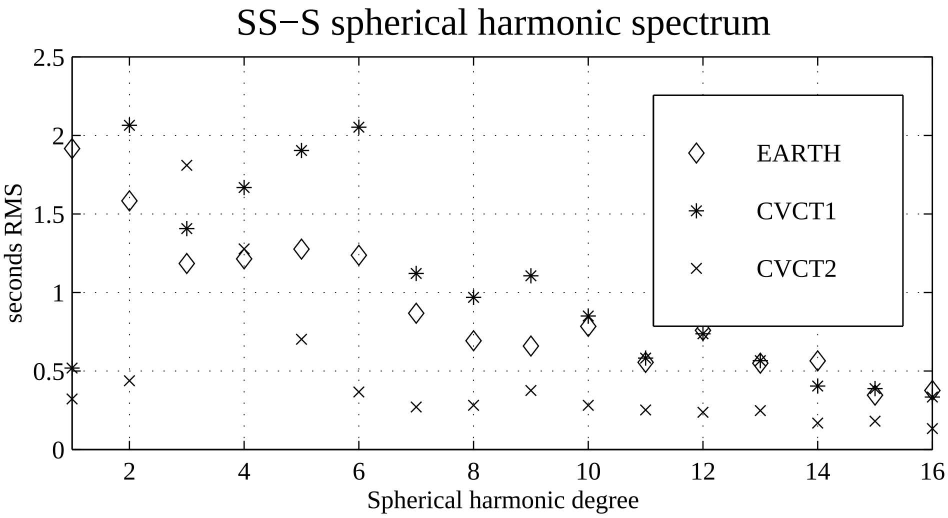


Figure 3.12: Spectra of synthetic and observed $SS - S$ and $ScS - S$ residuals at their surface bounce points fit by spherical splines and expanded in spherical harmonics to degree 16.

Surface Waves

Surface waves are sensitive to the structure of the earth extending only a few hundred kilometers in depth. Lower frequencies see deeper, and in general the Rayleigh wave sensitivity changes the most with frequency, as shown in figure 3.13. The structures of the upper mantles of CVCT1 and CVCT2 therefore have a clear effect on the surface wave spectra (see figures 3.15 and 3.4.2); both the Rayleigh and Love waves see the same spherical harmonic spectrum, and this spectrum varies little with frequency.

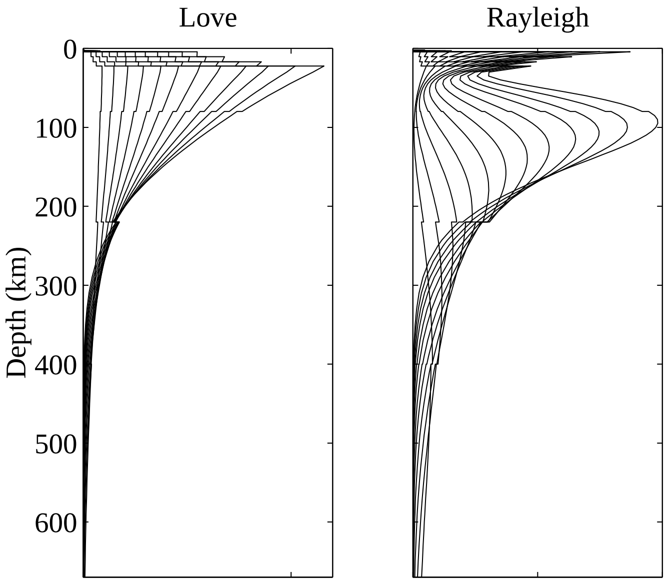


Figure 3.13: Sensitivity of Love and Rayleigh waves with depth and frequency. Both plots show the sensitivity of surface wave equivalent modes nearest integer frequencies in the range 4 – 15 mHz. The shallowest sensitivities are for 15 mHz; the deepest for 4mHz. The depth range covers the upper mantle.

The actual data, however, indicate a fairly different earth. The shallow structure seen by the Love waves has much more amplitude in degrees 1, 4 and 5 than the synthetic data. The structure seen by the Rayleigh waves varies with depth (i.e., changes with frequency) and again exhibits more amplitude in degrees 1,4 and 5.

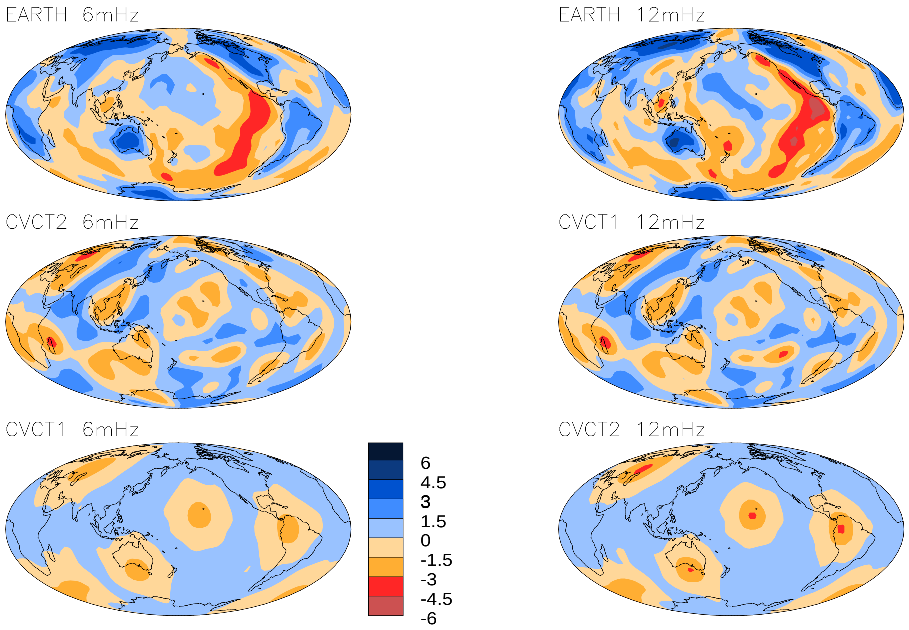


Figure 3.14: Relative phase velocity perturbation of Love waves calculated from CVCT1 and CVCT2 and estimated from actual data. Since the patterns vary continuously with frequency from 4 – 15 mHz, only two frequencies are shown. The maximum spherical harmonic degree is 16. These waves see very shallow structure and their sensitivity kernels or representers are very similar across frequencies.

It is clear from the surface wave data that there are significant differences between the Earth, CVCT1 and CVCT2. The Earth does not exhibit such a monolithic upper mantle as either of the convection simulations. The traveltime data indicate that the deep mantle of the Earth possesses a “boundary layer” of long wavelength, higher amplitude structure. While CVCT1 exhibits such structure, it is completely missing in CVCT2.

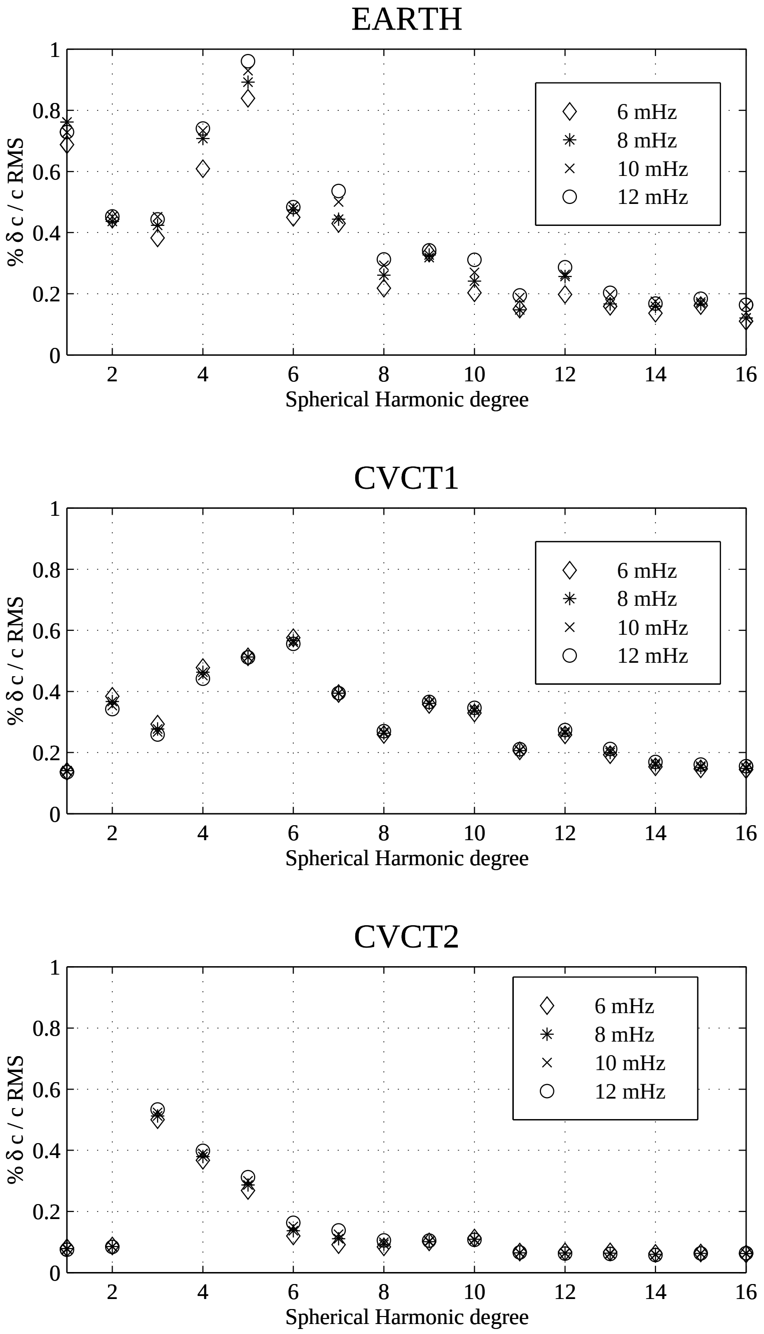


Figure 3.15: Spherical harmonic amplitude spectra of the relative phase velocity perturbation of Love waves calculated from CVCT1 and CVCT2 and estimated from actual data at several frequencies which are used in the inverse problem.

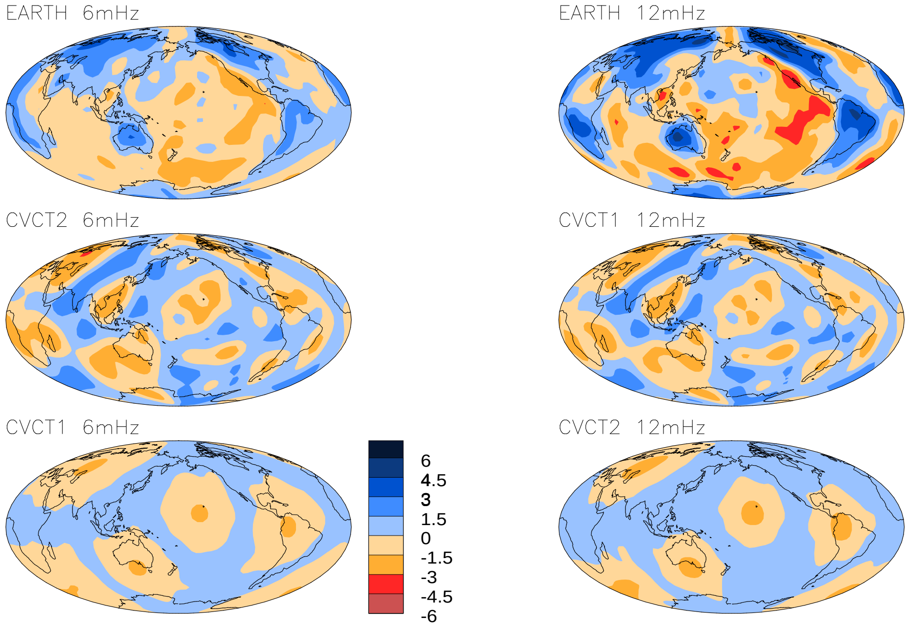


Figure 3.16: Relative phase velocity perturbation of Rayleigh waves calculated from CVCT1 and CVCT2 and estimated from actual data. The maximum spherical harmonic degree is 16. Note that the synthetic maps at 6mHz and 12mHz are much more similar to each other than the maps made from actual data; the apparent structure of the earth changes considerably across the upper mantle, while CVCT1 and CVCT2 are largely the same throughout the upper mantle.

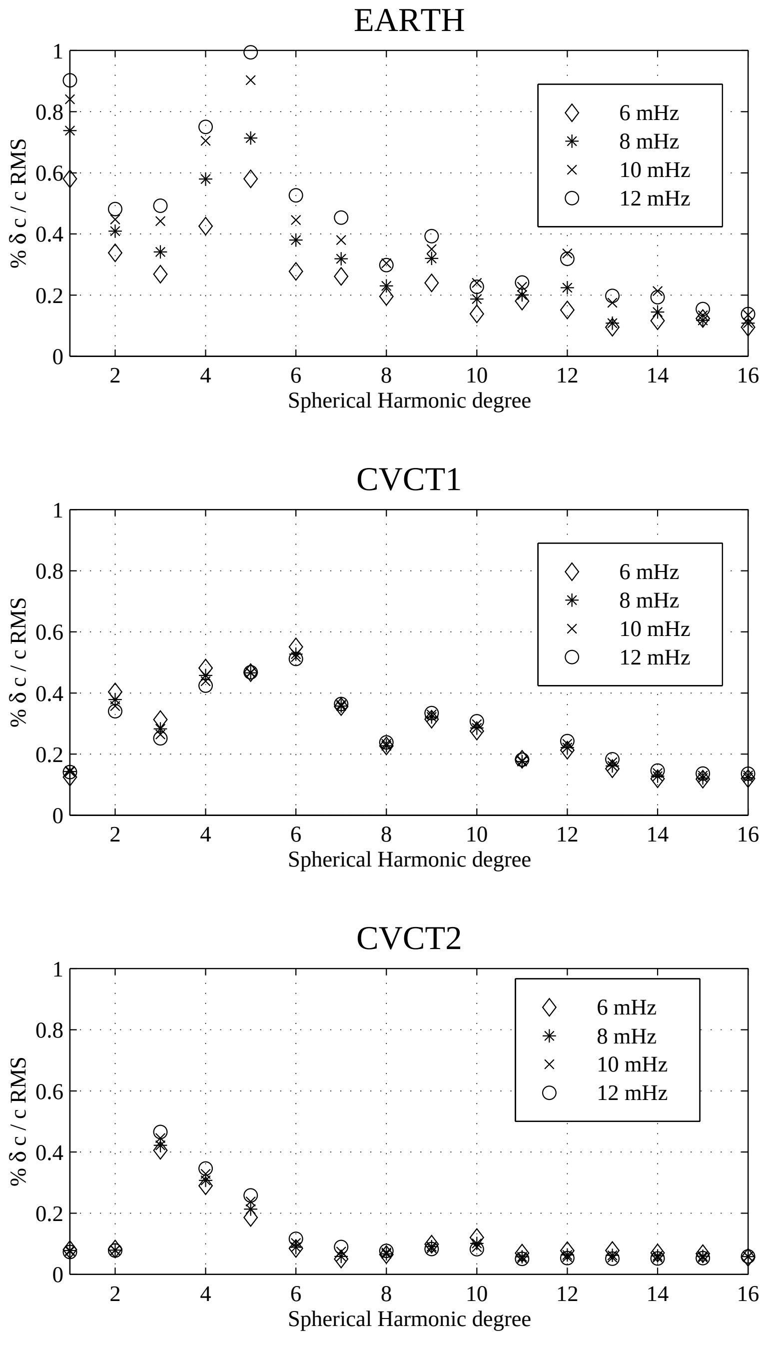


Figure 3.17: Spherical harmonic amplitude spectra of the relative phase velocity perturbation of Rayleigh waves calculated from CVCT1 and CVCT2 and estimated from actual data.

3.5 Inversion of Synthetic Data

As discussed in chapter 2, the inversion of seismic data is undertaken by a process of constrained least squares. The misfit of the data is set at a desired level, and the "smoothest" model fitting the data at that level is found. The construction of the model is a well defined process, and we use the QR technique described in the appendix of chapter 2.

Following the nomenclature of chapter 2, the smoothing operator is constructed from separate lateral $\mathbf{D}_L$ and radial $\mathbf{D}_R$ parts so that

$$\partial^T \partial = \mathbf{D} = \mathbf{D}_L + \alpha \mathbf{D}_R, \tag{3.5.0.3}$$

and the roughness of model m is $m^T \partial^T \partial m$. α is an adjustable parameter whose effect we wish to investigate in our inversions. To this end, we generate a large number of models at a given set of misfits (χ^2/N), and vary α from small to large values. We will then compare inversion results across a range of both χ^2/N and α. Clearly, models obtained with larger values of α will be radially smoother.

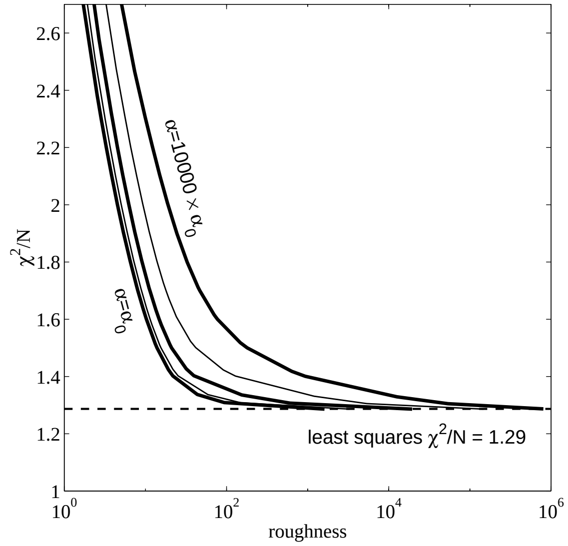


Figure 3.18: Tradeoff of model roughness $m^T \partial^T \partial m$ and data misfit $\chi^2/N = |d - Gm|^2/N$ for the inversion of CVCT1 synthetic data.

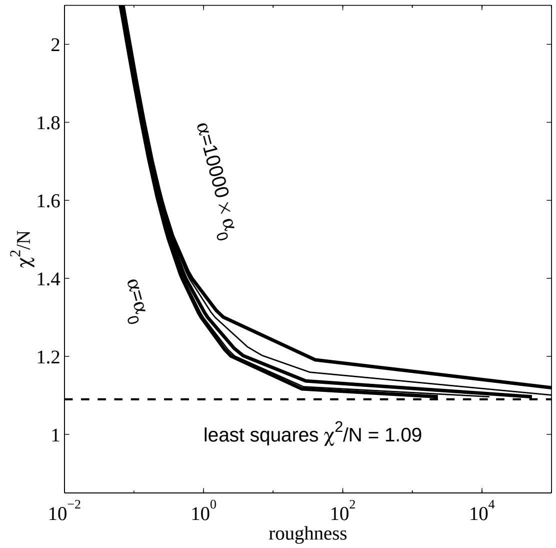


Figure 3.19: Same as figure 3.18, but for synthetic data obtained from CVCT2.

The trade-off of roughness and misfit is depicted in figures 3.18 and 3.19. Each of the curves is asymptotic to the least-squares χ^2/N, which is at some level due to the parameterization of the model. For CVCT1, the models we have obtained extend from $\chi^2/N = 1.3$ to $\chi^2/N = 2.3$ in steps of 0.1. At $\chi^2/N = 1.3$, the resulting model is essentially the least squares model, and at $\chi^2/N = 2.3$, the models are very smooth. We also vary α from α_0 to $10000 \times \alpha_0$ in steps of 100, and the results for each α are presented in figures 3.20, 3.21 and 3.22. For CVCT2, the models we have obtained extend from $\chi^2/N = 1.1$ to $\chi^2/N = 2.1$ in steps of 0.1. Again, at $\chi^2/N = 1.1$, the resulting model is essentially the least squares model. The results for the same α as for the CVCT1 models are presented in figures 3.23, 3.24 and 3.25.

The RCF's of the resulting inversions of CVCT1 synthetic data reveal that all models (even the least squares results) exhibit a region of radially correlated structure in the upper mantle. The location and sharpness of the change in structure at 670km depth depends upon α and the misfit level, but the upper mantle is imaged more or less as a layer. The same is true of the lowermost mantle.

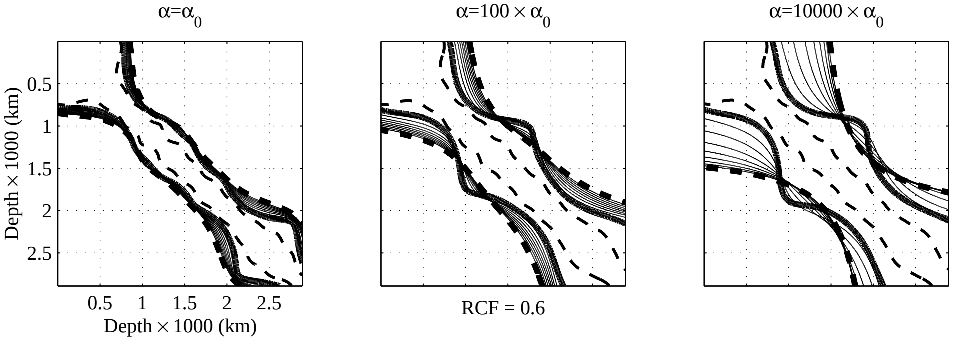


Figure 3.20: The 0.6 contours of the RCF's of models obtained by inversion of synthetic data. The lighter dashed line near the diagonal corresponds to $\chi^2/N = 1.3$, or essentially the least squares fit. The heavy solid line corresponds to $\chi^2/N = 1.4$, and the heavy dashed line to $\chi^2/N = 2.3$. The lighter lines are intermediate values in the range $\chi^2/N = 1.4 - 2.3$. The 0.6 contour of the RCF is chosen as it is representative of the general shape of the entire RCF.

The mid-mantle behaves differently, so that we must carefully choose intermediate values of misfit and α to bring out the mid-mantle layer of downwellings (see figure 3.20).

As long as we avoid models too close to the least squares model, the RMSD's of the models resulting from inversion of the CVCT1 data are very consistent as a function of α. By comparison to figure 3.3, the RMSD is well estimated, although the peak in amplitude of CVCT1 near 670km depth is smoothed out.

The correlation of the tomographic models with CVCT1 (figure 3.22) reveals that the mid-mantle suffers from the worst imaging. As in the case of the reproduction of the RCF's, the best correlations are obtained for intermediate values of α and at misfits "slightly above" the least squares misfit.

Although the RCF's of models resulting from the inversion of CVCT2 synthetic data are not particularly interesting in general, figure 3.23 shows that the

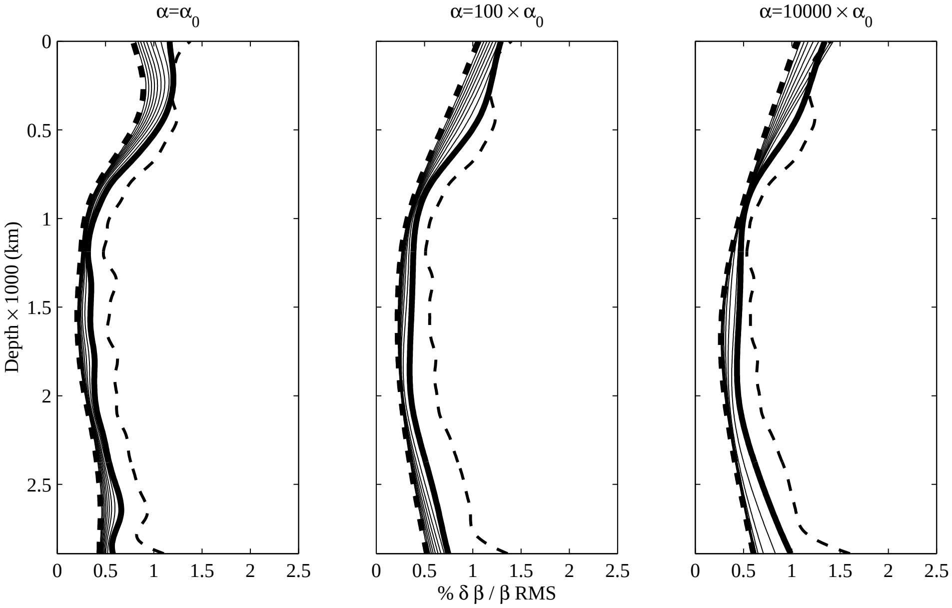


Figure 3.21: The RMSD's of models obtained by inversion of synthetic data. The levels of χ^2/N for each line are as described in figure 3.20. All of these results are quite similar, and indicate that the all tomographic images of CVCT1 feature high amplitudes throughout the upper mantle.

RCF, although everywhere at a very high value, is somewhat degraded in the lower mantle. The RMSD's of these models are very robust away from the least squares solution, and compare well with figure 3.4. The correlation of the models with the original convection frame (figure 3.25) show very good results at values χ^2/N near the least squares value, but quickly degrading results in the lower mantle for smoother solutions. Again, we know from the SNR of the $ScS-S$ data (table 3.2) that the lower mantle will be poorly constrained in the tomographic results.

The general quality of tomographic imaging attainable is quite good, as shown in figures 3.26 and 3.27.

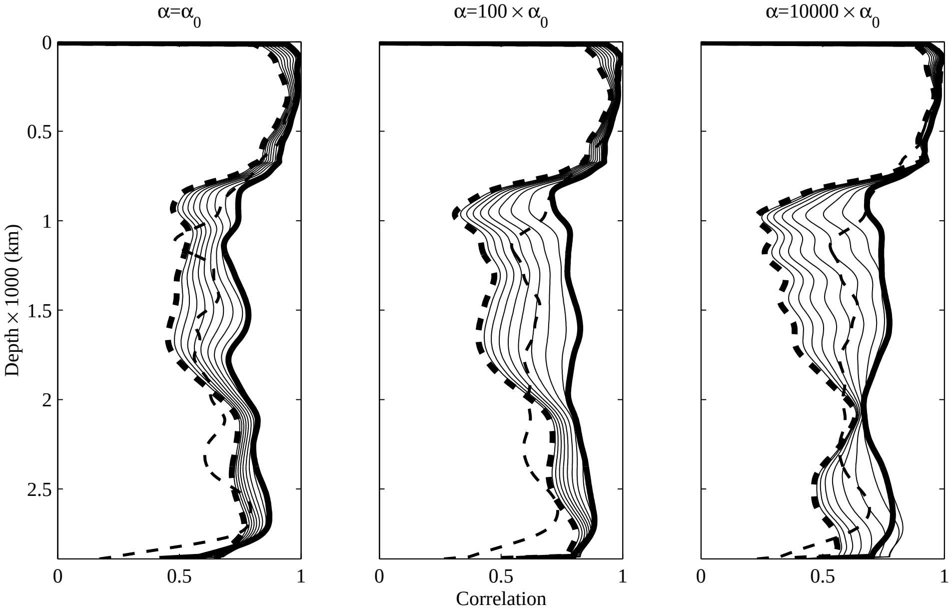


Figure 3.22: The correlation (see equation 3.3.0.2) with depth of models obtained by inversion of synthetic data with CVCT1 truncated to degree 14. Again, the lines are as described in figures 3.21 and 3.20. The correlation remains significant at the 95% level if above .36; some of the radially smooth models (for large α) fall below this level just below 670km depth. Note that the smoother models are "better" images than the least-squares images.

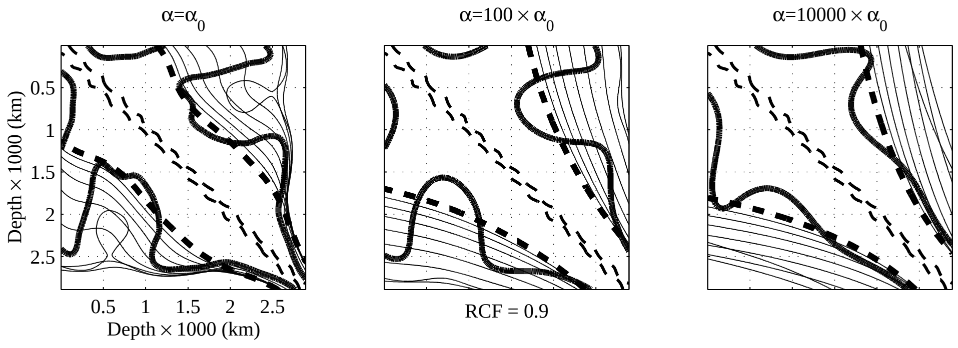


Figure 3.23: The 0.9 contours of the RCF's of models obtained by inversion of CVCT2 synthetic data. The lighter dashed line near the diagonal corresponds to $\chi^2/N = 1.1$, or essentially the least squares fit. The heavy solid line corresponds to $\chi^2/N = 1.2$, and the heavy dashed line to $\chi^2/N = 2.1$. The lighter lines are intermediate values in the range $\chi^2/N = 1.3 - 2.0$. The 0.9 contour of the RCF is chosen as it is representative of the general shape of the entire RCF.

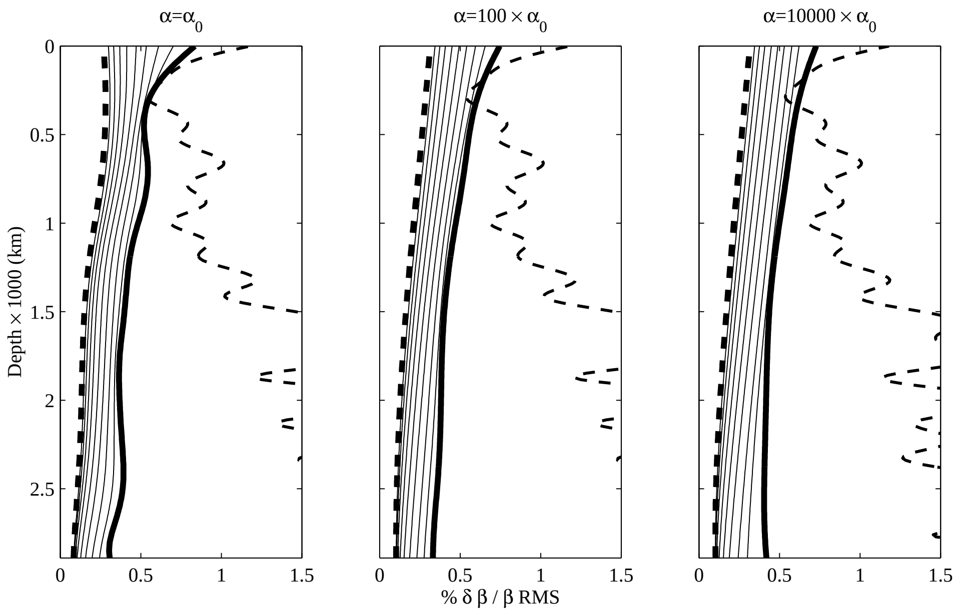


Figure 3.24: The RMSD's of models obtained by inversion of CVCT2 synthetic data. The levels of χ^2/N for each line are as described in figure 3.23.

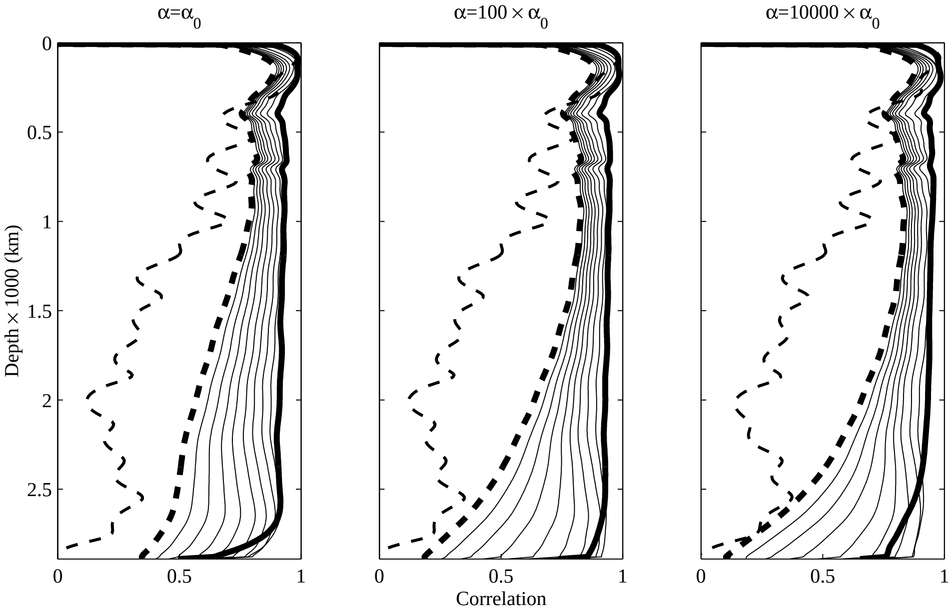


Figure 3.25: The correlation (see equation 3.3.0.2) with depth of models obtained by inversion of synthetic data with CVCT2 truncated to degree 14.

