

# An Automatic, Adaptive Algorithm for Refining Phase Picks in Large Seismic Data Sets

by C. A. Rowe,\* R. C. Aster, B. Borchers, and C. J. Young

**Abstract** We have developed an adaptive, automatic, correlation- and clustering-based method for greatly reducing the degree of picking inconsistency in large, digital seismic catalogs and for quantifying similarity within, and discriminating among, clusters of disparate waveform families. Innovations in the technique include (1) the use of eigenspectral methods for cross-spectral phase estimation and for providing subsample pick lag error estimates in units of time, as opposed to dimensionless relative scaling of uncertainties; (2) adaptive, cross-coherency-based filtering; and (3) a hierarchical waveform stack correlation method for adjusting mean intercluster pick times without compromising tight intracluster relative pick estimates. To solve the systems of cross-correlation lags we apply an iterative, optimized conjugate gradient technique that minimizes an  $L_1$ -norm misfit. Our repicking technique not only provides robust similarity classification—event discrimination without making a priori assumptions regarding waveform similarity as a function of preliminary hypocenter estimates, but also facilitates high-resolution relocation of seismic sources. Although knowledgeable user input is needed initially to establish run-time parameters, significant improvement in pick consistency and waveform-based event classification may be obtained by then allowing the programs to operate automatically on the data. The process shows promise for enhancing catalog reliability while at the same time reducing analyst workload, although careful assessment of the automatic results is still important.

## Introduction

Earthquake location and many other travel-time-based seismological applications historically have depended critically on the ability of human analysts to estimate arrival times of body waves. Standard network operations generally involve the manual measuring of  $P$ -wave and  $S$ -wave arrivals or, more recently, computer identification of these phases using software autopickers. The most common human or computer picking approach is done one event at a time, with records from several or all recording stations. These methods, although well suited for the near-real-time processing demands of network operation, do not necessarily produce consistent phase arrival times (picks), because path effects, signal-to-noise conditions, and source radiation pattern differences within the network of receivers may be large. Picks obtained from the resulting heterogeneous suite of waveforms may thus be highly inconsistent for even very similar events. These inconsistent picks are then used to calculate the hypocenter, and the event is archived. Seldom are any

but the most egregious picking errors noted and corrected prior to moving on to the next earthquake; hence, picking inconsistencies between similar events remain unresolved.

Such routine network operations have produced very large sets of hypocenter locations (e.g., the online SCEC database yields more than 430,000 southern California earthquakes having more than 28 million picks from 1981–2001) with location error estimates of a few to a few 10s of kilometers for regional-scale networks with good azimuthal coverage. Standard catalog locations have served to document general seismicity levels and the gross geometry of comparably scaled seismogenic features; however, within the diffuse scatter of locations that result from picking inconsistencies, fine details remain unresolved. Significant improvement in the precision of hypocenter location and the resulting delineation of the details of seismic source regions has sometimes been achieved through careful, painstaking visual cross-correlation and repicking of phases for preliminarily located events (e.g., Phillips *et al.*, 1997; Phillips, 2000), but this is generally a time- and cost-prohibitive undertaking that will necessarily be limited to small, focused subsets of the larger catalogs.

---

\*Present address: Department of Geology and Geophysics, University of Wisconsin–Madison, 1215 W. Dayton St., Madison, Wisconsin 53706 (char@geology.wisc.edu).

Quantitative, waveform-correlation-based phase repicking and relative doublet and multiplet relocations have produced some impressive resolution of seismogenic structures within families of similar events. Fremont and Malone (1987) used relative relocations based on cross-correlation lags of multiplets at Mount St. Helens to delineate source regions on the order of a few 10s of meters. Deichmann and Garcia-Fernandez (1992) used cross-correlation methods to identify relative arrival time differences among similar Alpine events for precise relative location. Got *et al.* (1994) relocated seismicity at Kilauea volcano, Hawaii, using multiplets chosen based on cross-spectral coherency. Slunga *et al.* (1995) used relative arrival times calculated from Fourier-interpolated cross-correlation functions to determine precise relative locations and improved absolute locations in clusters of similar microearthquakes in Iceland by incorporating the cross-correlation lags into a modified joint hypocentral determination (JHD) application. Gillard *et al.* (1998) used cross-correlation methods and multiplet analysis at Kilauea, Hawaii, to reduce quasilinear ‘cigars’ of microearthquake foci into precise relative relocations delineating pencil-thin lines of seismicity. Rubin *et al.* (1999) applied similar methods to identify tight multiplets on the Hayward Fault in California.

Dodge *et al.* (1995) developed an automatic, computer-based correlation approach which calculates individual pick-time inconsistencies for event pairs and uses the weighted lag constraints to adjust for consistent picks. This approach is a significant improvement over earlier efforts that rely on a master event method, in that master-event bias is reduced, highly similar multiplets are not required, and dissimilar event pairs will have limited influence over the pick adjustments. Shearer (1997) demonstrated significant improvement in delineating the seismogenic features associated with the Whittier Narrows aftershock sequence in California by also invoking a pick lag estimation method, after first segregating events into groups meeting minimum cross-correlation criteria (e.g., Aster and Scott, 1993). The above methods, although successively more quantitative and efficient, still rely to some significant degree on user interaction and have been so far applied to specific studies of catalog subsets chosen either through spatial restrictions or genetic assumptions (e.g., a limited box of data, a specific aftershock sequence), or they operate by preliminary exclusion of individual events failing to meet very high ( $\sim 0.9$ ) cross-correlation criteria. A certain a priori selection to ensure correlatability has therefore been invoked. We describe a method that combines the most desirable features of techniques already available with additional adaptability and portability, in an automatic package that can be implemented for large catalogs such that a wide variety of applications may be addressed with little time-consuming customization. Further, as with the Dodge *et al.* (1995) or Shearer (1997) approaches, our method provides corrected picks. These may be used not only for event relocation, using either standard single-event location methods or more sophisticated

joint location methods such as JHD (e.g., Pujol, 1992), JHD-collapsing (e.g., Fehler *et al.*, 2000) or the Hypo-DD technique (Waldhauser and Ellsworth, 2000), but are available also for other applications such as seismic tomography (e.g., Kissling, 1988).

We first outline the technique with a discussion of catalog segregation (clustering) and identification of similar event families. This is followed by a description of the signal-processing tools in our algorithm and an explanation of how we apply them to the data. We then outline the final calculation of lags and standard errors from the interevent constraints.

## Technique

Our method may be outlined as follows:

- Grooming the catalog
- Preliminary cross-correlation for waveform similarity matrix
- Clustering catalog based on waveform similarity
- Adaptive window-length cross-correlation within clusters
- Solving for consistent pick lags within clusters
- Stacking re-aligned waveforms within clusters
- Cross-correlating stacks to obtain intercluster pick adjustments.

### Preliminary Data Organization

Prior to processing, waveforms have preliminary *P* or *S* picks (or both) produced by an analyst or autopicker. These parameters are read from the trace headers, along with other potentially useful parametric data such as preliminary hypocenter coordinates.

As with other quantitative waveform correlation methods for precise earthquake relocation, we simultaneously analyze traces from many events on a station-by-station basis (e.g., Dodge *et al.*, 1995; Shearer, 1997; Waldhauser and Ellsworth, 2000). This data regrouping allows one to improve the consistency of pick times among similar events by exploiting waveform resemblance for similar source-receiver raypaths.

### Clustering

Once the data are properly formatted, we begin by obtaining preliminary cross-correlation values to divide the catalog into clusters of highly similar events. We have found that using cross-correlation results to adjust picks in a heterogeneous catalog provides unsatisfactory results by making inappropriate comparisons among inconsistent waveforms. Downweighting relative lags based on preliminary interhypocentral distance may incorrectly associate or dissociate constraints in the case of mislocated events. Solving the matrix of first differences to obtain consistent pick adjustments may be compromised by effectively zeroing some constraints without adjusting degrees of freedom appropri-

ately. By decoupling the system into highly similar subgroups, each can be solved for consistency among closely related waveform correlation lags (Rowe, 2000).

This subdivision, or clustering (discussed in detail below), is followed by relative lag estimation among member traces within each cluster. Individual picks are adjusted and the realigned waveforms are stacked to provide a composite representative waveform for the cluster. These stacks may then be cross-correlated to obtain optimal pick adjustments between clusters, providing improved intercluster as well as intracluster hypocentral relationships. We will outline details of these steps later in this article.

Many candidate approaches for similarity clustering exist. Among those used successfully in other seismological applications are signal envelope cross-correlation (Carr *et al.*, 1999a,b), sonogram pattern recognition (Joswig, 1995) (sometimes referred to as *spectral fingerprinting*), and multistation median waveform cross-correlation (e.g., Aster and Scott, 1993). Lees (1998) applied equivalence class analysis to events that had been segregated by preliminary hypocenter location. We have chosen the waveform cross-correlation coefficient for all interevent pairs as our catalog clustering criterion. The clustering may be performed on correlation values for a single station, the median correlation value for a suite of stations, or some other criterion best suited to the catalog being evaluated. Performing clustering early in the analysis boosts the efficiency of our technique, as time and memory requirements for correlation and lag estimation on resulting subclusters decrease quadratically with the number of events.

An agglomerative, dendrogram-based hierarchical pair-group clustering algorithm (e.g., Lance and Williams, 1967; Sneath and Sokal, 1973; Ludwig and Reynolds, 1988) has been chosen for catalog segregation. Available clustering options include cluster centroid mean or median, single link, complete link, or flexible combinational weighting. Our technique was implemented following the flexible method in the MATLAB-based seismic analysis package, MATSEIS (Harris and Young, 1997; Young *et al.*, 2001). We outline the algorithm and its implementation for seismic waveform clustering later in this article.

