A FREQUENCY CRITERION FOR SMOOTHING WITH OPTIMAL CUBIC SPLINES

J. Schleicher and R. Biloti

email: *js@ime.unicamp.br*
keywords: *smoothing, cubic splines, approximation*

ABSTRACT

When smoothing a function with high-frequency noise by means of optimal cubic splines, it is often not clear how to choose the number of nodes. The more nodes are used, the closer the smoothed function will follow the noisy one. In this work, we demonstrate that more nodes mean a better approximation of Fourier coefficients for higher frequencies. Thus, the number of nodes can be determined by specifying a frequency up to which all Fourier coefficients must be preserved. A comparison of the corresponding smoothing results with those obtained by filtering using a moving average of corresponding length and a lowpass with corresponding high-cut frequency show that optimal cubic splines yield better results as they preserve not only the desired low-frequency band but also important high-frequency characteristics.

INTRODUCTION

The smoothing of functions that contain high-frequency noise is a task that is frequent to seismic applications. One of the most obvious examples is the seismic trace that contains incoherent noise which can be reduced by some kind of filtering. Other examples are velocity fields as obtained from some processing or imaging technique applied to the seismic data, or any other parameters that are the result of seismic data processing or inversion.

Recently, a new technique of spline approximation has been proposed (Biloti, 2002; Biloti et al., 2002, 2003). This technique differs from standard spline interpolation in the fact that the node positions are chosen optimally in a least-squares sense. In this way, the so-called *optimal cubic splines* approximation will always find that particular smooth function with the smallest deviation from the original function, given a fixed number of nodes.

This difference between conventional and optimal cubic splines implies a difference in the choice of the number of nodes. When working with conventional splines, the interval of interest is often divided into a fixed number of equal subintervals using regularly spaced nodes. The size of such a subinterval can be determined by some frequency criterion such as a Nyquist condition for the frequencies to be preserved after smoothing.

For optimal cubic splines, however, it is not entirely clear how to choose an equivalent criterion. If the same number of nodes as for a conventional spline interpolation is chosen, the smoothed function might actually be oversampled because of the optimality of the solution. The quality of the approximation is not a simple function of the number of nodes. Sometimes the approximation can be improved dramatically by adding a single node, while in other cases, even the addition of several nodes cannot improve the quality of the approximation.

In this work, we investigate how the number of nodes in optimal cubic splines acts on the reconstruction of the Fourier coefficients of the original function. We show that more nodes lead to the reconstruction of Fourier coefficients up to higher frequencies. Thus, the number of nodes can be determined by specifying a frequency up to which all Fourier coefficients must be preserved. We discuss the results of optimal

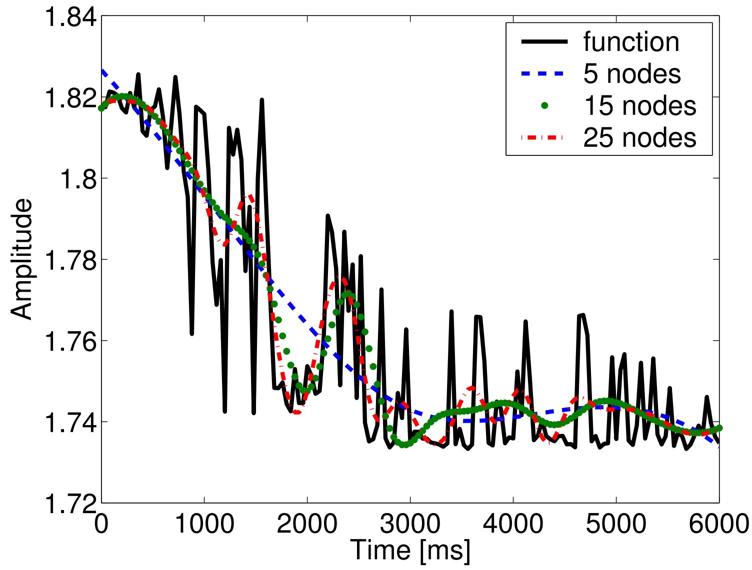

Figure 1: A noisy function together with three different smoothed versions as obtained by optimal cubic splines with 5, 15, and 25 nodes.

cubic-splines smoothing using the so-determined number of nodes and compare them with those obtained by filtering using a moving average of corresponding length and a lowpass with corresponding high-cut frequency.

