A TWO CHANNEL, BLOCK-ADAPTIVE AUDIO SEPARATION TECHNIQUE BASED UPON TIME-FREQUENCY INFORMATION

Daniel Smith, Jason Lukasiak and Ian S Burnett.

Whisper Laboratories, SECTE
University of Wollongong, NSW, Australia.
dsmith@titr.uow.edu.au, {jasonl, i.burnett}@elec.uow.edu.au

ABSTRACT

TIFROM [1, 2] is a two channel separation technique, which is well suited to separating audio signals, and in particular, dependent signals that fall outside the scope of conventional BSS applications [1]. One problem with TIFROM however, is degraded performance due to inconsistent estimation of the mixing system. To reduce these inconsistencies, we present a modified algorithm that incorporates k-means clustering [3] and normalised variance, improving upon TIFROM estimation results significantly. To improve TIFROM data efficiency we also include a weighting (running average) function for mixing column estimates. This transforms our modified algorithm into a block based adaptive algorithm with the ability to track a slowly time-varying mixture.

1. INTRODUCTION

Blind Signal Separation (BSS) techniques attempt to separate a set of unknown signals s, from observations x, which are a result of the signals s being mixed by system A. In addition to the observations x, assumptions need to be made with regards to the statistical properties of the signals in order for blind separation to be successful.

Conventional BSS techniques make the assumption that signals are statistically independent [4]; this assumption is invalid for audio signals that possess dependencies (i.e. singer and musical instruments playing in the same tone [1]). BSS techniques have thus utilised an alternate mechanism, their separation is based upon the sparsity [5, 6, 7] or equivalently disjoint orthogonality [8] of the signals (i.e. signals that do not overlap in the transform domain of the mixture). Although this approach encompasses dependent audio signals, it requires a high degree of sparsity within the chosen transform domain [9]; an assumption that is often only weakly satisfied by audio signals. TIFROM is a two channel BSS technique that reduces the influence of such strong assumptions of sparsity [1, 2], by assuming signals are sparse across only localised time frequency regions (as opposed to the entire signal representations used in [5, 6, 7, 8]).

Although TIFROM's assumptions are well suited to audio signals, and accommodate the separation of dependent signals [1], the performance of the current TIFROM algorithm can suffer as a result of poor mixing system estimation (discussed in Section 3). In this paper we attempt to modify the TIFROM algorithm to remedy this problem. In particular we incorporate normalised variance measures and k-means clustering [3] into the algorithm and, as a consequence, show improved mixing system estimation. A weighting function for mixing column estimates is also incorporated into our algorithm, transforming TIFROM into a block based adaptive algorithm with improved data efficiency and the potential to track instantaneous time-varying mixtures.

The paper is organised as follows. In Section 2, we introduce the TIFROM algorithm. We then discuss specific limitations associated with TIFROM mixing system estimation and ways to remedy them in Section 3. In Section 4, modifications to TIFROM are detailed and experiments are conducted to compare the performance of our modified algorithm and TIFROM, in terms of mixing system estimation quality and data efficiency. The conclusions of our work are presented in Section 5.

2. TIFROM APPROACH TO SEPARATION [1, 2]

Our mixture consists of two linear instantaneous mixed observations of two real signals $s_j(n)$:

$$\begin{aligned} x_1(n) &= a_{11}s_1(n) + a_{12}s_2(n) \\ x_2(n) &= a_{21}s_1(n) + a_{22}s_2(n) \end{aligned} \tag{1}$$

where a_{ij} are the mixing coefficients of the mixing system A.

TIFROM employs a simple approach to separation in this mixed system, by estimating each signal's mixing columns $C_1 = \frac{a11}{a21}$ and $C_2 = \frac{a12}{a22}$ from the time-frequency (TF) information of the observations $x_j(n)$.

