ELSEVIER

Neurocomputing 49 (2002) 79–86

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Blind deconvolution using temporal predictability

James V. Stone*

Psychology Department, Sheffield University, Sheffield S10 2UR, UK

Received 10 March 2001; accepted 22 October 2001

Abstract

A measure of *temporal predictability* is defined, and used for blind deconvolution of sound signals. The method is based on the observation that physical environments act as smoothing filters, and therefore increase the predictability of signals. These smoothing effects can be reversed by a deconvolution filter which minimises a measure of temporal predictability. This filter is obtained as the closed form solution to an eigenvalue problem which scales as $O(N^3)$, where N is the number of filter coefficients. It is proven that the method minimises mutual information in a gaussian channel with feedback. Results are presented for sound signals. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Deconvolution; Blind source separation; Unsupervised learning

1. Introduction

A signal can be altered by the effects of the environment in which it is measured. For example, a room typically causes speech signals to be heard with attenuated high frequencies and with multiple echoes. Thus, the room effectively acts as a filter on the signal. Such effects can be reversed using appropriate deconvolution filters. Here, it is shown how deconvolution filters can be obtained using a measure of temporal predictability. This measure has previously been used for blind source separation (BSS) [12], and unsupervised learning [10,11].

Consider a filtered version x of a signal s such that $x = \alpha * s$, where α is a filter $\alpha = [\alpha_1, \alpha_2, \ldots]$. If a deconvolution filter $\beta = [\beta_1, \beta_2, \ldots]$ exists for α then $y = s$ can be recovered as $y = \beta * x$. In practice, the correct deconvolution filter cannot be obtained exactly, so that $y \approx s$.

* Tel.: +0114-222-6522; fax: +0114-276-6515.
E-mail address: j.v.stone@sheffield.ac.uk (J.V. Stone).
URL: http://www.shef.ac.uk/~pc1jvs

0925-2312/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0925-2312(02)00520-9

Conventional blind deconvolution methods (e.g. minimum entropy methods [13,6]) adjust the coefficients of β in order to maximise a normalised cumulant (e.g. kurtosis) of a deconvolved signal $y = \beta * x$. In this respect, such methods are similar to the projection pursuit source separation methods [7], which maximise the kurtosis of extracted signals. Both types of method implicitly assume that the original signal s has a super-gaussian (e.g. highly kurtotic) probability density function (pdf), and the coefficients of a deconvolving filter β are adjusted to provide a deconvolved signal $y \approx s$ with a pdf consistent with this assumption. These methods work because the central limit theorem ensures that any smoothed signal $x = \alpha * s$ tends to have a gaussian pdf. In the particular case that s has a super-gaussian pdf, a filter which makes the pdf of $y = \beta * x$ super-gaussian tends to recover the signal $y \approx s$. In contrast to the method presented here, these methods assume that contiguous values of s are independent.

1.1. A strategy for deconvolution

The strategy used to estimate the deconvolution filter β will be described for a simple example. Consider a low-pass infinite impulse response (IIR) filter α, in which each filter coefficient α_n is defined $\alpha_n = \lambda_{\mathrm{D}}^n$, where $(0 \leqslant \lambda_{\mathrm{D}} \leqslant 1)$ defines a half-life h_{D} for the filter ($\lambda_{\mathrm{D}} = 2^{-1/h_{\mathrm{D}}}$). The effects of such IIR filters can be reversed using a finite impulse response (FIR) deconvolution filter $\beta = [1, -\lambda_{\mathrm{D}}]$, with $N = 2$ coefficients. A low-pass filter, such as α, effectively smooths a signal s to yield $x = a * s$. This smoothing operation inevitably makes x *more predictable* [1] than s. In order to emphasize this, consider the limit ($\lambda_{\mathrm{D}} \to 1$). In this limit, x has a constant value, and is therefore very predictable. It follows that x becomes increasingly predictable as λ_{D} increases from zero (i.e. no smoothing) to unity, and that a deconvolution filter β which makes x *less predictable* may be able to recover s. This observation is the key to the proposed method: β is defined as that filter which minimises a measure of predictability of $y = \beta * x$.

2. The method

The method relies on the same mathematics as previous work on blind source separation [12]. In order to place the current work in the context of that previous work, the convolution operations defined above will be reformulated in terms of mathematically equivalent vector–matrix operations.

