An Empirical Approach to Subspace Detection

by Sarah A. Barrett and Gregory C. Beroza

INTRODUCTION

Waveform cross-correlation detection methods, or match filter techniques (Van Trees, 1968), exploit waveform similarity for proximal events, and have proven to be a powerful approach for detecting and characterizing seismic events. Among the advantages over more traditional techniques is the ability to detect in adverse observational conditions either when events are closely spaced in time, such that waveforms overlap, or when the signal-to-noise ratio (SNR) is much less than unity because signal strength is low. Correlation methods can also be used for precise measurements of relative arrival time (e.g., Schaff *et al.*, 2004) and relative amplitude (Rubinstein and Ellsworth, 2010) in support of precise relative location and relative size measurements.

A complete and precise measurement of seismic events is critical for understanding earthquake processes, and as a result, the use of correlation-based detection methods is growing rapidly. They have been used on mining-induced seismicity (Gibbons and Ringdal, 2006), in nuclear test ban treaty verification research (Rowe *et al.*, 2012), to understand tectonic tremor as a swarm of low-frequency earthquakes (Shelly *et al.*, 2006, 2007), and to increase the number of detected early aftershocks by more than an order of magnitude (Peng and Zhao, 2009). They have also been applied to detect previously unobserved earthquakes in situations of suspected induced seismicity (van der Elst *et al.*, 2013).

Waveform matching is not always applicable because it requires prior identification of template events that are highly similar to the event(s) of interest. The subspace detection method is a generalization of template detection that exploits features common to a design set of earthquakes (Harris, 2006; Harris and Paik, 2006). Subspace detection uses the singular value decomposition (SVD) of the design set of waveforms to find an orthonormal representation (left singular vectors) that efficiently captures the important common characteristics of the waveforms. If we rank the singular vectors by the size of their corresponding singular value, then the first singular vector contains the information common to the design set events with the greatest power to describe the set of waveforms. The second singular vector contains the dominant remaining information that is common to the design set events once the contribution of the first singular vector is removed, and so on, with each successive vector containing increasingly smaller contributions to describing the design set. These vectors can be used in an approach similar to that of template matching to find events correlating above a given threshold, but with the detection statistic developed from a projection over a subspace of the design set, rather than of a single template vector.

We find the first singular vector closely approximates the average waveform, or stack of the design set. In the case of densely clustered earthquakes, the second singular vector strongly resembles the time derivative of the design set stack. We believe that this reflects the fact that the second singular vector represents information related to the variations produced by slight offsets in earthquake location of the design set earthquakes. Successive singular vectors do not have a clear physical representation (i.e., the third singular vector does not resemble the second derivative of the stack).

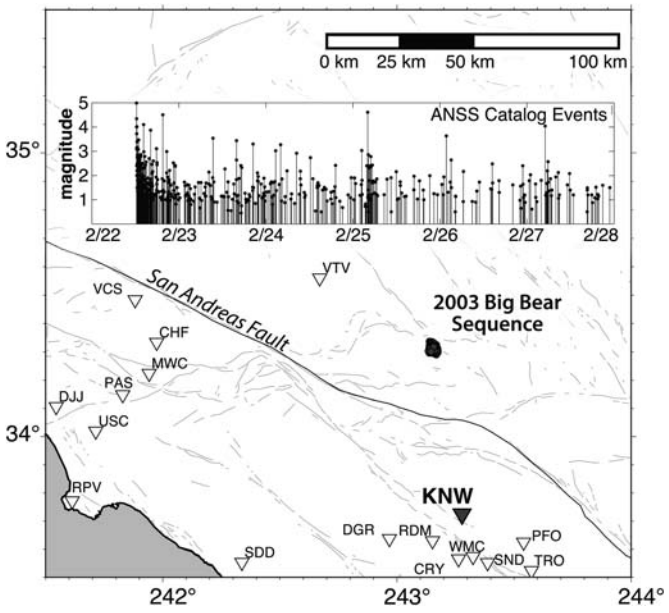
We use a matrix of two rows: the stack and the derivative of the stack as an empirical representation of the subspace method and compare its ability to detect missing earthquakes from the catalog with the template matching and subspace projection methods. We find that the empirical subspace approach provides substantial advantages over simple template matching methods in that it increases the detection of previously undetected and uncataloged events, including events with overlapping waveforms. In addition, it avoids an increase in false detections.