OPTIMAL SPLINES

The method to determine the approximation of a given function by optimal cubic splines detects the best positions of the given number of nodes. The penalty function is the deviation of the smoothed function from the original function to be approximated. The optimization solver employed is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) conjugate-gradient algorithm (Ueberhuber, 1997; Dennis and Schnabel, 1996). It is a quasi-Newton method which builds up an approximation to the second derivatives of the objective function using the difference between successive gradient vectors. By combining the first and second derivatives, the algorithm is able to take Newton-type steps towards the function minimum, assuming quadratic behavior in that region.

NUMBER OF NODES

The number of nodes is a crucial parameter for any spline interpolation or approximation. The actual degree of smoothing or preservation of details of the original function depends largely on the number of nodes. Thus, it has to be chosen in accordance with the actual needs.

As an example, consider the noisy function of Figure 1. Also shown are three smoothed versions of that function, obtained by optimal cubic splines with 5, 15 or 25 nodes, respectively. We see the different degree of detail preserved by the different approximations. Executing the smoothing by optimal cubic splines using more nodes results in a better fit of the smoothed to the unsmoothed function.

However, none of the three choices might be adequate. Even if we decide we need the detail preserved by 25 nodes, it might well be possible that less nodes can provide the very same amount of detail.

A quantitative measure of this better fit can be obtained in the frequency domain. Figure 2 depicts the total deviation, i.e., the square root of the sum over the squared deviations, between the Fourier coefficients of the original and smoothed functions, as a function of the number of nodes. We see that although the general tendency is a decay with increasing number of nodes, there are node numbers where the decay is particularly strong and there are node numbers where little or nothing is gained from adding one more node.

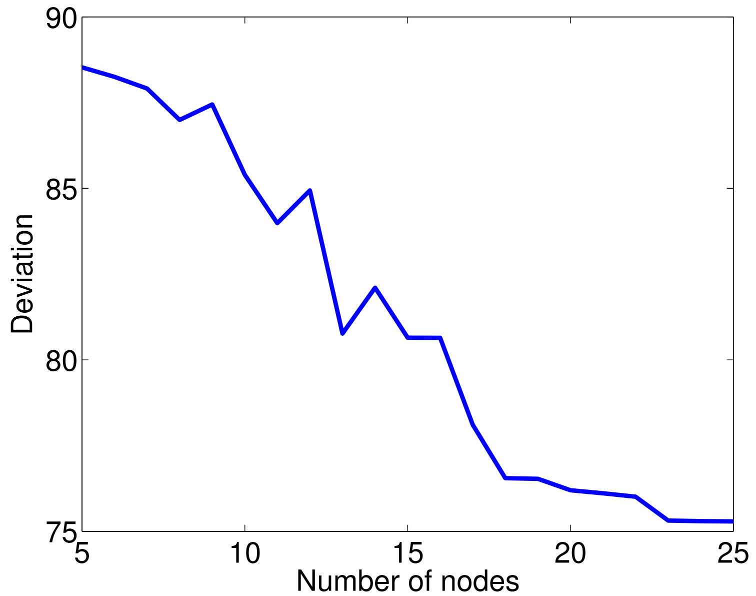

Figure 2: Total deviation between the Fourier coefficients of the original and smoothed functions, as a function of the number of nodes.

To investigate the quality of detail preserved by a certain number of nodes, let us take a look at the Fourier coefficients of the original time series as well as its smoothed versions. These are depicted in Figure 3. We see that the approximation with 5 nodes just preserves the zero-frequency coefficient (not inside the choice of axes in Figure 3) and the first coefficient to each side. All other coefficients are approximately zero. With 15 nodes, we have already 6 fitting coefficients to each side of frequency zero. Finally, with 25 nodes, all coefficients depicted in Figure 3, i.e., 11 to each side, have been reasonably well preserved.

The above observations lead to the idea of how to actually choose the number of nodes. Since the frequency up to which the Fourier coefficients match the original ones increases with an increasing number of nodes, one can just specify a frequency up to which such a match is required. From this condition, the necessary number of nodes can be determined.

In Figure 4, two smoothing results are compared that were obtained by the application of this criterion. The smoothed function obtained with 8 nodes (dashed curve in Figure 4) is the first to preserve all Fourier coefficients up to 3.1 Hz (corresponding to a period of 2000 ms), while the one with 22 nodes (dashed-dotted curve in Figure 4) is the first to preserve all frequencies up to 7.9 Hz (800 ms). Here, we considered a Fourier coefficient as matched, i.e., the corresponding frequency as preserved, when its error was below 10%.

