Information extraction methods and the use of power spectra

Mair Allen-Williams
s0454371

December 3, 2004

Abstract

We discuss ways of extracting information from data signals which have missing or incomplete data, or when the information is hidden. We describe independent component analysis, and its limitations, and then focus on techniques using the Fourier transform of the data or the power spectrum.

1 Introduction

Given a signal representing some data, it is frequently the case that in some way the signal doesn't explicitly encode the information we need to extract from it. The signal may be a combination of several input signals we'd like to separate (the "cocktail party" problem), it may have missing data (for example, audio sources taken from scratched media), or it may have been quantised at some point (as in JPEGs). In general, we face a problem of deciding which of several possible underlying sets of data could have generated the available signal. There are a number of techniques available for doing this, ranging from highly specialised and domain specific to statistical methods which try and seek patterns in the data without using external or domain-specific knowledge (although some prior information or assumptions may be supplied). We focus on the latter, looking for methods which are extensible or as generally applicable as possible.

In general, the aim is to find some kind of structure, or pattern, in the data available in order to determine which of the possible sets of underlying data was most likely. This usually means encoding some kind of prior assumption into the algorithm. One way of doing this is independent component analysis (ICA). Leino [15] describes the work done by Aapo Hyvärinen and Erkki Oja [12] on this, referring to the source separation (or cocktail party) problem. Roweis [16] points out that this needs a number of samples, or sensors, and considers an approach when there is only one sensor. Computer scientists taking advice from nature is nothing new, and Roweis notes that amongst other things, humans use energy spectra when "solving" the cocktail party problem. He discusses one technique for doing this. Bach and Jordan [3] suggest another. Moving on to cases where data is missing, Storkey [19] describes a technique based on the fast Fourier transform—recall that the power spectrum of a signal is the square of its Fourier transform, and so techniques which can be applied to Fourier transforms could also be applied to power spectra. Gregory [8] describes a slightly different approach to Bayesian analysis of a discrete Fourier transform. Storkey's techniques are based on the use of spectral priors, and in a joint paper with Allan [18] he discusses the use of such priors in reconstruction of quantised data.

2 Independent Component Analysis

Leino Leino [15] gives a mathematical overview of the ICA algorithm. Given a vector $\mathbf{x}$ of observations, we can write $\mathbf{x} = \mathbf{As}$. $\mathbf{A}$ is a *mixing matrix*; for example in the cocktail party problem it might depend on the distances of the speakers from the sensor. The components of $\mathbf{s}$ are the independent components, the original signals we wish to extract. ICA consists of finding $\mathbf{A}$ and $\mathbf{s}$, subject to the constraint that the components of $\mathbf{s}$ are independent. Given a collection of observations, $\mathbf{x}$, we can treat the x_i as random variables, so that what we are trying to find is a probability distribution for the s_i, given the distribution that the observations appear to be a sample of.

The independence constraint is not sufficient to uniquely determine A and s; in particular, the variances of s cannot be determined, since any multiplier of an s_i can be cancelled by dividing column i of $\mathbf{A}$ by the same value. Further, it is impossible to determine an order for the independent components; the columns of $\mathbf{A}$ can be reordered along with any reordering of s. We set the variances to be 1 for simplicity, but note that we still cannot determine the sign of the vector s. We can simplify calculations further by translating the data linearly so that its mean is zero.

There is one final constraint; the independent components may not have Gaussian distributions, for the practical reason that this results in a completely symmetric distribution, and it would be impossible to make any inferences about the directions of the columns of A. This practical reason is supported by the principle that the less Gaussian the components are, the more independent they are. This result is a consequence of the central limit theorem, from which we can infer that the sum of two independent random variables has a more Gaussian distribution than the original variables. We use this principle to seek the independent components, attempting to find those components which are as non-Gaussian as possible.

Leino describes a number of measures of non-Gaussianity. Kurtosis, based on the fourth moment, is simple, but non-robust. Negentropy, based on the information-theoretic notion that a Gaussian variable has the largest entropy (carries most information) among all random variables with the same variance, is accurate, but difficult to compute. Some approximation combining kurtosis and negentropy is frequently used.