We demonstrate its capabilities by applying the empirical subspace method, and comparing the results with traditional template matching and with subspace projection, for the 2003 M_w 5.0 Big Bear sequence (Fig. 1), which began on 22 February 12:19:10 (UTC) at 34.31° N 116.85° W, near the Helendale fault. Using recordings from a single SCSN station we find ~500 events in the aftershock sequence—far more than are present in the CISN catalog. The power of this approach suggests the potential to extract much more information from earthquake and tremor sequences using subspace approaches.

SUBSPACE DETECTION

Standard network-based earthquake detection relies on a combination of arrival-time picks to obtain possible first-arrival times, and associated algorithms to determine whether those arrival times are mutually consistent with a source within the Earth. Short-term average/Long-term average-based arrival-time detection has the important advantage of being applicable to events without *a priori* knowledge of an event's location or mechanism. This works well for isolated earthquakes with impulsive, high SNR first arrivals, and is particularly effective at locating and characterizing earthquakes that occur within a seismic network. Conversely, these detectors may run into trouble when earthquakes are closely spaced, when arrivals are emergent, or when earthquakes are sparsely recorded. On this last point, an earthquake usually is not cataloged if there are not

doi: 10.1785/0220130152



▲ **Figure 1.** Seismicity during the Big Bear swarm as recorded by the Advanced National Seismic System (ANSS) Catalog from 22 February 2003 through 28 February 2003 (circles) on AZ and CI broadband stations (triangles) station KNW, is the highest quality station used in this analysis (darkened triangle). Inset ANSS Catalog events during the swarm: the sequence had no activity prior to the mainshock (22 February 2003), whereas aftershocks persisted for a few weeks.

enough stations (typically four) providing independent information to constrain the hypocenter.

Waveform-based detectors are less general in that they require assumptions about the seismic source. In the case of template matching, the assumption is that the shape of the waveforms is known (Gibbons and Ringdal, 2006) or that the source waveform repeats (Brown *et al.*, 2008). Subspace detectors can be considered a generalization of template matching to multiple dimensions. Subspace detectors are constructed from the SVD of a matrix that consists of a set of design waveforms (Harris, 2006). Subspace detectors have been applied to the study of tremor with some success (Maceira *et al.*, 2010), but have not seen wide application in earthquake monitoring. The subspace technique works with a set of design waveforms that are selected from the total population of recorded, cataloged events and are intended to represent the characteristics of potential events of interest. In particular, the subspace detection algorithm assumes that undetected events can be represented as a linear combination of the largest singular vectors in the design set matrix subspace.

APPLICATION OF THREE DETECTION ALGORITHMS TO THE 2003 BIG BEAR SEQUENCE

The 2003 M 5.1 Big Bear earthquake occurred in roughly the same region as the 1992 M 6.5 Big Bear earthquake (the largest aftershock of the 1992 M 7.3 Landers earthquake) in the eastern California shear zone. This region is characterized by a series of northwest-trending faults with northeast-trending conjugate faults, bounded by the frontal thrust of the San Bernardino Mountains in the north, and the San Andreas fault in the south (Jones *et al.*, 1993). In both earthquake sequences, the aftershocks were prolific and diverse. Aftershocks occur primarily on the northwest-trending plane; however, there are significant populations on conjugate faults (e.g., Jones *et al.*, 1993; Chi and Hauksson, 2006). Dip-slip faults bound the region of aftershocks, and in both the 1992 and 2003 earthquakes, produced at least one moderate-sized reverse-faulting event. This sequence is a good test case for the empirical subspace methods; it is a rich and complex aftershock sequence with events of varied focal mechanism (Yang *et al.*, 2012) as well as overlapping events. The highest quality nearby station for which continuous data are available during the 2003 sequence is station KNW from the ANZA network, which is approximately 60 km away from the Big Bear sequence. This station recorded nearly all cataloged events. Waveforms from these events were downloaded using the Standing Order for Data (SOD) request tool from the Incorporated Research Institutions for Seismology (IRIS) Data Management Center (DMC; Owens *et al.*, 2004).