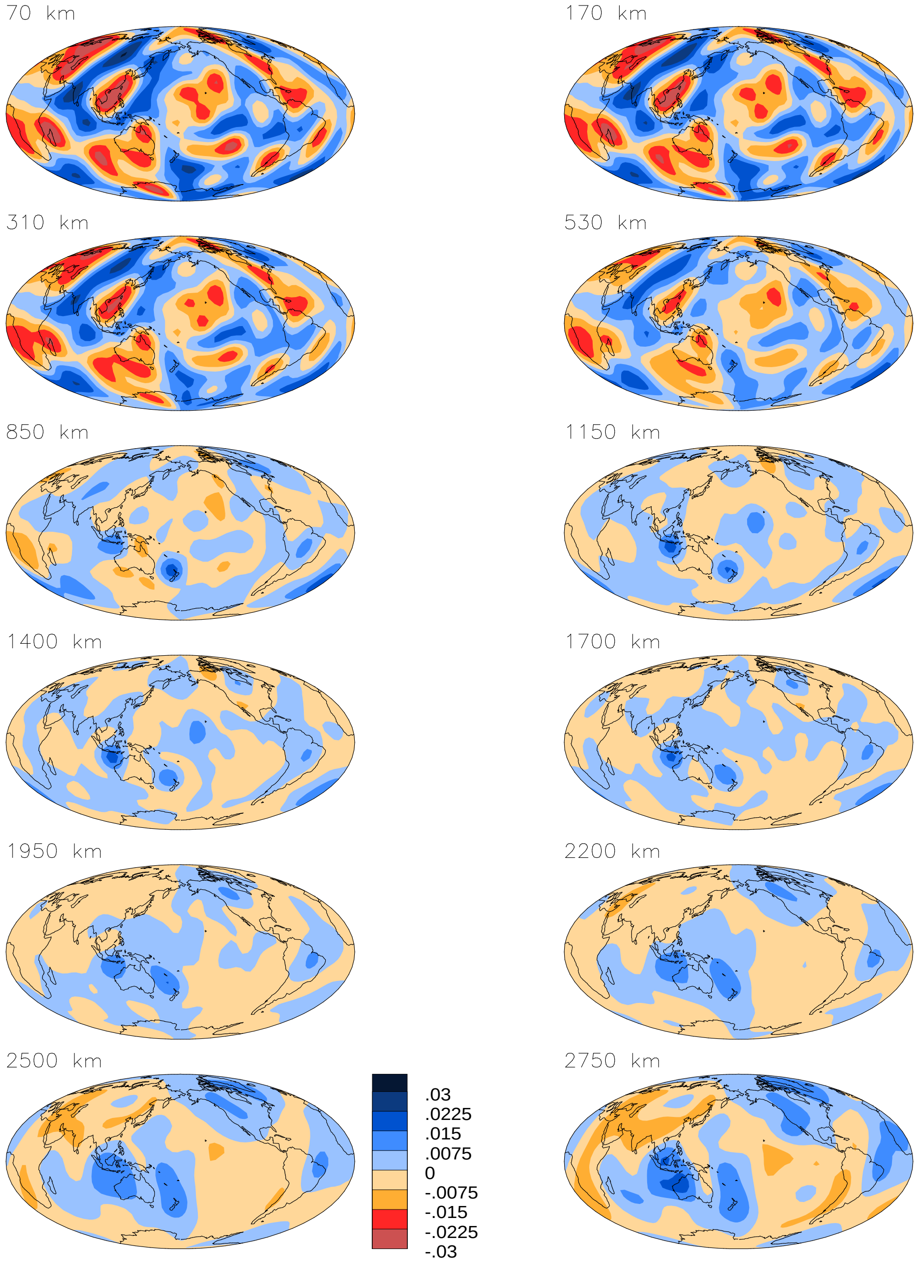


Figure 3.26: Twelve depth slices of an inversion of synthetic data generated from CVCT1 (see figure 3.1). Contoured values are shear velocity perturbation relative to the global model average at a given depth. Continents are for scale only. This model, TCVCT1, is obtained by setting $\chi^2/N = 1.5$, and $\alpha = 100 \times \alpha_0$.

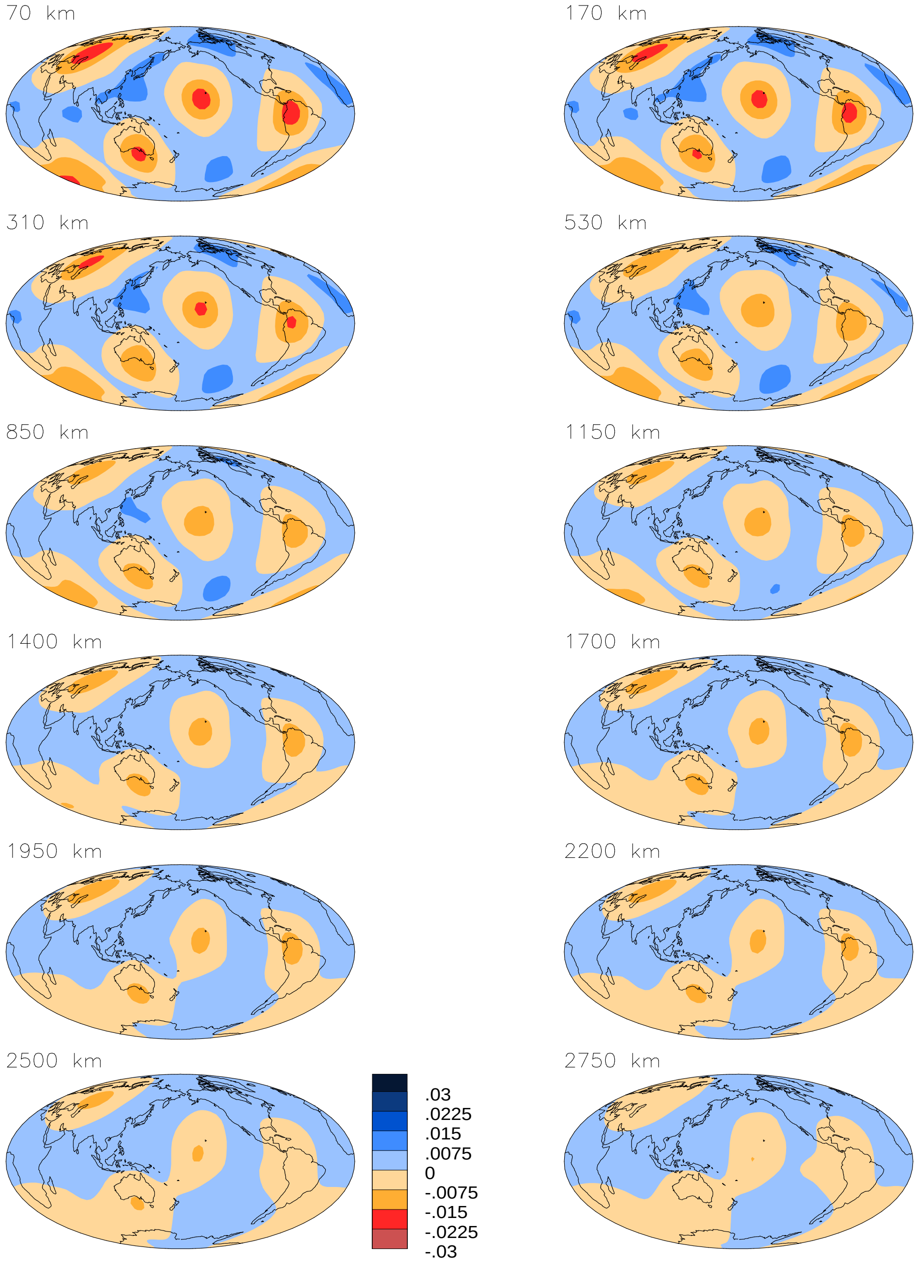


Figure 3.27: Twelve depth slices of an inversion of synthetic data generated from CVCT2 (see figure 3.2). Contoured values are shear velocity perturbation relative to the global model average at a given depth. Continents are for scale only. This model, TCVCT2, is obtained by setting $\chi^2/N = 1.3$, and $\alpha = 100 \times \alpha_0$.

3.6 Inversion of Actual Data

We proceed similarly for the actual data, and obtain another suite of models. The tomographic images of the earth are also generated for values of α ranging from α_0 to $10000 \times \alpha_0$. As the trade-off curves (figure 3.28) indicate, the least squares misfit is somewhat different, and we obtain models for $\chi^2/N = 1.4$ to $\chi^2/N = 2.4$. Again, the results for every other α are presented in figures 3.29 and 3.30.

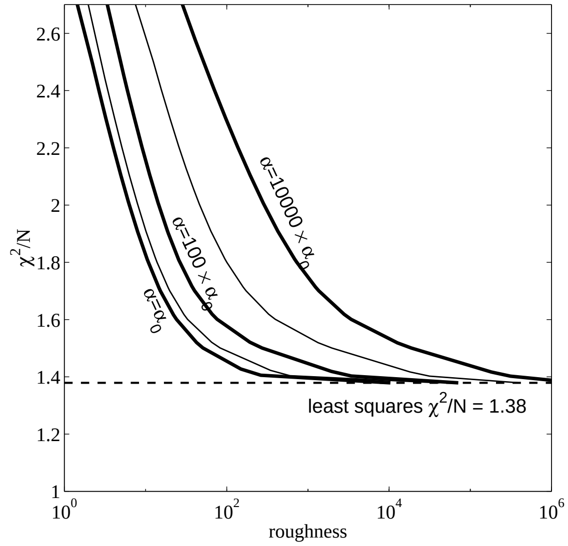


Figure 3.28: Tradeoff of roughness $m^T \partial^T \partial m$ and misfit $|d - Gm|^2/N$ for models in the inversion of actual data.

The RCF's (see figure 3.29) of all models resulting from inversions of actual data exhibit a narrow diagonal maximum and a strong decorrelation in the upper mantle, which is at odds with the RCF's of tomographic images of the convection results. All of the RMSD's, provided we avoid models too near the least squares solution, reveal high amplitude structure only in the upper 300km of the mantle (see figure 3.30). There is some hint of a slight amplitude change at depths of 500 – 1000km, but it is not robustly imaged.

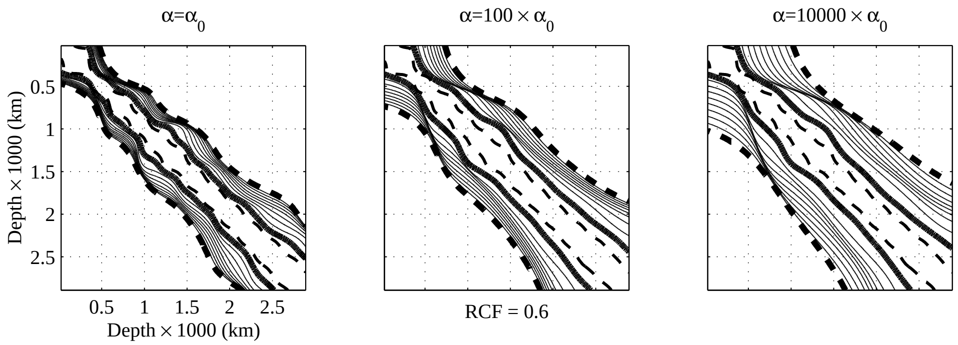


Figure 3.29: The 0.6 contours of the RCF's of models obtained by inversion of actual data. The lighter dashed line near the diagonal corresponds to $\chi^2/N = 1.4$, or essentially the least squares fit. The heavy solid line corresponds to $\chi^2/N = 1.5$, and the heavy dashed line to $\chi^2/N = 2.4$. The most prominent feature in all RCF's is the decorrelation in the shallow mantle. The lighter lines are intermediate values in the range $\chi^2/N = 1.5-2.4$. The most prominent feature in all RCF's is the decorrelation in the shallow mantle.

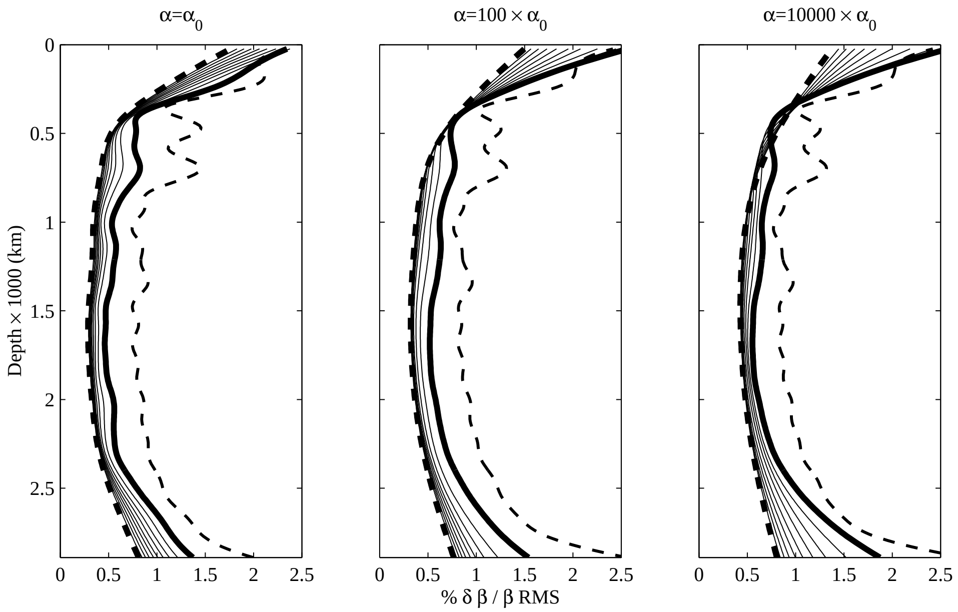


Figure 3.30: The RMSD's of models obtained by inversion of actual data. The levels of χ^2/N for each line are as described in figure 3.29.

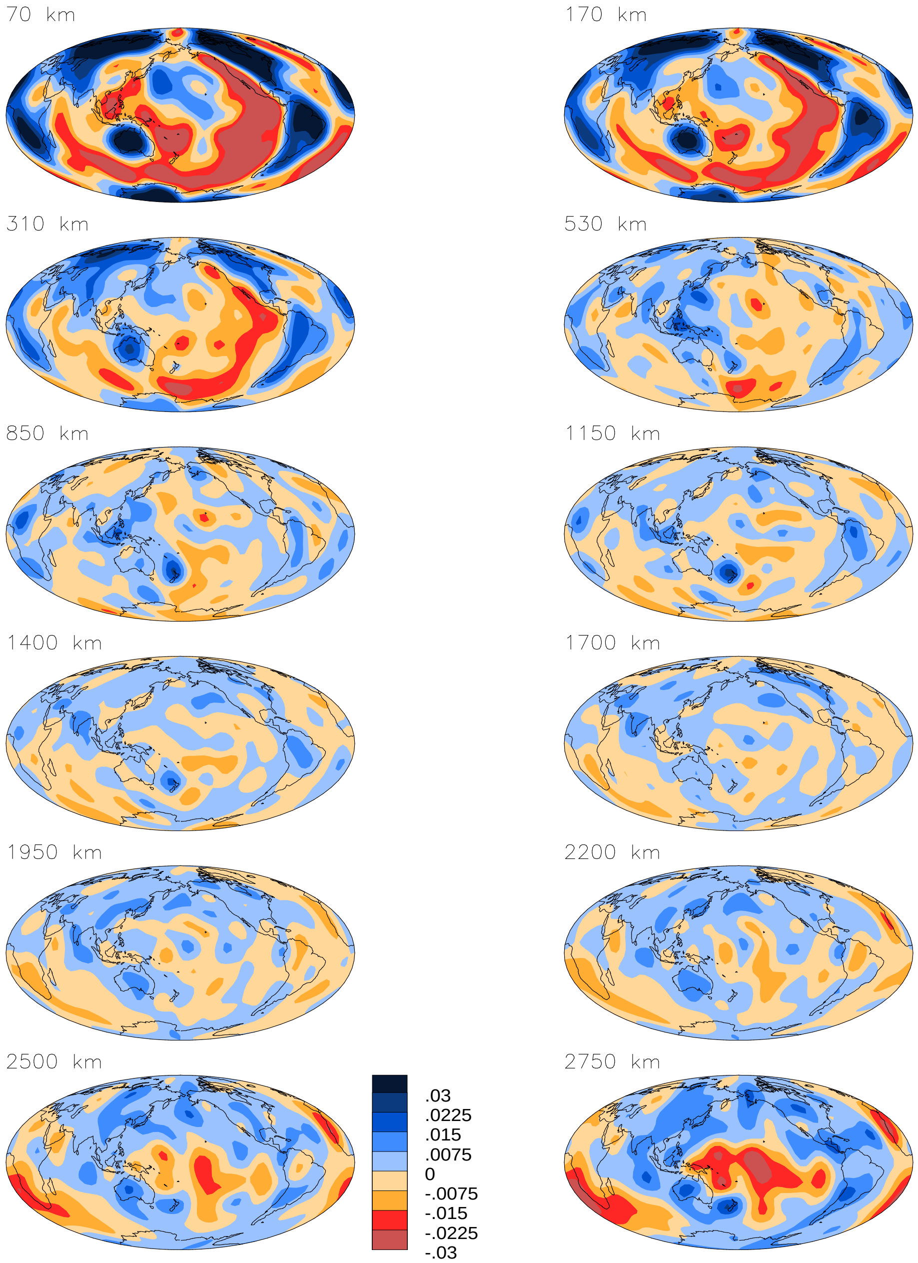


Figure 3.31: Twelve depth slices of an inversion of actual data. Contoured values are shear velocity perturbation relative to the global model average at a given depth. This model, TEARTH, is obtained by setting $\chi^2/N = 1.6$, and $\alpha = 100 \times \alpha_0$.

3.7 Conclusions

In order to summarize the results, we plot the RCF and RMSD diagnostics of several tomographic images. These reflect the comparison of the RMSD in figures 3.21, 3.24 and 3.30, and the RCF's in figures 3.20 and 3.29. The RMSD of CVCT1, a tomographic image of CVCT1 (TCVCT1) and a tomographic image constructed from actual data (TEARTH) are plotted in figure 3.32. This comparison strongly suggests that the Earth does not possess the upper mantle "layer" obtained in CVCT1. This agrees with the conclusion made by inspection of synthetic traveltimes and surface wave data above. The RCF's of TCVCT1 and TEARTH support the same conclusion (see figure 3.34).

The apparent high-amplitude structure near the surface and CMB of the actual Earth is possible evidence for compositional variations in those regions. Very high heat flow from the core would be required to generate the the deep structure of TEARTH by thermal effects alone. Additionally, the effects of continents with deep roots would push the flow pattern to low spherical harmonic degree and provide large seismic velocity variations in the shallow mantle (personal communication, P. Tackley).

While the endothermic phase change at 670 km depth in CVCT1 appears to be a way to generate large scale structure from high Rayleigh number convection, the resulting mantle does not agree in detail with what we see in the actual earth. The principal difference is in the strong change in structure at 670km depth; the Earth does not exhibit such a change. This suggests that other factors are very important in controlling mantle convection, such as the presence of tectonic plates (Mégnin, et al., 1997) and/or the presence of continents with deep roots..

The RMSD of CVCT2, the tomographic image TCVCT2 and TEARTH are shown in figure 3.33. The radial profile of TCVCT2 is much too smooth as compared to TEARTH. TCVCT2 is missing the increased amplitude in both the shallowest and deepest mantle. Again, this suggests that factors other than

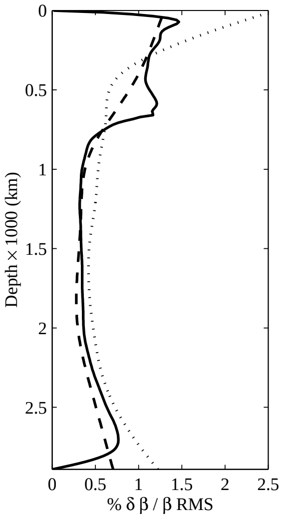


Figure 3.32: RMS amplitude of velocity heterogeneity in CVCT1 (solid - truncated to degree 14), TCVCT1 (dashed) and TEARTH (dotted) relative to the spherical average velocity at each depth. Even though the inversion process has smoothed out the sharp change in amplitude near 670km depth, TCVCT1 still exhibits high RMSD across the upper mantle. None of the models resulting from a similar inversion of actual data (e.g., TEARTH), exhibit this behavior.

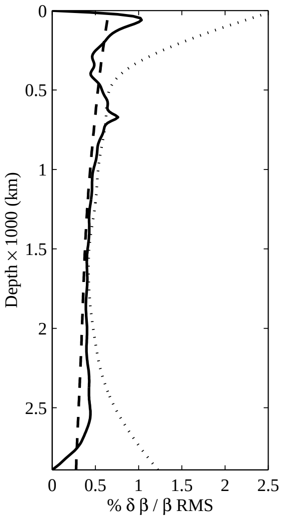


Figure 3.33: RMS amplitude of velocity heterogeneity in CVCT2 (solid - truncated to degree 14), TCVCT2 (dashed) and TEARTH (dotted) relative to the spherical average velocity at each depth.

those modelled in CVCT1 or CVCT2 are important in controlling the details of convection in the actual mantle.

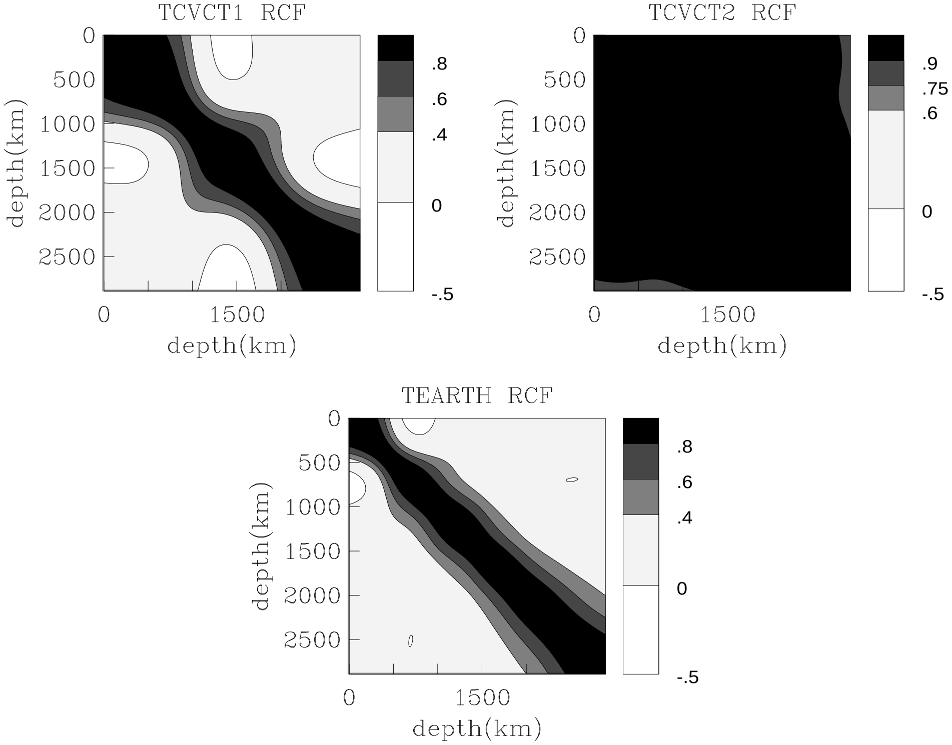


Figure 3.34: Radial correlation functions (RCF) of TCVCT1, TCVCT2 and TEARTH. The three-layer structure of CVCT1 is still evident in the TCVCT1 RCF. Inversions of actual data (e.g., TEARTH) show no such features, the only decorrelation being in the shallow mantle.

3.8 References

Anderson, D. 1989 *Theory of the Earth.* Blackwell Scientific Publications, Boston.

Bolton, H. 1996 *Long Period Travel Times and the Structure of the Mantle.* Doctoral dissertation, University of California, San Diego.

Dziewonski, A.M. 1984 Mapping the lower mantle: determination of lateral heterogeneity in P velocity up to degree and order 6. *J. Geophys. Res.* **89**, 5929–5952.

Edmonds, A. 1960 *Angular Momentum in Quantum Mechanics.* Princeton University Press, Princeton.

Jordan, T., Glatzmaier, G. & Tackley, P. 1995 Comparisons of seismic earth structures and mantle flow models using radial correlation functions. *Science* **261**, 1427 – 1431.

Kennett, B. & Engdahl, E. 1991 Traveltimes for global earthquake location and phase identification. *Geophy. J. Int.* **105**, 429 – 465.

Laske, G. & Masters, G. 1996 Constraints on global phase velocity maps by long-period polarization data. *J. Geophys. Res.* **101**, 16059 – 16075.

Li, X. & Tanimoto, T. 1993 Waveforms of long-period body waves in a slightly aspherical Earth model. *Geophys. J. Int.* **112**, 92–102.

Masters, G., S. Johnson, G. Laske, and H. Bolton 1996 A shear velocity model of the mantle. *Phil. Trans. R. Soc. Lond.* **354A**, 1385–1411.

Mégnin, C., Bunge, H.-P., Romanowicz, B. & Richards, M. 1997 Imaging 3-D spherical convection models: What can seismic tomography tell us about mantle dynamics?. *Geophysical Research Letters* **24**, 1299–1302.

Parker, R. 1994 *Geophysical Inverse Theory.* Princeton University Press, Princeton.

Pulliam, J. & Stark, P. 1994 Confidence regions for mantle heterogeneity. *Journal of Geophysical Research* **99, B4**, 6931–6943.

Smith, G. & Ekstrom, G. 1996 Improving teleseismic event locations using a three-dimensional Earth model. *Bull. Seis. Soc. Am.* **86**, 788 – 796.

Smith, M.F. & Masters, G. 1989 Aspherical structure constraints from free oscillation frequency and attenuation measurements. *J. Geophys. Res.* **94**, 1953–1976.

Stark, P. & Nikolayev, D. 1993 Toward Tubular Tomography. *Journal of Geophysical Research* **98, B5**, 8095–8106.

Tackley, P., Stevenson, D., Glatzmaier, G. & Schubert, G. 1993 Effects of an endothermic phase transition at 670 km depth on a spherical model of convection in the Earth's mantle. *Nature* **361**, 699–704.

Wang, Z. & Dahlen, F. 1995 Validity of surface-wave ray theory on a laterally heterogeneous earth. *Geophys. J. Int.* **123**, 757–773.

Woodhouse, J. & Dahlen, F. 1978 The effect of a general aspherical perturbation on the free oscillations of the Earth. *Geophys. J. R. Astr. Soc.* **53**, 335–354.

Woodward, R. & Masters, G. 1992 Global upper mantle structure from long-period differential travel times and free oscillation data. *J. Geophys. Res.* **109**, 275 – 293.

Chapter 4

H^0 and H^1 spectral FEM solutions of the frequency domain forced linearly attenuating elastic wave equation

4.1 Abstract

We analyze convergence of two variable order (spectral) finite element method (FEM) approximations to the variational solution of the forced linearly attenuating frequency domain elastic wave equation (LAFDWE) in 1-D, and extend results to higher dimensions. Convergence of the H^1 FEM is shown to be asymptotically optimal in both the $H^0(\Omega)$ and $H^1(\Omega)$ norms given the proper regularity of the forcing. Furthermore, the H^1 FEM solution converges identically to the best approximation in $H^1(\Omega)$. The conditions for asymptotic convergence to be attained are shown. The computational cost (in memory and FLOPS) of the H^1 FEM is explored. In order to extrapolate results to any element size h, the generalized eigenproblem is investigated for the H^1 FEM. Numerical results for the

FEM confirm the convergence predictions and reveal quantitatively the relative error as a function of mean grid intervals per wavelength. The same type of error analysis extends to higher dimensions.

Convergence of the discontinuous or H^0 FEM is asymptotically optimal as well, based on numerical study of a limiting error bound relative to the best approximation. In 1-D, given that the solution is smooth in the neighborhood of the boundary, the H^0 FEM converges in $H^0(\Omega)$ and the boundary space $H^0(\partial\Omega)$. In higher dimensions, the same type of error analysis also applies. The computational considerations for the H^0 FEM are essentially the same as for the $H^1(\Omega)$ FEM.

For both versions of the FEM, it is found that the higher order than linear ($p > 2$) FEM yields useful levels of accuracy at much less cost than $p = 2$ FEM; higher p is essential for accurate solutions. However, arbitrarily high order FEM does not yield less cost for a given level of solution accuracy. The best p for practical levels of solution accuracy is probably less than 10 (in 1-D).

In higher dimensions, it is found that sources of interest in seismology belong to the space $H^{-3}(\Omega)$, and approximation in the solution space $H^{-1}(\Omega)$ is problematic.

4.2 Introduction

We will investigate some variational methods for the solution of a "linearly attenuating" frequency domain wave equation (LAFDWE). Variational methods result in the formation and solution of equations like $A(u,v) = L(v)$, $\forall v \in S_2$, $u \in S_1$, where S_1 and S_2 are spaces of distributions with appropriate smoothness properties, and u is the solution. While in many cases the integro-differential operator A is self-adjoint ($A(u,v) = \overline{A(v,u)}$), it need not be to result in useful approximations to u. Indeed, this is the case for the LAFDWE. Variational methods for the solution of problems in complicated media (e.g., composed of discontinuous elastic moduli and complicated boundaries) are of practical use because of their ability to approximate some boundary conditions accurately and simply. This is accomplished by making the solution space S_1 as large as possible while still retaining the existence and uniqueness of the solution. The same enlargement of the solution space allows us to account for singular sources, although we will show that sources of typical use in seismology are so singular as still to involve additional study.

Some recent literature in seismology has attempted to address the nature of the variational solutions of the LAFDWE. That work (Geller & Ohminato, 1994; Cummins, et al., 1994a,b; Geller & Takeuchi, 1995) has proceeded even to the modification of the variational operators to decrease computational effort for a given accuracy (Geller & Takeuchi, 1995). However, no attention has been paid to the fundamental nature of the solution, much less to the nature of convergence to that solution. These authors also attempt to design some finite elements to better approximate the solution in the presence of a singular source, but an understanding of elliptic partial differential equations is a prerequisite for such an undertaking. Additionally, the attempt to parameterize the solution in a sphere by a combination of linear elements and spherical harmonics will always result in relatively dense matrix algebra in the presence of heterogeneities in elastic moduli, limiting the efficiency of computational resources. Clearly, a knowledge gap exists in the use of variational and finite element methods in the seismological literature. This chapter aspires to fill some of this knowledge gap, especially as this knowledge applies to practical considerations of computational effort.

The methods presented herein assume that we have decided that a full wavefield solution is required and that our problem can be cast as a LAFDWE. While this is the case to some level of approximation for materials in the Earth (Liu, et al., 1976), it is not so for most materials. Our reasons for exploring this method as opposed to simply implementing a finite difference solution (Viriuex, 1986) are as follows. First, inclusion of the linear attenuation mechanisms may yield additional understanding of the real effects of attenuation. Second, the use of a finite element method easily allows for complicated geometry in the solution domain. Third, many applications in seismology entail computing solutions for long time periods; the elastic energy traverses many wavelengths in the solution domain. In that traversal, any method will introduce artificial dispersion into the solution, causing the solution quality to vary with solution frequency. In solving the LAFDWE, however, we should be able to tune our mesh to the frequency of solution desired. For earth-like parameters of the linear attenuation mechanisms in the LAFDWE, the solution at a given frequency does not vary over an unlimited range of spatial scales. This advantage of a frequency domain solution was mentioned by Marfurt (1984). Fourth, the analysis of seismological problems in the context of the theory of elliptic partial differential equations is essentially absent from the literature, and this theory is a very powerful way of describing such problems in the presence of singular sources.

The analysis presented herein is based primarily on the exhaustive analysis of very general (e.g., non-selfadjoint) variational methods discussed in Babuška & Aziz (1972), Babuška (1971) and Roitberg (1996). Problems very similar to the one at hand are considered in Demkowicz & Oden (1994) and Ihlenburg & Babuška (1995).

Although most symbols used in this chapter are described as the discussion proceeds, appendix G contains a comprehensive list of symbols used.