An example of ICA operating on sound signals is shown in figure 1. The FastICA[1] code package was used for the example.

Hyvärinen and Oja Leino's coverage of the mathematical principles is mostly accurate. However, Leino's overview does not include examples or experiments, although he refers to some of the uses of ICA. For more examples, we look to Hyvärinen and Oja's work [12]. Hyvärinen and Oja provide examples from medicine—MEG data, cashflow analysis, and image denoising. They do not provide or describe quantities of test data, merely some specific pictorial examples which clearly demonstrate the flexibility of the technique.

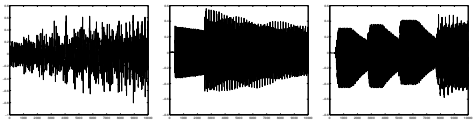
Limitations ICA is a powerful technique. It does not rely on any prior knowledge of the data or the probability distributions involved. However, it can only be used if the sources are indeed non-Gaussian, and if they are independent. Where there is dependence between signals ICA may find the underlying signals which give rise to the dependent signals (figure 3). We could consider what happens with more complex dependencies. An example is shown in figure 4; the analysis makes a brave attempt but fails to detect the original harp signal. This could cause problems using ICA when we incorrectly assume independence.

Note also that the algorithm described above does not take noise into account. The algorithm can perform with a small amount of noise, but the independent components it finds are also noisy (figure 2). Hyvärinen and Oja [12] discuss some ways of denoising data. One in particular which takes advantage of the statistics of the data is *Sparse Code Shrinkage* [11], which is a maximum likelihood estimation, closely related to ICA (the same assumptions about the data are made as in ICA, with the addition of noise).

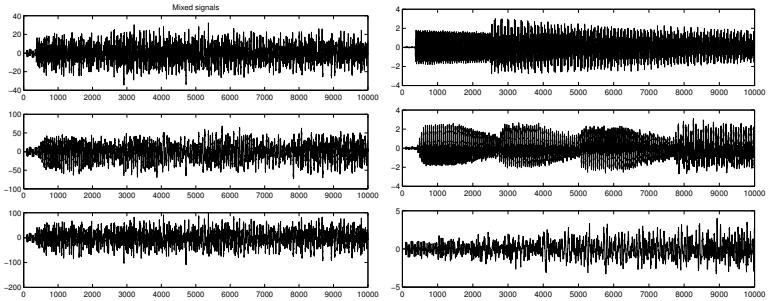
Another limitation of ICA is the necessity of providing sufficient sample data to extract useful statistics - in general at least N samples are needed if N signals are to be separated via ICA. Choi, Cichocki and Belouchrani [6] describe a technique which estimates A in the presence of white noise, and which works even if the sources are Gaussian, provided that the signals are temporally correlated with time varying variances. They present experimental data showing the success and robustness of their method on audio signals; comparing it against a number of existing algorithms and demonstrating that it is at least as effective as existing algorithms in the presence of several Gaussian signals, and more robust than the existing algorithms in the presence of noise. As mentioned earlier, power spectra can form the basis of a powerful alternative technique, relying on some prior information, but only requiring one sensor. We discuss this below.

Other Work Knuth [14] derives an ICA algorithm from Bayesian principles, and demonstrates how the Bayesian approach can be used to incorporate prior information. Wong et al. [20] describe and compare

[1] http://www.cis.hut.fi/projects/ica/fastica/

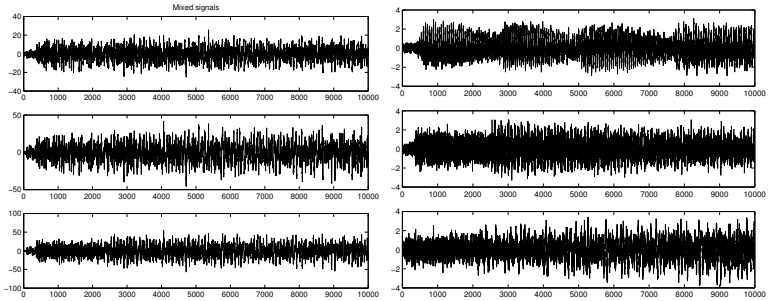


(a) Source signals: Left, a harp; centre, a tinkling sound; right, extract from Beethoven's fifth symphony