A short-time Fourier transform (STFT) of the mixed observations $x_j(n)$ is computed. $X_j(m,k)$ represents the TF window of the mixed observations, centered on short time window m and frequency k. TIFROM computes the ratio $\xi(m,k)$ between the corresponding TF windows of each mixed observation such that:

$$\xi(m,k) = \frac{X_1(m,k)}{X_2(m,k)} = \frac{a_{11}s_1(m,k)+a_{12}s_2(m,k)}{a_{21}s_1(m,k)+a_{22}s_2(m,k)} \tag{2}$$

From (2) we see that for only a single source $s_j(n)$ present in a TF window (m,k), $\xi(m,k)$ will correspond to the source's mixing column $C_j = \frac{a_{1j}}{a_{2j}}$. TIFROM exploits this property of signal sparsity under the assumptions:

1. For each signal s_j, there exist at least some adjacent TF windows (m,k) where either s_j occurs alone or where $s_j \gg s_i$. This will be referred to as TIFROM's sparsity assumption.
2. Signals should be time-varying across a set of adjacent TF windows (m,k).

Under these two conditions, for only one signal present (or dominant) across adjacent TF windows (m,k), the ratio $\xi(m,k)$ is approximately constant. However, if there are two signals in the window set (m,k), $\xi(m,k)$ will vary across those windows.

As a consequence of this property, TIFROM uses the variance across a series of TF windows to estimate each mixing column. Within a series (Υ_u,k) of time-adjacent windows of the ratio $\xi(m,k)$, the mean $me(\Upsilon_u,k)$ and variance $var(\Upsilon_u,k)$ are computed. All series are then searched for the lowest value of $var(\Upsilon_u,k)$. The $me(\Upsilon_u,k)_1$ corresponding to the $var(\Upsilon_u,k)_{min}$ is chosen as a mixing estimate C_{je}. The second mixing column estimate C_{ie} is found to be $me(\Upsilon_u,k)_2$ corresponding to the $var(\Upsilon_u,k)_{min}$ from the set $Q \in |me(\Upsilon_u,k) - me(\Upsilon_u,k)_1| > T$. T is a threshold set to determine the minimum difference between the ratios. Finally the mixing columns are used to estimate the separation matrix A^{-1} as in [1].

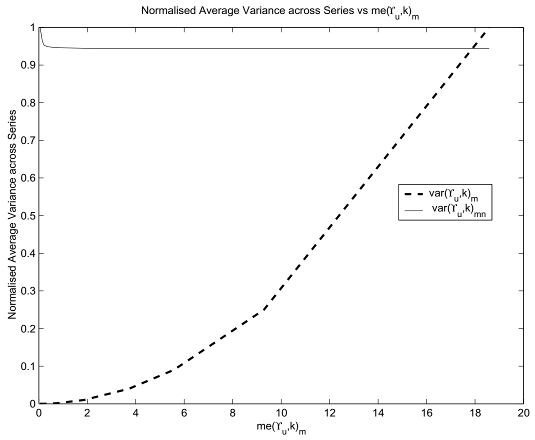


Figure 1: The relationship between $var(\Upsilon_u,k)_m$ and $var(\Upsilon_u,k)_{mn}$ across $me(\Upsilon_u,k)_m$ for a 2 signal mixture.

3. LIMITATIONS OF TIFROM

Even under the condition that signals comply with TIFROM's assumptions, inconsistency in mixing column estimation can be experienced. This is primarily because TIFROM uses a series of minimum variances to determine mixing column estimates. Thus, factors that increase the variance inequality between series of different mixing columns will degrade estimation. Two such factors that create variance inequality are now detailed.

3.1 Variance across series with different sized ratios

Inequality in $var(\Upsilon_u,k)$ is created across series of different sized $me(\Upsilon_u,k)$ ratios. This inequality across series can be demonstrated by altering a mixing column C$_j$ between $\{0.1...10\}$ (with C$_i$ fixed) across all series of a 2 signal mixture. At each increment of C$_j$, we computed the mean across all $var(\Upsilon_u,k)$ and $me(\Upsilon_u,k)$ of the mixture as $var(\Upsilon_u,k)_m$ and $me(\Upsilon_u,k)_m$ respectively. The relationship between $var(\Upsilon_u,k)_m$ and $me(\Upsilon_u,k)_m$, shown as the dotted line in Figure 1, illustrates that the absolute variance monotonically increase with larger ratio size, despite the fact that the signals exhibit identical sparsity relationships for all ratio sizes.

