EXPLOITING SENSITIVITY OF NONORTHOGONAL JOINT DIAGONALISATION AS A SECURITY MECHANISM IN STEGANOGRAPHY

B.R.Matam, David Lowe

Aston University, UK
{matambr,d.lowe}@aston.ac.uk

ABSTRACT

We have recently proposed the framework of independent blind source separation as an advantageous approach to steganography. Amongst the several characteristics noted was a sensitivity to message reconstruction due to small perturbations in the sources. This characteristic is not common in most other approaches to steganography. In this paper we discuss how this sensitivity relates the joint diagonalisation inside the independent component approach, and reliance on exact knowledge of secret information, and how it can be used as an additional and inherent security mechanism against malicious attack to discovery of the hidden messages. The paper therefore provides an enhanced mechanism that can be used for e-document forensic analysis and can be applied to different dimensionality digital data media. In this paper we use a low dimensional example of biomedical time series as might occur in the electronic patient health record, where protection of the private patient information is paramount.

Index Terms— ICA, Sensitivity, Watermarking, Nonorthogonal Joint Diagonalisation

1. INTRODUCTION

The use of embedding hidden messages in digital data for the protection of private information (such as in the electronic patient health record) or for forensic analysis (including proof-of-ownership, or last-access logs) relies on algorithms with in-built security. However most steganographic approaches to data embedding have no explicit security in that at least the location of the embedded message, if not the actual message itself can be discovered or removed by a skilled attacker [1, 2, 3, 4]. Hence a steganographic system which automatically embeds last-access details of any user would lose its relevance if a skilled attacker can discover and hence change the secret message. By 'skilled' we mean that any steganographic system must accept the 'Kerckhoff principle' in cryptography. Specifically, the security of the system must accept that the 'attacker' has full knowledge of the methods, design and implementation details of the system. The only missing information to the attacker is some extra private information or a secret key mechanism. For a good steganographic system, we additionally require that the attacker needs this key to high accuracy (ideally perfectly) in order to reconstruct the message.

We follow [5] in our use of the term 'security', in that we discriminate between robustness of a watermark (the ability of a watermark to withstand standard signal processing operations), and security (the ability of a watermark to withstand intentional tampering). In the electronic patient health record where a patient's private and confidential data is the content of the embedded message, it is important an attacker is unable to locate enough bits of the message to attempt a reconstruction. The attacker will only have access to the watermarked cover and possibly have access to similar, though not identical, examples of non-watermarked covers (and hence would be able to assess statistics of the covers for example). In this sense we focus on the issue of security in a watermarked-only-attack [6].

In this paper we wish to focus on a specific embedding approach and expand on a recently observed phenomenon of inherent sensitivity in one specific class of data hiding methods. We will show the source of this sensitivity, and provide a practical illustration of the use of this sensitivity as additional security to an attack to try and recover a random secret message. Since most steganographic methods are designed for images in which there is significant redundancy in data, we deliberately focus on time series data since it is much more difficult to embed secret messages in one-dimensional data streams. We have a specific interest in the protection of personal and private data in biomedical time series such as EEG and ECG time series as part of an individual's electronic patient health care record, but our approach has relevance to audio and other low dimensional data streams. The methods can also be applied to higher dimensional data such as images and video without modification.

Blind Source Separation (BSS) is a well known signal processing technique used in analysing a mixed set of data generated from multiple sources. The application of one of the popular BSS techniques, the ICA (Independent Component Analysis) method, for watermarking applications has previously been presented in [7, 8], and has also been used as an attack framework in [3] for images by exploiting the information that the embedded message is usually statistically independent of the natural statistics of the cover.

It relies on a simple linear model in which an i-th observation $\mathbf{x}_i \in \mathbb{R}^N$ is generated from a set of P underlying latent sources $\{\mathbf{s}_p \in \mathbb{R}^N\}$ combined by a set of mixing coefficients $\{a_{ip}\}$: $\mathbf{x}_i = \sum_p a_{ip}\mathbf{s}_p$. With a set of observations this is equivalent to $\mathbf{X} = \mathbf{AS} + \mathbf{n}$ where n is a possible single additive Gaussian noise source. The source separation problem cannot be solved if there is no knowledge of $\mathbf{A}$ or $\mathbf{S}$ from just the observation data (the covertext) $\mathbf{X}$. Simultaneous estimation of $\mathbf{A}$ and $\mathbf{S}$ is the blind source separation problem, and the independent component analysis solution to the BSS problem assumes that (1) the sources are statistically independent, (2) at most one of the sources is Gaussian distributed, and (3) the mixing matrix is full rank.