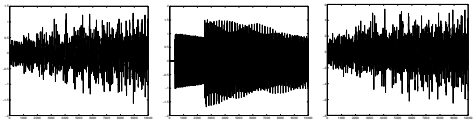
(b) Left: Signals mixed with a random mixing matrix giving three output signals. Right: Independent components from the mix. The characteristics of the individual signals are clear,

Figure 1: Independent component analysis of music

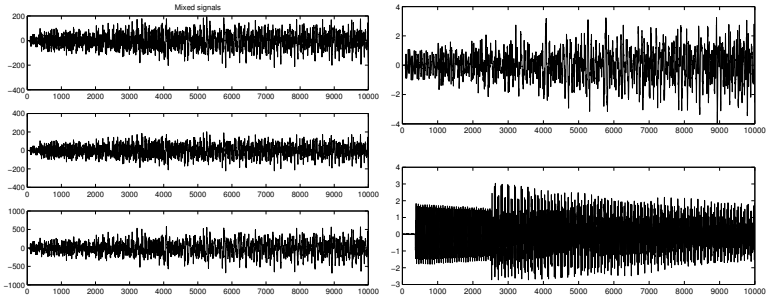


(a) Left: The same musical signals mixed with a random mixing matrix giving three output signals, with Gaussian noise of mean 0 and variance 1 added. Right: Independent components from the mix. The characteristics of Beethoven's fifth are still visible.

Figure 2: Independent component analysis with Gaussian noise

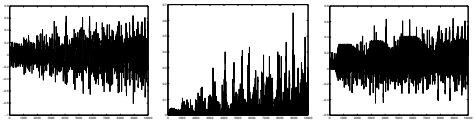


(a) Source signals: Left, a harp; centre, a tinkling sound; right, combination of harp and tinkle

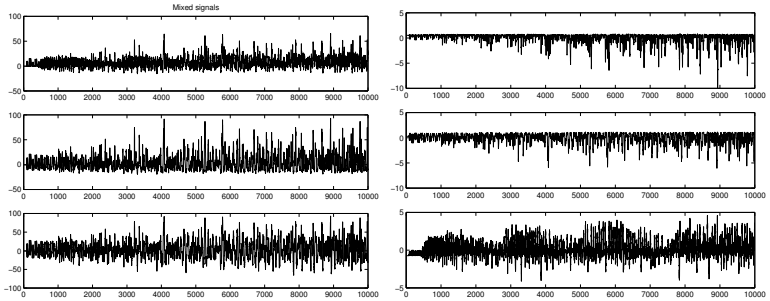


(b) Left: Signals mixed with a random mixing matrix giving three output signals. Right: Independent components from the mix; the harp and the tinkle.

Figure 3: Independent component analysis of dependent components



(a) Source signals: Left, a harp; centre, the square of the harp signal; right, Beethoven's fifth mixed with harp



(b) Left: Signals mixed with a random mixing matrix giving three output signals. Right: Independent components found by ICA.

Figure 4: Independent component analysis of dependent components

several algorithms for performing ICA, and compare them with principal component analysis. Zibulevsky and Pearlmutter [21] describe a separation algorithm based on sparsity of time signals.

3 Cocktail party using power spectra

3.1 Roweis

Roweis [16]'s work focuses on *refiltering*—extracting a single speaker from the sample. One can repeat this process as many times as necessary to extract all the speakers. This method requires training on the individual speakers before they can attempt to demix, and only applies to separation of known speakers. Roweis also uses heuristics—"cues" based on features of speech. Examples of generic cues include continuity (nearby points are likely to belong to the same signal) and common fate cues (various elements such as offset/onset which have common time variation). Speech-related cues include pitch estimation. In addition to hard-coding these features, Roweis discusses the possibility of unsupervised learning of such things. Although such methods could not be used on data which didn't have any features of this type, it seems likely that unsupervised methods could be used to learn features of types of data other than auditory, this is not discussed.

