Geophysical Journal International

*Geophys. J. Int.* (2013) **195,** 2023–2036
Advance Access publication 2013 October 3

doi: 10.1093/gji/ggt373

# Stacking with dual bootstrap resampling

Jun Korenaga

*Department of Geology and Geophysics, Yale University, New Haven, 06520, CT, USA. E-mail: jun.korenaga@yale.edu*

Accepted 2013 September 16. Received 2013 September 16; in original form 2013 May 23

## SUMMARY

A new kind of stacking scheme, based on the hypothesis testing of signal significance and coherence, is proposed. The significance of stacked data is evaluated by running two kinds of bootstrap resampling, one for standard bootstrap and the other for preparing noise stacks by scrambling relative time-shifts between traces. This dual bootstrap procedure allows us to formulate a two-sample problem for signal significance, which is shown to be more reliable than standard bootstrap estimates. The statistics of noise obtained in dual bootstrap resampling is also used when assessing the coherence of data with the empirical distribution function, in which the effect of noise is deconvolved by rescaling. Unlike conventional non-linear stacks such as *N*th-root stack and phase-weighted stack, the new stack can recover signals even when the signal-to-noise ratio (S/N) is low, and compared to simple linear stack, the number of traces required for unambiguous signal detection is reduced by up to two orders of magnitude. The new scheme, called dual bootstrap stack, could facilitate a range of geophysical data processing when trying to detect subtle signals by stacking low S/N data.

**Key words:** Time-series analysis; Probability distributions; Computational seismology.

## 1 INTRODUCTION

Stacking is one of elementary data processing techniques for noise reduction and signal detection. Its use has a long history in active-source seismology (e.g. Yilmaz 1987; Sheriff & Geldart 1995) but has also become common in the passive-source counterpart (e.g. Shearer 1991; Rost & Thomas 2002). The premise of stacking comes from the statistics for the average of a random sample. Suppose that $X_1(t), X_2(t), \ldots, X_n(t)$ represent time-series data recorded at $n$ receivers. They may be regarded as the sum of two components as

$$X_i(t) = Y(t) + Z_i(t), \tag{1}$$

where $Y$ denotes a signal component, which does not vary across receivers, and $Z_i$ is a noise component, which is assumed here as a random variable with zero mean and variance $\sigma_{\mathrm{N}}^2$. The so-called linear stack is a simple arithmetic average of $X_i(t)$, which has expectation and variance as

$$\mathrm{E}[\overline{X}(t)] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i(t)\right] = Y(t), \tag{2}$$

and

$$\mathrm{var}[\overline{X}(t)] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(Z_i) = \frac{\sigma_{\mathrm{N}}^2}{n}. \tag{3}$$

As the number of receivers increases, therefore, the standard deviation of the noise component decreases in proportion to $n^{-1/2}$. Unlike filtering, stacking allows noise reduction without attenuating signal even when signal and noise are in the same frequency range.

A fairly large $n$ is required, however, to reduce noise sufficiently when the signal-to-noise ratio (S/N) is low. In reference to eqs (1)–(3), S/N is defined in this paper as

$$\mathrm{S/N} \equiv \frac{\max(|Y|)}{\sigma_{\mathrm{N}}}. \tag{4}$$

Fig. 1 shows synthetic seismic data with $n = 40$ and with S/N ranging from 10 to 0.5. The signal component of the synthetic data is composed of four Ricker wavelets with the peak frequency of 0.2 Hz, so the power of signal is contained largely in the frequency range of 0.1–0.4 Hz. The first and third Ricker wavelets are centred at 10 and 35 s, respectively, and their amplitudes are set to fluctuate from trace to trace by 1 and 30 per cent, respectively. This intrinsic amplitude variability is not formulated in eq. (1) for the sake of simplicity; it is introduced to make the synthetic example more realistic but does not affect the overall trend of stacking performance because the range of S/N considered here is fairly wide. The second wavelet, with an intrinsic amplitude variability of 5 per cent, has a constant moveout, migrating from 10 to 30 s through the traces. The fourth wavelet is centred at 50 s and has a different type of amplitude variability, flipping between positive and negative polarities; this wavelet is used to demonstrate the performance of stack for an incoherent data set. The noise component of the synthetic data is prepared by generating Gaussian white noise by a pseudo-random number generator, bandpass-filtering it through the range of 0.1–0.5 Hz, and then adjusting its root-mean-square amplitude to match a given S/N. Because both signal and noise share the same frequency

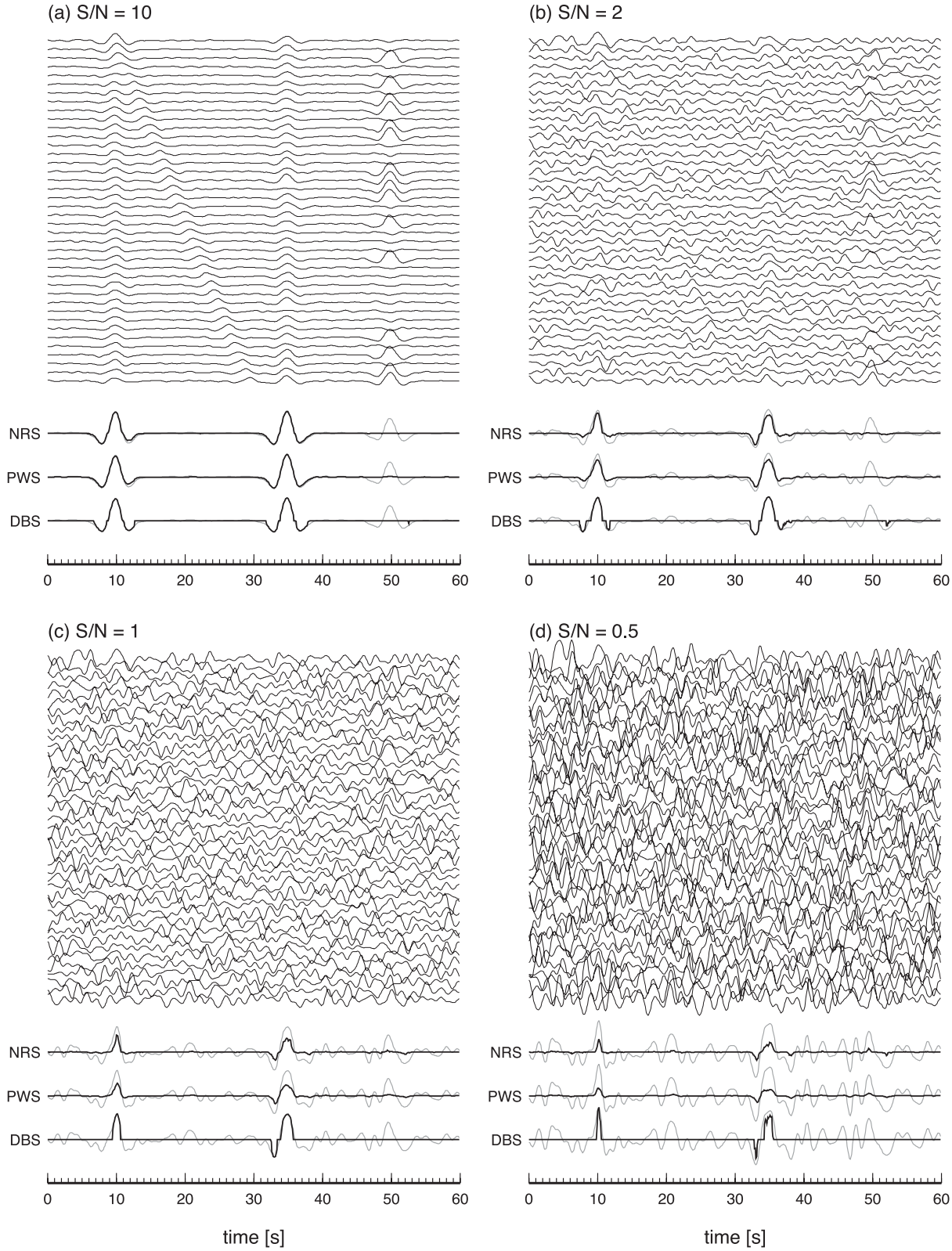


**Figure 1.** Performance of linear and non-linear stacks on synthetic data with the signal-to-noise ratio of (a) 10, (b) 2, (c) 1 and (d) 0.5. The signal component of the data is the same for all cases; only the amplitude of the noise component is different. Three non-linear stacks, *N*th-root stack with the power of 3 (NRS), phase-weighted stack with the order of 2 (PWS) and dual bootstrap stack (DBS), are shown with linear stack (in grey).

range, filtering cannot be used for noise reduction. It can be seen that, in the case of $n = 40$, residual noise after stacking has non-trivial amplitudes when S/N becomes lower than ∼2. Also, being simple averaging, linear stack can exhibit substantial amplitudes for an incoherent data set, regardless of S/N (Fig. 1, $t = 50$ s).

These deficiencies can be alleviated to some extent by non-linear stack such as *N*th-root stack (NRS; Muirhead 1968; Kanasewich *et al.* 1973; McFadden *et al.* 1986) and phase-weighted stack (PWS; Schimmel & Paulssen 1997). NRS is defined as

$$\overline{X}_{\rm NRS} = {\rm sgn}(V)|V|^N, \tag{5}$$

where sgn($x$) is a sign of $x$ and

$$V = \frac{1}{n}\sum_{i=1}^{n} {\rm sgn}(X_i)|X_i|^{1/N}. \tag{6}$$

The power $N$ is usually an integer greater than 1. Taking *N*th root reduces the amplitude differences of samples, and as a result, coherent samples (i.e. samples with the same sign) would survive in NRS whereas incoherent samples would be considerably attenuated (Fig. 1). PWS achieves a similar effect by utilizing the phase information of the analytic signal corresponding to given data. There is no practical difference between the performance of these two non-linear stacking schemes (Fig. 1); PWS is slightly more involved because one has to calculate the analytic signal by the Fourier transform. Throughout this paper, NRS is done with the power of 3, and PWS is done with the order of 2; results shown in this paper remain virtually the same with different exponents.