Projecting the observation vectors $\mathbf{X}$ onto the inverse of the mixing matrix, $\mathbf{W}$ yields the latent sources in the ideal case $\mathbf{WX} \rightarrow \mathbf{S}$.

We have previously [10, 11, 12] demonstrated for time series and images, the use of independent component analysis as a useful framework for information hiding. We have shown that hiding information in the independent sources has good robustness and data integrity. This is intuitively clear since the focus of such methods is to decompose a signal into its statistically independent underlying sources. If the sources are statistically independent they do not interfere with each other, unlike for example, in an orthogonal decomposition of a system such as a Fourier analysis. Hence slightly changing one of these independent components should not distort the other components since they are independent of each other.

However, we have also previously observed that there is a sensitivity to message reconstruction. Specifically, in attempting to recover the hidden message in the ICA approach it is important to use exactly the same separating matrices obtained in constructing the source vectors. This extreme sensitivity to message reconstruction is not intuitively obvious, since we do not observe the same sensitivity when we use an equivalent PCA or spectral approach for example. However, very recently [13] a possible mathematical explanation for this observation has been presented which we exploit in this paper as a significant advantage of the ICA approach to information hiding from a security and forensics perspective.

We will see that although the ICA method is stable to small perturbations in the observations in extracting the underlying statistically equivalent sources, it is not stable in message reconstruction. Hence an attacker who knows the method and attempts to reconstruct the correct but unknown separating matrix themselves from similar *but not identical* observation data (such as the exact set of original patient EEG data for example) will be unable to reconstruct the hidden message, such as private patient information or previous access data.

2. STATEMENT OF THE SENSITIVITY PROBLEM

A secret message is embedded in the transform domain defined by the ICA framework using a specific unmixing matrix $\mathbf{W}$ calculated explicitly from a specific original covertext. The unmixing matrix remains as a private key and the *exact* covertext is not publically available.

Using this unmixing matrix allows the construction of the latent sources, and hence have access to the basis vectors modified to incorporate the message.

In most ICA based watermarking techniques the set of mixed observations $\mathbf{X}$ are taken from a cover $\mathbf{c}$, and $\mathbf{W}$ is used as the key to obtain a transformed set of vectors from $\mathbf{X}$, $\mathbf{S}$. $\mathbf{S}$ is used as the embedding space for $\mathbf{m}$. $\mathcal{F}$ defining the embedding function, $\mathbf{k}$ samples of $\mathbf{S}$ are selected as the probable embedding locations of $\mathbf{m}$. The selection of $\mathbf{k}$ is based on the application of $\mathbf{c}$ and $\mathbf{m}$.

$$\mathcal{F}(\mathbf{S}(\mathbf{k}), \mathbf{m}) \rightarrow \tilde{\mathbf{S}}.$$

At the decoder the modified source $\tilde{\mathbf{S}}+\mathbf{W}\epsilon_{\mathbf{c}}$ is used to retrieve an estimate of $\mathbf{m}$, $\hat{\mathbf{m}}$.

From an attacker's perspective this provides both a forward and a backward problem. The forward problem is: *Given an estimate of* $\mathbf{W}$, $\hat{\mathbf{W}}$, *how closely can the mixing matrix and the latent sources be estimated?*. The important backward problem is then: *Given an estimate of the mixing matrix. how accurately can the unmixing matrix be determined and is this sufficient to locate the hidden message?*.

The backward problem is the important one because the attacker only has access to the watermarked cover (and possibly representative examples of unwatermarked covers), so an unmixing matrix needs to be applied to recover the basis vector sources which carry the secret message. Once the basis vectors are uncovered, traditional methods can be applied to the sources to hunt for the message. The essential ingredient is whether a small error in knowledge of the covertext leads to small errors in construction of the mixing matrix, and whether small errors in the mixing matrix lead to small errors in the unmixing matrix. This is the crux of the sensitivity problem.

