To appear in Proc. of International Joint Conference on Neural Networks, Washington, 1999.

Input Variable Selection Using Independent Component Analysis

Andrew D. Back and Thomas P. Trappenberg
RIKEN Brain Science Institute,
The Institute of Physical and Chemical Research
2-1 Hirosawa, Wako–shi, Saitama 351-0198 Japan.

Abstract

The problem of input variable selection is well known in the task of modeling real world data. In this paper, we propose a novel model-free algorithm for input variable selection using independent component analysis and higher order cross statistics. Experimental results are given which indicate that the method is capable of giving reliable performance and that it outperforms other approaches when the inputs are dependent.

1 Introduction

In many real world modeling problems, for example in the context of biomedical, industrial, or environmental systems, a problem can occur when developing multivariate models and the best set of inputs to use are not known. Input variable selection (IVS) is aimed at determining which input variables are required for a model. The is task is to determine a set of inputs which will lead to an optimal model in some sense. Problems which can occur due to poor selection of inputs include the following:

- As the input dimensionality increases, the computational complexity and memory requirements of the model increase.
- Learning is more difficult with unrequired inputs.
- Misconvergence and poor model accuracy may result from additional unrequired inputs.
- Understanding complex models more difficult than simple models which give comparable results.

Methods of input variable selection can be categorized into *model-based* and *model-free* methods [1]. *Model-based* methods typically involve selecting a model, choosing the inputs to use, optimizing the parameters, and then measuring some cost function. The inputs are changed and then the procedure is repeated. A test is used to choose which inputs to use based on these results. *Model-free* methods are based on performing a statistical test between the subsets of input *variable(s)* and the desired output(s) from the model.

For n possible input variables it is necessary to test all possible $2^n - 1$ subsets of inputs in order to determine the optimal subset [1]. Hence efficient algorithms are of vital importance as are any methods which may overcome the scaling problem. In the following sections we describe an algorithm for selecting input variables which offers some advantages over previous approaches.

2 An ICA Input Variable Selection Algorithm

2.1 Assumptions

The usual IVS problem can be described mathematically as follows. A system F_o receives input from the variables $\mathbf{x}(t) = [x_1(t) \cdots x_p(t)]'$ and produces an output $y(t)$:

$$y(t) = F_o(\mathbf{x}(t)) \tag{1}$$

It is assumed the system F_o can be approximated arbitrarily well by a linear or nonlinear functional map. To estimate F_o, measurements $\mathbf{z}(t) = [z_1(t) \cdots z_n(t)]'$ are taken with the assumption that

$$\mathbf{x}(t) \subseteq \mathbf{z}(t). \tag{2}$$

The usual model building approach is to apply an input variable selection procedure to obtain a set of model inputs $\mathbf{z}_a(t) \subseteq \mathbf{z}(t)$ where ideally $\mathbf{z}_a(t) = \mathbf{x}(t)$. Hence a model can be written as

$$\hat{y}(t) = F(\mathbf{z}_a(t); \boldsymbol{\theta}) \tag{3}$$

where F is a functional map parametrized by $\boldsymbol{\theta}$.

However, the assumptions indicated above can be quite significant and it is likely that the measuredments taken are not a strict superset of inputs to the actual system. We may observe data which has been *filtered* in some manner, relative to the true inputs. Consider a simple filtering of the measured multivariate data:

$$\mathbf{z} = \mathbf{A}(z)\mathbf{x}_L \quad , \tag{4}$$

where $\mathbf{x}(t) \subseteq \mathbf{x}_L(t)$ and $\mathbf{A}(z)$ is a multivariate filter. Each measured term z_i can be the result of the weighted and delayed sum of x_j. A simpler variation to consider is when the filters are first order only, giving

$$\mathbf{z} = \mathbf{A}\mathbf{x}_L \quad . \tag{5}$$

where in this case $\mathbf{A}$ is as mixture matrix containing scalar terms only. Note that in this situation, there does not exist any subset $\mathbf{z}_a(t) \subseteq \mathbf{z}(t)$ such that $\mathbf{z}_a(t) = \mathbf{x}(t)$. Clearly a different approach is required or else any input variable selection method will fail by overestimating the number of inputs required. What is required is an algorithm which would permit the reverse transformation before the IVS procedure is applied:

$$\begin{aligned} \mathbf{x}_L &= \mathbf{W}\mathbf{z} && (6) \\ \mathbf{x} &= \mathbf{G}\mathbf{x}_L && (7) \end{aligned}$$

where $\mathbf{W} = \mathbf{A}^{-1}$ is the inverse of the assumed mixture matrix and $\mathbf{G}$ is a sparse matrix which selects the desired subset of inputs to the model. To go further with this approach requires either *a priori* knowledge of $\mathbf{A}$ or some algorithm to estimate $\mathbf{A}$ and hence $\mathbf{W}$.