These non-linear stacks, however, perform well only when S/N is greater than ∼1. They rely on the coherence of samples, so when data are dominated by the noise component (i.e. S/N < 1), the overall coherence decreases, reducing the power of signal and noise altogether in the stack (Figs 1c and d). Herein lies a dilemma. In terms of recovering the amplitude of signals, simple linear stack still provides good results, but with an insufficient number of low S/N traces, it does not reduce noise sufficiently, so signal detection would be unreliable. Linear stack is also vulnerable for incoherent data. NRS or PWS can reduce random noise as well as incoherent signals efficiently even with a relatively small number of samples, but they do not preserve signal power when S/N is low.

The purpose of this paper is to introduce a new stacking scheme, which can mimic the performance of NRS and PWS even for low S/N data. The new scheme is based on statistical hypothesis testing, and because of its use of two kinds of bootstrap resampling, it is referred to as dual bootstrap stack (DBS) in this paper. In what follows, the new scheme is first described using the synthetic data shown in Fig. 1. More systematic tests of the new scheme are then presented, and its performance is quantified in comparison with linear stack, NRS and PWS. An example with real seismic data is also given. Possible extensions of DBS are discussed at the end.

## 2 METHOD

The new stacking scheme is implemented by scaling linear stack on the basis of two statistical hypothesis tests, the one for the significance of a stacked result, and the other for the coherence of samples to be stacked. Though previous non-linear stacks such as NRS and PWS do not distinguish between these issues, they are different statistical problems and thus are better handled separately. Before describing the new method and relevant statistics, however, it may be instructive to first consider a standard statistical approach based on bootstrap resampling (e.g. Efron 1982). In this approach, the actual data set $X_1, X_2, \ldots, X_n$ is regarded as an empirical distribution function (EDF), from which a bootstrap replicate $X_1^*, X_2^*, \ldots, X_n^*$ can be constructed by randomly sampling $n$ times with replacement. By generating a large number of bootstrap replicates, one can estimate a probability distribution function for the sample average $\overline{X}$, based on which confidence intervals may be drawn. Calculating bootstrap confidence intervals for stacked data is common in seismology (e.g. Revenaugh & Meyer 1997; Margerin & Nolet 2003; Hutko *et al.* 2008).

The number of bootstrap replicates has to be on the order of $10^3$ for accurate confidence intervals (e.g. Efron & Tibshirani 1993), and the 95 and 99 per cent confidence intervals based on 2000 bootstrap replicates for the synthetic data presented in Section 1 are shown in Fig. 2. Based on such probability distribution, one may devise a weighted stack as

$$\overline{X}_\alpha = \overline{X}\,\max\left\{0, 1 - \frac{p[{\rm sgn}(\overline{X}^*) \neq {\rm sgn}(\overline{X})]}{\alpha}\right\}, \tag{7}$$

where $p$ stands for probability and $\alpha$ is the critical significance level. For $\alpha = 0.01$, for example, the weighted stack takes zero amplitude if more than 1 per cent of bootstrap averages $\overline{X}^*$ have a different sign than the original average $\overline{X}$. Two examples, $\overline{X}_{0.05}$ and $\overline{X}_{0.01}$, are shown in Fig. 2; whereas this weighting scheme fails to reduce the incoherent data stack around $t = 50$ s, it does remove a substantial fraction of residual noise elsewhere without attenuating the peak amplitude of the true signal. Bootstrap-based weighting thus appears to be promising, though its performance is not quite satisfactory for low S/N data. Even with the tight significance level of $\alpha = 0.01$, noise removal is not perfect, and as more systematic tests will show (Section 3), the residual noise level of this weighting scheme is as high as ∼20–30 per cent. When the size of a random sample is not large, it is possible for the majority of the sample to have the same sign, resulting in false statistical significance by bootstrap resampling.

Fortunately, the significance of stacked data can be assessed more reliably by formulating an appropriate two-sample problem. The coherence of samples can be evaluated by formulating another kind of statistical test. The new stacking scheme, DBS, is a doubly weighted stack based on these statistical tests as

$$\overline{X}_{\rm DBS} = \overline{X}\; w_1\, w_2, \tag{8}$$

where $w_1$ and $w_2$ denote weighting factors based on, respectively, signal significance and coherence. In what follows, how to compute these weighting factors are explained in turn.

### 2.1 Significance of stacked data

Stacking in its simplest form is merely calculating the mean of a sample. If the stacked data $\overline{X}$ is different from zero, then, testing its statistical significance is equivalent to asking whether the non-zero $\overline{X}$ is obtained by chance or not. Testing the significance of a non-zero mean is a classical problem in statistics, and examples in textbooks are often taken from pharmaceutical applications, in which the effect of a new drug is tested by comparing the statistics of a treatment group with that of a control group. In the same spirit, the significance of stacked data may be better quantified not by looking at the statistics of $\overline{X}$ alone, but by comparing it to the statistics of noise. Estimating the statistics of noise, however, is challenging for at least two reasons. First, it is nearly impossible

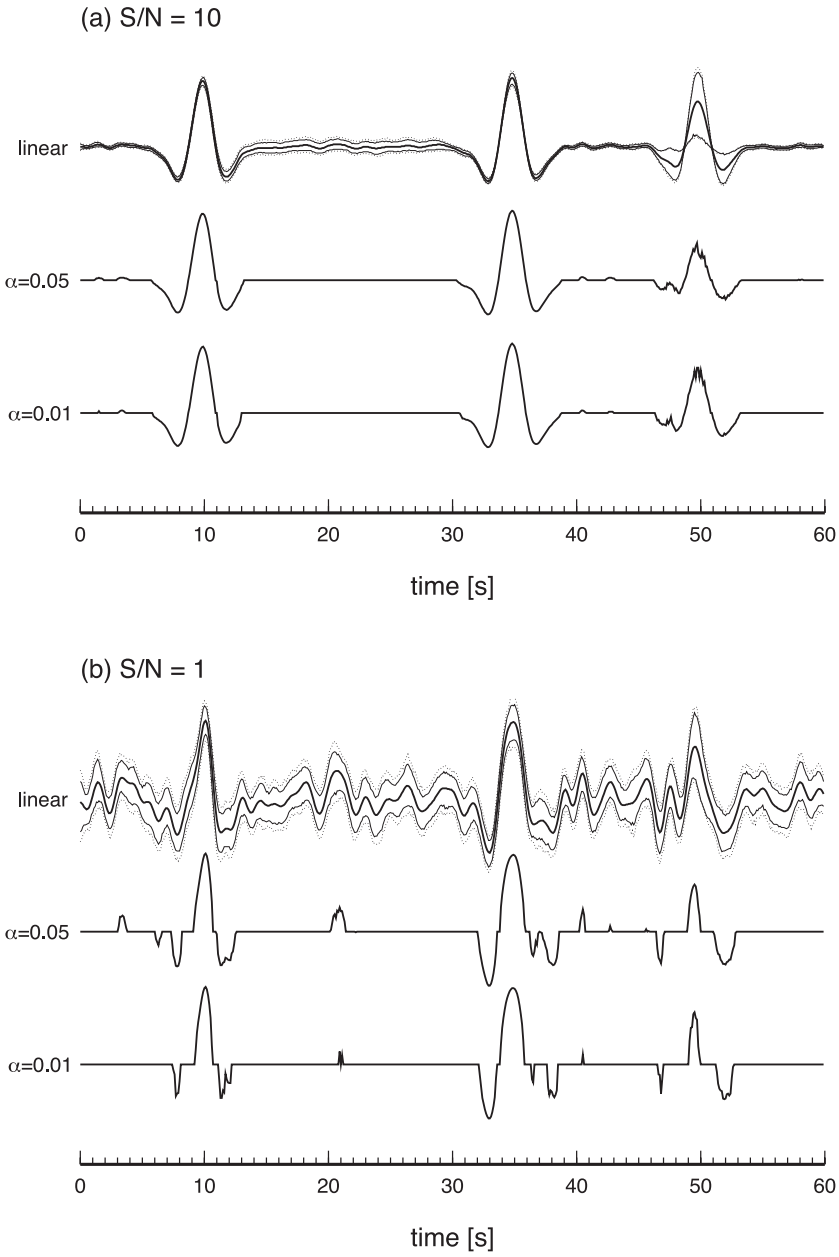


**Figure 2.** Linear stack (thick solid) with bootstrap confidence intervals (thin for 95 per cent and dotted for 99 per cent) for the synthetic data shown in Fig. 1. The weighted stack according to eq. (7) is shown for $\alpha = 0.05$ and $\alpha = 0.01$.

to tell *a priori* which part of data contains only noise when S/N is low, but without being able to do so, the statistics of noise cannot be estimated. Secondly, this difficulty is even more compounded by the non-stationary nature of signal-generated noise in seismic data.

A control group can still be prepared for stacking, by scrambling relative time-shifts between different traces (Fig. 3). If there is a signal contained in a given data set, the original (zero) time-shift between traces is optimal to stack the signal, and by randomizing the time-shift, the signal would be lost by stacking. In other words, a noise stack can easily be generated just by scrambling traces. This is somewhat reminiscent of block resampling and phase scrambling, both of which are used for the bootstrap resampling of time-series data (e.g. Davison & Hinkley 1997). One may note that, if the original data set contains a signal with non-zero moveout (e.g. the one in the range of $t = 10$–$30$ s in Fig. 1a), there is a finite probability for it to be aligned straight in a scrambled set so that a supposedly noise stack can yield a clear signal. This does not pose a problem because a two-sample test for this case would indicate that the original stack does not contain a statistically significant signal, which coincides with a desired diagnosis.

