GEOPHYSICS, VOL. 75, NO. 3 (MAY-JUNE 2010); P. V39–V49, 11 FIGS.
10.1190/1.3428749

High-amplitude noise detection by the expectation-maximization algorithm with application to swell-noise attenuation

Maïza Bekara[1] and Mirko van der Baan[2]

ABSTRACT

High-amplitude noise is a common problem in seismic data. Current filtering techniques that target this problem first detect the location of the noise and then remove it by damping or interpolation. Detection is done conventionally by comparing individual data amplitudes in a certain domain to a user-controlled local threshold. In practice, the threshold is optimally tuned by trial and error and is often changed to match the varying noise power across the data set. We have developed an automatic method to compute the appropriate threshold for high-amplitude noise detection and attenuation. The main idea is to exploit differences in statistical properties between noise and signal amplitudes to construct a detection criterion. A model that consists of a mixture of two statistical distributions, representing the signal and the noise, is fitted to the data. Then it is used to estimate the probability (i.e., likelihood) that each sample in the data is noisy by means of an expectation-maximization (EM) algorithm. Only those samples with a likelihood greater than a specific threshold are considered to be noise. The resulting probability threshold is better adapted to the data compared to a conventional amplitude threshold. It offers the user, through the probability threshold value, the possibility to quantify the confidence in whether a large amplitude anomaly is considered as noise. The method is generic; however, our work develops and implements the method for swell-noise attenuation. Initial results are encouraging, showing slightly better performance than an optimized conventional method but with much less parameter testing and variation.

INTRODUCTION

Seismic data are always corrupted with various types of noise that reduce data quality. Noise attenuation is therefore an important step in seismic data processing to aid interpretation. Signal and noise can be separated by means of model-based signal processing.

This approach assumes that signal and noise have different characteristics that can be captured by a specific mathematical model. The model can be based on a law in physics such as wave theory, which is used in many demultiple methods (Verschuur et al., 1992). Alternatively, the model may exploit a deterministic feature such as a different trace moveout (dip) (Freire and Ulrych, 1988; Bekara and van der Baan, 2007) or a statistical property such as independence to differentiate signal and noise (van der Baan, 2006).

We are concerned with the problem of high-amplitude noise attenuation. Consider a simple noise model where all data samples larger than a certain threshold are likely to be noise. The threshold can be specified in many forms but is often formulated as a threshold factor times some data statistic such as the mean, the median, or the rms value computed within the analysis window. The threshold factor and the data statistic are user-defined parameters. This model is invoked because of its simplicity (Anderson and McMechan, 1989; Cambois and Frelet, 1995; Elboth et al., 2008). However, its main disadvantage is the frequent change of the threshold value needed to match the variation of noise power across the data.

We propose an automatic threshold-determination technique for detecting large-amplitude noise in a domain of choice. Attenuation of swell noise in the frequency-offset (f-x) domain is developed as an example, but the result can be generalized to other types of noise such as diffracted multiple noise and spikes (Liu et al., 2009), where large-amplitude samples in a time or time-offset (t-x) window are detected and then removed by interpolation. Our technique is tested on real marine data; it shows better data adaptability than the conventional method.

First, we briefly review the problem of swell noise. Then, we outline our new method for automatic threshold determination. Finally, we show some real data examples.

Manuscript received by the Editor 30 July 2009; revised manuscript received 24 November 2009; published online 26 May 2010.
[1]Petroleum Geo-Services, Weybridge, United Kingdom. E-mail: maiza.bekara@pgs.com.
[2]University of Alberta, Department of Physics, Edmonton, Alberta, Canada. E-mail: mirko.vanderbaan@ualberta.ca.

SWELL-NOISE ATTENUATION

Swell noise is caused by rough weather conditions and is a frequent problem in acquiring marine seismic data. It has an adverse effect on seismic data quality and may even lead to temporary suspension of acquisition. It is characterized by large-amplitude and predominantly low-frequency content, as shown in Figure 1.

Elboth et al. (2009) discuss several possible mechanisms for swell-noise generation but conclude that for modern, foam-filled streamers, the most likely causes are (1) hydrostatic-pressure fluctuations resulting from vertical motion of the ocean because of strong sea waves or (2) dynamic pressure variations along the surface of the streamer that result from the presence of a turbulent layer surrounding the streamer.

Conventional techniques to attenuate swell noise first compute the f-x amplitude or power spectrum of the data within a sliding window. All spectral values at a given frequency within the window that exceed some threshold are considered noise. These noisy samples are then attenuated (Elboth et al., 2008) or interpolated (Soubaras, 1995; Schonewille et al., 2008). A threshold value must, however, be defined in this process.

For a given frequency f, we define $S_n = \{r_1, r_2, \ldots, r_n\}$ with $r_k = |D_k(f)|^2$, where $D_k(f)$ is the Fourier transform of the kth trace in the data window. Thus, S_n represents power-spectrum values of a set of traces at a given frequency. The threshold value is usually computed as

$$\text{threshold} = \alpha \hat{s}(S_n), \tag{1}$$

where $\hat{s}(S_n)$ is a statistical measure computed from the samples in S_n and α is a positive factor, referred to as the threshold factor. The threshold value represents an upper bound on the possible values of the signal amplitudes. Therefore, any amplitude above it is considered to be genuine noise.

The statistical measure $\hat{s}(S_n)$ can be the median, the mean, the rms value, or any percentile of the set S_n. The selection of an appropriate statistical measure depends on the distribution of the samples in S_n.

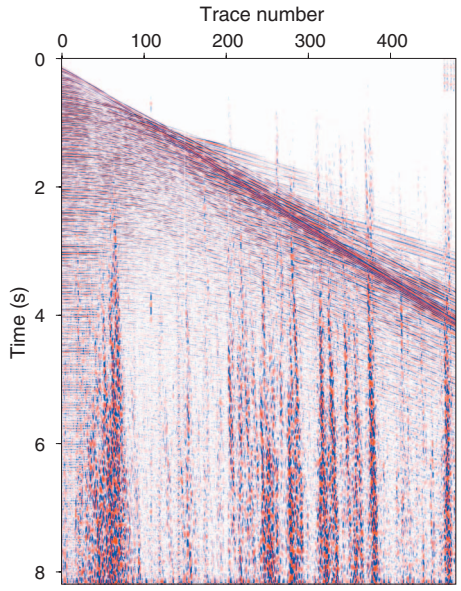

Figure 1. A marine seismic gather, showing clear swell-noise contamination (vertical stripes). Data courtesy of Fugro.

The mean and the rms value can be biased by large noise amplitudes, so the median is often used for its robustness to extreme values. The parameter α controls the strength of the threshold level and is adjusted by the processing specialist to obtain the best results.

Despite the fact that the threshold value in equation 1 is adapted to the data through the statistical measure $\hat{s}(S_n)$, practical experience shows that α is frequently updated if the noise power varies considerably across the entire data set.

THEORY

This section describes an automatic method to determine the appropriate threshold by considering the problem of outliers detection.

Outliers detection

Consider a data set $S_n = \{r_1, r_2, \ldots, r_n\}$, where the samples r_k are assumed independent and identically distributed (iid), generated from a probability density function (PDF) $g(r)$. Finding abnormal samples in set S_n is known in applied statistics as *outliers detection* (Rousseeuw and Leroy, 1987). The objective is to find those samples that show different statistical properties from the rest of the data (known as regulars). In our problem, the outliers (noise + signal) have larger amplitudes than the regular data (signal only). Therefore, the population of outliers differs from that of regular data in terms of some distinguishing statistical measures that can be captured by using a mixture of PDF modeling.