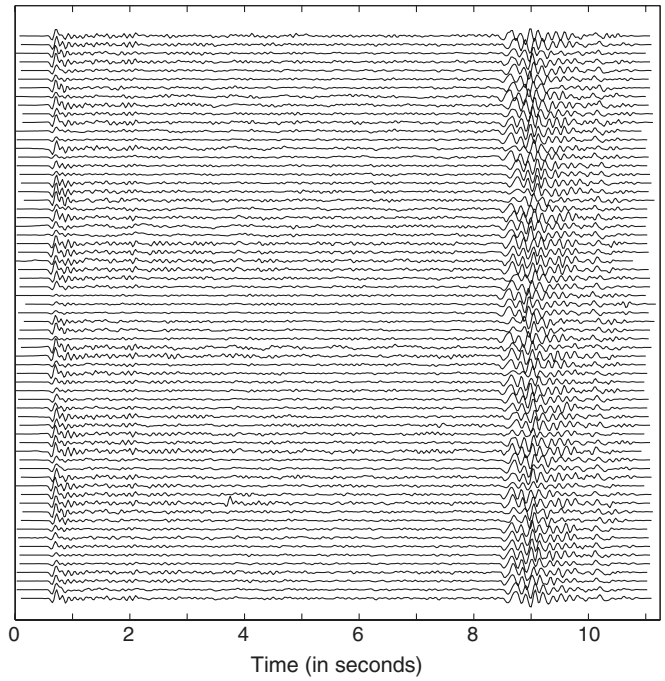
Basic processing on the waveforms from this station yields the subspace dimension of interest. We arrange the window for the design set to include the P arrival (1 s before the automatically picked arrival) and a substantial portion of the S-wave coda (10 s following the P arrival, or about 3 s after the S arrival). The mean of the signal is removed, and events are filtered using a band-pass of 1–10 Hz. Finally, a cosine taper is applied to 5% of the signal on each end. We cross correlate each event, pairwise, then perform a single-linkage cluster analysis. The largest five groups with normalized cross-correlation coefficients of at least 0.875 are included in the design set (Fig. 2). We select a master event with high-correlation values with respect to the other design set events to define lag times for P-wave alignment using cross correlation. We construct a matrix of aligned waveforms using these lag times, with each seismogram comprising a row of the **A** matrix (Fig. 3).

The SVD of matrix **A** provides the left singular vectors, which are the columns of the orthonormal **U** matrix that form a basis for the design event waveforms.

$$\mathrm{SVD}(\mathbf{A}) = \mathbf{U}\Lambda\mathbf{V}^T. \quad (1)$$

These vectors are used as subspace detectors for some sufficient dimension N.

Following Harris (2006), we examine the consequence of using different dimensions of the subspace for detection by calculating the fractional energy capture. In the Big Bear example, we find that the first two singular vectors are sufficient to capture all design set events at a threshold of 0.125 (subspace dimension of $N = 2$). However, not all cataloged events are adequately represented by only two singular vectors at this threshold. Including additional singular vectors will better

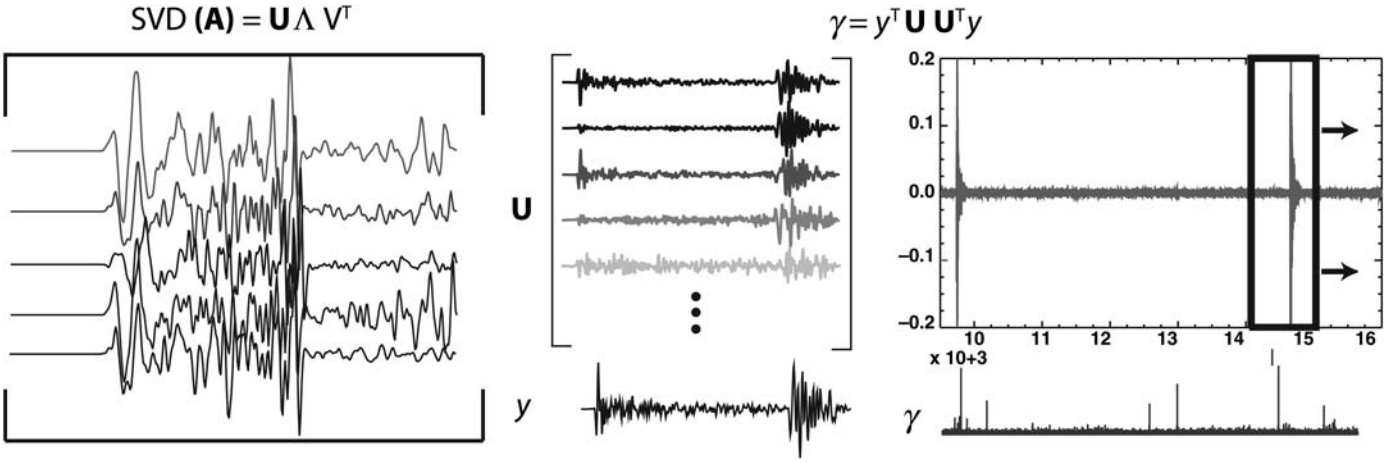


▲ **Figure 2.** Waveforms of the 66 events included in the design set cut 1 s before to 10 s after the *P*-wave arrival. Events are aligned through cross correlation to a master event on the *P*-wave arrival.