There are an infinite number of ways to scramble traces and generate a noise stack. It may thus be done most conveniently by incorporating it into bootstrap resampling. That is, we create a large number of noise stacks by repeating trace scrambling. The significance of stacked data can then be assessed by testing a null hypothesis that the original stack and these noise stacks are drawn from the same distribution. By combining the standard bootstrap procedure for testing such a two-sample problem (Efron & Tibshirani 1993) with trace scrambling, a bootstrap test statistic for stacking at the time of $t_0$ may be implemented as follows:

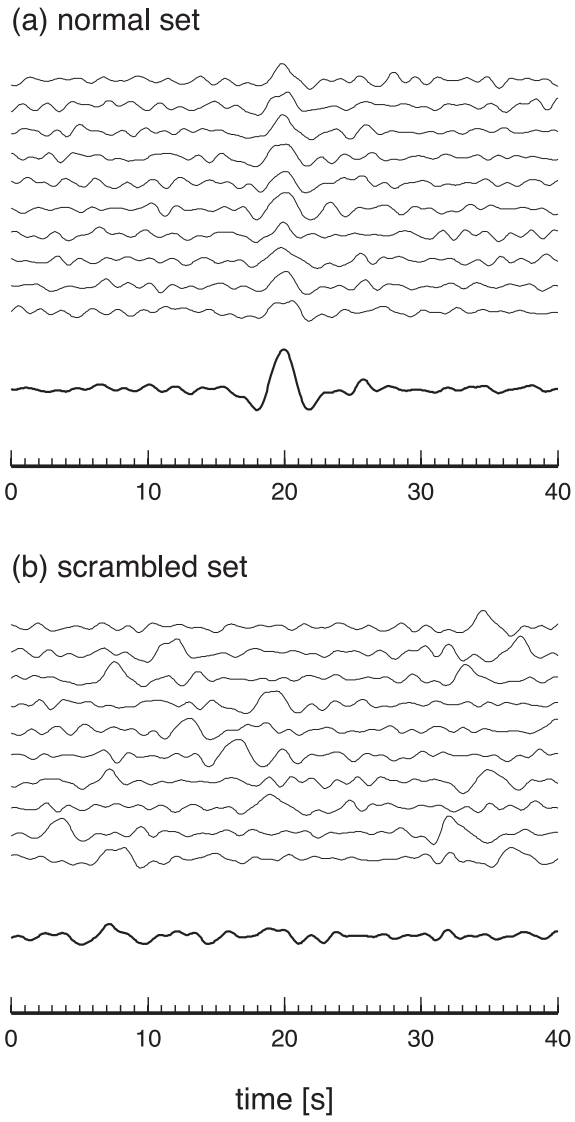


**Figure 3.** The significance of a normal stack (a) can be assessed by comparing it to some kind of noise stack, which can be generated by scrambling relative time offsets between traces as shown in (b).

(1) Randomly draw $n$ integers, $i_1, i_2, \ldots, i_n$, from $\{1, 2, \ldots, n\}$ with replacement, and construct a bootstrap replicate, $\mathbf{X}^*{}_b = \{X_{i_1}(t_0), X_{i_2}(t_0), \ldots, X_{i_n}(t_0)\}$.

(2) Randomly draw $n$ real numbers, $r_1, r_2, \ldots, r_n$, from the interval $[-1, 1]$, and construct a scrambled bootstrap replicate, $\mathbf{Z}^*{}_b = \{X_{i_1}(t_0 + Tr_1), X_{i_2}(t_0 + Tr_2), \ldots, X_{i_n}(t_0 + Tr_n)\}$, where $T$ is the maximum period for time-shifting. Also calculate its mean and call it $\overline{Z}^*_b$.

(3) Combine the above two replicates to form a data set of size $2n$ and randomize its order. Call the first $n$ data $\mathbf{x}^*$ and the remaining $n$ data $\mathbf{z}^*$.

(4) Evaluate the difference between the means of $\mathbf{x}^*$ and $\mathbf{z}^*$ as

$$D(\mathbf{X}^*{}_b) = \overline{x}^* - \overline{z}^*. \tag{9}$$

(5) Repeat (1)–(4) $B$ times and approximate the achieved significance level by

$$p_1 = \begin{cases} \#\{D(\mathbf{X}^*{}_b) > D_{\rm obs}\}/B, & \text{if } D_{\rm obs} > 0, \\ \#\{D(\mathbf{X}^*{}_b) < D_{\rm obs}\}/B, & \text{otherwise,} \end{cases} \tag{10}$$

where $\#\{\cdot\}$ denotes the number of occurrences for a given condition and the observed value of the statistic is given by

$$D_{\rm obs} = \overline{X} - \frac{1}{B}\sum_{b=1}^{B} \overline{Z}^*_b. \tag{11}$$

Here, $\overline{X}$ is the mean of $\mathbf{X} = \{X_1(t_0), X_2(t_0), \ldots, X_n(t_0)\}$, and $D_{\rm obs}$ measures how this linear stack deviates from the background noise level.

In the standard bootstrap procedure as considered at the beginning of this Method section, the $n$ data in a bootstrap replicate generated in the step (1) are simply stacked, and this step is repeated $B$ times to estimate the probability distribution function of a stack. In the dual bootstrap procedure above, the $n$ data in the bootstrap replicate is randomly mixed with another $n$ 'noise' data generated in the step (2), and the stack of the first $n$ and that of the second $n$ are compared in the step (3). This mixing with noise data is a key in the two-sample test. If the observed stack $\overline{X}$ is statistically significant, such significance would be destroyed by the random mixing with noise data in the step (3), leading to a low $p_1$ value in the step (5). Conversely, if the observed stack is indistinguishable from noise, it is likely for the bootstrap statistic $D(\mathbf{X}^*{}_b)$ to have a similar value to the observed value $D_{\rm obs}$, so the probability of the null hypothesis $p_1$ would become high; the observed stack would then be judged to be statistically insignificant as expected.

A weighted stack based on this two-sample test may be defined as

$$\overline{X}_{\rm TS} = \overline{X} w_1, \tag{12}$$

where the weight $w_1$ is equal to $\max(0, 1 - p_1/\alpha)$ and $\alpha$ is the critical significance level (typically set to 1 or 5 per cent) below which the null hypothesis is rejected. An example using the synthetic data with S/N of 1 is shown in Fig. 4. Only the portion of data up to $t = 45$ s is considered here because the coherence of data is out of the scope of the two-sample test and will be treated separately (Section 2.2). In this example, the number of bootstrap replicates $B$ is 2000, the maximum period $T$ is 20 s and the critical significance level is set to 0.01. The true signals centred at $t$ of 10 and 35 s are largely retained whereas residual noise in the linear stack is completely removed by this weighting. The residual noise that was difficult to remove on the basis of bootstrap confidence intervals (Fig. 2b) is clearly identified as insignificant by the two-sample test (Fig. 4c).

The above weighted stack has three control parameters, $\alpha$, $T$ and $B$. The critical significance level $\alpha$ should reflect the user's desired level of confidence in signal detection; lower $\alpha$ leads to a more stringent significance test. The maximum period for scrambling $T$ should be chosen to be large enough with respect to the dominant period of signals to be detected. Obviously, too small $T$ results in trivial scrambling, so there would not be much difference between $\mathbf{X}^*{}_b$ and $\mathbf{Z}^*{}_b$. As long as $T$ is greater than the dominant period of signals (about a few seconds in the case of the synthetic data shown in Fig. 1), the weighted stack would yield similar results (Fig. 5), though $T$ should not be too large for the two-sample testing to cope with the non-stationary nature of noise. Note that for this scrambling to work effectively, expected signals must not be characterized by two or more drastically different frequencies. In other words, data have to be band-limited, which can easily be achieved by filtering.

Given that the two-sample testing involves the estimate of a probability distribution, the number of bootstrap replicates $B$ should be on the order of $10^3$. A convergence test with the synthetic data indicates, however, that $B$ could be on the order of $10^2$ to obtain reasonable results (Fig. 6). This is probably a case-dependent issue,