For notation, the PDF of the outliers is denoted by $p(r|\theta_1)$ and the PDF of the regular data is denoted by $p(r|\theta_0)$. It is assumed for simplicity that both distributions belong to the same family of parametric PDFs but have different parameter values, i.e., $\theta_1 \neq \theta_0$. Let the scalar ϵ represent the fraction of outliers in set S_n. This parameter has the statistical meaning of being the a priori probability (i.e., predata modeling) that a sample r_i drawn randomly from S_n is an outlier. The data PDF is now structured as a mixture of two models:

$$g(r|\theta_0, \theta_1, \epsilon) = \epsilon p(r|\theta_1) + (1 - \epsilon) p(r|\theta_0). \tag{2}$$

One can estimate from S_n all parameters of the model defined in equation 2. Once the estimate is completed, the a posteriori probability (i.e., postdata modeling) that a given sample $r = r_i$ is an outlier can be computed using Bayes' rule. In Bayes' rule, A and B are two events and $P(A|B)$ is the probability of event A happening, given that event B has happened. It expresses the conditional likelihood of the two events, such as $P(A|B)P(B) = P(B|A)P(A)$ (Papoulis and Pillai, 2002). Here,

$$\Pr\{r \text{ is outlier} | r = r_i\} = \frac{\epsilon p(r_i|\theta_1)}{\epsilon p(r_i|\theta_1) + (1 - \epsilon) p(r_i|\theta_0)}. \tag{3}$$

The datum r_i is considered to be an outlier if its a posteriori probability is greater than a probability threshold value β, i.e.,

$$\Pr\{r \text{ is outlier} | r = r_i\} > \beta \text{ with } \frac{1}{2} \leq \beta < 1. \tag{4}$$

As an example, for $\beta = 0.5$, expression 4 means that an outlier is selected only if the model is at least 50% sure that it is an outlier. The choice of a probability threshold β in equation 4 is much more objective than an arbitrary threshold factor α in equation 1 and reflects

the statistical confidence we may require to classify any datum as an outlier.

Proposed threshold computation

When computing a threshold, three steps are involved: choosing a PDF, identifying the algorithm, and interpreting the results.

Choice of PDF

A parametric form for $p(r|\theta)$ in equation 2 needs to be chosen in order to use the detection criterion given by equation 3. The form of $p(r|\theta)$ should provide different decay rates when r increases. This is required to guarantee that the likelihood of large-amplitude values is greater in the population of outliers compared to that in the population of regular data. Using the approximation that the real and imaginary parts of a Fourier transform are asymptotically independent zero-mean Gaussian random variables with equal variance (Kay, 1988), one can show that the PDF of the power spectrum samples has an exponential form (Appendix A).

The exponential distribution has one parameter, $\theta = \lambda$, which is the mean of the distribution (Papoulis and Pillai 2002, their section 5.5):

$$p(r|\lambda) = \lambda^{-1}\exp\left(\frac{-r}{\lambda}\right), \quad \text{with } r \geq 0. \tag{5}$$

The model in equation 2 then becomes

$$g(r|\lambda_0,\lambda_1,\epsilon) = \frac{1-\epsilon}{\lambda_0}\exp\left(\frac{-r}{\lambda_0}\right) + \frac{\epsilon}{\lambda_1}\exp\left(\frac{-r}{\lambda_1}\right), \tag{6}$$

with $\lambda_1 > \lambda_0$ to reflect the fact that large amplitudes are more likely to be generated by the distribution of outliers.

At this point, one might question the goodness of fit of the model in equation 6 before using it to draw inferences. The answer to this question is nontrivial because it requires a statistical testing of the null hypothesis: data follow the model in equation 6. Mathematical models help us understand and describe the real world, and it is generally accepted that the true data-generating model is much more complex than any possible hypothetical model. Therefore, the issue of model correctness is not formally addressed, but a model-sensitivity test is discussed later in this paper.

Detection algorithm

The model in equation 6 has three unknown parameters $(\lambda_0,\lambda_1,\epsilon)$, which are to be estimated from S_n. We propose to use the maximum likelihood estimator (MLE), for its desirable statistical properties such as consistency and efficiency, in the following optimization problem:

$$(\hat{\lambda}_0,\hat{\lambda}_1,\hat{\epsilon}) = \arg\max_{\lambda_0,\lambda_1,\epsilon} \prod_{i=1}^{n} g(r_i|\lambda_0,\lambda_1,\epsilon). \tag{7}$$

There is no closed-form solution to this problem, so the parameter values are estimated using an iterative procedure proposed by Hasselblad (1969). This procedure belongs to a class of techniques known as the expectation-maximization (EM) algorithms, which are used to find the MLE of a model's parameter when data are missing or incomplete (Dempster et al., 1977). The iterative EM algorithm is given by the following steps:

1) Set initial values for the parameters $\lambda_0^{(0)}$, $\lambda_1^{(0)}$, $\epsilon^{(0)}$.
2) Update the parameters such that

$$\epsilon^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} A_i^{(t)}, \tag{8}$$

$$\lambda_1^{(t+1)} = \frac{\sum_{i=1}^{n} A_i^{(t)} r_i}{\sum_{i=1}^{n} A_i^{(t)}}, \tag{9}$$

and

$$\lambda_0^{(t+1)} = \frac{\sum_{i=1}^{n} (1-A_i^{(t)}) r_i}{\sum_{i=1}^{n} (1-A_i^{(t)})}, \tag{10}$$

where

$$A_i^{(t)} = \frac{p(r_i|\lambda_1^{(t)})\epsilon^{(t)}}{p(r_i|\lambda_1^{(t)})\epsilon^{(t)} + p(r_i|\lambda_0^{(t)})(1-\epsilon^{(t)})}. \tag{11}$$

Detailed derivation is presented in Appendix B. Convergence of the EM algorithm is quick in this particular application and is, to a large extent, independent of the initial conditions. The initial conditions are set such that $\epsilon^{(0)} = 0.1$, $\lambda_1^{(0)}$ equals the mean of the $[\epsilon^{(0)}n]$ largest amplitudes and $\lambda_0^{(0)}$ equals the mean of the rest of the data.

Combining equations 3, 4, and 6 while replacing the unknown parameters λ_0, λ_1, and ϵ with their estimates ($\hat{\lambda}_0$, $\hat{\lambda}_1$, and $\hat{\epsilon}$) leads to an amplitude-based threshold detection criterion. An amplitude r is an outlier with probability β if $r > r_{\text{exp}}$, where exp stands for exponent in the expression

$$r_{\text{exp}} = \frac{\hat{\lambda}_1}{\frac{\hat{\lambda}_1}{\hat{\lambda}_0} - 1}\left[\log\left(\frac{1-\hat{\epsilon}}{\hat{\epsilon}}\right) + \log\left(\frac{\beta}{1-\beta}\right) + \log\left(\frac{\hat{\lambda}_1}{\hat{\lambda}_0}\right)\right]. \tag{12}$$

Interpretation

The value $A_i^{(t)}$ in equation 11 is the a posteriori probability that the sample r_i is an outlier, computed at each iteration. It is used to weight the contribution of r_i toward the mean of outliers. Likewise, $(1 - A_i^{(t)})$ is the a posteriori probability that the sample r_i is a signal; it is used to weight the contribution of r_i toward the signal's mean. Therefore, one can view the EM algorithm in equations 8–11 as an iterative method to compute the signal mean, where the effect of noisy amplitudes is scaled down before the mean is computed.

The proposed threshold in equation 12 jointly takes into account the statistics of the data (through the estimates of the model's parameters in equation 6) and the user's confidence requirement through the value of β. The user's influence on the threshold is minimal

when $\log(\beta/(1-\beta))=0$, i.e.. $\beta=0.5$. This is consistent with the fact that for any binary decision to be made, a probability of 0.5 for each outcome represents the uninformative case in the sense that the a priori knowledge of this probability would not affect the decision. Therefore, a value of $\beta=0.5$ would be a preferable default for the proposed probability threshold.