The corresponding Fourier coefficients are depicted in Figure 5. We see that the approximation with 8 nodes preserves 4 Fourier coefficients to each side. Due to the size of the original time series, the frequency increment is $\Delta\omega = 1.04$ Hz. Therefore frequencies up to 4.16 Hz are actually preserved by this smoothed function, more than the requirement of 3.1 Hz. The reason is that the smoothed function with 7 nodes does only preserve 2 Fourier coefficients to each side, i.e., frequencies up to 2.08 Hz, which is less than required by the frequency criterion. By adding one more node, the preserved frequency actually doubled.

Correspondingly, the smoothed function with 22 nodes preserves 10 Fourier coefficients. Therefore, it actually preserves frequencies up to 9.36 Hz, more than actually required by the criterion. With 21 nodes (not depicted), only 6 coefficients are preserved with an error smaller than 10%, i.e., frequencies up to 6.24 Hz. Again, adding one more node adds the required precision to more than one Fourier coefficient.

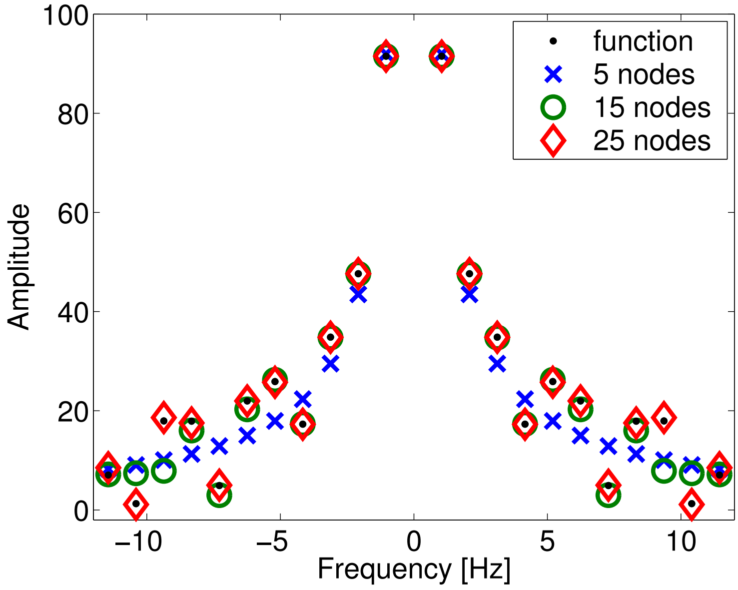

Figure 3: Fourier coefficients of the noisy function of Figure 1 and its three different smoothed versions as obtained by optimal cubic splines with 5, 15, and 25 nodes.

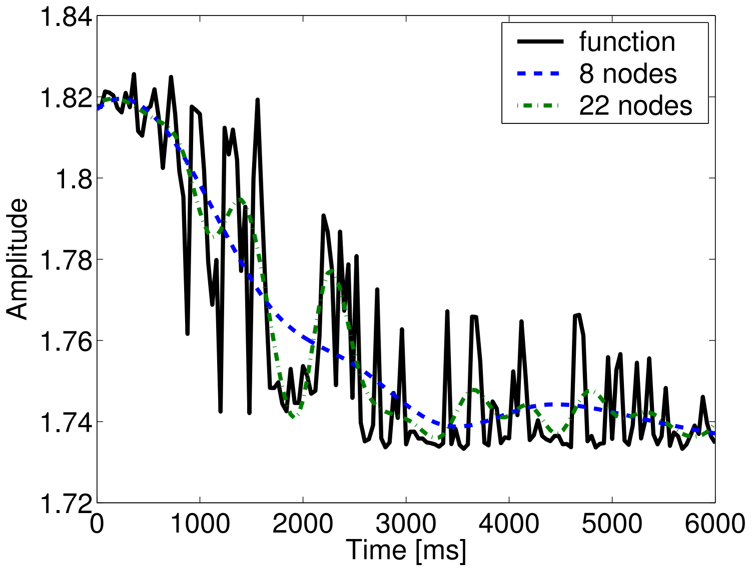

Figure 4: Smoothed functions for preservation of characteristic periods down to 2000 ms (dashed) and 800 ms (dash-dotted curve).