Consider a deconvolution filter $\beta = [\beta_1, \beta_2]$ with $N = 2$ coefficients, such that $y = \beta * x$. This can be re-written in vector–matrix notation $y = \beta \mathbf{x}$ by defining an N-dimensional vector variable $\mathbf{x}$ as $\mathbf{x} = \{x \,|\, xz^{-1}\}^{\mathrm{t}}$, where each row of $\mathbf{x}$ is a time-shifted version of x, and the superscript t is the transpose operator. The shift operator z^{-k} is defined by $x_t z^{-k} = x_{t-k}$. Now the convolution $y = \beta * x$ and the vector–matrix multiplication $y = \beta \mathbf{x}$ are exactly equivalent.

[1] A formal definition of predictability is given below.

2.1. Measuring signal predictability

The definition of signal predictability F used here is

$$F(\beta,\mathbf{x}) = \log \frac{V(\beta,\mathbf{x})}{U(\beta,\mathbf{x})} = \log \frac{\sum_{\tau}^{\tau_{\max}} (\bar{y}_\tau - y_\tau)^2}{\sum_{\tau}^{\tau_{\max}} (\tilde{y}_\tau - y_\tau)^2}, \tag{1}$$

where $y_\tau = \beta \mathbf{x}_\tau$ is the value of the signal y at time τ, and $\mathbf{x}_\tau$ is a vector of M signal values at times $\mathbf{x}_\tau = [x_\tau, x_{\tau-1}, x_{\tau-2}, \ldots, x_{\tau-(N-1)}]^{\mathrm{t}}$.

The term U reflects the extent to which y_τ is predicted by a short-term 'moving average' $\tilde{y}_\tau$ of values in y. In contrast, the term V is a measure of the overall variability in y, as measured by the extent to which y_τ is predicted by a long-term 'moving average' $\bar{y}_\tau$ of values in y.

The predicted values $\tilde{y}_\tau$ and $\bar{y}_\tau$ of y_τ are both exponentially weighted moving averages of signal values measured up to time $(\tau - 1)$. This exponential weighting ensures that recent values have a larger weighting than those in the distant past

$$\tilde{y}_\tau = \frac{1}{P_{\mathrm{S}}} \sum_{i=1}^{\infty} \lambda_{\mathrm{S}}^i y_{(\tau-i)}, \qquad \bar{y}_\tau = \frac{1}{P_{\mathrm{L}}} \sum_{i=1}^{\infty} \lambda_{\mathrm{L}}^i y_{(\tau-i)}, \tag{2}$$

where $(0 \leqslant \lambda_{\mathrm{S}} \leqslant 1)$ and $(0 \leqslant \lambda_{\mathrm{L}} \leqslant 1)$, and P_{S} and P_{L} are constants

$$P_{\mathrm{S}} = \sum_{i=1}^{\infty} \lambda_{\mathrm{S}}^i = \frac{\lambda_{\mathrm{S}}}{1-\lambda_{\mathrm{S}}}, \qquad P_{\mathrm{L}} = \sum_{i=1}^{\infty} \lambda_{\mathrm{L}}^i = \frac{\lambda_{\mathrm{L}}}{1-\lambda_{\mathrm{L}}}. \tag{3}$$

The half-life h_{L} of λ_{L} is much longer (typically 100 times longer) than the corresponding half-life h_{S} of h_{S}. The relation between a half-life h and the parameter λ is defined as $\lambda = 2^{-1/h}$.

Note that, in the limits $h_{\mathrm{S}} \to 0$ and $h_{\mathrm{L}} \to \infty$, $\tilde{y}_\tau = y_{\tau-1}$ and $\bar{y}$ is equal to the mean of y, so that Eq. (1) simplifies to

$$F(\beta,\mathbf{x}) = \log \frac{\sum_\tau (\bar{y} - y_\tau)^2}{\sum_\tau (y_{\tau-1} - y_\tau)^2}. \tag{4}$$

Simulations using this simplified expression yield results similar to those reported in this paper, but the resultant method may be less robust to noise.

If we were to minimise only V, then this would result in a low variance signal with no constraints on its temporal structure. In contrast, maximising only U would result in a signal with arbitrarily high amplitude. In both cases, trivial solutions would be obtained for β because V can be minimised by setting the norm of β to be zero, and U can be maximised by setting the norm of β to be large. In contrast, the ratio V/U can be minimised only if two constraints are both satisfied: (1) y has a small non-zero range (i.e. low variance), and (2) the values in y change 'quickly' over time (relative to h_{S}). Note also that the value of F is independent of the norm of β, so that only changes in the *direction* of β affect the value of F.