Preliminary cross-correlation of  $N$  events yields an  $N \times N$  symmetric event similarity matrix  $\mathbf{M}$ , whose  $i,j$  entries represent the cross-correlation maxima for the  $i$ th and  $j$ th events. From this the event dissimilarity matrix  $\mathbf{K}$  is constructed:

$$\mathbf{K}_{i,j} = 1.001 - \mathbf{M}_{i,j}. \quad (1)$$

The  $\mathbf{K}_{i,j}$  may be viewed as a measure of interevent distance in waveform similarity space for events  $i$  and  $j$ , where a value of  $\sim 0$  equates with colocation and a value of  $\sim 1$  represents infinite similarity distance. The use of 1.001 rather than 1.0 in equation (1) eliminates divide-by-zero errors in the rare instances where cross-correlation maxima are 1.0 to machine

precision. The algorithm constructs a hierarchical structure of event similarity for all events (Fig. 1).

First, the two events  $i$  and  $j$  ( $i \neq j$ ) with the smallest  $\mathbf{K}$  value (equation 1) are fused. A new vector  $\mathbf{k}^1$  is constructed whose entries are weighted by the  $\alpha_1$ ,  $\alpha_2$ , and  $\beta$  coefficients ( $\gamma = 0$ ) described in Table 1 (Lance and Williams, 1967). For  $N$  events,

$$\text{for } m = 1:N \mathbf{k}_m^1 = \alpha_1 \mathbf{K}_{i,m} + \alpha_2 \mathbf{K}_{j,m} + \beta \mathbf{K}_{i,j}. \quad (2)$$

This vector is added as a new row and column (plus a dummy diagonal value) to  $\mathbf{K}$ , whose dimension is now  $N + 1$ . The two rows and two columns,  $i$  and  $j$ , are then annihilated and the matrix is thus reduced to a matrix  $\mathbf{K}^g$  of dimension  $N - 1$ , where  $g$  denotes the fusion step and  $g = 1:N - 1$ . The matrix is again searched for minimum distance, and individual events may continue to be grouped by equation (1) until the shortest distance is found to belong either to a cluster with an individual event, or two clusters.

### CLUSTERING SCHEMATIC

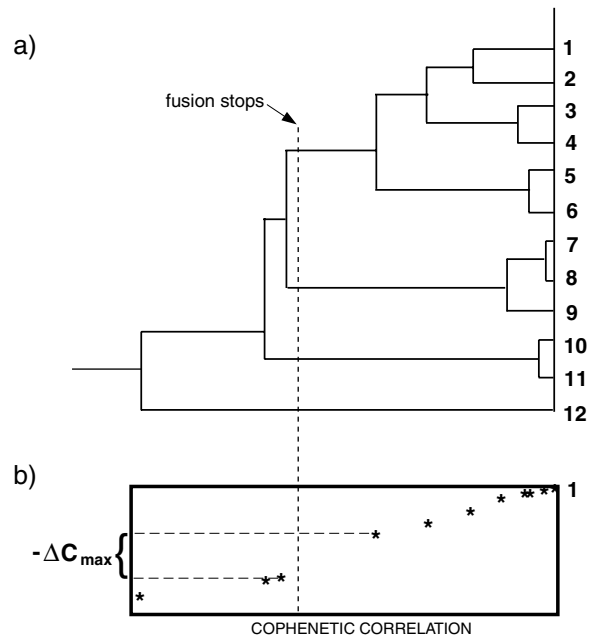


Figure 1. Schematic illustration of dendrogram-based, hierarchical pair-group clustering method. (a) The dendrogram is built beginning by selection of the most similar event pair. Its combination yields a new single cluster entity, which is then compared against all other events. Subsequent joinings may be between two individual events, one event and a pre-existing cluster, or between two clusters, depending on the values in the reduced similarity matrix. (b) The cophenetic correlation parameter of equation (4) is calculated with each fusion step. Fusion continues until all events have been associated; the retroactive segregation cutoff is chosen as the step prior to the greatest drop in cophenetic correlation value.

Table 1  
Cluster Combinational Weighting Parameters for Different Hierarchical Clustering Strategies

Strategy	$\alpha_1$	$\alpha_2$	$\beta$
Centroid (unweighted centroid)	$\frac{t(j)}{t(j,k)}$	$\frac{t(k)}{t(j,k)}$	$-\frac{t(j)t(k)}{t(j,k)}$
Centroid (weighted)/median	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$
Group mean/unweighted pair-grouping method	$\frac{t(j)}{t(j,k)}$	$\frac{t(k)}{t(j,k)}$	0
Flexible	0.625	0.625	-0.25

The number of entities in the  $j$ th and  $k$ th groups are represented by  $t(j)$  and  $t(k)$ , respectively, and the number of entities in the combined ( $j,k$ ) group is  $t(j,k)$ . (After Ludwig and Reynolds, 1988.)

The new combinational vector  $\mathbf{k}^g$  is derived using the linear combinational equation (Lance and Williams, 1967) of the form

$$\text{for } m = 1:N\mathbf{k}_m^g = \alpha_1 \mathbf{K}^{g-1}(i, h), m + \alpha_2 \mathbf{K}^{g-1}(j, h), m + \beta \mathbf{K}^{g-1}(i, j), \quad (3)$$

where the distance between the cluster  $\langle i,j \rangle$  and another entity  $\langle h \rangle$  may be computed from the known distances  $\mathbf{K}_{i,h}$  and  $\mathbf{K}_{j,h}$ , and the weighting parameters  $\alpha_1$ ,  $\alpha_2$ , and  $\beta$ . Note that  $\langle h \rangle$  may be an individual entity, or a previously joined cluster. These combinational steps are repeated, and the  $\mathbf{K}^g$  matrix reduced, until all events have been associated into a single group, requiring a total of  $N - 1$  cycles.

For the waveform similarity problem, we choose the *flexible* combinational weighting scheme. This weighting scheme has coefficients chosen so that the sum of the  $\alpha_1$ ,  $\alpha_2$ , and  $\beta$  parameters equals 1, which means that with successive joinings, the  $\mathbf{K}^g$  matrix values (distances between clusters) move monotonically in a single direction, either continually contracting or continually expanding, with no reversals of direction that would cause problems for an automated system. The value chosen for  $\beta$  can be shown (Sneath and Sokal, 1973; Ludwig and Reynolds, 1988) to govern the spatial relationships of the clustering hierarchy. When  $\beta$  approaches  $-1$  the system is dilated as in complete-linkage clustering schemes, whereas  $\beta \rightarrow 0$  contracts the space similar to single-linkage methods (Ludwig and Reynolds, 1988). In other words, after fusion the reconstructed matrix has distances that are, respectively, much greater or smaller than the original matrix. A choice near  $-0.25$  for  $\beta$  tends to minimize this distortion (e.g., Sneath and Sokal, 1973).

The final problem is determining at what point to terminate clustering in an automated manner. Carr *et al.* (1999a) have obtained good seismicity clustering results using the technique of cophenetic correlation (Sneath and Sokal, 1973), wherein the distances between elements in the

$\mathbf{K}^g$  matrix are compared at each clustering step with the event distances in the original  $\mathbf{K}$  matrix.

The cophenetic correlation for each pairwise combination (of entities or multievent clusters) is calculated after each fusion step using the original  $\mathbf{K}$  matrix and a cophenetic matrix  $\mathbf{K}^c$ . The cophenetic matrix begins as a duplicate of the original matrix, but with each successive grouping, all entries in  $\mathbf{K}^c$  associated with the clustering step (either individuals or all members of the clusters being addressed) are replaced with the current dissimilarity distance value. This reduces the similarity of  $\mathbf{K}^c$  to the original  $\mathbf{K}$  matrix. The  $g$ th value of the cophenetic correlation parameter (for the  $g$ th clustering step) is then

$$C_g = \frac{\sum_{j=1}^n \sum_{k=1}^n \mathbf{K}_{j,k}^c \mathbf{K}_{j,k}}{\left[ \sum_{j=1}^n \sum_{k=1}^n \mathbf{K}_{j,k} \sum_{j=1}^n \sum_{k=1}^n \mathbf{K}_{j,k}^c \mathbf{K}_{j,k} \right]^{1/2}}, \quad (4)$$

where  $j$  and  $k$  represent the entities being combined in the  $i$ th correlation step (either individuals or previously clustered groups). As larger groups are formed, the similarity between the  $\mathbf{K}^c$  matrix and the original  $\mathbf{K}$  matrix will continue to decrease as the original entries are replaced with the dissimilarity values calculated for the growing clusters. Overall values of  $C_g$  (equation 4) will thus decline, although the function decrease is not necessarily monotonic. We target the largest drop in cophenetic correlation as the point immediately before which fusion should stop, as this represents the transition where the greatest leap in disparity between  $\mathbf{K}^c$  and  $\mathbf{K}$  occurs.