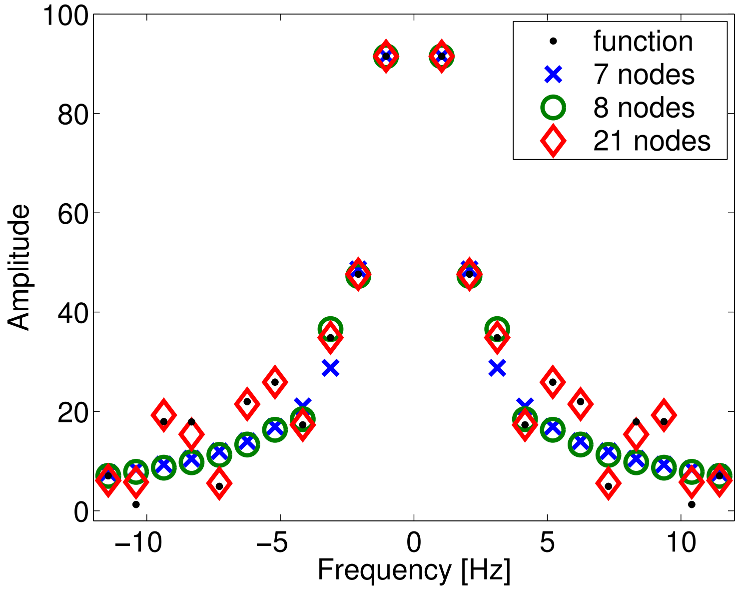

Figure 5: Fourier coefficients of noisy function and three different smoothed versions as obtained by optimal cubic splines with 8 and 22 nodes.

COMPARISON TO MOVING AVERAGE AND LOWPASS

The above procedure of determining the adequate number of nodes for optimal cubic splines leads to some natural questions. Since the criterion is just frequency preservation up to a given limit frequency, the first question is: Could a simple lowpass up to that frequency achieve the same degree of smoothing? Moreover, since this frequency corresponds to some minimum period that one would like to preserve, a second question is: Could a moving average with this length lead to a comparable smoothing result?

In this section, we address these questions by comparing the smoothed functions obtained with optimal cubic splines, moving average, and lowpass filtering.

Example 1

As the first example, we use the function that has already been utilized in the previous section. In Figure 6, we see the smoothed curve with 8 nodes of Figure 4, which preserves frequencies up to 3.1 Hz, i.e., periods down to 2000 ms, together with the smoothed functions resulting from a lowpass with the corresponding cutoff frequency of 3.1 Hz and from a moving average with the corresponding length of 2000 ms. The moving average was applied three times.

We see from Figure 6 that the lowpass has severe problems at the boundaries. The moving average is almost identical to the solution of the optimal cubic splines. The only significant difference between the two curves can be observed at the left side of the figure, where the moving average result clearly lies below the original function while optimal cubic splines seem to do a better job in staying within the range of values of the original function.

A corresponding comparison up to higher frequencies (7.9 Hz or 800 ms) is shown in Figure 7. As in Figure 6, also in this example involving higher frequencies, the lowpass presents severe problems at the boundaries. The most significant difference between the results of moving average and optimal cubic splines, however, has changed. Now, the latter solution preserves the valley at about 2000 ms much better than the former and, in fact, even better than the lowpass result. This observation indicates that optimal cubic splines will preserve characteristic features of the original function even if they involve higher frequencies than the chosen limit frequency. The moving average tends to smooth out such parts almost completely.

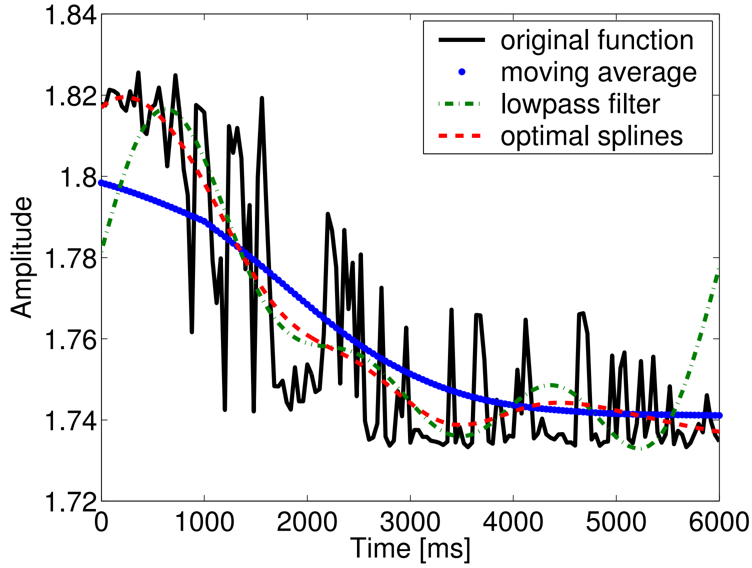

Figure 6: Comparison of results of optimal cubic-splines (dashed) for frequencies up to 3.1 Hz with corresponding moving average (dotted) and lowpass (dash-dotted line) results.