2.2. Deconvolution using temporal predictability

Given that $y = \beta\mathbf{x}$, Eq. (1) can be re-written as

$$F = \log \frac{\beta\bar{C}\beta^{\mathrm{t}}}{\beta\tilde{C}\beta^{\mathrm{t}}}, \tag{5}$$

where $\bar{C}$ is an $M \times M$ matrix of long-term covariances between time-shifted signals, and $\tilde{C}$ is a corresponding matrix of short-term covariances. The long-term covariance $\bar{C}_{ij}$ and the short-term covariance $\tilde{C}_{ij}$ between the ith and jth time-shifted signals are defined as

$$\tilde{C}_{ij} = \sum_{\tau} (x_{i\tau} - \tilde{x}_{i\tau})(x_{j\tau} - \tilde{x}_{j\tau}), \quad \bar{C}_{ij} = \sum_{\tau} (x_{i\tau} - \bar{x}_{i\tau})(x_{j\tau} - \bar{x}_{j\tau}). \tag{6}$$

$\tilde{C}$ and $\bar{C}$ need only be computed once for a given set of time-shifted signals, and the terms $(x_{i\tau} - \bar{x}_{i\tau})$ and $(x_{i\tau} - \tilde{x}_{i\tau})$ can be pre-computed using fast convolution operations.

Note that the function F is a ratio of quadratic forms. Therefore, F has exactly one global maximum and exactly one global minimum, with all other critical points being saddle points [3].

Gradient descent on F with respect to β could be used to minimise F, thereby minimising the predictability of y. The derivative of F with respect to β is

$$\nabla_{\beta} F = \frac{2\beta}{V}\,\bar{C} - \frac{2\beta}{U}\,\tilde{C}. \tag{7}$$

One optimisation procedure (not used here) would consist of iteratively updating β until a minimum of F is located: $\beta = \beta - \eta\nabla_{\beta}F$, where η is a small constant (typically, $\eta = 0.001$).

Note that Eq. (5) has the same form as Fisher's linear discriminant [2, p. 108]. In this case, the between-class variance of Fisher's discriminant corresponds to $\bar{C}$, the within-class variance corresponds to $\tilde{C}$, and the projection operator corresponds to β.

2.3. Deconvolution as an Eigenproblem

The gradient of F is zero at a solution where, from Eq. (7),

$$\beta\bar{C} = \frac{V}{U}\,\beta\tilde{C}. \tag{8}$$

Critical points in F correspond to values of β that satisfy Eq. (8), which has the form of a generalised eigenproblem [3]. Solutions for β can therefore be obtained as eigenvectors of the matrix $(\tilde{C}^{-1}\bar{C})$, with corresponding eigenvalues $\gamma = V/U$. As noted above, the first and last such eigenvectors define a maximum and a minimum in F, respectively, and each of the remaining eigenvectors define saddle points.

Note that eigenproblems have closed form solutions, and scaling characteristics of $O(N^3)$, where N is the number of deconvolution filter coefficients in β. The vector β can be obtained using a generalised eigenvalue routine. Results presented in this paper were obtained using the MatLab eigenvalue function $W = (\beta^1, \beta^2, \ldots, \beta^N) = eig(\bar{C}, \tilde{C})$. Each column if the $N \times N$ matrix W is a eigenvector β. The deconvolved signal can

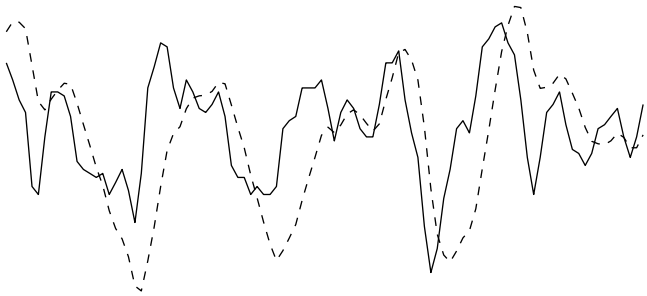

Fig. 1. Original signal s (—) and convolved signal x (- - -), plotted for 100 time steps. The correlation between s and x is $r = 0.430$.

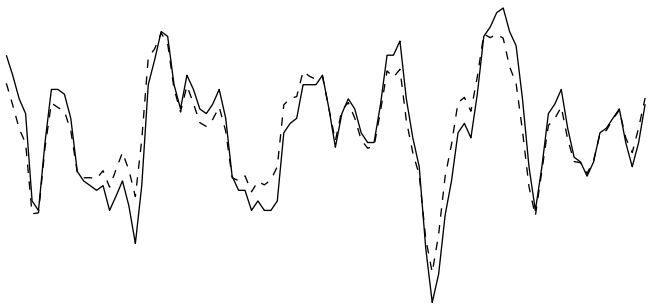

Fig. 2. Original signal s (—) and deconvolved signal $y = \beta x$ (- - -), plotted for 100 time steps. The correlation between the deconvolved signal $y = \beta * x$ and s is $r = 0.980$.

then be recovered by using the eigenvector β^N in W with the smallest eigenvalue to obtain $s \approx y = \beta^N \mathbf{x}$.