The primary technique in Roweis' system is the use of hidden Markov Models, trained from individual speakers. The mixed speaker model, called a *factorial hidden Markov Model*, consists of a combination of independent individuals. A key part of using this model is the observation that the log magnitude spectrogram of a mixture of sources is very nearly the elementwise maximum of the original spectrograms. This holds if the original spectra are fairly distinct; becoming less true as they overlap more.

3.2 Bach and Jordan

Bach and Jordan [3] describe an approach using a *spectrogram*, a redundant representation constructed from Fourier transforms of sections (windows) of the time-frequency diagram. The idea is to take this spectrogram and segment it into R disjoint subsets. Having done this, it is possible to find R signals such that each of the R spectrograms corresponds to one of the signals. In general, there is no one solution, owing to the redundancy in the spectrogram representation. The classical solution is to minimise the L_2 norm. Bach and Jordan adopt this solution without further discussion of its merits.

In order to perform the segmentation, a number of cues, of the kind described above, are used to construct a feature map, which takes the same shape as the spectrogram. By combining cues, segmentation can be performed effectively. Bach and Jordan describe how to build orthogonal affinity matrices based on these feature maps in order to define a segmenter. Such matrices tend to be very large, so for computational practicality it is necessary to approximate the matrix in some way.

This approach requires a quantity of training data taken from individual speakers Unlike Roweis' approach, the individuals need not be the speakers whose voices will be heard in the mix.

Bach and Jordan's method has some limitations—for example, it is necessary that speakers have distinct pitches, and that one pitch shouldn't be too close to twice another. The cues they use rely on the fact that they are operating on speech—in other words some domain-specific knowledge is built in. It should be possible to define sets of cues for other domains, and have the method work just as effectively. Although their techniques only need one sensor, the segmenter needs training data to run—still, it's to be expected that we have to feed some kind of information in to get anything useful out. They don't discuss how much training data is needed to construct an effective segmenter. Of course, this can to some extent be written off as a one-off cost, since the same data can be used for a whole series of similar separation problem.

Another limitation of Bach and Jordan's work is that it only applies to the separation of two signals. Because of the reliance on speech-related cues, it's not obvious that this can be applied to more than two sources by separating off one at a time. However, it may be possible to extend the technique to separate all the speakers at once. They also assume ideal conditions, in particular, no noise. Various de-noising solutions exist which can be pre-applied to the data, but as mentioned in the discussion of ICA, solutions which use features of the data would be preferred.

Bach and Jordan's methods are still computationally heavy and resource intensive—as an example, they took thirty minutes to separate four seconds of speech on a 1.8 GHz processor, with 1GB of RAM.

Comparison These spectrogram models have the advantage of only requiring one sensor; important for many real-life applications where several sensors may not be available. This advantage, however, is mitigated by the fact that both require training on individual speakers. Bach and Jordan's model, although it must be trained on some set of speakers, need not be trained on the same speakers as it is attempting to distinguish. Since human speech contains some random fluctuations, it seems likely that Roweis' method should be able to achieve some results given unknown speakers, especially if their voices are 'similar' to the known speakers. Experimentation would be interesting—if the speakers are very different from the input, can Roweis' technique achieve anything at all? It seems possible that the model, based on distinct Markov chains, would break down completely if a signal seemed to be "hopping" from one chain to the other. Roweis implies that since his techniques can involve almost entirely unsupervised learning, and training data is readily available, it is unnecessary to be concerned about this. Another shortcoming in Roweis' model is the use of the assumption that the log magnitude spectrogram of a mixture of sources is nearly the elementwise maximum of the original spectrograms, meaning that the system fail to work if the speakers are tonally similar.