We illustrate the clustering technique with a 12-object example dendrogram and cophenetic correlation function in Figure 1. Each joining on the dendrogram represents the identification of smallest distance in similarity space for the entries in the reduced matrix at each fusion step. In this example, objects 7 and 8 are most similar, and so are joined as a pair by equation (2). The matrix is reduced through annihilation of rows and columns 7 and 8, with a replacement row and column added whose entries are the new relationship of cluster  $\langle 7,8 \rangle$  to the remaining 10 members, governed by equation (2). We next join objects 10 and 11, then 5 and 6, then 3 and 4. The fifth combinational step finds the smallest distance in the reduced matrix to lie between cluster  $\langle 7,8 \rangle$  and object 9, so these are also joined under equation (3). The process continues until step 11, when all entities are united. The point at which to stop fusion is determined retroactively, using the maximum negative difference of the cophenetic correlation function,  $C_g$ . At each step we have calculated a value for the cophenetic correlation, displayed beneath the dendrogram in Figure 1. The greatest drop in cophenetic correlation occurs at the fusion of cluster  $\langle 5,6,3,4,1,2 \rangle$  with cluster  $\langle 7,8,9 \rangle$ , which implies that in the hierarchy of this dendrogram the greatest dissimilarity occurs at this fusion step. We therefore segregate the dataset

into cluster memberships as defined immediately prior to this fusion step, leaving three clusters,  $\langle 5,6,3,4,1,2 \rangle$ ,  $\langle 7,8,9 \rangle$ ,  $\langle 10,11 \rangle$ , and an orphan object  $\langle 12 \rangle$ , that is not very similar to any of the others. Other automatic decision-making techniques exist, such as comparing the inter- and intracluster variances (e.g., Ludwig and Reynolds, 1988). We find, however, that the cophenetic correlation seems well suited to catalogs that divide robustly into unrelated groups of distinct waveforms, such as the discrimination of mine blasts from different locations (e.g., Carr *et al.*, 1999a,b), or teleseisms from different source regions.

Efforts to apply the cophenetic correlation method to catalogs of earthquakes exhibiting continuous waveform variation, however, were less satisfactory (e.g., Rowe, 2000; Rowe *et al.*, 2002). Under such circumstances the cophenetic correlation function becomes erratic, and a derivative-based termination of fusion on such a function is unreliable. We have therefore modified the algorithm so that we may instead select a similarity threshold, below which fusion stops. We have found that using a threshold of 0.8 works well, although optimal results may vary from catalog to catalog. The threshold approach yields a large number of small (doublet and multiplet) similarity groups and may be overly aggressive in cases where large general earthquake families are present. The final cluster memberships under any segregation scheme depend strongly on the length of correlation window and the degree of filtering; hence, some interactive testing on a random catalog subset is advisable to determine such parameters.

Once a satisfactory division has been found, the catalog is separated into corresponding clusters, and individual phases within each cluster are cross-correlated to obtain relative pick lag estimates.

#### Relative Lag Estimation

Relative lag estimation between pairs of traces proceeds in two steps: a coarse discrete correlation step that provides an estimate of lag to the nearest time sample and a fine correlation step that provides a refinement to the subsample level. Among all events recorded at a given station, we compare each event pair for each phase ( $P,S$ ) to measure waveform similarity and to estimate lag. A user-specified  $M$ -sample time window, including a fractional prepick offset, is established about the preliminary picks for each pair of events to be compared. This window length is chosen based on sample rate and overall frequency content of the targeted phase. Generally speaking, two cycles of the dominant waveform is an acceptable length; however, the choice of correlation window length varies depending on the intended use of the correlation results. After cluster separation, intracluster cross-correlations are performed using a suite of correlation window lengths. This range is chosen such that the minimum window length includes one to two cycles of the highest-frequency component that may be consistently identified among a representative sample of waveforms; the longest window is chosen based on a conservative estimate of likely maximum pick error. These window length ranges

may be different for  $P$  waves and  $S$  waves and will generally vary among stations. The entire cluster membership for each phase is cross-correlated for each of the window lengths. The resulting systems of cross-correlation values, lags, and standard deviations are compared to identify the best overall cross-correlation values and smallest average lag standard deviations for each phase in question.

For stations with multiple components (usually three), we use polarization filtering to improve  $P$  and  $S$  signal-to-noise levels prior to waveform comparison. This provides the best function for subsequent correlation when the source–receiver geometry is not favorable for a particular component.

A mean covariance matrix (e.g., Aster *et al.*, 1990) is calculated from the sum of the energy-normalized multicomponent signal for each event in the pair:

$$\mathbf{C} = \frac{1}{2} \left( \frac{\mathbf{x}_1^T \mathbf{x}_1}{E_1} + \frac{\mathbf{x}_2^T \mathbf{x}_2}{E_2} \right), \quad (5)$$

where  $\mathbf{x}_j$  is the (usually) three-component matrix with columns that are individual component time series for event  $j$ , and

$$E_j = \text{trace}(\mathbf{x}_j^T \cdot \mathbf{x}_j)^{1/2}. \quad (6)$$

The diagonalization of the positive definite  $\mathbf{C}$  matrix gives the unit eigenvector  $\vec{\mu}_1$  characterizing the best data projection for mutually linearizing particle motion between the two traces. For two- or three-component seismograms, the eigenvalue and eigenvector decomposition of the signal covariance matrix may be calculated exactly; however, the use of four-component (or more) sensors in reservoir microearthquake studies (e.g., Baria *et al.*, 1999) requires an iterative approach to the signal decomposition. We have therefore made use of repeated Jacobi transformations (Press *et al.*, 1989) to find the principal components of the tensor. All subsequent analysis for the waveform pair is performed on the projected (1-dimensional) data

$$\mathbf{x}'_j = (\mathbf{x}_j \cdot \vec{\mu}_1) \vec{\mu}_1. \quad (7)$$

Each waveform pair is next transformed via fast Fourier transform (FFT) into the frequency domain

$$X_{jk} = \sum_{k=0}^{M-1} x'_{jk} e^{-ijk/N}. \quad (8)$$

The cross-spectrum,

$$s_k = X_{1k} X_{2k}^* \quad (9)$$

(where  $*$  denotes complex conjugate), and coherence,

$$c_k = \frac{\sum_{l=j-m}^{j+m} s_l}{\sum_{l=j-m}^{j+m} |s_l|}, \quad (10)$$

are calculated, where the coherence averaging width is  $2m + 1$  Rayleigh bins (incrementally decreasing near the zero and Nyquist frequencies). We choose  $m$  as a fraction of the window length,  $M$ , with a minimum of five Rayleigh bins.

Prior to cross-correlation, we use the coherence and signal power to uniformly prefilter the two seismograms under consideration to emphasize frequencies that have high signal-to-noise and high coherency (e.g., Rowe and Aster, 1999; Aster and Rowe, 2000; Rowe, 2000). The spectral weighting is

$$\gamma_k = (|X_{1k}|/|X_{2k}|)^{1/2} c_k. \quad (11)$$

This filtering thus downweights incoherent frequency bands while reducing removal of potentially useful signal (a risk in an a priori bandpass filter choice; Fig. 2). Although the most coherent frequencies will most likely be found in the lower frequencies of the spectrum, the adaptive nature of this approach enhances the comparisons of highly similar waveforms in a diverse catalog by passing more of the coherent spectrum to each interevent correlation function. At the same time, we can accommodate gross similarities among those event pairs whose higher frequency details may not correlate well. A similar method has been applied to processing of seismic array signals by Wassermann and Ohrnberger (2001). Use of coherency filtering provides an additional benefit of permitting robust cross-correlation of slightly to moderately clipped waveforms. The spurious spectral contributions that may arise from clipping generally exhibit poor coherency, so are downweighted in the integer correlation step.

The coherency-filtered signals, with spectra

$$Y_{jk} = X_{jk} \gamma_k, \quad (12)$$

are next cross-correlated (with zero padding to eliminate circular correlation wraparound) in the frequency domain. The filtered cross-spectrum is transformed back into the time domain and the maximum of this cross-correlation function is the estimate for the coarse interevent pick lag. To estimate a standard deviation for the coarse (integer sample) correlation lags,  $l_\alpha$ , we perform a set of narrow-band correlations (typically eight) and find the variance, where each term is weighted by the cross-spectral power in that band (Aster and Rowe, 2000; Rowe, 2000). This coarse correlation standard deviation,  $\sigma_\alpha$ , is used, with the coarse cross-correlation maximum, as a discriminator to determine which waveforms are sufficiently similar to merit being passed to the subsample cross-correlation step for further lag refinement.

In some instances, the desired level of relative lag res-

olution is less than the sample interval. For example, in an area where the  $P$ -wave velocity is approximately 5 km/sec and where data are sampled at 100 samples/sec, the integer sample arrival time resolution for even high-signal-to-noise data can be as poor as 0.005 sec, which may introduce a worst-case location error of up to 25 m. The ability to consistently pick waveforms to subsample precision can dramatically improve resolution of small-scale features if sufficiently high-quality data are available.

The cross-spectral method (e.g., Poupinet *et al.*, 1984) provides a means of determining subsample lag adjustments by estimating a continuous function, the zero-intercept slope of the cross-spectral phase ( $\phi$ ), where the subsample lag term estimate is

$$l_\beta = \frac{1}{2\pi} \frac{d\phi(f)}{df} \quad (13)$$

and

$$\phi_k = \text{atan} \left( \frac{\text{imag}(s_k)}{\text{real}(s_k)} \right). \quad (14)$$

In many applications of this technique (e.g., Poupinet *et al.*, 1984; Got *et al.*, 1994), the relative weights of the phase values used in slope estimation have been calculated using a coherency-based measure. This provides useful relative weights to the phase points, but requires ad hoc scaling to the standard deviations needed to estimate meaningful error statistics for the subsample lag. For re-estimation of phase picks and their subsequent inclusion in relocation or other applications, quantitative estimates of the pick standard deviations in time units are important.

