2014

Partial Information And Partial Weight - Two New Information Theoretic Metrics To Help Specify A Data-Based Natural System Model

Ashish Sharma

Rajeshwar Mehrotra

PARTIAL INFORMATION AND PARTIAL WEIGHT - TWO NEW INFORMATION THEORETIC METRICS TO HELP SPECIFY A DATA-BASED NATURAL SYSTEM MODEL

ASHISH SHARMA (1), RAJ MEHROTRA (1)
(1): School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales, Australia

How does one define or specify a system? This is a problem faced routinely in science and engineering, with solutions developed from our understanding of the processes inherent, to assessing the underlying structure based on observational evidence alone. In general, system specification involves identifying a few meaningful predictors (from a large enough set that is plausibly related to the response) and formulating a relation between them and the system response being modeled. For systems where physical relationships are less apparent, and sufficient observational records exist, a range of statistical alternatives have been investigated as a possible way of specifying the underlying form. Here, we present two new contributions that were recently published by Sharma and Mehrotra (2014) as a step towards an assumption free specification of a system using observational information alone. The first of these is the partial information (PI), a new means for specifying the system, its key advantage being the relative lack of major assumptions about the processes being modeled in order to characterize the complete system. The second is the concept of partial weights (PW) which use the identified predictors to formulate a predictive model that acknowledges the relative contributions varied predictor variables make to the prediction of the response. We assess the utility of the PI-PW framework using synthetically generated datasets from known linear, non-linear and high-dimensional dynamic yet chaotic systems, and demonstrate the efficacy of the procedure in ascertaining the underlying true system with varying extents of observational evidence available. We highlight how this framework can be invaluable in formulating prediction models for natural systems which are modeled using empirical or semi-empirical alternatives, and discuss current limitations that still need to be overcome.

INTRODUCTION

System specification is a problem common to many fields of science and technology. Given a sequence of data originating from a system, the goal is to identify its best possible drivers (inputs), formulate a model based on the identified drivers, determine the model parameters that provide the best fit to the data, and finally predict the system response for new (or yet to be observed) inputs. While identification of the drivers and their use in specifying the model is simplified if the processes involved are known in sufficient detail, underlying complexity in the processes and uncertainty in the data often require specification using empirical or data-based alternatives. Such a data based specification of the system is achieved

through two stages: (a) identifying system predictors (or drivers), and (b) casting identified predictors in a predictive framework. An approximation of the system form (into say a linear representation) greatly simplifies the above two steps, but can often be hard to justify. There is a need for options where the system can be fully specified (or predictors chosen and cast into a predictive framework) without resorting to simplified system representations. One such approach for doing so is presented here.

The presentation here is a summarized version of a recent extension to this logic, reported in Sharma and Mehrotra (2014). Many of the results and discussion have been reproduced from the paper. The methods presented in this paper are developed from earlier work reported in Sharma (2000a,b), Sharma et al (2000), with details about estimation in Sharma et al. (1998) and the regression method used in Lall and Sharma (1996). Readers are requested to read the referred papers for full details of the approach if sufficient information is not included due to page restrictions.

METHODS

Background

The Mutual Information (MI) (Fraser and Swinney, 1986) between two variables X and P is defined as:

$$\mathbf{MI}(X,P) = \int f_{X,P}(x,p)\log\left[\frac{f_{X,P}(x,p)}{f_X(x)f_P(p)}\right]dxdp \qquad (1)$$

where $f_X(x)$ and $f_P(p)$ are the marginal probability density functions (PDF) of X and P, and $f_{X,P}(x,p)$ is the joint PDF of X and P. The above expression can also be viewed as the expected value of $\log\left[\frac{f_{X,P}(x,p)}{f_X(x)f_P(p)}\right]$, leading to the following sample estimate:

$$\hat{\mathrm{MI}}(X,P) = \frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{f_{X,P}(x_i,p_i)}{f_X(x_i)f_P(p_i)}\right] \qquad (2)$$

where (x_i,p_i), i=1...n, are sample observations of (X,P).