3. Results

The method is demonstrated using 5000 samples of signal s which is the sound of a choir singing, sampled at 8192 Hz. The signal s was convolved with a low-pass filter with $n = 32$ coefficients $\alpha = [1, \lambda_{\mathrm{D}}^1, \lambda_{\mathrm{D}}^2, \ldots, \lambda_{\mathrm{D}}^{31}]$, where the half-life $h_{\mathrm{D}} = 8$ samples. As expected, the convolved signal $x = \alpha * s$ sounds very muffled. The convolved signal x was used as input to the method.

In this case, we know that a deconvolution filter with two coefficients is sufficient to recover s from x. Accordingly, the input data $\mathbf{x}$ was defined as $\mathbf{x} = \{x \,|\, xz^{-1}\}^{\mathrm{t}}$. Fig. 1 displays a short segment of s and x, and the correlation between s and x is $r = 0.430$. After applying the eigenvalue method described above to obtain the deconvolution filter β, the correlation between the deconvolved signal $y = \beta\mathbf{x} = \beta * x$ and s is $r = 0.980$. From Fig. 2, it can be seen that there is good agreement between s and the recovered signal y.

Given the known smoothing filter α, the signal x can be approximately deconvolved with a filter $\beta = [1, -\lambda_D] = [1, -0.917]$. The estimate of β found by the method is $[1, -0.999]$. Consequently, the resultant deconvolved signal y is over-whitened, relative to s. However, the deconvolved signal y is perceptually indistinguishable from the original s.

4. An information-theoretic interpretation

Insight into the method can be gained from two different, but related, proofs.

First, it was proved in [10,11] that maximising F maximises the power of low-(non-zero) frequency components passed by β. It follows that minimising F, as in the method described here, provides a filter β which minimises the power of low-frequency components. This tends to differentially amplify high-frequency components, which has the effect of deconvolving smoothed signals.

Second, it is proved here that the method provides a filter which minimises the mutual information between the input and output of a gaussian channel with feedback.

Consider a gaussian channel with feedback, such that $\tilde{y}$ is the channel input and $y = \tilde{y} + n$ is the channel output, where n is independent and identically distributed (iid) gaussian noise. The channel feedback implies that each channel input $\tilde{y}_t$ is a linear combination of previous values of n, and therefore that $\tilde{y}$ and n are not independent [4, p. 260]. The mutual information between $y = (\tilde{y} + n)$ and $\tilde{y}$ is $I(y; \tilde{y}) = H(y) - H(y|\tilde{y})$. If n is iid gaussian then this evaluates to

$$I(\tilde{y} + n; \tilde{y}) = \frac{1}{2} \log \frac{Var(\tilde{y} + n)}{Var(n)}, \tag{9}$$

where $E[\cdot]$ denotes expected value and Var denotes variance [4]. A special case of Eq. (9) obtains for the conventional case of a gaussian channel without feedback, where the input y_{ip} and noise n are independent,

$$I(y_{ip} + n; y_{ip}) = \frac{1}{2} \log \frac{Var(y_{ip}) + Var(n)}{Var(n)} \tag{10}$$

with equality between Eqs. (9) and (10) if $\tilde{y}$ and n are independent.

In order to relate this with the definition of predictability F given in Eq. (1) (reproduced here),

$$F = \log \frac{\sum_{\tau}^{\tau_{max}} (\bar{y}_\tau - y_\tau)^2}{\sum_{\tau}^{\tau_{max}} (\tilde{y}_\tau - y_\tau)^2}, \tag{11}$$

we can re-write Eq. (9) as

$$I(y; \tilde{y}) = \frac{1}{2} \log \frac{Var(y)}{Var(y - \tilde{y})}. \tag{12}$$

Ignoring constants $(1/\tau_{max})$, the numerators of (11) and (12) are equivalent in the limit $h_L \to \infty$, because then $\bar{y}$ is the mean of y (indeed, results are not noticeably affected if $\bar{y}$ is defined as the mean of y). Additionally, if the noise term $n = (y - \tilde{y})$

has zero mean then the denominators of Eqs. (11) and (12) are equivalent. Given these considerations and ignoring constants, Eq. (12) is equivalent to the measure of predictability defined in Eq. (11). Thus, F is the mutual information between a signal y and $\tilde{y}$, where $\tilde{y}$ is the input to a gaussian channel with feedback, and y is its output. The method proposed here, therefore, seeks that filter β which minimises the mutual information between the input and output of a gaussian channel with feedback.

