

Statistical Principles for Selection of Computer Vision Algorithms as Modules for Visual Perception - Show Me the Errors.

N. A. Thacker and P. A. Bromiley.

Last updated
16 / 11 / 2006

This document forms part of the **Recognition and Intelligence Series** available from www.tina-vision.net.

- 2007-001 Retinal Sampling, Feature Detection and Saccades:
A Statistical Perspective.
- 2006-008 Statistical Principles for Selection of Computer Vision Algorithms as
Modules for Visual Perception - Show Me the Errors.
- 1991-001 Designing a Layered Network for Context Sensitive Pattern Classification.
- 1997-002 Supervised Learning Extensions to the CLAM Network.
- 1996-003 Tutorial: Algorithms For 2-Dimensional Object Recognition.
- 1997-005 Speechreading Using Probabilistic Models.
- 2000-002 Solving Shape Based Object Recognition from a Computational Standpoint -
Practical and Physiological Constraints.
- 1995-004 Assessing the Completeness Properties of Pairwise Geometric Histograms.
- 1996-004 Robust Recognition of Scaled Shapes Using Pairwise Geometric Histograms.
- 1996-005 Multiple Shape Recognition Using Pairwise Geometric Histogram Based Algorithms.
- 2007-007 Automatic Identification of Morphometric Landmarks in Digital Images.
- 1999-002 A Feature Representation for Map Building and Path Planning.
- 2001-015 Colour Image Segmentation by Non-Parametric Density Estimation in Colour Space.
- 2001-006 What is Intelligence?: Generalised Serial Problem Solving.
- 1994-002 A Correlation Chip for Stereo Vision.
- 1995-001 Specification and Design of a General Purpose Image Processing Chip.



Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Abstract

This paper is intended as a discussion document and was written specifically for a workshop on Psychophysics and Computer Vision organised for the BMVA in November 2006. The views contained constitute a positional statement, outlining issues relating to identification of algorithms from the area of computer vision/pattern recognition as candidate models of human perception. The main motivation for this document is the observation that many problems tackled by those working in computer vision and pattern recognition can be seen as ill-posed (theoretically un-solvable). However intriguing this may sound, this can raise serious problems. This paper attempts to explore the consequences from the standpoint of conventional statistics.

*The work presented here focuses on attempts to extract meaningful low dimensional descriptions of high dimensional data. Such problems are considered key to topics such as machine/computer vision and more general aspects of data mining. The arguments enclosed are quite involved and assume more than a passing knowledge of standard methods (PCA, Kernel PCA). However, in this form it may serve as a consciousness raiser for those who are not aware of one important issue. Specifically, **it is incorrect to believe that data can generally be analysed in the absence of measurement errors.***

Any comments on the opinions presented here will be welcomed.

Introduction

It must be expected that ultimately the subject of computer vision should be able to explain how it is possible to extract meaningful information from images for practical applications. The majority of work done in this field is consistent with a modular approach, with multiple interacting components. Computer vision therefore has great potential to identify for us candidate models of human visual perception.

This raises the question of how we select specific algorithms as module candidates. It makes logical sense to start by assuming that evolution has endowed us with the ability to achieve any task up to the limits of the available data¹. This is just another way of saying that our interpretations of the world should be based upon quantitatively justifiable (statistical) processes. Researchers from areas other than computer vision, or new to this area, might initially expect that all algorithms should be intrinsically statistically valid. Unfortunately this appears not to be the case. Rather, most algorithms have (at best) a limited scope for valid application to real world data. In addition, even with good design methodology, it is difficult to get isolated modules to achieve a level of performance and reliability which matches human capabilities. It is reasonable to suggest that we need to start constructing systems of interacting modules in order to begin to do better. The case for suggesting that combination of modules also requires a statistically motivated approach has been made elsewhere [21]. To assume that solutions need not have statistical validity immediately requires us to accept any approach to data analysis as a candidate module, regardless of how sub-optimal it may be at extracting the required information. Aside from the immense number of candidate algorithms such lack of criticality allows, this also prevents us from later tackling very important scientific questions. Such as; how close does a given biological system (inevitably an approximation) come to solving a problem ‘correctly’? Also; does human visual perception involve making assumptions about the world which are entirely justifiable?

The suggestion here is simply this. We should attempt to identify candidate vision modules by selecting approaches which are known to have intrinsic statistical validity. These modules should make explicit the information required for a particular task and the accuracy with which any result can be obtained. Such predictions will then be useful for the construction of tests which support or refute the existence of equivalent computational modules in the human brain². If not in algorithmic detail at least in statistically equivalent form. Notice, that the statistical constraint does not require us to identify biological mechanisms. It is a self standing principle which we should be able to apply irrespective of the details of how we will ultimately interpret the computations supported by biological tissue. As an echo of Marr’s seminal work in the subject, once we understand what is needed, then we can consider how these modules are constructed in biology.