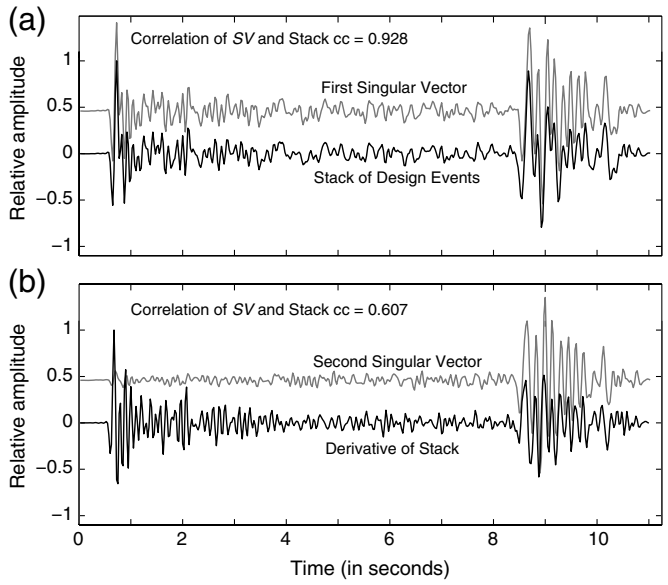
represent the total waveform variability within the sequence. Although this is not computationally prohibitive, we are interested in properties of the subspace detector for more strongly similar events and in particular the feasibility and utility of using only two singular vectors in the projection.

Examination of the singular vectors provides some insight into their physical representation. The largest singular vector very closely resembles the stack of the design set (Fig. 4a). Because the first singular vector should include the dominant information in all the events, this not a surprising result. The second singular vector (Fig. 4b) is strongly similar to the time derivative of the stack, especially for the *S*-wave arrival. T suggests the second singular vector represents variability in the seismogram that comes about from location differences in the events, or essentially variations in the *S*–*P* times. The *S*-wave arrival dominates the amplitude of this singular vector because the design set matrix **A** is constructed after *P*-wave alignment. If we examine a ~3 s window centered on the *S*-wave arrival, the correlation between the time derivative and second singular vector is even higher (~0.8). Successive singular vectors are not clearly related to higher order time derivatives and contain information, which is less representative of most of the events. Examination of the fractional energy capture suggests the first two singular vectors are sufficient to represent the design set.

We propose that the similarity of the first two singular vectors to the stack and time derivative of the stack can be exploited to improve template correlation techniques empirically without the need to construct a subspace detector through a SVD of the **A** matrix. To make this comparison, we apply three detectors to the continuous seismic record during the 2003 Big Bear sequence. The first detector is the stack of the design set events; this is intended to represent template matching, or a 1D subspace. The second is our empirical subspace detector; it is a matrix comprised of two rows, the stack, and the time derivative of the stack. Although some empirical weighing would be appropriate (analogous to subspace vectors being weighted by their respective singular values), the vectors in this example are weighted equally. Weighting may be more important for more dispersed seismicity. We normalize the stack detector and each row of the empirical detector to be of unit length. We suggest this is representative of the improvements that might be observed by using a matrix of a template waveform and its time derivative. The final detector is the 2D subspace detector determined using the SVD.



▲ **Figure 3.** A schematic of the subspace detector process. The matrix **A** is constructed as a matrix of cataloged waveforms, in which each row is an event. The singular value decomposition (SVD) of this matrix yields a matrix of left-singular vectors (**U**). These vectors are ordered by their contribution to the span of the events. In this work, we use only two dimensions of **U**. The projection of *y* onto **U** yields a normalized detection value (γ) that signifies how well the subspace detector represents the time series.