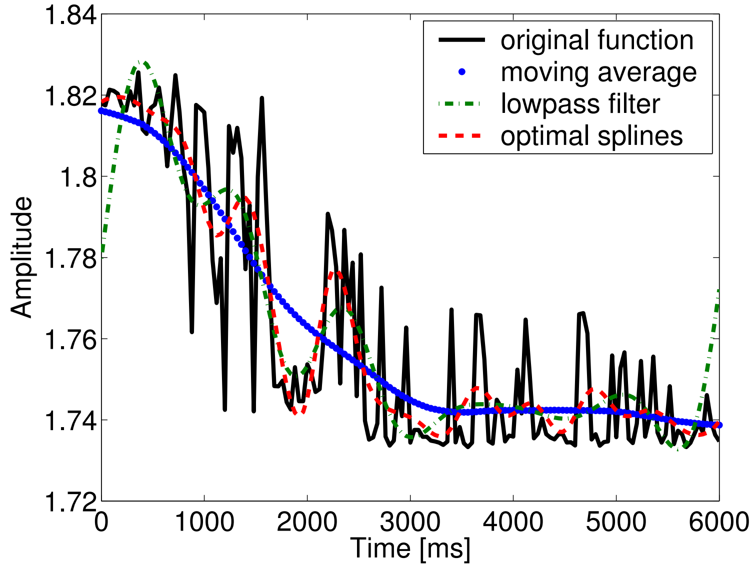

Figure 7: Comparison of results of optimal cubic splines (dashed) for frequencies up to 7.9 Hz with corresponding moving average (dotted) and lowpass (dash-dotted line) results.

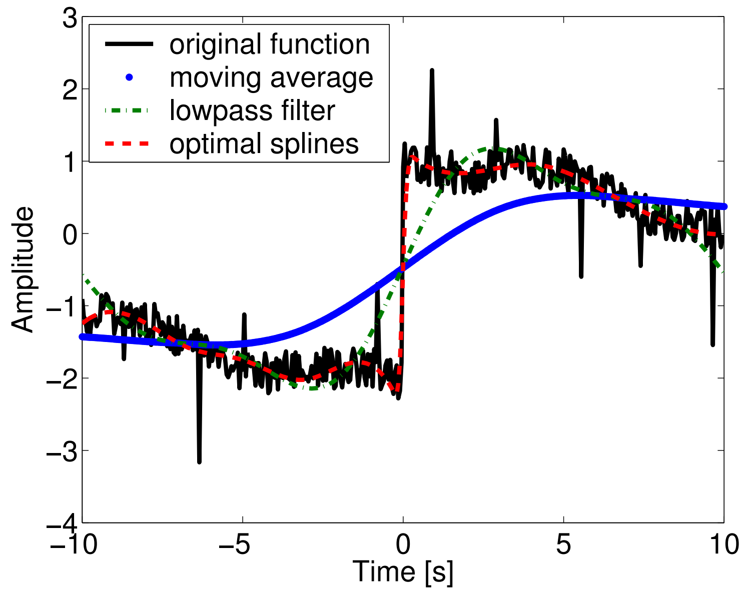

Figure 8: Comparison of results of optimal cubic splines (dashed) for frequencies up to 1.05 Hz (period 6 s) with corresponding moving average (dotted) and lowpass (dash-dotted line) results.

Example 2

The above observation can be confirmed in our second example. The original function is a simple sine with a frequency of 1 Hz (period of 2π s), to which a dipping linear function, a step function, and random noise have been added. Figure 8 compares the results of smoothing by optimal cubic splines, moving average, and lowpass. The limit period was chosen as 6 s, i.e., the limit frequency was 1.05 Hz. This requires 13 nodes for the optimal cubic splines. Again, all methods have been applied using the same length scales and frequencies. As we can see, the optimal cubic-splines approximation does a fantastic job in preserving the step while nicely smoothing out the noise. On the other hand, the lowpass and the moving average cannot preserve the step. Moreover, as before we note the problems of the lowpass at the ends of the approximation interval.

Even if the limit frequency is chosen too small to include the frequency of the underlying original sine function, optimal cubic-splines approximation does a much better job of preserving the main characteristics of the function than lowpass or moving average filtering (see Figure 9). Here, the filter length was chosen as 10 s, which corresponds to a frequency of 0.63 Hz. Therefore, the number of necessary nodes reduces to 7. For this filter length, the moving average completely destroys the properties of the function.

Example 3

The third example confirms the conclusions drawn from the previous ones. As we can see in Figure 10, the optimal cubic-splines approximation (here taken down to a period of 200 ms, i.e., up to a frequency of 31.4 Hz) can preserve a reasonable amount of detail from the original function while simultaneously smoothing out the noise, although the ideal number of nodes needed was just 7. In this case, we even note a more drastic failure of the lowpass filtering, which seems to even oscillate contrarily to the original function. The moving average, on the other hand, again smoothes out almost any detail, preserving not much more than the basic background trend. As before, close to the boundaries, the moving average result does not lie with the range of function values.