Analysis of the new threshold criterion defined in equation 12 reveals several points. First, when the user increases the value of β, the statistical confidence required to accept outliers increases, consequently increasing the threshold level. Therefore, β has a similar role to the threshold factor α for the conventional method. Second, under the statistical model, small values of $\hat{\epsilon}$ imply the existence of few outliers, and the threshold level increases to reflect that. On the other hand, when $\hat{\epsilon}\rightarrow 1$, the model infers that most of the data are outliers and the threshold level decreases as a consequence. Finally, when $\hat{\lambda}_0\approx\hat{\lambda}_1$, the assumption of a mixture of two models in equation 6 is not empirically supported by the data. Only a single population exists; therefore, the value of $\hat{\epsilon}$ has no statistical meaning. In this situation, we assume that the data are composed of regular samples only because the case of outlier-only samples is excluded. The threshold is automatically set to a large value, and therefore no outlier is selected.

To study the sensitivity of the proposed technique to the chosen statistical model, we consider also using a Rayleigh distribution instead of the exponential one in the data modeling. Derivation of the detection criterion is presented in Appendix C.

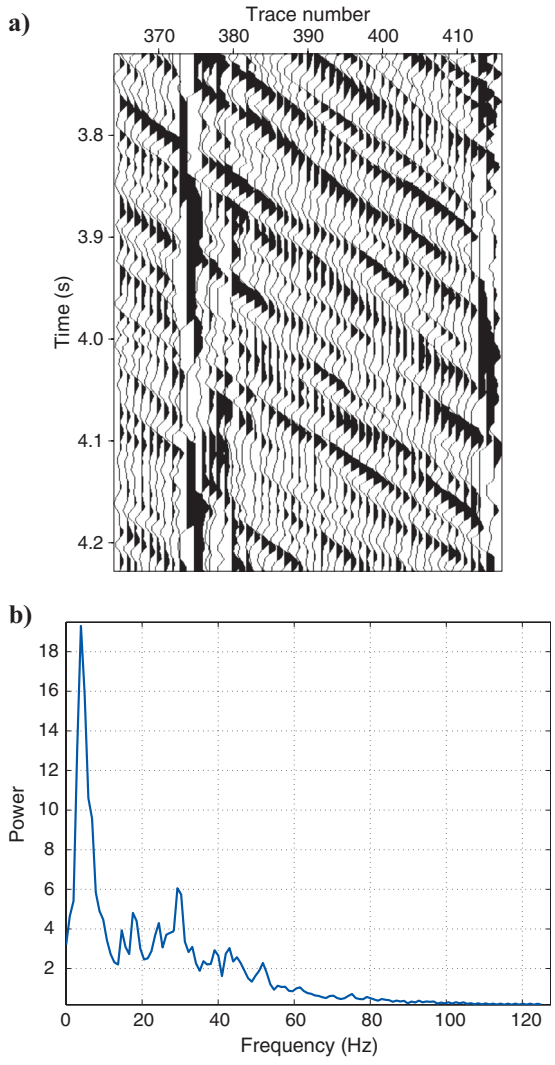

Figure 2. A small data window extracted from Figure 1. (a) Refracted arrivals mixed with swell noise. (b) Its average power spectrum shows that swell noise is concentrated in the frequency band between 2 and 15 Hz.

Outlier attenuation

Once the noisy amplitudes are identified by means of the threshold criterion in equation 12 (for the exponential model) or in equation C-2 (for the Rayleigh model), they can be removed from S_n (e.g., by setting $r_i=0$) or interpolated from neighboring samples (Soubaras, 1995). However, we choose a more conservative option by rescaling the noisy samples with a constant factor such that (1) the new mean of the noisy samples is equal to the mean of the regular data for the exponential model and (2) the new rms value of the noisy samples is equal to the rms value of the regular data for the Rayleigh model.

DATA EXAMPLE

This section contains a series of test examples using our proposed threshold with the exponential model. At the end, a comparative test between the exponential and the Rayleigh model is performed to investigate the sensitivity of the statistical approach to the choice of PDF used in the modeling.

Threshold computation

Consider the data window displayed in Figure 2a, which is extracted from the marine shot gather in Figure 1. This window consists of 50 traces that extend over a 512-ms time gate (128 time samples). Swell noise is visible from traces 370–380 and from trace 410 to the end of the window. The swell noise is localized in the frequency from 2 to 15 Hz with a peak near 4 Hz (Figure 2b). This data window is a good example of a typical noise-detection scenario because good-quality signal (albeit refractions and near-surface reverberations) is mixed in high-amplitude swell noise.

Our proposed detection criterion is applied to two spatial sequences (i.e., constant-frequency slices) obtained from the data in Figure 2a after transformation to the f-x domain. The first sequence is taken at frequency $f=4$ Hz (strong swell noise); the second sequence is taken at $f=20$ Hz (no swell noise). The case of no swell noise is considered to investigate whether our method preserves signal when no noise is present.

Figure 3a shows the power spectral values of each trace at $f=4$Hz and the estimated threshold level $r_{\exp}$ given by expression 12. Figure 3b displays the corresponding noise probability (expression 3) with probability threshold $\beta=0.5$. Large probabilities (>0.95) occur at trace numbers where swell noise has been identified visually (Figure 2a). Note the improved detection of swell noise at traces 377–380; their amplitudes are relatively low, but the computed noise probability is large.

The histogram of the data in Figure 3a is plotted with the estimated mixture of exponential densities (equation 6) and is shown in Figure 4a. The fitting is not globally perfect but is reasonably good for the low amplitudes (mainly signal). The EM algorithm (equations

8–11) converges quickly, as shown in Figure 4b and c; one can conclude the existence of two populations in the data ($\hat{\lambda}_1 \neq \hat{\lambda}_0$).

Likewise, Figure 5a shows the spectral power values of each trace at $f = 20$ Hz, along with the estimated threshold value. A logarithmic scale is used for the spectral power axis to allow plotting the large threshold along with the observed spectral power values. Unsurprisingly, the computed probabilities of the presence of swell noise are low, approximately 0.08 (Figure 5b). This is consistent with the fact that at $f = 20$ Hz, the swell noise is weak to absent (Figure 2a). The use of a more conventional thresholding scheme such as in equation 1 bears the risk that some of the largest power-spectrum values at $f = 20$ Hz could be mistakenly identified as noise if α is set too low.

To see why our method preserves the spectral amplitudes at $f = 20$ Hz, we examine the values of the model parameters ($\hat{\lambda}_0$, $\hat{\lambda}_1$, $\hat{\epsilon}$) throughout the iterations, as computed by the EM algorithm and shown in Figure 6b and c. The relative portion of outliers $\hat{\epsilon}$ quickly decreases to 0.08 (Figure 6b), and the respective mean values for the regulars $\hat{\lambda}_0$ and outliers $\hat{\lambda}_1$ converge to the same numerical values (Figure 6c). The statistical model concludes that all data at this fre-

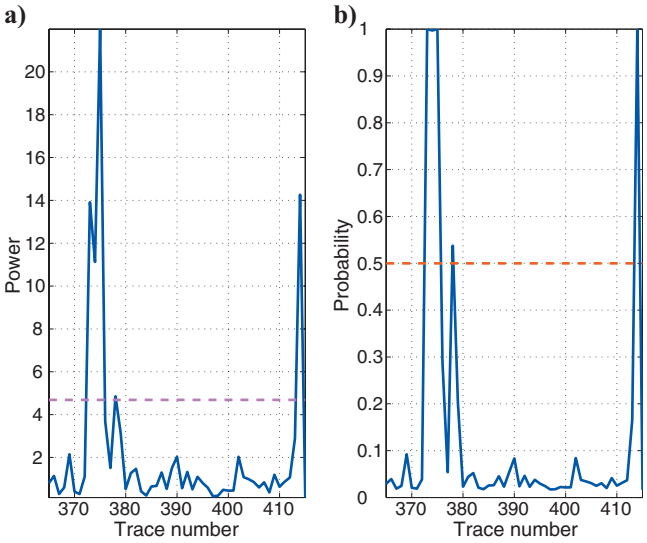

Figure 3. Swell noise detected at $f = 4$ Hz for the data window in Figure 2. (a) Power spectrum values (solid) and the proposed threshold (dashed). (b) Noise probability (solid) and the probability threshold β used in the detection (dashed). Notice the improved detection of swell noise at traces 377–380; their amplitudes are relatively low, but the computed noise probability is large.