This paper demonstrates how small perturbations to $\mathbf{X}$ can result in large perturbations in the estimation of $\mathbf{W}$. It will be demonstrated how the relationship between the perturbations in $\mathbf{X}$ and $\mathbf{W}$ provides a security mechanism for watermarking applications.

3. THE APPROXIMATE ICA APPROACH

Given observations $\mathbf{X}$, the ICA algorithm estimates the separating (unmixing) matrix $\mathbf{W}$ and the probable statistically independent sources $\mathbf{S}$. Assume $\check{\mathbf{S}}$ is an ideal true set of statistically independent sources which are not observable, then

$$\check{\mathbf{S}}_{p\times N} = [\check{\mathbf{s}}_1, \ldots, \check{\mathbf{s}}_P],$$

$$\boldsymbol{Pr}(\cap_{p=1}^{P}\check{\mathbf{s}}_p) = \prod_{p=1}^{P}\boldsymbol{Pr}(\check{\mathbf{s}}_p).$$

Let $\check{\mathbf{A}} \in \mathbb{R}^{P\times P}$, such that

$$\mathbf{X} = \check{\mathbf{A}}\check{\mathbf{S}},$$

where $\check{\mathbf{A}}$ is the unknown mixing matrix. (We are now assuming a square mixing problem for simplicity of notation).

Once the secret message is embedded in the sources, they are no longer statistically independent and the previous assumptions of separability of the source distributions is only approximate. This slight relaxing of the independence assumption has a strong consequence for algorithms attempting to diagonalise the resulting cumulants, as discussed later.

3.1. Estimation Problem

The estimated separating matrix $\hat{\mathbf{W}}$ by the ICA algorithm should ideally be the inverse of the unknown $\check{\mathbf{A}}$. Let ξ represent the threshold of the perturbation of $\mathbf{X}$, due to uncertainties in its exact determination by an attacker.

$$\mathbf{X} + \xi \rightarrow \hat{\mathbf{X}}.$$

$$\mathbf{X} \overset{ICA}{\rightarrow} \check{\mathbf{A}}\check{\mathbf{S}}.$$

$$\hat{\mathbf{X}} \overset{ICA}{\rightarrow} \hat{\mathbf{A}}\hat{\mathbf{S}}.$$

It was observed in [10] that if $||\hat{\mathbf{S}} - \check{\mathbf{S}}|| = \epsilon_{\mathbf{c}}$ and

$$\mathbf{D}_{Emb} + \epsilon_{\mathbf{c}} \gg \zeta.$$

where $\mathbf{D}_{Emb}$ is the cover distortion due to the embedded message, then the hypothesis of correct message retrieval would be zero $\mathcal{H}_{\epsilon_{\mathbf{c}}} = 0$.

The sensitivity problem is therefore defined as 'the problem of defining bounds for ξ which in turn will define the bounds for $\epsilon_{\mathbf{c}}$ which will affect the decision $\mathcal{H}_{\epsilon_{\mathbf{c}}}$'.

The query is thus, what affects the accurate reconstruction of $\mathbf{W}$ given only observations of data, or approximations of the moments of the data distribution?

4. MATRIX JOINT DIAGONALISATION

Many problems in signal processing can be posed as an issue of jointly diagonalising a set of matrices. Specifically, given a set of $n \times n$ symmetric matrices $\{\mathbf{C}_i\}, i = 1 \ldots N$, find a non-singular matrix $\mathbf{B}$ such that $\mathbf{B}\mathbf{C}_i\mathbf{B}^T$ are as diagonal as possible. When we construct these matrices from observed data, we do not expect to be able to exactly diagonalise all these matrices, and so we are dealing with an approximate joint diagonalisation problem.

Many domains consider a restricted form of *orthogonal* joint diagonalisation where the matrix is constrained to be orthogonal. ICA does not generally have this constraint unless an additional requirement of prewhitened covariance matrices are used, which then imposes an additional constraint on the possible diagonalisations of other matrices. So generally, ICA can be considered as a non-orthogonal joint matrix diagonalisation problem as follows.