If we go to higher frequencies or lower periods (62.8 Hz or 100 ms, see Figure 11), the number of nodes necessary to fulfill the frequency criterion increases to 11. We see that the optimal cubic-splines solution behaves basically the same as the lowpass one, except for the better approximation close to the ends of the interval.

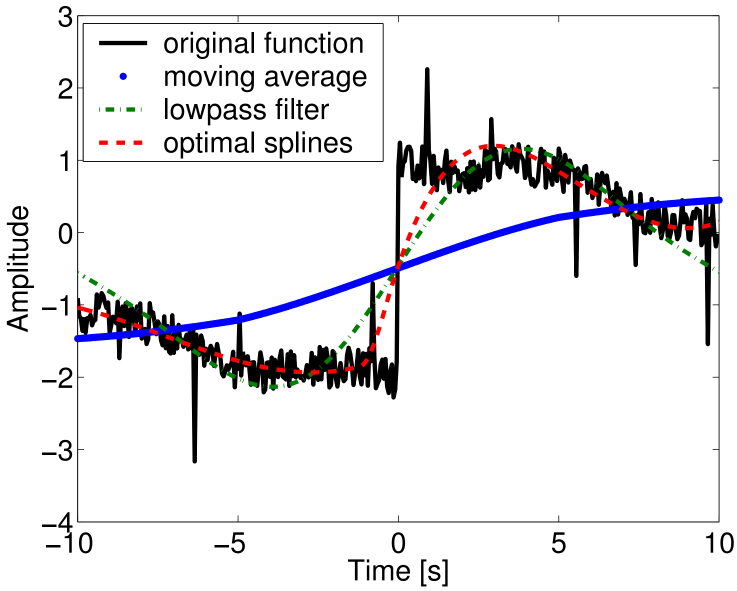

Figure 9: Comparison of results of optimal cubic splines (dashed) for frequencies up to 0.63 Hz (period 10 s) with corresponding moving average (dotted) and lowpass (dash-dotted line) results.

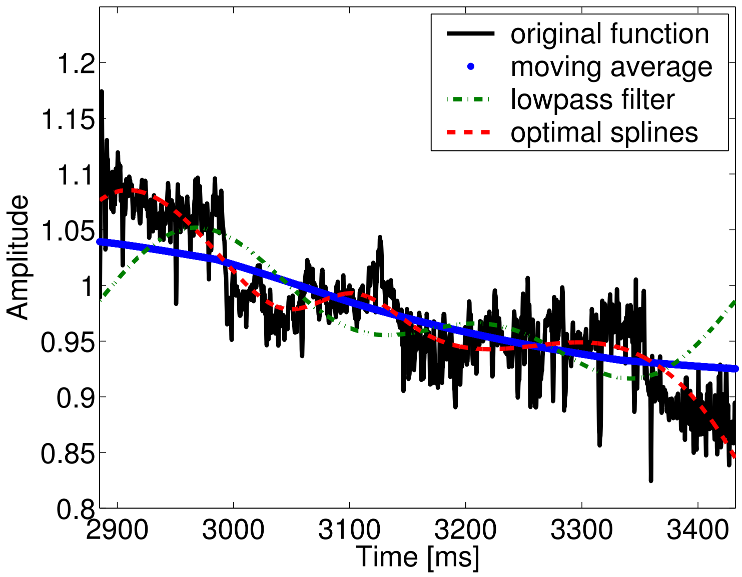

Figure 10: Comparison of results of optimal cubic splines (dashed) for frequencies up to 31.4 Hz (period 200 ms) with corresponding moving average (dotted) and lowpass (dash-dotted line) results.