5. Discussion

It has been demonstrated that a method, previously used for unsupervised learning [10,11] and BSS [12], can also be used for deconvolution. Whereas BSS was obtained by maximising a measure of predictability of recovered signals, deconvolution is obtained by minimising the same measure of predictability.

Whilst the two proofs given above provide some insight into the method, neither of these prove that the method deconvolves signals. However, the fact that the method works has been demonstrated in this paper, and the above proofs are an indication that the method is well-founded.

The predictions ($\bar{y}_\tau$ and $\tilde{y}_\tau$) of each signal value y_τ are based on a *constant* linear weighted sum of previous values $\{y\}$. Extensions to this method could involve defining predicted values of y_τ as general functions in terms of *variable* linear predictive coefficients, as in [9].

In common with other deconvolution methods (e.g. [1]), the deconvolved signal recovered by the method can be over-whitened. For instance, if the filter β has five coefficients in the above example, then the deconvolved signal contains too much power in the high frequencies. As discussed in [1], this type of problem can be reduced by using widely spaced filter taps in β, but it remains a potential source of error.

Deconvolution methods which depend explicitly on extremising a cumulant (e.g. [13,6]) or information-theoretic measures (e.g. [1]) of the pdf of deconvolved signals can be interpreted as maximum likelihood (ML) methods. Whilst ML methods are often successfully applied to non-white signals, they are nevertheless based on the assumption that temporally consecutive signal values are independent [5]. No such assumption is required here,[2] implying that the proposed method should work for non-white signals, as has been demonstrated in Section 3.

The main contribution of the preliminary results presented here is to demonstrate that deconvolution can be achieved by explicitly minimising a simple measure of signal predictability. More generally, minimising predictability may be viewed as maximising a simple form of Kolmogorov complexity [4]. This interpretation has been explored for the problem of blind source separation in [8]. Whilst the results presented here are preliminary, the general strategy of using measures of signal complexity may prove as useful for the general problem of blind deconvolution as it has for BSS.

[2] Although it is assumed that the residual n is iid gaussian (see Section 4).

An intriguing aspect of the proposed method is that it works by extremising a function F, which was originally intended as a model of unsupervised learning in perceptual systems [10,11], and was later used for BSS [12]. Whilst it may be imprudent to draw any general conclusions from this, it is nevertheless tempting to speculate that manipulating the temporal predictability of corrupted signals represents a powerful and generic strategy for recovering underlying signals in both artificial and biological systems.

Acknowledgements

Thanks to S. Isard, N. Hunkin, J. Porrill and two anonymous referees for comments on a previous draft of this paper.

References

[1] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.
[2] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, England, 1996.
[3] M. Borga, Learning Multidimensional Signal Processing, Linkoping University, 1998.
[4] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[5] S.C. Douglas, S. Haykin, Relationships between blind deconvolution and blind source separation, in: S. Haykin (Ed.), Unsupervised Adaptive Filtering, Blind Deconvolution, Vol. II, Wiley, New York, 2000, pp. 113–145.
[6] W. Gray, Variable norm deconvolution, Stanford University, California, April 1978 (http://sepwww.stanford.edu/oldreports/sep14/14_19_abs.html).
[7] P. Huber, Projection pursuit, Ann. Stat. 13 (1985) 435–475.
[8] A. Hyvarinen, Complexity pursuit: separating interesting components from time series, Neural Comput. 13 (2001) 883–898.
[9] B.A. Pearlmutter, L.C. Parra, A context-sensitive generalization of ica, in: International Conference on Neural Information Processing, Hong Kong, 1996 (http://www.cs.unm.edu/~bap/publications.html#journal).
[10] J.V. Stone, Learning perceptually salient visual parameters through spatiotemporal smoothness constraints, Neural Comput. 8 (7) (1996) 1463–1492.
[11] J.V. Stone, Learning perceptually salient visual parameters using spatiotemporal smoothness constraints, in: G. Hinton, T. Sejnowski (Eds.), Unsupervised Learning: Foundations of Neural Computation, MIT Press, London, England, 1999, pp. 71–100.
[12] J.V. Stone, Blind source separation using temporal predictability, Neural Comput. 13 (7) (2001).
[13] R.A. Wiggins, Minimum entropy deconvolution, Geoexploration 16 (1978) 21–35.