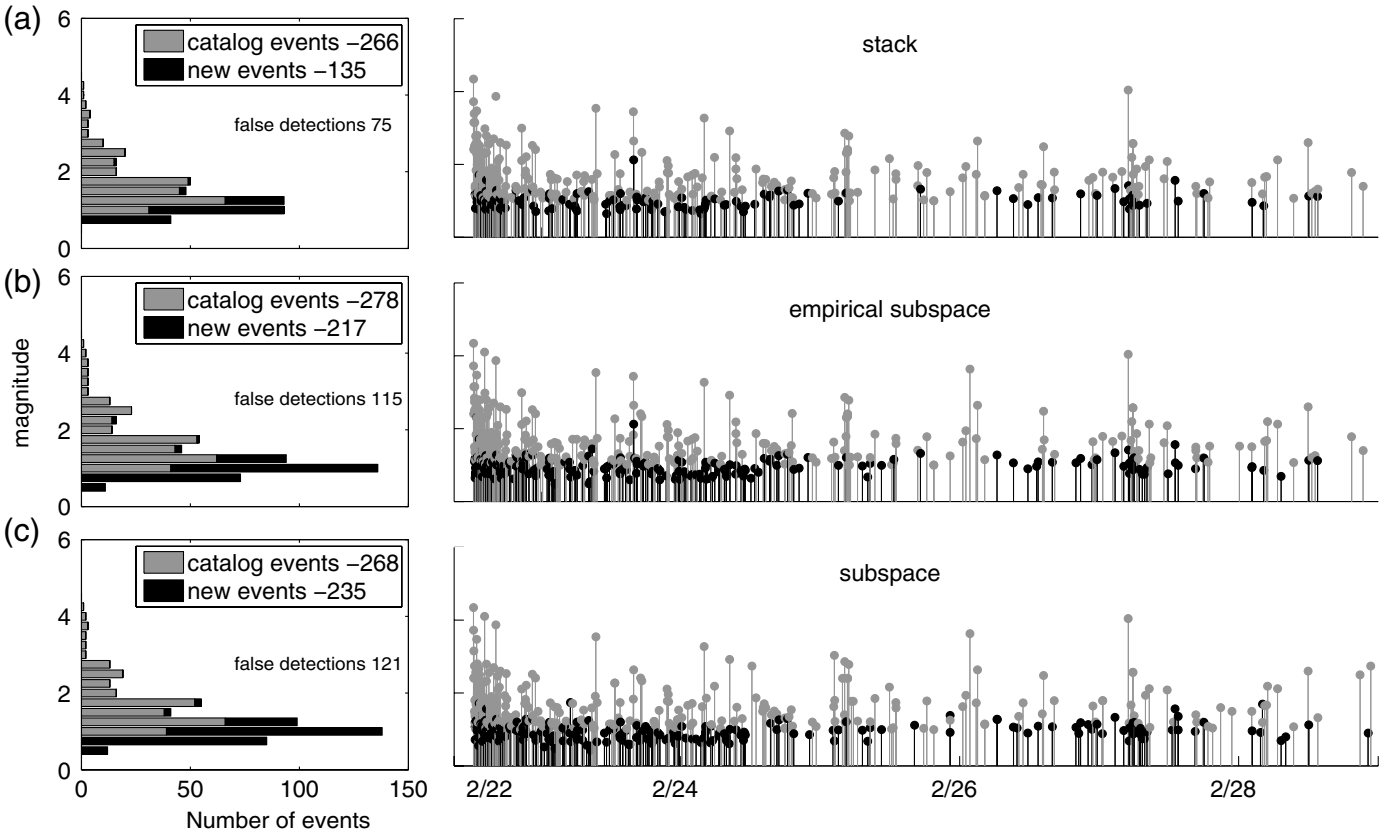


▲ **Figure 4.** (a) The largest singular vector shows (cc = 0.928) similarity to the stack of the 66 design set events. (b) The second largest singular vector resembles the time derivative of the design set stack (cc = 0.607), especially in the portion corresponding to the *S*-wave arrival. All waveforms are normalized to unit amplitude.

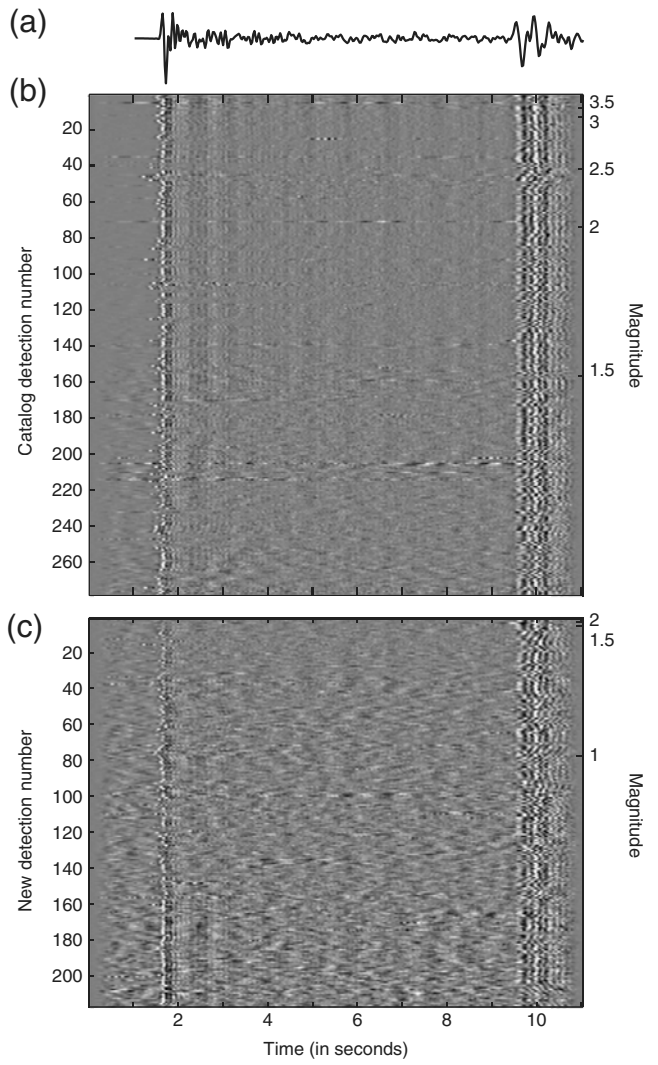
We apply the three detectors to seven days (the period for which there are more than 20 cataloged events per day) of the continuous 40 samples per second displacement record on the vertical channel at station KNW. The continuous signal is cut into windows, processed in an identical manner to the design set waveforms, and normalized to unit length. The dot product of the windowed continuous signal, and each detector is computed (γ) at each sample in the time domain. Values above an empirically determined threshold are recorded as detections and compared with the Advanced National Seismic System (ANSS) catalog for the sequence.