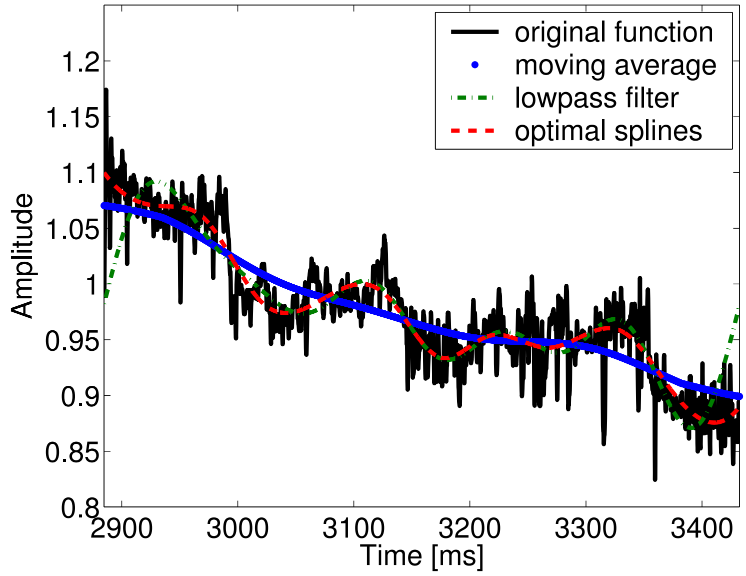

Figure 11: Comparison of results of optimal cubic splines (dashed) for frequencies up to 62.8 Hz (period 100 ms) with corresponding moving average (dotted) and lowpass (dash-dotted line) results.

Example 4

The fourth example is a synthetic seismic trace with a Ricker wavelet with a duration of 128 ms, i.e., a peak frequency of 49.1 Hz. The result of smoothing using these values as the parameter for the algorithms is depicted in Figure 12. The optimal cubic splines need 30 nodes to preserve this limit frequency. While the optimal cubic splines almost completely preserve the original wavelet, this can be said of neither of the moving average nor the lowpass filter.

Even when using a lower frequency, optimal cubic-splines smoothing preserves more of the information of the wavelet that the other two methods. Figure 13 shows the corresponding result for a limit frequency of 15.6 Hz (period of 400 ms), which is the peak frequency of the original noisy trace. For this limit frequency, we need only 17 nodes in the optimal cubic splines. We see that in this case, the moving average and lowpass smoothing just zero out the event, while optimal cubic-splines smoothing preserves the wavelet, although destroying its amplitude.

To have lowpass filtering yield a better result, we increased the frequency parameter to 78.5 Hz (period of 80 ms, see Figure 14). In this case, the optimal cubic splines need 36 nodes. However, a comparison to the optimal cubic-splines approximation with 30 nodes of Figure 12 reveals that almost nothing has been gained from the additional 6 nodes. Still, even for this filter length, the bandpass destroys much of the signal and the moving average almost completely eliminates the wavelet from the trace.

CONCLUSIONS

When smoothing a function with high-frequency noise by means of optimal cubic splines, it is often not clear how to choose the number of nodes. The more nodes are used, the more detail of the original function will be preserved by the smoothed function. However, the increase in quality is strongly nonlinear with the number of nodes. In this work, we have demonstrated that more nodes mean a better approximation of Fourier coefficients for higher frequencies. Thus, the number of nodes can be determined by specifying a frequency up to which all Fourier coefficients must be preserved.

To illustrate our findings, the corresponding smoothing results have been compared numerically to those obtained by filtering using a moving average of corresponding length and a lowpass with corresponding high-cut frequency. Our numerical examples show that there are cases where one or the other of the reference methods can achieve comparable quality to optimal cubic-splines smoothing. However, evaluating the results of the whole set of examples, we conclude that optimal cubic splines always yield a very good re-

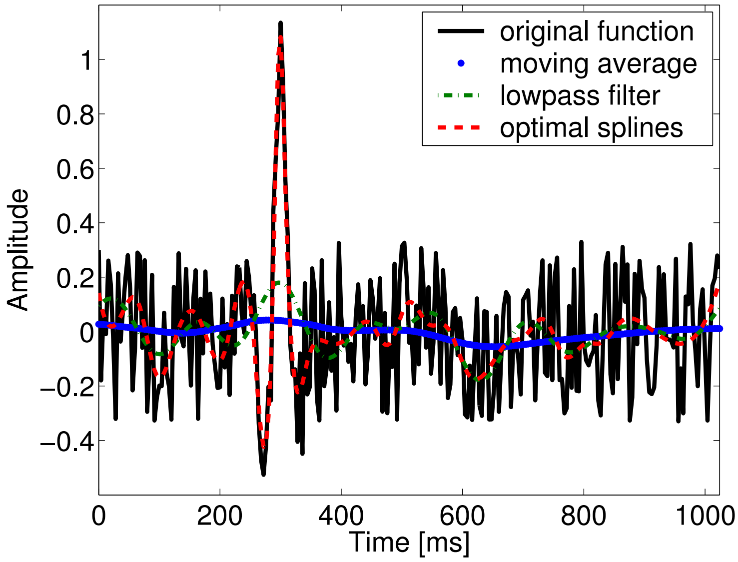

Figure 12: Comparison of results of optimal cubic splines (dashed) for frequencies up to 49.1 Hz (period 128 ms) with corresponding moving average (dotted) and lowpass (dash-dotted line) results.

