ON MINIMUM ENTROPY DECONVOLUTION

David Donoho
Western Geophysical Company

1. INTRODUCTION

Given a time series \( y \) which is a filtered version of a white noise \( x \),

\[ y = f \ast x, \tag{1.1} \]

the deconvolution problem is to find a filter \( b \) which recovers \( x \) from the observed series \( y \):

\[ x = b \ast y. \tag{1.2} \]

Standard second-order (spectrum, correlation) methods can solve this problem when the delay properties of \( f \) are known -- for example, prediction error filtering can be used when \( f \) is known to be a minimum delay filter. Unfortunately, the assumption that one knows the delay characteristics of a filter precisely is very restrictive, particularly for long filters (for a typical filter of length \( p+1 \), the number of filters with distinct delay characteristics but having the same second-order properties is \( 2^p \); see the appendix). Consequently, second-order methods are not able to solve the deconvolution problem in the generality proposed.

†Now at Department of Statistics, Harvard University.
Minimum Entropy Deconvolution (MED) is a technique introduced by Wiggins (1977) for deconvolution without making prior assumptions about the delay characteristics of the filter \( f \). The method may be described in general terms as follows: one has a trial estimate \( \tilde{b} \) of \( b \); the adequacy of \( \tilde{b} \) is evaluated by inspection of \( \tilde{x} = \tilde{b} \ast y \). If \( \tilde{x} \) has a "simple" appearance, then \( \tilde{b} \) is judged to be a good estimate of \( b \).

Figure 1 shows a time series generated according to Eq. (1.1) using a white noise \( x \) with a spiky appearance and a mixed-delay filter \( f \). The result \( y \) was then filtered using \( b^+ \), \( b^- \), and \( b \) -- filters all with the same second-order properties -- corresponding to deconvolution of \( y \) under maximum-, minimum- and (the correct) mixed-delay assumptions about \( f \). It is visually apparent that the correct solution to the deconvolution problem, \( x \), has a "simpler" appearance than either \( x^+ \) or \( x^- \). Note that these three series are all uncorrelated; the differences are apparent even though they have the same second-order statistics.

The eye, using a judgement of simplicity, can identify the correct solution to the deconvolution problem even though correlation/spectrum technologies could not.

Wiggins used this simple observation as the basis for a formal procedure for deconvolution. He found simple functions of the data that reproduced roughly the judgements of simplicity that the eye would make: compare Fig. 1 of Wiggins (1978). If \( O \) is one such objective function a formal estimate for the deconvolution problem is

\[
\hat{b} = \text{maximizer } O(\hat{b} \ast y),
\]

(1.3)
Heuristically, $\hat{b}$ is that filter which gives $\hat{x} = \hat{b} \ast y$ the simplest appearance among all series of the form $\hat{x} = b \ast y$.

Rules of the general form (1.3) have recently attracted the attention of a number of geophysicists. In published work, Wiggins's paper was followed up by Ooe and Ulrych (1979) with an extensive simulation study, while Claerbout (1978), Godfrey (1979) and Gray (1979) developed strong heuristic motivation for application of modified versions of the technique to seismic data. In "industrial" situations, a number of geophysical firms provide MED-type processing of seismic data for their clients.

Since every author and programmer seems to use a different objective function $Q$ (even in this paper some new objectives are proposed), one should regard rule (1.3) as describing a family of methods which will be called in this paper MED-type methods. The "Minimum Entropy" terminology is used because it is now customary in seismology: it could also be justified by some of the formal results stated below.

1.2 Notation

A time series is a doubly infinite sequence $x = (\ldots, x_{-1}, x_0, x_1, \ldots)$. A filter is a time series $f$ with $\sum f^2 < \infty$, and filtering is effected by the (two-sided) convolution operation $*: (f \ast x)_t = \sum u x_{t-u}$. A filter $f$ is invertible if there is a filter $b$ with $b \ast b \ast x = x$.

A white noise $x$ is a time series sampled from a sequence $(X_t)$ of independent and identically distributed copies of the random variable $X$.

In practical situations, one observes only finite-length series of data; the symbol $y^n$ will denote the sampled series $(\ldots, 0, y_1, \ldots, y_n, 0, \ldots)$. Then the practical deconvolution problem is really to obtain an estimate $\hat{b}^n$ of $b$ based on $y^n$. In a well-posed estimation problem there should be fewer parameters than observations, so that $b$ and $\hat{b}^n$ should be of length $p+1 < n$. A more precise version of rule (1.3) is then

$$\hat{b}^n = \text{maximizer } Q(\hat{b} \ast y^n),$$

$$\hat{b} : \text{length}(\hat{b}) = p+1$$

Further details will be given in section 3.
1.3 Objective Functions

Wiggins's original proposal called for the use of rule (1.3) with objective

$$O_2^n(x^n) = \frac{1}{n} \left( \frac{1}{n+2} \right)^2 \left( \frac{1}{n+1} \right)^2 \right)^2.$$  
(1.4)

This objective makes the filter of length 11 with output $y^n = \hat{b}^n * y^n$ having the largest kurtosis attainable by filtering.

$O_2^n$ is a special case of the general family

$$O_2^n(x^n) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_i}{x_{n+1}} \right)^r \left( \frac{x_{n+1}}{x_n} \right)^{r/s}$$  
(1.5)

introduced by Gray (1979), who recommended especially $O_2^n$.

Ooe and Ulych (1979) also considered objectives related to the $O_2^n$ family.

Claerhout (1978) and Godfrey (1978) took an information theoretic point of view and suggested objectives

$$O_2^n(x^n) = -\text{"entropy"} \left( \frac{x^n}{s_x} \right).$$  
(1.6)

where "entropy" is an appropriate finite-sample measure of entropy and $s_x$ is a scaling factor.

All these objective functions have two properties in common:

1. They are scale invariant: $O(ax) = O(x)$, $a \neq 0$.
2. They depend on the sample $x^n$ only as an unstructured batch of numbers: i.e., they depend only upon the empirical distribution of the numbers $(x_1, \ldots, x_n)$ viewed as a simple random sample.

Because of these properties, success of a MED-type procedure depends on there being a difference between the empirical distribution of $x$ and that of $\hat{x} = b \star y$ where $b \neq b$. But by the ergodic theorem, the empirical distribution of $(\hat{x}_1, \ldots, \hat{x}_n)$ tends, as $n$ increases, to the theoretical distribution of any one value, $\hat{x}_i$ say, under repeated realization. Thus the workability of the MED procedure depends on properties of the theoretical distributions, which are discussed in section 2 and applied to the deconvolution problem in section 3.