RESULTS

The resulting detections at a threshold of 0.125 for each of the detectors are shown in Figure 5. The stack finds a majority of cataloged events as well as many previously uncataloged events. The subspace detector finds more cataloged and uncataloged events; however, the number of false detections also increases slightly. The empirical subspace detector finds the most cataloged events of the three detectors and detects more uncataloged events than the stack detector alone with a similar failure-to-detect rate. The waveforms from the events identified by the empirical subspace detector are displayed in Figure 6.



▲ **Figure 5.** Results of the three detectors on one week of continuous data. (a) The results of the stack (representative of templates) detector. Left, distribution of cataloged (gray) and new (black) event detections, sorted by magnitude. False detections not automatically removed during processing are displayed. (b) The results of the empirical (representative of a template and its derivative) detector. (c) The results of the 2D subspace.



▲ **Figure 6.** Detections made using the Empirical Subspace Detector. Events are ordered by magnitude. (a) The stack of the design set events waveform. (b) Detections which already appeared in the catalog (278). (c) New detections (217) with false detections removed.

Each of these detections shown was verified by eye to have seismic characteristics (expected *P*- and *S*-arrival time and corresponding change in frequency content) consistent with events originating from the Big Bear region. The results for the empirical subspace detector are promising, as it detects more new events than would simple template matching, with an equivalent number of cataloged events, but does not increase the false detection rate. Consider the four possible outcomes for the three detectors.

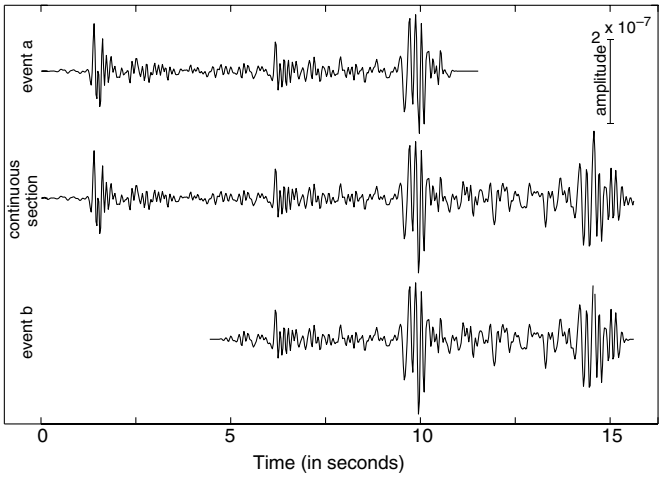
1. No detection in a waveform, which contains only noise (correct null hypothesis)
2. A detection of relevant signal in a waveform of only noise (Type I error)
3. A detection in which there is an event of interest (correct detection)
4. No detection in which there is an event of interest (Type II error)

The stack detector has a slightly lower percentage of Type I errors compared to both subspace-based detectors (~19% versus 23–24%). Exploring the use of an alternative detection threshold for the subspace-based detectors would mitigate this outcome. Following the fractional energy capture analysis, we chose to use a detection threshold of 0.125, which is sufficient to represent, at minimum, all the design set events. We could have adopted a higher detection threshold to minimize false detections; however, this would result in fewer detections, and we preferred to tolerate an increase in Type I errors in order detect as many earthquakes as possible.