Computer vision is necessarily a broad research topic, encompassing tasks as disparate as stereo, object recognition, motion analysis, etc., but it is widely accepted that at its core is the need to extract meaningful low dimensional descriptions of data from a high dimensional set of visual observations (ie: temporal sequences of images). Recently, there has been considerable excitement regarding solution of these problems in

¹Optical illusions provide us with ample evidence that human vision systems are not perfect, however it is hoped that these illusions are created due to inappropriate assumptions, which under normal circumstances would be valid.

²I accept here that there will be difficulties in designing psychophysical tests which will lever directly on the capabilities of individual modules.

real world applications, such as image retrieval. Much of this excitement has been generated by the ability to solve relatively large scale problems using relatively small amounts of computation on serial computers. In particular, techniques such as Kernel PCA [16], and SIFT [10] have proved to be highly popular. The fact that these solutions may or may not be currently in a serial form is unimportant. What is important is understanding their theoretical limitations. The following is therefore intended partly as a critical analysis of such ‘fast’ solutions from the standpoint of statistical validity.

This document will analyse the process of non-linear component analysis as a specific example of the technologies required for the construction of a modular vision system. In the process we will define the objectives of sought solutions with regard to statistical feasibility. We will then use the conclusions of this analysis to help understand the reasons for the breadth of literature and the principles we may apply in order to select viable (in this case also valid) approaches. As a consequence we hope to convince the reader that popular approaches in computer vision and pattern recognition cannot be relied upon to be a good starting point for understanding human perception. This suggests that a concerted effort is needed to identify those algorithmic approaches which would be relevant for this task.

Density Analysis of Non-Linear Data Distributions: The Questions

Cluster analysis and non-linear component analysis (eg: Factor Analysis ³, Latent Variable Analysis and Non-Linear PCA) have been a major research topic in the area of pattern recognition for decades. This raises the questions;

Why are there so many publications in this area?

Could any published pattern recognition method be considered as a valid solution?

Only when we have the answer to these questions will it be possible to identify a limited number of solutions as candidate models of visual perception. In order to address these questions, this document defines a prototypical data analysis problem for which there is a well defined statistical solution. Generalisations away from this simple problem can then be assessed according to the consequences of changes in data properties.

Problem Definition

The first observation which needs to be made regarding the published literature is that data is generally not regarded as being generated by a known measurement process. Rather, particularly in pattern recognition, it is often taken as an implicit assumption that solutions should be applicable to data from any source, and with unknown measurement characteristics. In my opinion, in taking this stance the pattern recognition researcher might be replacing a difficult practical problem with a logically impossible one. The following analysis explicitly defines the quantitative process of measurement and subsequent statistical analysis. This is done to help us gain an understanding of the aspects of data interpretation which are affected by ignoring measurement errors.

We define the process of data space construction as a vector v resulting from a measurement of n data variables with uniform IID Gaussian noise. The data is assumed to have originated from a fixed but (mathematically) unknown physical process with m degrees of freedom, so that vector measurements in the data space are constrained to lie on a manifold within the n dimensional data space which has a local intrinsic dimensionality of m . The problem we seek to solve is that of identifying a set of m variables in order to more concisely characterise the signal content of the data. This may be done for a variety of reasons, including;

- **(A)**, so that the noise component can be discarded (filtered),
- **(B)**, for the reduction of problem complexity (by dimensional reduction),
- **(C)**, for the construction of Bayesian decision systems,
- **(D)**, in order to help in the location of maximal densities of samples with which to characterise the data (eg: clustering of temporal behaviours),

³In this document I will break with conventions of grammatical English slightly and adopt the policy of referring to all analysis methods as real nouns. This is to try to avoid the potential confusion which can arise when taking these words to have the conventional spoken meaning.

- finally (**E**), to construct a meaningful definition of similarity between points on this manifold.

We should accept the possibility at this point, that although we can write this list there may be circumstances under which meaningful solutions do not exist. This comment is particularly relevant to **E**, as will be explained below. Notice also that algorithmic execution speed is not on this list, but may be a criterion which needs consideration. These reasons will be returned to during the consideration of candidate solutions to non-linear component analysis.

Analysis

Principle Component Analysis

What follows in this section is stated without proof and expected to be generally accepted by the academic community.