Furthermore, both approaches rely to some extent on features which are specific to audio processing, hence cannot immediately be applied to other fields. By defining equivalent feature sets it should be possible to migrate the approach to the separation of, for example, astronomical signals. Roweis argues that a "good" learning algorithm given sufficient data should be able to learn these cues, so that the method could be extended to any field, given sufficient data.

Neither paper includes a great deal of quantitative test data, although both include figures showing the systems at work. It would be interesting to see Roweis comparing different kinds of Markov model and seeing what extent of similarity in the speakers can be tolerated, and how much the speakers can vary. As discussed, there is also scope for more example data from Bach and Jordan. It would be interesting, too, to see performance comparisons over different types of audio data—Roweis introduces his paper with an analogy of many pianos playing, but mainly discusses speech. Could the cries of babies be distinguished? The barks of dogs? As discussed earlier, extensions to other fields, both supervised and unsupervised, would also be of interest.

3.3 Further work

Cai, Lu, Zhang and Cai [5] describe a method of feature extraction based on wavelets. They also mention the use of principal component analysis on the frequency spectrum in de-noising. Godsill [7] discusses techniques for Gaussian noise reduction and signal enhancement based on Bayesian computations using a Markov chain Monte Carlo simulation. His techniques exploit the fact that the signal is highly non-Gaussian while the noise is Gaussian. Smaragdis [17] describes an approach based on information theory, operating in the frequency domain.

We can also compare these approaches to the human ear. Sitting at a dinner table, humans can usually pick out one speaker from those immediately nearby. In speech separation problems, human processing is able to use transition probabilities—the likelihood that one word will be followed by another—as well as lower level cues [1]. It seems that humans also use cues similar to those described by Jordan and Bach. A baby recognises its mother's voice long before it learns the meaning of words. Like Jordan and Bach, humans require training data.

4 Missing data

4.1 Gregory

In both the work of Roweis, and that of Bach and Jordan, a Fourier transform is used in moving from a frequency-time representation to an energy spectrogram, against which a feature map can be matched. In some cases, however, the traditional Fourier transform is impossible because there is missing data. An

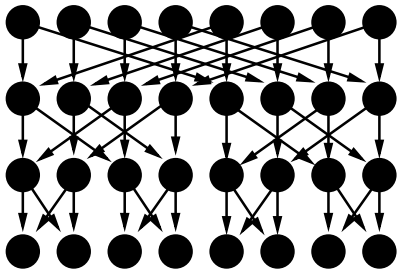


Figure 5: The belief network corresponding to the fast Fourier transform [19]

example is an audio signal emanating from a scratched medium. Gregory [8] describes a technique (due to Jaynes [13], in 1987) or estimating a Fourier transform with missing data, given prior probabilities. Jaynes' technique is based on a Bayesian derivation of the Fourier transform. Gregory describes various extensions to Jaynes' work. He refers to Bretthorst's work given strong prior information, and his own work on dealing with systems with no prior data. If there is no known prior information then we must at least assume a model; it is impossible to extract information from nothing. Gregory uses a family of histogram models to detect signals of unknown shape, backing up his work with examples from astronomy.

Gregory does not set out to discuss missing or inaccurate data, but rather to describe the reformulation of the traditional problem in Bayesian terms. Gregory's paper is mainly an overview, and so does not include details of the tests he refers to. However, he mentions several sets of results [9] [10] which seem to demonstrate that Bayesian methods are effective for spectral analysis.

Gregory's work does not say how well the system deals with noisy or missing data, although the astronomical examples would suggest that he can handle at least some noise. The work is clearly broadly applicable—he also mentions Bretthorst's applications within NMR [4], and one can imagine that it can be applied to any kind of signal. Gregory does not discuss the efficiency of his methods. In the next section we describe a technique for handling missing data, specifically designed with efficiency in mind.