The result of this relationship is the possibility of poor estimation of mixing columns. Given $C_i > C_j$, C_{ie} can be underestimated by series of smaller ratios that possess less variance than the series corresponding to the actual C_i. These smaller ratios, often correspond to weaker estimates of C_j (i.e. C_j ± T), which results in the C_{je} and C_{ie} being estimates of the same mixing column.

We overcome inequality of the $var(\Upsilon_u,k)$ series across different sized $me(\Upsilon_u,k)$ by normalisation of $var(\Upsilon_u,k)$:

$$var(\Upsilon_u,k)_n = \frac{var(\Upsilon_u,k)}{me(\Upsilon_u,k)^2} \tag{3}$$

To illustrate that $var(\Upsilon_u,k)_n$ possesses uniformity across $me(\Upsilon_u,k)$, the experiment conducted for the $var(\Upsilon_u,k)$ series, is now replicated for $var(\Upsilon_u,k)_n$ (using the same mixture series and thus equivalent levels of sparsity). If we compare the solid line of the normalised mean variance ($var(\Upsilon_u,k)_{mn}$) to the dotted line of $var(\Upsilon_u,k)_m$, it is evident that normalistion produces variance uniformity across $me(\Upsilon_u,k)$ and thus resolves estimation problems between mixing columns due to difference in ratio size.

3.2 Variance across series with varying degrees of compliance to TIFROM's sparsity assumption

Although normalisation creates variance uniformity across different sized ratio series with equivalent levels of sparsity, inequality in the $var(\Upsilon_u,k)$ of mixing columns is also created by the different degrees to which signals satisfy the sparsity assumption of TIFROM. If the TF series estimating C_j is sparser than the TF series estimating C_i, the $var(\Upsilon_u,k)$ of the C_j series is less than the C_i series. Thus C_i is estimated incorrectly if the TF series differ in their compliance of TIFROM's sparsity assumption. The result is that the C_i series estimates possess greater variance than series that are weak estimates of C_j i.e. in the vicinity of $C_j \pm T$.

Estimation of the second mixing column therefore relies upon choosing a series that possesses low variance, but more importantly, a series with a $me(\Upsilon_u,k)$ that is distinct from the first mixing column estimate. K-means clustering allows series to be partitioned into distinct clusters in $me(\Upsilon_u,k)$ space. Differentiation of series through clustering ensures that unique $me(\Upsilon_u,k)$ ratios are estimated for each mixing column. Implementation details of the k-means clustering will be discussed in Section 4.

4. NEW TIFROM SYSTEM

To improve mixing system estimation and resolve the estimation problems discussed in Section 3.1 and 3.2, we modify the TIFROM algorithm. This new architecture (TIFmod) consists of the original TIFROM algorithm, but with additional steps after calculation of the variance series:

1. The variance series is normalised to $var(\Upsilon_u,k)_n$ as in (3).
2. 2-D k-means clustering is conducted on series belonging to S, where $S \, \varepsilon \, var(\Upsilon_u,k)_n < v_{max}$. Series of S are partitioned into 2 clusters P_i with respect to $me(\Upsilon_u,k)$ and $var(\Upsilon_u,k)_n$ space.
3. The variance centroid $var(P_i)$ of both clusters are examined. If $var(P_i) < v_{min}$, then the centroid mean ($me(P_i)$) is chosen as a mixing system estimate. Otherwise ratios are estimated as in the original algorithm, but under the condition that if the first ratio is estimated by a series belonging to cluster P_i, all series belonging to P_i are excluded from the second ratio estimation.

4.1 TIFmod Results

To verify that TIFmod displays improvement over TIFROM for mixing matrix estimation, we apply the algorithms to 6 audio mixtures that are 2.5s in length and sampled at 8000 Hz. Each pair of audio signals were mixed by 24 different stationary mixing models. All mixtures were passed to the algorithms in data blocks sized:

$$\begin{aligned} blocksize = overlap * framesize * (fps + 1) \\ + overlap * framesize * (seriesnum - 1) \end{aligned} \tag{4}$$

where the $framesize$ is 20ms, $overlap$ is 0.5 of a frame, number of adjacent frames per series (fps) is 6 and number of series in each block of data ($seriesnum$) ranges from 1 to 181. Mixing ratio estimation and data block update are every 40ms. Threshold (T) is not given in [1, 2], but our empirical results indicate that a suitable value for T is 15% of the first ratio. For TIFmod, the v_{min} and v_{max} heuristics were obtained from an extensive empirical study of the variance of ratio estimates, with v_{min} being set to 0.0016 and v_{max} being set to 0.0123.