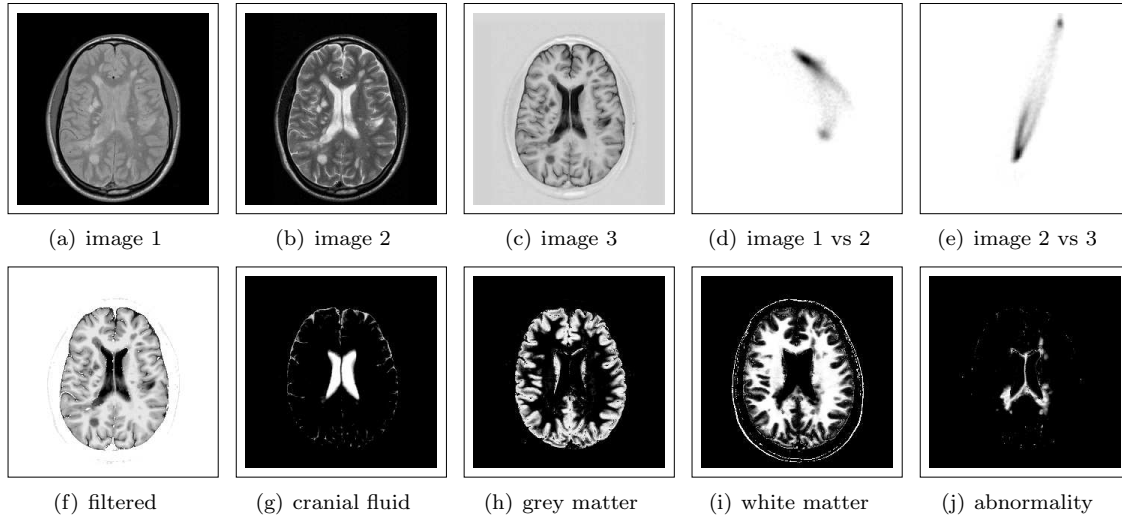
We start first with the simplest case of a linear data generation process. The standard solution for this is a planar constraint manifold and is well described in the literature. It corresponds to a Likelihood based fit of the hyper-plane parameters, which minimises the perpendicular distance from each datum to the constraint plane. This is solved directly by techniques such as Principle Component Analysis (PCA) [25, 12], which can be shown to estimate the position and orientation of a plane of fixed dimension m such that the sum of remaining off-plane squared residuals in the remaining $n - m$ dimensions is a minimum. **The Assumption of uniform IID Gaussian noise is fundamental to the intrinsic validity of this solution.** Note that in doing this it is not necessary to specify how data will actually be distributed within the first m maximally variable dimensions (ie: on the constraint manifold). Indeed the distribution on the manifold should have no systematic influence on the estimation of the hyper-plane. If we know the magnitude of the noise process then we can determine the dimensionality of the noise subspace if we did not know the value of m beforehand. Note also that the presence of homogenous measurement error has the result that the dimensions identified as describing the noise process are ‘degenerate’, meaning the ordering of values and specific details of the eigen-vectors are irrelevant beyond the identification of the noise subspace. The projection of data with homogenous IID Gaussian errors onto the m dimensional plane results in a homogenous measurement constraint for each data point within the plane, ie: the process is volume preserving (see below). This has two consequences; data densities constructed on the ‘hyper-plane’ are proportional to the probability of obtaining measurably distinguishable events [22], and similarity between locations on the hyper-plane can be justifiably written in simple ‘Euclidean’ form (ie: the least cost statistical path is a straight line). For this simple linear model all required characteristics (**A-E**) of our solution are available. It should be immediately obvious why and how these solutions are valid. In particular, distances between data points projected onto the surface of the fitted hyper-plane can be taken as noise free estimates of statistical similarity **E**.

An alternative justification of PCA can be made when we are analysing signal distributions which are known to be Gaussian. PCA can be said to be fitting the parameters of this multi-dimensional Gaussian. This is entirely consistent with a Likelihood based analysis of data, and is embodied by approaches such as mixture modelling using Expectation Maximisation (EM) [3]. However, assumptions regarding prior distributions represent a much stronger constraint than an m dimensional manifold. For arbitrary data samples we must consider whether any prior distribution assumption is even appropriate. As a simple example, this is equivalent to predicting a theoretical relationship between the volume and mass of an object, and then demanding that real world objects are always sampled from a pre-defined distribution of sizes in order to enforce this relationship. Some may argue that in many cases such an approach embodies an unjustifiable assumption which has little to do with the physical constraints we are trying to apply.

The above description emphasises the role of the noise sub-space in principle component analysis. Another common justification for PCA is the observation that the m dimensional signal will dominate the determination of the largest eigen values. It is a simple, yet erroneous, step from here to decide that the details of the noise-subspace are, in general, unimportant and can be neglected. The noise dominates in the $n - m$ dimensional noise subspace and can be amplified when we apply arbitrary non-linear transformations. In other words, **the ‘negligible noise’ argument is only valid for the linear data generation model.**

As an example of a real problem, we include here data from the area of medical image analysis. Figure 1 shows three MR scans of a human brain which have been aligned so that we can construct a 3 dimensional vector space from the corresponding grey levels. Figure 1(d) and 1(e) show scatter plots, from the region of brain tissue, for two combinations of variables. It is data such as this which we may wish to analyse to

extract clusters (**D**). The structure of the total distribution is quite complex, though physical constraints apply separately between pairs of tissues due to the process of partial voluming. Presumably, such complexity will also be manifest in all vision interpretation tasks, if we have an arbitrary sampling process, as we can expect any data set to be a union of many physical data generation processes. This adds an extra level of difficulty to any automated analysis. In order to apply any kind of component analysis we need a way of partitioning the data into groups associated with specific physical processes. This can be done for MR data using a mixture model. For this particular example the most meaningful similarity we can compute



(**E**) is a volumetric estimate of the partial volume of each tissue within each voxel. In fact this is also our definition of the Bayesian segmentation output (**C**) Figures 1(g)-(j). These results were computed using an EM based mixture model, comprising assumed distributions (which can be theoretically determined for the case of MR images). Figure 1(f) shows an attempt to reconstruct a noise free interpretation of the data in the region determined to be brain tissue (**A**), from the partial volume estimates. In this case it is well known that the partial volume process results in approximately linear behaviour. Pure tissue distributions are approximately Gaussian though the partial volume distributions are not. Note that task **B** appears to have no real meaning for this data, unless it is already compartmentalised into partial volume groupings.