4.3 Motivation

The accurate computation of the wavefield in Earth-like media is important. We need to know in detail how the wavefield is generated at the source and interacts with heterogeneities and interfaces to generate the observed seismogram. In fields such as marine seismology or exploration geophysics, complicated boundaries are present which clearly require accurate computation of the full wavefield. In global geophysics, approximate methods abound, and are certainly essential if we wish to make an inverse problem practical. However, it is also important to determine what bias, if any, is introduced into our interpretation of the seismogram by these approximations. The details of how these approximate methods might fail are unknown, and probably depend on what (hypothetical) media they are attempting to model. Extensions of these methods are perhaps less questionable, but exactly how much better are they?

In the context of global seismology, perturbation theory has yielded a number of approximations of the low-frequency response of the earth. Besides assuming small perturbations from a reference model, these methods are computationally inefficient once mode coupling is taken into account or the frequency of interest becomes sufficiently high. Mode summation or asymptotic mode techniques are presented in Park & Gilbert (1986), Woodhouse & Dziewonski (1984), Li & Romanowicz (1996), and Marquering & Snieder (1995). Methods such as the Born approximation also assume that heterogeneities are small perturbations (Snieder & Romanowicz, 1988), and it is clear that this approximation fails in some settings (Hudson & Heritage, 1981).

Ray-based methods assume that the wavelength of the seismic waves is much less than the wavelength of the heterogeneities present in the medium. Modifications of these high frequency methods attempt to extend ray theory to finite frequency. Maslov theory is expounded in Chapman & Drummond (1982). A comparison of Maslov theory to a finite difference result, as an example of the importance of full wavefield solutions, can be found in Huang, et al. (1996). Gaussian Beam summation (Červený, et al., 1982) and the Kirchoff method of Frazer & Sinton (1984) are other examples of ray-based approaches. All of these methods are sufficiently questionable that they require verification by more accurate methods.

Finite difference techniques are quite popular for solving the full wavefield problem (Viriuex, 1986). Finite difference techniques have been extended to very large problems in Igel & Weber (1995) and by others, and are beginning to be used to understand the bias of tomographic forward problems (Igel & Gudmundsson, 1996). In the context of marine geophysics, examples of the usefulness of finite difference techniques in media containing heterogeneities and very complicated interfaces are found in Dougherty & Stephen (1988), Dougherty & Stephen (1991) and Wilcock, et al. (1993). The pseudo-spectral methods (Wang & Tromp, 1996; Hung, et al., 1996; Fornberg, 1996) are very efficient ways of implementing finite difference schemes in domains with simple boundaries.

The methods analyzed in this chapter can take advantage of several useful aspects of other full wavefield techniques, combining the computational efficiency of pseudo-spectral methods with the flexibility of finite elements. Regardless of the method used, an accurate reference solution for the approximate methods listed above is needed. Since the problem of simulating wave propagation in the entire Earth is so computationally intensive, the evaluation of approximate forward problems in 2-D is a reasonable first step. 2-D simulations have had some success in isolating some surprising effects of waveform distortion due to finite frequency effects (personal communication, Guust Nolet). Of particular interest are complete tomography experiments in 2-D.

4.4 Mathematical Groundwork

The tools necessary to analyze the problem at hand are briefly described. The intention is to provide the minimum background to make the problem easy to analyze. While the tools developed are generally useful in any dimension, for simplicity the examples of details will be in one spatial dimension to prevent extraneous complication of nomenclature. In section 4.9, we comment in some detail about the 2-D and 3-D problems.

4.4.1 Smooth functions

The set of all complex valued functions on $\mathbb{R}^n$ with continuous derivatives up to and including order k are denoted $C^k(\mathbb{R}^n)$. The functions in $C^k_B(\mathbb{R}^n)$ are bounded in modulus as well. The subset of $C^k(\mathbb{R}^n)$ with compact support is denoted $C^k_0(\mathbb{R}^n)$. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain, and $\partial\Omega$ be its boundary, so that $\overline{\Omega} = \Omega \cup \partial\Omega$. The set of the restrictions of $C^\infty_0(\mathbb{R}^n)$ to $\overline{\Omega}$ is written as $C^\infty(\overline{\Omega})$. The functions in $C^\infty(\overline{\Omega})$ will be used to define all of the distributions on Ω and its boundary $\partial\Omega$. It is assumed that $\partial\Omega$ is sufficiently smooth that it makes sense to extend derivatives on Ω to $\partial\Omega$.

4.4.2 Distributions on Ω

Define the complex inner product on the domain Ω

$$(u, v) = \int\limits_{\Omega} \overline{u}v, \tag{4.4.2.1}$$

and the inner product on the boundary as

$$\langle u, v\rangle = \int\limits_{\partial\Omega} \overline{u}v. \tag{4.4.2.2}$$

A distribution is a functional which assigns a complex value to a test function ϕ. For example, a distribution f which is *regular* can be defined in terms of a function g such that

$$(f, \phi) = (g, \phi), \forall \phi \in C^\infty(\overline{\Omega}). \tag{4.4.2.3}$$

The operation of differentiation is defined for all distributions by integration by parts, or more generally by Green's theorem. For example, on the 1-D domain $\Omega = (0, 1)$,

$$(D_z f, \phi) = f\phi|_0^1 - (f, D_z\phi). \tag{4.4.2.4}$$

4.4.3 The Sobolev Spaces $H^s(\Omega)$

Let Ω be a subset of $\mathbb{R}^n$, and let $k = \{k_1, k_2, \ldots, k_n\}$, where the k_j are positive integers and define $|k| = k_1 + k_2 + \ldots k_n$. Define the derivative operator as

$$D^k = \frac{\partial^{k_1}}{\partial {x_1}^{k_1}} \frac{\partial^{k_2}}{\partial {x_2}^{k_2}} \cdots \frac{\partial^{k_n}}{\partial {x_n}^{k_n}}. \tag{4.4.3.5}$$

The Sobolev space $H^s(\Omega)$ is defined as the completion of $C^\infty(\overline{\Omega})$ under the H^s norm

$$\|u\|_s = \left(\sum_{|k|=0}^{s} \left\|D^k u\right\|_0^2 \right)^{\frac{1}{2}}, \; u \in C^\infty(\overline{\Omega}), s \geq 0, \tag{4.4.3.6}$$

where the L^2 (H^0) norm is

$$\|v\|_0 = (u, u)^{\frac{1}{2}} = \left(\int_\Omega \overline{u} u \right)^{\frac{1}{2}}. \tag{4.4.3.7}$$

The negative Sobolev spaces are similarly defined as the completions of $C^\infty(\overline{\Omega})$ under the norm

$$\|u\|_s = \sup_{v \in C^\infty(\overline{\Omega})} \frac{|(u, v)|}{\|v\|_{-s}}, \quad u \in C^\infty(\overline{\Omega}), \quad s < 0. \tag{4.4.3.8}$$

The nature of the negative Sobolev spaces will become more evident in the next section. The completion of $C^\infty(\overline{\Omega})$ includes the limits $\hat{x}$ of all Cauchy sequences in $C^\infty(\overline{\Omega})$ in the completed space. The Cauchy sequences in $H^s(\Omega)$

are those $\{x_n\}, x_n \in C^\infty(\overline{\Omega})$ such that $p, q > n_0$ constrains $\|x_p - x_q\|_s < \epsilon$. Most importantly, $C^\infty(\overline{\Omega})$ is *dense* in $H^s(\Omega)$, i.e., for any $\epsilon > 0, \hat{u} \in H^s(\Omega)$, $\exists u \in C^\infty(\overline{\Omega}) | \, \|\hat{u} - u\|_s < \epsilon$. For example, let $\hat{u} = \lim_{m \to \infty} x_m$. Choose u to be the constant sequence $\{x_{n_0}\}$ such that $\|x_{n_0} - x_m\| < \epsilon$, which is always possible by the definition of a Cauchy sequence. Then $\|\hat{u} - x_{n_0}\|_s = \left\| \lim_{m \to \infty} (x_{n_0} - x_m) \right\|_s$. Since each $x_{n_0}, x_m \in H^s(\Omega)$, $\|\hat{u} - x_{n_0}\|_s = \lim_{m \to \infty} \|x_{n_0} - x_m\|_s < \epsilon$.

All of the Sobolev spaces (and their Fourier transforms) are Hilbert spaces. The inner product on the space $H^s(\Omega)$ will be referred to as $(\bullet, \bullet)_s$, so that $(\bullet, \bullet)_0 = (\bullet, \bullet)$. Further equivalent norms and inner products are defined in section 4.6.1.

4.4.4 Dense Inclusions

The Sobolev spaces are such that

$$C^\infty(\overline{\Omega}) \subset H^s(\Omega) \subset H^k(\Omega), \; s, k \in \mathbb{N}, \; s > k \; (\Omega \text{ bounded}), \tag{4.4.4.9}$$

where the inclusions are dense. The importance of these dense inclusions is that relationships between dense subsets can be *extended* to the superset. Define a continuous linear functional T such that $|T(u)| \leq C \|u\|_S, \forall u \in S$. C is a constant which does not depend on u.

Theorem 4.4.1 (Extension of Identities) *Let T_1 and T_2 be two continuous linear functionals on S, and let s be a dense subspace of S. If $T_1(u) = T_2(u)$, $\forall u \in s$, then $T_1(v) = T_2(v)$,$\forall v \in S$.*

Proof: s is dense in S, so that given $\epsilon > 0$, there exists a $u \in s$ such that $\|u - v\|_S < \epsilon$. T_1 and T_2 are bounded, so that $|T_1(u - v)| \leq C_1 \|u - v\|_S < C_1\epsilon$, and $|T_2(u - v)| \leq C_2 \|u - v\|_S < C_2\epsilon$. Since $|a - b| \leq |a| + |b|$, by linearity of T_1 and T_2, $|T_1(u) - T_1(v) - T_2(u) + T_2(v)| < (C_1 + C_2)\epsilon = \epsilon'$. By the condition on the operators, $|T_1(v) - T_2(v)| < \epsilon'$, $\forall v \in S$, and for any $\epsilon' > 0$. Since ϵ' is arbitrary, $|T_1(v) - T_2(v)| = 0$.

A similar analysis reveals the same can be said for inequalities, so we have

Theorem 4.4.2 (Extension of Inequalities) *Let* T_1 *and* T_2 *be two continuous linear functionals on* S*, and let* s *be a dense subspace of* S*. If* $T_1(u) \leq T_2(u)$*,* $\forall u \in s$*, then* $T_1(v) \leq T_2(v)$*,* $\forall v \in S$*.*

This allows us to better understand the definition of the negative Sobolev spaces. The completion process defines the norm of $u \in H^s(\Omega)$ as

$$\|u\|_s = \sup_{v \in C^\infty(\overline{\Omega})} \frac{|(u,v)|}{\|v\|_{-s}}, \quad s < 0. \tag{4.4.4.10}$$

This is clearly equivalent to the inequality

$$\|v\|_{-s} \|u\|_s \geq |(u,v)|, \forall v \in C^\infty(\overline{\Omega}). \tag{4.4.4.11}$$

Using theorem 4.4.2 and considering u a fixed element of $H^s(\Omega)$, we can extend this to the useful statement

$$\|v\|_{-s} \|u\|_s \geq |(u,v)|, \forall v \in H^{-s}(\Omega). \tag{4.4.4.12}$$

Thus, we can rewrite the definition of the negative Sobolev norm as

$$\|u\|_s = \sup_{v \in H^{-s}(\Omega)} \frac{|(u,v)|}{\|v\|_{-s}}, \quad s < 0. \tag{4.4.4.13}$$

The negative Sobolev spaces are therefore naturally paired with the positive Sobolev spaces; H^s is the *dual* of H^{-s}. The inner product on H^0 makes sense when it operates on such a pair of distributions.

4.4.5 The Sobolev Imbedding Theorem

The Sobolev spaces are imbedded in various other function spaces. We will make use of the following theorem which relates the square-integrability of some distributions to their properties at a point.

Theorem 4.4.3 (Sobolev Imbedding Theorem) *Let* $\Omega \subset \mathbb{R}^n$. *If* $s \geq n/2 + k$, $H^s(\Omega) \subset C^k_B(\Omega)$, *and for* $u \in H^s(\Omega)$,

$$\sup_{x\in\Omega} \left|(D^k u)(x)\right| \leq C \left\|u\right\|_s, \tag{4.4.5.14}$$

where C *does not depend on* u.

This theorem follows immediately from Lemma 5.15 and the proof of Corollary 5.16 of Adams (1975), p. 107-108. The discussion of the imbeddings of the Sobolev spaces in Adams (1975), chapter 5, is quite useful. Also see Theorem 5.5 of Al-Gwaiz (1992), p. 179, and the subsequent corollary, and Roitberg (1996), p. 61-62. The change of the classification of distributions by their continuity with dimension hinges on the fact that the volume element in the definition of the Sobolev norms changes with dimension.

The continuity properties of distributions in Sobolev spaces allows us to classify all of the delta distributions and their derivatives. We know that these distributions are in negative Sobolev spaces, since they must integrated against continuous functions to make sense. Using the definition of the negative Sobolev spaces,

$$\left\|\delta\right\|_s = \sup_{v\in H^{-s}(\Omega)} \frac{\left|(\delta, v)\right|}{\left\|v\right\|_{-s}}, \quad s < 0. \tag{4.4.5.15}$$

By definition, $(\delta, v) = v(0)$, so we have

$$\left\|\delta\right\|_{-s} = \sup_{v\in H^{-s}(\Omega)} \frac{\left|v(0)\right|}{\left\|v\right\|_s}, \quad s > 0. \tag{4.4.5.16}$$

This expression is bounded, by theorem 4.4.3, if we constrain $s > n/2$. The smallest Sobolev space to which δ belongs is therefore $H^{-[n/2+1]}$, where $[\cdot]$ is the integer part.

The classification of the derivatives of the delta distributions follows from the definition $(\partial_j\delta, v) = \partial_j v(0)$, where $1 \leq j \leq n$. Therefore

$$\left\|\partial_j\delta\right\|_{-s} = \sup_{v\in H^{-s}(\Omega)} \frac{\left|\partial_j v(0)\right|}{\left\|v\right\|_s}, \quad s > 0, \tag{4.4.5.17}$$

is bounded only if, by theorem 4.4.3, $s > n/2 + 1$, and therefore $\partial_j \delta \in H^{-[n/2+2]}$. Table 4.1 summarizes these results for all interesting dimensions. We note that equivalent forces of interest in seismology fall into the $\partial_j \delta$ category (see section 4.5 below).

Table 4.1: Classification of delta distributions and their derivatives in $\mathbb{R}^n$

Distribution	$n = 1$	$n = 2$	$n = 3$
δ	$H^{-1}(\Omega)$	$H^{-2}(\Omega)$	$H^{-2}(\Omega)$
$\partial_j \delta$	$H^{-2}(\Omega)$	$H^{-3}(\Omega)$	$H^{-3}(\Omega)$

See also Strang & Fix (1973), p. 73.

4.4.6 The trace spaces $H^{s-1/2}(\partial\Omega)$ and the spaces $\tilde{H}^s(\Omega)$

The *trace* of a distribution $u \in H^s$ on the boundary of the domain, $u|_{\partial\Omega}$, has a number of well defined properties depending upon s and the dimension. The norm in $H^{s-1/2}(\partial\Omega)$ is defined as

$$\langle\!\langle u \rangle\!\rangle_{s-1/2} = \inf_{\substack{\phi \in C^\infty(\overline{\Omega}) \\ \phi|_{\partial\Omega} = u}} \|\phi\|_s \,, \quad s = 1, 2, \ldots. \tag{4.4.6.18}$$

$H^{s-1/2}(\partial\Omega)$ is the completion of $C^\infty(\partial\Omega)$ in this norm. The dual trace spaces are defined so that

$$\langle\!\langle u \rangle\!\rangle_{-s+1/2} = \sup_{v \in H^{s-1/2}(\partial\Omega)} \frac{|\langle u, v \rangle|}{\langle\!\langle v \rangle\!\rangle_{s-1/2}}, \quad s = 1, 2, \ldots. \tag{4.4.6.19}$$

Likewise, $H^{-s+1/2}(\partial\Omega)$ is the completion of $C^\infty(\partial\Omega)$ in this norm. Analogously to equation 4.4.4.12, we therefore obtain

$$\langle\!\langle v \rangle\!\rangle_{-s+1/2} \langle\!\langle u \rangle\!\rangle_{s-1/2} \geq |\langle u, v \rangle| \,. \tag{4.4.6.20}$$

The trace norms (and more generally the non-integer Sobolev spaces) can be defined in various ways (see Adams (1975), Chapter 7). An equivalent norm to

4.4.6.18 for $n \geq 1$ is, for any real $s \geq 0$,

$$\langle\!\langle u \rangle\!\rangle_s^2 = \|u\|_{[s]}^2 + \sum_{[\alpha]=[s]} \int_{\partial\Omega}\int_{\partial\Omega} \frac{\left|D^\alpha u(x) - D^\alpha u(y)\right|^2}{\left|x-y\right|^{n-1+2(s-[s])}} dxdy, \tag{4.4.6.21}$$

where x and y each range over $\partial\Omega \subset \mathbb{R}^{n-1}$. $[\alpha]$ is the integer part of α. It should be understood that the norm will apply not precisely to u but to some set of smooth mappings of u to $\mathbb{R}^{n-1}$ (Adams, 1975, p. 215). See also the discussion of trace norms in Aubin (1972), chapter 6, section 3. The significant fact of the trace spaces is the loss of 1/2 a derivative relative to the related domain space.

Finally, we will make use of distributions with certain regularity properties both on a domain and its boundaries. The spaces of most general interest for second order differential operators are the $\tilde{H}^s(\Omega)$ defined as the completion of $C^\infty(\overline{\Omega})$ in the norm

$$\langle\!| u |\!\rangle_s = \left(\|u\|_s^2 + \langle\!\langle u|_{\partial\Omega} \rangle\!\rangle_{s-1/2}^2 + \langle\!\langle D_n u|_{\partial\Omega} \rangle\!\rangle_{s-3/2}^2 \right)^{1/2}. \tag{4.4.6.22}$$

For distributions with sufficient regularity, the $\tilde{H}^s(\Omega)$ are equivalent to the $H^s(\Omega)$. Using equation 4.4.6.18, it is clear that

$$\langle\!\langle u|_{\partial\Omega} \rangle\!\rangle_{s-1/2} \leq \|u\|_s, s = 1, 2, \ldots. \tag{4.4.6.23}$$

As well, $\langle\!\langle Du|_{\partial\Omega} \rangle\!\rangle_{s-1/2} \leq \|Du\|_s, s = 1, 2, \ldots$. From the definition of the Sobolev norm, $\|Du\|_s \leq \|u\|_{s+1}$. Substituting, we obtain

$$\langle\!\langle Du|_{\partial\Omega} \rangle\!\rangle_{s-3/2} \leq \|u\|_s, s = 2, 3, \ldots. \tag{4.4.6.24}$$

Therefore, if $s \geq 2$, then $\langle\!| u |\!\rangle_s \leq C\|u\|_s$ and we have that $\tilde{H}^s(\Omega) \equiv H^s(\Omega), s \geq 2$. For $s \leq 2$, these are not equivalent spaces; one or both of the trace elements becomes independent of the element on the domain. For that reason, we will most generally refer to members U of $\tilde{H}^s(\Omega)$ as the triple of distributions $U = \{u, u_0, u_1\}$ (after Babuška & Aziz (1972), p.73). Table 4.2 summarizes the nature of these triples, where we have chosen the normal derivative as in typical usage. Extensive discussion of the $\tilde{H}^s(\Omega)$ is found in Roitberg (1996), Ch. 2.

Table 4.2: Triples of distributions $U = \{u, u_0, u_1\}$ in $\tilde{H}^s(\Omega)$

$s \geq 2$	$u \in H^s(\Omega)$	$u_0 = u\|_{\partial\Omega} \in H^{s-1/2}(\partial\Omega)$	$u_1 = D_n u\|_{\partial\Omega} \in H^{s-3/2}(\partial\Omega)$
$s = 1$	$u \in H^1(\Omega)$	$u_0 = u\|_{\partial\Omega} \in H^{1/2}(\partial\Omega)$	$u_1 \in H^{-1/2}(\partial\Omega)$
$s = 0$	$u \in H^0(\Omega)$	$u_0 \in H^{-1/2}(\partial\Omega)$	$u_1 \in H^{-3/2}(\partial\Omega)$

In 1-D, the trace spaces become very transparent and simple. Consider the domain $\Omega = (0,1)$, and a sufficiently smooth function f. Clearly,

$$f(0) = f(t) - \int_0^t f'(\tau)d\tau$$

Using the triangle inequality and bounding an integral using the maximum value of the integrand, it is easy to show that

$$|f(0)| \leq \int_0^1 |f(t)| + \int_0^1 |f'(t)|dt.$$

By Hölder's inequality (Adams, 1975, chapter 2), we have therefore that

$$|f(0)|^2 \leq \int_0^1 |f(t)|^2 + \int_0^1 |f'(t)|^2 dt. \tag{4.4.6.25}$$

The same can be said for $|f(1)|^2$. Clearly, if we wish the trace of a 1-D distribution g to be bounded, we can either require that $g \in H^1(\Omega)$ or we must separately require that $|g(0)|^2 + |g(1)|^2 \leq \infty$. In the first case, equation 4.4.6.25 is sufficient to insure that $|g(0)|^2 + |g(1)|^2 \leq \|g\|_1^2 \leq \infty$. It is also evident how $|g'(0)|^2$ behaves, so that for integer s and k,

$$|D^k g(0)|^2 + |D^k g(1)|^2 \leq \|g\|_s^2, 0 \leq k < s. \tag{4.4.6.26}$$

A suitable norm for the traces of all 1-D distributions (regardless of their regularity) on the unit interval is therefore

$$\langle\langle u|_{\partial\Omega} \rangle\rangle_0 = \left(|u(0)|^2 + |u(1)|^2\right)^{1/2}. \tag{4.4.6.27}$$

The trace spaces in 1-D are therefore composed of all of the pairs of bounded complex numbers. This space is $H^0(\partial\Omega)$ (in 1-D). In 1-D, the duality of the trace spaces is very simple, since the integral L^2 inner product is replaced by the simple product. For any two distributions u and v in 1-D,

$$|\overline{u}v|_0^1| = |\overline{u}(1)v(1) - \overline{u}(0)v(0)| \leq \langle\!\langle u|_{\partial\Omega}\rangle\!\rangle_0 \langle\!\langle v|_{\partial\Omega}\rangle\!\rangle_0 . \tag{4.4.6.28}$$

The relationship of these quantities to the (Sobolev) norms on the domain follows from equation 4.4.6.26.

All of the results for higher dimensions (such as the following theorem on the regularity of the solution of 2nd order PDE's) are true in 1 dimension; however, all of the trace spaces and norms reduce to $H^0(\partial\Omega)$ and $\langle\!\langle\cdot\rangle\!\rangle_0$, respectively.

4.4.7 The Fourier Transform and Sobolev spaces

The Fourier transform is especially well-suited for the understanding of the Sobolev spaces and their traces. We consider several direct calculations which illuminate the nature of these spaces. The discussion of Al-Gwaiz (1992) amongst others is quite useful.

The Fourier transform $\mathcal{F}$ of a distribution is defined as an extension of these transforms for L^1 functions, where

$$\mathcal{F}(u)(\omega) = \int e^{-i\omega\cdot x} f(x) dx \tag{4.4.7.29}$$

The Fourier Transform $\mathcal{F}$ of the distribution u, or $\hat{u}$, is defined as an integral operation on the test functions, so that

$$(\hat{u}, \phi) = \left(u, \check{\phi}\right) \tag{4.4.7.30}$$

where the ˇ signifies evaluation of the Fourier transform at the reversed sign of the argument:

$$\check{\phi}(x) = \hat{\phi}(-x). \tag{4.4.7.31}$$

This peculiar reversal arises from the complex conjugate in the definition of the inner product. An example for an L^1 function u derives from a change in the order of integration (Al-Gwaiz, 1992, p.129):

$$\begin{aligned} (\hat{u}, \phi) &= \int d\omega \; \overline{\hat{u}}(\omega)\phi(\omega) \\ &= \int dx \; \overline{u}(x) \left(\int d\omega \; \phi(\omega) e^{i\omega \cdot x} \right) \\ &= (u, \check{\phi}) \end{aligned} \tag{4.4.7.32}$$

As for L^1 functions, the Fourier transform of a distribution possesses the property that

$$\mathcal{F}(D^k u) = (-i)^{|k|} \omega^k \mathcal{F}(u), \tag{4.4.7.33}$$

which is evident from some manipulation of the above definitions. In this case the multi-index exponent on ω is meant in the same fashion as for the exponent of the operator D (Al-Gwaiz, 1992, p. 16).

Using this derivative property, it is clear that we can rewrite the definition of the Sobolev norms in equation 4.4.3.6. Using Parseval's relation (Al-Gwaiz, 1992, equation 4.10),

$$\|u\|_s = C \|\hat{u}\|_s = \left(\sum_{|k|=0}^{s} \left\| \omega^k \hat{u} \right\|_0 \right)^{\frac{1}{2}} . \tag{4.4.7.34}$$

A little algebra shows that this is an equivalent norm to

$$\begin{aligned} \|u\|_s &= \left\| (1 + |\omega|^2)^{s/2} \hat{u} \right\|_0 \\ &= \int d\omega \;\; (1 + |\omega|^2)^s |\hat{u}(\omega)|^2) \end{aligned} \tag{4.4.7.35}$$

so that $u \in H^s$ if $\left\| (1 + |\omega|^2)^{s/2} \hat{u} \right\|_0 < \infty$. It also follows that, if $u \in H^s$, $D^k u \in H^{s-k}$. We can therefore think of the Sobolev spaces as distributions whose spectra have some bounding slope at large wavenumber $|\omega|$. The extension to any real s is evident as well.

The nature of the dual of a Sobolev space is also immediately evident. Using Parseval's relation again we see that

$$(u,v) = \int \overline{u}v = C\int \overline{\hat{u}}\hat{v} = \int (1+|\omega|^2)^{s/2}\overline{\hat{u}}(1+|\omega|^2)^{-s/2}\hat{v}, \tag{4.4.7.36}$$

and using Schwarz' inequality and the Sobolev norm definition we reproduce equation 4.4.4.12,

$$|(u,v)| \leq \|u\|_s \, \|v\|_{-s}\,. \tag{4.4.7.37}$$

Equation 4.4.6.20 for the boundary spaces also follows.

The relationship of the boundary Sobolev spaces to the Sobolev spaces on the domain is illuminated by the following argument. Consider the Fourier transform in $\mathbb{R}^{n-1}$ of a function $v \in L^1$ on $\mathbb{R}^n$. We evaluate the inverse Fourier transform on the plane at x_n, obtaining

$$\hat{v}|_{x_n}(\omega_1,\ldots,\omega_{n-1}) = \int d\omega_n e^{ix_n\omega_n}\hat{v}(\omega_1,\ldots,\omega_n). \tag{4.4.7.38}$$

In order to ascertain the Sobolev space to which $\hat{v}|_{x_n}$ belongs, we evaluate the norm on $H^t(\partial\Omega)$,

$$A = \int d\omega_1 \ldots d\omega_{n-1}\left|(1+\omega_1^2+\ldots+\omega_{n-1}^2)^{t/2}\int d\omega_n\hat{v}(\omega_1,\ldots,\omega_n)\right|^2 \tag{4.4.7.39}$$

and require that this be bounded. Evaluating at any convenient x_n and wedging in a dummy exponent s, we can bound the previous quantity by

$$\begin{aligned} A \leq \int d\omega_1 \ldots d\omega_{n-1}\Bigg|\int d\omega_n &\frac{(1+\omega_1^2+\ldots+\omega_{n-1}^2)^{t/2}}{(1+\omega_1^2+\ldots+\omega_n^2)^{s/2}} \\ &\cdot (1+\omega_1^2+\ldots+\omega_n^2)^{s/2}\hat{v}(\omega_1,\ldots,\omega_n)\Bigg|^2. \end{aligned} \tag{4.4.7.40}$$

Applying Hölder's inequality, we then arrive at

$$\begin{aligned} A \leq \int d\omega_1 \ldots d\omega_{n-1}\Bigg[\int d\omega_n &\frac{(1+\omega_1^2+\ldots+\omega_{n-1}^2)^{t}}{(1+\omega_1^2+\ldots+\omega_n^2)^{s}} \\ &\cdot \int d\omega_n (1+\omega_1^2+\ldots+\omega_n^2)^{s}\hat{v}^2\Bigg]. \end{aligned} \tag{4.4.7.41}$$

By the suggestion of Aubin (1972) in the proof of theorem 3-5, p. 197, we make the substitution of variables

$$\omega_n = \lambda(1 + \omega_1^2 + \ldots + \omega_{n-1}^2)^{1/2} \tag{4.4.7.42}$$

and rewrite the integral

$$\int d\omega_n \frac{(1 + \omega_1^2 + \ldots + \omega_{n-1}^2)^t}{(1 + \omega_1^2 + \ldots + \omega_n^2)^s}, \tag{4.4.7.43}$$

as

$$(1 + \omega_1^2 + \ldots + \omega_{n-1}^2)^{t+1/2-s} \int \frac{d\lambda}{(1 + \lambda^2)^s}. \tag{4.4.7.44}$$

When $s > 1/2$, this integral converges to $C(1 + \omega_1^2 + \ldots + \omega_{n-1}^2)^{t+1/2-s}$. This provides the bound

$$\begin{aligned} A \leq C \int d\omega_1 \ldots d\omega_n \hat{v}^2 (1 + \omega_1^2 + \ldots + \omega_n^2)^s \\ \cdot (1 + \omega_1^2 + \ldots + \omega_{n-1}^2)^{t+1/2-s}. \end{aligned} \tag{4.4.7.45}$$

Distinguishing the two cases $t + 1/2 - s > 0$ and $t + 1/2 - s \leq 0$ yields the general conclusion that

$$\|v|_{\partial\Omega}\|_{s-1/2} \leq C \|v\|_s, \text{ if } s \geq 1. \tag{4.4.7.46}$$

Likewise, it clearly follows that

$$\left\|\partial^k v|_{\partial\Omega}\right\|_{s-k-1/2} \leq C \|v\|_s, \text{ if } s > k + 1/2. \tag{4.4.7.47}$$

These results are contained in the general definitions of the boundary spaces.

The Sobolev spaces to which the delta distributions and their derivatives belong is also nicely demonstrated using the Fourier transform. Using equation 4.4.7.30, it is clear that

$$(\hat{\delta}, \phi) = (\delta, \check{\phi}) = \check{\phi}(0) = (1, \phi), \tag{4.4.7.48}$$

so that $\hat{\delta} = 1$. Using equation 4.4.7.33, for any partial derivative of the delta distribution,

$$\int (1 + |\omega|^2)^m |(\mathcal{F}(D^k\delta))(\omega)|^2 d\omega < C \int (1 + |\omega|^2)^m |(\omega)|^{2k} |\omega|^{n-1} d|\omega|, \quad (4.4.7.49)$$

which is bounded when $2m + 2k + n - 1 < -1$, or $m < -(k + n/2)$. This clearly confirms the above results (table 4.1).

4.4.8 Elliptic Regularity

We now have the necessary tools to define what the relationship between the data f, boundary conditions g_j and solution u are in a boundary value problem (BVP) of the form $Lu = f$, $B_j u|_{\partial\Omega} = g_j$. We assume that L is a second order differential operator so that the B_j are differential operators on the boundary of order 0 or 1. In other words, homogeneous Dirichlet boundary conditions correspond to $u|_{\partial\Omega} = B_0 u$, $g_0 = 0$, Neumann to $D_n u|_{\partial\Omega} = B_1 u$, and $g_1 = 0$. (D_n is the normal derivative.) Additionally, assume that the coefficients of the operators L and the B_j are $C^\infty(\overline{\Omega})$ functions.