When we consider the generation of observed data vectors from the independent sources $\mathbf{x} = \mathbf{A}\mathbf{s}$, the requirement of independence of the $\mathbf{s}$ imposes constraints on the form of the possible matrices $\mathbf{A}$ that can be estimated from knowledge of the distribution of $\mathbf{x}$. In particular the covariance of $\mathbf{x}$, $\mathbf{R}_{\mathbf{x}}$ satisfies $\mathbf{R}_{\mathbf{x}} = \mathbf{A}\Lambda_s\mathbf{A}^T$ where Λ_s is the diagonal covariance of $\mathbf{s}$ — diagonal, since the source vectors are independent. Similarly, it is a property of independent random variables that all higher order cumulant matrices of independent sources must also be diagonal.

Hence, given a set of observation vectors from which we can construct a finite set of higher order cumulants, the ICA method is equivalent to finding a decomposition of $\mathbf{X}$ such that its projection jointly diagonalises all higher order cumulants.

Hence ICA as a problem is generically equivalent to the problem of non-orthogonal joint diagonalisation of matrices. What does this have to do with the sensitivity issue in steganography?

Recently [14, 13], Afsari analysed the mathematics of *approximate* non-orthogonal joint diagonalisation. He concluded that the exact joint diagonalisation becomes particularly sensitive when noise is added to the 'clean' matrices, or when the sought joint diagonaliser is close to being ill-conditioned. He showed that the problem is particularly severe in the case that the number of matrices used is small and high dimensional. We also note that this issue of sensitivity surrounding joint diagonalisation in ICA methods was also previously discussed by Hori and Manton [15] in the context of a critical point analysis.

We reproduce the principal results here [13]. Specifically, consider the case of *estimating* the cumulants from observed data. Then the cumulant matrices are approximate.So consider the perturbation problem where the true cumulants are perturbed by uncertainties

$$C_i = A\Lambda_i A^T + \alpha N_i \quad \alpha \in [-\delta, \delta]$$

where Λ_i are diagonal, N_i are symmetric error noise matrices and α denotes a small noise contribution. For $\alpha = 0$ the problem has an exact joint diagonaliser of A^{-1}. However for $\alpha \neq 0$ we have the approximate diagonaliser:

$$B \approx (I + \alpha\Delta)A^{-1}$$

where $\Delta \in \mathbb{R}^{n\times n}$ and $diag(\Delta) = 0$. $||\Delta||$ measures the sensitivity of the joint diagonalisation problem to noise.

For an orthogonal joint diagonaliser, it can be shown that

$$\Delta_{kl} = \frac{S_{kl}}{\sum_{i=1}^{N}(\lambda_{ik} - \lambda_{il})^2}$$

where $S = -\sum_{i=1}^{N}[(A^T N_i A)^o, \Lambda_i]$, with $[X, Y] \equiv XY - YX$ and $X^o \equiv X - diag(X)$. The term in S provides the amplification of the noise, and since A is assumed orthogonal, this term is finite.

However, for a non-orthogonal diagonaliser, a perturbation analysis shows that Δ satisfies

$$M_{kl}\Delta_{kl} = \mathcal{T}_{kl}, \;\; 1 \leq k < l \leq N$$

where $\mathcal{T} = -\sum_{i=1}^{N}(A^{-1}N_i(A^{-1})^T)^o\Lambda_i$ and

$$M_{kl} = \gamma_{kl}\begin{bmatrix} 1/\eta_{kl} & \rho_{kl} \\ \rho_{kl} & \eta_{kl} \end{bmatrix} \; 1 \leq k \neq l \leq N$$

and

$$\gamma_{kl} = (\sum_{i=1}^{N}\lambda_{ik}^2)^{1/2}(\sum_{i=1}^{N}\lambda_{il}^2)^{1/2}, \;\; \eta_{kl} = \frac{(\sum_{i=1}^{N}\lambda_{ik}^2)^{1/2}}{(\sum_{i=1}^{N}\lambda_{il}^2)^{1/2}}$$

and

$$\rho_{kl} = \frac{\sum_{i=1}^{N}\lambda_{ik}\lambda_{il}}{(\sum_{i=1}^{N}\lambda_{ik}^2)^{1/2}(\sum_{i=1}^{N}\lambda_{il}^2)^{1/2}}$$

The point about the above is that, the noise contribution of Δ is magnified by $\mathcal{T}$ which can be very large if A is ill-conditioned, since $\mathcal{T}$ is controlled by the estimated unmixing matrix which can have large errors. So a small perturbation in the estimated cumulant matrices can lead to large errors in the joint diagonaliser.