Knowledge of the expected tissue distributions and the physical processes underlying the generation of data makes it possible to define exactly the model we need to describe the distribution of grey level vector measurements. This gives us the ground truth for the best interpretation of MR images. In particular, MR data has locally varying dimensionality (0 dimensions for pure tissue, 1 dimensional for partial volume and potentially 2 or more dimensions for combinations of 3 or more tissues).

Consideration of MR images shows us that we are not able to feed a random sample of data into any analysis based upon assumptions of a fixed dimensional manifold with no self-intersections and expect useful results. These problems will arise whenever we have multiple physical generation processes for data, such as we would also find for real world vision tasks such as colour, texture and 3D shape recognition. We can therefore see analysis of MR data as a good test for any dimensional reduction algorithm. It is also generally accepted that most data analysis tasks in computer vision are also not consistent with a linear model. Unless, like with MR images, the way data is represented within the human brain is somehow a special case, these observations preclude conventional PCA based approaches from being a viable model of human visual perception.

Non-Linear PCA

Now consider the problem of a non-linear data generation model. Under these circumstances researchers often talk in terms of a non-linear mapping of the original data as a new vector in a *latent variable space*. It is assumed that there exist latent spaces in which the required manifolds are again linear, so that the standard linear solutions can be applied. This approach exemplifies recent pattern recognition algorithms, such as Kernel PCA (KPCA). However, even though such latent spaces exist, our solution for linear systems would no longer be strictly applicable, as in the process of transforming our measured vector to the latent space we have modified the expected measurement distribution so that it is no longer uniform IID Gaussian.

We therefore *cannot* use PCA in the latent space to; estimate correctly the orientation of the hyper-plane in this space, remove noise **A**, construct valid density models **D**, or construct simple similarity measures **E**. Thus, **Kernel PCA** [16] and associated techniques [26] are inconsistent with a Likelihood based analysis of data. Indeed, in the absence of an externally imposed constraint (see MDS below), the major axes of variation identified in the latent space may be nothing more than non-linear amplification of noise, and have little to do with the intrinsic signal. This process can make the problem of determining the intrinsic dimensionality of the data m meaningless. The remaining tasks **B**, **C** are addressed to some extent, but we can have no confidence that the solution we obtain is in any statistical sense optimal.

Following the alternative likelihood interpretation of PCA described above, assumptions are often made regarding what the data distributions we are dealing with should look like, for example Gaussian-Process Latent Variable Models (GP-LVM) [8]. In this work, the data in the latent space is transformed so that the data distribution matches the assumed distribution by minimising a Kullback-Liebler divergence. However, forcing data to map onto a particular distribution, particularly after applying an adjustable non-linear transformation, precludes learning anything regarding the intrinsic distribution of the original data (**D**). We will never, for example, discover that data is bi-modal if we do everything we can to force it into a single peak. If we wish to use clustering approaches to identify significant locations then any non-linear transformation needs to be constrained to produce a meaningful data density. This could be achieved either through theoretical arguments (for example data transformations which achieve specific invariances), or for example (in the case of homogenous IID input data) by enforcing conservation of volume.

Biologically inspired ‘local’ approaches, such as the Kohonen Net [7], perform exactly the kind of function mapping required (for 2D systems; ie, $m = 2$), with noise suppression capabilities consistent with local eigen-vector analysis. With only a small amount of care these ‘memory intensive’ algorithms can be quite easily adapted to form solutions which are entirely consistent with statistical notions of pattern similarity [19]. However, these techniques cannot be expected to extract generally valid low dimensional representations (**B**). Not only are the topology and dimensionality of the manifold fixed beforehand (see MR data characteristics above), but also the ordinal values of generated topographic mappings are arbitrarily related to the information content (ie: measurement characteristics) of input data.

In order to continue with an approach which follows our preferred justification from Likelihood, we have two choices, either propagate the effects of noise from the input through to the latent space and take appropriate account of it (which precludes use of techniques such as PCA). Or, generally much more straightforwardly, construct our Likelihood measure not in the latent space but in the original data domain, where we know and understand our measurement noise. Unfortunately, this rules out the possibility of a closed form solution almost immediately⁴ and we must expect to have to formulate the problem as a numerical optimisation.