All three detectors show similar Type II error rates—missing approximately the same number of ANSS cataloged seismic events. We believe this results from the variation in focal mechanism and fault perpendicular aftershocks, observed by Chi and Hauksson (2006). For example, an early aftershock (M_L 4.5) was determined to have a reverse mechanism. We would not expect such a source to produce a similar waveform, which would be represented in the chosen design set or well represented by the subspace of singular vectors that results, and thus its detection is not observed. Chi and Hauksson (2006) also show moderate populations of aftershocks occurring on planes perpendicular to the main fault plane. These variations in source are perhaps not well represented by the explored subspace.

Overlapping events are most prevalent early in the seismic sequence, and such events are often uncataloged because they occur within the coda of a mainshock or large aftershock, or when earthquake swarms get particularly active. Overlapping events are particularly problematic for energy-based detectors, and correlation detectors are preferred in such situations because they rely on the full waveform rather than detection of an abrupt onset. We observe several occasions in which the subspace-based detectors outperform the template detector (Fig. 7). Seismic events with overlapping waveforms may not even be detected using the stack alone unless the design set specifically includes a near repeat of that event, as in the case of the example shown.

The subspace detector has some disadvantages. For example, an analyst must assemble and design the population of waveform to input in matrix **A**, although automation of this process should be possible (Rowe *et al.*, 2012). The selection of the design set has direct consequences for the ability of a detector to perform well across diverse events. As illustrated by the Type II errors, the selection of a simple, single design set may not be optimal for a diverse population of events. For earthquake sequences of spatially dense seismicity, we advocate the empirical subspace, in which an event template and its derivative form a matrix that is scanned through the continuous record much like a traditional template. This eliminates the need to construct a design set matrix and perform an SVD. The use of the template time derivative acts as an approximation for the second singular vector and increases the variably in available templates, thereby improving detection of events that



▲ **Figure 7.** An example of pair (events a and b) overlapping events as detected by the empirical subspace detector. The *P* wave of event b arrives before the *S* wave of event a. The top and bottom panels show the arrivals of events a and b, respectively. The center panel illustrates the entire time series.

are similar, but not identical. In the example, we detail a near doubling of cataloged events using the empirical detector, with a substantial improvement when compared to template detection.

CONCLUSIONS

Template matching is an effective approach to detect small events. Its broad range of applicability includes earthquakes (Peng and Zhao, 2009), low-frequency earthquakes within tremor (Shelly *et al.*, 2007), induced seismicity (van der Elst *et al.*, 2013), and Test Ban Treaty verification (Gibbons and Ringdal, 2006). Our development of the empirical subspace detector has the potential to extend and improve upon template detection for the same range of seismic applications.

For the 2003 Big Bear sequence, continuously recorded data was not retained for most stations in southern California and is only available for events that are more recent. Station KNW, although distant from the sequence, is a high-performing station that recorded most of the events in the ANSS catalog for the sequence. Other stations in the region were found to be unsuitable based on a lack of continuous records, source to station distance, or low SNRs. Our detailed results are based on only a single component at a single station. We verify the results using a horizontal channel at station KNW and observe similar relative performances of each of the detectors. Data sets that are more extensive could be used for demonstrating the detection approach. The use of other available channels in combination (Harris, 2006) as well as validating detections across an array would improve performance and reduce both Type I and Type II errors. We show that although we use a single component at a station in isolation, the empirical subspace detector correctly identifies numerous uncataloged events.

The empirical detector ought to be especially useful for regions with sparse coverage. The empirical detector doubles the number of cataloged events we were able to identify (when compared to the template performance), yet Station KNW is ~60 km away from the Big Bear events. There are many regions of low background earthquake activity, such as the central and eastern United States, where station spacing will likely remain sparse. The proposal to keep approximately one in four of the EarthScope Transportable Array stations as a permanent network (Leith *et al.*, 2011) would improve current coverage. However, station spacing would still be on the order of ~140 km (maximum station–event distance ~70 km). This distance is similar to that for the example presented in this paper, and suggests empirical subspace detectors might provide an effective and important extension to standard template matching practices for earthquake monitoring.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation (NSF) Grant Number EAR-1045684. It was also supported by the Southern California Earthquake Center (SCEC). SCEC is funded by NSF Cooperative Agreement EAR-0529922 and U.S. Geological Survey (USGS) Cooperative Agreement 07HQAG0008. The SCEC contribution number for this paper is 1800.

Sarah A. Barrett is supported by the Northern California Chapter of the Achievement Rewards for College Scientists Foundation.