2. THE PARTIAL ORDER \( \vdash \)

The time series studied in this paper are filtered versions of white noises: $y = f \star x$ is explicitly of this form, and $x = b \star y$ can be written as $x = (b \star f) \star x$. Consequently, each series is sampled from random variables which are linear combinations of independent and identically distributed random variables: $x_1$ is sampled from $\sum u_i x_{t-i}$ where $g = b \star f$. This section discusses the distributional properties of linear combinations and the implications in model (1.1) - (1.2).

Let $X$ be a set of random variables with finite variances which is closed under linear combinations, so that $\sum u_i X_i + c$ is in $X$ \(^\ddagger\) when the $X_i$ are.

\(^\ddagger\) In section 3, more specific properties of $X$ are introduced.
Two random variables \(X\) and \(Y\) will be regarded as equivalent, written \(X \equiv Y\), if for some constants \(c\) and \(a \neq 0\), \(aX + c\) has the same probability distribution as \(Y\). \(\equiv\) is an equivalence relation on \(X\).

**Definition:** \(X \succ Y\) means that for appropriate constants \((a_i)\) with \(\sum a_i^2 < \infty\),
\[
Y \equiv \sum a_i X_i,
\]
where the \(X_i\) are independent copies of \(X\). \(\succ\) is short for "\(\succ\) but not \(\equiv\)."

\(\succ\) is a partial order on \(X^*\), because of these two properties:

(a) **Transitivity:** If \(X \succ Y\) and \(Y \succ Z\) then \(X \succ Z\).

(b) **Asymmetry:** Let \(X\) and \(Y\) have finite variances.
If \(X \succ Y\) and \(Y \succ X\) then \(X \equiv Y\).

**Proof:** Property (a) is immediate from the definition, since if \(Z \equiv \sum b_i Y_i\) and \(Y \equiv \sum a_i X_i\), then
\[
Z \equiv \sum_{i,j} a_i b_j X_{ij}.
\]

Property (b) is more subtle, and requires the following characterization (a consequence of Theorem 5.6.1 of Kagan, Linnik, and Rao (1973)): Z is a Gaussian random variable if and only if Z has finite variance and
\[
Z \equiv \sum a_i Z_i
\]
holds for some coefficients \((a_i)\) which are nontrivial (at least two \(a_i\)'s \(> 0\)) and square-summable (\(\sum a_i^2 < \infty\)).

To use this result, suppose \(Y \equiv \sum a_i X_i\) and \(X \equiv \sum b_j Y_j\). Then \(X \equiv \sum a_i b_j X_{ij}\). This is precisely a relation of the form (2.1), and so if either \((a_i)\) or \((b_j)\) is nontrivial, then \(X\) is a Gaussian random variable. If so, then the characterization implies that any linear combination of \(X_i\)'s -- e.g., \(Y\) -- is also Gaussian, in which case \(X \equiv Y\). If the coefficients \((a_i)\) and \((b_j)\) are both trivial, of course, then \(Y \equiv X\) is immediate.

The characterization result provides that if \(Z\) is Gaussian, then \(Z \equiv X\) for any \(X\), since any linear combination of copies of \(Z\) is Gaussian, and hence equivalent to \(Z\). On the other hand, if \(X\) has zero mean and finite variance, the central limit theorem says that \(S_n = n^{-1/2} \sum X_i\) converges in distribution to a Gaussian random variable \(Z\). Since \(X \equiv S_n\) and \(S_n \overset{d}{\to} Z\), it is tempting to conclude that \(X \equiv Z\) for every \(X\) in \(X\). In this paper, this is taken as a definition; • remains a partial order with this additional stipulation.

**Result 2.1:** For \(X\) in \(X\) and \(Z\) Gaussian
\[
X \equiv \sum a_i X_i \equiv Z;
\]
this order is strict unless either (a) \(X\) is Gaussian (then \(X \equiv \sum a_i X_i \equiv Z\)) or (b) \(X\) is not Gaussian, but the linear combination is trivial (no two \(a_i\)'s nonzero) (then \(X \equiv \sum a_i X_i \equiv Z\)).

Relation (2.2) says that linear combinations of independent random variables are "more nearly" Gaussian than the individual components of the combination. Because of this, a mnemonic for "\(X \succ Y\)" is the phrase "\(Y\) is more Gaussian than \(X\)."
2.1 \( \rightarrow \) and Filtering

Equation (2.2) has a straightforward translation to a time series setting. For example, a white noise \( x \) is naturally associated with the random variable \( X \) from which \( x_t \) is sampled; and a filtered version of \( x \), \( g \ast x \), is associated with \( \{g \ast x\}_{t} \). Then for filtered white noises \( x \) and \( y \), \( x \ast \rightarrow y \) is defined to mean that \( X \ast \rightarrow Y \) for the associated random variables. Relation (2.2) becomes: if \( x \) is a white noise sampled from copies of \( X \) in \( X \), and if \( z \) is a Gaussian time series, then for any filter \( g \)

\[
x \ast \rightarrow g \ast x \ast \rightarrow z;
\]

(2.3)

the order is strict unless either (a) \( X \) is Gaussian (then \( x \ast \rightarrow g \ast x \ast \rightarrow z \)) or (b) \( X \) is non-Gaussian but \( g \) is a trivial filter \( g = (\ldots, 0, a, 0, \ldots) \) (then \( x \ast \rightarrow g \ast x \ast \rightarrow z \)).

In other words, every filtered version of a white noise is "more nearly" Gaussian than the white noise itself.

In the deconvolution setting, this has immediate applications. Since \( x = b \ast y \), putting \( g = b \ast f \) in (2.3) gives

\[
b \ast y \ast \rightarrow b \ast y \ast \rightarrow z
\]

(2.4)

Thus the filter \( b \) is characterized by giving the "least Gaussian" output of any filtered version of \( y \).

The extremal characterization of \( b \) by (2.4) makes the numerical maximization in rule (1.3) seem plausible. In fact it will be seen in Section 3 that if \( O \) "agrees" with the partial order \( \ast \rightarrow \), then rule (1.3) is a consistent way of estimating \( b \).

\[\dagger\] This "association" is that between the distribution of \( x \) and the limiting empirical distribution of \( \{X_1, \ldots, X_n\} \), which are the same with probability 1 because \( x \) is ergodic.

2.2 Well-Posedness of Model (1.1)

The relation (2.3) helps settle the question of uniqueness of the convolutional representation (1.1) \( \rightarrow \) (1.2).

Of course, equation (1.1) has an inherent nonuniqueness, since it does not fix the scale or time origin of \( x \). If \( y = f \ast x \) then also \( y = f' \ast x' \) where \( f' = cf \) and \( x' = x_0 - v \)

for some integer \( v \) and any \( c \neq 0 \). Such ambiguities are not usually serious, since in physical problems they can be resolved by assumptions of causality and energy conservation, and in other problems by "natural" choices. When the representation \( y = f \ast x \) is satisfied only by \( f \) and \( x \) (or by scaled/shifted versions of those), model (1.1) will be said to be essentially unique.