An iterative approach which minimises a valid Likelihood based upon distances computed in the input data space was suggested by the neural network community over a decade ago. The ‘counter propagation’ neural network uses a non-linear multilayered network architecture (such as an MLP) to generate a prediction of the input data on output [9]. The key part of this process is to define a restricted number of ‘neurons’ in the hidden layers, so that the representation of the data is forced through a ‘bottleneck’ which supports only a limited number of variables. Once the network has been trained, the final layers of the network can be discarded and the output is taken from the ‘bottleneck’ layer. This representation is effectively a non-linear re-representation of the input. There is considerable freedom within this framework for the selection of specific non-linear functions, and generally this selection must be optimised by determining which of the available choices exhibits the best generalisation performance. This is often performed via a cross-validation. Such a step is crucial if we are to believe that the selected non-linear model has any statistical validity, as we should already know that Likelihood, on its own, is incapable of performing model selection. Although these algorithms are presented in terms of ‘neural network’ terminology, the basic idea of minimising a Likelihood in order to determine the parameters of a functional mapping can be extended to any non-linear computational structure we care to choose, not just one composed of functional nodes and connection weights.

The output from a counter propagation network will still have several problems however. Unlike the linear PCA case, the new variables do not have the properties of uniform propagated errors. If the network has found a good model for the data then the trained output is a noise free estimate of the input data (**A**). Although we can say that the bottleneck variables have achieved some measure of dimensional reduction **B** and can also take ratios of data density as valid (such as in Bayesian calculations of conditional probability **C**), any data densities constructed in this latent space have arbitrary structure. As it is possible to define

⁴Only quadratic functions (such as those generated by solutions of sets of linear equations) have a unique minima. Arbitrary non-quadratic functions have arbitrary numbers of multiple minima, making direct solution generally impossible.

an infinite number of non-linear transformations to equivalent spaces (as might easily be obtained when training with different non-linear mapping functions within the neural network formulation) which will produce redefinitions of apparent data density, ruling out meaningful solutions to **D**. In addition, we cannot compute meaningful statistical measures of similarity (ruling out **E**). The first of these problems (and perhaps the second one too, but see next paragraph) arises because this latent space should loosely be considered as a metric space⁵, in which the local scaling of the space is related to the expected distribution of noise propagated from the input space. Therefore construction of a clustering analysis in this new space requires either the tools of Riemann geometry [17], or a separate stage to remap the new variables to ones of homogenous propagated error. The former would appear to be insufficient to deal with the requirements of a quantitative statistical approach, the latter of these two approaches would therefore be preferred, and also confers the property that measures of similarity can be computed along (Euclidean) straight lines. This is similar in essence to the aims of Sammon mapping [15], but with much more stringent requirements on the outcome. Those familiar with the difficulty of finding a good solution using Sammon mapping will probably realise that this is not going to be an easy problem to solve.

The question we need to ask now is whether distances computed on a manifold with homogenised errors provides a meaningful definition of similarity. In one respect the answer is clearly no. Consider the case where the manifold is allowed to fold back on itself so that two points which are distant on the manifold become proximal in the measurement space (such as when considering parts of cerebral cortex across a sulcus). In this circumstance we know that a poorly localised measurement may have ambiguous interpretation with regard to the manifold. A similarity estimate which constrains distances to be measured only along the manifold will be anomalously large for some purposes. Distances between points in the original input vector space may provide the only meaningful statistical definition. **We need to therefore consider whether the characteristics of data require similarities to be constructed with a manifold constraint.** In some circumstances we may wish to map locations on the manifold to a separate objective definition of distance (the measured position along the cortex in the previous example, or a pattern classification probability), in order to regain a meaningful definition of **E**. Common approaches include Multi-Dimensional-Scaling (MDS) [6]. In principle this can be done within the framework of conventional neural network training algorithms.

Allowable Generalisations

The analysis so far has assumed homogenous Gaussian IID noise, but now that we have the overall framework it is possible to consider the effects of more general noise models.

On the positive side, provided that we know the noise associated with each measurement, we can accommodate individual measurement errors, including data correlation and non-Gaussian data, should it become necessary. Once we have already accepted that the identification of non-linear latent spaces must be done via numerical optimisation, these additional extensions can be made directly. There are computational advantages to be gained from keeping the statistical behaviour of the data as simple as possible, if we can. This becomes a constraint for the way we perform feature selection. For example, if we wish to use conventional PCA, why not see to it that the input data has the properties of IID Gaussian measurement noise? **Consideration of statistical characteristics of input data can probably be identified as the most essential factor in algorithm selection, and also the first casualty of assuming measurement errors are not important.**

On the other hand, the identification of either a Riemann space or a space of homogenised errors, presupposes that there is a unique error process associated with any particular location in the input space. Any suggestion that the noise process could vary for repeated observation of the same vector immediately precludes the possible construction of a meaningful clustering solution. One example of this would be a data vector constructed from input image data with the property of scale invariance.

Answers

Question 1; Why are there so many publications in this area?