A further difficulty in applying the cross-spectral method arises from the inherent difficulty of characterizing the spectrum for a short time series. Spectral estimation on a sampled time series via FFT may be severely compromised when the length of the time window is shortened, because spectral leakage, which results from truncation of the windowing function, can bias the high-frequency rolloff of the spectral estimate both in amplitude and phase (Park *et al.*, 1987). For many seismological applications, however, including the cross-correlation-based repicking algorithm, it is desirable to use a fairly short time window to isolate the limited and correlatable (direct phase arrival) portion of an intrinsically nonstationary signal; lengthening the window to reduce spectral leakage introduces a higher proportion of background noise, as well as scattering contributions from the coda, and degrades the direct phase arrival comparison.

A standard approach to reducing spectral leakage is to apply a taper to the truncated time series, one that smoothly downweights data points toward zero at the ends of the window. This provides good results in terms of reduced spectral leakage, but causes a severely elevated variance for the spectral estimate. The Hann taper, for instance, discards approximately 5/8 of the statistical information of the time series (Park *et al.*, 1987). A further difficulty arises because of the

**SYNTHETIC EXAMPLE - ADAPTIVE PRE-FILTERING**

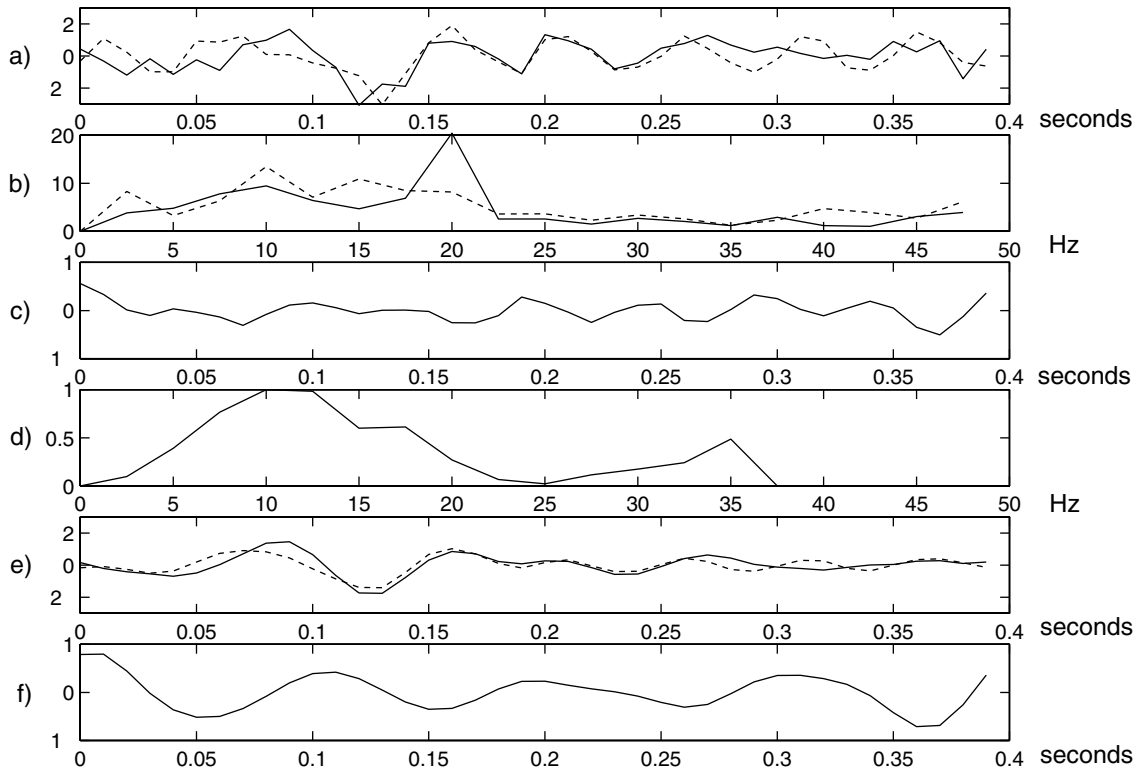


Figure 2. Example illustration of adaptive pre-filtering. (a) Two synthetic seismograms with added Gaussian noise. (b) Amplitude–frequency spectra of unfiltered seismograms from Fig. 1a. (c) Cross-correlation function for unfiltered noisy seismograms in Fig. 1a; the maximum cross-correlation coefficient is approximately 0.5. (d) Cross-coherency for spectra shown in Fig. 1b, plotted as a function of frequency. Although most coherent energy resides below 20 Hz, cross-coherency has a small peak at 35 Hz. A priori lowpass filtering may reject this energy, which could be an important common constituent to include in the cross-correlation. (e) Adaptively filtered seismograms from Fig. 1a; note significant reduction in the random noise constituent and overall similarity of resulting waveforms. (f) Recomputed cross-correlation function for the waveform pair, showing maximum cross-correlation coefficient > 0.9. Initial cross-correlation provided a zero sample lag; the filtered trace pair yields a lag of 2 samples (from Aster and Rowe, 2000).

intrinsic nonstationarity of the signal. Use of the Hann (or similar) taper tends to preferentially emphasize the spectral properties of that portion of the time series which falls in the central part of the window, while neglecting most of the information near the extrema, which may not be well represented by the center portion of the window.

One of the most successful approaches to solving these problems is the multitaper spectral estimation (Thomson, 1982), wherein discrete prolate spheroidal wave functions, which are eigenfunctions of the Dirichlet kernel, are employed. The eigenfunctions, denoted by  $U_k(N, W; f)$ ,  $k = 0, 1, \dots, N - 1$  are solutions to

$$\int_{-W}^W \frac{\sin N\pi(f - f')}{\sin\pi(f - f')} U_k(N, W; f') df' = \lambda_k(N, W) \cdot U_k(N, W; f), \quad (15)$$

where  $W$  ( $0 < W < 1/2$ ) is a bandwidth normally of the order  $1/N$ . The functions are ordered by their eigenvalues:

$$1 > \lambda_0(N, W) > \lambda_1(N, W) > \dots > \lambda_{N-1}(N, W). \quad (16)$$

The first  $2NW$  eigentapers have eigenvalues that are extremely close to 1. Of all functions that are the Fourier transform of an index-limited sequence, the discrete prolate spheroidal wave function has the greatest fractional energy concentration within the bandwidth of  $(-W, W)$  (Thomson, 1982). These eigenfunctions are orthogonal over the interval  $(-W, W)$  and are orthonormal over  $(-1/2, 1/2)$ . Their Fourier transforms provide the discrete prolate spheroidal sequences, also known as prolate eigentapers, with which we can window the time series prior to estimating its spectrum (Park *et al.*, 1987). We illustrate the five lowest-order multitaper functions for a 128-sample window in Figure 3.

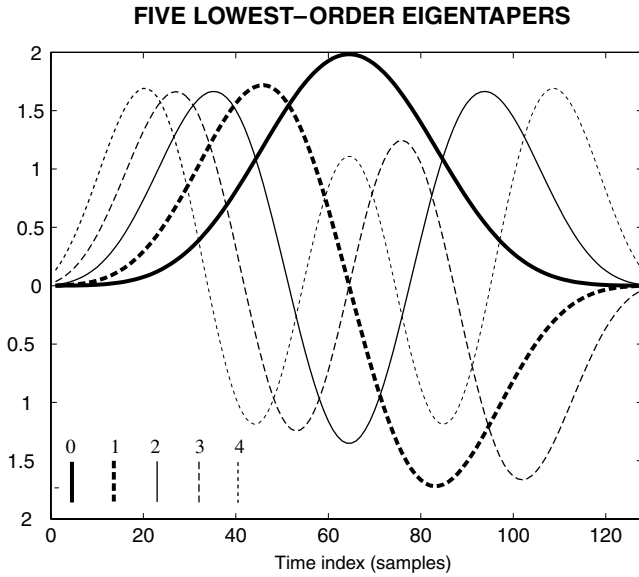


Figure 3. First five prolate spheroidal eigentapers for a time-bandwidth product of 4. Each taper recovers a different portion of the windowed seismogram; note that higher-order tapers have increasingly steep initial slopes; hence, spectral leakage becomes greater with higher-order tapers. The window is 128 samples long.

The products of the time series with each of the eigentapers are mutually orthogonal, as are their Fourier transforms; hence, we obtain a linearly independent series of eigenspectra for the time series, which may be combined in a weighted sum to estimate its true spectrum. The orthogonality of the eigenspectra further permits us to calculate estimates of error statistics for the summed spectrum in units of time. Use of the eigenspectral method therefore also enables us to address the question of dimensionally meaningful estimates of subsample pick lag errors, in lieu of dimensionless weights derived from ad hoc scaling of coherency, discussed earlier.

We precompute the tapering functions for a particular data length  $N$  and a specified time-bandwidth product,  $W$ .  $W$  is typically chosen to be 4, where  $2W$  approximately specifies the resolution of the resulting spectral estimates in Rayleigh bins (Thompson, 1982; Park *et al.*, 1987).

In our application, we calculate multitaper estimates of the cross-spectrum from two seismograms through conjugate multiplication (equation 9) of corresponding multitaper spectra (Fig. 4). Figure 4a shows two synthetic seismograms of length 32 samples. Each waveform is multiplied by the six lowest-order eigentaper functions, providing six linearly independent tapered realizations of each trace (Fig. 4b). Each is then transformed to the frequency domain, and we compute six linearly independent cross-spectral phase estimates from the tapered functions (Fig. 4c).