Result 2.2: The model

\[
y = f \ast x
\]

(2.5)

a. \( x \) white noise sampled from \( X \) in \( X \);

b. \( f \) an invertible filter;

is essentially unique if and only if \( X \) is a non-Gaussian random variable.

Proof: Suppose there are two representations satisfying (2.5a) and (2.5b):

\[
y = f \ast x
\]

Now \( f \ast x \) and \( f' \ast x' \) have the same power spectrum, so both \( f \) and \( f' \) are invertible. Thus one can write

\[
x' = (f' \ast f) \ast x
\]

\[
x = (f' \ast f) \ast x'.
\]

Both \( x \ast \rightarrow x' \) and \( x' \ast \rightarrow x \). From relation (2.3) it follows that either (a) \( X \) is Gaussian; or (b) \( f' \ast f' \) is
trivial, so that the two representations differ only in scaling and choice of time origin. #

Figure 2 illustrates the nonuniqueness of the convolutional model when $x$ comes from a Gaussian process. Here the experiment of Figure 1 is repeated with Gaussian input. The three results $x^+$, $x^-$ and $x$ are all uncorrelated and statistically indistinguishable, since in the Gaussian case, uncorrelatedness implies independence. Moreover, the eye cannot see any qualitative differences between the results. One might claim that the eye makes the same sort of qualitative distinctions as *; if this is so, then Result 2.2 would imply that only for Gaussian inputs will the eye be unable to make distinctions among different phase assumptions in deconvolution. Jon Claerbout has suggested that perhaps the eye itself follows a MED-type procedure for focusing; it is not clear whether this analogy has any physiological validity. If there is any basis in this suggestion, we should feel lucky that the scenes we view in everyday life do not come from a Gaussian distribution, or else we might get headaches from squinting in vain!

3. CONSISTENCY

In this section, the estimation rule (1.3) is checked for consistency. Specifically, it is asked under what conditions on $y$ and $0$ will $b^n$ (defined in section 1.2) converge in probability to $b$, in the sense that for any $c > 0$,

$$\text{Prob}\{|b^n - b| > c\} \to 0 \text{ as } n \to \infty.$$  \hspace{1cm} (3.1)
Consistency is a standard "test of metal" for an estimation rule. Any "reasonable" technique should, in a stationary setting, give answers which improve as the sample size increases. Consequently, an estimator that does not satisfy (3.1) is not making good use of the available data and does not appear to be a good procedure.

3.1 Preliminaries

The major assumption being made in a consistency proof is the assumption that the model holds: in this case, that means that the observed series \( y^n \) is a segment of \( y=f_{x} \), where \( x \) is a white noise and \( f \) is an invertible filter with an inverse \( b \) of known length \( p+1 < n \). Whether this model does hold is an empirical matter, of course.

Equation (3.1) should be interpreted with care. Because, as discussed in section 2.2, the scale and time origin of \( x \) are unknown, the "true" \( b \), i.e. the filter that satisfies (1.2), is known only up to a time shift and/or scale factor. The comparison made in (3.1) must be interpreted as being between normalized versions of \( b \) and \( b^n \). A convenient set of normalized filters is

\[
B_{p} = \{ \tilde{b} : \tilde{b}_{t}=0 \text{ if } t<0 \text{ or } t>p, \tilde{b}_{0}^2=1, \tilde{b}_{0} > 0 \}.
\]

The proof requires two things of the objective functions of interest. First, they must be scale invariant; this is natural restriction, since the scale of \( x \) and \( b \) is unknown. Second, they must be continuous functions of the distribution of their argument; continuity is essential if rules like (1.3) are to make sense. Objectives (1.4)-(1.6) are continuous where they are well-defined. For example, \( O_2 \) is continuous if the restriction is made that all random variables in the "universe" \( X \) have finite fourth moments.

3.2 Characterization of Admissible Objectives

The technical details involved in proving consistency are discussed in the appendix. The basic idea is simple. By the ergodic theorem, the empirical distribution of \( x^n \) converges as \( n \) increases to the distribution of the random variable \( x \) associated with \( x \). Now, \( \tilde{x} \) is a good estimate of \( x \) to the same extent that \( \tilde{x} \) resembles \( X \) rather than a linear combination \( \sum a_i X_i \). So the objective function \( O \) has to discriminate between the distribution of \( X \) and that of \( \sum a_i X_i \). In other words, \( O \) must do roughly the same thing as \( \rightarrow \).

Definition: \( O \) agrees with \( \rightarrow \) on \( X \) if

\[
x \rightarrow y \text{ implies } O(x) > O(y)
\]

(3.2)

for every \( x,y \) which are filtered white noises sampled from random variables in \( X \).

Result 3.1: Let \( O \) be a scale-invariant, continuous functional. The sequence \( (b^n) \) defined by (1.3') is a consistent sequence of estimates of \( b \) for every non-Gaussian \( x \) of the form (2.5) if and only if \( O \) agrees with \( \rightarrow \) on \( X \).

One drawback of standard techniques such as maximum likelihood is that one must make very specific assumptions about the distribution of the random component of one's data (e.g.

\[
O(x) \text{ is the limit of } O(x^n) \text{ as } n \rightarrow \infty.
\]

It is well-defined because \( x \) is ergodic and \( O \) is continuous.
x). Result 3.1 shows that an objective which agrees with \( \cdot \cdot \cdot \) will give consistent estimates no matter what the distribution of \( x \), as long as it is non-Gaussian. This property is useful because the filtering operation \( y = f \ast x \) distorts distributions and makes it hard to determine the properties of \( x \).

3.3 Examples

While condition (3.2) is not trivial to verify, for certain objectives this has already been done; for others, numerical verification is possible. Here are four examples.

Example 1: Standardized Cumulants. Granger (1976) has discussed the order-like properties of standardized cumulants. The mth cumulant of the random variable \( X \) is defined by

\[
C_X^m = \left( -i \frac{d}{dt} \right)^m \log(e^{itX})
\]

If independent random variables are added, the corresponding cumulants add:

\[
C_{X+Y}^m = C_X^m + C_Y^m.
\]

The standardized mth cumulant is defined by

\[
K_X^m = C_X^m / (C_X^{2m/2});
\]

after standardization, additivity is lost. Instead if \( Y = \sum a_i X_i \), where the \( X_i \) are independent copies of \( X \),

\[
K_Y^m = K_X^m \left( \frac{\sum a_i}{\sum a_i^2} \right)^{m/2};
\]

so for \( m > 2 \), \( |K_Y^m| \leq |K_X^m| \). Therefore

\[
X \preceq Y \implies |K_X^m| \geq |K_Y^m| \quad m > 2. \quad (3.3)
\]

Now (3.3) says that the standardized cumulants of order \( m \) agree with \( \cdot \cdot \cdot \). In order to get the consistency condition (3.2), the ordering in (3.3) must be strict. One way to do this is to use the power of fiat: let \( \chi \) be a linear space of random variables having finite mth cumulant and which contains no non-Gaussian random variables with a zero mth cumulant. Then \( C_K^m(\chi) \), the absolute value of the standardized mth cumulant of the sample \( (\chi_1, \ldots, \chi_n) \), agrees with \( \cdot \cdot \cdot \) on \( \chi \), and so gives a consistent estimation rule.