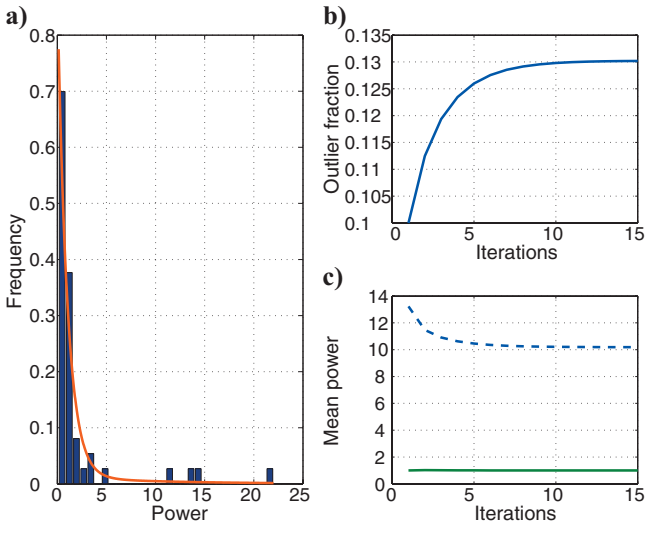

Figure 4. (a) Histogram of the data in Figure 3a, plotted with the estimated mixture of exponential distributions. Evaluation of the estimated parameters through the iterations of the EM algorithm: (b) fraction of outlier $\hat{\epsilon}$; (c) mean power $\hat{\lambda}_1$ (dashed) and $\hat{\lambda}_0$ (solid). One can conclude the existence of two populations in the data ($\hat{\lambda}_1 \neq \hat{\lambda}_0$).

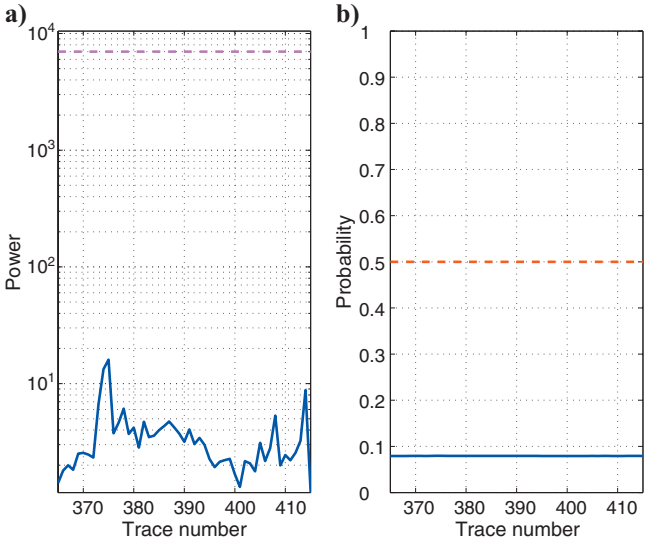

Figure 5. Swell noise detected at $f = 20$ Hz for the data window in Figure 2. (a) Power-spectrum values (solid) and the estimated threshold value (dashed). (b) Noise probability (solid) and the probability threshold β used in detection (dashed). Note a logarithmic scale is used in (a). The statistical model in equation 6 gives no empirical evidence supporting the presence of swell noise at this frequency. All computed probabilities are well below the probability threshold $\beta = 0.5$.

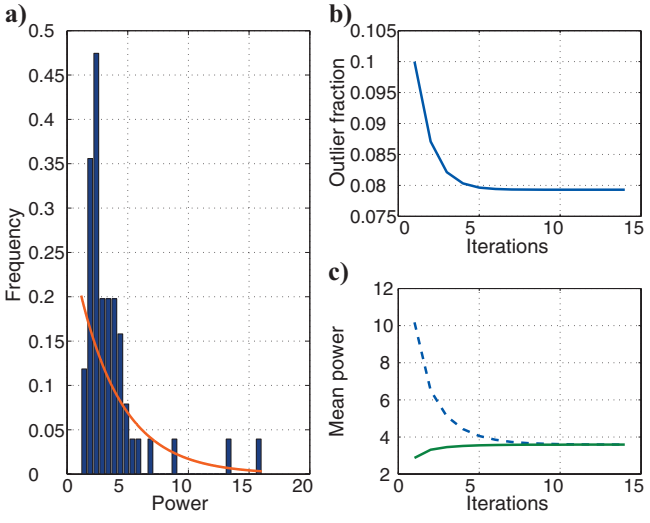

Figure 6. (a) Histogram of the data in Figure 5a plotted with the estimated mixture of exponential distributions. Evaluation of the estimated parameters through the iterations EM algorithm: (b) fraction of outlier ϵ; (c) mean power $\hat{\lambda}_1$ (dashed) and $\hat{\lambda}_0$ (solid). The relative portion of outliers $\hat{\epsilon}$ converges to 0.08, and the population means $\hat{\lambda}_0$ and $\hat{\lambda}_1$ converge to identical values, indicating that the data likely consist only of a single population without noise contamination.

quency result from a single distribution (signal only) instead of a double population (signal and noise). As a result, no samples are classed as swell noise. The estimated mixture of the exponential distribution model does not fit the data's histogram (Figure 6a), a clear indication of model misspecification. We argue, based on other tests, that when a misfit is significant, the estimated model parameters are such that $\hat{\lambda}_0 \approx \hat{\lambda}_1$. This means no data will be classified as noisy, making the technique safer.

Swell-noise attenuation in individual windows

The threshold criterion in equation 12 is implemented in combination with the outlier rescaling methodology, mentioned earlier, to attenuate swell noise in the shot domain. Tests are done on two different data windows (shallow and deep) extracted from Figure 1. For comparison, we use the conventional detection criterion in equation 1, with the median as the statistical measure. Detected noisy samples are then scaled down to the threshold level $\alpha\hat{s}(S_n)$. Two different values for α are used to test, respectively, a mild and a harsh noise filter. Our approach is implemented with a fixed value of $\beta = 0.5$.

The first data window is shown in Figure 2a. This shallow, far-offset window includes refracted arrivals and swell noise. A mild noise filter ($\alpha = 3$) preserves the signal, but it also attenuates less swell noise (Figure 7a and b). On the other hand, a harsh noise filter ($\alpha = 1$) distorts the refracted arrivals but attenuates much of the swell noise (Figure 7c and d). The proposed method (Figure 7e and f) achieves results similar to the conventional mild filter.