There are two main aspects to the answer; firstly, as explained above, non-linear PCA is required for a variety of reasons, some researchers may wish to focus only on a small number of them and may quite

⁵I use the word ‘loosely’ here, as the process of an unconstrained non-linear mapping is likely to generate correlations between parameters in the latent space unless steps are taken to avoid them. Such correlations cannot be accommodated within the mathematical framework supported by a conventional ‘metric’.

correctly ignore the consequences of their designs with regard to other tasks. The most obvious of these is seen when performing Bayesian classification **C**, where ratios of probability density can be supported, sufficient to perform classification tasks via Bayes theorem, even if the absolute density distributions have arbitrary quantitative meaning. This would be more impressive, if it were not for the fact that Bayesian approaches have a habit of introducing arbitrary prior probabilities into algorithms; thereby making any results non-quantitative and solutions non-unique [4].

Secondly, I believe it is also necessary to consider this question in the light of assuming that our input data errors are unknown. From a conventional statistical perspective, a lack of knowledge of measurement errors is equivalent to allowing any transformation (and consequent statistical re-weighting) of the input data. This implies that there can be **no unique solution** to any specific analysis. The problem as specified is ill-posed⁶ and there are a multitude of assumptions which we may feel can be justifiably made in order to regain a unique ‘solution’. Generally these assumptions will be driven by specific application domains. There are therefore as many techniques as there are data sets.

It seems to me that the negligible noise and Bayes classification interpretations, are the only ways a non-linear component analysis which neglects measurement error can be considered consistent with a measurement based analysis. Whatever we think of the legitimacy of these arguments, neither of these interpretations can provide solutions consistent with the requirements of **D** and **E**.

Ultimately, **it is impossible to empirically prove that any unique solution to an ill-posed problem is fundamentally correct** (ie: that our choice of prior distribution or implicit measurement scale is theoretically valid), so the door is always open for another publication .

Question 2; **Could any published pattern recognition method be considered as a valid solution?**

We have identified here that the method of counter-propagation neural networks, published over a decade ago, together with minor modification of the estimated non-linear latent space to regain homogenous errors (eg: MDS), could be considered as a principled statistical approach, which allows the construction of **unique** density based clustering algorithms, for well characterised measurement systems. The key characteristics are the propagated effects of the noise perturbation process on;

- the formulation of the parameter optimisation process,
- the definition of density estimates in the latent space,
- and the construction of similarity measures in the latent space.

This does at least provide a straw man for the definition of an appropriate statistical solution to categories **A-D**, though this approach raises several new questions, specifically regarding practicality. One observation which casts more light on the problems of getting a solution to **E**, is that we have made an implicit assumption when defining non-linear component analysis that we can represent an input data vector as a single point on the manifold. In the presence of noise and for arbitrary (self-intersecting) manifolds this is clearly an invalid assumption ⁷. The dependence of any representational ambiguity on the measurement process provides another cause for concern given the design of many algorithms. It is therefore difficult to see how we could ever expect to find an algorithm which guarantees a solution to **E** in all circumstances. In particular, the decision whether or not to compute distances along the extracted manifold will involve problem dependant information which some researchers might prefer not to require.

Approaches which assume prior distributions, as criticised in the previous section, can also be modified to explicitly incorporate measurement error. In the case of EM mixture modelling this would correspond to modifying the assumed prior distribution to account for broadening due to the expected noise propagated from the measurement into the latent space, thereby providing solutions to **A-C** [14].

Beyond the basic requirement of making sure the theory used to derive a method is quantitatively valid, the key problem with getting any non-linear analysis working is that of identifying the most appropriate non-linear model and the difficulty of estimating its parameters. Others may therefore prefer a more elegant (closed form) algorithm. Many methods have been suggested in the interest of computational efficiency with the sole aim of getting something which might work at ‘some level’ in a short time. **The basic theoretical**

⁶Not the original meaning of attempting to construct a mathematical description of a non-physical system, but in the more common usage that we have insufficient information to compute a solution to the problem as posed.

⁷If we are considering things in terms of physical spaces then we have just specified the requirements for a wormhole, so we can see why in general physicists are going to have to work hard to encounter the same problem when they use Riemann geometry.

inadequacies of some approaches (Fuzzy Logic, Boosting, Support Vector Machines, Kernel PCA) are understood and accepted by some but rarely discussed. However, I do not believe that execution speed should be the only criterion, and certainly we should not under any circumstances, ignore the issue of statistical validity. Otherwise we are simply developing very fast ways of getting wrong answers. For myself, I would like to see solutions to the problem of Non-Linear PCA which have practical utility but conform to my expectations of a valid approach.

As even the simplest of non-linear functions (such as a loop) can provide a real challenge for iterative algorithms, it is difficult to believe that ‘global’ problems of arbitrary complexity in our n dimensional space can seriously be solved in this way.

Counterpoint

There are several counter arguments to these conclusions which I would predict from observations of research in computer vision over the last decade.

The first is; *if we must ultimately measure the generalisation capabilities of our Likelihood solutions anyway, then why can't we ignore the statistical arguments and rely on cross-validation to select good answers? This would then open the door to unconstrained use of any method we like.* While this is a possibility, we should not cling to the belief that these methods have intrinsic theoretical validity, or that this process is likely to be reliably effective. I would also cite this as the major cause of the criticism levelled at computer vision algorithms over the recent decades, that performance of algorithms cannot be predicted outside of the data sets with which they were developed.