The possible existence and uniqueness of the solution to this BVP are covered by Theorem 3.8.1 of Babuška & Aziz (1972):

Theorem 4.4.4 (Elliptic Regularity) *Let L be a second order elliptic differential operator and $f \in H^{s-2}(\Omega)$, and B_0, B_1, g_0 and g_1 represent compatible boundary conditions, where s is any integer. Define the null spaces N of $\{L, B_0, B_1\}$ and N^* of its adjoint $\{L^*, B_0^*, B_1^*\}$ as*

$$N = \{v | Lv = 0, B_0 v = 0|_{\partial\Omega}, B_1 v = 0|_{\partial\Omega}\} \quad (4.4.8.50)$$

and

$$N^* = \{v | L^* v = 0, B_0^* v = 0|_{\partial\Omega}, B_1^* v = 0|_{\partial\Omega}\}. \quad (4.4.8.51)$$

If $N = N^* = \{0\}$, then

$$\{L, B_0, B_1\} : \tilde{H}^s(\Omega) \to H^{s-2}(\Omega) \times H^{s-1/2}(\partial\Omega) \times H^{s-3/2}(\partial\Omega) \quad (4.4.8.52)$$

is a one-to-one, onto map, and

$$\langle\!|u|\!\rangle_s \leq C\left[\|f\|_{s-2} + \langle\!\langle g_0\rangle\!\rangle_{s-1/2} + \langle\!\langle g_1\rangle\!\rangle_{s-3/2}\right]. \tag{4.4.8.53}$$

The proof of this theorem is not conceptually difficult. The space to which the solution u belongs is examined by integrating the differential equation $(Lu, \phi) = (f, \phi), \forall u, \phi \in C^\infty(\overline{\Omega})$ by parts and requiring that the resulting expressions make sense (i.e., are bounded). This results in the extension of the differential equation to hold on the space of distributions $\tilde{H}^s(\Omega)$. The fact that the null spaces of the operators $\{L, B_0, B_1\}$ and $\{L^*, B_0^*, B_1^*\}$ are empty results in a $1:1$, onto mapping. In other words, the solution exists and is unique. Extension of this theorem to systems of equations is discussed in section 4.9.

4.4.9 Local Regularity

We will make use of a special case of elliptic regularity to make some error estimates more concrete in some seismological applications. In particular, we typically have forcing which is zero (i.e., in $C^\infty(\overline{\Omega})$) away from the source point. In such cases, the solution itself is smoother away from the source.

Suppose we construct a smooth function χ that vanishes in the neighborhood of the subdomain $(\Gamma_1 \subset \partial\Omega) \cup (\Omega_1 \subset \Omega)$. Under all of the assumptions leading to the BVP for which elliptic regularity holds, we have the following theorem.

Theorem 4.4.5 (Local Regularity) *Let* $u \in \tilde{H}^s(\Omega)$ *be the solution to the above BVP with* $f \in H^{s-2}(\Omega)$, $g_0 \in H^{s-1/2}(\partial\Omega)$ *and* $g_1 \in H^{s-3/2}(\partial\Omega)$. *Then, if* $\chi f \in H^{t-2}(\Omega)$, $\chi g_0 \in H^{t-1/2}(\partial\Omega)$ *and* $\chi g_1 \in H^{t-3/2}(\partial\Omega)$, *where* $t > s$, *then* $\chi u \in \tilde{H}^t(\Omega)$. *Moreover,*

$$\langle\!|\chi u|\!\rangle_t \leq C\left[\|\chi f\|_{t-2} + \langle\!\langle \chi g_0\rangle\!\rangle_{t-1/2} + \langle\!\langle \chi g_1\rangle\!\rangle_{t-3/2} + \langle\!|\chi u|\!\rangle_s\right]. \tag{4.4.9.54}$$

The proof of this theorem is presented in Roitberg (1996), p. 216. This theorem is of some use in the discussion of the convergence of the $H^0(\Omega)$ FEM. In particular, if the forcing is identically zero in some region including the boundary, then $\chi f \in C^\infty(\overline{\Omega})$, and therefore $\chi u \in C^\infty(\overline{\Omega})$. We can therefore measure $u|_{\partial\Omega}$ in any convenient boundary norm on $H^{t-1/2}(\partial\Omega)$, where $t > s$. Sufficient smoothness in the vicinity of the boundary can assure that the traces of u and its derivatives are determined by u itself. This follows from table 4.2 (also see Roitberg (1996), p. 160).

4.4.10 Approximation Properties of $H^k_{h,p}$

The errors derived for the FEM method will be stated relative to the "best" approximation error. The best approximation is the projection of a function onto a set of approximating functions. The best approximation is best relative to the norm associated with that projection. The error in the approximation of functions by interpolation provides an upper bound on the error in the best approximation. Let $p-1$ be a maximum polynomial degree, and h be an interval length, and let $H^k_{h,p} \in H^k(\Omega)$ be a piecewise polynomial subspace of a Sobolev space.

Theorem 4.4.6 (Approximation Theorem) *For every* $g \in H^n$, $n \geq 0$, $0 \leq s \leq \min(n,k)$, *there exists a* $\phi \in H^k_{h,p}$ *such that for sufficiently small* h,

$$\|g-\phi\|_s \leq Ch^\kappa \|g\|_n\,, \quad \kappa = \min(p-s, n-s). \tag{4.4.10.55}$$

The constant C is not a function of g or h. This is a basic property of finite element spaces and is discussed at length in chapter 4 of Babuška & Aziz (1972). See also Strang & Fix (1973).

In higher dimensions, h is an element diameter, and the above theorem holds assuming that the elements are of the same geometry as h is varied. Note that H^n can refer to the regularity of functions *within* the elements. For example, theorem 4.4.6 holds with any positive integer n if discontinuities in the derivatives of $g \in H^1(\Omega)$ (in 1-D) are placed at element boundaries. Even when we are not guaranteed that the interpolate of g makes sense, the above theorem holds (Strang & Fix, 1973, p. 145).

4.5 The problem

The simplest wave equation of seismological interest is closely related to the linearized (small displacement) equation of motion Backus & Mulcahy (1976), equation 6.4a. After introducing time dependent elastic parameters to model attenuation, that wave equation becomes, after Fourier transformation in time

$$-\rho(\mathbf{r})\omega^2 u_i - \partial_j a_{ijkl}(\mathbf{r},\omega)\partial_k u_l = f_i(\omega) \ , \tag{4.5.0.1}$$

where $\mathbf{r}$ is spatial position, ω is frequency, and the vector displacement u_i and data f_i are distributions by the necessity of considering singular forcing. The effects of gravity and prestress have been completely ignored for simplicity. The parameters of the medium in question are assumed to be piecewise smooth. The medium is assumed to be bounded by a smooth surface $\partial\Omega$ with normal components n_j, and the Neumann B.C. is

$$[n_j(\mathbf{r})a_{ijkl}(\mathbf{r},\omega)\partial_k u_l] = \mathbf{0}, \ \ \mathbf{r} \in \partial\Omega \ . \tag{4.5.0.2}$$

The medium is assumed to be entirely composed of an (linearly attenuating) elastic solid, although in general fluid and solid regions are of interest. The elastic parameters a_{ijkl} obey certain symmetries and are, for isotropic media, given by

$$a_{ijkl}(\mathbf{r},\omega) = \lambda(\mathbf{r},\omega)\delta_{ij}\delta_{kl} + \mu(\mathbf{r},\omega)(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) \ . \tag{4.5.0.3}$$

The moduli are frequency dependent and given by an expression such as Park (1985), equations 1.50,

$$\mu(\mathbf{r},\omega) = \mu_0(\mathbf{r},\omega)\left\{1 + Q_\mu^{-1}(\mathbf{r})\left[(2/\pi)\ln(\omega/\omega_0) - i\right]\right\} , \tag{4.5.0.4}$$

and

$$\lambda(\mathbf{r},\omega) = \lambda_0(\mathbf{r},\omega)\left\{1 + Q_\lambda^{-1}(\mathbf{r})\left[(2/\pi)\ln(\omega/\omega_0) - i\right]\right\}, \tag{4.5.0.5}$$

where i is the imaginary unit and ω_0 is a reference frequency. In materials found in the solid mantle of the earth, Q_μ generally ranges from 50 to 1000.

We shall assume that the f_i are smooth everywhere except at a finite number of points internal to Ω, which of course are the locations of the sources. In the case of a so-called "ideal fault" (i.e., earthquake) source,

$$f_i = \Gamma_{ij}\partial_j\delta(\mathbf{s}), \tag{4.5.0.6}$$

where $\mathbf{s}$ is the source location. In the less singular case of the Green's functions for the problem,

$$f_i = \delta(\mathbf{s}). \tag{4.5.0.7}$$

Due to the complexity of higher dimensions, systems of equations, and variable coefficients ρ and a_{ijkl}, discussion of equation 4.5.0.1 is postponed until we consider the 1-D version of the problem in detail. The 1-D problem is more than sufficient to understand the problem well, and the theoretical results can be generalized to higher dimensions (see section 4.9).

4.6 The Standard FEM: Solutions for $f \in H^{-1}(\Omega)$

The BVP in equations 4.5.0.1 and 4.5.0.2 without the complication of higher dimensions and non-smooth coefficients ρ and a_{ijkl} is now studied. Equation 4.5.0.1 becomes

$$L_k u = f, \tag{4.6.0.1}$$

where

$$L_k(u(z)) = k^2 u(z) + \gamma D_z^2 u(z), \tag{4.6.0.2}$$

where $\gamma = 1 - i\epsilon$, $k, \epsilon \in \mathbb{R}^+$, $0 < \epsilon \ll 1$ and $z \in (0,1)$. The wavenumber parameter is $k^2 = w^2/v^2 = \rho w^2/\mu_0$, where v is the shear velocity.

The (Neumann) boundary conditions are

$$D_z u|_{\partial\Omega} = 0. \tag{4.6.0.3}$$

Equations 4.6.0.1 and 4.6.0.3 define a second order elliptic boundary value problem. The possible existence and uniqueness of the solution to this BVP are covered by Theorem 4.4.4.

Since our interest is in data f as singular as $f \in H^{-2}(\Omega)$ (in 1-D), theorem 4.4.4 distinguishes two cases. For $f \in H^{-1}(\Omega)$, i.e. for data no more singular than the delta distribution in 1-D, $u \in H^1(\Omega)$. The method of approximating the solution to this problem is covered in this section, and will be referred to as the standard of H^1 FEM, since it is the typical case for most applications. For $f \in H^{-2}(\Omega)$, as in the case of a faulted 1-D medium, theorem 4.4.4 requires $u \in H^0(\Omega)$. The method of solution for this case will be referred to as the discontinuous or H^0 FEM (since $H^0(\Omega)$ contains discontinuous distributions in 1-D), and is analyzed in section 4.8.

4.6.1 The sesquilinear form $B(u, v)$

If $f \in H^{-1}(\Omega)$, theorem 4.4.4 asserts that there is a unique solution u_0 to $Lu_0 = f$, $D_z u_0|_{\partial\Omega} = 0$ and that $u_0 \in H^1(\Omega)$. The distributional form of the differential equation is

$$(L_k(u_0), \phi) = (f, \phi), \quad \forall \phi \in C^\infty(\overline{\Omega}). \tag{4.6.1.4}$$

Consider the sesquilinear form

$$B(u, v) = k^2 (u, v) - \gamma (D_z u, D_z v), \tag{4.6.1.5}$$

and the linear functional

$$F(v) = (f, v). \tag{4.6.1.6}$$

A sesquilinear operator is linear in one argument and conjugate linear in the other. Define the inner products

$$\begin{aligned} (\!(u, v)\!)_0 &= k^2 \, (u, v) \, , & (4.6.1.7) \\ (\!(u, v)\!)_1 &= k^2 \, (u, v) + (D_z u, D_z v) & (4.6.1.8) \end{aligned}$$

and the associated norms $|\!|\!| \bullet |\!|\!|_0$ and $|\!|\!| \bullet |\!|\!|_1$. Physically, these norms (squared) correspond to the kinetic and total energy, respectively, in the perfectly elastic medium (i.e., $\gamma = 1$). In the linearly attenuating medium ($\gamma \neq 1$), these norms are still a sensible way to measure error. These norms are equivalent to the Sobolev norms in that

$$\begin{aligned} k \left\| u \right\|_0 = & \; |\!|\!| u |\!|\!|_0 \quad , & (4.6.1.9) \\ \min(1, k) \left\| u \right\|_1 \leq & \; |\!|\!| u |\!|\!|_1 \quad \leq \max(1, k) \left\| u \right\|_1 . & (4.6.1.10) \end{aligned}$$

The limits in the second expression are evident after applying the fact that

$$\left\| u \right\|_s \leq \left\| u \right\|_t , \quad s \leq t, u \in H^t. \tag{4.6.1.11}$$

Suppose $f \in H^{-1}(\Omega)$ so that $u \in H^1(\Omega)$. It is easy to show, using both the Schwarz and triangle inequalities that $|B(u, v)| \leq C_1 |\!|\!| u |\!|\!|_1 |\!|\!| v |\!|\!|_1$, where $C_1 = 3 + 2 |\gamma|$. Therefore B is continuous. Likewise, equation 4.4.4.12 indicates that $|L(v)| \leq \|f\|_{-1} \|v\|_1$, so that F is continuous. By theorem 4.4.1, if $B(u, v) = F(v)$, $\forall v \in C^\infty(\overline{\Omega})$, then $B(u, v) = F(v)$, $\forall v \in H^1(\Omega)$, since $B(u, v)$ and $F(v)$ are continuous and $C^\infty(\overline{\Omega})$ is a dense subset of $H^1(\Omega)$.

Let $u_0 \in H^1(\Omega)$ be the unique solution of $L(u_0) = f$, $D_z u_0|_{\partial\Omega} = 0$. Integrating $B(u_0, v)$ by parts

$$B(u_0, v) = k^2(u_0, v) + \gamma(D_z^2 u_0, v) - \gamma(D_z u_0) v|_0^1, \ \forall v \in C^\infty(\overline{\Omega}). \tag{4.6.1.12}$$

By definition, however, $D_z u_0|_{\partial\Omega} = 0$ and

$$k^2(u_0, v) + \gamma(D_z^2 u_0, v) = (L(u_0), v) = (f, v) = F(v), \forall v \in C^\infty(\overline{\Omega}). \qquad (4.6.1.13)$$

Therefore, extending by continuity,

$$B(u_0, v) = F(v), \forall v \in H^1(\Omega), \qquad (4.6.1.14)$$

so that the solution of the BVP in equations 4.6.0.1 – 4.6.0.3 is the solution of equation 4.6.1.14. Does equation 4.6.1.14 uniquely determine u_0? Conditions ii) and iii) on the operator B given in Theorem D.1 address this question. We have only to check that the null spaces of B and its adjoint B^* are empty. If so, B is then referred to as being *coercive*. For our sesquilinear operator, this is easily shown to be true, as follows.

Consider equation 4.6.1.14 and assume there exists a $\zeta \in H^1(\Omega)$ such that

$$B(\zeta, v) = 0, \forall v \in H^1(\Omega), \qquad (4.6.1.15)$$

i.e., ζ is in the null space of B. It is equivalent to state that 4.6.1.15 holds on a dense subset of $H^1(\Omega)$, so that $B(\zeta, \phi) = 0, \forall \phi \in C^\infty(\overline{\Omega})$. Integrating by parts, we have

$$k^2(\zeta, \phi) + \gamma(D_z^2\zeta, \phi) - \gamma(D_z\zeta)\phi|_0^1 = 0, \ \ \forall \phi \in C^\infty(\overline{\Omega}). \qquad (4.6.1.16)$$

Since this must be true even for those ϕ which vanish on the boundary, it must separately be that

$$k^2\zeta + \gamma D_z^2\zeta = 0 \ \text{ and } \ D_z\zeta|_{\partial\Omega} = 0. \qquad (4.6.1.17)$$

In this case, the ζ are the functions $cos(n\theta)$, and in general will be $C^\infty(\overline{\Omega})$ functions given smooth coefficients of the operators in the BVP. Substituting into 4.6.1.17 we obtain that $k^2 - \gamma n^2\pi^2 = 0$ if any ζ satisfying equation 4.6.1.15 exists.

The null space of the adjoint is investigated similarly. Let us suppose there is a $\zeta \in H^1(\Omega)$ such that

$$B(u, \zeta) = 0, \forall u \in H^1(\Omega), \tag{4.6.1.18}$$

i.e., ζ is in the null space of B^*. The same analysis as above reveals that $k^2 - \overline{\gamma} n^2 \pi^2 = 0$ if any ζ satisfying equation 4.6.1.18 exists.

It is evident therefore that if $k^2 \neq \gamma n^2 \pi^2$ and $k^2 \neq \overline{\gamma} n^2 \pi^2$ (for integer n) then $B(u, v)$ is coercive on $H^1(\Omega) \times H^1(\Omega)$, i.e., if $u \in H^1(\Omega)$ and $v \in H^1(\Omega)$.

The relationship between the statement $L(u_0) = f$ and $B(u_0, v) = L(v)$ is different in the case of Dirichlet boundary conditions. In that case, the solution u_0 is still unique, and equation 4.6.1.12 still holds. However, since $D_z u_0|_{\partial\Omega} \neq 0$, equation 4.6.1.14 does not follow unless v satisfies the Dirichlet boundary conditions as well. The Dirichlet conditions must also be satisfied by the solution space.

Observation: The solution of $B(u, v) = F(v), u \in H^1(\Omega), \forall v \in H^1(\Omega)$ contains no explicit reference to the Neumann boundary conditions. *The trial functions u and the test functions v need not satisfy the Neumann boundary conditions.* This can be is very convenient in applications.

We can therefore abandon $(L(u), \phi) = (f, \phi)$, $D_z u|_{\partial\Omega} = 0$, $\forall \phi \in C^\infty(\overline{\Omega})$ in favor of $B(u, v) = F(v), \forall v \in H^1(\Omega)$, with the proviso that $f \in H^{-1}(\Omega)$. The latter form is one of a number of useful *weak* forms of a BVP.

4.6.2 Error bounds in $H^1_{h,p} \subset H^1(\Omega)$

Theorem D.2 addresses the nature of the convergence of approximate solutions to the true solution. We know that the operator B is continuous and coercive. Theorem D.1 tells us that there is a unique solution $u_0 \in H^1(\Omega) | B(u_0, v) = F(v), \forall v \in H^1(\Omega)$. As is shown above, it is also true that, if $f \in H^{-1}(\Omega)$, the same u_0 satisfies the BVP $L(u_0) = f$, $D_z u_0|_{\partial\Omega} = 0$.

Theorem D.2 addresses the convergence of the approximation $\hat{u}_0 \in H^1_{h,p}$ to $u_0 \in H^1(\Omega)$. The error bound provided by this theorem depends on the behavior of the constant $\hat{C}_2(h,p)$ in condition ii), where

$$\inf_{\substack{u \in H^1_{h,p} \\ \|\!|u|\!\|_1 = 1}} \sup_{\substack{v \in H^1_{h,p} \\ \|\!|v|\!\|_1 \neq 0}} |B(u,v)| / \|\!|v|\!\|_1 = \hat{C}_2(h,p). \tag{4.6.2.19}$$

$\hat{C}_2(h,p)$ is determined as discussed in appendix E. That analysis generalizes the approach of Demkowicz & Oden (1994) for a similar problem. First, a convenient basis for the finite dimensional space $H^1_{h,p}$ ($N = \dim\left(H^1_{h,p}\right) = (p-1)/h + 1$) is chosen. Second, the method of Lagrange multipliers is used to find the extrema of $|B(u,v)| / \|\!|v|\!\|_1$ subject to the constraint that $\|\!|u|\!\|_1 = 1$.

Define

$$\{\lambda_n(h,p) \in \mathbb{R}, e_n \in H^1_{h,p}\} \,|\, (D_z u, D_z e_n) = \lambda_n(h,p)\,(u, e_n)\,, \forall u \in H^1_{h,p}. \tag{4.6.2.20}$$

The $\{\lambda_n(h,p), e_n\}, n = 1, \ldots, N$ are therefore the spectrum of a generalized eigenproblem on the FEM space. This eigenproblem is discussed thoroughly in the appendices. Appendix E reveals that

$$\hat{C}_2(h,p) = \min_{1 \le j \le N} \frac{|k^2 - (1 - i\epsilon)\lambda_j(h,p)|}{|k^2 + \lambda_j(h,p)|}. \tag{4.6.2.21}$$

$\hat{C}$ (see Theorem D.2) has the same (non-zero) value, satisfying condition iii) of that theorem. $\hat{C}_2$ has some non-zero value regardless of h or p, and its behavior is shown in figure 4.1. Theorem D.2 therefore bounds the error between $\hat{u}_0 | B(\hat{u}_0, v) = F(v), \forall v \in H^1_{h,p}$ and $u_0 | B(u_0, v) = F(v), \forall v \in H^1(\Omega)$ as

$$\boxed{\|\!|u_0 - \hat{u}_0|\!\|_1 \le \left(1 + \frac{C_1}{\hat{C}_2(h,p)}\right) \inf_{w \in H^1_{h,p}} \|\!|w - u_0|\!\|_1 . \qquad (4.6.2.22)}$$

Using Theorem 4.4.6,

$$\|\!|u_0 - \hat{u}_0|\!\|_1 \le C' h^{p-1} \left(1 + \frac{C_1}{\hat{C}_2(h,p)}\right) \|u_0\|_p . \tag{4.6.2.23}$$

Tighter bounds can be placed on the convergence in $H^1(\Omega)$. At the same time, the origin of the error estimate of theorem D.2 can be clearly shown. Following the discussion of Ihlenburg & Babuška (1995), consider the following. Let s be the best approximation of the solution u_0 in $H^1(\Omega)$: $(\!(u_0, v)\!)_1 = (\!(s, v)\!)_1$, $\forall v \in H^1_{h,p}$. Then $B(\hat{u}_0 - s, v) = B(\hat{u}_0 - u_0, v) + B(u_0 - s, v)$, $\forall v \in H^1_{h,p}$. However, $B(\hat{u}_0 - u_0, v) = B(\hat{u}_0, v) - B(u_0, v) = F(v) - F(v) = 0$, $\forall v \in H^1_{h,p}$. Therefore, $B(\hat{u}_0 - s, v) = B(u_0 - s, v)$, $\forall v \in H^1_{h,p}$. Since B is bounded, $|B(\hat{u}_0 - s, v)| \leq C_1 \,|\!|\!| u_0 - s |\!|\!|_1 \,|\!|\!| v |\!|\!|_1$. Upon closer inspection, B can be written $B(u, v) = k^2(u, v) - \gamma\,[(\!(u, v)\!)_1 - k^2(u, v)]$. Therefore, $B(u, v) = (1 + \gamma)k^2(u, v) - \gamma\,(\!(u, v)\!)_1$. By the definition of the best approximation, then $B(\hat{u}_0 - s, v) = B(u_0 - s, v) = (1 + \gamma)k^2(u_0 - s, v)$, $\forall v \in H^1_{h,p}$. Using Schwarz's inequality, $|B(\hat{u}_0 - s, v)| \leq |(1 + \gamma)| \,|\!|\!| u_0 - s |\!|\!|_0 \,|\!|\!| v |\!|\!|_0$. Applying *Nitsche's trick* (see below) to the operator $(\!(u, v)\!)_1$ and an associated BVP, $|\!|\!| u_0 - s |\!|\!|_0 \leq ch \,|\!|\!| u_0 - s |\!|\!|_1$. Since $|\!|\!| v |\!|\!|_0 \leq |\!|\!| v |\!|\!|_1$, it follows that

$$|B(\hat{u}_0 - s, v)| \leq Ch \,|\!|\!| u_0 - s |\!|\!|_1 \,|\!|\!| v |\!|\!|_1 \,, \quad \forall v \in H^1_{h,p}. \tag{4.6.2.24}$$

By the definition of the constant $\hat{C}_2(h, p)$, however,

$$\sup_{\substack{v \in H^1_{h,p} \\ |\!|\!| v |\!|\!|_1 \neq 0}} \frac{|B(u, v)|}{|\!|\!| u |\!|\!|_1 \,|\!|\!| v |\!|\!|_1} \geq \hat{C}_2(h, p), \quad \forall u, v \in H^1_{h,p}. \tag{4.6.2.25}$$

Certainly, then,

$$\sup_{\substack{v \in H^1_{h,p} \\ |\!|\!| v |\!|\!|_1 \neq 0}} \frac{|B(\hat{u}_0 - s, v)|}{|\!|\!| \hat{u}_0 - s |\!|\!|_1 \,|\!|\!| v |\!|\!|_1} \geq \hat{C}_2(h, p). \tag{4.6.2.26}$$

Substituting 4.6.2.24 yields $|\!|\!| \hat{u}_0 - s |\!|\!|_1 \leq \Big(Ch / \hat{C}_2(h, p)\Big) \,|\!|\!| u_0 - s |\!|\!|_1$. Finally, using the triangle inequality, $|\!|\!| \hat{u}_0 - u_0 |\!|\!|_1 \leq |\!|\!| \hat{u}_0 - s |\!|\!|_1 + |\!|\!| u_0 - s |\!|\!|_1$, so that

$$\boxed{|\!|\!| u_0 - \hat{u}_0 |\!|\!|_1 \leq \left(1 + \frac{Ch}{\hat{C}_2(h, p)}\right) \inf_{w \in H^1_{h,p}} |\!|\!| w - u_0 |\!|\!|_1 \,. \qquad (4.6.2.27)}$$

In other words, the FEM solution converges identically to the best approximation in $H^1(\Omega)$, given that $\hat{C}_2(h,p)$ approaches some constant independent of h as $h \to 0$. It does not appear than the same can be said in the $H^0(\Omega)$ norm. (In fact, such a result would contradict the numerical results of section 4.6.4).

Although $H^1(\Omega)$ is the most reasonable norm in which to measure the error, it is interesting to bound the error in $H^0(\Omega)$ as well. The following approach is referred to as *Nitsche's trick* (Strang & Fix, 1973). Consider the solution of the BVP $L^*u = u_0 - \hat{u}_0$, $D_z u|_{\partial\Omega} = 0$, where L^* is the adjoint of L. Theorem 4.4.4 applies and

$$\|u\|_2 \leq C\,\|u_0 - \hat{u}_0\|_0\,. \tag{4.6.2.28}$$

The corresponding problem in $H^1(\Omega)$ is

$$B(v,u) = (u_0 - \hat{u}_0, v), \forall v \in H^1(\Omega). \tag{4.6.2.29}$$

Since $u_0 - \hat{u}_0 \in H^1(\Omega)$, let $v = u_0 - \hat{u}_0$, therefore

$$B(u_0 - \hat{u}_0, u) = \|u_0 - \hat{u}_0\|_0^2\,. \tag{4.6.2.30}$$

Let $\hat{u} \in H^1_{h,p}$, then $B(u_0 - \hat{u}_0, \hat{u}) = B(u_0, \hat{u}) - B(\hat{u}_0, \hat{u}) = (f, \hat{u}) - (f, \hat{u}) = 0$, so that

$$B(u_0 - \hat{u}_0, u - \hat{u}) = \|u_0 - \hat{u}_0\|_0^2\,. \tag{4.6.2.31}$$

Since B satisfies the conditions of Theorem D.1,

$$C\,\|u_0 - \hat{u}_0\|_1\,\|u - \hat{u}\|_1 \geq \|u_0 - \hat{u}_0\|_0^2\,. \tag{4.6.2.32}$$

Choosing $\hat{u}$ to be the best approximation of u in $H^1_{h,p}$, and using Theorem 4.4.6, $\|u - \hat{u}\|_1 \leq Ch\,\|u\|_2$. Therefore, by equation 4.6.2.28, $\|u - \hat{u}\|_1 \leq Ch\,\|u_0 - \hat{u}_0\|_0$. Substituting into equation 4.6.2.32,

$$\|u_0 - \hat{u}_0\|_0 \leq Ch\,\|u_0 - \hat{u}_0\|_1\,. \tag{4.6.2.33}$$

or , using Theorem 4.4.6,

$$\|u_0 - \hat{u}_0\|_0 \leq C'h^p \left(1 + \frac{C_1}{\hat{C}_2(h,p)}\right) \|u_0\|_p \,, \tag{4.6.2.34}$$

which should be compared to equation 4.6.2.23.

The asymptotic convergence of the FEM eigenvalues $\lambda_j(h,p)$ is well known. For $H^1_{h,p}$, applying Fried (1979) equation 6.114,

$$\lambda_j(h,p) - \lambda_j \leq Ch^{2(p-1)}\lambda_j^p, \;\; j \ll N, \tag{4.6.2.35}$$

where $N = dim[H^1_{h,p}] = (p-1)/h + 1$. Clearly, for sufficiently small h, or equivalently, sufficiently small j or k^2, $\lambda_j(h,p) \to \lambda_j$ and $\hat{C}_2(h,p) \to C_2$, so that

$$\|u_0 - \hat{u}_0\|_0 \leq C(p)h^p, \;\; h \to 0, \;\; k^2 \text{ fixed,} \tag{4.6.2.36}$$

$$\|u_0 - \hat{u}_0\|_1 \leq C(p)h^{p-1}, \;\; h \to 0, \;\; k^2 \text{ fixed.} \tag{4.6.2.37}$$

The solution converges in both norms at the highest possible rate for sufficiently small h, so that the method is *asymptotically optimal.* Eigenvalue errors computed from the results of appendix C confirm the asymptotic results above (see figure 4.16), and can be used to compute values of $\hat{C}_2$ (see figure 4.1) Those parameter values where $\hat{C}_2$ becomes independent of h are particularly of interest, since convergence is then asymptotically optimal by equation 4.6.2.23.

It is evident from the definition of $\hat{C}_2(h,p)$ (equation 4.6.2.19) that this constant is smallest in the limit $h \to 0$ when $k^2 = k_n^2$ is an exact eigenfrequency. In that case, it is also evident that the limiting value of $\hat{C}_2(h,p) \propto \epsilon$. Therefore the difference between the FEM solution and the best approximation in $H^1(\Omega)$ is (by equation 4.6.2.27) bounded in norm by some constant multiple of ϵ^{-1}. This is examined further in section 4.6.3.

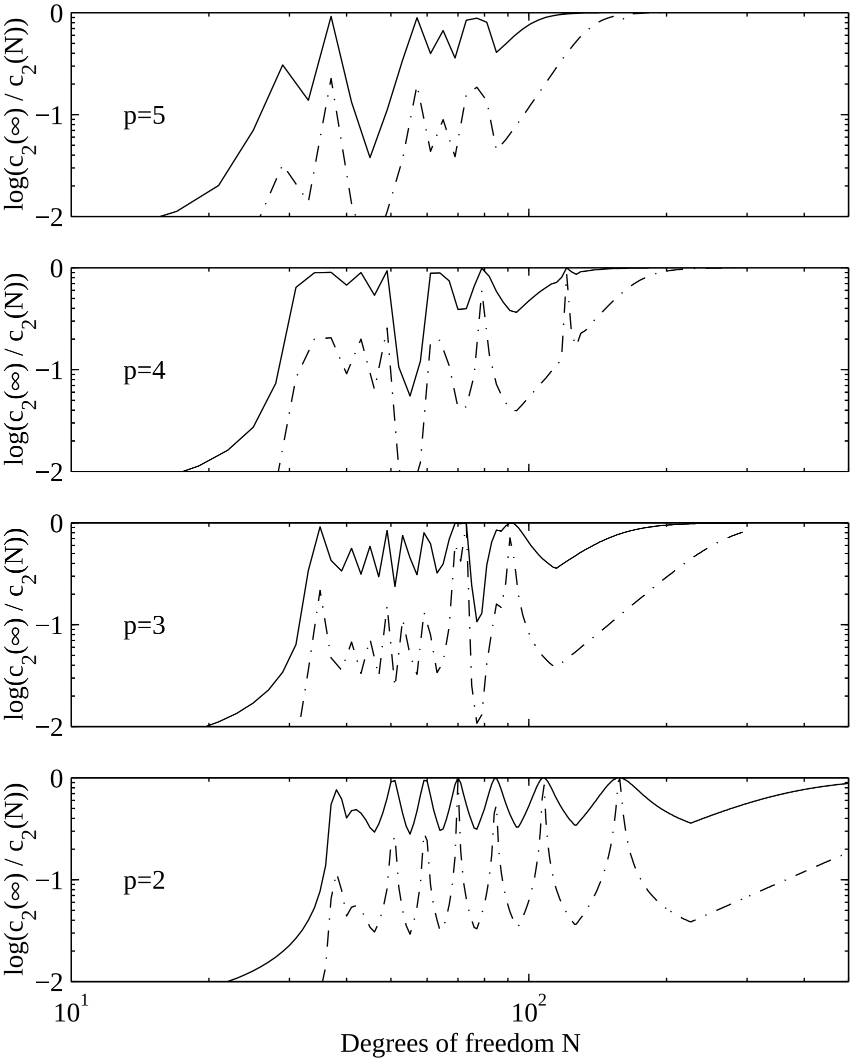


Figure 4.1: Convergence of $1/\hat{C}_2(h,p)$ to its limiting value as $h \to 0$. Since the numerical eigenvalues converge to their exact values for small h, equation 4.6.2.21 indicates that $1/\hat{C}_2(h,p) \to 2/\epsilon$ as $h \to 0$ when evaluated at $k^2 = k_j^2$. For all other values of k^2, $1/\hat{C}_2(h,p)$ converges to a smaller value. Results are shown for several element orders. The solid curves correspond to $\epsilon = .01$, the dashed to $\epsilon = .001$. Once $1/\hat{C}_2(\infty,p)$ is attained, the FEM method converges optimally, or at the same rate as the best approximation. $1/\hat{C}_2(h,p)$ rises initially until near the resolution limit of 2 mean grid intervals per wavelength ($N = 40$ in this case), then peaks sharply as approximate FEM eigenvalues $\hat{k}_m^2$ approach k_j^2. The larger dips are due to the jump in approximate eigenvalues near the lower end of the highest mode branch. Evaluated at a k^2 away from any k_j^2, $1/\hat{C}_2(h,p)$ approximates $1/\hat{C}_2(\infty,p)$ at a lower N.