The second data window is extracted from a deep, near-offset section, as shown in Figure 8 along with its average power spectrum. An aggressive noise filter with $\alpha = 1$ (Figure 9c and d) provides better

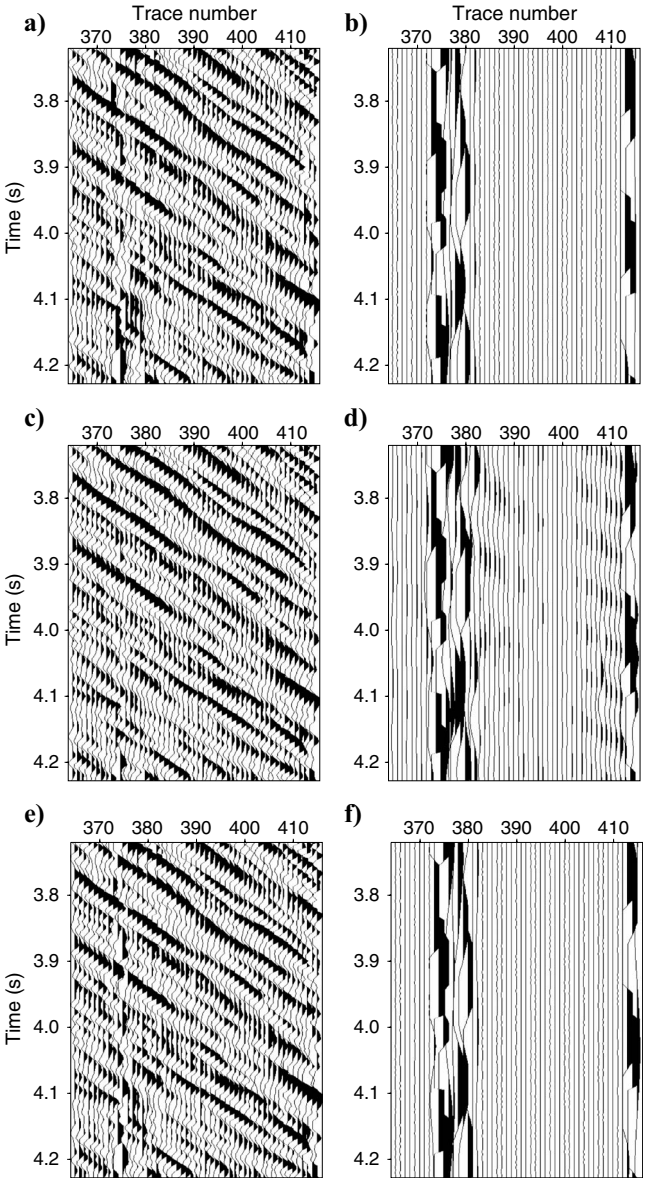

Figure 7. Swell-noise filtering applied to the data in Figure 2. (Left column) Result after swell-noise attenuation. (Right column) Difference section. (a, b) Conventional method with $\alpha = 3$ (mild filter); (c, d) conventional method with $\alpha = 1$ (harsh filter); and (e, f) proposed method with $\beta = 0.5$. The mild filter preserves the refracted arrivals better but attenuates swell noise less well than the harsh filter. The proposed method is similar to the mild filter.

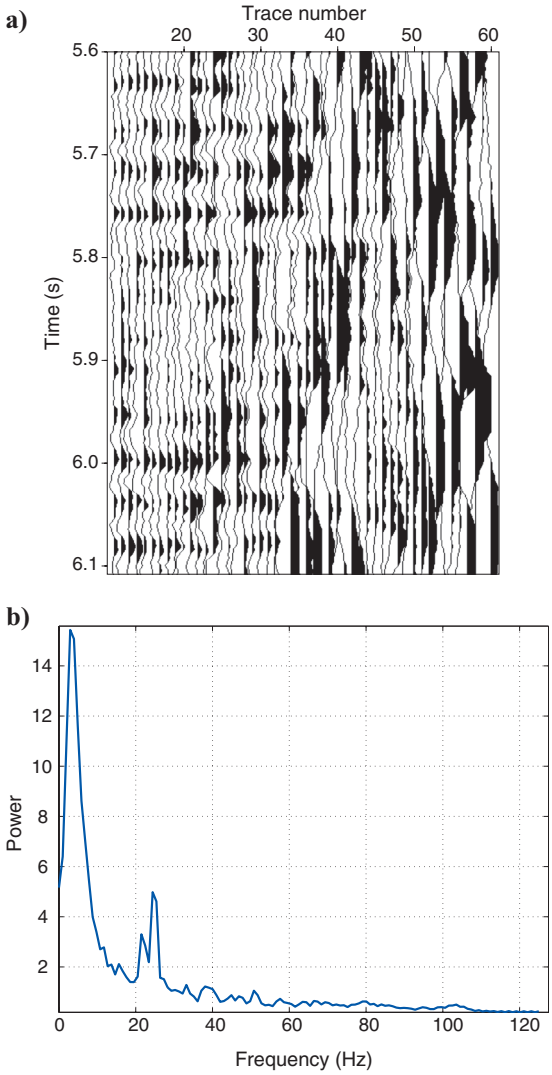

Figure 8. A small data window extracted from Figure 1. (a) The data window contains some weak flat multiples mixed with strong swell noise. (b) The average power spectrum of the data shows that swell noise is concentrated in the frequency band from 2 to 15 Hz.

results compared to a more conservative one with $\alpha = 2$ (Figure 9a and b). The proposed method shown in Figure 9e and f performs similarly to the aggressive filter.

These tests demonstrate that α in the conventional method may need to be adjusted in different parts of a single gather to obtain optimum results, e.g., if the noise characteristics are nonstationary in time and position. Our statistical method automatically achieves a more satisfactory balance between noise attenuation and signal preservation for a single parameter choice.

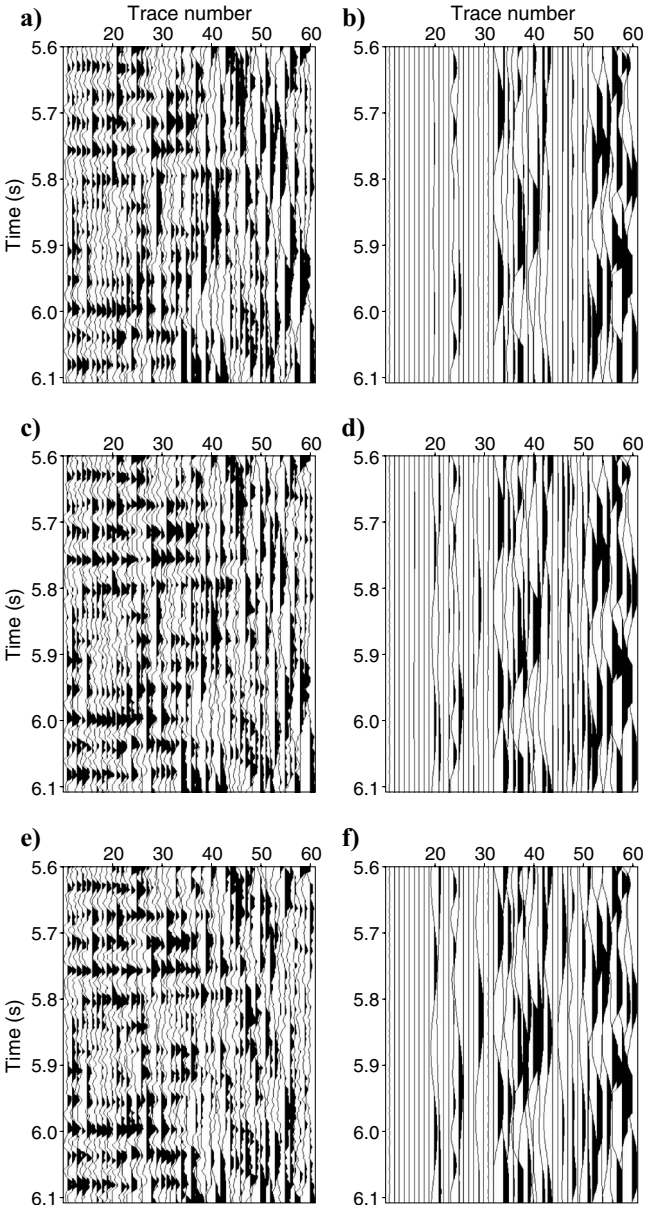

Figure 9. Swell-noise filtering applied to the data in Figure 8. (Left column) Result after swell-noise attenuation. (Right column) Difference section. (a, b) Conventional method with $\alpha = 3$ (mild filter), (c, d) conventional method with $\alpha = 1$ (harsh filter), and (e, f) proposed method with $\beta = 0.5$. A harsh filter (c, d) provides better results than a mild filter (a, b). The proposed method performs similarly to the harsh filter.