It turns out that if \( x \) has zero mean, \( O_4^4 = |O_4^2 - 3| \), where \( O_4^2 \) is Wiggins's original proposal (1.4); this shows that the original MED procedure is consistent whenever in model (1.1) \( x \) has a kurtosis greater than 3. This relation between \( O_4^4 \) and \( O_4^2 \) also shows that if \( x \) has a kurtosis smaller than 3, minimizing, not maximizing, \( O_4^4 \) is the appropriate strategy -- a somewhat unexpected finding.

Example 2: Measures of Information and Entropy. Fisher's Information (for location) is defined for a random variable \( X \) with a smooth probability density \( p_X \) by

\[
I(X) = -\int \sum \frac{1}{u^2} \log(p_X(u)) p_X(u) \, du. \quad (3.4)
\]

The (Shannon) Entropy of \( X \) is defined by

\[
H(X) = -\int \log(p_X(u)) p_X(u) \, du. \quad (3.5)
\]

Using variational calculus, it is not hard to show that if \( \int \sum a_i^2 \), \( 1 \) and \( X \) has finite variance,

\[
\frac{1}{\sum a_i^2} = 1
\]
The inequalities being strict unless \( X \) is Gaussian or the linear combination is trivial. Therefore, if \( O_E \) and \( O_I \) are objectives with

\[
O_E(x^n) = -E(\bar{X})/\text{Var} \bar{X} \quad \text{and} \quad O_I(x^n) = I(\bar{X})/\text{Var} \bar{X}, \quad n = \infty,
\]

then \( O_E \) and \( O_I \) agree with \( \rightarrow \) on a space \( X \) of random variables with finite variances and smooth density functions.

Thus, terminology such as "Minimum Entropy Deconvolution" does describe valid techniques for deconvolution.

Example 3: Ideal Metrics. The interpretation of (2.2) was that linear combinations of independent variables are more nearly Gaussian than the individual variables. Using Zolotarev's (1979) ideal metrics, this qualitative statement can be made precise.

An ideal metric of order \( s \), \( \nu^s \), obeys the triangle inequality

\[
\nu^s(X,Z) \leq \nu^s(X,Y) + \nu^s(Y,Z)
\]

and the identification property \( \nu^s(X,Y) = 0 \) only if \( X \rightarrow Y \), as well as the convolution property \( \nu^s(X*Y,Z*Y) \leq \nu^s(X,Y) \) and the homogeneity property

\[
\nu^s(cX,cY) = |c|^s \nu^s(X,Y).
\]

A result of Zolotarev (1976) implies that if \( \{X_i\} \) are independent copies of \( X \), then for an ideal metric of order \( s \),

\[
\nu^s\left( \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n a_i^2}}, Z \right) \leq \nu^s(X,Z) \left( \frac{\sum_{i=1}^n a_i}{\sqrt{\sum_{i=1}^n a_i^2}} \right)^s / 2
\]

where \( Z \) denotes a Gaussian random variable. Therefore, for \( n > 2 \), on the set \( X \), where \( \nu^s(X,Z) < \infty \),

\[
X \rightarrow Y \implies \nu^s(X,Z) > \nu^s(Y,Z).
\]

In section 2, the phrase "\( Y \) is more nearly Gaussian than \( X \)" was suggested as an mnemonic for the relation \( X \rightarrow Y \). If the reader will use the mnemonic "distance of \( X \) from the Gaussian" for \( \nu^s(X,Z) \), then (3.7) sounds particularly simple when read aloud!

Gray (1979) suggested that a certain MED-type method worked by driving its output as "far as possible" from Gaussianity. When one uses an ideal metric as objective function, such a description of the output of the procedure would be particularly apt.

Example 4: Numerical Verification. Sample \( x \) from copies of \( X \) where \( \text{Var}(X) < \infty \); then \( x \) has the Fourier representation

\[
x_t = \text{Re} \int_{-\pi}^\pi e^{it\omega} dZ_x(\omega),
\]

where \( Z_x \) is a spectral measure with uncorrelated increments. Let \( \phi \) be a given function, and define \( x^\alpha \) for \( \alpha > 0 \) by

\[
x_t^\alpha = \text{Re} \int_{-\pi}^\pi e^{it\omega + it\alpha \phi(\omega)} dZ_x(\omega);
\]

thus \( x^\alpha \) is a "phase shifted" version of \( x \).

Note that \( x^0 \rightarrow x \), and since phase shifting corresponds to a filtering operation,

\[x^0 \rightarrow x^\delta, \ \delta > 0.\]
Based on the additive property $e^{\alpha+\delta} = e^{\alpha}e^{\delta}$, it should be true that

$$x^\alpha \ast \sim x^{\alpha+\delta}, \quad \delta > 0.$$  \hspace{1cm} (3.9)

Hence if $0$ satisfies (3.2), $O(x^\alpha)$ should be a decreasing function of $\alpha$.

This is an assertion which can be checked on a digital computer, since for finite length series, the analog of representation (3.8) involves summation rather than integration.

Figure 3 shows $x^\alpha$ for various values of $\alpha$ using $\phi(\omega) = \sin(4\pi\omega)$; it is apparent in a rough way that $x^\alpha$ gets more entropic or Gaussian-like as $\alpha$ increases. In Figure 4, $O(x^\alpha)$ is plotted as a function of $\alpha$ for the objective functions $O_2^0$, $O_1^0$, $O_E$, and $O_1$. The monotone decreasing property is seen to hold for all these objectives, as required by consistency (3.2) and the assertion (3.9). If it were not known in advance that (3.2) held, this numerical technique might provide a useful check.

### 3.3 Summary

There are many ways of attacking the deconvolution problem using rule (1.3); approaches based on "simplicity" (cumulants), "information," or "distance from the Gaussian" all give consistent ways of estimating the deconvolution filter. In general, any well-behaved functional which agrees with the order $\sim$ will give a consistent MAD-type estimator under rule (1.3).
4. PRECISION OF MED-TYPE ESTIMATES

Section 3 shows that (1.3) is a consistent estimation rule under a wide variety of objective functions. Some of these objectives perform better than others, in the sense of yielding more precise estimates in finite samples. In this section the precisions of several MED-type methods are evaluated and compared using asymptotic techniques.