In the context of steganography and an attacker trying to discover the hidden message, the attacker knows everything about the method but is faced with trying to reconstruct a joint diagonaliser from similar but not identical data to the ones used to construct the original unmixing matrix $\mathbf{W}$. This matrix is large and there are only a few matrices (this depends on the optimisation algorithm used, but many blind source separation algorithms only consider low order cumulants and hence only a small number of matrices are approximately diagonalised) and hence the problem is ill conditioned.

This is not the case for fixed feature space methods which do not aspire to independence, but only to orthogonalisation for example (such as PCA or Fourier methods). Hence we would expect to see a difference in the ability of an attacker to recover a secret hidden message in a system using statistically independent basis functions versus a system using orthogonal basis functions to hide the message.

We now demonstrate this difference.

5. TIME SERIES EXAMPLE

We show how gradual increases in the magnitudes of perturbations to a signal lead to smooth changes in the condition number of the relevant covariance matrices (as would be expected), but at a critical threshold the ability to recover a random message abruptly switches for the method based on independent components for small perturbation values. This is due to a step-function change in the structure of the separating matrix. It will be shown that much larger perturbation values are required for an orthogonal basis vector expansion method,hence an attacker will not need to estimate the secret key separating matrix as exactly for the orthogonal method as opposed to the independence method.

The experimental protocol is as follows:

1. A one-dimensional EEG signal $\mathbf{c}$ of twenty seconds duration is considered as the cover work. The EEG time series is transformed into a matrix of observation vectors, $\mathbf{X}$ using the dynamical embedding method described in [12]. (Hence the finite observation vectors $\mathbf{x}_i$ are overlapping windows of samples from $\mathbf{c}$.)

2. The ICA approach is used to decompose $\mathbf{X}$ into a set of independent sources $\mathbf{S}$ and a separating matrix $\mathbf{W}$. $\mathbf{W}$ is used as one of the secret keys needed to retrieve the watermark.

3. One of the sources, $\mathbf{s}_{wm}$ thus obtained is watermarked using the QIM method of message embedding. (the QIM method is not crucial to the argument. Other message embedding approaches can be used.) Let $\tilde{\mathbf{s}}_{wm}$ represent the watermarked source. $\tilde{\mathbf{S}}$ represents the watermarked source matrix.

4. The watermarked EEG $\tilde{\mathbf{c}}$ is reconstructed from $\tilde{\mathbf{X}}$ which is obtained by applying the inverse of the separating matrix, $\mathbf{A} = \mathbf{W}^{-1}$ to $\tilde{\mathbf{S}}$.

5. To estimate the sensitivity of the ICA method $\mathbf{c}$ is perturbed by a zero mean random noise signal, η to obtain $\bar{\mathbf{c}}$. The variance of the noise signal represents $\epsilon_{\mathbf{c}}$ and several increasing values of noise power are used to simulate the effect of an attacker using different data to estimate the separating matrix.

6. $\bar{\mathbf{c}}_l$ obtained for each value of $\epsilon_{\mathbf{c}}$ is transformed into a matrix of observation vectors, $\bar{\mathbf{X}}_l$ where l indexes the different noise levels.

7. The condition number of the covariance of $\mathbf{X}$ and each $\bar{\mathbf{X}}_l$ is calculated.

8. The norm of the difference between $\mathbf{X}$ and its perturbed version $\bar{\mathbf{X}}_l$, ξ is calculated.

9. The ICA is applied to each $\bar{\mathbf{X}}_l$ to obtain a set of modified 'best-guess' sources $\bar{\mathbf{S}}_l$ and separating matrices $\bar{\mathbf{W}}_l$. The ICA is initialised using the eigenvectors of the covariance of $\mathbf{X}$ in order to obtain the same order of the estimated sources.