Whole section

We apply both detection techniques on the whole section shown in Figure 1. The methods are implemented over a sliding time-space window of 512 ms in length and 50 traces with a 50% overlap in time and space. The conventional method is implemented with a variable threshold factor: from time t of 0–2.80 s, $\alpha = 3.0$; t of 2.80–4.00 s, $\alpha = 2.0$; t of 4.00–4.64 s, $\alpha = 1.2$; and t of 4.64–8.20 s, $\alpha = 1.0$. This thresholding scheme with a decreasing factor α provides the best performance for this gather. On the other hand, our method uses a fixed value $\beta = 0.5$ throughout.

The filtering result of the conventional method (Figure 10a) shows that a large amount of swell noise has been removed; however, some weak components of the refracted arrivals have been attenuated around traces 100 and 400 (Figure 10b). The proposed technique is slightly better, particularly at traces 80–95 and 250–350 (Figure 10c). The seabed event is also slightly better preserved (Figure 10d). More importantly, the statistical method needs no parameter testing and optimization, unlike the conventional method.

The result of using the Rayleigh distribution (Appendix C) is compared with the result obtained previously using the exponential

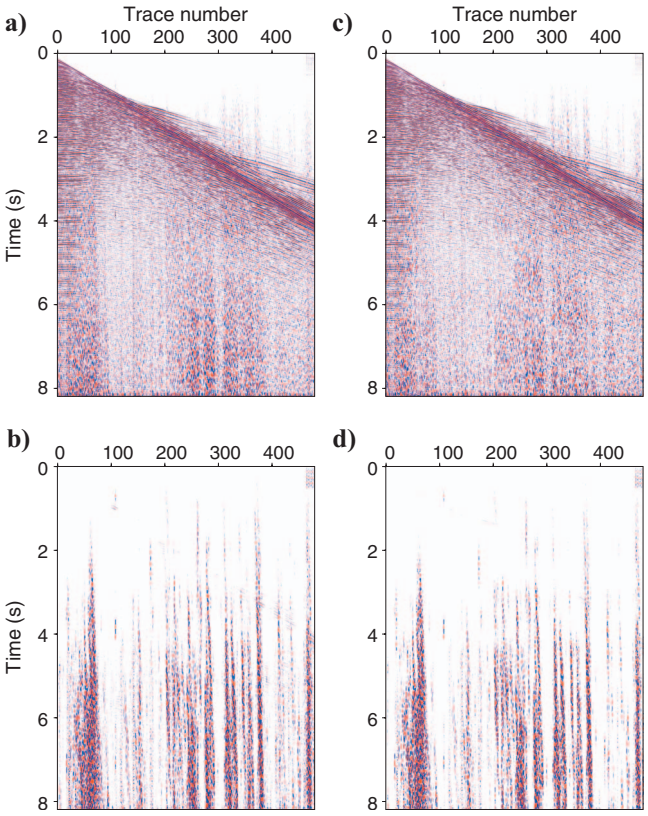

Figure 10. Comparison between the result of (a, b) an optimized conventional method and (c, d) the proposed method for swell-noise attenuation applied on the whole gather in Figure 1. (Top row) Output after swell-noise attenuation. (Bottom row) Difference section. The conventional method was implemented with a time-varying threshold factor α and the proposed method with a constant $\beta = 0.5$ throughout the section. The proposed method gives results comparable to the optimized conventional method (though it is slightly better at traces 80–95 and 250–350) but with much less parameter testing and variation.

distribution (Figure 11). Both results clearly are close, indicating a relatively weak sensitivity to the choice of PDF used in the modeling.

DISCUSSION

Model-based signal processing is a powerful tool to extract signal from noisy observations. The example of using a two-class mixture of PDFs to model the amplitudes of the data demonstrates that a statistical parametric approach may provide better results than a conventional nonparametric approach. The statistical approach leads to an adaptive noise detector applicable across a wider range of varying noise power.

The performance of model-based signal processing depends on the appropriateness of the chosen model. The goodness of fit of S_n to the mixture of exponential model in equation 6 is not assessed because it is beyond the scope of this paper. However, in our experience when the misfit is significant, the estimated model parameters are such that $\hat{\lambda}_0 \approx \hat{\lambda}_1$. As a result, no data are classified as noise and the technique is quite safe, even when the underlying statistical assumptions are violated.

The sensitivity of our method to the choice of the statistical distribution has been investigated by considering a Rayleigh distribution in addition to the exponential distribution. The result shows a relatively weak sensitivity, but generalization to other distributions is yet to be tested. We claim, however, that for data classification, the knowledge of specific statistical moments is enough to make a reliable decision as long as the statistical moments can be considered as reliable discriminators. (The statistical moment of order s for the PDF $p(x)$ is defied as $m_s = \int x^s p(x) dx$.) The fact that swell-noise amplitudes are larger than signal amplitudes implies that possible discriminating statistical moments include the mean (first-order moment) and the rms value (square root of the second-order moment). Hence, if a PDF can capture a discriminating statistical moment through its parameters, then we would expect the result of the detection to be comparable to the one obtained with the exponential or the Rayleigh distribution.

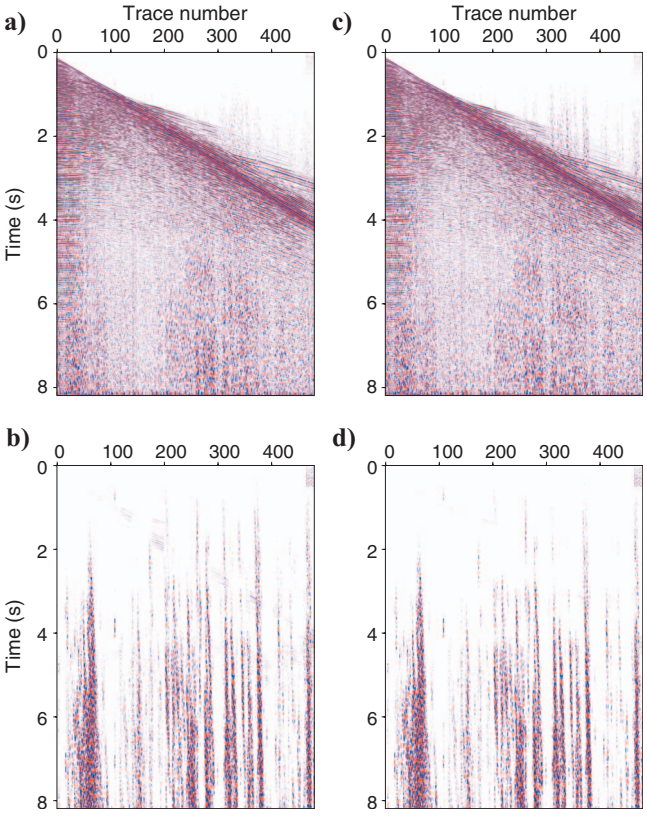

Figure 11. Sensitivity of the proposed method to the choice of the statistical distribution. Comparison between (a, b) a Rayleigh PDF and (c, d) an exponential PDF for swell-noise attenuation applied on the whole gather in Figure 1. (Top row) Output after swell-noise attenuation. (Bottom row) Difference section. Results are similar, suggesting little sensitivity to assumed PDF.