The second is; *insistence that data has been provided in a way that measurement details must remain unknown.* **Although many would certainly prefer to believe that there are generic reasons to ignore the measurement process, logical consideration of the facts appears to arrive at the opposite conclusion.** I therefore see persistence in this view as a combination of laziness and wishful thinking. In my opinion, it is more productive to consider how this information might be estimated from the data available. Some approaches in pattern recognition, such as Factor Analysis [2], make this an explicit goal. Aside from the conventional analytic approaches, such as error propagation, there are several relatively simple methods for doing this, such as repeatability measurements or estimation using continuity or smoothness constraints. One could reasonably argue that any analysis task for which the measurement error can be estimated should simply not be analysed in a way which ignores it. We must also expect trained scientists to hold the view that to persist in this approach runs counter to basic principles.

The third is; *problems in computer vision are well known to be ill-posed, and techniques such as Bayesian approaches are fundamental to their solution.* Many computer vision problems are indeed ill-posed and with care Bayesian methods may be used to enforce prior knowledge. **However, algorithms which generate arbitrary results with no intrinsic quantitative validity (even if presented as Bayesian approaches) should be considered unsatisfactory.** Perhaps these tasks are better interpreted as the test of a hypothesis with respect to specific prior assumptions, rather than a unique answer. We should prefer to identify the data which is required to solve a problem, and where we might obtain it, rather than pre-suppose that everything can be solved with a common prior assumption.

Conclusions

This document motivates the selection of candidate modules of human perception on the basis of statistical validity. A central theme of this document is that many pattern recognition algorithms ignore the measurement characteristics of the data they are applied to. While this approach is justifiable in some cases, in general it is not. We have attempted to illustrate this by discussing the process of dimensional reduction in a case where it is possible to formulate a statistical solution that takes explicit account of measurement. We find that the popular claim that measurement error can be neglected is not applicable in the non-linear case. Also the parameters of the extracted manifold will be unreliable so that even Bayesian classification will be compromised. Any technique which purports to estimate the parameters of a non-linear manifold should be based upon Likelihood (or its **quantitative** equivalents [1]), and should take explicit account of the errors. Those who need to use these techniques should therefore make an effort to understand the noise characteristics of their measurement process, rather than persist in ignoring them. This lesson would appear to apply to any data processing task, not just non-linear component analysis, see for example [11] for algorithms designed as solutions to linear equations, [22] for constructions in probability such as ‘information’

and Likelihood and finally [24] for an outline of the problems of taking a statistical interpretation of ‘shape theory’, as exemplified by the use of Riemann geometry.

There is a specific need for those who consider themselves to be in a pattern recognition or computer vision ‘user-group’ such as cognitive scientists, to try to understand the limitations of published techniques. They should be aware that algorithms are often tolerated in application driven areas even though the principles which underpin them are flawed. Paradoxically, publications and approaches are likely to become popular in the area of computer vision specifically because they appear to work quickly or achieve the impossible. The emphasis here is on the word ‘appear’. Unfortunately, the continued popularity of totally unjustifiable approaches demonstrates that we cannot rely entirely on experimental testing to show us when there are problems and thereby convince everyone of what constitutes best practice.

In the absence of convincing empirical validation, it makes sense where possible to appeal to logical analysis. For example; when attempting non-linear component analysis, the assumption that errors can be neglected is not the same as saying there are none, and the assumption that a particular data set has a specific prior distribution is highly specific. Both of these examples illustrate that correct selection of a method is conditional on an understanding of the data. We can therefore not expect to use techniques, which are in effect a sophisticated statistical analysis, as though they are black boxes. To insist on taking this approach may result in increased publication rates but can be regarded as both poor science and poor engineering. **Our models of vision will need to make explicit the details of how data is obtained and its statistical character** [13].

Finally, I would like to mention once again the difficulty raised by algorithms which claim to provide unique solutions to ill-posed problems. The assumption that we can ignore errors, and the use of arbitrary prior distributions in Bayesian solutions, is expected to result in an infinite number of arbitrary solutions (and potential publications), not just for non-linear component analysis but any data analysis task. If we understand enough about a problem to be able to say that it is ill-posed then we should not accept any published work which claims to provide a solution as a theoretical basis for a scientific study. As a consequence, we must decide if such solutions can ever be legitimately accepted as candidate models of human perception. If this argument has failed to convince the reader, there is another way of considering this problem. Ill-posed problems are those which have insufficient data for a unique solution. Any quantitative estimation process should therefore make this explicit by reporting the uncertainty. Even ill-posed problems are not intrinsically a bad thing provided we are ‘honest’ regarding the un-reliability of any result. This is just another way of saying we need to know the errors on the output [5, 18, 20]. Both the issue of statistical design and ill-posed problems can therefore be seen as two facets of the same requirement. It can be summarised in one demand which researchers should make when selecting computer vision algorithms as candidate models of visual perception;

Show me the errors!