The lowest-order eigentapers (especially 0 and 1) have very low spectral leakage outside of the specified spectral

resolution bandwidth ( $k - W, k + W$ ). Higher-order tapers, however, have progressively worse spectral leakage characteristics, which may unacceptably flatten the estimated phase slope and hence underestimate the subsample correlation lag (e.g., Aster and Rowe, 2000; Rowe, 2000). We therefore use the average of the two lowest-order tapers to obtain the spectral values, while using the standard deviation of the six lowest-order tapers to estimate standard deviations on each phase point. This provides an acceptable trade-off between the advantages of the multitaper method (particularly quantitative error bars) and the need to reduce spectral leakage and resulting underestimation of the subsample lag (Aster and Rowe, 2000; Rowe, 2000). We illustrate by showing the average cross-spectral phase in Figure 4d, calculated from the 0th- and 1st-order spectra represented by solid lines in Figure 4c. Error bars in Figure 4d were calculated using all six of the eigen-cross-spectra shown in Figure 4c. We further downweight the most uncertain phase values by stretching the standard deviations of the multitaper estimates,  $\sigma_k$ , using the mapping

$$\sigma'_k = \tan(\sigma_k), \quad (-\pi/2 + \varepsilon < \sigma_k < \pi/2 - \varepsilon), \quad (17)$$

where  $\varepsilon = 0.01$ , and removing phase points from the phase slope estimation if  $\sigma_k$  is outside the range specified in equation (17). The phase slope and its standard deviation  $\sigma_\beta$  are estimated using the  $L_2$  (least-squares) zero-intercept linear regression for the  $K$  usable data points with

$$\frac{d\phi}{df} = \frac{\sum_{i=1}^K \phi_i / \sigma'_i}{\sum_{i=1}^K f_i / \sigma'_i} \quad (18)$$

and

$$\sigma_\beta(d\phi/df) = \frac{\left(\sum_{i=1}^K 1/\sigma'_i\right)^{1/2}}{\sum_{i=1}^K f_i / \sigma'_i}. \quad (19)$$

Because we use a preliminary integer cross-correlation step that adjusts the traces to the nearest sample prior to invoking the cross-spectral phase slope method, we do not need to concern ourselves with phase unwrapping when estimating the slope of the cross-spectral phase. Event pairs that are sufficiently similar to be passed to the subsample lag estimator will have a maximum phase lag of  $\pm\pi$ . Since slight to moderate waveform clipping does not affect the signal phase, this subsample lag estimation is relatively robust even when applied to clipped waveforms (e.g., Poupinet *et al.*, 1984).

The final interevent cross-correlation lag  $d_i$  for event pair  $i$  is determined by summing the coarse, integer lag value  $l_\alpha$  and the subsample lag estimate  $l_\beta$ . Total lag standard de-

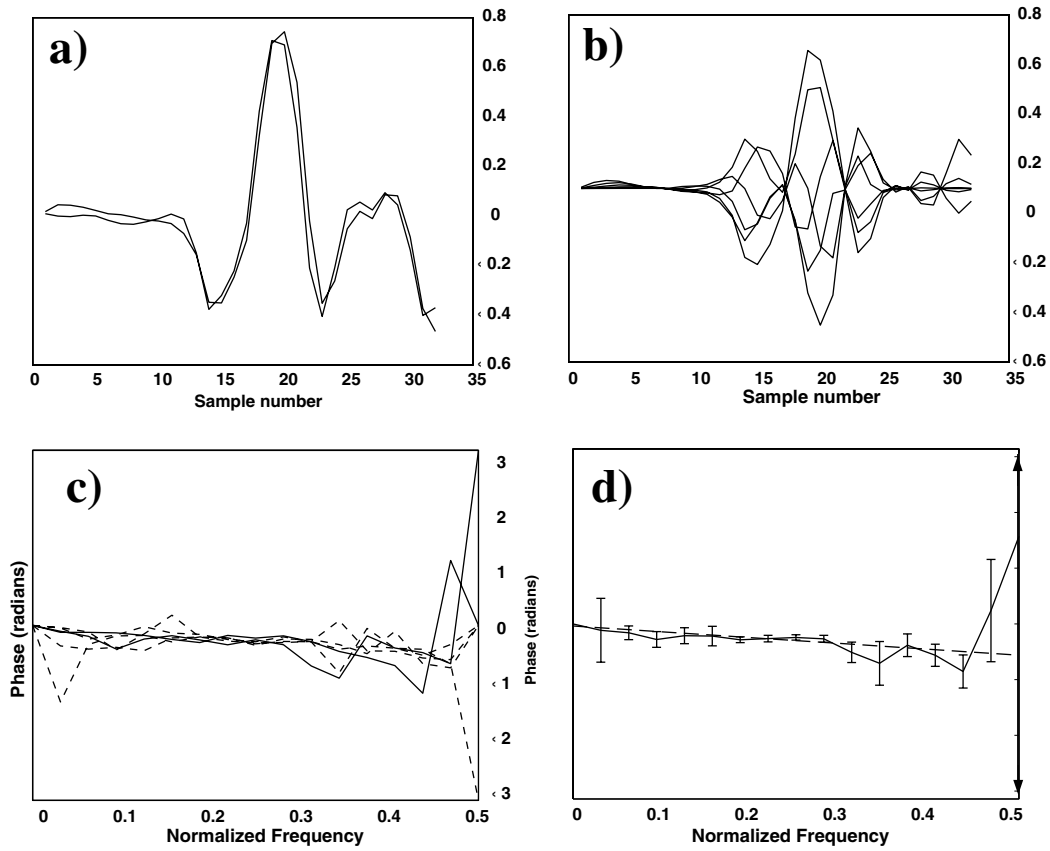


Figure 4. Example of estimation of cross-spectral phase slope and associated standard deviation. (a) Two synthetic seismograms that have been aligned to the nearest sample with integer correlation. (b) Six tapered representations of one of the synthetic traces; each function is the result of using one of the six lowest-order eigentapers to weight the time series. (c) Six linearly independent estimates of the cross-spectral phase for the traces in panel a. Each trace was windowed with the six lowest-order eigentapers (as in panel b), then six corresponding cross-spectra were calculated. Solid lines represent the cross-spectra corresponding to the two lowest-order eigentapers. The dashed functions are cross-spectral phases estimated from the third through sixth tapers. (d) Mean cross-spectral phase function estimated using two lowest-order cross-spectra. Dashed line represents the phase slope estimate; vertical error bars show standard deviations for each Nyquist bin, estimated from all six cross-spectra. Final phase-frequency point with arrowed error bar has a standard deviation large enough to disqualify the point in the phase slope fitting (after Aster and Rowe, 2000).

viation estimates  $\sigma'$  are the quadrature sum of the coarse and fine lag standard deviations:

$$\sigma'_i = \sqrt{\sigma_{a,i}^2 + \sigma_{\beta,i}^2}. \quad (20)$$

Solving the Systems of Constraints for Outlier-Resistant Pick Corrections

The desired  $N$ -vector of pick adjustments,  $\vec{b}$ , is the solution to

$$\mathbf{G} \vec{b} = \vec{d}, \quad (21)$$

where  $\vec{d}$  is a  $M$  (up to  $N(N-1)/2$ -length) vector of weighted interevent lags,

$$d_i = \frac{l_{a,i} + l_{\beta,i}}{\sigma'_i} \quad (i \neq M), \quad (22)$$

and the system matrix,  $\mathbf{G}$ , is a weighted first-difference operator on  $\vec{b}$  of the form

$$\mathbf{G} = \begin{bmatrix} -1/\sigma'_{2,1} & 1/\sigma'_{2,1} & 0 & 0 & 0 & 0 & \dots \\ -1/\sigma'_{3,1} & 0 & 1/\sigma'_{3,1} & 0 & 0 & 0 & \dots \\ -1/\sigma'_{4,1} & 0 & 0 & 1/\sigma'_{4,1} & 0 & 0 & \dots \\ 0 & -1/\sigma'_{3,2} & 1/\sigma'_{3,2} & 0 & 0 & 0 & \dots \\ 0 & -1/\sigma'_{4,2} & 0 & 1/\sigma'_{4,2} & 0 & 0 & \dots \\ 0 & -1/\sigma'_{5,2} & 0 & 0 & 1/\sigma'_{5,2} & 0 & \dots \\ 0 & 0 & -1/\sigma'_{4,3} & 1/\sigma'_{4,3} & 0 & \dots & \dots \\ 0 & 0 & -1/\sigma'_{5,3} & 0 & 1/\sigma'_{5,3} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix} \quad (23)$$

(Aster and Rowe, 2000; Rowe, 2000).  $\mathbf{G}$  is sparse; its  $M \times N$  dimension results in  $MN \propto N^3$  entries, of which only  $2M \propto N^2$  entries are nonzero. This system sparseness can be exploited to reduce greatly computer storage and solution time. We parameterize  $\mathbf{G}$  by two  $M$ -length index vectors,  $\vec{A}^-$  and  $\vec{A}^+$ , which contain the row indices of the negative and positive entries for each constraint, and by an  $M$ -length vector,  $\vec{\sigma}'$ , containing the lag standard deviations. This storage scheme is capable of easily representing very large (tens of millions of elements)  $\mathbf{G}$  systems within currently available workstation memory limits.

Straightforward linear approaches to solving equation (21) for a least-squares residual ( $L_2$ ) solution include Cholesky factorization or other techniques of solving the normal equations or involve singular-value decomposition (e.g., Press *et al.*, 1989). Such methods, however, require calculation of the (nonsparse)  $\mathbf{G}^T \mathbf{G}$  (which contains  $M^2 \propto N^4$  entries) or other large intermediary objects, which eliminate computational and storage advantages associated with our  $(\vec{A}^+, \vec{A}^-)$  representation of the sparse  $\mathbf{G}$ . Additionally, the  $L_2$  solution has the undesirable property of being strongly perturbed by outliers (e.g., Parker and McNutt, 1980; Shearer, 1998).