The assumption of independent and identically distributed sampling used in the statistical model (equation 2) is not totally correct for the signal amplitudes. It also is not correct for the noise amplitudes because swell noise contaminates a set of neighboring traces at once in the shot domain. This assumption is used to simplify the mathematical formulation of the problem but may result in a lower detection performance. The conventional noise-detection method in equation 1 assumes implicitly an iid model because the statistic $\hat{s}(S_n)$ does not depend on any data-correlation measure. If we consider a random reordering of the data in S_n, then a pseudo-iid model can be achieved by removing possible correlation between the samples without affecting $\hat{s}(S_n)$. The proposed method is also invariant under random data reordering.

Reducing the number of samples n used to fit the statistical model affects the quality of parameter estimation and hence reduces the effectiveness of noise detection. This issue is shared with the conventional method; the quality of the estimated data statistics $\hat{s}(S_n)$ in equation 1 also depends on n. But because $\hat{s}(S_n)$ involves estimating only one unknown, compared with fitting the parametric model in equation 6 with three unknowns, conventional methods are expected to suffer less when n is small. Therefore, when applying the proposed method on a short spatial window (e.g., 10 traces), one should consider analyzing a packet of frequencies at a time rather than a single frequency.

We anticipate that better swell-noise attenuation results could be obtained if noisy samples were interpolated rather than just rescaled. However, the focus of this paper is on the detection of noisy samples, which is a prerequisite of any target-oriented interpolation scheme. Furthermore, the described rescaling can provide an initial estimate for any f-x interpolation based on autoregressive modeling (Canales, 1984; Soubaras, 1995).

CONCLUSION

Automatic detection of large, noisy amplitudes can be achieved using statistical modeling. The resulting statistical detection method is equivalent to an optimized conventional one but with much less parameter testing and variation. Tests on swell-noise attenuation show consistent results for a wide range of varying noise levels across the data. The method is generic and can be generalized to other types of anomalous high-amplitude noise such as spikes and diffraction multiples noise.

ACKNOWLEDGMENTS

The authors thank Fugro Multi-clients for providing the real data set and for permission to publish it. Special thanks go to Thomas Elboth for fruitful discussions and to anonymous reviewers for their

helpful comments and suggestions. This work was part of a research collaboration between the University of Leeds and Fugro when the authors were employed at the University of Leeds.

APPENDIX A

DISTRIBUTION OF POWER SPECTRUM

Let X_1 and X_2 be two independent zero-mean Gaussian random variables with equal variance σ^2. We are interested in finding the PDF of the random variable $Z = X_1^2 + X_2^2$. For the problem considered in this paper, X_1 and X_2 represent the real part and imaginary part of the Fourier transform and Z is the power spectrum. The cumulative density function is defined as

$$F(z) = \Pr\{Z \le z\} = \Pr\{X_1^2 + X_2^2 \le z\} = \int\int_{x_1^2 + x_2^2 \le z} (2\pi\sigma^2)^{-1} \exp\left(-\frac{x_1^2 + x_2^2}{2\sigma^2}\right) dx_1 dx_2. \tag{A-1}$$

By changing the variables $x_1 = r\cos(\theta)$ and $x_2 = r\sin(\theta)$, the double integral in equation A-1 can be rewritten as

$$F(z) = \int_0^{2\pi}\int_0^{\sqrt{z}} (2\pi\sigma^2)^{-1} r \exp\left(\frac{-r^2}{2\sigma^2}\right) dr d\theta = 1 - \exp\left(\frac{-z}{2\sigma^2}\right). \tag{A-2}$$

The PDF of Z, $p(z) = dF(z)/dz = (2\sigma^2)^{-1}\exp(-z/2\sigma^2)$, is an exponential distribution with a location parameter $\lambda = 2\sigma^2$.

APPENDIX B

EXPECTATION-MAXIMIZATION ALGORITHM

The EM algorithm is used to find the MLEs of parameters in a statistical model, where the model depends on some unobserved or missing variables. Given a statistical model defined by its PDF, $P(\mathbf{D},\mathbf{Z};\boldsymbol{\theta})$, where $\mathbf{D}$ is the observed data, $\mathbf{Z}$ is the unobserved or missing data, and θ is the vector of parameters, the MLE is defined by the marginal likelihood of $\mathbf{Z}$ (Redner and Walker, 1984):

$$\boldsymbol{\theta}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta};\mathbf{D}) = \arg\max_{\boldsymbol{\theta}} \int P(\mathbf{D},\mathbf{Z};\boldsymbol{\theta}) d\mathbf{Z}. \tag{B-1}$$

However, $L(\boldsymbol{\theta};\mathbf{D})$ is often difficult to compute. The EM algorithm seeks to find the MLE by applying two steps iteratively:

1) *Expectation step (E-step)*. Calculate the expected value of the log-likelihood function with respect to the conditional distribution of $\mathbf{Z}$, given $\mathbf{D}$ under the current estimate of the parameters $\boldsymbol{\theta}^{(t)}$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\mathbf{Z}|\mathbf{D},\boldsymbol{\theta}^{(t)}}\{\log L(\boldsymbol{\theta};\mathbf{D},\mathbf{Z})\}. \tag{B-2}$$

Note that maximizing $L(\boldsymbol{\theta};\mathbf{D})$ is equivalent to maximizing $\log L(\boldsymbol{\theta};\mathbf{D})$.

2) *Maximization step (M-step)*. Find the parameter that maximizes this quantity, i.e.,

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}). \tag{B-3}$$

EM is particularly useful when the likelihood function has an exponential form. The E-step becomes the sum of expectations of some easy statistics, and the M-step involves maximizing a linear function. In such a case, it usually is possible to derive closed-form updates for each step.

Example: Exponential mixture

Let $R_n = \{r_1, r_2, \ldots, r_n\}$ be a sample realization from independent observations generated by a mixture of two exponential distributions defined as

$$g(r|\epsilon,\lambda_0,\lambda_1) = \epsilon p(r|\lambda_0) + (1-\epsilon)p(r|\lambda_1), \tag{B-4}$$

where

$$p(r|\lambda_j) = \lambda_j^{-1}\exp\left(\frac{-r}{\lambda_j}\right), \quad \text{with } r \ge 0, \quad \lambda_j > 0, \quad j = 0,1. \tag{B-5}$$

Here, λ_j represents the mean of the jth distribution and ϵ is the mixing ratio. We introduce the set of missing variables $\mathbf{Z}_n = \{z_1, z_2, \ldots, z_n\}$ that determines the component from which each observation $\{r_i\}_{i=1}^n$ originates, i.e.,

$$g(r_i|z_i = 1) = p(r|\lambda_1) \text{ and } g(r_i|z_i = 0) = p(r|\lambda_0),$$

with

$$\Pr(z_i = 1) = \epsilon \text{ and } \Pr(z_i = 0) = 1 - \epsilon. \tag{B-6}$$

The aim is to estimate the unknown parameters $\boldsymbol{\theta} = (\epsilon,\lambda_0,\lambda_1)$. The MLE using the observed and missing data obtained by maximizing the log-likelihood function is

$$\begin{aligned}\log L(\boldsymbol{\theta};R,\mathbf{Z}) &= \log\prod_{i=1}^n \left[\frac{\epsilon}{\lambda_1}\exp\left(\frac{-r_i}{\lambda_1}\right)\right]^{z_i} \times \left[\frac{1-\epsilon}{\lambda_0}\exp\left(\frac{-r_i}{\lambda_0}\right)\right]^{(1-z_i)} \\ &= \log\left(\frac{\epsilon}{\lambda_1}\right)\sum_{i=1}^n z_i - \sum_{i=1}^n \frac{z_i r_i}{\lambda_1} + \log\left(\frac{1-\epsilon}{\lambda_0}\right)\sum_{i=1}^n (1-z_i) - \sum_{i=1}^n \frac{(1-z_i)r_i}{\lambda_0}.\end{aligned} \tag{B-7}$$