Recently, a family of mathematical techniques known as *independent component analysis* (ICA) has been shown to give exactly the solution to this type of problem under some assumptions [2]. The aim of ICA is to use an algorithm applied only to measured multivariate data and, based on the assumption of either temporal or spatial independence of the channels, estimate a demixing matrix to give outputs which meet the specified criteria. A common approach is to assume the input channels are statistically independent. Then an algorithm based on estimating the mutual information between the inputs permits the estimation of $\mathbf{W}$ to give outputs which are maximally independent. We consider this approach further as a means of providing an improved method of input variable selection.

2.2 A Statistical Test

In this paper, we propose to use ICA to make the input variables as mutually independent as possible. Moroever, using ICA allows us to derive model-free IVS algorithms based on statistical dependence tests. In the following section we examine the use of ICA in the problem of input variable selection. The basic strategy we suggest is to apply ICA to estimate the independent inputs $\mathbf{s}$ from $\mathbf{x}$ and then derive a statistical test to determine the desired subset of input variables $\mathbf{s}_a$.

One approach to determining statistical dependence is to use of mutual information between two signals x and y, given by

$$I(x;y) = \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx dy \tag{8}$$

This is also known as the cross-entropy or the Kullback-Leibler divergence between the joint pdf of (x,y) given by $p(x,y)$ and the product of the marginal pdfs $p(x)$, $p(y)$. It may be implemented by estimating the pdfs in terms of the cumulants of the signals[1].

Since we only require a relatively simple binary decision to be made about the dependence or otherwise of signals, it is not necessary to compute a precise value for the mutual information. Instead, the higher order cross cumulants[2] of multiple variables can be used directly up to some suitable order, to determine the statistical dependence of the signals.

2.3 The ICAIVS Algorithm

The particular IVS test we propose is based on using all possible cross cumulants up to a specified order among the individual terms. This statistical measure can be used to establish the independence or otherwise of non-Gaussian signals. These cumulants are defined as

$$\begin{aligned} \mathbf{C}_{xy}(k) &= [c_{x_1y}(k), ..., c_{x_ny}(k)]^T \\ &\vdots \\ \mathbf{C}_{x^py}(k) &= \left[c_{x_1^py}(k), ..., c_{x_n^py}(k)\right]^T \\ \mathbf{C}_{x_1x_2y}(k) &= [c_{x_1x_2y}(k), ..., c_{x_1x_2y}(k)]^T \\ &\vdots \\ \mathbf{C}_{x_1^px_2^py}(k) &= \left[c_{x_1^px_2^py}(k), ..., c_{x_1^px_2^py}(k)\right]^T \\ &\vdots \\ \mathbf{C}_{x_1^p..x_n^py}(k) &= \left[c_{x_1^p..x_n^py}(k), ..., c_{x_1^p..x_n^py}(k)\right]^T \end{aligned}$$

where the cross cumulant vectors are between the inputs $x_1, x_2, ..., x_n$ at time $t-k$ and the output y. For convenience, we will use the notation $C_m(k) = C_{x_1...x_my}(k)$. There are two broad cases that can be considered at this point:

1. Models with only instantaneous inputs. In this case we use $\mathbf{C}_m(0)$.

[1]Note that although a truncated expansion is often used, to approximate the pdf exactly could require an infinite number of terms. The usual difficulties of approximating with polynomials apply in this case also [3].

[2]In contrast to the often quoted first order cross cumulant measure $c_{x_1\cdots x_n}(\tau_1,\ldots,\tau_n)$, of n variables or the mth order cumulant of a single variable. Since we are seeking to determine the statistical dependence between variables not just the correlation between variables, it is necessary to use higher order cross cumulants.

2. Models with delayed inputs. Here we test input variable $\{x_i(t-k)\}$ $k = 0, ..., p$, $i = 1, ..., n$ against the system output $y(t)$, by examining the points in the cross-cumulant space given by elements of the vector $\mathbf{C}_m(k)$.