Waveforms and continuous data for this research were obtained using the Standing Order for Data (SOD, Owens *et al.*, 2004) from the Incorporated Research Institutions for Seismology (IRIS) Data Management Center at www.iris.edu (last accessed June 2013). Figure 1 was made in part using the Generic Mapping Tools (GMT, Wessel and Smith, 1998).

REFERENCES

Brown, J. R., G. C. Beroza, and D. R. Shelly (2008). An autocorrelation method to detect low frequency earthquakes within tremor, *Geophys. Res. Lett.* **35,** L16305, doi: 10.1029/2008GL034560.

Chi, W.-C., and E. Hauksson (2006). Fault-perpendicular aftershock clusters following the 2003 M_w = 5.0 Big Bear, California, earthquake, *Bull. Seismol. Soc. Am.* **33,** L07301, doi: 10.1029/2005GL025033.

Gibbons, S. J., and F. Ringdal (2006). The detection of low magnitude seismic events using array-based waveform correlation, *Geophys. J. Int.* **165,** 149–166.

Harris, D. (2006). Subspace detectors: Theory, *Lawrence Livermore Natl. Lab. Rep. UCRL-TR-222758*, Lawrence Livermore National Laboratory, Livermore, California.

Harris, D., and T. Paik (2006). Subspace detectors: Efficient implementation, Theory, *Lawrence Livermore Natl. Lab. Rep. UCRL-TR-223177*, Lawrence Livermore National Laboratory, Livermore, California.

Jones, L. E., S. E. Hough, and D. V. Helmberger (1993). Rupture process of the June 28, 1992 Big Bear Earthquake, *Geophys. Res. Lett.* **20,** 1907–1910.

Leith, W. S., H. M. Benz, and R. B. Herrmann (2011). Improved earthquake monitoring in the central and eastern United States in

support of seismic assessments for critical facilities, *U.S. Geol. Surv. Open-File Rept. 2011-1101*, 29 pp.
Maceira, M., C. A. Rowe, G. Beroza, and D. Anderson (2010). Identification of low-frequency earthquakes in non-volcanic tremor using the subspace detector method, *Geophys. Res. Lett.* **37,** L06303, doi: 10.1029/2009GL041876.
Owens, T. J., H. P. Crotwell, C. Groves, and P. Oliver-Paul (2004). SOD: Standing order for data, *Seismol. Res. Lett.* **75,** 515–520.
Peng, Z., and P. Zhao (2009). Migration of early aftershocks following the 2004 Parkfield earthquake, *Nat. Geosci.* **2,** 877–881.
Rowe, C. A., R. J. Stead, M. L. Begnaud, and E. A. Morton (2012). Seismic signal analysis for event detection and categorization, in *2012 Monitoring Research Review: Ground-Based Nuclear Explosion Monitoring Technologies* 312–320.
Rubinstein, J. L., and W. L. Ellsworth (2010). Precise estimation of repeating earthquake moment: Example from Parkfield, California, *Bull. Seismol. Soc. Am.* **100,** 1952–1961.
Schaff, D. P., G. H. R. Bokelmann, W. L. Ellsworth, E. Zanzerkia, F. Waldhauser, and G. C. Beroza (2004). Optimizing correlation techniques for improved earthquake location, *Bull. Seismol. Soc. Am.* **94,** 705–721.
Shelly, D. R., G. C. Beroza, and S. Ide (2007). Non-volcanic tremor and low-frequency earthquake swarms, *Nature* **446,** 305–307.
Shelly, D. R., G. C. Beroza, H. Zhang, C. H. Thurber, and S. Ide (2006). High-resolution Subduction zone seismicity and velocity structure beneath Ibaraki Prefecture, Japan, *J. Geophys. Res.* **111,** B06311–B06320.
van der Elst, N. J., H. M. Savage, K. M. Keranen, and G. A. Abers (2013). Enhanced remote earthquake triggering at fluid-injection sites in the midwestern United States, *Science* **341,** no. 6142, 164–167.
Van Trees, H. L. (1968). *Detection, Estimation and Modulation Theory*, Vol. 1, John Wiley and Sons, New York.
Wessel, P., and W. H. F. Smith (1998). New, improved version of the Generic Mapping Tools released, *Eos Trans. AGU* **79,** 579.
Yang, W., E. Hauksson, and P. M. Shearer (2012). Computing a large refined catalog of focal mechanisms for southern California (1981–2010): Temporal stability of the style of faulting, *Bull. Seismol. Soc. Am.* **102,** 1179–1194.

Sarah A. Barrett
Gregory C. Beroza
Department of Geophysics
Stanford University
Stanford, California 94305-2215 U.S.A.
sabarrett@stanford.edu