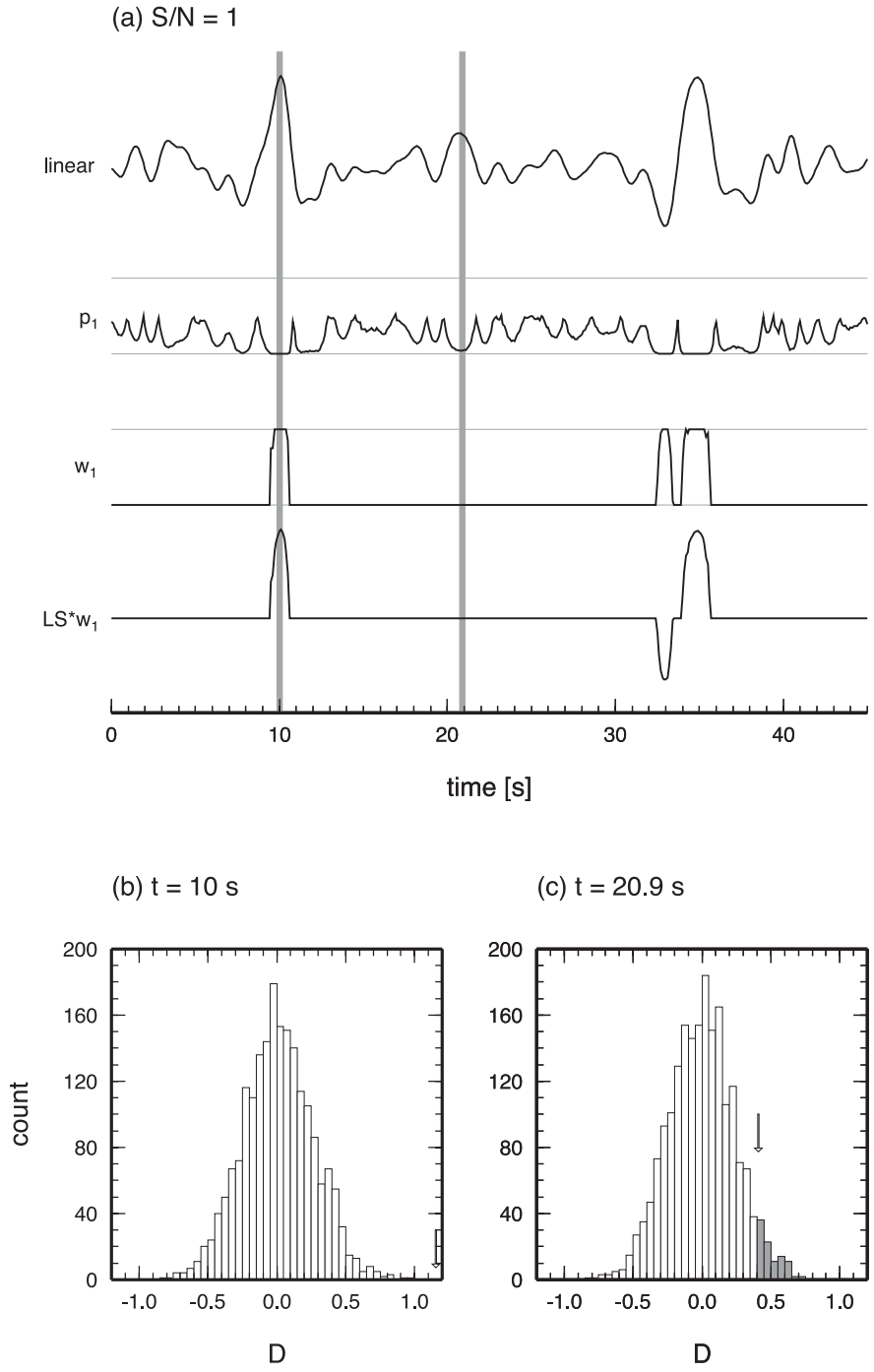


**Figure 4.** (a) Weighting of linear stack according to the two-sample test (eq. 12) using the synthetic data shown in Fig. 1(c). Thin horizontal lines for $p_1$ and $w_1$ denotes their entire range, that is [0,1]. Grey vertical bars correspond to (b) $t = 10$ s and (c) 20.9 s, at which the histogram of the test static $D$ is given. Arrows denote the location of $D_{\rm obs}$.

warranting a more careful look in specific applications in future, because an order-of-magnitude reduction in computation may be achieved by properly choosing $B$.

### 2.2 Statistical measure of coherence

An incoherent data set does not mean a statistically insignificant signal, so a different statistical measure is needed if one wants to attenuate the influence of such data on a stacked result. How NRS and PWS attenuate incoherent data is based simply on the statistics of signal polarity. A data set is said to be coherent when the majority of data have the same polarity, and such statistics can be depicted most simply by the EDF of a given data set (Fig. 7)

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} H(x - X_i), \tag{13}$$

where $H(x)$ is the Heaviside step function. A coherent data set may thus be defined as a data set whose EDF is contained mostly in either

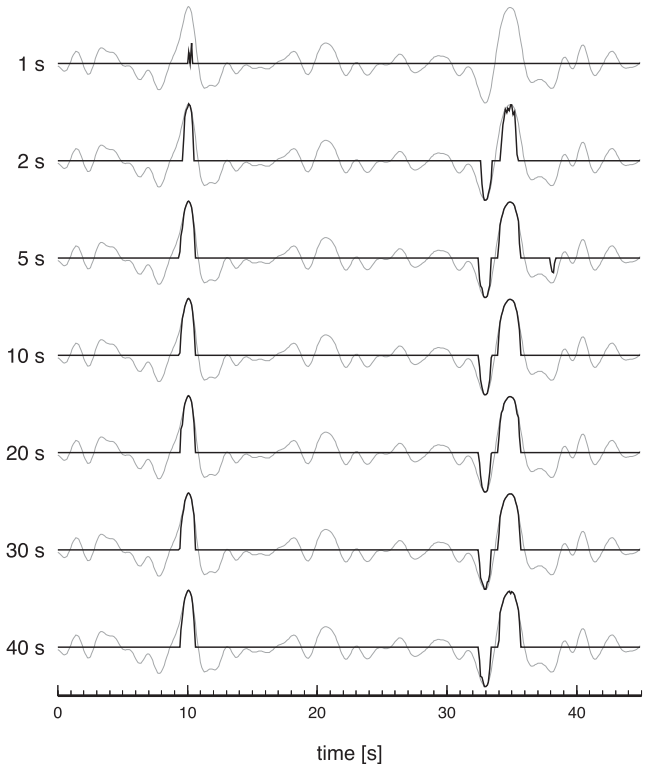


**Figure 5.** Effect of varying the maximum period $T$ for the weighted stack considered in Fig. 4. The other control parameters are fixed as $\alpha = 0.01$ and $B = 2000$. Shown in grey is linear stack for comparison.

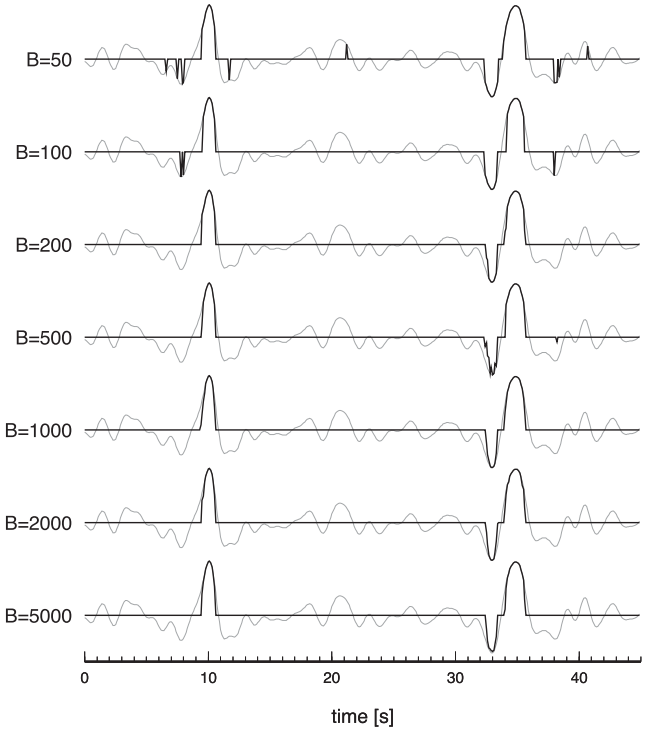


**Figure 6.** Effect of varying the number of bootstrap replicates $B$ for the weighted stack considered in Fig. 4. The other control parameters are fixed as $\alpha = 0.01$ and $T = 20$ s. Shown in grey is linear stack for comparison.

$x > 0$ or $x < 0$, and the significance level of the null hypothesis that a given data set is incoherent may be computed as

$$p_2 = \begin{cases} F(0), & \text{if } \overline{X} > 0, \\ 1 - F(0), & \text{otherwise.} \end{cases} \tag{14}$$

The direct use of EDF is appropriate, however, only for high S/N data. When S/N is low, EDF becomes more diffuse as a result of convolution with a noise EDF, and the distinction between coherent and incoherent data sets is blurred (Fig. 7). NRS and PWS fail for low S/N data for the same reason; the direct polarity statistics cannot tell the coherence of a signal buried deep in noise.

One way to remedy the situation is to apply deconvolution. Assuming Gaussian noise, the cumulative distribution function corresponding to the signal–noise model of eqs (1)–(3) may be written as

$$F(x) = \int_{-\infty}^{\infty} H(\xi - Y)\Phi_{(0,\sigma_{\mathrm{N}}^2)}(x - \xi)\,\mathrm{d}\xi, \tag{15}$$

where $\Phi_{(0,\sigma_{\mathrm{N}}^2)}(x)$ denotes the cumulative distribution function for the normal distribution with zero mean and variance $\sigma_{\mathrm{N}}^2$. When a signal itself has the intrinsic variability of $\sigma_{\mathrm{S}}^2$, it may be generalized to

$$F(x) = \int_{-\infty}^{\infty} \Phi_{(Y,\sigma_{\mathrm{S}}^2)}(\xi)\Phi_{(0,\sigma_{\mathrm{N}}^2)}(x - \xi)\mathrm{d}\xi, \tag{16}$$

in which the signal variability is assumed to be Gaussian as well. It is possible to deconvolve the normal cumulative distribution function by the method of simulation extrapolation Stefanski & Bay (1996), if an estimate on $\sigma_{\mathrm{N}}^2$ is available. The results of simulation extrapolation deconvolution are shown in Fig. 7 (dashed line), for which $\sigma_{\mathrm{N}}^2$ is estimated by averaging the variances of scrambled bootstrap replicates $\mathbf{Z}^*{}_{\mathrm{b}}$. It is found that the deconvolution works well only when S/N is sufficiently high. The unsatisfactory performance may be explained by an insufficient number of samples as well as an inaccurate estimate of $\sigma_{\mathrm{N}}^2$. Also, simulation extrapolation deconvolution is unattractive from the perspective of computational cost; it requires the computation of error-inflated cumulative distributions and post-processing such as isotonic regression, which may be too much to be conducted for each instance of $t$.