4.6.3 Eigensolution analysis

Comparing the solution of the BVP in terms of exact and approximate eigenfunctions allows the extrapolation of errors evaluated at a fixed h to any h. This comparison process, however, requires that one assume $\epsilon \ll 1$.

Let $\{\hat{k}_m^2, \hat{e}_m\} | \hat{k}_m^2 (e_m, v) = -(D_z e_m, D_z v)$, $\forall v \in H^1(\Omega)$, and the e_m are normalized so that $(e_m, e_j) = \delta_{mj}$. For $\Omega = (0,1)$, it is clear that $\{k_m^2, e_m\} = \{m^2\pi^2, c_m cos(m\pi z)\}, m = 0, \ldots, \infty$, where $c_m = 1, m = 0$, $c_m = \sqrt{2}, m \neq 0$. Let E be the space spanned by the e_m, $m = 0, \ldots, \infty$. Since E is dense in $H^1(\Omega)$ (see Aubin (1972), p. 37), the solution of the BVP $B(u_0, v) = F(v), \forall v \in H^1(\Omega), u_0 \in H^1(\Omega)$ can be replaced by

$$B\left(\sum_{j=0}^{\infty} \alpha_j e_j, e_m\right) = F(e_m), \quad m = 0, \ldots, \infty, \quad \alpha_m \in \mathbb{C}. \tag{4.6.3.1}$$

Using the definition of B in equation 4.6.1.5,

$$\alpha_m = \frac{(f, e_m)}{k^2 - \gamma k_m^2}, \tag{4.6.3.2}$$

so that the mode solution is

$$u = \sum_{j=0}^{\infty} \frac{(f, e_j)\, e_j}{\left(k^2 - \gamma k_j^2\right)}, \tag{4.6.3.3}$$

Likewise, the solution of the generalized eigenvalue problem in the FEM space $H_{h,p}^1$ yields the FEM approximation

$$\hat{u} = \sum_{j=0}^{N-1} \frac{(f, \hat{e}_j)\, \hat{e}_j}{\left(k^2 - \gamma \hat{k}_j^2\right)}, \tag{4.6.3.4}$$

where $\{\hat{k}_m^2, \hat{e}_m\} | \hat{k}_m^2 (\hat{e}_m, v) = -(D_z \hat{e}_m, D_z v)$, $\forall v \in H_{h,p}^1$, and the $\hat{e}_m$ are normalized so that $(\hat{e}_m, \hat{e}_j) = \delta_{mj}$. The approximation error can be written as $E_{0,1}^2 = \|\hat{u} - u\|_{0,1}^2$. Let $s_0 | (s_0, \hat{e}_m) = (u, \hat{e}_m), m = 1, \ldots, \infty$ and $s_1 | \left((s_1, \hat{e}_m)\right)_1 = \left((u, \hat{e}_m)\right)_1, m = 1, \ldots, \infty$ be the best approximations of the solution u in the subspace $H_{h,p}^1$ with respect to the norms on $H^0(\Omega)$ and $H^1(\Omega)$. Therefore,

$$E_0^2 = \|(\hat{u} - s_0) - (u - s_0)\|_0^2 = \|(\hat{u} - s_0)\|_0^2 + \|(u - s_0)\|_0^2, \tag{4.6.3.5}$$

and

$$E_1^2 = \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert (\hat{u} - s_1) - (u - s_1) \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_1^2 = \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert (\hat{u} - s_1) \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_1^2 + \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert (u - s_1) \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_1^2 , \tag{4.6.3.6}$$

where we have used orthogonality to isolate contributions to the norms. Substituting the mode sums in equations 4.6.3.3 and 4.6.3.4 into equations 4.6.3.5 and 4.6.3.6 and doing a bit of algebra results in relatively simple expressions for all of the required errors. The errors in the FEM space are

$$\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert (\hat{u} - s_0) \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_0^2 = k^2 \sum_{j=0}^{N-1} \left| \sum_{m=0}^{\infty} \frac{(e_m, \hat{e}_j)\,(f, e_m)\gamma \left(\hat{k}_j^2 - k_m^2\right)}{(k^2 - \gamma k_m^2)\left(k^2 - \gamma \hat{k}_j^2\right)} \right|^2 \tag{4.6.3.7}$$

and

$$\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert (\hat{u} - s_1) \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_1^2 = \sum_{j=0}^{N-1} \left(k^2 + \hat{k}_j^2\right) \left| \sum_{m=0}^{\infty} \frac{(e_m, \hat{e}_j)\,(1+\gamma)\,k^2 \left(\hat{k}_j^2 - k_m^2\right)(f, e_m)}{\left(k^2 + \hat{k}_j^2\right)(k^2 - \gamma k_m^2)\left(k^2 - \gamma \hat{k}_j^2\right)} \right|^2 \tag{4.6.3.8}$$

The best approximation errors are

$$\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert (u - s_0) \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_0^2 = \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert u \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_0^2 - k^2 \sum_{j=0}^{N-1} \left| \sum_{m=1}^{\infty} \frac{(e_m, \hat{e}_j)\,(f, e_m)}{k^2 - \gamma k_m^2} \right|^2 \tag{4.6.3.9}$$

and

$$\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert (u - s_1) \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_1^2 = \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert u \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_1^2 - \sum_{j=0}^{N-1} (k^2 + \hat{k}_j^2) \left| \sum_{m=1}^{\infty} \frac{(k^2 + k_m^2)\,(e_m, \hat{e}_j)\,(f, e_m)}{(k^2 - \gamma k_m^2)(k^2 + \hat{k}_j^2)} \right|^2 \tag{4.6.3.10}$$

The solution norms are

$$\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert u \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_0^2 = k^2 \sum_{m=0}^{\infty} \left| \frac{(f, e_m)}{k^2 - \gamma k_m^2} \right|^2 , \tag{4.6.3.11}$$

and

$$\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert u \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_1^2 = \sum_{m=0}^{\infty} (k^2 + k_m^2) \left| \frac{(f, e_m)}{k^2 - \gamma k_m^2} \right|^2 . \tag{4.6.3.12}$$

In deriving the above terms, it is assumed that both the true and FEM eigenfunctions are normalized so that $(e_m, e_m) = (\hat{e}_m, \hat{e}_m) = 1$. The fact that

$(D_z\hat{e}_j, D_z e_m) = k_m^2\,(\hat{e}_j, e_m)$ was used in the derivation of the errors in the $H^1(\Omega)$ norm.

Equations 4.6.3.7-4.6.3.12 are fairly intractable, so two approximations are made to obtain useful information. First, assume the error expressions above are to be evaluated at an exact eigenfunction so that $k^2 = k_n^2$. Since the above series converge due to the result of section 4.6.2 and choice of f, the sums over m can be restricted to the n^{th} term by reducing ϵ to a sufficiently small value. Second, it is clear that equations 4.6.3.7 and 4.6.3.8, when normalized by equations 4.6.3.11 and 4.6.3.12, peak when $k^2 = \hat{k}_j^2$, not when $k^2 = k_n^2$. In order that these equations do not under predict the error, require that $1/(k_n^2 - \gamma\hat{k}_n^2) \simeq 1/(\hat{k}_n^2 - \gamma\hat{k}_n^2)$. This is equivalent to requiring that $\delta k_n^2/(\epsilon k_n^2) \ll 1$.

Hence, the results of appendix C can be used to evaluate equations 4.6.3.7-4.6.3.12 numerically, and three sources of error are revealed in equations 4.6.3.7 - 4.6.3.10 : best approximation error, eigenfunction error, and eigenvalue (or dispersion) error. The best approximation error behaves as section 4.6.2 predicts, and is dominated by those terms in equations 4.6.3.9 and 4.6.3.10 where $j = n$. The eigenfunction error is due to the contribution of terms in equations 4.6.3.7 and 4.6.3.8 when $j \neq n$. The eigenfunction error is bounded by the best approximation error for all p and k^2. The best approximation error and the eigenfunction errors are very insensitive to ϵ. Dispersion error is due entirely to errors in the eigenvalues. Evaluating equations 4.6.3.7 and 4.6.3.8 with $k^2 = k_n^2$, $j = n$, requiring $\delta k_m^2 = \hat{k}_m^2 - k_m^2 \ll \epsilon k_m^2$ and allowing $\epsilon \to 0$ indicates that the dispersion error is

$$E_d \simeq \delta k_n^2/(\epsilon k_n^2), E_d \ll 1, \tag{4.6.3.13}$$

regardless of the norm used to measure the error. Finally, the error peaks at the eigenvalues, so that by evaluating the error at $k^2 = k_n^2$, an upper bound for the error in the vicinity of k_n^2 is found.

The convergence of the eigenvalues and eigenvectors is such that, for sufficiently small j and h, Strang & Fix (1973)

$$\begin{aligned} \|e_j - \hat{e}_j\|_0^2 &\simeq ch^{2p}k_j^{2p}, \\ \hat{k}_j^2 - k_j^2 &\simeq c'h^{2(p-1)}k_j^{2p}. \end{aligned} \qquad (4.6.3.14)$$

These asymptotic estimates are confirmed in the analysis of appendix C. Substituting equations 4.6.3.14 into equations 4.6.3.7 - 4.6.3.12 and assuming $\delta k_n^2/k_n^2 \ll \epsilon \ll 1$, the total relative error resulting from dispersion and best approximation error yields

$$\frac{\|\hat{u} - u\|_0^2}{\|u\|_0^2} \lesssim c_1 h^{2p} k^{2p} + c_3 \epsilon^{-2} h^{4(p-1)} k^{4(p-1)}, \qquad (4.6.3.15)$$

$$\frac{\|\hat{u} - u\|_1^2}{\|u\|_1^2} \lesssim c_2 h^{2(p-1)} k^{2(p-1)} + c_3 \epsilon^{-2} h^{4(p-1)} k^{4(p-1)}. \qquad (4.6.3.16)$$

where approximate equality holds when $k^2 = k_n^2$. The results of appendix C can be used to make specific predictions of these relative errors (see figures 4.2 and 4.3). With one exception, for all p and norms the best approximation error dominates for sufficiently small h or k. The exception is the measurement of error in the $H^0(\Omega)$ norm with linear elements, where dispersion error is dominant for all h and k.

4.6.4 Numerical Tests

Numerical simulations of the problem presented in section 4.6.2 allow a number of useful statements to be made about the computational effort required to compute solutions at a given relative accuracy. The numerical solutions presented in figures 4.2 -4.5 are based on the FEM solution of equation 4.7.1.7 for w, given constant μ and ρ. More explicitly,

$$k^2(u, v) - \gamma(D_z u, D_z v) = -(k^2/\gamma)(H, v), \;\; \forall v \in H_{h,p}^1 \qquad (4.6.4.1)$$

where $\gamma = 1 - i\epsilon$, and $\epsilon \ll 1$. H, the Heaviside step function, is shifted so that the source is not at the origin but is at $z = .75$ on the 1-D medium $(0, 1)$. The solutions of equation 4.6.4.1 are compared to very accurate reference solutions u computed with the same method, but with much higher element order p_{ref}. The best approximations $\hat{w}_1 | \left(\!\left(\hat{w}_1, v\right)\!\right)_1 = \left(\!\left(u, v\right)\!\right)_1, \forall v \in H^1_{h,p}$, and $\hat{w}_0 | (\hat{w}_0, v) = (u, v), \forall v \in H^1_{h,p}$ are computed from the reference solution as well.

The measure of the mesh refinement for the presentation of most numerical results is mean grid intervals per wavelength, or m. For a solution dominated by a cosine of wavelength k, $m = C(p)/(kh)$, where the constant $C(p)$ depends on p in some manner determined by the mesh construction. The predictions of the error, both asymptotic (section 4.6.2) and for small ϵ (section 4.6.3) have simple dependence on m. From equations 4.6.3.15 and 4.6.3.16,

$$\left\|\hat{u} - u\right\|_0^2 / \left\|u\right\|_0^2 \lesssim c_1 m^{-2p} + c_3 \epsilon^{-2} m^{-4(p-1)}, \tag{4.6.4.2}$$

$$\left\|\hat{u} - u\right\|_1^2 / \left\|u\right\|_1^2 \lesssim c_2 m^{-2(p-1)} + c_3 \epsilon^{-2} m^{-4(p-1)}, \tag{4.6.4.3}$$

where the terms dependent on ϵ are dispersion error. These error estimates are borne out by the computations.

The effect of the use of different p on computational efficiency is presented in figures 4.4 and 4.5. The reduction in memory requirements from using higher p is quite impressive, even at rather high relative error. The FLOP requirements are not as strongly effected, and it is probable that most of the gain in using higher p (measured in FLOP reduction) is seen in the transition from linear to quadratic and cubic elements. For very accurate results, higher order elements are essential.

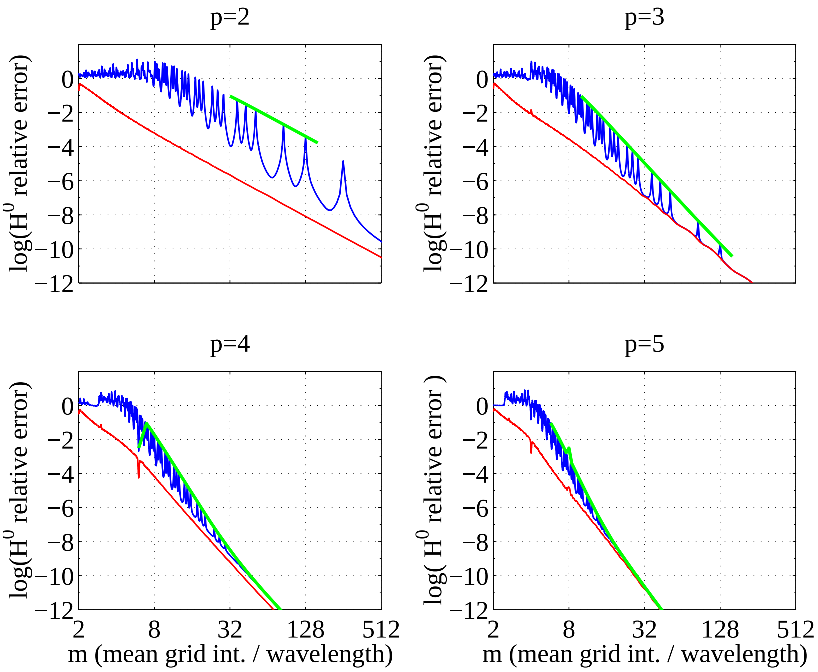


Figure 4.2: Relative errors of the FEM solution $\hat{u}$ of equation 4.6.4.1 using the spaces $H^1_{h,2}$ - $H^1_{h,5}$, i.e. linear - quartic elements. The blue lines indicate the solution error $\|\hat{u} - u\|_0^2 \,/\, \|u\|_0^2$ and the red lines the best approximation error $\|\hat{w}_0 - u\|_0^2 \,/\, \|u\|_0^2$, both computed from the FEM using a reference solution u with a much higher p. The green lines are the error predictions of the eigensolution analysis (equations 4.6.3.7 - 4.6.3.12 for $\epsilon \ll 1$), interpolated between eigenfrequencies. While both error plots are computed with approximately 128 DOF, the eigensolution error envelope applies for any h, as is indicated by the m-axis. The modulation of the mode peaks is due to location of the source; every fourth mode is not excited. In this case, $\epsilon = .01$. Note that the error peaks at numerical eigenfrequencies, and departs from the best approximation due to dispersion error, as indicated in equation 4.6.3.15. The convergence of the linear elements as measured in $H^0(\Omega)$ is particularly poor; dispersion error dominates for all k. The eigensolution error is only displayed at relative error values below which it is expected to bound the error: $E_d \ll 1$ (see equation 4.6.3.13).

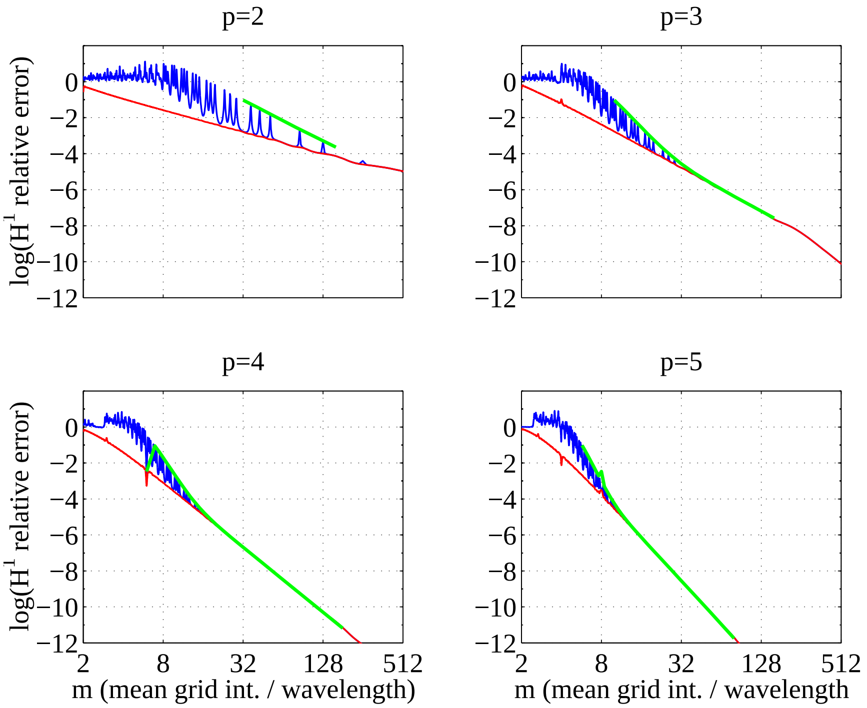


Figure 4.3: Relative errors of the FEM solution $\hat{u}$ of equation 4.6.4.1 using the spaces $H^1_{h,2}$ - $H^1_{h,5}$, i.e. linear - quartic elements. The blue lines indicate $|||\hat{u} - u|||_1^2 \,/\, |||u|||_1^2$ and the red lines the best approximation error $|||\hat{w}_1 - u|||_1^2 \,/\, |||u|||_1^2$, both computed from the FEM using a reference solution with a much higher p. The green lines are the error predictions of the eigensolution analysis (equations 4.6.3.7 - 4.6.3.12 for $\epsilon \ll 1$), interpolated between eigenfrequencies. While both error plots are computed with approximately 128 DOF, the eigensolution error envelope applies for any h, as is indicated by the m-axis. As in figure 4.2, $\epsilon = .01$. Note that the error peaks at numerical eigenfrequencies, and departs from the best approximation due to dispersion error, as indicated in equation 4.6.3.16. This figure differs most notably from figure 4.2 in that all FEM solutions converge to the best approximation. The eigensolution error is only displayed at relative error values below which it is expected to bound the error: $E_d \ll 1$ (see equation 4.6.3.13).

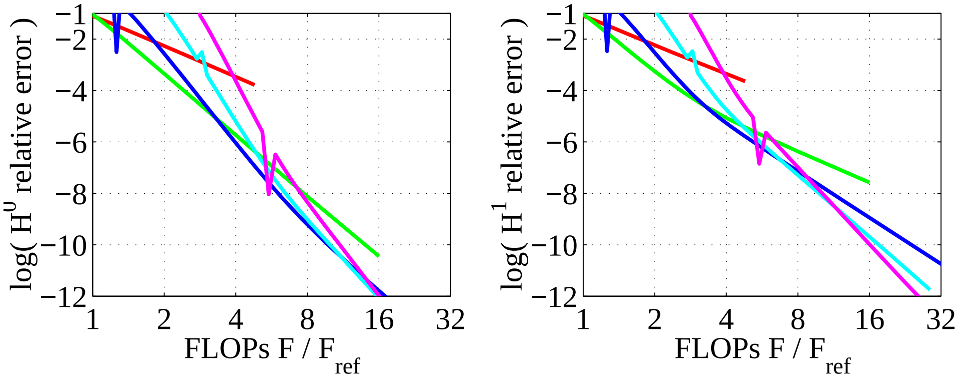


Figure 4.4: FLOP requirements for the FEM solution at a target solution frequency k. The spaces $H^1_{h,2}$ (red), $H^1_{h,3}$ (green), $H^1_{h,3}$ (blue), $H^1_{h,5}$ (cyan) and $H^1_{h,6}$ (magenta) are used. The error envelopes (i.e., errors at the eigenfrequencies) are computed from the eigensolution analysis of section 4.6.3 in order to extrapolate to any h. This figure is simply a transformation in the x-axis of figures 4.2 and 4.3 using the FLOP requirements of algorithm NROB (Carey & Oden, 1984) as an estimate of the FLOP requirements of algorithm ZGBSVX (Anderson, et al., 1992): $F \approx N(p-1)(2(p-1)+1)$, where $N = (p-1)/h + 1$. This algorithm is not the most efficient FLOPwise for higher p, since a full band is operated upon. Selecting a different target frequency only changes the absolute FLOP requirements (i.e., F_{ref}). The higher order spaces yield better performance in terms of FLOPS only if the required error is sufficiently small. However, $p = 3$ and $p = 4$ elements yield better performance over linear elements at fairly high levels of relative error (e.g., near .1). In these calculations, $\epsilon = .01$.

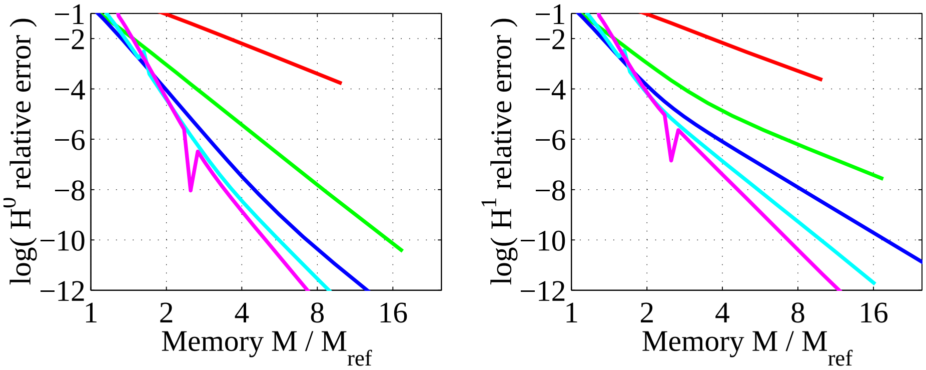


Figure 4.5: Memory requirements for the FEM solution at a target solution frequency k. The spaces $H^1_{h,2}$ (red), $H^1_{h,3}$ (green), $H^1_{h,3}$ (blue), $H^1_{h,5}$ (cyan) and $H^1_{h,6}$ (magenta) are used. The error envelopes (i.e., errors at the eigenfrequencies) are computed from the eigensolution analysis of section 4.6.3 in order to extrapolate to any h. This figure is simply a transformation in the x-axis of figures 4.2 and 4.3 using the memory requirements of algorithm ZGBSVX (Anderson, et al., 1992): $M \approx N(3(p-1)+1)$, where $N = (p-1)/h + 1$. This algorithm is not the most efficient memorywise for higher p, since a full band is operated upon. Selecting a different target frequency only changes the absolute memory requirements (i.e., M_{ref}). The higher order spaces yield better performance in terms of memory requirements, even at high levels of relative error.In these calculations, $\epsilon = .01$.

4.7 Modification of singular data

4.7.1 Modification of $H^{-2}(\Omega)$ data in 1-D

It is possible to obtain asymptotically optimal convergence given some data in $H^{-2}(\Omega)$ for the 1-D problem. Although the discontinuous FEM addresses the solution given data in $H^{-2}(\Omega)$ in a general fashion, we include a brief discussion of the removal of discontinuous sources. Other authors (Cummins, et al., 1994b) have presented this approach without adequate consideration.

Consider the distributional statement of the second order differential equation with variable but smooth coefficient μ

$$(L(u), \phi) = (\rho\omega^2 u, \phi) + \gamma(D_z(\mu D_z u), \phi), \quad \forall\phi \in C^\infty(\overline{\Omega}), \tag{4.7.1.1}$$

where $\mu \in C^\infty(\overline{\Omega})$. Inserting a Heaviside step function H into this equation and utilizing the rule for differentiation of distributions in equation 4.4.2.4,

$$(L(H), \phi) = (\rho\omega^2 H, \phi) - \gamma\mu(0)\phi'(0), \forall\phi \in C^\infty(\overline{\Omega}), \tag{4.7.1.2}$$

where the definition $(\delta, \phi) = \phi(0)$ was used. Now consider the problem

$$(L(u), \phi) = (\delta', \phi) = -\phi'(0), \forall\phi \in C^\infty(\overline{\Omega}). \tag{4.7.1.3}$$

Clearly, $\delta' \in H^{-2}(\Omega)$, and $u \in H^0(\Omega)$. However, let

$$(L(w + \nu H), \phi) = -\phi'(0), \forall\phi \in C^\infty(\overline{\Omega}). \tag{4.7.1.4}$$

so that

$$(L(w), \phi) = -\phi'(0) - \nu(L(H), \phi), \forall\phi \in C^\infty(\overline{\Omega}). \tag{4.7.1.5}$$

Substituting from equation 4.7.1.2, get

$$(L(w), \phi) = -\phi'(0) + \nu\gamma\mu(0)\phi'(0) - \nu(\rho\omega^2 H, \phi), \forall\phi \in C^\infty(\overline{\Omega}). \tag{4.7.1.6}$$

Now set $\nu = 1/(\gamma\mu(0))$ to get

$$(L(w), \phi) = -(\gamma\mu(0))^{-1}(\rho\omega^2 H, \phi), \forall\phi \in C^\infty(\overline{\Omega}). \tag{4.7.1.7}$$

Therefore, given Neumann boundary conditions on w, Theorem 4.4.4 applies. Since $H \in H^0(\Omega)$, $w \in H^2(\Omega)$, and the convergence results of section 4.6.2 apply. Asymptotically optimal convergence will then follow if an element boundary is placed at the source to allow for the discontinuity in the second derivative of w. After solving for w, νH can be added to recover the complete solution.

4.7.2 Modification of $H^{-3}(\Omega)$ data in higher dimensions

In 2-D and 3-D, the practicality of removing the source singularity is not clear. It is certainly possible to solve some statics problems (i.e., with $\omega = 0$ in equation 4.5.0.1) analytically for sources of interest in seismology . Examples are given in Kumari, et al. (1992). It may be that these statics solutions can be added to a continuous solution, as in the 1-D case just discussed, effectively removing the source. However, the solutions to these statics problems are not as simple as in the 1-D case and it is not clear that the PDE resulting from the substitution of the statics solution results in a manageable problem. In any case, we briefly examine how, in principle, one would go about removing the source singularity in higher dimensions.

Let us assume that we wish to solve equation 4.5.0.1 with the boundary condition of equation 4.5.0.2. Let us also suppose that for some tensor of elastic constants $\overline{c}_{ijkl}$ and for some set of boundary conditions, we can analytically solve the problem

$$\partial_j \overline{c}_{ijkl} w_{k,l} = f_i, \tag{4.7.2.8}$$

where it is likely that $f_i \in H^{-3}$. Such a solution exists for spatially constant (and real) $\overline{c}_{ijkl}$ in an infinite domain (Kumari, et al., 1992). We require that the true medium can be described by the the tensor

$$c_{ijkl} = \overline{c}_{ijkl} + \tilde{c}_{ijkl}, \tag{4.7.2.9}$$

where $\tilde{c}_{ijkl}$ vanishes in some region $\Gamma \subset \Omega$ near the source. Equation 4.5.0.1 becomes

$$\partial_j c_{ijkl} s_{k,l} + \rho\omega^2 s_i = f_i, \tag{4.7.2.10}$$

where $s_i = w_i + u_i$, and we mean this equation to hold in the sense of distributions. The boundary conditions are

$$n_j c_{ijkl} s_{k,l}|_{\partial\Omega} = 0. \tag{4.7.2.11}$$

Inserting equations 4.7.2.8 and 4.7.2.9 into equation 4.7.2.10, we obtain

$$\partial_j c_{ijkl} u_{k,l} + \rho\omega^2 u_i = -\partial_j \tilde{c}_{ijkl} w_{k,l} - \rho\omega^2 w_i, \tag{4.7.2.12}$$

and the boundary conditions become inhomogeneous:

$$n_j c_{ijkl} u_{k,l}|_{\partial\Omega} = -n_j c_{ijkl} w_{k,l}|_{\partial\Omega}. \tag{4.7.2.13}$$

Since $w_i \in C^\infty(\overline{\Omega})$ away from the source (on $\Omega \backslash \Gamma$), if we additionally assume that the $\tilde{c}_{ijkl} \in C^\infty(\overline{\Omega})$ then the first term on the right hand side of equation 4.7.2.12 is smooth. If $f_i \in H^{-3}$, elliptic regularity (for systems of equations - see section 4.9 below) and equation 4.7.2.8 imply that $w_i \in H^{-1}$. Therefore equation 4.7.2.12 indicates that $u_i \in H^1$, so that we can utilize the standard FEM. The sesquilinear form for this inhomogeneous Neumann boundary condition problem is

$$\rho\omega^2 (v_i, u_i) - (v_{i,j}, c_{ijkl} u_{k,l}) + \langle v_i|_{\partial\Omega}, g_i \rangle = (v_i, h_i), \tag{4.7.2.14}$$

where

$$g_i = -n_j c_{ijkl} w_{k,l}|_{\partial\Omega}, \tag{4.7.2.15}$$

and the less singular data is now

$$h_i = -\partial_j \tilde{c}_{ijkl} w_{k,l} - \rho\omega^2 w_i. \tag{4.7.2.16}$$

While this looks attractive in principle, the difficulty in evaluating (v_i, h_i) and $\langle v_i|_{\partial\Omega}, g_i \rangle$ may be considerable.

For $f \in H^{-2}$, the following section introduces a variational method which can converge to the solution without such manipulations of the source.

4.8 The Discontinuous FEM: Solutions for $f \in H^{-2}(\Omega)$

The singularity of sources present in seismological problems compels us to look for solutions with less regularity than solutions obtained using the standard FEM, although seismological sources are even more singular than the method to be described allows. The problem of more singular data (than $f \in H^{-1}(\Omega)$) is discussed in Douglas & Dupont (1977), Babuška & Aziz (1972) and Rachford & Wheeler (1974). The extension of the standard FEM to handle less continuous sources is quite straightforward in the context of the theory presented in section 4.4. The error analysis is somewhat more difficult, but we will present some fairly convincing evidence that the discontinuous FEM converges very much like the standard FEM.

4.8.1 The Sesquilinear form $C(u, v)$

We wish to solve the 1-D Neumann elliptic BVP in equations 4.6.0.1 - 4.6.0.3,

$$L_k u = f, \quad D_z u|_{\partial\Omega} = 0, \tag{4.8.1.1}$$

when $f \in H^{-2}(\Omega)$. Our method of section 4.6.1 is no longer applicable, since the test space must now be the dual space of f or $H^2(\Omega)$, and the solution space (by elliptic regularity) must be $\tilde{H}^0(\Omega)$. Consider the sesquilinear form

$$C(u, v) = k^2 (u, v) + \gamma \left(u, D_z^2 v\right) - \gamma \overline{u} D_z v|_0^1, \tag{4.8.1.2}$$

and, again the linear functional

$$F(v) = (f, v). \tag{4.8.1.3}$$

As before, $\gamma = 1 - i\epsilon$. Define the inner products

$$((u, v))_0 \quad = \quad k^2 (u, v), \tag{4.8.1.4}$$

$$((u, v))_2 \quad = \quad k^2 (u, v) + (D_z u, D_z v) + \left(D_z^2 u, D_z^2 v\right) \tag{4.8.1.5}$$

and the associated norms $|||\bullet|||_0$ and $|||\bullet|||_2$. These norms are equivalent to the usual Sobolev norms $\|\bullet\|_0$ and $\|\bullet\|_2$, respectively.