To measure the quality of each algorithms mixing ratio estimation we used the Interference Measurement (IM) as a criteria:

$$IM = \frac{1}{2} \sum_{j=1}^{2} (p_j^T * p_j - max(p_j)^2)^{\frac{1}{2}} \tag{5}$$

where p is the product of the separation and mixing matrix, and p_j is a column of p. IM is a performance measure of mixing system identification, measuring p's average distance from a scaled, permuted diagonal matrix corresponding to perfect estimation of the mixing channel. It is related to the measure used in [10].

Figure 2 shows TIFROM and TIFmod average average log distortion ($10 * log_{10} IM$) across 6 pairs of audio mixtures and 24 stationary mixtures, with respect to *seriesnum*. Although TIFmod outperforms TIFROM for mixing column estimation across all *seriesnum*, it is evident that TIFmod has a greater advantage over TIFROM for larger *seriesnum*. This is because k-means clustering in TIFmod requires a larger sample size to produce significant estimation improvement. As signals are more likely to have TF regions of high sparsity with larger sample size, clustering will be conducted on $me(\Upsilon_u, k)$ ratios that are accurate estimates of all mixing columns. With less data however, the signals are likely to have reduced or no compliance with TIFROM's sparsity assumption, and thus clustering will be conducted on $me(\Upsilon_u, k)$ ratios that are poorer estimates of mixing columns.

4.2 Modification of TIFROM into block-based adaptive algorithm

Figure 2 indicates that TIFmod's ratio estimation in Section 4 is poor for smaller data blocks (*seriesnum* < 10). This problem arises from the fact that sparsity is less likely to occur in smaller block sizes and thus inaccurate mixing matrix estimates will result. We propose a weighting (running average) function that uses the confidence of the estimate to determine the update weight for the mixing columns. As we can measure our confidence in the accuracy of the $C_{je}(t)$ estimate from its variance (v_{je}), the weighting function we utilise is:

$$\begin{aligned} C_{jwe}(t) &= C_{je}(t) \quad if\ v_{je} \leq v_{min} \qquad (6) \\ C_{jwe}(t) &= (1 - Vw) * C_{je}(t-1) + Vw * C_{je}(t) \\ & if\ v_{min} < v_{je} < v_{max} \\ C_{jwe}(t) &= C_{je}(t-1) \quad if\ v_{je} \geq v_{max} \end{aligned}$$

where $Vw = \frac{v_{max} - v_{je}}{v_{max} - v_{min}}$. Poor estimates of $C_{je}(t)$ are thus penalised or excluded in $C_{jwe}(t)$. The modifications in Section 4 and this weighting function are combined to form the block-based adaptive algorithm (adTIFmod) evaluated in the following section.

4.3 adTIFmod Results

The experiment from Section (4.1) was repeated for the adaptive algorithm (adTIFmod). Figure 2 shows that adTIFmod achieves improvement upon TIFmod for mixing column estimation for *seriesnum* < 81, and most importantly, significantly better performance as *seriesnum* decreases. Therefore, the adTIFmod algorithm offers superior estimation performance to TIFROM across all series sizes in Figure 2, and in particular, a higher data efficiency i.e. a much higher quality of ratio estimation with smaller block sizes. This highlights the potential that adTIFmod has to estimate time varying mixtures in real time.

A second experiment was conducted to demonstrate that the adTIFmod algorithm has the ability to track a slowly time-varying mixture and offer superior performance to TIFROM. Both algorithms were applied to five different pairs of audio signals, 10s in length and sampled at 8kHz. Each pair of audio signals were mixed by a time-varying system A1 with transitions every 1.66s. The actual track of the mixing columns, corresponding to A1, are shown as the solid lines in Figure 4a and 4b. The T, v_{max} and v_{min} heuristics, *framesize*, *overlap*, *seriesnum* range and *fps* were the same as in the previous experiment. Ratio estimates and data blocks were again updated every 40ms.