As a more computationally efficient and potentially more robust alternative, the following rescaling of the EDF is considered:

$$F'(x) = \frac{1}{n}\sum_{i=1}^{n} H(x - X_i'), \tag{17}$$

where

$$X_i' = \overline{X} + (X_i - \overline{X})\left[\frac{\min(0, \sigma_X^2 - \sigma_{\mathrm{N}}^2)}{\sigma_X^2}\right]^{1/2}, \tag{18}$$

and $\sigma_X^2$ denotes the variance of $\mathbf{X}$. This rescaling is equivalent to deconvolution, when both signal and noise follow the normal distribution and when they are uncorrelated. That is, it tries to recover a normal distribution with the mean of $\overline{X}$ and the variance of $\sigma_{\mathrm{S}}^2$, by scaling down the original spread of data. The minimum function is used in the above to reject the case of $\sigma_{\mathrm{N}}^2$ being greater than $\sigma_X^2$, which could happen if the estimate of noise variance by trace scrambling is not accurate enough. The results of rescaling are also shown in Fig. 7 (dotted line). Rescaling works particularly well for coherent data sets (e.g. at $t$ of 10 s) and, although not expected to restore non-Gaussian cumulative distributions accurately, it still serves the purpose to detect incoherent data sets (e.g. at $t$ of 50 s) because the extent of variance reduction tends to be limited for such data, thereby leading to high $p_2$ values.

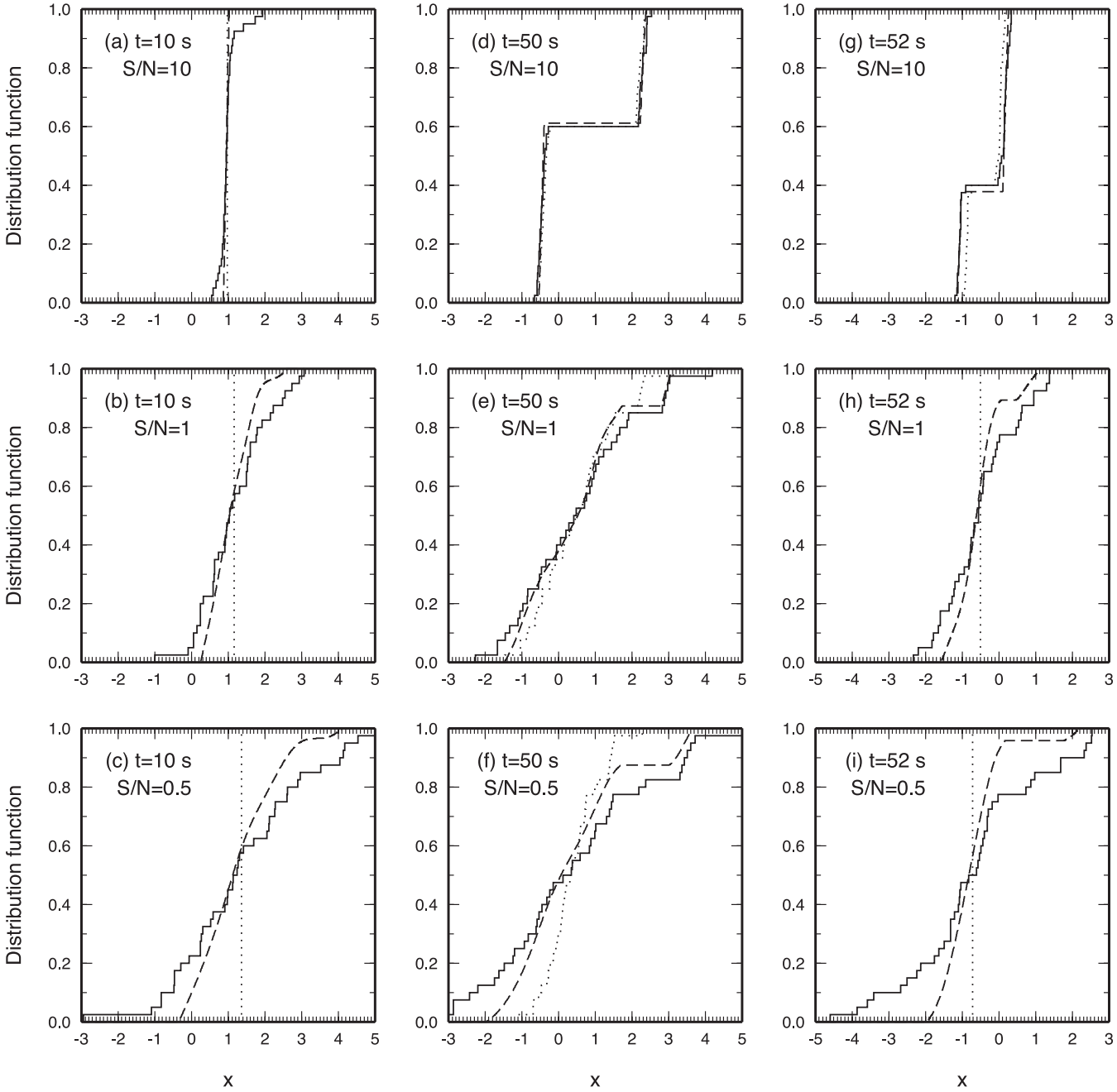


**Figure 7.** Empirical distribution functions (solid) for the synthetic data of Fig. 1, at $t = 10$ s (left-hand column), 50 s (middle column) and 52 s (right-hand column). Also shown are deconvolved distribution functions by simulation extrapolation (dashed) and rescaling (dotted).

### 2.3 Dual bootstrap stack

The implementation of DBS is now laid out in full. The weighting factor $w_1$ in eq. (8) is the same as in eq. (12), and $w_2$ is equal to $\max(0, 1 - p_2/\alpha)$, where $p_2$ is calculated by eq. (14) but with the rescaled EDF of eq. (17). Unless noted otherwise, DBS in this paper is done with $\alpha$ of 0.01, $T$ of 20 s and $B$ of 2000. Its application to the synthetic data is already shown in Fig. 1, and the details of weighting can be seen in Fig. 8. Because the product $w_1 w_2$ is what matters, the second weight $w_2$ is calculated only when $w_1$ is non-zero; otherwise it is set to unity. It can be seen that the incoherent data around $t$ of 50 s pass the significance test but fail the coherence test.

It is also seen that an incoherent data set can sometimes pass the coherence test when its amplitude is small (Fig. 8a, $t = 52.5$ s); this is because the rescaling of EDF becomes sensitive to the subtle balance between $\sigma_X^2$ and $\sigma_N^2$ (Fig. 7, right-hand side column). Also, the weight $w_1$ becomes zero where a signal itself approaches zero, leading to the slight distortion of a weighted waveform even when S/N is high. These minor deficiencies may be mended by taking into account the temporal continuity of a detected signal when computing weights. For example, where the product of the weights $w_1 w_2$ is mostly unity but is disrupted with occasional zero values (e.g. $t$ of 7–13 s in Fig. 8a), those zero weights may be lifted to unity if corresponding linear stacks are of small amplitude. Alternatively, one may want to apply an averaging filter to

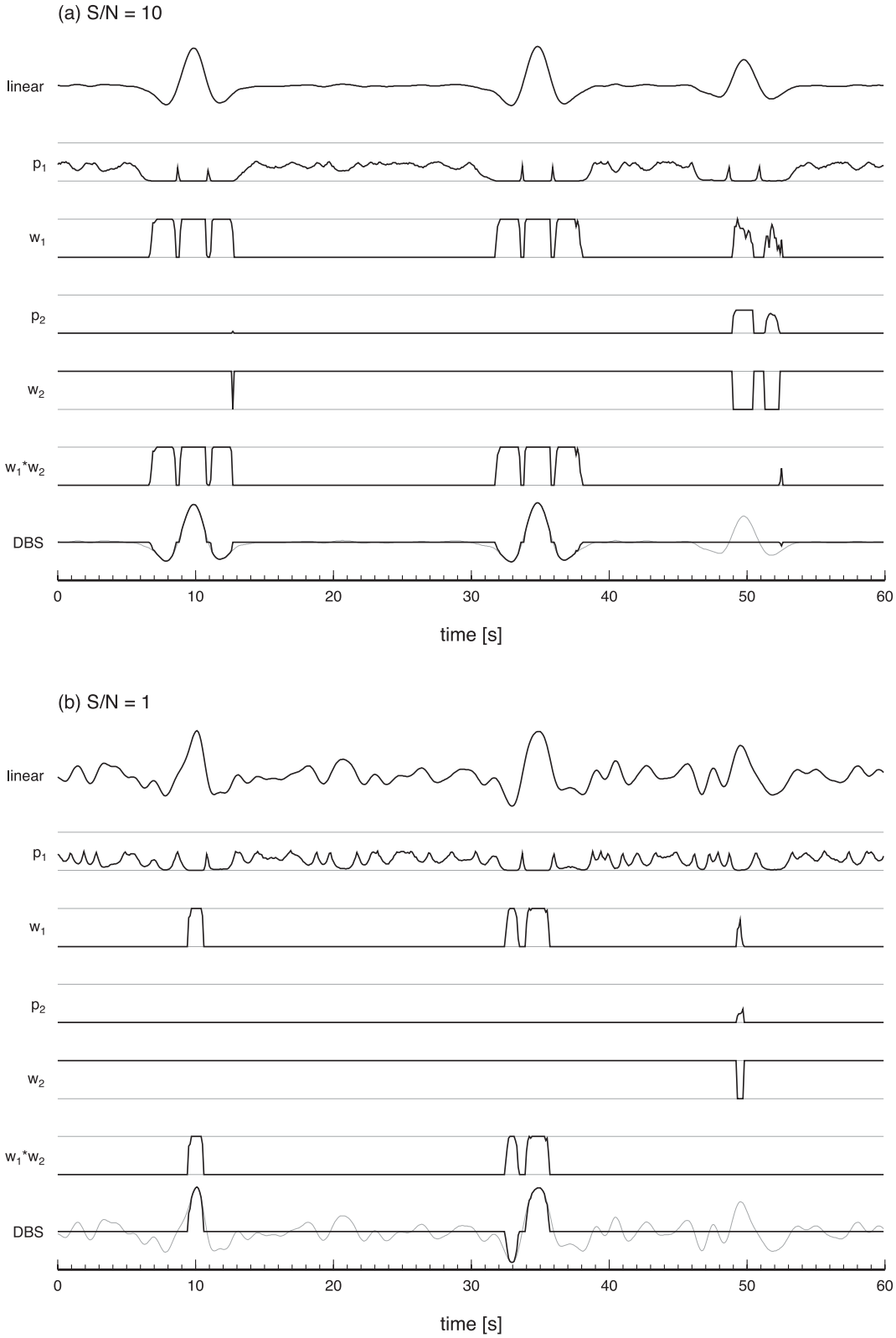


**Figure 8.** Step-by-step decomposition of dual bootstrap stack, using the synthetic data of Fig. 1. As in Fig. 4, thin horizontal lines for probabilities and weights denote the interval [0,1]. DBS is shown with linear stack in grey.

smooth out a rapidly changing weight. Comparing different possibilities for better waveform preservation is not pursued here because the success of a particular implementation is likely to be case-dependent.