Equation 4.4.4.12 indicates that $(u, v) \leq C_a \|u\|_0 \|v\|_0$ and $(u, D_z^2 v) \leq C_b \|u\|_0 \|D_z^2 v\|_0$, Since, by the definition of the Sobolev spaces, $\|D_z^2 v\|_0 \leq \|v\|_2$, and $\|v\|_0 \leq \|v\|_2$, the first two terms of $C(u, v)$ are bounded for $u \in H^0(\Omega)$ and $v \in H^2(\Omega)$. In order that $uD_z v|_0^1$ make sense, additional constraints must be placed on the trace of u. In 1-D, equation 4.4.6.28 indicates that $|\overline{u} D_z v|_0^1| \leq \langle\langle u|_{\partial\Omega} \rangle\rangle_0 \langle\langle D_z v|_{\partial\Omega} \rangle\rangle_0$. Since $v \in H^2(\Omega)$, equation 4.4.6.26 indicates that $\langle\langle D_z v|_{\partial\Omega} \rangle\rangle_0 \leq \|v\|_2$. It is not true that $\langle\langle u|_{\partial\Omega} \rangle\rangle_0 \leq \|u\|_0$, so we must constrain the trace of u such that $u|_{\partial\Omega} \in H^0(\partial\Omega)$. Summarily, $C(u, v) \leq C(|||u|||_0^2 + \langle\langle u|_{\partial\Omega} \rangle\rangle_0^2)^{1/2} |||v|||_2$, so that $C(u, v)$ is continuous on $(H^0(\Omega) \times H^0(\partial\Omega)) \times H^2(\Omega)$. As well, the linear functional $F(v)$ is bounded for $f \in H^{-2}(\Omega)$, by equation 4.4.4.12. Therefore, we can extend

$$C(u, v) = F(v), \forall v \in C^\infty(\overline{\Omega}), \tag{4.8.1.6}$$

to

$$C(u, v) = F(v), \forall v \in H^2(\Omega), \tag{4.8.1.7}$$

provided that $f \in H^{-2}(\Omega)$. To what BVP does this correspond?

Proceeding as in section 4.6.1, let

$$U = \{u_0 \in H^0(\Omega), u_1 = D_z u_0|_{\partial\Omega} \in H^0(\partial\Omega)\}, \tag{4.8.1.8}$$

be the unique solution of the Neumann BVP above: $L_k(u_0) = f$, $D_z u_0|_{\partial\Omega} = 0$. The existence of this solution is assured by elliptic regularity. In higher dimensions, a solution of the same form is assured if f is smooth in some neighborhood of the boundary. Substituting into the equation, $C(U, v) = F(v), \forall v \in C^\infty(\overline{\Omega})$, and integrating by parts twice yields

$$k^2 (u_0, v) - \gamma \overline{u}_1 v|_0^1 + \gamma \left(D_z^2 u_0, v\right) = F(v), \forall v \in C^\infty(\overline{\Omega}) \tag{4.8.1.9}$$

Since u_1 satisfies the boundary conditions, this is simply the BVP we wished to solve:

$$k^2(u_0, v) + \gamma\left(D_z^2 u_0, v\right) = F(v), \forall v \in C^\infty(\overline{\Omega}). \qquad (4.8.1.10)$$

Also as in section 4.6.1 the sesquilinear form $C(u,v)$ is coercive if $L_k(u) = 0$ has no solutions and if the adjoint problem $L_k^*(u) = 0$ has no solutions. Again, we have insured that this is the case by setting $\epsilon > 0$. The form $C(u,v)$ allows the solution (and therefore the approximating subspaces) to be completely discontinuous, and also does not require the solution space to satisfy any boundary conditions. This added flexibility, however, is gained at the expense of assuring a more continuous test space. Appendix B discusses how this is accomplished. We know turn to the nature of $C(u,v)$ on the FEM solution and test space pair $H^0_{h,q-2}(\Omega) \times H^0(\partial\Omega) \times H^2_{h,q}(\Omega)$.

4.8.2 Error bounds in $H^0(\Omega) \times H^0(\partial\Omega)$ using $H^0_{h,q-2}(\Omega) \times H^0(\partial\Omega) \times H^2_{h,q}(\Omega)$

The error in the solution to the present problem is sensibly measured (in 1-D) only in $H^0(\Omega) \times H^0(\partial\Omega)$. In higher dimensions, the sesquilinear form will involve integral norms on the boundary resulting, for example, from the application of Green's Theorem. This will result in half-integral norms on the boundary, specifically that the trace of u belong to $H^{-1/2}(\partial\Omega)$. In the case that we have placed our source away from the boundary of the domain, we can use theorem 4.4.5 to replace the space $H^{-1/2}(\partial\Omega)$ with a smoother space of boundary distributions, $H^0(\partial\Omega)$. This makes the constant $\hat{C}^2(h,q)$ directly computable, so that the error computations can proceed as for the $H^1(\Omega)$ FEM, even in higher dimensions. We discuss this briefly.

Suppose that we can choose the function χ in theorem 4.4.5 so that $\chi = 1$ in some neighborhood of the boundary. If $\chi f \in C^\infty(\overline{\Omega})$, then $u|_{\partial\Omega} \in C^\infty(\partial\Omega)$, and we can measure $u|_{\partial\Omega}$ in any convenient Sobolev boundary norm. In particular, we have that $u|_{\partial\Omega} \in H^0(\partial\Omega)$. Since $\langle\!\langle u|_{\partial\Omega}\rangle\!\rangle_{-1/2} \leq \langle\!\langle u|_{\partial\Omega}\rangle\!\rangle_0$, we have that

$$| \langle u, D_z v\rangle \, | \leq \langle\!\langle u|_{\partial\Omega}\rangle\!\rangle_{-1/2} \, \langle\!\langle D_z v|_{\partial\Omega}\rangle\!\rangle_{1/2} \leq \langle\!\langle u|_{\partial\Omega}\rangle\!\rangle_0 \, \|v\|_2 \qquad (4.8.2.11)$$

It is still not true that $\langle\!\langle u|_{\partial\Omega}\rangle\!\rangle_0 \leq \|u\|_0$, but we have introduced a directly calculable boundary norm.

Due to the nature of the solution, there is no particularly good reason to restrain the boundary values of u in a FEM approximation to have anything to do with the solution in the domain; in general, they are not necessarily related (see the definition of the $\tilde{H}^s(\Omega)$ in section 4.4.6). For the purposes of computation, we will specify the boundary values as two complex numbers, independently of the polynomial-based expansion on the interval $(0,1)$. This is, of course, the (approximation) space $H^0(\partial\Omega)$. In any dimension, we will typically use FEM spaces for which $\langle\!\langle u|_{\partial\Omega}\rangle\!\rangle_0$ is directly computable; for Lagrange interpolants on the domain, it is a simple linear combination of the boundary nodal values. Calculating $\langle\!\langle u|_{\partial\Omega}\rangle\!\rangle_0$ follows from the definition of the surface integral over $\partial\Omega$.

Summarizing, we can state,

$$|C(u,v)| \leq C(\|u\|_0^2 + \langle\!\langle u|_{\partial\Omega}\rangle\!\rangle_0^2)^{1/2} \, \|v\|_2 \, , \qquad (4.8.2.12)$$

where $u \in H^0(\Omega) \times H^0(\partial\Omega)$, and $v \in H^2(\Omega)$. In the case of smooth forcing in the neighborhood of the boundary, and in general for 1-D, the space $H^0(\Omega) \times H^0(\partial\Omega)$ also contains the solution. We note that the solution space is a Hilbert space with the inner product $(\!(u,v)\!)_0 + \langle u|_{\partial\Omega}, v|_{\partial\Omega}\rangle$, where the boundary inner product reduces to $\overline{u}(1)v(1) + \overline{u}(0)v(0)$ in 1-D. We use equation 4.8.2.12 to extend C to a bounded linear functional on $H^0(\Omega) \times H^0(\partial\Omega) \times H^2(\Omega)$, and we apply the standard FEM error estimate (appendix D) to the resulting method. The determination of $\hat{C}_2(h,q)$ follows the technique of appendix E. In particular, we search for $\hat{C}_2(h,q)$

such that

$$\inf_{\substack{u \in H^0(\Omega)_{h,q-2} \\ u|_{\partial\Omega} \in H^0_{h,q-2}(\partial\Omega) \\ \|u\|_0^2 + \langle\langle u|_{\partial\Omega} \rangle\rangle_0^2 = 1}} \sup_{\substack{v \in H^2(\Omega)_{h,q} \\ \|v\|_2 \neq 0}} |C(u,v)| / \|v\|_2 = \hat{C}_2(h,q), \tag{4.8.2.13}$$

similarly to equation 4.6.2.19. Assuming that we will place the source at an element vertex, we can expect convergence to the solution as follows. Applying theorem 4.4.6 to both the domain and boundary spaces,

$$\|u - \hat{u}\|_0^2 + \langle\langle (u - \hat{u})|_{\partial\Omega} \rangle\rangle_0^2 \leq \left(1 + \frac{C(q)}{\hat{C}_2(h,q)}\right)^2 \left(C_d^2(q) + C_b^2(q)\right) h^{2(q-2)}, \tag{4.8.2.14}$$

where the constants $C_d(q)$ and $C_b(q)$ include sums of the element local domain and boundary norms of the solution. In the 1-D case when the approximation on the boundary is independent of the approximation in the domain (as it is in the numerical tests below) $C_b^2(q) = 0$.

The usefulness of the solution of an associated generalized eigenproblem is not evident, as was the case for the $H^1(\Omega)$ FEM. In the case that we have chosen a solution space for which the boundary values are independent of the domain approximation, the associated eigenproblem makes no sense (since the mass matrix has no dependency on the boundary values). Even for approximation spaces where the generalized eigenproblem makes sense, neither of the norm operators on the FEM spaces are diagonalized in the process of solving that eigenproblem. Therefore the constant $\hat{C}_2(h,q)$ does not attain its extrema at the generalized eigenfunctions of the operator $C(u,v)$. Instead, equation E.0.9 must be used to obtain the extrema. It is not clear (in a rigorous fashion, at least) how the eigenvalues of expression E.0.9 should behave as a function of the mesh parameters h and q. However, it is quite easy to solve this eigenproblem numerically, especially since only one eigenvalue is needed. The results of such a numerical evaluation are shown in figure 4.6. It is worth noting that the plots in figure 4.6 are essentially an exact overlay with the appropriate plots of figure 4.1 for high N.

It will be assumed that the limiting value of $\hat{C}^2(h,q)$ as $h \to \infty$ is reflected by the limiting value of these plots, so that asymptotically optimal convergence is attained once the (apparent) limiting value is reached. For the purpose of numerical evaluations of the error, we will assume that we can use this property of higher order solutions to establish highly accurate reference solutions.

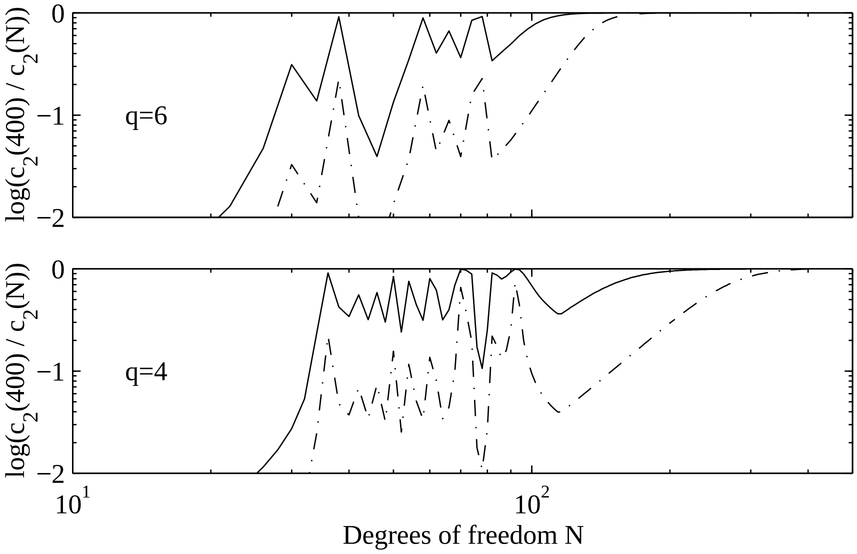


Figure 4.6: Convergence of $1/\hat{C}_2(h,q)$ to its limiting value as $h \to \infty$ when evaluated at $k^2 = k_j^2$, an exact eigenfrequency of the 1-D medium. Results are shown for several element orders. The solid curves correspond to $\epsilon = .01$, the dashed to $\epsilon = .001$. Once $1/\hat{C}_2(\infty,q)$ is attained, the FEM method converges optimally, or at the same rate as the best approximation. Compare the plot for $q = 4$ to the plot for $q = 3$ in figure 4.1 and the plot for $q = 6$ to the $p = 5$ case in figure 4.1; they are essentially identical at higher values of N.

Error estimates can also be established using test spaces for which $D_z v|_{\partial\Omega}$ is sufficiently small. (In 1-D, it is trivial to construct trial spaces with first derivatives that are identically zero on the boundaries; this is not so in higher dimensions.) This is the approach taken in the analysis of convergence in a two-dimensional case discussed in Douglas & Dupont (1977). This is clearly important for the approximation of solutions that are not smooth in the vicinity of the boundary in dimensions higher than 1. This approach was not examined in this work since it is not clear how this would effect the eigenproblem associated with the determination of $\hat{C}_2$ in dimensions higher than 1.

4.8.3 Numerical Tests

As in the case of the H^1 FEM, we compare solutions obtained on a given set of h elements to solutions obtained on the same elements but with much higher element polynomial order. The problem we solve is the 1-D faulted medium with free ends, described by the equation

$$k^2 u + (1 - i\epsilon) D_z^2 u = \delta'(z_0), \quad D_z u|_{\partial\Omega} = 0. \tag{4.8.3.1}$$

The fault at z_0 is placed at an element boundary, and the solution found for a variety of h and q, with reference solutions determined at a q several times higher. A typical solution is shown in figure 4.7.

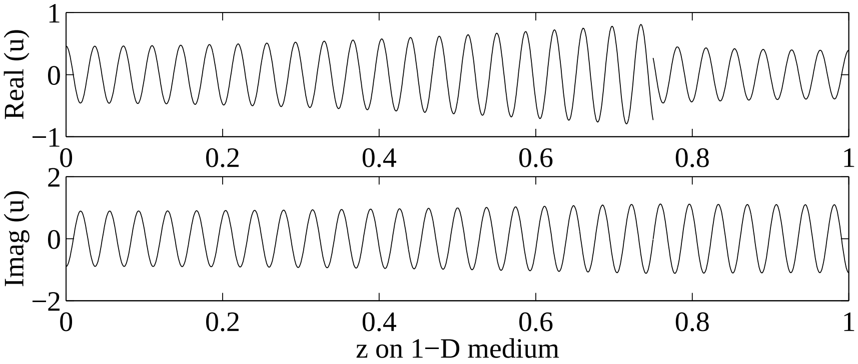


Figure 4.7: Solution u to equation 4.8.3.1, using the discontinuous FEM. An ideal fault is present at $z = .75$, corresponding to a body force of the form $f = \delta'$. Neither the continuous test functions nor the discontinuous solution functions are required to obey any boundary conditions at the ends of the string or at the source. The solution at the source is truly discontinuous. If modeling this solution with a continuous set of solution functions, we would find ringing in the solution resulting from the discontinuity. The solution shown is essentially the exact solution due to the high order of the elements.

The general predictions of relative error for the $H^1(\Omega)$ FEM carry over to the $H^0(\Omega)$ FEM. For small h, as long as $\hat{C}_2$ behaves as figure 4.6 indicates, the $H^0(\Omega)$ relative error will converge as the best approximation. For larger h, we expect dispersion error to dominate as in the H^1 case. We find that the experimentally determined errors (see figure 4.8) can be bounded by

$$\|\hat{u} - u\|_0^2 \,/\, \|u\|_0^2 \;\lesssim\; c_4 m^{-2(q-2)} + c_5 \epsilon^{-2} m^{-4(q-2)}, \tag{4.8.3.2}$$

where the term dependent on ϵ behaves like dispersion error in the H^1 case. Therefore the H^0 FEM should behave as the H^1 FEM (with error measured in H^1), except that the effective element order is one higher in the H^0 FEM.

The FLOP and Memory considerations for the $H^0(\Omega)$ FEM are therefore already documented in figures 4.4 and 4.5, at least in the right hand plots for error in $H^1(\Omega)$. The lines for $p = 3$ in figures 4.4b and 4.5b apply to the $q = 4$ $H^0(\Omega)$ FEM, and the lines for $p = 5$ apply to the $q = 6$ $H^0(\Omega)$ FEM. The same arguments for the advantages of higher order elements apply.

4.9 Higher dimensions and systems of equations

To consider higher dimensional vector PDE's, which is our actual interest, the above results must be generalized. The generalization to systems of equations is discussed in Roitberg (1996), chapter 10. We have already classified the data of interest in section 4.4.5. In the case of $n = 3$ dimensions, we can write, for the Green's function (see table 4.2)

$$f \in H^{-2}(\Omega) \times H^{-2}(\Omega) \times H^{-2}(\Omega) = \prod_{j=1}^{n} H^{-2}(\Omega), \tag{4.9.0.1}$$

where the product of spaces is over the vector components of f. For the ideal fault or explosive equivalent force, we must write, from equation 4.5.0.6 and table

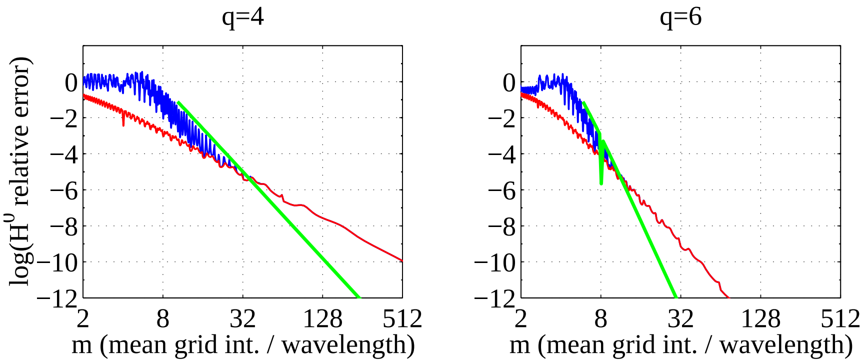


Figure 4.8: Relative errors of the H^0 FEM solution $\hat{u}$ of equation 4.8.3.1 in 1-D using $H^0_{h,q-2}(\Omega) \times H^0(\partial\Omega) \times H^2_{h,q}(\Omega)$ with $q = 4$ and $q = 6$. The blue lines indicate the solution error $\left[|\!|\!| \hat{u} - u |\!|\!|_0^2 + \langle\!\langle (\hat{u} - u)|_{\partial\Omega} \rangle\!\rangle_0^2 \right] / \left[|\!|\!| u |\!|\!|_0^2 + \langle\!\langle u|_{\partial\Omega} \rangle\!\rangle_0^2 \right]$ and the red lines the best approximation error $\left[|\!|\!| \hat{w}_0 - u |\!|\!|_0^2 + \langle\!\langle (\hat{w}_0 - u)|_{\partial\Omega} \rangle\!\rangle_0^2 \right] / \left[|\!|\!| u |\!|\!|_0^2 + \langle\!\langle u|_{\partial\Omega} \rangle\!\rangle_0^2 \right]$, both computed from the FEM using a reference solution u with a much higher p. The green lines are dispersion error predictions computed in a way similar to that for the H^1 FEM. Although the addition of the boundary norm makes the eigenproblem slightly different (note the oscillations at small h), these plots are essentially the same as the rightmost two plots in figure 4.3. The appropriate lines in figures 4.4 and 4.5 therefore predict the error as a function of FLOPS and memory for the H^0 FEM as well.

4.2,

$$f \in \prod_{j=1}^{n} H^{-3}(\Omega). \tag{4.9.0.2}$$

The generalization of elliptic regularity (see section 4.4.8) is given by (Roitberg, 1996, equation 10.1.9, p. 325), and the PDE and associated boundary values of section 4.5.0.1 yield a mapping of the form

$$\{\ell, b_0, b_1\} : \prod_{j=1}^{n} \tilde{H}^{s}(\Omega) \to \prod_{j=1}^{n} H^{s-2}(\Omega) \times H^{s-1/2}(\partial\Omega) \times H^{s-3/2}(\partial\Omega) \tag{4.9.0.3}$$

where the $\{\ell, b_0, b_1\}$ are systems of differential operators and boundary operators. Again, on the condition that the null spaces of $\{\ell, b_0, b_1\}$ and its adjoint are empty, this is again a 1:1, onto map (Roitberg, 1996, Theorem 10.1.2).

The finite element formulations of the $H^1(\Omega)$ FEM and $H^0(\Omega)$ FEM follow simply from the application of Gauss's formula,

$$\int\limits_{\Omega} dV\ \overline{v}_i \partial_j c_{ijkl} \partial_l u_k = \int\limits_{\partial\Omega} dS\ n_j \overline{v}_i c_{ijkl} \partial_l u_k - \int\limits_{\Omega} dV\ c_{ijkl} \partial_j \overline{v}_i \partial_l u_k. \tag{4.9.0.4}$$

The summation convention again applies, and we have assumed that $c_{ijkl} \in C^{\infty}(\overline{\Omega}), \forall \{i, j, k, l\}$. Using the same arguments as in section 4.6.1, the $H^1(\Omega)$ FEM results in the sesquilinear form

$$\mathcal{B}(u, v) = \omega^2 \int\limits_{\Omega} \rho \overline{v}_i u_i - \int\limits_{\Omega} dV\ c_{ijkl} \partial_j \overline{v}_i \partial_l u_k. \tag{4.9.0.5}$$

The $H^1(\Omega)$ FEM is therefore

$$\mathcal{B}(u, v) = \mathcal{F}(v), \forall v \in \prod_{j=1}^{3} H^1(\Omega), u \in \prod_{j=1}^{3} H^1(\Omega), \tag{4.9.0.6}$$

where the restriction of u is necessary for $\mathcal{B}$ to be continuous, and

$$\mathcal{F}(v) = \int\limits_{\Omega} dV\ f_i v_i. \tag{4.9.0.7}$$

This same form is stated in Geller & Ohminato (1994). Of course, this extension of $\mathcal{B}$ requires that $f \in \prod_{j=1}^{n} H^{-1}(\Omega)$, which does not even extend to the Green's function for 2 or 3 dimensions. An error analysis similar to that performed for the 1-D H^1 FEM is completely valid. The convergence will again depend on the behavior of the constant $\hat{C}_2$.

The $H^0(\Omega)$ FEM produces the sesquilinear form

$$\mathcal{C}(u, v) = \omega^2 \int\limits_{\Omega} \rho \overline{v}_i u_i + \int\limits_{\Omega} dV\ u_i \partial_j c_{ijkl} \partial_l \overline{v}_k - \int\limits_{\partial\Omega} dS\ n_j u_i c_{ijkl} \partial_l \overline{v}_k \tag{4.9.0.8}$$

and the $H^0(\Omega)$ FEM is

$$\mathcal{C}(u, v) = \mathcal{F}(v), \forall v \in \prod_{j=1}^{n} H^2(\Omega), u \in \prod_{j=1}^{n} H^0(\Omega) \times H^{-1/2}(\partial\Omega). \tag{4.9.0.9}$$

This extension of $\mathcal{C}$ requires that $f \in \prod_{j=1}^{n} H^{-2}(\Omega)$, which does extend to the Green's function for 2 or 3 dimensions, but does not extend to typical seismic sources. If f is smooth in the neighborhood of the boundary, the $H^0(\Omega)$ FEM becomes

$$\mathcal{C}(u,v) = \mathcal{F}(v), \forall v \in \prod_{j=1}^{n} H^2(\Omega), u \in \prod_{j=1}^{n} H^0(\Omega) \times H^0(\partial\Omega). \tag{4.9.0.10}$$

Again, analysis of the convergence of this case will be analogous to the 1-D H^0 FEM case.

Explosive and ideal fault seismic sources obviously require that $v \in \prod_{j=1}^{n} H^3(\Omega)$, the dual space of $f \in \prod_{j=1}^{n} H^{-3}(\Omega)$. The solution is, by elliptic regularity, $u \in \prod_{j=1}^{n} \tilde{H}^{-1}(\Omega)$. Surprisingly enough, the solution itself is not even square-integrable. One is faced with the construction of useful FEM subspaces of H^{-1} and H^3.

4.10 Conclusions and further work

The 1-D H^0 and H^1 FEM converge to the solution of the linearly attenuating frequency domain wave equation (LAFDWE) in their respective norms at the optimal rate given sufficient mesh refinement. For the H^0 FEM, the appropriate norm in which the solution converges includes a norm on the boundary, although other FEM spaces may converge in H^0 alone (Douglas & Dupont, 1977). Utilizing high order test and solution spaces realizes a large payoff in computational efficiency. The Green's function of the LAFDWE can be found dependably in any dimension up to 3 using the H^0 FEM.

The extension of the H^0 FEM to triangular parameterizations might best be developed from the high-order FEM triangular bases of Mavripilis & van Rosendale (1993) or Sherwin & Karniadakis (1996). These authors develop high-order FEM bases and quadratures from tensor-product Lagrange interpolants on the triangle (and tetrahedra). It should not be difficult to extend these to Hermite interpolants and thereby generate C^1 elements for the test space of the H^0 FEM. Triangular elements yield a powerful way of paving irregular domains in two dimensions. The combination of a flexible, high-order accurate solution warrants further exploration. At the time of writing, the Center for Fluid Mechanics at Brown University is working on discontinuous Galerkin methods for spectral elements (on the web: *http://www.cfm.brown.edu/people/cbq/*).

A Appendix: The $H^1_{h,p}$ FEM Approximation Space

The discussion of this section follows that of Patera (1984) very closely. A basis for $H^1_{h,p}$ requires a set of functions which are continuous at element boundaries but have no continuity requirements on higher-order derivatives at these boundaries. One easy way to construct such a basis is to begin with any of the Lagrange interpolators, functions L_m which exhibit the property

$$L_m(y_n) = \delta_{mn}, \tag{A.0.1}$$

at a set of knots y_n, where $0 \leq n \leq p-1$ and $0 \leq m \leq p-1$. $p-1$ is the maximum order of the polynomials comprising the Lagrange interpolants. The knots y_n are chosen for convenience, but it is generally true that for higher orders, numerical stability demands that the knots cluster near the ends of the elements. Chebyshev and Legendre polynomials are typical, and Legendre elements are preferred if the computational effort to construct the mass and stiffness matrices is important. Since such problems are not of interest in this work, the Chebyshev basis is chosen for its ease of use. A convenient set of knots for the Chebyshev-based Lagrange interpolants are the p Gauss-Lobatto Chebyshev points Patera (1984)

$$y_n = -\cos\left(\frac{n\pi}{N}\right), 0 \leq n \leq p-1, \tag{A.0.2}$$

upon which the Chebyshev polynomials effortlessly generate Lagrange interpolants on the interval $[-1, 1]$

$$L_j(y) = \begin{cases} \frac{2}{p-1}\sum_{n=0}^{p-1}\frac{T_n(y_j)}{c_j c_n}T_n(y) & \text{if } -1 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{A.0.3}$$

where T_m is the Chebyshev polynomial of degree m, $0 \leq m \leq p-1$, defined by

$$T_m(y) = cos(m \arccos(y)) \tag{A.0.4}$$

and

$$c_m = \begin{cases} 1 & \text{if } m \neq 0, p-1 \\ 2 & \text{if } m = 0, p-1 \end{cases} \tag{A.0.5}$$

A basis for $H^1_{h,p}$ is now easily constructed by paving the medium with elements composed of Lagrange interpolators, and forcing adjoining elements to have the same boundary coefficient. More concisely,

$$f(z) = a_m f_m\left(\frac{2\underline{h}\,(z - z_k)}{Z}\right), 1 \leq m \leq \underline{h}(p-1) + 1, \tag{A.0.6}$$

where the summation convention applies,and

$$z_k = \frac{Zk}{\underline{h}}, \tag{A.0.7}$$

is the minimum z coordinate of the k^{th} element, $k = \text{floor}\left(\frac{m}{p}\right)$, and $\underline{h} = Z/h$. The individual FEM basis functions are simple combinations of Lagrange interpolants

$$f_m(y) = \begin{cases} L_0\left(|y| - 1\right) & \text{if } m \bmod p = 0 \\ L_{(m \bmod p)}\left(y - 1\right) & \text{otherwise} \end{cases}, \tag{A.0.8}$$

where the fact that $L_0(-y) = L_p(y)$ has been used. This construction assures continuity of the function to be represented at the element boundaries, and allows the derivatives at the boundaries to be discontinuous.

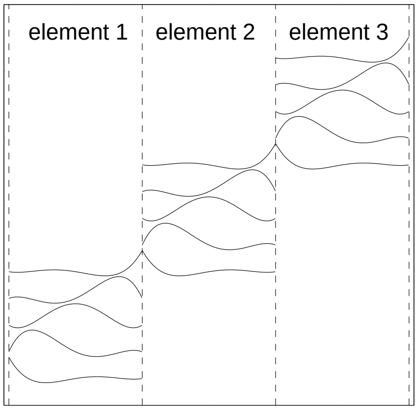


Figure 4.9: Lagrange interpolant basis for $H^1(\Omega)$. Three elements are shown, each containing 5 basis functions. Each basis function is a linear combination of Chebyshev polynomials. This basis is referred to as $H^1_{1/3,5}$.

The expressions for the mass and stiffness matrices are evaluated by performing the necessary integrals analytically for Chebyshev polynomials, then taking linear combinations of these integrals in accordance with the definition of the Lagrange interpolants. Since the intervals of definition of the elements do not overlap, the total mass and stiffness matrices can be formed from the appropriate sums of the element mass and stiffness matrices (so-called "direct stiffness" summation).

The element mass matrix is

$$M^e_{ij} = (L_i, L_j) = \int_{-1}^{1} dy \; L_i(y) L_j(y) \tag{A.0.9}$$

and the stiffness matrix is

$$K^e_{ij} = ((L_i, L_j)) = \int_{-1}^{1} dy \; \partial_y L_i(y) \; \partial_y L_j(y). \tag{A.0.10}$$

Substituting the expression for L (equation A.0.3) into these expressions yields

$$M^e_{ij} = \frac{4}{N^2} \sum_{m=0}^{N} \sum_{n=0}^{N} \frac{T_n(y_j)}{c_j c_n} \frac{T_m(y_i)}{c_i c_m} \int_{-1}^{1} dy \; T_n(y) T_m(y), \tag{A.0.11}$$

and

$$K^e_{ij} = \frac{4}{N^2} \sum_{m=0}^{N} \sum_{n=0}^{N} \frac{T_n(y_j)}{c_j c_n} \frac{T_m(y_i)}{c_i c_m} \int_{-1}^{1} dy \; \partial_y T_n(y) \; \partial_y T_m(y), \tag{A.0.12}$$

The integrals in these expressions are easily evaluated, resulting in two triple matrix products to form M^e and K^e. To form the global matrices K and M, it is sufficient to translate the element matrices by $N-1$ along the diagonal and sum for each of the $\underline{h}$ elements. The overlap of the element matrices by one provides C^0 continuity. An example of such a sum is shown in figure 4.13. It is also necessary to incorporate the change from local element to global variable. Writing the translated sum (or direct stiffness sum) as $\sum'$ (after Patera (1984)), the global matrices are

$$M_{ij} = \frac{Z}{2\underline{h}} \sum_{n=1}^{\underline{h}}{}' \, M^e_{ij} = \frac{Z}{2\underline{h}} \hat{M}_{ij}, \tag{A.0.13}$$

and

$$K_{ij} = \frac{2h}{Z} \sum_{n=1}^{h}{}' K^e_{ij} = \frac{2h}{Z} \hat{K}_{ij}, \tag{A.0.14}$$

where the matrices $\hat{M}_{ij}$ and $\hat{K}_{ij}$ will be used elsewhere.