4.1 Asymptotic Variance

Parameter estimates usually satisfy the "root $n$" law: if $\hat{\theta}^n$ is an estimate of $\theta$ based on $n$ samples of data, the estimation error $|\hat{\theta}^n - \theta|$ will be of order $1/\sqrt{n}$. This law can be expressed in rigorous form as

$$\sqrt{n}(\hat{\theta}^n - \theta) \sim \text{Gauss}(0, \sigma^2), n \rightarrow \infty,$$

meaning that the scaled deviation $\sqrt{n}(\hat{\theta}^n - \theta)$ has approximately a zero mean Gaussian distribution with variance $\sigma^2$. If this relation holds, then $(\hat{\theta}^n)$ is called asymptotically Gaussian with asymptotic variance $\sigma^2$.

Since most commonly-encountered estimators give asymptotically Gaussian results, a standard way of comparing the performances of estimators is simply to compare asymptotic variances. This is the approach used here. Be warned that asymptotics do not tell the whole story: they give approximations valid in large samples. However, they often give adequate results in small samples; furthermore, in the analysis of nonlinear technologies like MED, asymptotics may be the only analytic results available.

Below, the normalization $b_0 = b_0 = 1$ is used, rather than $b_0 = b_0 = 1$ which is useful in proving consistency. Then
the variance of an estimate \( \hat{b}_1, \ldots, \hat{b}_p \) of \( b_1, \ldots, b_p \) taken
the form of a \( p \times p \) matrix \( \Sigma \), whose entries indicate the
covariances of the corresponding estimation errors:
\[
\Sigma_{ij} = \text{Cov}(\hat{b}_i, \hat{b}_j).
\]
Thus the asymptotic properties of a M-E-D-type estimator are
summarized by
\[
\sqrt{n}(\hat{b} - b) \Rightarrow N(0, \Sigma), \quad n \to \infty,
\tag{4.1}
\]
and the performance of \( \hat{b} \) as an estimate of \( b \) is measured by
the asymptotic variance matrix \( \Sigma \).

4.2 (M) Estimators

A convenient strategy for deriving \( \Sigma \) is to study the
analytically tractable objective functions resulting from:

\[
\frac{1}{2} \sum_{i=1}^{p} \psi_{\theta}^R(x_i) y_{t-1} + R_{t-1}^n, \quad i = 1, \ldots, p \tag{4.2}
\]

where \( R^n = b \ast y^n \), \( \psi_{\theta} \) converges to a function \( \psi \) as \( n \to \infty \),
and \( R_{t-1}^n \) converges to 0 as \( n \to \infty \).

For objectives of type (M), gradients can easily be computed
from (4.2), and so numerical solution of (1.3) is feasible. Moreover, if (M)
holds, there is a simple asymptotic theory for the sequence \( \{\hat{b}^n\} \) along the lines of Martin (1970).
Thus there are practical and theoretical justifications for
assumption (M).

A simple (M) objective is Wiggins' proposal \( O_2 \). Straight-
forward differentiation as in (4.2) gives \( R_{t-1}^n = 0 \) and

\[
\psi_{\theta}^{O_2}(u) = \frac{4n^2}{(\frac{x_t}{u})^2} \left[ u^3 - \frac{x_t^4}{x_t^2} u \right]
\]

Properties of \( \psi \). The consistency of rule (1.3) requires
that \( O(b \ast y) \) have a maximum at \( \hat{b} = b \); from (4.2) and the
requirement that derivatives with respect to \( \hat{b} \) vanish at \( b \)
come the constraints
\[
0 = \frac{3}{\theta} \psi'(x_t)(\hat{b} = b) = \lim_{n \to \infty} \frac{1}{n} \psi(x_t) y_{t-1}, \quad i = 1, \ldots, p \tag{4.3}
\]

Regularity conditions sufficient to ensure (4.3) are

\[(R1) \quad E[\psi(X)] = 0, \tag{R1}
\]
\[(R2) \quad E[\psi'(X)] > 0, \tag{R2}
\]
\[(R3) \quad E[X \psi(X)] > 0. \tag{R3}
\]

Conditions (R1) and (R2) are familiar from Huber's (M)-esti-
mation theory; see Huber (1977). If \( X \) is assumed to be a
symmetric random variable, they will be satisfied by any
increasing, skew-symmetric function \( \psi \). Condition (R3) is
a result of the causality assumption for \( b \) (i.e. \( b_t = 0, t < 0 \)),
a choice of time origin which generally requires \( f \) to be a
causal filter. Because (4.3) requires that \( \frac{1}{n} \psi(x_t) y_{t-1} x_t \) and
similar sums tend to zero, (R3) arises if \( f \) is causal.

Note that if \( \psi \) satisfies (R1) and (R2), the third con-
tion is satisfied after "orthogonalization" -- i.e. by

\[
\psi_{new}(u) = \psi(u) - \frac{E[\psi(X)]}{\text{Var}(X)} (u - EX) \tag{4.4}
\]

In fact, \( \psi_{new} \) defined above can be interpreted as the application
of (4.4) to \( \psi(u) = u^3 \), under the assumption that \( X \) has
mean 0.
4.3 The Formula

RESULT 4.1: The asymptotic variance for objectives of type (M) with \( \psi \) satisfying (R1)-(R3) is

\[
\Sigma = \frac{V(\psi, X)}{\text{Var}(X)} R_1^{-1} R_0^{-1} R_1 R_0
\]

(4.5)

where

\[
V(\psi, X) = \left( \frac{\psi^2(X)}{(\psi'(X))^2} \right)
\]

and the matrices \( R_0, R_1 \) are defined in the appendix.

To a good approximation (see the appendix),

\[
\Sigma = \frac{V(\psi, X)}{\text{Var}(X)} R_1^{-1}
\]

(4.6)

where \( R \) is the \( p \times p \) autocorrelation matrix of \( y \):

\[
R_{ij} = \text{Cov}(y_{t+i}, y_{t+j})/\text{Var}(y_t) = \frac{\sum f_{t+i} f_{t+j}}{\sum f_t^2}.
\]

Formula (4.6) also occurs in Martin (1979). It splits the asymptotic variance nicely into two factors: (i) \( R_1^{-1} \), which depends only on the filter \( f \); and (ii) \( V(\psi, X)/\text{Var}(X) \), which depends only on the objective function (via \( \psi \)) and the probability model (via \( X \)). Consequently, the quantity

\[
A(\psi, X) = \frac{V(\psi, X)}{\text{Var}(X)}
\]

is fundamental for performance comparisons; for example, the asymptotic precisions under (4.6) of two different (M)-type estimators are related by

\[
\Sigma_1 = \frac{A(\psi, X)}{A(\psi, X)} \Sigma_2.
\]

4.4 Comparisons

In this section, the figure of merit \( A(\psi, X) \) is used to evaluate the performance of various MED-type methods. It will be useful to introduce some baselines against which the calculated values may be compared.