This observation is consistent with a systems design methodology for vision system construction, whereby each module is selected in order to match the statistical characteristics of input and output data to the assumptions used in module design [21]. However, if you think that the ideas presented in this document are at all obvious and perhaps did not require stating, please take a look at [23] which discusses a diverse set of the published literature with regard to this and related statistical design principles. Meanwhile, we leave it to the reader to take another look at other popular approaches, such as SIFT [10], and to decide for themselves how they live up to this demand.

Acknowledgements

We would like to thank Chris Rose for his helpful comments regarding this manuscript.

References

- [1] H.Akaike, ‘A New Look at Statistical Model Identification’, IEEE Trans. on Automatic Control, **19**, 716, (1974).
- [2] A. Basilevsky, Statistical Factor Analysis and Related Methods. Wiley, New York, 1994.
- [3] C.M.Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 66 ff, 1995.

- [4] P.A. Bromiley, M.L.J. Scott, M. Pokrić, A.J. Lacey and N.A. Thacker, Bayesian and Non-Bayesian Probabilistic Models for Magnetic Resonance Image Analysis, *Image and Vision Computing, Special Edition; The use of Probabilistic Models in Computer Vision.*, 21, 851-864, 2003.
- [5] P.A.Bromliey, M.Pokric and N.A.Thacker, Computing Covariances for Mutual Information Co-registration, *Proc. MIUA*, 77-80, London, 2004.
- [6] M.L. Davison, *Multidimensional Scaling*. John Wiley & Sons, New York, 1993.
- [7] T. Kohonen, The Self-Organising Map, *Proc. IEEE*, 78, 9, 1464-1480, 1990.
- [8] N. Lawrence, Probabilistic Non-Linear Principle Component Analysis with Gaussian Process Latent Variable Models, *Jou. Mach. Learn. Res.*, 6, 1783-1816, 2005.
- [9] D. Lowe, M.E.Tipping, Feed-Forward Neural Networks and Topographic Mappings for Explanatory Data Analysis, *Neural Comp. and App.*, 4, 83, 1996.
- [10] D.G.Lowe, Distinctive Image features from Scale-Invariant Key-points, *Int. Jou. Comp. Vis*, 2004.
- [11] A. Nayak, E. Trucco and N.A. Thacker, "When are simple LS estimators enough? An empirical study of LS, TLS and GTLS", *IJCV*, 68-2, 203-216, 2005.
- [12] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, "Numerical recipes in c" 2nd Ed. Cambridge University Press: Cambridge, 1992.
- [13] V Ramesh. Performance Characterisation of Image Understanding Algorithms. *PhD Thesis*, University of Washington, 1995.
- [14] D.B. Rubin and D.T.Thayer, EM Algoritihms for ML Factor Analysis, *Psychometrica*, 47, 1, 69-76, 1982.
- [15] J.W.Sammon, A Non-Linear mapping for Data Structure Analysis, *IEEE Trans. Comp.*, C-18, 5, 401-409, 1969.
- [16] B. Scholkopf, A. Smola and K-R, Muller, Non-linear Component Analysis as a Kernel Eigenvalue Problem. *Neur. Comp.* 10, 1299-1319, 1998.
- [17] C. G. Small, *The Statistical Theory of Shape*, Springer, 1996.
- [18] N.A.Thacker and J.E.W.Mayhew. 'Optimal Combination of Stereo Camera Calibration from Arbitrary Stereo Images', *Image and Vision Computing*. 9(1),27-32, 1990.
- [19] N.A.Thacker, I.A.Abraham and P.Courtney, 'Supervised Learning Extensions to the CLAM Network.' *Neural Networks Journal*, 10, 2, pp.315-326, 1997.
- [20] N.A.Thacker, A.Jackson, Mathematical Segmentation of Grey Matter, White Matter and Cerebral Spinal Fluid from MR image Pairs, *British Journal of Radiology*, 74, 234-242, 2001.
- [21] N.A.Thacker, A.J.Lacey, P.Courtney and G. S. Rees, An Empirical Design Methodology for the Construction of Machine Vision Systems. Tina memo, 2002-005, www.tina-vision.net, 2002.
- [22] N.A.Thacker, P.Bromiley, The Equal Variance Domain: Issues Surrounding the use of Probability Densities for Algorithm Construction. Tina memo, 2004-005, www.tina-vision.net, 2004.
- [23] N. A. Thacker, A. F. Clark, J. Barron, R. Beveridge, C. Clark, P. Courtney, W.R. Crum, V. Ramesh, Performance Characterisation of Machine Vision Algorithms; A Guide to Best Practices. Tina-memo, 2005-009, www.tina-vision.net, 2005.
- [24] N. A. Thacker, A Critical Assessment of the Use of Riemann Manifolds for Shape Analysis. Tina-memo, 2006-006, www.tina-vision.net, 2006.
- [25] M.E.Tipping and C.M.Bishop, Probabilistic Principle Component Analysis, *Jou. Royal Stat. Soc.*, B, 6, 3, 611-622, 1999.
- [26] K.C.I.Williams, On a Connection between Kernel PCA and Metric Multi-dimensional Scaling, *Machine Learning*, 46, 11-19, 2002.