B Appendix: The $H^2_{h,p}$ and $H^0_{h,p}$ FEM spaces

Bases for $H^2_{h,p}$ and $H^0_{h,p}$ can be easily built from piecewise polynomials. The construction of discontinuous functions for $H^0_{h,p}$ follows from the previous section; it is not necessary to overlap the parameterization from element to element in order to enforce any continuity requirements. A basis for $H^0_{h,p}$ is shown in figure 4.10. The construction of C^1 functions as a basis for $H^2_{h,p}$ is somewhat more complicated. The additional continuity requirements between elements is enforced by the construction of Hermite interpolants.

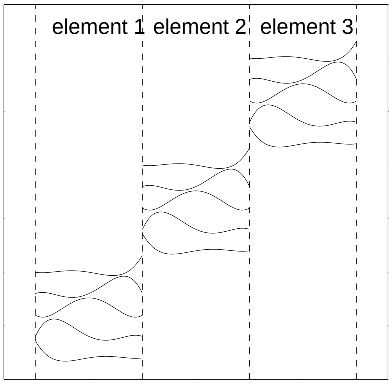


Figure 4.10: Lagrange interpolant basis for $H^0(\Omega)$. Three elements are shown, each containing 5 basis functions. Note that no continuity is imposed across element boundaries. Each basis function is a linear combination of Chebyshev polynomials. This basis is referred to as $H^0_{1/3,5}$.

Hermite interpolants extend the Lagrange interpolant to include derivative values at the nodes. The Hermite interpolants are defined by the simple requirements

$$g_m(y_k) = \delta_{mk}, \quad g'_m(y_k) = 0, \; m,k = 0, \cdots, N \tag{B.0.1}$$

$$h'_m(y_k) = \delta_{mk}, \quad h_m(y_k) = 0, \; m,k = 0, \cdots, N \tag{B.0.2}$$

so that the g_m interpolate function values and the h_m interpolate derivatives at the $N+1$ nodes y_k. The g_m and h_m are expanded in Chebyshev polynomials as

$$g_m(x) = \sum_{n=0}^{2N+1} T_n(x) a_{mn}, \tag{B.0.3}$$

$$h_m(x) = \sum_{n=0}^{2N+1} T_n(x) b_{mn}. \tag{B.0.4}$$

Inserting into the definition of the Hermite interpolants and evaluating at the Gauss-Lobatto Chebyshev points of appendix A,

$$\begin{aligned} g_m(x_k) &= \sum_{n=0}^{2N+1} T_n(x_k) a_{mn} &&= \delta_{mk}, && \text{(B.0.5)} \\ g'_m(x_k) &= \sum_{n=0}^{2N+1} T'_n(x_k) a_{mn} &&= 0, && \text{(B.0.6)} \end{aligned}$$

and

$$\begin{aligned} h_m(x_k) &= \sum_{n=0}^{2N+1} T_n(x_k) b_{mn} &&= 0, && \text{(B.0.7)} \\ h'_m(x_k) &= \sum_{n=0}^{2N+1} T'_n(x_k) b_{mn} &&= \delta_{mk}. && \text{(B.0.8)} \end{aligned}$$

where $m, k = 0, \cdots, N$. The derivatives of the Chebyshev polynomials follow from equation A.0.4 and are

$$T'_n(x_k) = \begin{cases} n^2, & k = 0, \\ n^2(-1)^{n-1}, & k = N, \\ \frac{n \sin \frac{n\pi k}{N}}{\sin \frac{\pi k}{N}}, & k = 1, \cdots, N-1. \end{cases} \tag{B.0.9}$$

The Chebyshev coefficients a_{mn} and b_{mn} are therefore easily determined by matrix inversion.

The spaces $H^2_{h,p}$ and $H^0_{h,p}$ are related by linear transformation to the Chebyshev polynomials, so that mass and stiffness matrices are determined by the appropriate operations on Chebyshev polynomials. In order to assure C^1 continuity, the mass and stiffness matrices are summed from element matrices with an

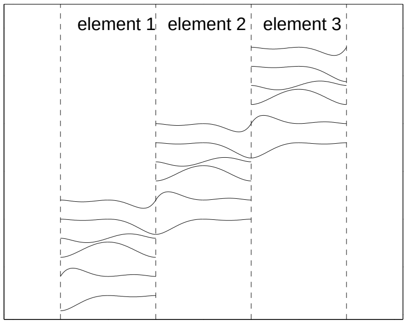


Figure 4.11: Hermite interpolant basis for $H^2(\Omega)$. Three elements are shown, each containing 6 basis functions. The two combined functions at the ends of the elements enforce continuity of the function represented and continuity of the first derivative. Each basis function is a linear combination of Chebyshev polynomials. This basis is referred to as $H^2_{1/3,6}$.

overlap to assure that the derivative and function values at the element boundaries are each parameterized by a single coefficient. Figure 4.11 demonstrates an example of $H^2_{h,p}$.

The use of $H^2_{h,p}$ and $H^0_{h,p}$ as test and solution spaces for the H^0 FEM yields some freedom in the choice of spaces to yield square stiffness and mass matrices. An extra degree of freedom is consumed at each end of an element for each continuity requirement. This suggests that one utilize a test space of degree 2 greater than that of the solution space. The resulting disparity in solution and test space degrees of freedom can be accommodated quite naturally in the H^0 FEM by the introduction of an independent boundary parameter. An example of this is shown in figure 4.12.

C Appendix: The generalized eigenproblem on the $H^1_{h,p}$ FEM solution space

The finite element eigenproblem has been the subject of much analysis in the literature. Primary interest has focused on the asymptotic convergence properties of the finite element method, stemming from the application of the method to stability analysis, where only the lowest eigenvalues and vectors are of interest. However, the entire FEM spectrum $\left\{\left\{\hat{k}_j^2, \hat{e}_j\right\}, j = 1 \dots N\right\}$ for all h, p and j is

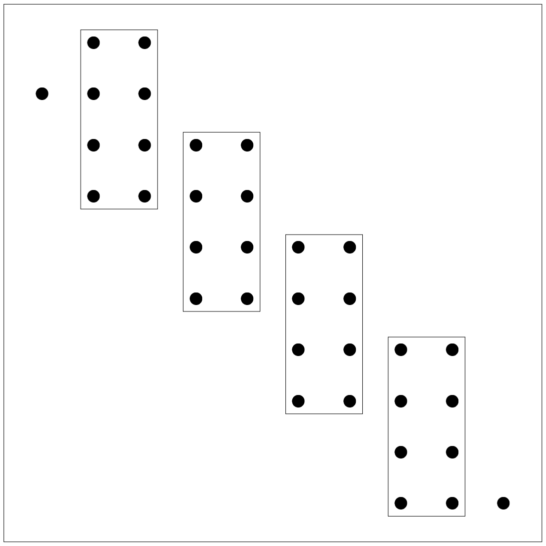


Figure 4.12: Stiffness matrix in the discontinuous FEM. In order that the matrix be square, the test and solution spaces have different element orders. The internal elements represent parameterizations of the form $H^2_{h,4} \times H^0_{h,2}$. The extra components near the corners of the matrix represent the boundary term present in the H^0 FEM. The overlap in the column space enforces the C^1 continuity of the test functions. It is presumed that discontinuities in the solution (at the fault in 1-D) will be placed at element boundaries. We will refer to the FEM solution using these two spaces using the order of the test space: $q = 4$.

of interest in the present problem.

The solutions to the generalized eigenproblem are desired in $H^1_{h,p}$. Expanding in terms of the FEM basis functions $f_j \in H^1_{h,p}$ (see equation A.0.8),

$$\hat{e}_m = {}_m a_j f_j \tag{C.0.1}$$

the generalized eigenproblem can be written using the stiffness and mass matrices as

$$\hat{k}^2_m \, M_{ij} \, {}_m a_j = K_{ij} \, {}_m a_j, \tag{C.0.2}$$

where the summation convention has been used. Solution of equation C.0.2 is a fairly straightforward problem numerically, but in order to understand the solution better, a solution is obtained analytically. It is also convenient to have a more numerically robust technique for solving the eigenproblem than a direct numerical technique. The closed form solution of equation C.0.2 has been worked out for the case of linear ($p = 2$) elements in Strang & Fix (1973), and the results are repeated briefly below. The solutions for higher order problems follow in all their gory detail.

C.1 The linear eigenproblem

For linear elements, the mass matrix is

$$\mathbf{M} = \frac{1}{6\underline{h}} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & n-1 & n & n+1 & N-1 & N \end{array} \\ \begin{pmatrix} 2 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & 4 & \ddots & 0 & \cdots & \cdots & \vdots \\ 0 & \ddots & \ddots & 1 & 0 & \cdots & \vdots \\ \vdots & 0 & 1 & 4 & 1 & 0 & \vdots \\ \vdots & \vdots & 0 & 1 & \ddots & \ddots & 0 \\ \vdots & \vdots & \vdots & 0 & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 & 2 \end{pmatrix} \end{array} \tag{C.1.3}$$

and the stiffness matrix is

$$\mathbf{K} = \underline{h} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & n-1 & n & n+1 & N-1 & N \end{array} \\ \begin{pmatrix} 1 & -1 & 0 & \cdots & \cdots & \cdots & 0 \\ -1 & 2 & \ddots & 0 & \cdots & \cdots & \vdots \\ 0 & \ddots & \ddots & -1 & 0 & \cdots & \vdots \\ \vdots & 0 & -1 & 2 & -1 & 0 & \vdots \\ \vdots & \vdots & 0 & -1 & \ddots & \ddots & 0 \\ \vdots & \vdots & \vdots & 0 & \ddots & \ddots & -1 \\ 0 & \cdots & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix} \end{array} \tag{C.1.4}$$

where it has been assumed that the 1-D medium has unit length. As before, $\underline{h}$ is the number of elements, and $N = \underline{h} + 1$. Rewrite equation C.0.2 as the set of difference equations

$$\frac{k_m^2}{6\underline{h}^2}\left(4\,{}_m a_j + {}_m a_{j+1} + {}_m a_{j-1}\right) = 2\,{}_m a_j - {}_m a_{j+1} - {}_m a_{j-1},\ 0 \le j \le \underline{h} \tag{C.1.5}$$

with the additional constraints to handle the boundaries

$${}_m a_{-1} = {}_m a_1, {}_m a_{\underline{h}-1} = {}_m a_{\underline{h}+1}. \tag{C.1.6}$$

Equation C.1.5 is solved by the interpolants of the true eigenfunctions

$$ {}_m a_j = \cos\left(\frac{mj\pi}{\underline{h}}\right), 0 \leq j \leq \underline{h}, 0 \leq m \leq \underline{h}. \tag{C.1.7}$$

Substituting equation C.1.7 into C.1.5 and using some trigonometric identities yields the linear FEM eigenvalues:

$$k_m^2 = 6\underline{h}^2 \left(\frac{1-\cos\left(\frac{mj\pi}{\underline{h}}\right)}{2+\cos\left(\frac{mj\pi}{\underline{h}}\right)}\right). \tag{C.1.8}$$

C.2 Higher order eigenproblems

The solution of higher order ($p \geq 2$) eigenproblems is more complicated, since the stiffness and mass matrices are no longer tridiagonal, but consist of the overlapped sums of the element matrices. Solution of these problems is accomplished by separating the problem into two sets of equations, one of which is tridiagonal and can be solved by interpolants of cosines. The method is suggested by *static condensation*, which is also a recipe for parallelization (Patera, 1984). The matrix equation C.0.2 can be directly rewritten as

$$\begin{aligned} K_{BB}a_B + K_{BI}a_I &= \lambda\left(M_{BB}a_B + M_{BI}a_I\right) && \text{(C.2.9)} \\ K_{IB}a_B + K_{II}a_I &= \lambda\left(M_{IB}a_B + M_{II}a_I\right) && \text{(C.2.10)} \end{aligned}$$

where the subscripts I and B refer to internal and boundary (of a finite element) entries in the matrices of equation C.0.2. This is simply a partitioning of the matrices K and M and solution vector a; equations C.2.10 are the element boundary rows of equation C.0.2, equations C.2.10 the element internal rows. Collecting terms,

$$\begin{aligned} \left(K_{BB} - \lambda M_{BB}\right)a_B &= -\left(K_{BI} - \lambda M_{BI}\right)a_I && \text{(C.2.11)} \\ \left(K_{IB} - \lambda M_{IB}\right)a_B &= -\left(K_{II} - \lambda M_{II}\right)a_I. && \text{(C.2.12)} \end{aligned}$$

Now solve for the element internal values in terms of the element boundary values using the second equation:

$$a_I = -\left(K_{II} - \lambda M_{II}\right)^{-1}\left(K_{IB} - \lambda M_{IB}\right) a_B, \tag{C.2.13}$$

and substitute into the first equation to get

$$\begin{aligned} &\left[\left(K_{BB} - \lambda M_{BB}\right)\right. \\ &\left.- \left(K_{BI} - \lambda M_{BI}\right)\left(K_{II} - \lambda M_{II}\right)^{-1}\left(K_{IB} - \lambda M_{IB}\right)\right] a_B = 0. \end{aligned} \tag{C.2.14}$$

The matrix in square brackets in equation C.2.14 is a tridiagonal one. The resulting difference equation is in terms of element boundary coefficients, so that, in the case of higher order elements, the eigenvectors interpolate cosines at the element boundaries. The equations at the ends of the medium can be handled in the same manner as in the linear case. In order to set down precisely the components of C.2.14, figure 4.13 shows a partitioning of a general high order FEM (1-D) stiffness or mass matrix.

Using the submatrices in figure 4.13, the difference equation resulting from equation C.2.14 can be written

$${}_m a_{j-1}\left(v - \underline{x}^T\theta^{-1}x\right) + {}_m a_j\left(v - 2x^T\theta^{-1}x\right) + {}_m a_{j+1}\left(v - \underline{x}^T\theta^{-1}x\right) = 0, \tag{C.2.15}$$

where the values corresponding to those in the figure are formed from the individual mass and stiffness matrices as follows:

$$v = v_K - \lambda v_M \tag{C.2.16}$$

$$u = u_K - \lambda u_M \tag{C.2.17}$$

$$x = x_K - \lambda x_M \tag{C.2.18}$$

$$\theta = \theta_K - \lambda \theta_M. \tag{C.2.19}$$

Again, the difference equation (C.2.15) is solved by

$${}_m a_j = \cos\left(\frac{mj\pi}{\underline{h}}\right), 0 \le j \le \underline{h}, 0 \le m \le \underline{h}, \tag{C.2.20}$$

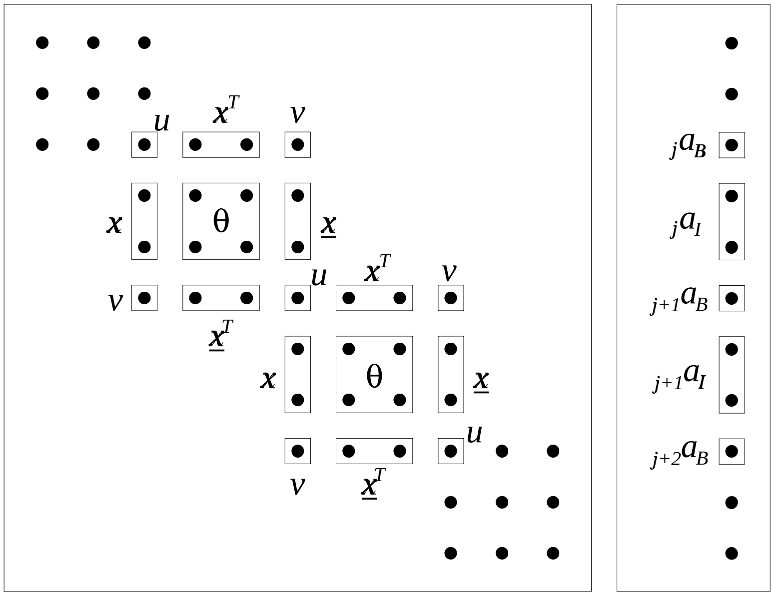


Figure 4.13: Partitioning of a FEM mass or stiffness matrix and an eigenvector for the purpose of solving the generalized eigenvalue problem. Dots correspond to non-zero entries in the matrix or vector. Each 4x4 element matrix is symmetric about both diagonal axes, so that $\underline{x}$ is x with its elements in reverse order. The vector is simply decomposed into boundary and internal degrees of freedom. The FEM space corresponding to this diagram is $H^1_{h,4}$.

resulting in the equation for λ

$$F_0 = 2\ v\ cos\left(\frac{m\pi}{\underline{h}}\right) - 2\ \underline{x}^T\theta^{-1}x\ cos\left(\frac{m\pi}{\underline{h}}\right) + u - 2\ x^T\theta^{-1}x = 0. \tag{C.2.21}$$

The above equation must yield $\underline{h}(p-1)+1$ eigenvalues. One of these eigenvalues is 0 with the corresponding constant eigenfunction. There are $\underline{h}+1$ ways of choosing m, so there must be multiple eigenvalues for each m. Inspection of equations C.2.21 and C.2.19 reveals that there are several values of λ at which the term θ^{-1} is singular. In particular, the generalized eigenvalue problem for the element internal degrees of freedom

$$[\theta_K - {}_I\lambda\theta_M]\,a_I = 0, \tag{C.2.22}$$

has eigenvalues ${}_I\lambda_n, n = 1,\ldots,p-2$. These ${}_I\lambda$ subdivide the total eigenvalue range into $p-1$ mode branches, with special cases occurring at $m=0$ and $m=\underline{h}$. Figure 4.14 shows this behavior in the value of the indicial equation C.2.21.

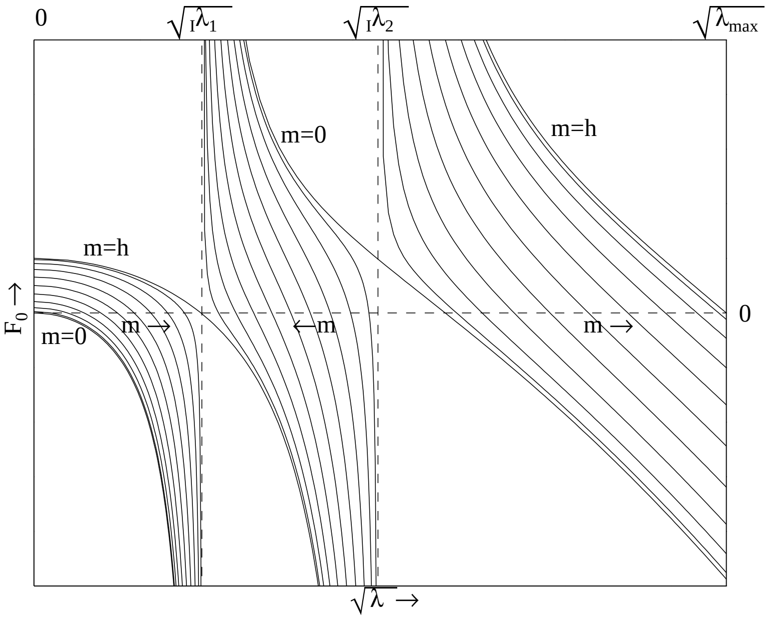


Figure 4.14: Plot of the values of the function F_0 of equation C.2.21. The element internal eigenvalues ${}_I\lambda$ divide the full eigenvalue range into $p-1$ branches. Each solid curve corresponds to a value of $m, m=0,\ldots,\underline{h}$, and adjacent curves have m values differing by 1. An eigenvalue lies at the intersection of each curve with 0. The FEM space corresponding to this diagram is $H_{1/10,4}$.

Equation C.2.21 is quite well behaved once the ${}_I\lambda$ are known, and is easily solved using a common iterative root finding algorithm. The entire process is easily automated since the element internal eigenvalues are obtained by solving the generalized eigenvalue problem (equation C.2.22) for a relatively small $(p-2)\times(p-2)$ system. Determination of the maximum eigenvalue is also quite simple, as shall soon be revealed.

Once the eigenvalues are found, it is a straightforward matter to compute the element-internal values of the eigenfunctions, since the boundary values are already known to be interpolates of cosines. Rewriting equation C.2.13 using the submatrices of figure 4.13,

$$ {}_j a_I = -\theta^{-1}\; x \;\; {}_j a_B - \theta^{-1}\; \underline{x} \;\; {}_{j+1} a_B, \; j = 0, \ldots, \underline{h}, \tag{C.2.23}$$

in other words, the internal coefficients of an element only depend on that element's boundary values.

Finally, the eigenvalues and eigenvectors for any $H^1_{h,p}$ can be computed quite handily, and a number of interesting facts about the nature of the eigenproblem become apparent. First, return to the definition of the mass and stiffness matrices from equations A.0.13 and A.0.14. Using these definitions, write the eigenvalue problem (on the space $H^1_{h,p}$) in equation C.0.2 as

$$ {}^p_n\Lambda^{\underline{h}}_m \; \hat{M}_{ij} \; {}_n a_m = \hat{K}_{ij} \; {}_n a_m, \tag{C.2.24}$$

where

$$ {}_n\hat{k}^2_m = \frac{4\underline{h}^2}{Z^2} \; {}^p_n\Lambda^{\underline{h}}_m, \tag{C.2.25}$$

where the eigenvalues are indexed by their branch number n and the cosine interpolant index m. This is somewhat sloppy considering that some modes lie on their own branch, but these modes do not vary with $\underline{h}$ in any case. Consider the values of Λ. If p is fixed, then the element submatrices of equation C.2.24 are the same for all $\underline{h}$. This means that the gaggle of submatrices in the eigenproblem analysis, u,v,x,$\underline{x}$ and θ are also unchanged as $\underline{h}$ is varied. The only difference in equation C.2.21 as $\underline{h}$ is increased is that the term $\cos\left(\frac{m\pi}{\underline{h}}\right), m = 0, \ldots, h$ interpolates the interval $[-1, 1]$ more finely. The distribution of the modes amongst the $n, n = 1, \ldots, p-1$ branches remains the same, and the eigenfrequencies within each branch are ordered by m, either increasing or decreasing, depending on the branch (see figure 4.14). This implies that the spectra of meshes in $H^1_{h,p}$ simply interpolates some continuous limiting spectral function ${}^p_n\Lambda^\infty(\cdot)$. For the space

$H^1_{\underline{h},p}$,

$$ {}^p_n\Lambda^{\underline{h}}_m = {}^p_n\Lambda^\infty\left(\frac{m}{\underline{h}}\right), 0 \le m \le \underline{h}, n = 0, \ldots, p-1, \tag{C.2.26}$$

so that ${}^p_n\Lambda^\infty(\cdot)$ is defined on the interval $[0, 1]$. The eigenvalues of the approximate medium are therefore

$$ {}_n\hat{k}^2_m = \frac{4\underline{h}^2}{Z^2}\, {}^p_n\Lambda^\infty\left(\frac{m}{\underline{h}}\right). \tag{C.2.27}$$

For comparison, the true eigenvalues sample the parabola

$$\begin{aligned} k^2_j &= \frac{j^2\pi^2}{Z^2} \\ &= \frac{(N-1)^2\pi^2}{Z^2}\left(\frac{j}{N-1}\right)^2, \\ &= \frac{\underline{h}^2(p-1)^2\pi^2}{Z^2}\left(\frac{j}{N-1}\right)^2, j = 0, \ldots, N-1, \end{aligned} \tag{C.2.28}$$

where $N = \underline{h}(p-1)+1$ is the number of degrees of freedom in the space $H^1_{\underline{h},p}$.

Equations C.2.27 and C.2.28 are used below to compare the approximate and true eigenvalues for any number $\underline{h}$ of finite elements of order p. The eigenvalues for several p are shown in figure 4.15. One consequence of equation C.2.26 is that the maximum eigenvalue (always at $m = 0$ or $m = \underline{h}$), just as the location of the mode branches, can be determined from the generalized eigenvalue problem for a single element and one of its submatrices.

C.3 Fourier series for higher order FEM eigenfunctions

Some effort is now made to understand some of the properties of the approximate eigenfunctions.

The expansion of each $\hat{e}_j$ as a combination of e_m follows directly from writing the $\hat{e}_j$ as the convolution of a local element interpolating function with a sampled cosine, then taking the proper Fourier transforms. Collecting terms proportional

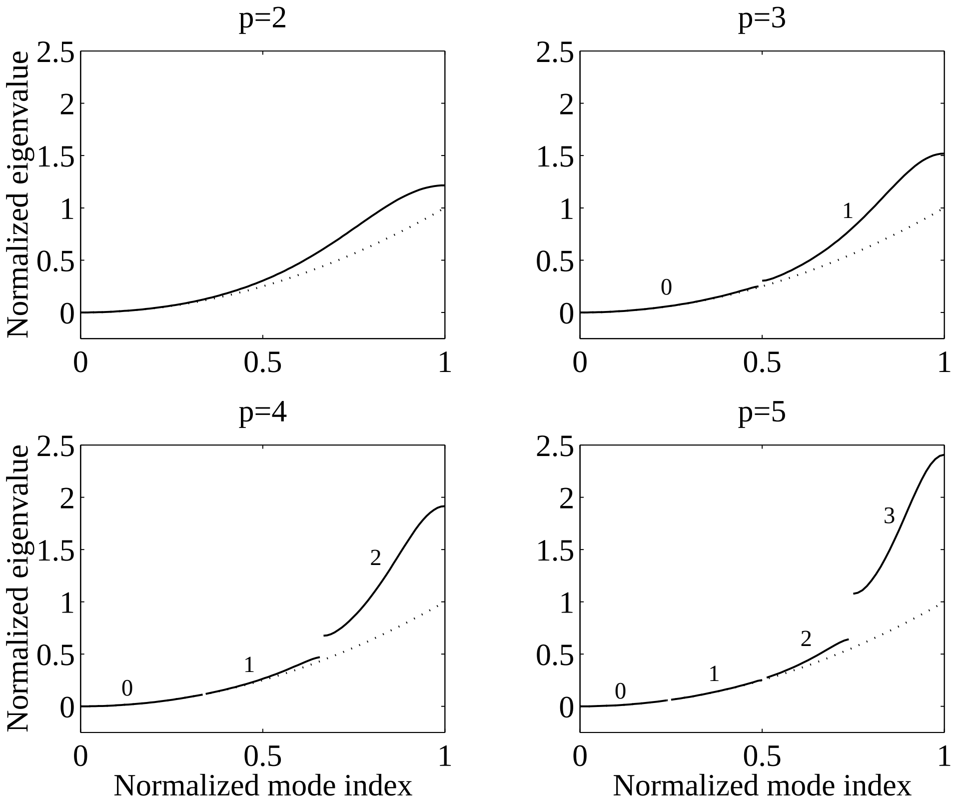


Figure 4.15: FEM (solid line) and exact (dotted line) eigenvalues for several element orders. Equations C.2.27 and C.2.28 have been used to scale the eigenvalues for comparison. In all plots, the true eigenvalue at the highest mode index has amplitude 1. The normalized mode index is obtained by sorting the eigenvalues in ascending order and counting from zero, then dividing by the maximum index. As the number of elements is varied, the solid lines are interpolated more or less finely. Each plot exhibits $p-1$ mode branches, as numbered from 0. The higher order elements yield more accurate lower eigenvalues and possess higher branches of very inaccurate eigenvalues.

to the sample value at a single element boundary in equation C.2.23 yields an expansion for one of the approximate eigenfunctions as

$$ {}_l\hat{e}_j(z) = \sum_{n=0}^{\underline{h}} \cos(\frac{jn\pi}{\underline{h}})\, {}^l_j g_n \left(\frac{2\underline{h}\,(z - z_n)}{Z} \right), \tag{C.3.29} $$

where $0 \leq l \leq p-1$ is the branch index and the ${}^l_j g_n$ are related to the original f_m (equation A.0.8) by

$$ {}^l_j g_n(y) = \sum_{k=-(p-1)}^{(p-1)} {}^l_j w_k f_{(k+np+1)}(y). \tag{C.3.30} $$

The weights ${}^l_j w_k$ are obtained from equation C.2.23 and in column vector form are

$$ {}^l_j w = \begin{bmatrix} -{}^l_j\theta^{-1}\, \underline{x} \\ 1 \\ -{}^l_j\theta^{-1}\, x \end{bmatrix}, \tag{C.3.31} $$

where k is the row index. A brief check of the properties of the $f_m(y)$ reveals that the symmetry of the ${}^l_j w$ produces a ${}^l_j g_n(y)$ which is symmetric about $y = 0$.

Using equation C.2.19 and equations A.0.13 , A.0.14 , C.2.24, C.2.25, and C.2.26

$$ {}^l_j\theta = \frac{2\underline{h}}{Z} \left(\theta_{\hat{K}} - {}^p_l\Lambda^{\infty} \left(\frac{j}{\underline{h}} \right)\, \theta_{\hat{M}} \right) \tag{C.3.32} $$

$$ {}^l_j x = \frac{2\underline{h}}{Z} \left(x_{\hat{K}} - {}^p_l\Lambda^{\infty} \left(\frac{j}{\underline{h}} \right)\, x_{\hat{M}} \right), \tag{C.3.33} $$

so that the components of the weight vector are

$$ {}^l_j w = \begin{bmatrix} -\left[\left(\theta_{\hat{K}} - {}^p_l\Lambda^{\infty} \left(\frac{j}{\underline{h}} \right)\, \theta_{\hat{M}} \right) \right]^{-1} \left(\underline{x}_{\hat{K}} - {}^p_l\Lambda^{\infty} \left(\frac{j}{\underline{h}} \right)\, \underline{x}_{\hat{M}} \right) \\ \\ 1 \\ \\ -\left[\left(\theta_{\hat{K}} - {}^p_l\Lambda^{\infty} \left(\frac{j}{\underline{h}} \right)\, \theta_{\hat{M}} \right) \right]^{-1} \left(x_{\hat{K}} - {}^p_l\Lambda^{\infty} \left(\frac{j}{\underline{h}} \right)\, x_{\hat{M}} \right) \end{bmatrix}, \tag{C.3.34} $$

This expression is derived to show that the ${}^{l}_{j}w$ do not depend explicitly on $\underline{h}$, but only on the ratio $j/\underline{h}$. Therefore the ${}^{l}_{j}g_n(y)$ are also dependent on $j/\underline{h}$, not $\underline{h}$. Rewrite equation C.3.29 using the nomenclature of Bracewell (1986) as

$$ {}_{l}\hat{e}_j(z) = \left[\cos\left(\frac{j\pi z}{Z}\right)\frac{\underline{h}}{Z}\mathrm{III}\left(\frac{\underline{h}z}{Z}\right)\right] * {}^{l}_{j}g\left(\frac{2\underline{h}z}{Z}\right), \tag{C.3.35} $$

where $*$ refers to convolution and III is the sampling function

$$ \mathrm{III}(x) = \sum_{n=-\infty}^{\infty} \delta(x-n). \tag{C.3.36} $$

While the function in equation C.3.35 is defined over $(-\infty, \infty)$, it agrees exactly with the approximate eigenfunction in the interval $[0, Z]$.

Using the Fourier transform pairs

$$ \mathrm{III}(x) \Leftrightarrow \mathrm{III}(s) \tag{C.3.37} $$

$$ \cos(\pi x) \Leftrightarrow \frac{1}{2}\delta\left(s+\frac{1}{2}\right) + \frac{1}{2}\delta\left(s-\frac{1}{2}\right) \tag{C.3.38} $$

$$ {}_{l}\hat{e}_j(x) \Leftrightarrow {}_{l}\hat{\mathrm{E}}_j(s) \tag{C.3.39} $$

$$ {}^{l}_{j}g(x) \Leftrightarrow {}^{l}_{j}G(s), \tag{C.3.40} $$

and the Similarity theorem (Bracewell (1986), p. 101) the Fourier transform of equation C.3.35 is

$$ {}_{l}\hat{\mathrm{E}}_j(S) = \left[\left[\frac{1}{2}\delta\left(S+\frac{j}{2\underline{h}}\right) + \frac{1}{2}\delta\left(S-\frac{j}{2\underline{h}}\right)\right] * \mathrm{III}\left(S\right)\right] {}^{l}_{j}G\left(2S\right). \tag{C.3.41} $$

where the change of variable to $S = \frac{sZ}{\underline{h}}$ has been made to understand the effects of changing the number of elements $\underline{h}$. Since ${}^{l}_{j}g(x)$ is symmetric about the origin, ${}^{l}_{j}G(2S)$ is real and symmetric, so that C.3.41 represents a sum of cosines. The wavenumbers of these cosines are the wavenumbers of the aliases of $\cos\left(\frac{j\pi z}{Z}\right)$ on the set of boundary knots. This is shown schematically in figure 4.17.