On Minimum Entropy Deconvolution

Baseline 1: Minimum Delay Wavelet

If \( b \) were known to be minimum delay, then least-squares autoregression (i.e., predictive deconvolution) would be a consistent estimator of \( b \), with asymptotic variance

\[
\Sigma_{ls} = R_1^{-1};
\]

(4.7)

see Martin (1979). Comparing (4.6) with (4.7), it is evident that if \( A \) is much larger than 1.0, then MED is an inefficient procedure to use in situations where minimum delay data are present.

Baseline 2: A Lower Bound

Variational techniques can be used to show that if \( \psi \) and \( X \) satisfy (R1)-(R3),

\[
A(\psi, X) \geq H(X) = (I(X)\text{Var}(X) - 1)^{-1};
\]

(4.8)

here \( I(X) \) is Fisher's Information (3.5). This lower bound has two properties based on those of \( I \):

(i) \( X \rightarrow Y \) implies \( B(X) < B(Y) \); the deconvolution problem (1.1)-(1.2) is intrinsically harder nearer the Gaussian.

(ii) As \( X \) approaches the Gaussian, \( B(X) \rightarrow 0 \); the deconvolution problem is intractable near the Gaussian (Result 2.1 showed that the problem is undefined at the Gaussian).

Performance at the Exponential Power Family

Gray (1979) was interested in the behavior of MED-type methods for data distributed according to the exponential power laws

\[
p_X(u) = \exp(-|u|^\alpha).
\]

Figure 5 displays the performance index \( A(\psi, X) \) as a function of \( \alpha \) for the objectives \( O_2^4 \) and \( O_2^1 \); the baselines \( B(X) \) and 1.0 are plotted as well. Note that near the Gaussian law \( (\alpha = 2) \) the methods perform poorly, as they must \( B(\text{Gaussian}) = 0 \).
Both objectives do well if $\alpha$ is far from 2: $O_{2}^{4}$ is optimal (it attains the lower bound $B$) for $\alpha = 4$, while $O_{2}^{1}$ is optimal for $\alpha = 1$.

**Performance at a Mixture Family**

Godfrey (1979) considered the two-parameter family of densities

$$p_{\xi, \varphi}(u) = (1-\xi)\phi(u) + \xi\phi(u/\varphi)$$

where $\phi$ is the Gaussian density with mean 0 and variance $\varphi$. $p_{\xi, \varphi}$ can be described as a mixture of typical and atypical data occurring in proportions $(1-\xi)$ and $\xi$ with variances $\varphi$ and $s^2$. Godfrey interpreted the typical data as seismic background and the atypical data as seismic events.

Figure 6 shows the performance of $O_{2}^{4}$, $O_{1}^{2}$ and $O_{BW}$ for $s = 5$ and $0 \leq \xi \leq 0.5$. Here $O_{BW}$ is derived from Tukey's biweight $\psi$ via (4.4); the biweight (bisquare) is defined in Hogg (1979).

For this family, Wiggins' and Gray's objectives perform poorly, while the biweight-based objective seems to do well. This can be explained by the fact that the biweight was in a sense designed for cases such as this mixture family, while the other objectives were designed for exponential power family data.

4.5 **Asymptotic optimality -- Minimum Entropy!**

The asymptotically best objective function under (4.6) can be derived using variational methods: the lower bound $B(X)$ of $A(\psi, X)$ is attained by

$$\psi^*(u) = \psi_{X}(u) - (u - EX)/\text{Var}(X).$$

(4.9)
On Minimum Entropy Deconvolution

\[ \psi_X(u) = -\frac{d}{du} \log p_X(u) \] is Fisher's score function (for location) for \( X \). An objective whose 'derivative' is \( \psi_X \) (recall (4.2)) is

\[ O^*(X) = \int \log(p_X/\psi_X) p_X ; \quad (4.10) \]

Here the density of \( X \) is \( p_X \) and \( \psi_X \) is Gaussian with the same mean and variance as \( p_X \). \( \exp(-O^*(X)) \) measures the asymptotic probability that a sample from \( p_X \) is mistaken by the most powerful test to be a sample from \( \psi_X \); see for example Kullback (1959). It follows from this information-theoretic interpretation that the best MED-type estimator minimizes the probability that \( \hat{x}^n = \hat{u}^n \cdot y^n \) could be mistaken for a Gaussian time series.

The integral in (4.10) is a measure of entropy; ignoring constants, \( O^*(X) = -E(X/\sqrt{\text{Var}(X)}) \), where \( E \) is the Shannon entropy functional (3.4). Thus the asymptotically optimal MED method is minimum entropy deconvolution. Wiggins's choice of terminology now appears prescient.

The optimal objective given by (4.10) is to some extent a mathematical fiction, since the density \( p_X \) appearing in the integral is in principle unknown. In practice, a density "guess" \( \hat{p}_X \) is necessary. There are two ways to make this guess.

1. Based on the data. For example, a simple histogram estimate of the density can be made, using bins \( k \) samples wide. This leads to the objective

\[ O^k(\hat{x}^n) = (-k/n) \sum_{i=1}^{n/k} \log(\hat{x}_{ik:n} - \hat{x}_{ik-k:n}) + (k/2) \log(\hat{x}_{i:n}^2) ; \]

the notation \( x_{i:n} \) means the \( i \)th largest of the \( n \) values \( x_1 \).
\[ x_n \] Claerbout (1978) argued heuristically in favor of \( \hat{Q}^1 \), but got unencouraging results. However, a histogram with bins of size \( k = 1 \) is an inconsistent density estimate; one might anticipate, based on general density estimation theory (e.g., Van Ryzin (1973)), that \( k \geq \sqrt{n} \) might be a better starting point for testing Claerbout's idea.

(11) Based on prior assumptions. In small samples, adaptive methods may work poorly, and a non-adaptive estimator based on a fixed guess may be preferred. Minimax theory (Huber (1977)) gives one way to choose \( \hat{p}_X \). One assumes that the "true" density lies in a neighborhood \( P \) of probability densities, and chooses \( \hat{p}_X \) in \( P \) to minimize the worst case loss (measured by \( A(\hat{p}, X) \)) of such a choice.