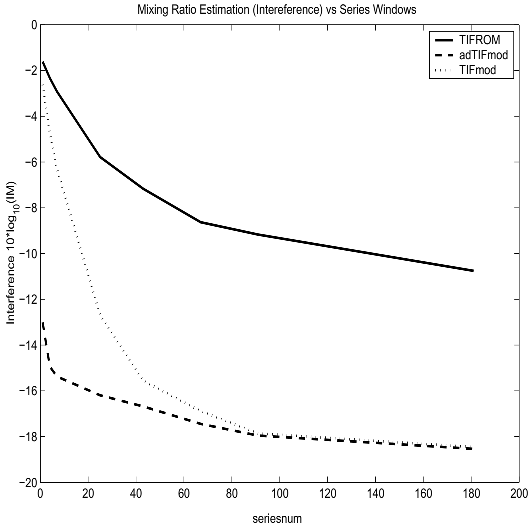


Figure 2: The average ($10 * log_{10} IM$) interference vs *seriesnum* for TIFROM, TIFmod and adTIFmod, across 6 audio mixtures and 24 stationary mixing matrix.

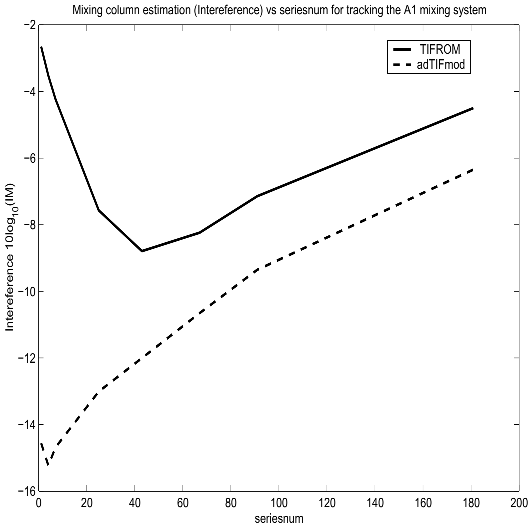
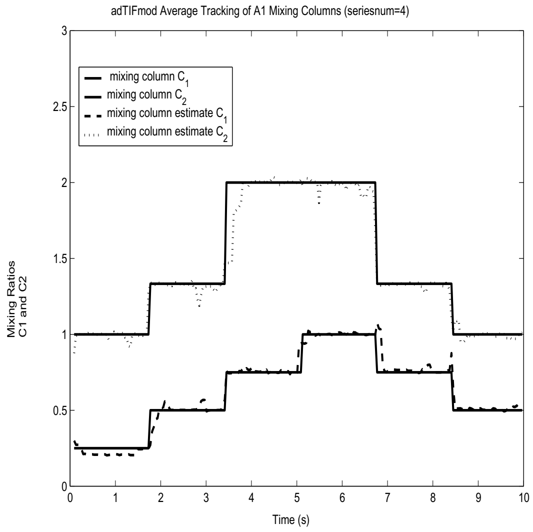


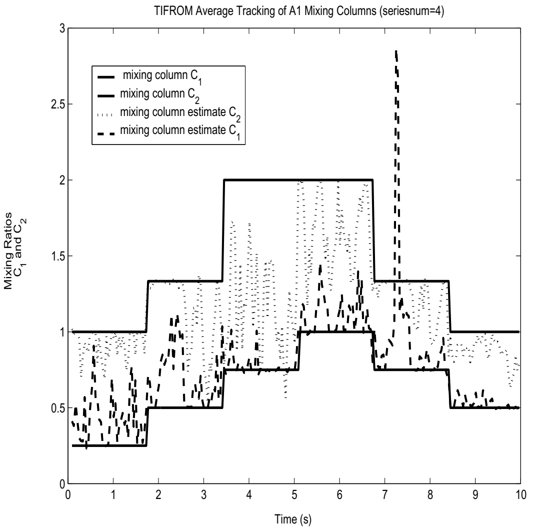
Figure 3: The average interference ($10 * log_{10} IM$) vs *seriesnum* for TIFROM and adTIFmod, across 5 audio mixtures and a time-varying mixture A1.