E-step

Given the current estimate of the parameter $(\lambda_0^{(t)},\lambda_1^{(t)},\epsilon^{(t)})$, the conditional expectation of the binary variable z_i is given by Bayes' theorem:

$$A_i^{(t)} = \mathbb{E}_{z_i|r_i,\lambda_0^{(t)},\lambda_1^{(t)},\epsilon^{(t)}}\{z_i\} = \Pr\{z_i = 1|r_i;\epsilon^{(t)},\lambda_0^{(t)},\lambda_1^{(t)}\}$$

$$= \frac{\dfrac{\epsilon^{(t)} \exp\left(\dfrac{-r_i}{\lambda_1^{(t)}}\right)}{\lambda_1^{(t)}}}{\dfrac{\epsilon^{(t)} \exp\left(\dfrac{-r_i}{\lambda_1^{(t)}}\right)}{\lambda_1^{(t)}} + \dfrac{(1-\epsilon^{(t)})\exp\left(\dfrac{-r_i}{\lambda_0^{(t)}}\right)}{\lambda_0^{(t)}}}. \quad \text{(B-8)}$$

Thus, the E-step is

$$Q(\epsilon,\lambda_0,\lambda_1|\epsilon^{(t)},\lambda_0^{(t)},\lambda_1^{(t)}) = \log\left(\frac{\epsilon}{\lambda_1}\right)\sum_{i=1}^{n} A_i^{(t)} - \sum_{i=1}^{n}\frac{A_i^{(t)} r_i}{\lambda_1} + \log\left(\frac{1-\epsilon}{\lambda_0}\right)\sum_{i=1}^{n}(1-A_i^{(t)}) - \sum_{i=1}^{n}\frac{(1-A_i^{(t)})r_i}{\lambda_0}. \quad \text{(B-9)}$$

M-step

The expression in equation B-9 has a linear form with respect to $A_i^{(t)}$, and this makes the optimization for the unknown parameters $\epsilon,\lambda_0,\lambda_1$ straightforward. The M-step results in

$$\epsilon^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} A_i^{(t)}, \quad \text{(B-10)}$$

$$\lambda_1^{(t+1)} = \frac{\sum_{i=1}^{n} A_i^{(t)} r_i}{\sum_{i=1}^{n} A_i^{(t)}}, \quad \text{(B-11)}$$

and

$$\lambda_0^{(t+1)} = \frac{\sum_{i=1}^{n} (1-A_i^{(t)}) r_i}{\sum_{i=1}^{n} (1-A_i^{(t)})}, \quad \text{(B-12)}$$

where

$$A_i^{(t)} = \frac{p(r_i|\lambda_1^{(t)})\epsilon^{(t)}}{p(r_i|\lambda_1^{(t)})\epsilon^{(t)} + p(r_i|\lambda_0^{(t)})(1-\epsilon^{(t)})}. \quad \text{(B-13)}$$

APPENDIX C

ALTERNATIVE PDF

The Rayleigh distribution is defined as

$$p(r|\sigma) = \frac{r}{\sigma^2}\exp\left(\frac{-r^2}{2\sigma^2}\right), \quad \text{with } r \geq 0, \quad \text{(C-1)}$$

where σ represents the mode of the distribution. The equivalent amplitude threshold is r_{Ray} (where Ray stands for Rayleigh) is

$$r_{\text{Ray}}^2 = 2\frac{\hat{\sigma}_1^2}{\dfrac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2} - 1}\left[\log\left(\frac{1-\hat{\epsilon}}{\hat{\epsilon}}\right) + \log\left(\frac{\beta}{1-\beta}\right) + \log\left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}\right)\right]. \quad \text{(C-2)}$$

The unknown parameters $(\sigma_0,\sigma_1,\epsilon)$ are computed iteratively, in a similar way to the parameters of the exponential model, using the EM algorithm:

$$\epsilon^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} A_i^{(t)}, \quad \text{(C-3)}$$

$$\sigma_1^{(t+1)} = \sqrt{\frac{1}{2}\frac{\sum_{i=1}^{n} A_i^{(t)} r_i^2}{\sum_{i=1}^{n} A_i^{(t)}}}, \quad \text{(C-4)}$$

and

$$\sigma_0^{(t+1)} = \sqrt{\frac{1}{2}\frac{\sum_{i=1}^{n} (1-A_i^{(t)}) r_i^2}{\sum_{i=1}^{n} (1-A_i^{(t)})}}, \quad \text{(C-5)}$$

where

$$A_i^{(t)} = \frac{p(r_i|\sigma_1^{(t)})\epsilon^{(t)}}{p(r_i|\sigma_1^{(t)})\epsilon^{(t)} + p(r_i|\sigma_0^{(t)})(1-\epsilon^{(t)})}. \quad \text{(C-6)}$$

The values σ_0 and σ_1 represent, respectively, $2^{-1/2}$ times the rms value of the signal and the noise amplitudes.

REFERENCES

Anderson, R. G., and G. A. McMechan, 1989, Automatic editing of noisy seismic data: Geophysical Prospecting, **37**, 875–892.

Bekara, M., and M. van der Baan, 2007, Local singular value decomposition for signal enhancement of seismic data: Geophysics, **72**, no. 2, V59–V65.

Cambois, G., and J. Frelet, 1995, Can we surgically remove swell noise?: 65th Annual International Meeting, SEG, Expanded Abstracts, 1381–1384.

Canales, L., 1984, Random noise reduction: 54th Annual International Meeting, SEG, Expanded Abstracts, 525–527.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977, Maximum likelihood from incomplete data via the EM algorithm: Journal of the Royal Statistical Society, Series R, **39**, 1–38.

Elboth, T., H. Qaisrani, and T. Hertweck, 2008, De-noising seismic data in the time-frequency domain: 73rd Annual International Meeting, SEG, Expanded Abstracts, 2622–2626.

Elboth, T., B. A. P. Reif, and O. Andreassen, 2009, Flow and swell noise in marine seismic data: Geophysics, **74**, no. 2, Q17–Q25.

Freire, S. L. M., and T. J. Ulrych, 1988, Application of singular value decomposition to vertical seismic profiling: Geophysics, **53**, 778–785.

Hasselblad, V., 1969, Estimation of finite mixtures of distributions from the exponential family: Journal of the American Statistical Association, **64**, 1459–1471.

Kay, S., 1988, Modern spectral estimation: Theory and application: Prentice-Hall, Inc.

Liu, Y., C. Lui, and D. Wang, 2009, A 1D time-varying median filter for seismic random, spike-like noise elimination: Geophysics, **74**, no. 1, V17–V24.
Papoulis, A., and S. U. Pillai, 2002, Probability, random variables and stochastic processes, 4th ed.: McGraw-Hill Book Co.
Redner, R. A., and H. F. Walker, 1984, Mixture densities, maximum likelihood and the EM algorithm: SIAM Review, **26**, 195–239.
Rousseeuw, P., and A. Leroy, 1987, Robust regression and outlier detection: Wiley Interscience.
Schonewille, M., A. Vigner, and A. Ryder, 2008, Swell-noise attenuation using an iterative *f-x* prediction filtering approach: 78th Annual International Meeting, SEG, Expanded Abstracts, 2647–2651.
Soubaras, R., 1995, Prestack random and impulsive noise attenuation by *f-x* projection filtering: 65th Annual International Meeting, SEG, Expanded Abstracts, 711–714.
van der Baan, M., 2006, PP/PS wavefield separation by independent component analysis: Geophysical Journal International, **166**, 339–348.
Verschuur, D. J., A. J. Berkhout, and C. P. A. Wapenaar, 1992, Adaptive surface-related multiple elimination: Geophysics, **57**, 1166–1177.