Our aim is to combine various cumulant measurements and thereby apply a decision function to determine the relative dependence of each input subset on the model output(s). To do this, it is useful to normalize[3] each of the cumulants in order to compare and combine them in a reasonable manner. Hence, we apply the following normalization steps to the cumulants:

1. Zero mean signals: $x_i = x_i - E[x_i]$
2. For cumulants using second and higher even order terms, normalize as: $x_i^{2n} = x_i^{2n} - E[x_i^{2n}]$ ... $n = 1, 2, ...$
3. Bipolarize data: $x_i = \frac{x_i}{|x_i|}$
4. Renormalize individual cumulants. Assuming perfectly correlated data, we may use the normalization: $\mathbf{C}'_{xyz..}(k) = \mathbf{C}(k)_{xyz..}/\mathbf{C}(k)_{xxx..}$

Doing these normalizations will allow us to combine various cumulant estimators for the decision function. A simple test for dependence can be obtained as follows.

For each cross-cumulant statistic, determine the average level of dependence implied by the magnitude of the statistic. Compare each input in turn to this average. The inputs which are significantly different from the average value are candidates for inputs to the model. That is, for ICA transformed inputs $\tilde{x}_i$ $i = 1, .., n$, considering the subsets of inputs, $\tilde{\mathbf{x}} = [x_1, ..., x_n, x_1x_2x_3, ..., x_1 \cdots x_n]^T$ where $\tilde{\mathbf{x}}_j$ is the jth element of $\tilde{\mathbf{x}}$, Hence we obtain the rule:

$$\mathbf{S}_{xy}(i,k) = \begin{cases} 1 & \begin{array}{l}\sum_i |\mathbf{C}_{xy}(i,k)| - \\ E\left[\sum_i |\mathbf{C}(i,k)|\right] > K_{xy}(i,k)\end{array} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$$\tilde{\mathbf{x}}_a(i,k) = \begin{cases} \tilde{\mathbf{x}}(i,k) & \left(\begin{array}{c} \mathbf{S}_{xy}(k)\oplus \\ \mathbf{S}_{xxy}(k)\oplus \\ \mathbf{S}_{xxy}(k)... \end{array}\right)_{i,k} > 1 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where $\mathbf{C}_{xyj}(i,k)$ is the i th element of $\mathbf{C}_{xy}(i,k)$ at time $t-k$ and $K_{xy}(i,k)$ is a threshold value chosen to suit the level of precision required in determining the independence of the signals, chosen for each subset. The values of K_{xy} is problem dependent and is chosen according to how the level of dependence that can be tolerated between the measured variables.

This rule means that if any of the cross cumulants for any given input subset are above a set threshold level, that particular input subset is deemed to be required for the model.

[3] For example, correlation functions normalized to a maximum value of 1 are independent of the actual magnitude scaling of the input variables. Similarly, we would like to normalize the cumulants such that the cumulant slices will not change if the input variables are scaled in asimilar manner.

The test is applied across all subsets of inputs, however in the equations above, only one input is actually shown.

2.4 Computational Complexity

For a cross cumulant between p variables, we need to test for all orders of cumulants up to the second power, which will be 2^p tests. For n input variables, this implies testing all possible subsets up to the set of n variables which leads to a total number of tests given by

$$N_T = \sum_{i=1}^{n} \binom{n}{i} 2^i \quad (11)$$

$$= 3^n - 1 \quad (12)$$

While the test scales poorly, in practice, we may not need to test all possible combinations of inputs to establish conclusively whether a variable is required or not. Low order terms can be tested initially and them higher orders.

Remarks

1. PCA is often used to select inputs, but this is not always useful, since the variance of a signal is not necessarily related to the importance of the variable. The features selected may have nothing to do with the problem.
2. In contrast to the use of PCA for input variable selection [4, 5], the variables we remove are those which are independent of the output, which is quite different from removing those with low variance.
3. Spurious correlations or dependencies may exist between unrelated variables and hence could lead to falsely included inputs, eg: generated by coupled systems.

3 Simulation Examples

3.1 Example 1

In this example, we show the effect of using higher order cross cumulants as a means of detecting dependence among variables. Here 15 mutually independent binary iid signals were generated. Three signals, x_2, x_6 and x_9 were used as inputs to a system with output y. The nonlinear model is described by

$$y = x_2^3 + \cos(x_6) + 0.3\sin(x_9) \quad (13)$$

We assume that the measured data has been recovered from the ICA processing and test the HOS directly. Some results are shown in Fig. 1, where it can be noted that the required input signals are readily discerned. Examination of the normalized cross cumulants shows that the dependent inputs have markedly higher values than those which are independent of the output y.

It is especially interesting to observe that the dependence is not always obvious with the second order statistics for example, but the higher order cumulants serve a role in identifying all the required inputs.