10. The condition number of the covariance of $\mathbf{W}$ and each $\bar{\mathbf{W}}_l$ is calculated.

11. The norm of the difference between $\mathbf{W}$ and its perturbed version $\bar{\mathbf{W}}_l$ is calculated.

12. An estimate of the sources $\bar{\mathbf{S}}_w$, $\bar{\mathbf{S}}_l$ is obtained by applying $\mathbf{W}$ and each $\bar{\mathbf{W}}_l$ respectively to $\tilde{\mathbf{X}}$. An estimate of the embedded message is retrieved from $\tilde{\mathbf{s}}_{wm}$ of $\bar{\mathbf{S}}_w$ and each $\bar{\mathbf{S}}_l$.

13. The Hamming distance between the original embedded watermark and the estimated watermark is noted.

The same protocol is repeated, except that rather than use the ICA framework to construct basis vectors, we use a standard PCA approach applied to the same observation matrices $\mathbf{X}$, and the same quantities calculated for comparison.

6. RESULTS

Figure 1 indicates the monotonically increasing change in the observation matrix $\bar{\mathbf{X}}$ with respect to the original matrix as the noise power on the original cover time series is increased. This is an attack which represents the attacker using very similar but not identical data to estimate the only remaining unknown: the separating matrix (or the matrix of PCA coefficients). Distorting the cover will induce a matrix norm distortion in the data matrices used in the ICA algorithm. For smooth and small distortions, if the overall method is *not* sensitive, then the message should be recoverable since approximate knowledge should be sufficient.

We would naively expect that the message would remain recoverable to an expert attacker until the noise power significantly distorted the original cover.

Figure 2 shows the change in structure of the PCA coefficient projection matrix and the ICA separation matrix as the noise power is increased. The flat structure of the PCA approach is due to the whitening effect and the orthogonalisation which preserves the structure. However the separating matrix changes its structure for quite small values of noise power, as exhibited by the increase in the matrix norm. We would expect this change in matrix structure to be reflected in the ability of an attacker to recover the hidden message.

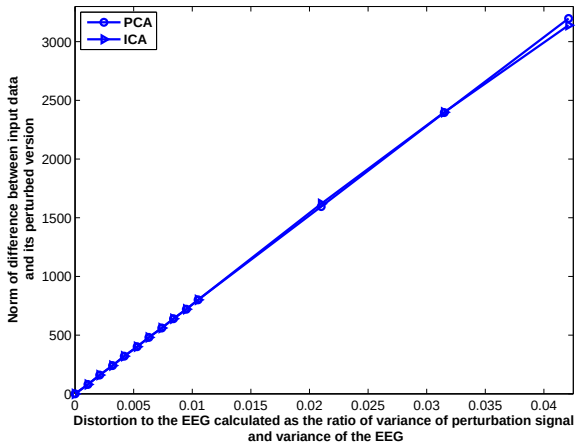

Fig. 1. Change in the perturbed input data matrix $\mathbf{X}$ due to increased noise power on the original cover signal for both the PCA and ICA experiments. The slight differences are due to different random realisations of the noise.

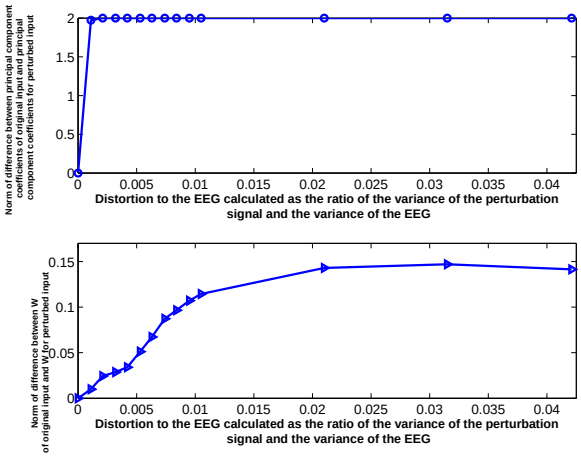

Fig. 2. Matrix norm of the difference between the original and the perturbed coefficient projection matrix for PCA and the separating matrix for ICA due to increasing noise power on the original cover signal. The PCA difference remains flat. The ICA difference increases showing a change in structure of the separating matrix. This is expected for small noise.