The author has considered the neighborhood \( P \) consisting of unit variance densities with most probability concentrated in the interval \((-\epsilon, \epsilon)\); this neighborhood has some seismological plausibility (sec. 5) and permits easy calculations using the techniques of Vandelinde (1979). The minimax answer, however, is involved and gives an objective similar to the biweight objective used in sec. 4.4.

4.6 Summary

Formula (4.6) simplifies asymptotic comparisons of precision. Wiggins' and Gray's objectives offer satisfactory performance in certain situations, but theory predicts that they can be outdone -- either by adaptive entropy-like objectives or by non-adaptive minimax objectives.

5. FURTHER ISSUES

There are many questions raised by the results of this paper which can only be briefly discussed here; the most important seem to be:

(1) The extent of non-Gaussianity in real data. In view of results 2.2 and 3.1, MED is only useful to the extent that one has non-Gaussian data in convolutional form.

Claerbout and Godfrey of the Stanford Exploration Project have argued that real seismic data are non-Gaussian. Godfrey (1978) found that seismic reflectivity logs (i.e., direct measurements of \( x \)) look roughly like white noises following distributions in the exponential power family with exponent \( a \) between .75 and 1.5; the exponent would be 2 if the data were Gaussian.

I have applied the fourth-order spectrum technique (see Brillinger (1965)) to the analysis of marine seismic data and measured standardized fourth cumulants in excess of 12 for "bright spot" data and in excess of 8 for ordinary data; under Gaussianity, values near 0 would be expected.

My own opinion is that non-Gaussianity is a pervasive phenomenon in seismology, and probably in all branches of time series analysis. However, the effect of filtering is to mask the distributional characteristics of the data; one must properly deconvolve one's data before the non-Gaussianity will be apparent.

(ii) The need for an independence assumption. This paper has leaned heavily on the assumption that the desired output of a MED-type procedure is a white noise. Is this assumption really important?
Scargle (1977), and also in this volume, has suggested that independence is the key property that allows phase estimation procedures to work. Stanford project workers such as Gray (1979) have also used the independence assumption in the derivation of their "Bussgang" heuristics.

Wiggins (1978), on the other hand, claimed that "... the MWD process makes no assumptions about the phase characteristics of the 'seismic wavelet', nor does it assume that the reflection series \( x \) is white ...". Ooe and Ulrych (1979) repeated Wiggins's assertion; however, neither of these papers discussed in any detail the recovery of a colored \( x \) given \( y = f^*x \).

My own opinion is that Wiggins is at least partly right. An example indicating that sufficiently extreme non-Gaussianity is sufficient to determine the success of a MWD-type scheme is the following: let \( x \) be a nonconstant, periodic time series with values \( 1,1 \). Then (see the appendix), if \( g \) is any filter shorter than the period of \( x \), \( O_2^4(x) \leq O_2^4(f^*x) \), with strict inequality unless \( g \) is a trivial filter. Thus \( x \) can be recovered (up to choice of sign and of time origin) from \( y \).

This example shows that MWD can work even in a deterministic setting; the probabilistic "white noise" framework makes for simple proofs, but the MWD idea actually can be used with some broader class of signals. My own feeling is that the characterization of such signals is a difficult problem; I have made some progress towards a characterization of the periodic signals \( x \) which can be recovered from \( y = f^*x \).

On Minimum Entropy Deconvolution

(iii) The effects of observational noise. Equation (1.1) can be made more general with the addition of a noise term:

\[
y = f^*x + e.
\]

If the noise level is moderate, Wiggins (1977) showed that the MWD technique gives stable and useful results. If the noise level is heavy, however, convolutional estimates of \( x \) (i.e. \( \hat{x} = b^*y \)) will be outperformed by Bayesian, nonlinear estimates of \( x \). In informal tests, a few such estimators have done well; see also Godfrey's thesis (1979).

(iv) The accuracy of the asymptotics. The asymptotic theory presented here depends on the assumption that there are many more observations than there are parameters to be estimated, so that \( p/n \) is small. This is often not the case in seismic practice, so these results need to be checked by Monte Carlo methods on artificial series with lengths typical of real data. Such tests have been tried on a small scale and seem to agree qualitatively with the asymptotic predictions.

(v) Relation to other work. Wiggins (1977) and Ooe and Ulrych (1979) emphasized the connection between \( O_2^4 \) and the varimax approach to factor analysis. A closer connection to multivariate statistics is that between MWD and Friedman and Tukey's (1974) projection pursuit.

In the projection pursuit method, one has a function \( I \) which measures how "interesting" a sample of data looks. Given an \( n \times p \) matrix \( Y \), one finds that projection \( Yb \) of \( Y \), where \( b \) is a \( p \times 1 \) vector, that is "most interesting"; i.e., one maximizes \( I(Yb) \) as a function of \( b \).

If projection pursuit is applied to the matrix
where the $y_t$ are samples from the time series $y$, then the "most interesting" projection $Yb$ is a filtered version of $y$:

$$Yb_t = \sum_{j}^b y_{t-j}.$$

In fact, this "filter pursuit" method is a MED-type method, with objective function I. The word "interesting" plays the same role in discussion of projection pursuit that the word "simple" plays in motivating MED methods. The reader may find a comparison between Wiggins (1978) and Friedman and Tukey (1974) quite interesting.

(vi) Application in other fields. Obviously, the problems treated in this paper can occur in many fields related to signal processing. The convolutional equation (1.1) describes the deblurring problem in image processing, where the convolutions are two-dimensional. In that setting, the results of this paper show that a picture $x$ which has been blurred by an unknown linear filter $f$ can be recovered, provided that the intensity of the pixels in $x$ does not follow a Gaussian distribution. I do not know of any applications of this idea, however.

In fields unrelated to signal processing, where convolutional structure is unlikely, MED can still be used as an exploratory data analysis tool. It can reveal "interesting"

or "simple" structure in unexpected places, and provoke thought and discussion in situations where one might have thought that there was not much to be said.

6. CONCLUSION

Proponents of MED can "walk proud", since the method has strong justification when the model (1.1)-(1.2) holds. Of course, the adequacy of such a model can be called into question, and so empirical study of the statistical structure of seismic time series is an important next step. One and Ulrych have claimed that MED represents the first significant breakthrough in deconvolution technology in many years. At least, MED represents a conceptual breakthrough, a tool not based on the usual correlation and spectrum approach to time series analysis.

APPENDIX

First, apologies to the technically minded persons who have been frustrated by the language I used in this paper. I did not use essential concepts like distribution function nor did I use proper qualifiers (e.g. "almost every"). My intention was to use a terminology that would be accessible to the applied worker.

This appendix is a collection of comments on the important technical points behind the main results.