We instead solve equation (21) by implementing an iterative Polak–Ribiere conjugate gradient minimization (Polak, 1971; Press *et al.*, 1989) formulated to operate efficiently with the  $(\vec{A}^+, \vec{A}^-)$  sparse storage scheme. This can also be implemented for the more robust minimum one-norm residual ( $L_1$ ) solution. The functional to be minimized is

$$f = \mu^{(1)} = \sum_{i=1}^M \frac{|d_i - d_{i,\text{pred}}|}{\sigma_i}, \quad (24)$$

and the gradient of  $f$  at a general solution space point,  $x$ , is

$$\nabla f = \text{sgn}(\mathbf{G} \cdot \vec{x} - \vec{d}), \quad (25)$$

where the sign function operates on each element of a vector, returning 1 if the argument is positive,  $-1$  if the argument is negative, and 0 if the argument is zero. Although it has superior resistance to outliers, implementation of the  $L_1$  residual minimization becomes problematic when any of the residuals becomes too small, as the derivative function becomes discontinuous. We have successfully addressed this difficulty by modifying the misfit function for values of  $f$  that lie within the region  $-\varepsilon < 0 < \varepsilon$  for small  $\varepsilon$ :

$$\text{if } |f_i| > \varepsilon, f_i = \frac{|d_i - d_{i,\text{pred}}|}{\sigma_i}, \quad (26)$$

and  $\nabla f$  is as described in equation (25) (Aster and Rowe, 2000; Rowe, 2000).

$$\text{If } |f_i| \leq \varepsilon, f_i = \frac{\langle d_i - d_{i,\text{pred}} \rangle^2}{2\sigma_i \varepsilon} + \frac{\varepsilon}{2} \quad (27)$$

$$\text{and } df_i = \frac{\text{sgn}(d_i - d_{i,\text{pred}})}{\sigma_i \varepsilon}. \quad (28)$$

This modification has a theoretical drawback, insofar as the smallest misfit we may obtain is  $\varepsilon/2$ , as opposed to zero; however, this poses no practical difficulty. We are currently obtaining satisfactory results using a value of  $\varepsilon = 0.1$ . Calculation of the solution probability (outlined next) may be done by recomputing  $f$  with the exact  $L_1$  formulation, although this will not be the true minimum because of our approximation for small  $f_i$  (Rowe, 2000).

From a probabilistic viewpoint, the  $L_1$  solution is the maximum likelihood under the assumption of exponentially distributed data errors, described by

$$P^{(1)}(x) = \frac{1}{\sigma} \exp(-|x| - m|l/\sigma|). \quad (29)$$

Parker and McNutt (1980) describe the statistics of  $\mu^{(1)}$  (equation 24) under an assumption of Gaussian data errors, which we invoke as a useful quality-of-fit measure to assess whether the relative lags estimated by our  $L_1$  solutions form a consistent set of first-difference constraints on  $\vec{b}$  (equation 21). The  $L_1$  analogue to the  $L_2$  ( $\chi^2$ )  $q$ -statistic for  $M = K - N$  degrees of freedom is approximated by a third-moment expression for the probability that a greater value of  $\mu^{(1)}(M)$  than the observed one (equation 24) could have occurred:

$$q(f, M) = P(x) - \frac{\gamma}{6} Z^{(2)}(x), \quad (30)$$

where  $P(x)$  is the cumulative probability integral

$$P(x) = \frac{1}{\sigma_1(2\pi)^{1/2}} \int_{-\infty}^x \exp(-t^2/\sigma_1^2) dt \quad (31)$$

for a zero-mean Gaussian distribution with the variance of  $\sigma_1 = \sigma^2(\mu^{(1)})$ , where

$$\sigma_1^2 = (1 - 2/\pi) M. \quad (32)$$

The second term is proportional both to the skewness of  $\mu^{(1)}$ :

$$\gamma = \frac{2 - \pi/2}{(\pi/2 - 1)^{3/2} M^{1/2}}. \quad (33)$$

and to

$$Z^{(2)}(x) = \frac{1}{(2\pi)^{1/2}} (x^2 - 1) \exp(-x^2/2), \quad (34)$$

where

$$x = \frac{f - \bar{\mu}}{\sigma_1} \quad (35)$$

and

$$\bar{\mu} = (2/\pi)^{1/2} M. \quad (36)$$

We have found that we can further improve the solution by conservatively rejecting correlation outliers. The first step is to discard lag constraints whose cross-correlation maxima are sufficiently poor that there is little likelihood of constructive contribution to the solution. We have adopted an *a priori* threshold of 0.8; any constraints which fail to meet this minimum are rejected prior to the first attempt at solving the system. This tolerance will vary depending on the quality of the catalog being addressed and the desired similarity threshold in the application.

Preliminary clustering helps to ensure that disparate families of earthquakes are not being compared, but outliers and poorly correlating events resulting from correlation cycle skips, excessive noise or grossly inaccurate initial picks may still remain. To eliminate the influence of these outliers, we first calculate the  $L_1$  solution to equation (21) and its misfit measure,  $f$  (equation 24), using the full constraint and data set for the cluster (minus the *a priori* rejections). If there are data outliers or if the system is otherwise highly inconsistent, a large value of  $\gamma$  (equation 33) will produce a highly unlikely (very small) value of  $q(f, M)$  (equation 30). We then successively cull constraints and corresponding data from the system, using a binary search mechanism (e.g., Aster and Rowe, 2000; Rowe, 2000). The data misfit vector is sorted and the constraints corresponding to the worse half of the misfit estimate are discarded. Discarding, rather than down-weighting, the highest misfit constraints ensures correct probability calculations for subsequent solutions by appropriately adjusting the degrees of freedom of the system. The reduced system is solved, and the value of  $q(f^i, M^i)$  is recalculated for the  $i$ th bisection step under the new degrees of freedom. If this value is too good ( $q(f^i, M^i) > 0.02$ ), we assume that too many constraints have been discarded and we restore a portion of them. We recompute  $q(f^i, M^i)$  and restore or discard constraints again, as appropriate. This process generally converges to a satisfactory value of  $q(f^i, M^i)$  within 10 steps, depending on predetermined thresholds for convergence and bisection step size parameters. After convergence has been achieved, we obtain a final pick adjustment solution for the reduced system of  $M'$  constraints, and calculate  $1-\sigma$  error bars for each element of the final solution via Monte Carlo propagation of Gaussian data errors. An a posteriori zero-mean constraint is applied to the final set of pick adjustments.

Figure 5 illustrates the solution process for an  $m = 6$ -event synthetic cluster, initially constrained by a full set of  $M = (6)(5)/2 = 15$  interevent lag estimates and 1 zero-mean constraint. The true pattern of pick adjustments was chosen arbitrarily to be a zero-mean half-period sine function with

an amplitude of 1 time unit. The 15 first-difference data points were randomized by adding a Gaussian error term with a standard deviation of 0.2 time units. Outliers were introduced to the system by adding large random terms to data points 3 and 7.

Figures 5a and 5b show the  $L_1$  solution and data fit for the entire data and constraint set, where the recovery of true pick adjustments has been skewed by the data outliers, and the probability of a worse misfit is  $q \approx 0$  to single precision. After automatically rejecting the two system constraints with the largest residual contributions, as we have already described, and re-solving the problem, we obtain a revised solution (Fig. 5c) with an acceptable data misfit (Fig. 5d) of  $q(9.22, 8) \approx 0.06$  and an  $L_1$  misfit improvement between solution and true model,

$$f' = \sum_{i=1}^{M'} \frac{|x_i - x_{i,\text{true}}|}{\sigma_1}, \quad (37)$$

of 59% with generally tighter  $1-\sigma$  error bars.

#### Absolute versus Relative Locations

Introduction of the new picks for each cluster provides precise relative event relocations within clusters, but the question of improved intercluster locations is not addressed in this fashion. Within any given cluster it is commonly observed that analyst picks do not always scatter about a zero mean; they are often systematically late in instances of low signal/noise. The resulting adjusted picks for a particular phase may therefore exhibit a significant bias. Such biases will result in the relative mislocation of mean cluster centroids, an artifact that is carried forward from the preliminary catalog to the intracluster relative relocations (Rowe, 2000; Rowe *et al.*, 2002).

To correct for relative cluster centroid mislocations, we compare waveform similarity among the clusters. Relative pick lags within clusters are estimated as we have outlined, adjusting the phase picks accordingly. Waveforms for each phase within each cluster are then aligned on their adjusted picks and stacked (Fig. 6). Each stack is then treated as a composite earthquake trace for the cluster. Ensembles of stacked seismograms are then cross-correlated and relative pick lags determined between the composite earthquakes using the  $L_1$ -norm conjugate gradient solver as before. The resulting intercluster lags are used to adjust mean picks within each cluster. In this way the very tightly constrained relative adjustments for intracluster associations are preserved, with no risk of degrading these relative locations by including uncorrelated events, and the overall intercluster relationships are adjusted according to the composite waveform cross-correlation lags (Rowe, 2000; Rowe *et al.*, 2002).