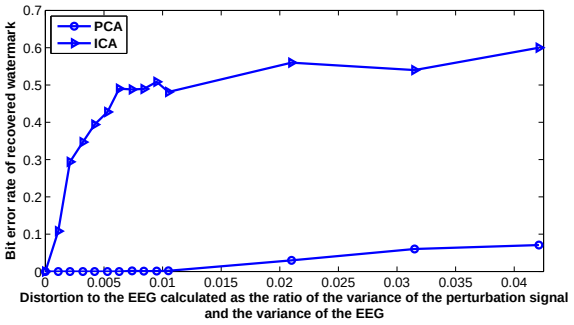

Fig. 3. The probability of bit error between the estimated recovered message as determined by the expert attacker and the original message. Note the significant difference between the ICA and PCA approaches. The message becomes randomly unintelligible when the added noise power increases beyond a small threshold. The message embedded in the PCA vectors remains recoverable for much higher values of noise which equates to a much less accurate estimation of the PCA projection matrix.

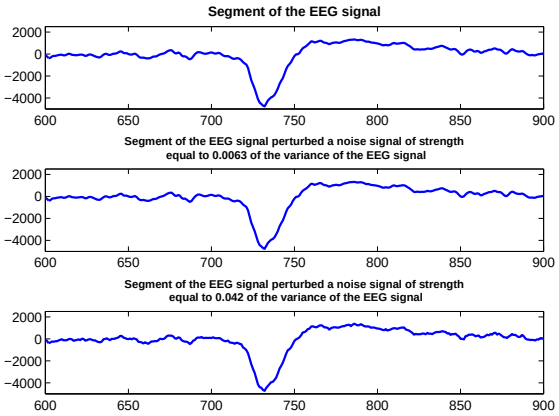

Fig. 4. This figure illustrates the very small distortion considered in this experiment. The figure plots a small segment of the original EEG time series and compares with the distorted EEG. The middle figure is for a variance where the ICA error becomes randomised, but the PCA error remains at zero. The bottom figure reflects the EEG distortion for higher noise power when the PCA method starts to deteriorate,

Figure 3 shows the probability of bit error of the estimated message compared to the original random message embedded in the clean data. This figure depicts the marked contrast between the orthogonalised basis vector approach and the independence basis vector approach. We note that at the point where the structure of the separating matrix is changing, the message basically becomes randomly correlated with the true message: the attacker is unable to recover the message for the independent vector approach.

Figure 4 is for reference only, indicating that for the noise power we are considering sufficient to destroy the hidden message, the structure of the original cover (EEG time series) is essentially unchanged. As far as a clinician's perception is concerned, no physiological characteristics of the EEG have been destroyed at this level of signal degradation. Hence, the attacker would need to use data at least as accurate as the distorted data shown in this figure if she is to have an success at recovering the embedded private information by using the data to estimate the separation matrix.

These results support the theoretical discussion in the early part of the paper. We had hypothesised that a steganographic method which relied on the non-orthogonal joint diagonalisation of a small set of matrices would have an inherent sensitivity in that it would be ill-conditioned. We have now observed this effect for the ICA approach used in this one-dimensional time series example. In addition, we also surmised that a method which relied on a secret key derived from a constraint of orthogonalised basis vectors would not have the same sensitivity. This is also shown in the comparative example of the PCA approach. (Note, the approach based on fourier transforms is trivially open to an attacker since it is not data-derived and the attacker is assumed to have all knowledge of the method).

Although this paper has not discussed other features of the nonorthogonal independent basis vector expansion of signal space, we have previously argued that it should also have better tradeoffs of data rate robustness and imperceptibility in the trade-off triangle.

This additional feature of sensitivity to estimation can be used as a security measure for privacy protection, or alternatively considered as a better forensic tool in that we can also embed different messages in different independent components. We can choose these independent components for their stability characteristics. Embedding last-access information of users who may have opened a digital document can now be made more secure by using a private key based on the separating matrix of the ICA approach. As we have demonstrated, the approximate non-orthogonal joint diagonalisation is expected to be difficult in the ICA example, and hence even an expert attacker is unlikely to be able to estimate the separating matrix accurately enough to recover and therefore modify the embedded message without destroying the cover itself.