(a) Second-order equivalence of filters. Let $f$ be a filter of length $p+1$; its $z$-transform is a polynomial of degree $p$ and can be factored as $w_0 \prod_k (w_k + z)$. Consequently $f$ can be factored into a convolution of dipoles $f = w_0 \cdot (w_1 \cdot$
defined as \( x_t = X_t(\omega) \), where \( (X_t) \) is a sequence of independent and identically distributed random variables defined on \( \Omega \). The series \( \hat{x} = \sum \hat{x}_t \) is a function of shifts of \( x \), which is ergodic; thus \( \hat{x} \) is also ergodic.

For almost every \( \omega \), the sequence of empirical distributions \( (F_{b,n}) \) of \( \hat{x} \) converges weakly to a limit \( F_b \). The distribution of \( \check{x}_t = \sum_{i=1}^{n} 1_{t-i = u} \) where \( g = b \ast f \). Moreover, if \( R(F_0) \) and \( R(F_{b,n}) \) are (a.e.) finite, then \( (F_{b,n}) \) converges to \( F_b \) in the stronger topology (a2). Therefore, \( O(b^*y^n) \) converges almost surely as \( n \to \infty \) to a well-defined limit \( O(b^*y) \). This limit function is a continuous function of \( b \), because \( F_b \) is a weakly continuous function of \( b \) (i.e.,

\[ \mathbb{E} \left( F(\cdot/e_1) \right) \ast F(\cdot/e_2) \ast \ldots \) is weakly continuous).

(a4) Uniform Convergence of Empirical Distributions.

Let \( B \) be a set of filters of length \( p+1 \) that is compact when viewed as a subset of \( (p+1) \)-dimensional Euclidean space. Then, in probability, \( (F_{b,n}) \) converges to \( F_b \) uniformly in \( b \) in the sense that for any \( \delta > 0 \),

\[ \text{Prob} \left\{ \sup_{b \in B_p} \sup_n \| F_{b,n} - F_b \| > \delta \right\} \to 0 \]

as \( n \to \infty \). Here \( \| \cdot \| \) is a metric generating the weak topology (see Huber (1977)) or the stronger topology of (a2) (see Zolotarev (1979)).

(a5) Convergence in Probability. Consider only objectives taking on a unique maximum, so rule (1.3) is well-defined. Then \( \hat{b}^n \) converges in probability to the maximizer \( \hat{b} \) of \( O(F_b) \).

Proof: Fix \( \epsilon > 0 \). It is required to establish (3.1).
On Minimum Entropy Deconvolution

(a7) Asymptotic Variance Formula. The general formula for the asymptotic variance of an M-estimate is

$$\lim_{n \to \infty} n \times \frac{1}{n} \mathbb{E} \left[ \mathbf{R}_n^T \mathbf{R}_n^{-1} \right]$$

where $\mathbf{R}_n$ is the gradient of $O(b \ast y)^n$ at $\hat{b} = b$, and $\mathbb{E}$ is the expected value of the Hessian there. This formula assumes $\mathbb{E} \mathbf{X}_t = 0$,

$$\mathbb{E} \left[ \mathbf{R}_n^T \mathbf{R}_n^{-1} \right] = \mathbb{E} \left[ \frac{1}{n} \sum \psi(x_i) y_{t-1} \right] \mathbb{E} \mathbf{X}_t^2 \times \left[ \mathbf{R}_t + e_{t+1} U(\psi, X) \right],$$

after some algebra, this can be reduced to

$$n^{-1} \times \left( \mathbf{I}^T \mathbf{I} \right) \mathbf{E} \psi^2(x_t) \mathbf{E} \mathbf{X}_t^2 \times \left[ \mathbf{R}_t + e_{t+1} U(\psi, X) \right],$$

where $\mathbf{R}_t = \mathbf{I} \mathbf{I}^T$, $e_{t+1} = \mathbf{I} \mathbf{I}^T$, $\mathbf{E} \psi^2(x_t)$, and $U(\psi, X) = 1 - (\mathbf{E} \psi^2(x_t) \mathbf{X}_t^2) / (\mathbf{E} \psi^2(x_t) \mathbf{X}_t^2)$. The term in braces above is the 1j-th entry of the matrix $\mathbf{R}_t$.

The Hessian is analyzed similarly:

$$\left( \mathbf{R}_t^T \mathbf{R}_t^{-1} \right)_{1j} = \mathbb{E} \left[ \frac{1}{n} \sum \psi(x_i) y_{t-1} y_{t-j} \right] \mathbb{E} \mathbf{X}_t^2 \times \left[ \mathbf{R}_t + e_{t+1} W(\psi, X) \right],$$

where the term in braces is the 1j-th entry of $\mathbf{R}_t$, and

$$\mathbb{E} \psi^2(x_t) = 1 - (\mathbf{E} \psi^2(x_t) \mathbf{X}_t^2) / (\mathbf{E} \psi^2(x_t) \mathbf{X}_t^2).$$

The approximation (4.6) comes from the fact that trace $\mathbf{R} = n$ and trace $\mathbf{W} = n (n-1) / 2$, and the assumption that $\mathbf{U}$ and $\mathbf{W}$ are not large.

(a8) If $x$ is any sequence of -1's and +1's, then $Q_{\psi}^4(x) = 1$. It is easy to check that this is the smallest possible value of $Q_{\psi}^4(x)$ and that it is only attained by a sample of numbers all

...
with the same absolute value. Therefore, filtering x must
increase $O_2^4$ unless the output of the filter has constant
absolute value. Is there any nontrivial filter whose output
has constant absolute value?

Suppose \( x \) is periodic with period \( q+1 \). The circular
\( z \)-transform of \( x \) is a polynomial of degree \( q \) with coeffi-
cients \( +1 \). Any filter which does not increase $O_2^4$ maps this
polynomial into another polynomial with coefficients \( +a \)
and has a \( z \)-transform which is the ratio of two such poly-
nomials. In particular, if such a filter is nontrivial,
its \( z \)-transform is not a polynomial of degree less than \( q \)
(this substantiates the claim made in the text that filters
of length less than \( q \) must increase $O_2^4$ in this case). More-
over, fixing the amplitude, \( a \), of the output series, there
are at most a finite number of such filters, since there are
at most a finite number of series of length \( q \) with values \( +a \)
\( (2^q \) series to be exact). Therefore, for "almost every" filter
\( f \), \( x \) is recoverable from \( y \cdot f \ast x \).

The interested reader can already see how if one studies
an objective like $O_2^4$ with nice algebraic structure (the
ratio of two polynomials), the set of recoverable inputs can
be characterized using algebraic techniques.

ACKNOWLEDGEMENTS

Thanks to Ken Larner for suggesting and supporting this
work.

I was supported during the writing of this paper by a
NSF graduate fellowship, by NSF contract MCS-79-08685, and
by ONR contract N00014-79-C-0512.