4.2 Storkey

Storkey's treatment, also Bayesian, exploits the graphical structure of the Fast Fourier Transform (FFT), treating the FFT graph (figure 5) as a belief network. Probabilities can then be propagated through the network. Once again, we cannot analyse out of a vacuum. While the algorithms of Bach and Jordan or Roweis require individual signals to learn from, here we require prior information about the likelihood of the power spectra. There is no learning involved once we have these priors, and there is no domain restriction for the application of the algorithms.

Given a stream of data and the priors, Bayes' rule can be applied to estimate the likelihood of finding a particular data point at a particular point in a stream. Jordan and Bach mentioned efficiency as an important criterion in their analyses, struggling to find a balance between accuracy and efficiency. Storkey's graphical representation is intended to be efficient. He doesn't go into detail about the theoretical accuracy; the approach appears to work on real world data, but there is no discussion of borderline or pathological cases.

The main subtlety in this approach is deciding on a suitable belief propagation technique, particularly given that the graph contains loops. Storkey describes an iterative message-passing algorithm which he asserts (with the support of tests) tends to converge fairly well in practice, although convergence cannot be guaranteed. It is possible to resort to damping if the method doesn't seem to converge.

Storkey does not go into theoretical detail about the limitations of his approach, but provides tests and examples. In particular, because neighbouring nodes in the network are not in general directly connected to neighbouring input or output data, continuity features of the data may not be picked up on by this system. However, the approach worked well on some real world audio samples of laughter with missing data. Given that this approach is intended to be fairly generic, it would be interesting to see it tried on examples other than audio data—image data, for example, which contains different types of features. It may also be worth

trying to combine this approach with the feature map systems, considering whether there's some way of spotting a "gap" in a feature. If this were possible, it might help to compensate for the limitations of the belief propagation methods.

Although the FFT has good scaling properties, Storkey mentions that the number of iterations needed for the the belief propagation to converge depends on the data size, affecting the efficiency of the technique. His investigations suggest that it does not scale badly, but more tests will be needed to determine the scaling factors. He also recommends further investigation into propagation algorithms. Finally he suggests that priors on phase information or Fourier coefficients could be investigated. Further work could also be done on applying the technique to noisy data—estimating the likelihood of a particular point being erroneous, for example.

Accurate estimation of priors is also important here. Storkey doesn't discuss techniques for learning or inputting priors. We assume that they can be decided manually, or learnt (unsupervised) from "clean" data. It should also be possible to use the gappy or inaccurate data to learn priors—more useful in situations where pure data may not be available.

4.3 Other work

Bayesian analysis using spectral priors is used in other cases where there is missing or potentially inaccurate data. The JPEG algorithm uses the discrete cosine transform—a Fourier transform on an odd function. After performing the DCT, the data is quantised, sometimes causing artifacts at decompression time. Storkey and Allan [18] discuss how to use information about spectral priors on the quantised data to estimate the original DCT coefficients. Atwal and Bialek [2] discuss the use of priors given noisy data.

5 Conclusions

5.1 Summary

We describe the ICA technique for performing separation of signals and explain why it is not optimal in some cases. Inspired by nature, spectrogram-based approaches can be used. However, these techniques assume a complete data set. We describe a technique for estimating the missing data when performing a Fourier transform, based on Bayesian analysis. We also mention the use of Bayesian analysis in handling missing or noisy data in other cases.

Accurate testing of separation work is difficult—what's a quantitative measure of a recognisable speech signal? However, visual observation of results can usually determine whether a method is working or not. It is also possible to evaluate methods based on their applicability to other fields. ICA is generally applicable anywhere where the key assumptions (independence, non-Gaussianity, appropriate mixing matrix) are relevant. The separation techniques of Bach and Jordon and of Roweis (a) need training data (but if learning is unsupervised, one can simply acquire the relevant training data and feed it to the algorithm) and (b) use heuristics based on voice data; this could be extended to other fields by finding equivalent heuristics. One could also consider the effectiveness of the techniques without the use of such cues. Roweis' Markov model might work, whereas Bach and Jordan's algorithm relies on cues for separation. Storkey's FFT algorithms are generally applicable to anyone working with Fourier transforms. We describe a variety of applications for Gregory's Bayesian techniques.