Even in the present form, however, DBS suffices the purpose of signal detection. DBS makes a clean separation between signal and noise by keeping the entire amplitude of a stacked sum if it passes two statistical tests and by setting it to zero otherwise. Though the

synthetic data used in this paper assume Gaussian noise, estimating the statistics of noise by trace scrambling should be applicable to non-Gaussian noise as well. An important notion is that, by randomly time-shifting traces, it is easy to generate a physically meaningless stack, which may be treated as a noise stack. The bootstrap resampling itself is quite general, not restricted to Gaussian-based statistics (e.g. Efron & Tibshirani 1993; Davison & Hinkley 1997). Also in DBS, the statistics of noise is estimated at every time instance by locally scrambling traces, so the non-stationarity of noise, if present, can be automatically handled.

## 3 ESTIMATE ON SIGNAL RECOVERY RATE AND RESIDUAL NOISE LEVEL

The synthetic data used so far are of a fixed size ($n = 40$), and only four cases are considered with one particular noise realization. In this section, a more comprehensive test of DBS is conducted by generating a range of synthetic data in a systematic manner. Two kinds of tests, one on signal recovery and the other on noise reduction, are considered.

For the signal recovery test, one Ricker wavelet with the peak frequency of 0.2 Hz is embedded in a 30-s-long record, and as done for the previous synthetic data, the Gaussian noise that is band-limited in the range of 0.1–0.5 Hz is added. In this way, the power spectra of signal and noise are completely overlapped, so the synthetic data may be regarded to have been already filtered and be devoid of easily removable noise components characterized with frequencies outside the bandwidth of interest. Though only one peak frequency is considered here, results to follow can be translated to other cases with different peak frequencies simply by rescaling the time axis. The synthetic data are thus of reasonably general nature. For each of the permutations of the following three parameters, 10 different synthetic data sets are created by using different seeds for the generation of pseudo-random numbers: (1) 11 different S/N values ranging from 10 to 0.1, (2) four different intrinsic signal variabilities (1, 10, 20 and 40 per cent), and (3) five different sample sizes (20, 40, 80, 160 and 320). Thus, the total of 2200 different synthetic cases are prepared.

For each of these synthetic data, NRS, PWS and DBS are performed, and the signal recovery rate is measured as

$$R_S = \frac{\max(\overline{X}_{NLS})}{\max(Y)}, \tag{19}$$

where NLS is either NRS, PWS or DBS. The use of the maximum of a waveform here, instead of the power of the entire waveform, reflects that waveform distortion is severe for low S/N (e.g. Fig. 1d).

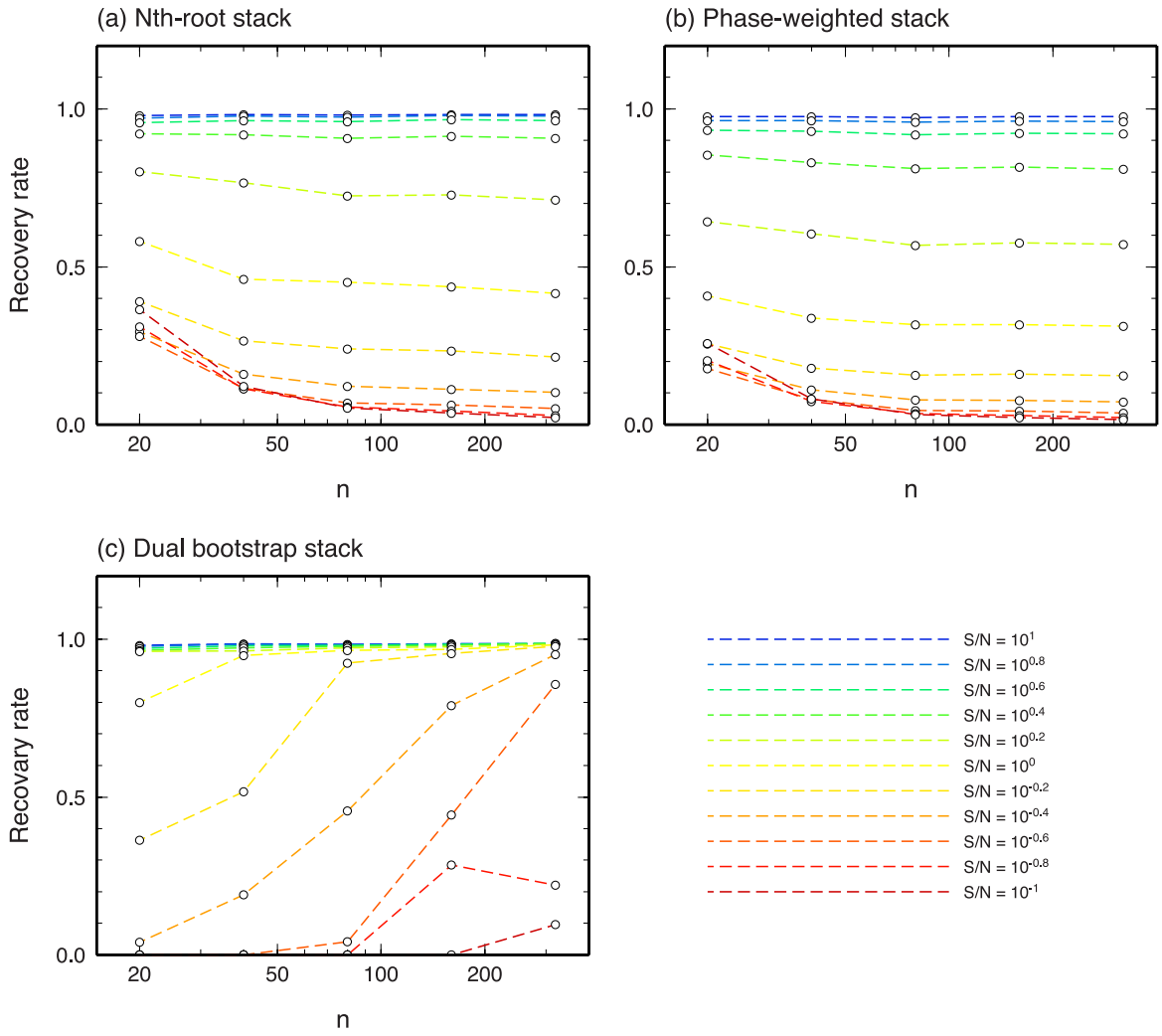


**Figure 9.** Average signal recovery rate (eq. 19) as a function of the number of traces $n$, for (a) NRS, (b) PWS and (c) DBS. Only the case of $\sigma_S = 0.01$ is shown here. Different line colours correspond to different signal-to-noise ratios.

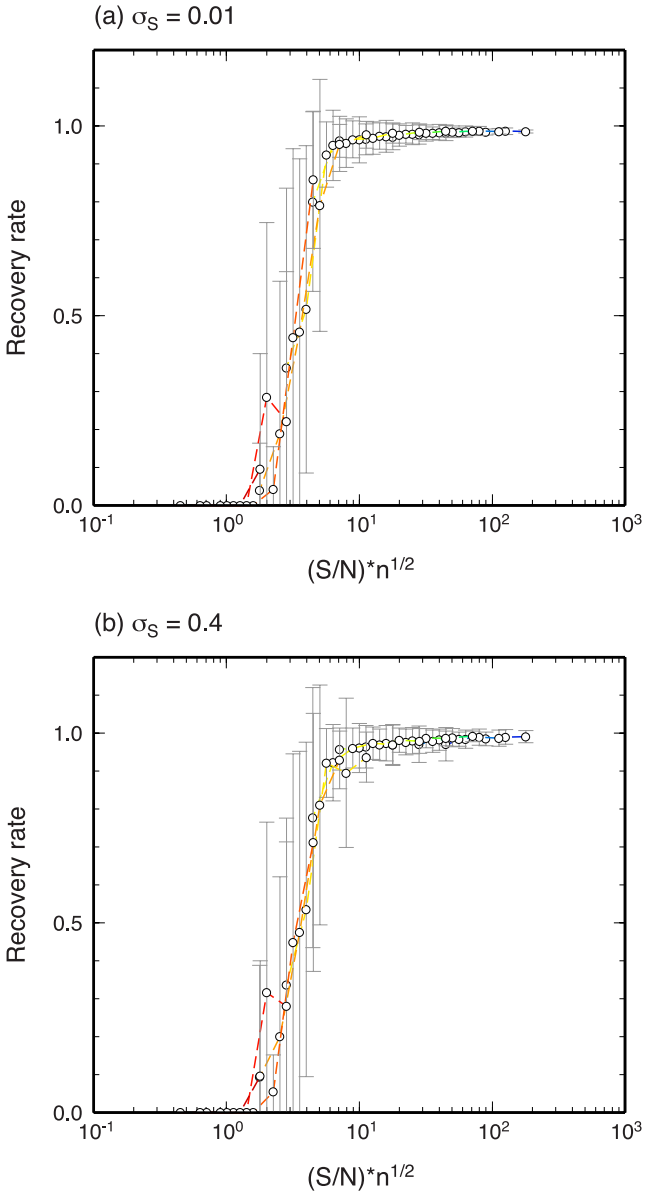


**Figure 10.** Data collapse for the signal recovery rate of DBS for the cases of (a) $\sigma_S = 0.01$ and (b) $\sigma_S = 0.4$. As in Fig. 9, different line colours correspond to different signal-to-noise ratios. Error bars denote one standard deviation.