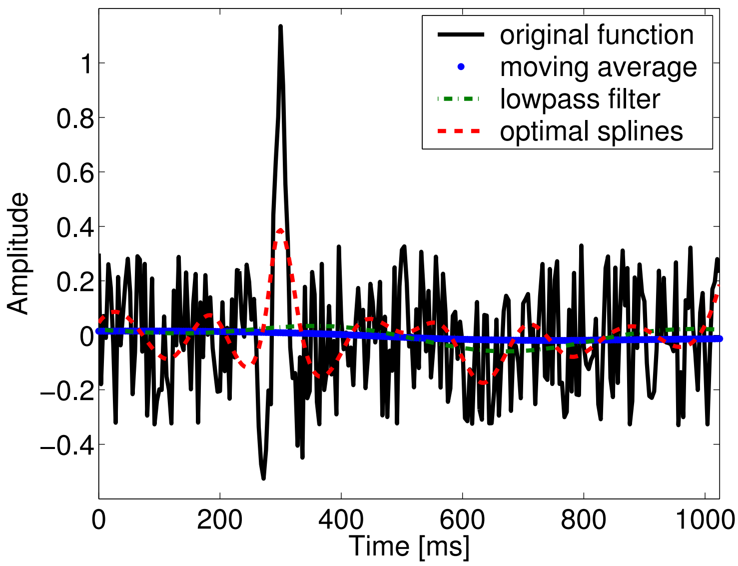

Figure 13: Comparison of results of optimal cubic splines (dashed) for frequencies up to 15.6 Hz (period 400 ms) with corresponding moving average (dotted) and lowpass (dash-dotted line) results.

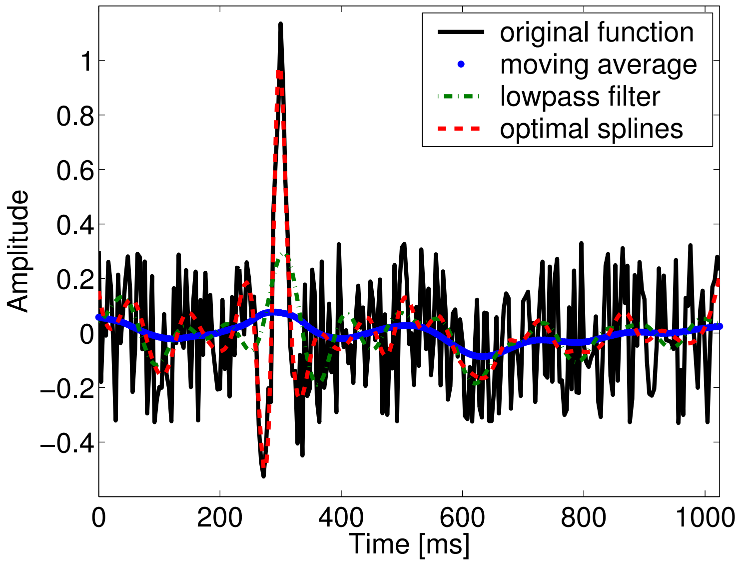

Figure 14: Comparison of results of optimal cubic splines (dashed) for frequencies up to 78.5 Hz (period 80 ms) with corresponding moving average (dotted) and lowpass (dash-dotted line) results.

sults while both moving average and lowpass filtering fail occasionally. The reason is optimal cubic splines preserve not only the desired low-frequency band but also important high-frequency characteristics.

We stress that smoothing is a method to reduce high-frequency noise. Neither of the methods compared in this work, be it optimal cubic splines, moving average, or lowpass filtering, can be used to eliminate low-frequency noise.

ACKNOWLEDGMENTS

This work was kindly supported by the Brazilian research agencies CNPq and FAPESP, as well as the sponsors of the *Wave Inversion Technology (WIT) Consortium*.

REFERENCES

Biloti, R. (2002). *Tomographic Inversion of the velocity model using multiparametric traveltimes*. PhD thesis, State University of Campinas, Brazil. In Portuguese.

Biloti, R., Santos, L. T., and Tygel, M. (2002). Multiparametric traveltime inversion. *Studia Geophysica et Geodaetica*, pages 177–192.

Biloti, R., Santos, L. T., and Tygel, M. (2003). Automatic smoothing by optimal splines. *Brazilian Journal of Geophysics*, 21(2):173–178.

Dennis, Jr., J. E. and Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial and Applied Mathematics.

Ueberhuber, C. W. (1997). *Numerical Computation*, volume 2. Springer.