5.2 State of the art and future work

Spectrogram-based approaches provide a good way of analysing signals. However, these approaches are frequently resource-intensive. Storkey investigates an efficient approach to Fourier-based analysis with missing data. In general the best approaches to the separation problem involve domain-specific knowledge. In the work described this takes the form of finding features in the data. Given sufficient data, unsupervised learning technqiues can be used to find good features so that pre-existing domain knowledge is not necessary.

Future work should investigate learning techniques for features and for spectral priors. Algorithms exist which handle noisy data, but they focus on de-noising so are not generally applicable for missing or quantised

data. More work could be done on combining algorithms since it is often beneficial to perform the de-noising or other analysis in combination with the signal analysis in order to make the best use of the available information. Storkey's or Gregory's work on estimating missing data could possibly be adapted to estimate the noise on the data. It might also be interesting to consider how Bayesian analyses of missing or noisy data could be extended to the signal separation problem, although there isn't a way of doing so directly. Further work should also be done on improving the efficiency of algorithms such as Bach and Jordan's, either by finding more efficient techniques or by determining good approximating assumptions. There is scope for creating new efficient algorithms such as Storkey's. In a slightly different direction, it would also be possible to investigate the extension of Storkey's FFT algorithm to two or more dimensions.

References

[1] Barry Arons. A review of the cocktail party effect.

[2] Gurinder S. Atwal and William Bialek. Ambiguous model learning made unambiguous with 1/f priors. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[3] Francis R. Bach and Michael J. Jordan. Blind one-microphone speech separation: A spectral learning approach, 2004.

[4] G. L. Bretthorst. Bayesian Analysis. I. Parameter Estimation Using Quadrature NMR Models. *Journal of Magnetic Resonance*, 88:533–551, 1990.

[5] Rui Cai, Lie Lu, Hong-Jiang Zhang, and Lian-Hong Cai. Improve audio representation by using feature structure patterns. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004.

[6] Seungjin Choi, Andrezej Cichocki, and Adel Belouchrani. Blind separation of second-order nonstationary and temporally colored sources. pages 444–447, 2001.

[7] S. Godsill. Bayesian enhancement of speech and audio signals which can be modelled as ARMA processes, 1997.

[8] P.C. Gregory. A Bayesian revolution in spectral analysis. 1997.

[9] P.C. Gregory. Bayesian periodic signal detection. I. Analysis of 20 years of radio flux measurements of the x-ray binary lsi+61 303. *Astrophysical Journal*, 1999.

[10] P.C. Gregory. Bayesian periodic signal detection. II. Discovery of periodic phase modulation in lsi+61 303 radio outbursts. *Astrophysical Journal*, 1999.

[11] A. Hyvärinen. Sparse code shrinkage: Denoising of nonGaussian data by maximum likelihood estimation. Technical Report A51, Helsinki University of Technology, Laboratory of Computer and Information Science., 1998.

[12] Aapo Hyvärinen and Erkki Oja. Independent component analysis: A tutorial. April 1999.

[13] E. Jaynes. Bayesian spectrum and chirp analysis, 1987.

[14] Kevin H. Knuth. A Bayesian approach to source separation. In *ICA99 Proceedings, Aussois, France*, pages 283–288, 1999.

[15] Antti Leino. Independent component analysis: An overview, 2004.

[16] Sam T. Roweis. One microphone source separation. In *NIPS*, pages 793–799, 2000.

[17] Paris J. Smaragdis. Information theoretic approaches to source separation, 1997.

[18] Amos Storkey and Michael Allan. Cosine transform priors for enhanced decoding of compressed images, 2004.

[19] Amos J. Storkey. Generalised propagation for fast Fourier transforms with partial or missing data. In *NIPS*, 2003.

[20] P.M. Wong, S.Choi, and Y. Niu. A comparison of PCA/ICA for data preprocessing in a geoscience application. 2001.

[21] Michael Zibulevsky and Barak A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4):863–882, 2001.