The aliasing of C.3.41 yields a fairly succinct expression for the FEM eigenvectors:

$$ {}_{l}\hat{e}_j\left(z\right) = \sum_{m=1}^{\infty} \cos\left[\frac{\pi z}{Z}\left(-j+2\underline{h}m\right)\right] {}^{l}_{j}G\left[2\left(-\frac{j}{2\underline{h}}+m\right)\right] $$

$$+\sum_{m=0}^{\infty} \cos\left[\frac{\pi z}{Z}(j+2\underline{h}m)\right] \, {}^{l}_{j}G\left[2\left(\frac{j}{2\underline{h}}+m\right)\right] \qquad \text{(C.3.42)}$$

Finally, using the orthogonality relation

$$\left(\cos\left(\frac{m\pi z}{Z}\right), \cos\left(\frac{n\pi z}{Z}\right)\right) = \int_0^Z dz \, \cos\left(\frac{m\pi z}{Z}\right)\cos\left(\frac{n\pi z}{Z}\right) = \frac{Z}{2}\delta_{mn}, \qquad \text{(C.3.43)}$$

write

$$(e_n, {}_l\hat{e}_j) = \begin{cases} \frac{Z}{2} \, {}^{l}_{j}G\left(\frac{n}{\underline{h}}\right), & \begin{cases} \text{if } n = j + 2\underline{h}m, & m = 0, 1, 2, \ldots \\ \text{or} & \\ \text{if } n = -j + 2\underline{h}m, & m = 1, 2, 3, \ldots \end{cases} \\ 0, & \text{otherwise.} \end{cases} \qquad \text{(C.3.44)}$$

Note that equation C.3.44 contains no explicit reference to $\underline{h}$, only to the ratios $j/\underline{h}$ and $n/\underline{h}$. Equation C.3.34 demonstrates that the ${}^{l}_{j}g(y)$ (and therefore the ${}^{l}_{j}G(2S)$) are determined entirely by $j/\underline{h}$. Assuming that the ${}^{l}_{j}G(2S)$ are fairly well behaved, it is reasonable to expect that if the inner products $(e_n, {}_l\hat{e}_j)$ are computed for a convenient value of $\underline{h}$, the inner products for a higher value of $\underline{h}$ will simply interpolate the conveniently computed inner products. This is as expected, since changing the number of elements does not change the nature of the periodicity of the mesh. Non-zero values of the matrix $(e_n, {}_l\hat{e}_j)$ are depicted in figure 4.18.

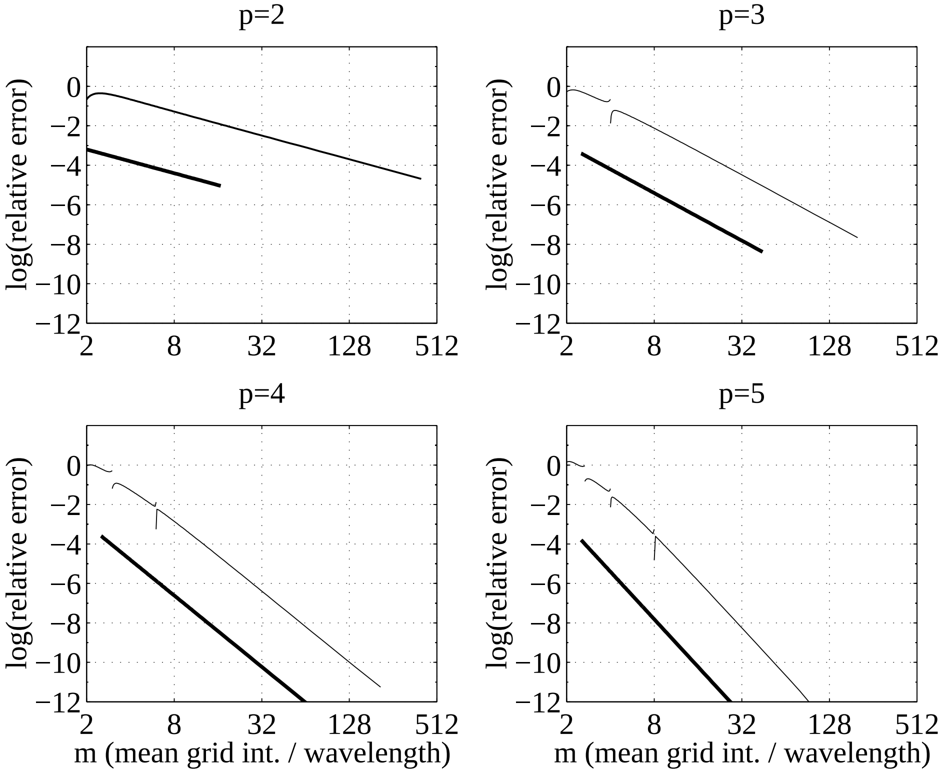


Figure 4.16: Difference of FEM and exact eigenvalues (also see figure 4.15) for several element orders. The upper fine lines are the eigenvalue error normalized by the exact eigenvalue to yield relative errors. The m axes are mean grid intervals per wavelength of the exact solution (which is a cosine). The fact that the eigenvalue error does not depend on the number of elements is a result of the periodicity of the mesh (see appendix C). Since the Chebyshev grid intervals are irregular, the mean grid interval is used. The slopes of the bold lines (not the actual values) are asymptotic predictions of the error from equation 4.6.2.35 and are proportional to $m^{-2(p-1)}$. The excursions of the error away from the main trend are due to anomalous modes at the ends of the dispersion branches. Special care has been taken in the evaluation of these eigenvalues so that the plot shows the maxiumum excursion from the main trend.

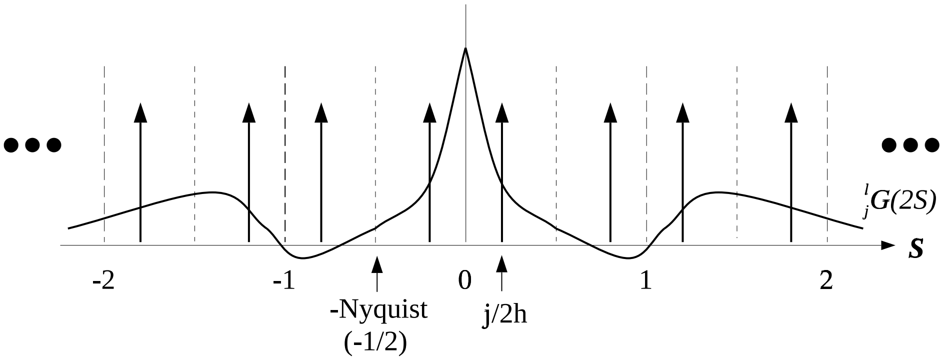


Figure 4.17: A schematic of the operation of equation C.3.41, which defines the spectrum of the FEM eigenfunction ${}_l\hat{\mathrm{E}}_j(S)$. Each pair of delta functions symmetric about the origin defines a cosine, whose amplitude is given by the value of the element spectrum $G(2S)$ at that (positive or negative) S. As the number of elements is increased, the above plot remains fixed, but $G(2S)$ is interpolated more finely, since $0 \leq j \leq \underline{h}$. $G(2S)$ does not correspond to any actual element spectrum, but is merely meant to demonstrate a real, even function.

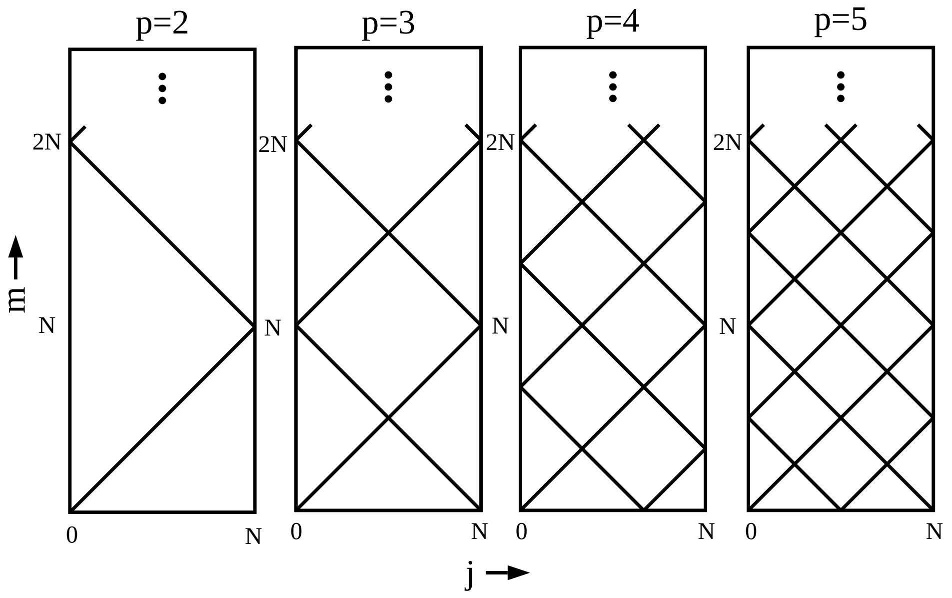


Figure 4.18: Non-zero entries in the matrix $(e_m, \hat{e}_j)$. These values all peak on the first diagonal $j = m$ and decay at large m as m^{-2}. The value $N + 1$ is the number of degrees of freedom in the FEM space: $N = \underline{h}(p-1)$. m, the cosine index, continues to infinity.

D Appendix: Convergence of the FEM

Estimates and conditions of the convergence of finite element methods associated with bilinear forms are given in Babuška (1971). These estimates are not difficult to prove, and stem from the condition that the linear problem is both 1 : 1 and *onto*.

Theorem D.1 *Let the bilinear form B on the Hilbert spaces $H_1 \times H_2$ satisfy:*

$$(i) |B(u,v)| \leq C_1 \|u\|_{H_1} \, \|v\|_{H_2}, \ \forall u \in H_1, \forall v \in H_2, C_1 < \infty,$$

$$(ii) \sup_{\substack{v \in H_2 \\ v \neq 0}} |B(u,v)| / \|v\|_{H_2} \geq C_2 \|u\|_{H_1}, \ \forall u \in H_1, C_2 > 0,$$

$$(iii) \sup_{\substack{u \in H_1 \\ u \neq 0}} |B(u,v)| / \|u\|_{H_1} \geq C \|v\|_{H_2}, \ \forall v \in H_2, C > 0.$$

Then, for $f \in H_2^*$, a given linear functional on H_2, there exists a unique $u_0 \in H_1$ such that

$$B(u_0, v) = f(v), \forall v \in H_2,$$

and

$$\|u_0\|_{H_1} \leq \frac{1}{C_2} \|f\|_{H_2^*}.$$

Any B satisfying condition i) is bounded or continuous. If B satisfies conditions ii) and iii), it is referred to as *coercive.*

Theorem D.2 *Let the assumptions of Theorem D.1 be satisfied. Given the closed subspaces $(V_1 \in H_1, V_2 \in H_2)$, if*

$$(i) \sup_{\substack{v \in V_2 \\ v \neq 0}} |B(u,v)| / \|v\|_{V_2} \geq \hat{C}_2 \|u\|_{V_1}, \forall u \in V_1, \hat{C}_2 > 0,$$

$$(ii) \sup_{\substack{u \in V_1 \\ u \neq 0}} |B(u,v)| / \|u\|_{V_1} \geq \hat{C} \|v\|_{V_2}, \forall v \in V_2, \hat{C} > 0.$$

then there exists a unique $\hat{u}_0 \in V_1$ such that

$$B(\hat{u}_0, v) = f(v), \forall v \in V_2,$$

and

$$\|u_0 - \hat{u}_0\|_{H_1} \leq (1 + \frac{C_1}{\hat{C}_2}) \inf_{w \in V_1} \|w - u_0\|_{H_1}.$$

The last quantity, $\inf\limits_{w \in V_1} \|w - u_0\|_{H_1}$, is the "best approximation error". The source of this estimate is an obvious variation on the derivation of equation 4.6.2.27. The spaces V_1 and V_2 will be FEM spaces of finite degrees of freedom, in our case $H^1_{h,p}$, $H^2_{h,p}$ and $H^0_{h,p}$.

E Appendix: The constant $\hat{C}_2$

With reference to appendix D, the most important quantity to establish is the constant $\hat{C}_2$. Knowledge of this constant as a function of our FEM parameterization enables us to bound the maximum error relative to the best approximation of the solution.

In general, we assume that the elements $u \in V_1$ and $v \in V_2$ are parameterized by complete sets $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, $i = 1, \ldots, N$. These will be the local FEM basis functions themselves or some properly chosen set of eigenfunctions on the FEM spaces. We write $u = \sum\limits_{i=1}^{N} u_i \boldsymbol{u}_i$ and $v = \sum\limits_{i=1}^{N} v_i \boldsymbol{v}_i$, where N is the number of degrees of freedom in these spaces. In accordance with appendix D, the norms on these spaces are denoted $\|u\|_{V_1}$ and $\|v\|_{V_2}$. The necessary quantities in the computation of $\hat{C}_2$ can be written

$$\|u\|^2_{V_1} = (u, u)_{V_1} = \sum_i \sum_j \overline{u}_i u_j (\boldsymbol{u}_i, \boldsymbol{u}_j)_{V_1}, \tag{E.0.1}$$

$$\|v\|^2_{V_2} = (v, v)_{V_2} = \sum_i \sum_j \overline{v}_i v_j (\boldsymbol{v}_i, \boldsymbol{v}_j)_{V_2}, \tag{E.0.2}$$

and

$$B(u, v) = (u, v)_B = \sum_i \sum_j \overline{v}_i u_j (\boldsymbol{v}_i, \boldsymbol{u}_j)_B. \tag{E.0.3}$$

We will also write these in terms of the matrices U, V, and B as

$$\|u\|_{V_1}^2 = u^H U u, \quad \|u\|_{V_2}^2 = v^H V v, \quad \& \quad B(u,v) = v^H B u. \tag{E.0.4}$$

The notation u^H is the Hilbert transpose, or the complex conjugate transpose of the vector u. Due to the nature of each operator, matrices U and V are both real.

An equivalent expression to the condition given in appendix D is

$$\inf_{\substack{u \in V_1 \\ \|u\|_{V_1}=1}} \sup_{\substack{v \in V_2 \\ v \neq 0}} |B(u,v)|^2 / \|v\|_{V_2}^2 = \hat{C}_2^2. \tag{E.0.5}$$

We establish the value of $\hat{C}_2$ by the method of Lagrange multipliers. The quantity $|B(u,v)|^2/\|v\|_{V_2}^2$ possesses all of its derivatives when $v \neq 0$, and we therefore write

$$D_x \left(|B(u,v)|^2 / \|v\|_{V_2}^2 \right) + \beta D_x \left(\|u\|_{V_1}^2 \right) = 0, \tag{E.0.6}$$

where β is a Lagrange multiplier for the single constraint that $\|u\|_{V_1} = 1$. The derivatives are to be taken with respect to the real and imaginary parts of all coefficients of u and v. In other words, x is to range over $\mathcal{R}e(u_i)$, $\mathcal{I}m(u_i)$, $\mathcal{R}e(v_i)$ and $\mathcal{I}m(v_i)$ for all i. Performing all of these derivatives is straightforward and yields two expressions. After elimination of the Lagrange multiplier using the fact that $\|u\|_{V_1} = 1$, the vanishing of the variations with respect to the components of u yields

$$B^H v = \overline{B(u,v)} U u. \tag{E.0.7}$$

Likewise, the vanishing of variations with respect to the components of v reveals that

$$\|v\|_{V_2}^2 B u = B(u,v) V v. \tag{E.0.8}$$

Solving equation E.0.7 for v and substituting into equation E.0.8, we obtain

$$u = \lambda B^{-1} V \left(B^{-1} \right)^H U u, \tag{E.0.9}$$

where

$$\lambda = \frac{|B(u,v)|^2}{\|v\|_{V_2}^2}. \tag{E.0.10}$$

We have used the fact that $\left(B^{-1}\right)^H = \left(B^H\right)^{-1}$. The Cholesky decomposition of the positive definite matrix U as $C^T C$ reveals that this is indeed an eigenvalue problem for a positive definite Hermitian matrix:

$$\tilde{u} = \lambda C B^{-1} V \left(B^{-1}\right)^H C^T \tilde{u}, \tag{E.0.11}$$

where $\tilde{u} = Cu$. Therefore the eigenvalue λ must be real and positive. Each eigenvalue λ_i, $i = 1, \ldots, N$, corresponds to a possible value of $\hat{C}_2^2$. In fact, we need only choose the minimum. We therefore arrive at the succinct expression

$$\hat{C}_2^2 = \min_{i=1,\ldots,N} \lambda_i. \tag{E.0.12}$$

Although the eigenproblem of equation E.0.9 in general would appear to yield little analytical insight into the behavior of $\hat{C}_2$, in some cases it reduces to a very simple expression. In particular, consider the case in which $V_1 = V_2 = H^1_{h,p}$ and the operator B is given by equation 4.6.1.5. Let $\{k_n^2, e_n\}, k_n^2 \in \mathbb{R}, e_n \in H^1_{h,p}, n = 1, \ldots, N$ such that $(D_z u, D_z e_n) = k_n^2(h,p)\,(u, e_n)$, $\forall u \in H^1_{h,p}$. Also normalize e_j so that $(e_j, e_j) = 1$. These (generalized) eigenvectors form a complete, orthonormal basis of $H^1_{h,p}$ since $(D_z u, D_z v)$ is a self-adjoint matrix. Now proceed with the above determination of $\hat{C}_2$, with $\boldsymbol{u}_i = \boldsymbol{v}_i = e_i$. Equations E.0.1 - E.0.3 become

$$\|u\|_1^2 = \sum_{j=1}^{N} \left(k^2 \overline{u}_j u_j + k_j^2 \overline{u}_j u_j\right), \tag{E.0.13}$$

$$\|v\|_2^2 = \sum_{j=1}^{N} \left(k^2 \overline{v}_j v_j + k_j^2 \overline{v}_j v_j\right), \tag{E.0.14}$$

and

$$B(u,v) = \sum_{j=1}^{N} \left(k^2 \overline{v}_j u_j - (1 - i\epsilon) k_j^2 \overline{v}_j u_j\right). \tag{E.0.15}$$

The matrices U,V and B become $U = k^2 I + \Lambda$, $V = k^2 I + \Lambda$, and $B = k^2 I - (1 - i\epsilon)\Lambda$, where I is the identity matrix and Λ is the matrix with the eigenvalues k_i^2 on its diagonal. Equation E.0.9 then gives directly that

$$\hat{C}_2^2 = \min_{i=1,\ldots,N} \frac{|k^2 - (1 - i\epsilon)k_i^2|^2}{|k^2 + k_i^2|^2}. \tag{E.0.16}$$

F Appendix: Integrals of products of Chebyshev polynomials and their derivatives

The numerical tests depend on the evaluation of integrals of various products of Chebyshev polynomials. All of the integrals used in the analysis of the H^1 and H^0 FEM are given below.

We assume that the moduli and density are expanded in Chebyshev polynomials so that the required integrals result in triple products of Chebyshev polynomials and their derivatives. The mass matrix results in integrals of the form

$$\int_{-1}^{1} dz T_l(z) T_m(z) T_n(z) = \begin{cases} \frac{1}{2}\Big[\frac{1}{1-(n+m+l)^2} + \frac{1}{1-(l-n-m)^2} \\ + \frac{1}{1-(n-m+l)^2} + \frac{1}{1-(l-n+m)^2}\Big], & n+m+l \text{ even} \\ 0, & n+m+l \text{ odd} \end{cases}$$

This integral is given in Korczak & Patera (1986).

The stiffness matrix results in the integral

$$\int_{-1}^{1} dz T_l(z) T_m'(z) T_n'(z) = \begin{cases} nm\left[\alpha(n+m+l) + \alpha(l-n-m)\right. \\ \left. -\alpha(n-m+l) - \alpha(l-n+m)\right], & n+m+l \text{ even} \\ 0, & n+m+l \text{ odd} \end{cases}$$

where

$$\alpha(j) = \sum_{k=1}^{|j|/2} \frac{1}{2k-1},$$

and $'$ denotes differentiation with respect to z. The stiffness matrix for the $H^0(\Omega)$ FEM can be written in terms of the above stiffness matrix after an integration by parts.

The inner product on H^2 results in the integral

$$\int_{-1}^{1} dz T_n'' T_m'' = \begin{cases} \frac{nm+n^2m^2}{4}\beta(n-m) - \frac{nm-n^2m^2}{4}\beta(n+m) & \\ -\frac{(n-n^2)(m-m^2)}{8}\beta(n+m+2) & \\ -\frac{(n+n^2)(m+m^2)}{8}\beta(n+m-2) & \\ +\frac{(n-n^2)(m+m^2)}{8}\beta(n-m+2) & \\ +\frac{(n+n^2)(m-m^2)}{8}\beta(n-m-2), & n+m \text{ even} \\ 0, & n+m \text{ odd} \end{cases},$$

where

$$\beta(k) = -\frac{9 - 10k^2 + k^4}{6}\alpha(k).$$

This integral is considerable work to evaluate. The author utilized Mathematica to piece together a solution.

G Appendix: Mathematical symbols and abbreviations

Symbol	comments
inf	greatest lower bound; for a finite set, the minimum (see Dieudonné (1960), section 2.3)
sup	least upper bound; for a finite set, the maximum (see Dieudonné (1960), section 2.3)
$[\cdot]$	integer part
$\overline{a}$	the complex conjugate of a
$\in$	a member of
$\exists$	there exists
$\Omega, \partial\Omega$	a domain and its boundary
$\mathbb{N}$	the natural numbers (positive and negative integers)
$\forall$	for all
$\subset$	subset of
∂	partial derivative
$\vert$	such that or evaluated on
$\{\}$	set definition
L^*	the adjoint of L
$a : b \rightarrow c$	definition of the mapping a
$\cup$	union of sets
$\vert \cdot \vert$	absolute value or modulus
$\propto$	proportional to
$\Vert\cdot\Vert_s$	norm on the Sobolev space H^s
$\vert\vert\vert\cdot\vert\vert\vert$	scaled norm
$(\cdot,\cdot)$	L^2 inner product on a domain
$\langle\cdot,\cdot\rangle$	L^2 inner product on the boundary of a domain
$\langle\langle\cdot\rangle\rangle_s$	boundary norm
h	size of a finite element
$\underline{h}$	number of finite elements
FLOP	floating point operation
$\prod$	product
u^H	the Hermitian transpose
$\mathbb{R}$	the real numbers

4.11 References

Adams, R.A. 1975 *Sobolev Spaces.* Academic Press, New York.

Al-Gwaiz, M.A. 1992 *Theory of Distributions.* Marcel Dekker, New York.

Anderson, E., Bai, Z., Bischof, C., Demmel, J., Dongarra, J., DuCroz, J., Greenbaum, A., Hammarling, S., McKenney, A., Ostrouchov, S., Sorenson, D. 1992 *LAPACK User's Guide.* SIAM, Philadelphia.

Aubin, Jean-Pierre 1972 *Approximation of Elliptic Boundary-Value Problems.* Wiley-Interscience, New York.

Babuška, I. & Rheinboldt, W.C. 1978 Error Estimates for Adaptive Finite Element Computations. *SIAM J. Numer. Anal.* **15,4**, 736.

Babuška, I. 1971 Error-Bounds for Finite Element Method. *Numer. Math.* **16**, 322-333.

Babuška, I. & Aziz, A.K. 1972 Survey lectures on the mathematical foundations of the finite element method. *in, The mathematical foundations of the finite element method with applications to partial differential equations. Academic Press, New York* , 1-359.

Backus, G.E. & Mulcahy, M. 1976 Moment tensors and other phenomenological descriptions of seismic sources - II. discontinuous displacements . *Geophys. J. R. Astr. Soc.* **47**, 301-329.

Bracewell, R.N. 1986 *The Fourier Transform and Its Applications.* McGraw-Hill, New York.

Carey, Graham F. & Oden, J. Tinsley 1984 *Finite Elements, Volume III, Computational Aspects.* Prentice-Hall, New Jersey.

Červený, V., Popov, M., Pšenčík, I. 1982 Computation of wave fields in inhomogeneous media – Gaussian beam approach. *Geophysical Journal of the Royal Astronomical Society* **70**, 109 –128.

Davis, P.J., & Rabinowitz, P. 1967 *Numerical Integration.* Blaisdell Publishing Co., Waltham, MA.

Demkowicz, L. & Oden, J.T. 1994 Recent progress on application of *hp* -adaptive BE/FE methods to elastic scattering. *International Journal for Numerical Methods in Engineering* **37**, 2893-2910.

Dieudonné, J. 1960 *Foundations of Modern Analysis.* Academic Press, New York.

Douglas, J. & Dupont, T. 1977 H^{-1} -Galerkin methods for problems involving several space variables. In: *Topics in Numerical Analysis III, John J. Miller, ed., Academic Press, New York*, pp.125 – 141.

Douglas, J. & Dupont, T. 1979 A family of C^1 finite elements with optimal approximation properties for various Galerkin methods for 2nd and 4th order problems. *R.A.I.R.O, Analyse Numérique* **13, No. 3**, 227 – 255.

Dubiner, M. 1991 Spectral methods on triangles and other domains. *Journal of Scientific Computing* **6,** No. 4, 345 – 390.

Chapman, C. & Drummond, R. 1982 Body-wave seismograms in inhomogeneous media using Maslov asymptotic theory. *Bull. Seism. Soc. Am.* **72**, 227 – 317.

Dougherty, M. & Stephen, R. 1991 Seismo/acoustic propagation through rough seafloors. *Journal of the Acoustical Society of America* **90** , 5, 2637 – 2651.

Dougherty, M. & Stephen, R. 1988 Seismic energy partitioning in laterally heterogeneous ocean crust. *PAGEOPH* **128**, 195 – 229.

Fornberg, B. 1996 *A practical guide to pseudospectral methods.* Cambridge University Press.

Frazer, L. & Sinton, J. 1984 A Kirchoff method for the computation of finite-frequency body wave synthetic seismograms in laterally inhomogeneous media. *Geophysical Journal of the Royal Astronomical Society* **78**, 413 – 429.

Fried, Isaac 1979 *Numerical Solution of Differential Equations.* Academic Press, New York.

Geller, R.J. & Ohminato, T. 1994 Computation of synthetic seismograms and their partial derivatives for heterogeneous media with arbitrary natural boundary conditions using the Direct Solution Method. *Geophysical Journal International* **116**, 421-446.

Cummins, P.R., Geller, R.J. & Takeuchi, N. 1994 DSM complete synthetic seismograms: P-SV, spherically symmetric, case. *Geophysical Research Letters* **21,** 15, 1663-1666.

Cummins, P.R., Geller, R.J., Hatori, T. & Takeuchi, N. 1994 DSM complete synthetic seismograms: SH, spherically symmetric, case. *Geophysical Research Letters* **21,** 7, 533-536.

Geller, R.J. & Takeuchi, N. 1995 A new method for computing highly accurate DSM synthetic seismograms. *Geophysical Journal International* **123**, 449-470.

Geller, R., Igel, H., Cummins, P. & Romanowicz, B. 1996 Comparison and validation of techniques for computing synthetic seismograms for 3-D global Earth models. *Supplement, EOS Transactions AGU* **77, No. 46**, 490.

Gottlieb, D., Hussaini, M.Y., and Orszag, S.A. 1984 Theory and Applications of Spectral Methods. *Spectral Methods for Partial Differential Equations* **SIAM, Philadelphia**, 1-54.

Huang, X., Kendall, J-M., Thomson, C. & West, G. 1996 A comparison of the Maslov integral seismogram and the finite difference method. *Geophysical J. Int.*, submitted.

Hudson, J. & Heritage, J. 1981 The use of the Born approximation in seismic scattering problems. *Geophysical J. R. astr Soc.* **66, 1**, 221-240.

Hung, S., Forsyth, D., Fischer, K. & Parmentier, E. 1996 Modeling anisotropic wave propagation in oceanic lithosphere using the multi-domain pseudo-spectral method. *Supplement, EOS Transactions AGU* **77, No. 46**, 525.

Igel, H. & Gudmundsson, O. 1996 Frequency-dependent effects on travel times and waveforms of long-period S waves: implications for the scale of mantle heterogeneity. *Supplement, EOS Transactions AGU* **77, No. 46**, 474.

Igel, H. & Weber, M. 1995 SH wave propagation in the whole mantle using high-order finite differences. *Geophysical Research Letters* **22** , 6, 731-734.

Ihlenburg, F. & Babuška, I. 1995 Dispersion analysis and error estimation of Galerkin finite element methods for the Helmholtz equation. *International Journal for Numerical Methods in Engineering* **38**, 3745-3774.

Korczak, K.Z. & Patera, A.T. 1986 An Isoparametric Spectral Element Method for Solution of the Navier-Stokes Equations in Complex Geometry. *Journal of Computational Physics* **62**, 361-382.

Kumari, G., Singh, S.J., & Singh, K. 1992 Static deformation of two welded elastic half-spaces caused by a point dislocation source. *Physics of the Earth and Planetary Interiors* **73**, 53-76.

Li, X. & Romanowicz, B. 1996 Global mantle shear velocity model developed using nonlinear asymptotic coupling theory. *JGR* **101, 10**, 22245-22272.

Liu, H., Anderson, D.L. & Kanamori, H. 1976 Velocity dispersion due to anelasticity; implications for seismology and mantle composition. *Geophys. J. R. Astr. Soc.* **47**, 41 – 58.

Marfurt, Kurt J. 1984 Accuracy of finite-difference and finite-element modeling of the scalar and elastic wave equation. *Geophysics* **49,**No. 5, 533 – 549.

Marquering, H. & Snieder, R. 1995 Surface-wave mode coupling for efficient forward modelling and inversion of body-wave phases. *Geophys. J. Int.* **120**, 186 – 208.

Mavripilis, C. & van Rosendale, J. 1993 Triangular spectral elements for incompressible fluid flow. *ICASE Report 93-100, NASA Langley Research Center, Hampton, VA..*

Park, J.J. 1985 *Application of the Galerkin Formalism in the Coupling of the Earth's Free Oscillations.* Dissertation, Scripps Institution of Oceanography, UCSD.

Park, J. & Gilbert, F. 1986 Coupled free oscillations of an aspherical, dissipative, rotating earth: Galerkin theory. *Journal of Geophysical Research* **91** , 7, 7241 – 7260.

Patera, A.T. 1984 A Spectral Element Method for Fluid Dynamics: Laminar Flow in a Channel Expansion. *Journal of Computational Physics* **54**, 468-488.

Rachford, H. & Wheeler, M. 1974 An $H^{-1}(\Omega)$-Galerkin procedure for the two-point boundary value problem. In: *Mathematical aspects of finite elements in partial differential equations, Carl de Boor, ed., Academic Press, New York*, pp.353–382.

Roitberg, Yakov 1996 *Elliptic Boundary Value Problems in the Spaces of Distributions.* Kluwer Academic Publishers, Dordrecht.

Sherwin, S. & Karniadakis, G. 1996 Tetrahedral *hp* finite elements: Algorithms and flow simulations. *Journal of Computational Physics* **124**, 14 – 45.

Sherwin, S. & Karniadakis, G. 1995 A new triangular and tetrahedral basis for high-order (*hp*) finite element methods. *International Journal for Numerical Methods in Engineering* **38**, 3775 – 3802.

Sherwin, S. & Karniadakis, G. 1995 A triangular spectral element method; application to the incompressible Navier-Stokes equations. *Computer Methods in Applied Mechanics and Engineering* **123**, 189 – 229.

Sneider, R. & Romanowicz, B. 1988 A new formalism for the effect of lateral heterogeneity on normal modes and surface waves - I: isotropic perturbations, perturbations of interfaces and gravitational perturbations. *Geophysical Journal* **92**, 207 – 222.

Strang, G. & Fix, G. 1973 *An Analysis of the Finite Element Method.* Prentice Hall, Englewood Cliffs, N.J..

Virieux, J. 1986 P-SV wave propagation in heterogeneous media: velocity-stress finite-difference method. *Geophysics* **51** , 4, 889-901.

Wang, Z. & Tromp, J. 1996 Modeling wave propagation in complex media: a pseudo-spectral approach. *Supplement, EOS Transactions AGU* **77, No. 46**, 523.

Wilcock, S., Dougherty, M., Solomon, S., Purdy, G. & Toomey, D. 1993 Seismic propagation across the East Pacific Rise: finite difference experiments and implications for seismic tomography. *Journal of Geophysical Research* **98** , 11, 19913 – 19932.

Woodhouse, J. & Dziewonski, A. 1984 Mapping the upper mantle: three-dimensional modeling of earth structure by inversion of seismic waveforms. *Journal of Geophysical Research* **89** , 7, 5953 – 5986.