7. CONCLUSION

We have discussed how the observed sensitivity of the ICA method derived from the problem of joint diagonalisation of estimated matrices leads to a sensitive estimation of the unmixing matrix $\mathbf{W}$. Without almost exact knowledge of the precise original data used to construct the true unmixing matrix, the estimation of the matrix is an ill-conditioned problem. Since in our method, the unmixing matrix is used as a key for extraction of the hidden message, the inability of an attacker to estimate this key accurately is sufficient to prevent recovery of the message. This system allows for Kerckhoff's principle, in that the attacker is allowed to know everything about the method except the key, and we have seen that the key cannot be accurately estimated enough even with knowledge of how it was generated.

It was seen that for the orthogonal basis vector approach as typified by the PCA method, the message could be recovered with a much less rigorous 'error bar' on the corresponding projection matrix of PCA coefficients. However, for the method based on independent basis vectors, a very accurate estimate of the separating matrix was required. When this threshold of matrix difference norm was crossed, the degradation of the message was a sharp transition, and not one of gradual degradation.

The paper is a preliminary investigation only. We have not explored all variants of independent component analysis or alternative algorithm variants which might be more stable to the perturbation sensitivity identified here. Similarly we have not investigated specific attacks which might compromise the new suggested security mechanism proposed here. These are topics for the future.

We have been motivated by this use of the ICA method as a steganographic framework for exploitation in the future electronic patient healthcare record, for security of personal and confidential patient information likely to be embedded inside the actual raw medical data, and also for possible forensics of retaining a record of last access logs on data files. However the method has a much wider generality of application.

8. REFERENCES

[1] B.R.Matam and David Lowe, "A signal processing approach to countering the security of dm-qim as a steganographic principle," in *Eighth IMA International Conference on Mathematics in Signal Processing*, 2008.

[2] B.R. Matam and David Lowe, "Watermarking: How secure is the DM-QIM embedding technique," in *Proceedings DSP2009*, 2009, p. These Proceedings.

[3] P. Bas and J. Hurri, "Vulnerability of DM watermarking of non-IID host signals to attacks utilising the statistics of independent components," *IEEE Trans. Information Forensics and Security*, vol. 153, no. 3, pp. 127 –139, 2006.

[4] Francois Cayre and Patrick Bas, "Kerckhoffs-based embedding security classes for WOA data-hiding," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 1–15, 2008.

[5] I. J. Cox, M. L. Miller, and J. A. Bloom, ," in *Digital Watermarking*. Morgan Kaufmann Publishers, 2002.

[6] F. Cayre, C. Fontaine, and T. Furon, "A theoretical study of watermarking security," *Proc. Information Theory (ISIT)*, pp. 1868 –1872, 2005.

[7] S. Bounkong, B. Toch, D. Saad, and D. Lowe, "ICA for watermarking," *J. Machine Learning Research*, vol. 4, no. 7-8, pp. 1471–1498, 2004.

[8] C. Jin and L. Pan T. Su, "Multiple digital watermarking scheme based on ICA," *Proc. IEEE* 8^{th} *Int. Workshop on Image Analysis for Multimedia Interactive Services*, pp. 70–73, 2007.

[9] A. Hyvarinen, J. Karhunen, and E. Oja, ," in *Independent Component Analysis*. Wiley-Interscience, 2001.

[10] B. Toch, D. Lowe, and D. Saad, "Watermarking of audio signals using independent component analysis," 3^{rd} *International Conference on WEB Delivering of Music (WEDELMUSIC'03)*, pp. 71–74, 2003.

[11] B. Toch and D. Lowe, "Watermarking of medical signals," *Proc.* 2^{nd} *Int. Conf. Computational Intelligence in Medicine and Healthcare*, pp. 231–236, 2005.

[12] B. R. Matam and D. Lowe, "Steganography, biopatterns and independent components," *Proc.* 7^{th} *Int. Conf. Mathematics in Signal Processing*, pp. 206–209, 2006.

[13] Bijan Afsari, "Sensitivity analysis for the problem of matrix joint diagonalization," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1148–1171, 2008.

[14] Bijan Afsari, "What can make joint diagonalization difficult?," in *Proceedings ICASSP2007*, 2007, pp. III–1377 – III–1380.

[15] Gen Hori and Jonathan H. Manton, "Critical point analysis of joint diagonalization criteria," in *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, April 2003, pp. 1095–1100.