The recovery rate averaged over 10 random ensembles is shown in Fig. 9 for the case of 1 per cent signal variability. The recovery rates of NRS and PWS behave in a similar manner; both of them fall below 0.5 when S/N is lower than unity, and this systematics does not improve as the sample size increases. This is expected because, when S/N is low, both NRS and PWS are dominated by noise, and having a larger sample does not rectify the situation. The performance of NRS and PWS actually slightly deteriorates with increasing $n$, because the stack of a noisy data set is recognized more conclusively as noise. In contrast, the recovery rate of DBS improves with increasing $n$ (Fig. 9c), and compared at the same $n$, the recovery rate of DBS is almost always higher than those of NRS and PWS.

The dependence of DBS recovery rate on the sample size and S/N may be seen more clearly by performing data collapse, and as shown in Fig. 10, all test results are found to lie approximately on a single trend when plotted as a function of $(S/N)n^{1/2}$. Nearly perfect signal recovery is guaranteed when $(S/N)n^{1/2} > 5$, and signal detection is possible when $(S/N)n^{1/2}$ is greater than ∼2. This trend does not change much even when the intrinsic signal variability $\sigma_S$ is as high as 40 per cent (Fig. 10b). The values of $(S/N)n^{1/2}$ for the four cases shown in Fig. 1 are, in order of decreasing S/N, ∼63, ∼13, ∼6 and ∼3. So the successful signal recovery by DBS in Fig. 1(d) is not unexpected but not something always guaranteed.

The signal recovery test is done with a single wavelet, but based on this, it is also possible to predict the recovery of relative amplitudes between different wavelets because DBS is a purely time-domain method. The recovery of an individual wavelet depends simply on $(S/N)n^{1/2}$, where S/N is defined with the given wavelet. If all wavelets under consideration have $(S/N)n^{1/2}$ greater than ∼5, therefore, their relative amplitudes would be retained nearly perfectly. This may be confirmed with the examples given in Fig. 1; a single Ricker wavelet can be considered as one positive peak paired with two negative peaks of about half an amplitude. The values of $(S/N)n^{1/2}$ for the negative peaks in the four cases shown in Fig. 1 are ∼28, ∼5.6, ∼2.8 and ∼1.4. So the waveform of the Ricker wavelet is thus severely deformed in the last two cases.

Note that the signal recovery rate of linear stack is always unity (or close to unity), but this does not mean that linear stack is better. What is tested with these non-linear stacks is their ability to distinguish between signal and noise. To detect signals unambiguously in linear stack, residual noise should be sufficiently small; for example, the $2\sigma$-amplitude of stacked noise may need to be less than 10 per cent of signal amplitude (i.e. $2\sigma_N/\sqrt{n} < 0.1S$). This requirement is equivalent to $(S/N)n^{1/2}$ being greater than 20, which is far more strict than that for DBS. In terms of the minimum number of traces needed for signal detection, DBS requires 4–25 times $(S/N)^{-2}$ whereas linear stack requires 400 times $(S/N)^{-2}$. It may be said that DBS is 20–100 times more effective than linear stack.

The above signal recovery test would not be fully meaningful without testing for the likelihood of a false alarm, that is, checking the reliability of detected signals. In the second test, therefore, pure noise data are stacked by NRS, PWS and DBS, and the root-mean-square amplitude of stacking results is compared with that of linear stack. The length of each trace is 60 s in this test, and 100 different random ensembles with the band-limited Gaussian noise are created for 5 different sample sizes (20, 40, 80, 160 and 320). The total of 500 synthetic noise cases are thus prepared, and the residual noise level is measured as

$$R_N = \left[ \frac{\int \overline{X}_{NLS}(t)^2 dt}{\int \overline{X}(t)^2 dt} \right]^{1/2}. \quad (20)$$

Results are summarized in Fig. 11, which also shows the result for the standard bootstrap approach mentioned at the beginning of Section 2. The residual noise level of DBS is found to be only ∼1 per cent on average, regardless of the sample size; NRS and PWS can achieve similarly low noise levels only when the sample size is sufficiently large ($n > 100$). As mentioned earlier, the simple use of bootstrap confidence intervals results in high residual noise levels. The consistently low noise level attained by DBS is owing to the tight significance level ($\alpha = 0.01$) used to reject the null hypotheses on signal significance and coherence; linear stack is simply weighted down to zero if it does not pass both of these hypothesis tests.

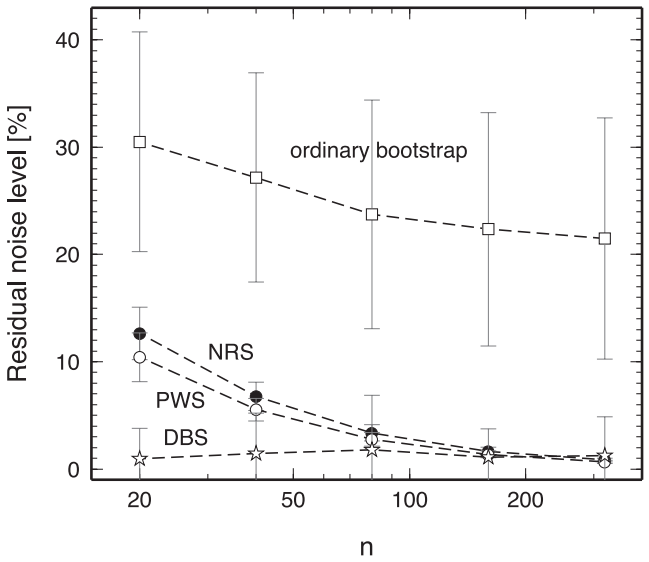


**Figure 11.** Residual noise level (eq. 20) as a function of the number of traces $n$, for NRS (solid circle), PWS (open circle), DBS (star) and the bootstrap approach of eq. (7) with $\alpha = 0.01$ (square). Error bars denote one standard deviation.

## 4 AN EXAMPLE WITH REAL DATA

How DBS performs with real seismic data is briefly discussed here, by applying it to the computation of a slant stack (also known as vespagram). The source is the event of 1991 December 17 at 0638 UT (latitude 47.39°, longitude 151.50°, depth 157 km and magnitude $m_b$ of 5.8), and the receivers are the German Regional Seismic Network and the Gräfenberg array. This source–receiver pair is a familiar one in the study of the lowermost mantle (e.g. Thomas *et al.* 2002) and is also used for vespagram examples in a review article on array seismology by Rost & Thomas (2002). It is thus deemed suitable for the comparison of the new stacking scheme with the existing ones.

The data shown in Fig. 12(a) are bandpass-filtered through 0.1 and 0.5 Hz, aligned on the direct $P$ arrival, and normalized to the amplitude of the first maximum of the arrival. The theoretical slowness of the $P$ arrival at the centre of the stations is 5.56 s deg$^{-1}$. The direct stack of the data as shown enhances the $P$ arrival, whereas stacking with different slownesses would enhance other arrivals if present. In this range of epicentral distance, the $PcP$ arrival is expected to be very weak and is indeed barely visible, but the data also show a conspicuous precursor called $PdP$, which is a reflection off the top of the $D''$ layer. The vespagram is computed with linear stack as well as three non-linear stacks, NRS, PWS and DBS (Figs 12b–e). As in the previous sections, NRS is done with the power of 3, PWS with the order of 2 and DBS with $\alpha = 0.01$, $T = 20$ s and $B = 2000$. Using different settings for NRS and PWS would not modify the results considerably; for the cases of NRS with the power of 4 and PWS with the order of 4, see figs 5 and 10 of Rost & Thomas (2002), respectively. The linear vespagram is characterized by poor slowness resolution, which calls the significance of subtle features into question. The NRS and PWS vespagrams have better slowness resolution, and they also suggest that most of local maxima seen in the linear vespagram are artefacts. The DBS vespagram has a considerably sharper appearance than these conventional vespagrams, because the stacked sum is set simply to zero if it does not pass the two kinds of statistical tests. Though it is a minor point, a small negative peak before the $PdP$ arrival is stacked better by DBS than by NRS and PWS, and given that all features in the DBS vespagram are those have passed the two stringent hypothesis tests, even minor ones are significant at least statistically. Resolving the origin of these small-scale signals, however, requires us to explore additional data and also to experiment with other methods such as migration. Indeed, a more interesting example can be created by applying DBS to teleseismic migration, and it will be presented elsewhere.