We illustrate this hierarchical correlation and stacking method in Figure 6 for three synthetic clusters. In Figures 6a, 6b, and 6c we show preliminary and repicked waveform alignments for each of the three clusters. These synthetic

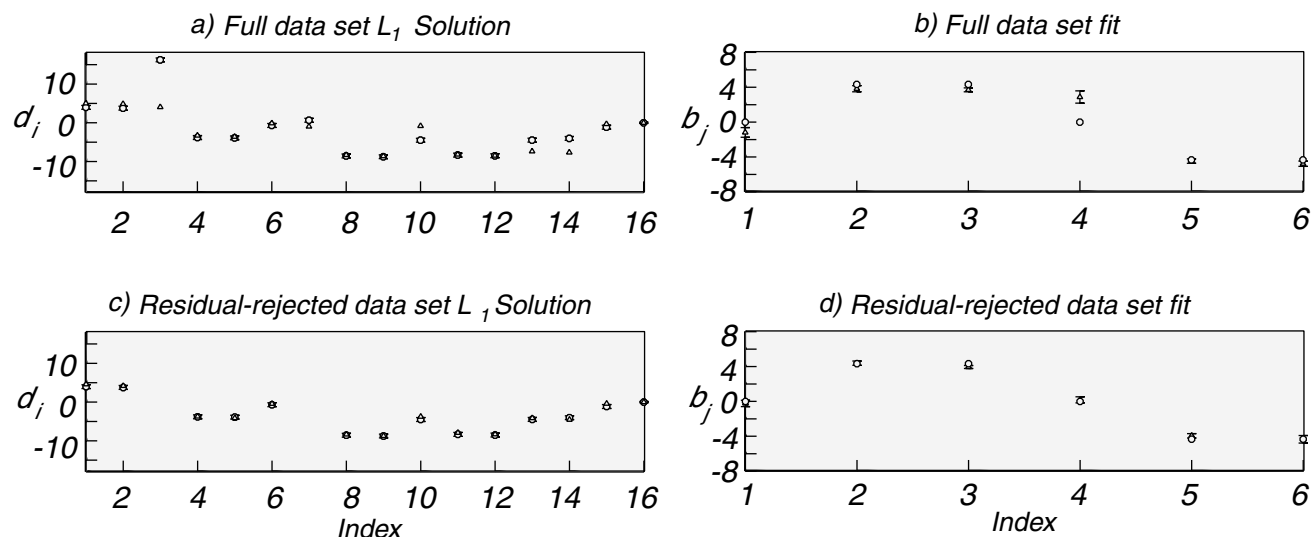


Figure 5. Improved solution robustness using the  $L_1$  residual minimization and iterative residual-based system constraint rejection. (a) Full data set solution (triangles) with Monte Carlo-estimated standard deviations, compared to the true model (circles). (b) Predicted data (triangles) and actual data (circles) accompanied by standard errors. (c) Reduced dataset with outliers removed by residual-based rejection. (d) Refined solution following removal of outliers (after Aster and Rowe, 2000).

clusters were generated by isolating one cluster of similar waveforms and variously perturbing the preliminary picks to generate different pick variances and means among the three examples. Although within each of the aligned clusters the new picks are consistent, note that the mean adjusted pick (horizontal dashed line) occurs at somewhat different times on the resulting alignment stack. This is an artifact of the differing preliminary pick distributions among each of the clusters. Figure 6d illustrates the problem that results if we assume that preliminary cluster centroids are accurate: in the upper panels we show on the left all member events of the three clusters, aligned on their preliminary picks, with the resulting stack shown to the right. In the lower panels of Figure 6d we show the traces aligned on their revised intracluster picks, and the resulting stack. Although the waveforms have clearly been well aligned for intracluster consistency, a serious misalignment is exhibited among the three clusters in terms of the resulting mean pick estimates.

If we subsequently cross-correlate the stacked waveforms, however, and determine relative lags among the stacks (Fig. 6e), we can then apply the additional pick correction to each of the member traces and adjust all events by their relative intercluster lags (Fig. 6f). This bilevel cross-correlation approach still does not address the question of overall analyst bias for an entire catalog, but any remaining picking artifact could be handled through individual station corrections determined through JHD or other joint location methods. We note, however, that this approach cannot remove the artifacts from the picks themselves. Hence, other applications that rely on the picks will be unable to separate picking bias from actual travel-time residuals at the receive-

ers. This overall bias exists in the preliminary as well as the relocated catalog. Furthermore, cluster stacks and individual events that do not correlate well with other events or clusters remain uncorrected with this technique. Addressing these orphans, as well as addressing overall catalog bias, is the subject of ongoing work.

## Summary

We have developed an automatic, adaptive algorithm for adjusting phase picks for consistency among similar events within large digital seismic waveform catalogs. Innovations include automatic, adaptive, cross-coherency and polarization filtering, and the use of eigenspectral methods for estimating subsample phase lags and dimensionally meaningful lag standard deviations. After initial cross-correlation the catalog is clustered using a hierarchical, dendrogram-based pair-group classification scheme with segregation based either on the cophenetic correlation function or on a predetermined cross-correlation threshold. Resulting clusters are solved for consistent intracluster pick lags using an  $L_1$ -norm minimizing, outlier-resistant, iterative, conjugate gradient method formulated to minimize memory and computation requirements. Intracluster seismograms are aligned on the zero-mean adjusted repicks, then stacked to provide a composite waveform. Ensembles of stacked seismograms are then cross-correlated among clusters to determine intercluster pick lag adjustments; this corrects for possible analyst biases and provides for consistent intercluster pick (and location) relationships within the

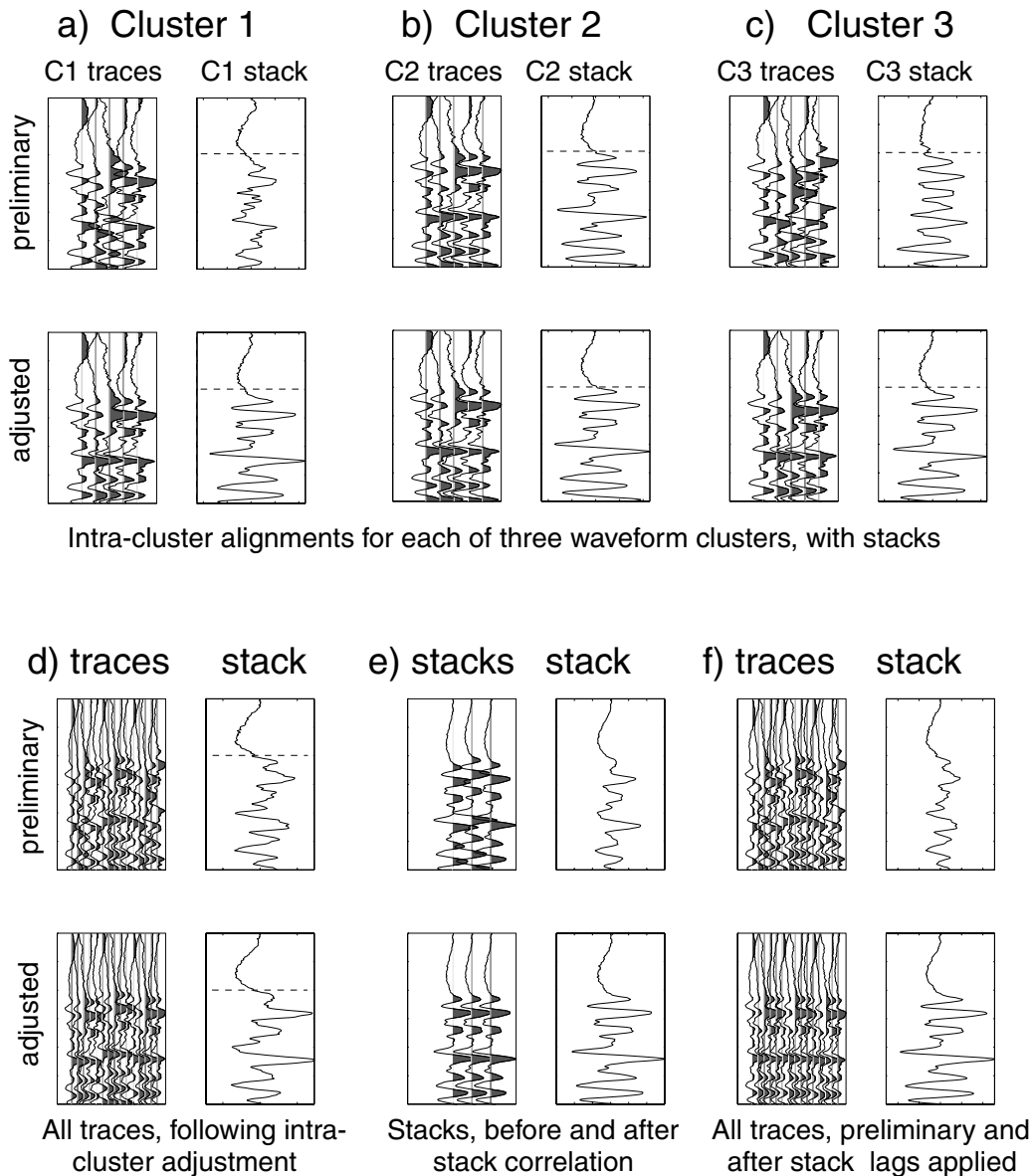


Figure 6. Hierarchical clustering, lag adjustment, and stacking method (a–c). Three hypothetical clusters of five events each. Upper panels show traces aligned on preliminary picks and associated waveform stack; lower panels show traces aligned on adjusted picks, with associated stack. Horizontal dashed lines indicate pick times on the stacked trace in each. (d) Clusters from a, b, and c combined to show initial scatter (upper panels) and relationships among intracenter adjustments (lower panels). (e) Stacks from clusters of a, b, and c, showing preliminary alignment (upper panels) and appropriately shifted stacks, following hierarchical stacking and cross-correlation (lower panels). (f) The same three clusters showing initial misalignments (upper panels) and final, corrected alignments (lower panels) after both intracenter and intercenter lags have been applied.

catalog, with no dependence on preliminary hypocenters. We stress that, although the method is termed automatic, this does not imply black box or completely hands-off, insofar as application requires some intelligent choices on the part of the user to tune the parameters in the algorithm so that it then may be allowed to process the data automatically. The goal of this software is to provide substantial improvement

in pick consistency and accuracy while reducing the burden on analysts; it goes without saying that problematic events will require human intervention, and assessment of the success of the automatic processing will be required. We are continuing development of the procedure to further minimize user intervention and to implement near-real-time functioning.