The specification of MI in (1) is easily extendable to more than two variables. Consider the vector (X,P,Z) where X is the system response, P is a variable that could potentially be a predictor to the response, and Z is a variable (or $\boldsymbol{Z}$, a vector of variables) that represents a pre-existing or pre-identified predictors. Readers should note that if $\boldsymbol{Z}$ is a vector, the product in the denominator of the MI estimate in (2) comprises of the product of the marginal probability density of individual variables in $\boldsymbol{Z}$ (and not the joint probability of $\boldsymbol{Z}$). If P is independent of X, it can be shown than $MI(X,P,Z) = MI(X,Z)$. Similarly, if P is not independent, it can be shown that $MI(X,P,Z) > MI(X,Z)$. We use these two properties in the development of the Partial Information (PI), our proposed measure of partial dependence.

Partial Information (PI)

MI measures information common to two or more variables. Specification of a system requires identification of system predictors from a list of variables that could be included in this category. Such identification usually starts from a null model (with no predictors), and involves adding variables such that the predictive uncertainty is reduced. If predictors are included one at a time, the selection process is referred to as stepwise selection. As the MI does not measure incremental improvements in the system specification as each predictor is included, it is of limited use in identifying predictor variables.

Partial Information (PI) is an extension of the MI developed to measure partial dependence, thus making it useable for system predictor identification. The PI is expressed as PI($X,P|\mathbf{Z}$), where P denotes the new predictor variable under consideration, and $\mathbf{Z}$ is a pre-identified predictor variable or vector. Extending the rationale for MI to the case of conditional or partial dependence, PI($X,P|\mathbf{Z}$) can be written as:

$$\mathbf{PI}(X, P \mid \mathbf{Z}) = \int f_{X,P|\mathbf{Z}}(x, p \mid \mathbf{z}) \log\left[\frac{f_{X|\mathbf{Z},P|\mathbf{Z}}(x, p \mid \mathbf{z})}{f_{X|\mathbf{Z}}(x \mid \mathbf{z}) f_{P|\mathbf{Z}}(p \mid \mathbf{z})}\right] dx dp d\mathbf{z} \qquad (3)$$

The partial information in (3) represents the partial dependence of X on P conditional to the pre-selected predictor set $\mathbf{Z}$. This is a measure of association that represents the joint dependence of all variables (X, P, and $\mathbf{Z}$) from which one has removed the dependence, the pre-existing predictor $\mathbf{Z}$ has on the other two variables (X and P).

Following (2), a sample estimate of the PI in (3) can be formulated as:

$$\hat{\mathrm{PI}}(X, P \mid \mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{f_{X|\mathbf{Z},P|\mathbf{Z}}(x_i, p_i \mid \mathbf{z}_i)}{f_{X|\mathbf{Z}}(x_i \mid \mathbf{z}_i) f_{P|\mathbf{Z}}(p_i \mid \mathbf{z}_i)}\right]$$

Representing the above conditional PDFs as ratios of the respective joint and marginal PDFs, the above expression can be re-arranged to be:

$$\hat{\mathrm{PI}}(X, P \mid \mathbf{Z}) = \hat{\mathrm{MI}}(X, P, \mathbf{Z}) - \hat{\mathrm{MI}}(X, \mathbf{Z}) - \hat{\mathrm{MI}}(P, \mathbf{Z}) + \hat{\mathrm{MI}}(\mathbf{Z}) \qquad (4)$$

For the above measure to work, it should be able to eliminate predictor variables that (a) are not related to the response, and (b) are predictable by the pre-existing predictor set, thus making them useless in a predictor identification context. It can be shown that the above measure of PI collapses to zero for the two cases mentioned.

Partial informational correlation (PIC)and Partial Weight (PW)

Mutual information can be transformed to a 0 to 1 scale, where 0 represents no dependence and 1 represents perfect dependence. The rescaled statistic is called Informational Correlation (IC) (Linfoot 1957), and is expressed as:

$$\hat{\mathrm{IC}} = \sqrt{1 - \exp(-2\hat{\mathrm{MI}})}$$

This transformation can also be applied to the partial information in (4); the resulting dependence measure being referred to as the Partial Informational Correlation (PIC):

$$\hat{\mathrm{PIC}} = \sqrt{1 - \exp(-2\hat{\mathrm{PI}})} \qquad (5)$$

Similarly, consideration of additional predictor variables in evaluating Euclidean distance metrics requires an assumption that all predictors are equally relevant in the estimation of the distance. This is not the case, and a partial weight metric is presented which ascertains the relative importance each predictor has in the final prediction. Readers are referred to Sharma and Mehrotra (2014) for details on the metric.