Figure 3 shows TIFROM's and adTIFmod's average log distortion according to (5) across five audio mixtures for the mixing system A1 in relation to *seriesnum*. It indicates that the adTIFmod algorithm outperforms TIFROM in tracking the slowly time-varying mixture A1 across the range of *seriesnum*. In particular, adTIFmod has a greater performance advantage over TIFROM for tracking with a smaller number of series (*seriesnum* < 20) than a larger number of series (*seriesnum* > 40). We attribute this to:

1. TIFROM and adTIFmod tracking A1 poorly for a larger number of series (*seriesnum* > 40). For a larger number of series, after mixing system A1 changes, outdated buffered series will have



4a. adTIFmod



4b. TIFROM

Figure 4: adTIFmod and TIFROM average track of mixing system A1 for 5 audio mixtures.

influence in the data block for a greater number of estimates, resulting in TIFROM and adTIFmod being slower to update the A1 mixing columns.

2. TIFROM continuing to track A1 poorly while adTIFmod shows strong tracking performance for a smaller number of series (*seriesnum* $<$ 20). TIFROM's mixing column estimation is poor for smaller amounts of data (as detailed in Section 4.1), however the improved data efficiency of adTIFmod ensures A1 is successfully tracked in real time.

Figure 4 illustrates adTIFmod's superior performance for tracking a time-varying system A1 with a small number of series (*seriesnum*=4). The adTIFmod and TIFROM algorithm's average tracking estimates of the A1 columns are shown as the dotted lines of Figure 4a and Figure 4b respectively. Whilst adTIFmod estimates trace the actual A1 mixing columns for their duration, the average TIFROM estimates are more oscillatory, and therefore track the A1 mixing columns poorly. Mixing columns are often estimated too close together in TIFROM, thus corresponding to estimates of the same mixing column. This is a consequence of the difference in the size of series ratios (see Section 3.1) and the difference in the signal's compliance with TIFROM's sparsity assumption (see Section 3.2). The adTIFmod algorithm overcomes these problems, through k-means clustering, or in the presence of a poor estimate, a weighting function that reduces its influence in the current block.

5. CONCLUSION

The TIFROM framework was modified to resolve inconsistencies regarding mixing column estimation. As a consequence of these modifications, our algorithm was shown to offer significant improvements in estimation performance and data efficiency compared to the original algorithm. These improvements enabled us to demonstrate that our modified architecture could operate in real time, tracking a slowly time-varying instantaneous mixture. The underlying assumptions of TIFROM, that appear well suited to audio signals and accommodate dependent signals, combined with the improved performance of the algorithm, suggest that a robust algorithm for separation of signals in the audio domain has been developed.

REFERENCES

[1] F.Abrard and Y.Deville, "Blind separation of dependent signals using the 'time-frequency ratio of mixtures' approach," in *Proc.7th International Symposium on Signal Processing and its Applications (ISSPA 2003)*, 2003.

[2] F.Abrard, Y.Deville, and P.White, "From blind source separation to blind source cancellation in the undetermined case: A new approach based on time-frequency analysis," in *Proc.3rd International Conference on Independent Component Analysis and Blind Source Separation (ICA2001)*, 2001, pp. 734–739.

[3] A.Gersho and R.Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.

[4] A.Hyvarinen, J.Karhunen, and E.Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[5] P.Bofill and M.Zibulevsky, "Blind source separation of more sources than mixtures using sparsity of their short-time fourier transform," in *Proc.2nd International Workshop on Independent Component Analysis and Blind Signal Seperation (ICA)*, 2000, pp. 87–92.

[6] T.-W. Lee, M.Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Processing Letters*, vol. 6, no. 4, pp. 87–90, 1999.

[7] C.Choi, "Real time binaural blind source separation," in *Proc.4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003)*, 2003, pp. 567–572.

[8] A.Jourjine, S.Rickard, and O.Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proc.IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP2000)*, 2000, vol. 5, pp. 2985–2988.

[9] S.Rickard and O.Yilmaz, "On the w-disjoint orthogonality of speech," in *Proc.IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2002)*, 2002.

[10] A.Cichoki and S.Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, John Wiley & Sons, 2002.