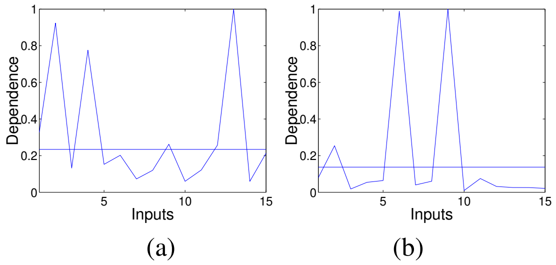

Figure 1: Input variable selection Experiment 1 results: (a) without using ICA the selected signals are incorrect, (b) using ICA the required inputs 2,6 and 9 are correctly selected.

3.2 Example 2

Suppose we observe a multivariate time series $\mathbf{x}(t) = [x_1(t) \ldots x_n(t)]$ consisting of dependent variables $\{x_i(t)\}$, and an output $y(t)$ from a system to be identified. For the purposes of this example, suppose $n = 15$ and the output is the result of a functional model given by

$$y(t) = F_o\left(v_2(t), v_7(t-5), v_9(t-11)\right) \quad . \qquad (14)$$

where $F_o(a, b, c)$ is defined in this example, as

$$\begin{aligned} y &= (2a-1.6)^3 - 2a + \\ & \quad (3b-3.5)^2 - 3b - (0.5c-0.8)^3 - 1.2c + 4 \end{aligned}$$

The input dependencies in $\mathbf{x}$ arise due to some true signals $\mathbf{v}(t) = [v_1(t) \ldots v_n(t)]$ becoming mixed, according to

$$\mathbf{v} = \mathbf{A}\mathbf{x} \quad . \qquad (15)$$

The results from this experiment are observed in Fig. 2 where the necessary inpts to the model are easily selected. Although not shown here, when ICA is not used, *all* inputs were identified in this case as being required. Thus, the algorithm successfully identified just the inputs required from the measured data.

4 Conclusions

To effectively model and predict multivariate time series data it is important to use only inputs actually required and remove those inputs not required. If unrequired inputs are used significant problems can occur, especially in problems of high input dimensionality. Invariably it will be considerably more difficult to estimate a given model and the accuracy of the model will also suffer. Computational burden

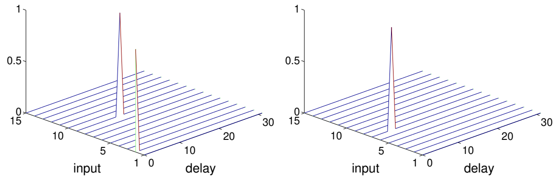

Figure 2: Results for Experiment 2 on input variable selection. Here the cumulants are shown after thresholding according to the test described in the text. Shown are the thresholded cumulants $c_{xy}(k)$, $c_{xxy}(0, k)$, $k = 1, \ldots, p$ respectively.

will also be increased dramatically due to the increased difficulty in learning. The problem of input variable selection normally assumes that it is possible to select the required optimal set of inputs directly from the set of measured data. However we showed that this assumption is easily violated. When this occurs, overestimation of the number of inputs can occur.

In this paper, we proposed a new method for performing input variable selection which helps to solve the above problem. The method is based on the recently introduced method of independent component analysis. This approach permits a relatively straightforward statistical test to be derived for model free input variable selection. We applied the proposed algorithm to some examples which showed that it is capable of successfully isolating the inputs required to a model, even when the measured data itself is mixed and would normally lead to overestimating the number of inputs required. It is apparent that ICA provides a useful tool for accurately estimating the inputs required in building complex models.

References

[1] B. V. Bonnlander and A. S. Weigend, "Selecting input variables using mutual information and nonparamteric density estimation," in *Proceedings of the 1994 International Symposium on Artificial Neural Networks (ISANN'94)*, Tainan, Taiwan, 1994, pp. 42–50.

[2] C. Jutten and J. Herault, "Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, pp. 1–10, 1991.

[3] A. Stuart and K. Ord, *Kendall's Advanced Theory of Statistics: Sixth Edition Volume 1, Distribution Theory*, Edward Arnold, New York, 1994.

[4] Asriel U. Levin, Todd K. Leen, and John E. Moody, "Fast pruning using principal components," in *Advances in Neural Information Processing Systems*, Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, Eds. 1994, vol. 6, pp. 35–42, Morgan Kaufmann Publishers, Inc.

[5] Nandakishore Kambhatla and Todd K. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1493–1516, 1997.