An application of this technique to a large seismic catalog can be found in Rowe *et al.* (2002), in which the algorithm is applied to microearthquakes associated with injection experiments at the Soultz, France, hot dry rock geothermal site.

### Acknowledgments

Very useful discussions of the method were provided by H. Moriya and R. Jones. Helpful reviews of the manuscript were provided by M. Fehler, C. Thurber, H. Tobin, J. Schlue, and N. Deichmann. We also appreciate thorough reviews by S. Moran and H. Asanuma. This work was supported under a grant from Sandia National Laboratories, Albuquerque, New Mexico. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy under Contract Number DE-AC04-94AL85000. Funding was also provided by Niitsuma Laboratories, Tohoku University, Sendai, Japan, and by National Science Foundation Office of Polar Programs Grant Number 9419267.

### References

- Aster, R. C., and C. A. Rowe (2000). Automatic phase pick refinement and similar event association in large seismic data sets, in *Advances in Seismic Event Location*, C. Thurber and N. Rabinowitz (Editors), Kluwer, Amsterdam, 231–263.
- Aster, R. C., and J. Scott (1993). Comprehensive characterization of waveform similarity in microearthquake data sets, *Bull. Seism. Soc. Am.* **83**, 1307–1314.
- Aster, R. C., P. M. Shearer, and J. Berger (1990). Quantitative measurements of shear wave polarization at the Anza seismic network, southern California: implications for shear wave splitting and earthquake prediction, *J. Geophys. Res.* **95**, 12,449–12,473.
- Baria, R., J. Baumgartner, A. Gerard, R. Jung, and J. Garnish (1999). European HDR research programme at Soultz-sous-Forêts (France) 1987–1996, *Geothermics* **28**, 655–669.
- Carr, D., C. Young, R. Aster, and X. Zhang (1999a). Cluster Analysis for CTBT Seismic Event monitoring, 21st Annual Seismic Research Symposium on Monitoring a CTBT, 285–293.
- Carr, D., C. Young, J. Harris, R. Aster, and X. Zhang (1999b). Cluster Analysis for CTBT Seismic Event monitoring (abstract), *Seism. Res. Lett.* **70**, 227–228.
- Deichmann, N., and M. Garcia-Fernandez (1992). Rupture geometry from high-precision relative hypocentre locations of microearthquake clusters, *Geophys. J. Int.* **110**, 501–517.
- Dodge, D. A., G. C. Beroza, and W. L. Ellsworth (1995). Foreshock sequence of the 1992 Landers, California, earthquake and its implications for earthquake nucleation, *J. Geophys. Res.* **100**, 9865–9880.
- Fehler, M., W. S. Phillips, R. Jones, L. House, R. Aster, and C. Rowe (2000). A method for improving relative earthquake locations, *Bull. Seism. Soc. Am.* **90**, 775–780.
- Fremont, M.-J., and S. D. Malone (1987). High precision relative locations of earthquakes at Mount St. Helens, Washington, *J. Geophys. Res.* **92**, 10,223–10,236.
- Gillard, D., A. M. Rubin, and P. Okubo (1998). Highly concentrated seismicity caused by deformation of Kilauea's deep magma system, *Nature* **384**, 343–346.
- Got, J.-L., J. Fréchet, and F. W. Klein (1994). Deep fault plane geometry inferred from multiplet relative relocation beneath the south flank of Kilauea, *J. Geophys. Res.* **99**, 15,375–15,386.
- Harris, M., and C. Young (1997). MatSeis: a seismic GUI and tool-box for MATLAB, *Seism. Res. Lett.* **68**, 267–269.
- Joswig, M. (1995). Automated classification of local earthquake data in the BUG small array, *Geophys. J. Int.* **160**, 262–285.
- Kissling, E. (1988). Geotomography with local earthquake data, *Rev. Geophys.* **26**, 659–698.
- Lance, G. N., and W. T. Williams (1967). A general theory for classificatory sorting strategies. 1. hierarchical systems., *Comput. J.* **10**, 271–276.
- Lees, J. M., (1998). Multiplet analysis at Coso geothermal, *Bull. Seism. Soc. Am.* **88**, 1127–1143.
- Ludwig, J. A., and J. F. Reynolds (1988). *Statistical Ecology: A Primer on Methods and Computing*, John Wiley & Sons, New York.
- Park, J., C. R. Lindberg, and F. L. Vernon III (1987). Multitaper spectral analysis of high-frequency seismograms, *J. Geophys. Res.* **92**, 12,675–12,684.
- Parker, R., and M. McNutt (1980). Statistics for the one-norm misfit measure, *J. Geophys. Res.* **85**, 4429–4430.
- Phillips, W. S. (2000). Precise microearthquake locations and fluid flow in the geothermal reservoir at Soultz-sous-Forêts, France, *Bull. Seism. Soc. Am.* **90**, 212–228.
- Phillips, W. S., L. S. House, and M. C. Fehler (1997). Detailed joint structure in a geothermal reservoir from studies of induced microearthquake clusters, *J. Geophys. Res.* **102**, 11,745–11,763.
- Polak, E. (1971). *Computational Methods in Optimization*, Academic Press, New York.
- Poupinet, G., W. L. Ellsworth, and J. Fréchet (1984). Monitoring velocity variations in the crust using earthquake doublets: an application to the Calaveras Fault, California, *J. Geophys. Res.* **89**, 5719–5731.
- Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1989). *Numerical Recipes in C*, Cambridge Univ. Press, New York.
- Pujol, J. (1992). Joint hypocentral location in media with lateral velocity variations and interpretation of the station corrections, *Phys. Earth Planet. Inter.* **75**, 7–24.
- Rowe, C. (2000). Correlation-Based Phase Pick Correction and Similar Earthquake Family Identification in Large Seismic Waveform Catalogs, *Ph.D. Thesis*, New Mexico Institute of Mining and Technology, Socorro.
- Rowe, C. A., and R. C. Aster (1999). Application of automatic, adaptive filtering and eigenspectral techniques to large digital waveform catalogs for improved phase pick consistency and uncertainty estimates (abstract), *EOS* **80**, F660.
- Rowe, C. A., R. C. Aster, W. S. Phillips, R. H. Jones, B. Borchers, and M. C. Fehler (2002). Relocation of induced microseismicity at the Soultz geothermal reservoir using automated, high-precision repicking, *Pure Appl. Geophys.* **159**, 563–596.
- Rubin, A. M., D. Gillard, and J.-L. Got (1999). Streaks of microearthquakes along creeping faults, *Nature* **400**, 635–641.
- Shearer, P. M. (1997). Improving local earthquake locations using the  $L_1$  norm and waveform cross correlation: application to the Whittier Narrows, California, aftershock sequence, *J. Geophys. Res.* **102**, 8269–8283.
- Shearer, P. M. (1998). Evidence from a cluster of small earthquakes for a fault at 18 km depth beneath Oak Ridge, Southern California, *Bull. Seism. Soc. Am.* **88**, 1327–1336.
- Slunga, R., S. T. Rognvaldsson, and R. Bodvarsson (1995). Absolute and relative locations of similar events with application to microearthquakes in southern Iceland, *Geophys. J. Int.* **123**, 409–419.
- Sneath, P. H. A., and R. R. Sokal (1973). *Numerical Taxonomy*, W. H. Freeman & Company, San Francisco.
- Thomson, J. D. (1982). Spectrum estimation and harmonic analysis, *Proc. IEEE* **70**, 1055–1096.
- Waldhauser, F., and W. L. Ellsworth (2000). A double-difference earthquake location algorithm: method and application to the northern Hayward Fault, California, *Bull. Seism. Soc. Am.* **90**, 1353–1368.
- Wassermann, J., and M. Ohrnberger (2001). Automatic hypocenter determination of volcano induced seismic transients based on wavefield coherence: an application to the 1998 eruption of Mt. Merapi, Indonesia, *J. Volcanol. Geotherm. Res.* **110**, 57–77.
- Young, C. J., B. J. Merchant, and R. C. Aster (2001). Comparison of cluster analysis methods for identifying regional seismic events, 23rd Annual DTRA/NNSA Seismic Research Review, 229–238.

Department of Earth and Environmental Science and Geophysical  
Research Center  
New Mexico Institute of Mining and Technology  
Socorro, New Mexico 87801  
char@geology.wisc.edu  
(C.A.R., R.C.A.)

Sandia National Laboratories  
P.O. Box 5800, MS 1138  
Albuquerque, New Mexico 87185-1138  
(C.J.Y.)

Department of Mathematics  
New Mexico Institute of Mining and Technology  
Socorro, New Mexico 87801  
(B.B.)

Manuscript received 22 August 2001.