## 5 DISCUSSION

One obvious drawback of DBS is its computational cost. Compared to linear stack, conventional non-linear stacks such as NRS are more time-consuming only by a factor of 3 or so, but DBS is more costly by three orders of magnitude. We can of course take the advantage of cheap computational resources; the parallelization of DBS is straightforward. Also, if one is willing to calculate bootstrap confidence intervals to begin with, DBS would not appear so computationally expensive. Nevertheless, intensive bootstrap resampling at every single point in the time axis may not be very appealing when applying DBS to vespagram and migration, in which numerous stacks have to be evaluated. As mentioned in Section 2.1, however, the required number of bootstrap replicates $B$ may not have to be as high as $10^3$, and this issue deserves careful consideration in each application of DBS. Some kind of adaptive approach may also be beneficial. For example, DBS may be done first with a small $B$, and if $p_1$ is found to exceed a threshold, DBS may be repeated with a larger $B$. The tests with synthetic data described in Section 3 are repeated with this adaptive approach (using the first $B$ of 100, the second $B$ of 2000 and the critical $p_1$ of 0.1), and it is found that the signal recovery rate is slightly impaired with the critical value of $(\mathrm{S/N})n^{1/2}$ for guaranteed recovery going up from 5 to 6.7. At the same time, the residual nose level turns out to be exactly zero for all of the 500 noise cases, so there seems to be a trade-off between signal recovery and noise reduction. At any rate, this preliminary experiment suggests a promising direction towards more efficient implementations of DBS. If $B$ can be on the order of 100, parallelized DBS on a PC with 20 cores (which has become commodities in recent years) may be only several times slower than conventional stacks. Also, what distinguishes DBS from the conventional stacks is that the statistical significance of stacked signals is already estimated when the stack is completed. Its computational cost may not be so high after all, if one considers the amount of additional computational work to quantify the significance of conventional stacks.

The two kinds of hypothesis tests considered in this paper are motivated by the performance of NRS and PWS, and one could formulate different hypothesis tests depending on the nature of a problem at hand. The notion of using a hypothesis test to design a weight provides a flexible framework to accommodate additional requirements. For example, extending for three-component seismograms (e.g. Kennett 2000) would be straightforward. Semblance or phase weights cannot correctly measure correlation between components when S/N is low, but the approach based on the EDF, which would be 2-D in this case, can still be used if deconvolution by rescaling is employed (Section 2.2).

NRS and PWS are sufficient when S/N is reasonably high, and ignoring the issue of coherence, simple linear stack would be adequate for low S/N data if a large number of traces are available. An intermediate situation, characterized by a limited number of low S/N data, is most troublesome but is also common. This is the

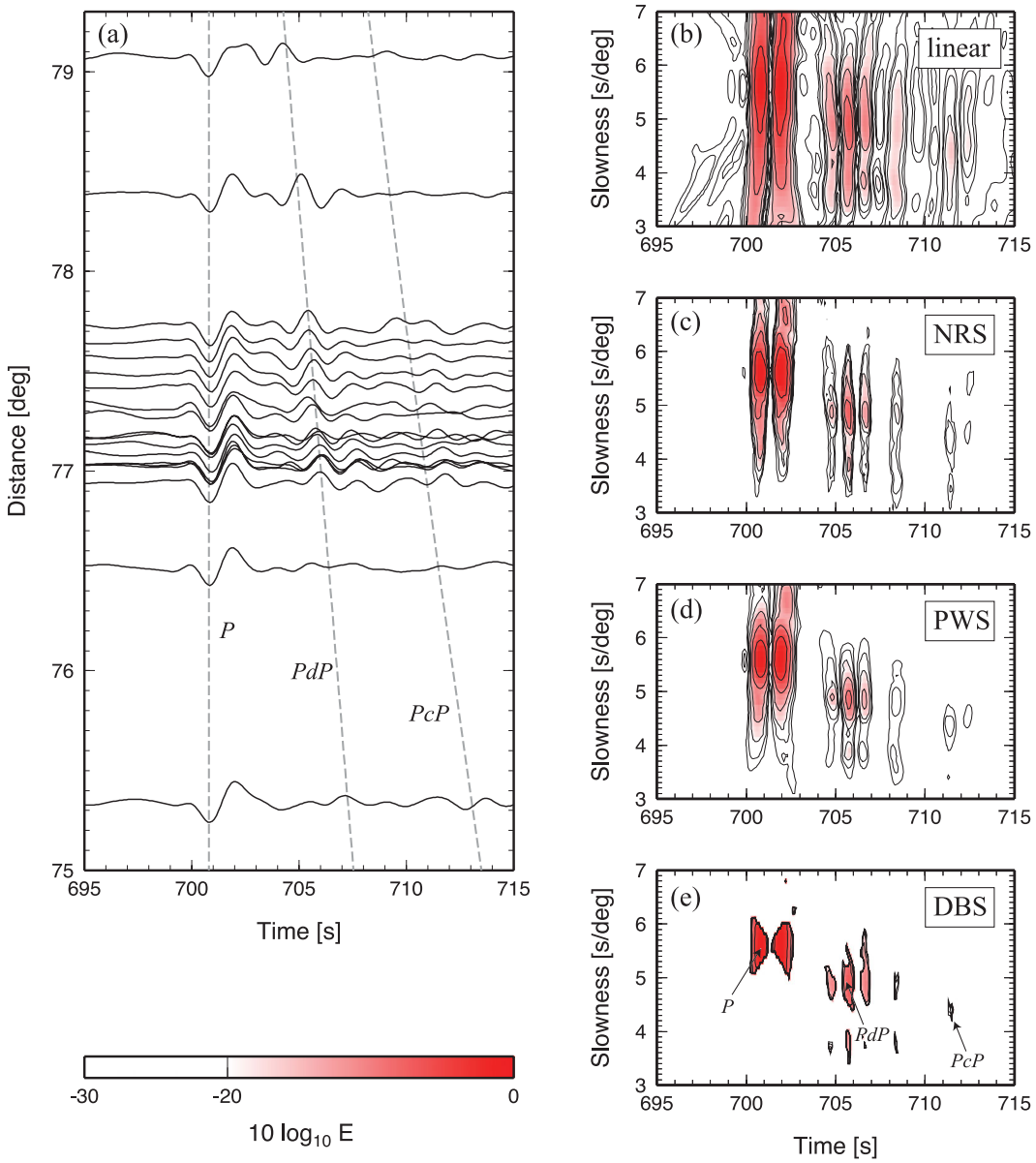


**Figure 12.** (a) Seismogram section for the 1991 December 17 event as described in the text. Records have been aligned on the direct *P* arrival and normalized to the amplitude of the first maximum of the arrival. The theoretical *PcP* arrival [based on IASP91 Kennett & Engdahl (1991)] as well as the empirical *PdP* arrival (based on vespagram reading) are also shown as dashed lines. (b)–(e) Corresponding vespagram with linear stack, NRS, PWS and DBS. Stacked energy is shown in the unit of dB, relative to the maximum energy of unity.

domain where the resolving power of DBS would be best appreciated. Stacking is ubiquitous in active-source as well as passive-source seismology, and the benefit of using DBS is yet to be examined in a variety of applications.

## ACKNOWLEDGEMENTS

This work was sponsored by the U.S. National Science Foundation under grant EAR-0842753. Figures were generated with GMT (Wessel & Smith 1995). Data used in this study were obtained from Das Seismologische Zentralobservatrium (SZGRF). This work was also supported in part by the facilities and staff of the Yale University Faculty of Arts and Sciences High Performance Computing Center. The author thanks Editor Duncan Agnew and two anonymous reviewers for comments and suggestions.

## REFERENCES

Davison, A.C. & Hinkley, D.V., 1997. *Bootstrap Methods and Their Applications,* Cambridge Univ. Press.

Efron, B., 1982. *The Jackknife, the Bootstrap and Other Resampling Plans,* SIAM.

Efron, B. & Tibshirani, R.J., 1993. *An Introduction to the Bootstrap,* Chapman & Hall.

Hutko, A., Lay, T., Revenaugh, J. & Garnero, E.J., 2008. Anticorrelated seismic velocity anomalies from post-perovskite in the lowermost mantle, *Science,* **320,** 1070–1074.

Kanasewich, E.R., Hemmings, C.D. & Alpaslan, T., 1973. *N*th-root stack nonlinear multichannel filter, *Geophysics,* **38,** 327–338.

Kennett, B.L.N., 2000. Stacking three-component seismograms, *Geophys. J. Int.,* **141,** 263–269.

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

Margerin, L. & Nolet, G., 2003. Multiple scattering of high-frequency seismic waves in the deep Earth: *PKP* precursor analysis and inversion for mantle granularity, *J. geophys. Res.,* **108,** 2514, doi:10.1029/2003JB002455.

McFadden, P.L., Drummond, B.J. & Kravis, S., 1986. The *N*th-root stack: theory, applications, and examples, *Geophysics,* **51,** 1879–1892.

Muirhead, K.J., 1968. Eliminating false alarms when detecting seismic events automatically, *Nature,* **217,** 533–534.

Revenaugh, J. & Meyer, R., 1997. Seismic evidence of partial melt within a possibly ubiquitous low velocity layer at the base of the mantle, *Science,* **277,** 670–673.

Rost, S. & Thomas, C., 2002. Array seismology: methods and applications, *Rev. Geophys.,* **40,** 1008, doi:10.1029/2000RG000100.

Schimmel, M. & Paulssen, H., 1997. Noise reduction and detection of weak, coherent signals through phase-weighted stacks, *Geophys. J. Int.,* **130,** 487–505.

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

Sheriff, R.E. & Geldart, L.P., 1995. *Exploration Seismology,* 2nd edn, Cambridge Univ. Press.

Stefanski, L.A. & Bay, J.M., 1996. Simulation extrapolation deconvolution of finite population cumulative distribution function, *Biometrika,* **83,** 407–417.

Thomas, C., Kendall, J.-M. & Weber, M., 2002. The lowermost mantle beneath northern Asia—I. Multi-azimuth studies of a D″ heterogeneity, *Geophys. J. Int.,* **151,** 279–295.

Wessel, P. & Smith, W.H.F., 1995. New version of the generic mapping tools released, *EOS, Trans. Am. geophys. Un.,* **76,** 329.

Yilmaz, O., 1987. *Seismic Data Processing,* Society of Exploration of Geophysicists.