RESULTS

Two threshold autoregressive order two models as described in *Tong* [1990, pp.99-101] and *Sharma* [2000a] with the following structures are used:

TAR1

$$x_i = \begin{cases} -0.9x_{i-3} + 0.5\varepsilon_i & if\ x_{i-3} \le 0 \\ 0.4x_{i-3} + 0.5\varepsilon_i & if\ x_{i-3} > 0 \end{cases} \qquad (6)$$

TAR2

$$x_i = \begin{cases} 0.6x_{i-1} - 0.1x_{i-2} + \varepsilon_i & if\ x_{i-2} > 0 \\ -1.1x_{i-1} + \varepsilon_i & if\ x_{i-6} \le 0 \end{cases} \qquad (7)$$

Here, the first fifteen lags of the data (e.g., $x_{i-1}, x_{i-2}, \ldots, x_{i-15}$) are considered as potential candidate predictors from which the final predictor set is estimated.

In addition, an example where the underlying relationship is nonlinear and uses three exogenous covariates is also used. This model, denoted NLR for nonlinear regression, is written as:

NLR

$$y_i = 0.8x_{1_i} + 0.4{x_{2_i}}^2 + 0.7{x_{3_i}}^3 + \varepsilon_i \quad (8)$$

where, the response y_i is a function of three exogenous covariates x_1, x_2 and x_3, all normally distributed with a zero mean and unit variance. The three variables x_1, x_2 and x_3 and the four Gaussian random variates with a zero mean and unit variance are considered as potential candidate predictors from which the final predictor set is estimated.

Table 1 presents the results of these three examples in terms of the total number of predictors identified and the number of times a predictor is identified out of 100 realisations using PI and a k-nn regression model. The results indicate that majority of times the PI logic is able to identify the non-linear nature of the system whereas an assumed linear regression model, as should be expected when applied to a nonlinear dataset, fails to identify the correct predictors of the system in most cases.

Table 1. Frequency of correct selections and length of the resulting predictor vector using the PI and a linear regression approach – nonlinear autoregressive datasets. Given that the 100 samples the two approaches are applied to originate from markedly nonlinear datasets, it is to be expected that the linear regression approach will perform poorly compared to the PI.

Test example	Percent of samples the predictor is selected			Percent of samples total number of predictors identified		
	Variable	PI	Regression	Number of variables	PI	Regression
TAR1	No variable	0	0	0	1	99
	1	0	0	1	90	1
	2	0	0	2	8	0
	3	99	1	3	1	0
	4	0	0	4	0	0
	5	1	0	5	0	0
	6	2	0	6	0	0
	7	0	0	7	0	0
	8	0	0	8	0	0
	9	1	0	9	0	0
	10	1	0	10	0	0
	11	0	0	11	0	0
	12	2	0	12	0	0
	13	1	0	13	0	0
	14	0	0	14	0	0
	15	0	0	15	0	0
TAR2	0	0	0	0	0	0
	1	100	13	1	0	87
	2	87	5	2	98	8
	3	14	100	3	2	5
	4	1	0	4	0	0
	5	0	0	5	0	0
	6	0	0	6	0	0
	7	0	0	7	0	0
	8	0	0	8	0	0
	9	0	0	9	0	0
	10	0	0	10	0	0
	11	0	0	11	0	0

	12	0	0	12	0	0
	13	0	0	13	0	0
	14	0	0	14	0	0
	15	0	0	15	0	0
NLR	0	0	0	0	0	0
	1	97	100	1	1	0
	2	66	0	2	31	100
	3	100	100	3	65	0
	4	03	0	4	3	0
	5	02	0	5	0	0
	6	01	0	6	0	0
	7	0	0	7	0	0

CONCLUSIONS

This paper presented a nonparametric framework of system specification based on partial information (PI), as detailed in Sharma and Mehrotra (2014). PI is a partial measure of dependence derived from mutual information theory, which is theoretically able to measure dependence without needing to make assumptions about the form of the random variables being studied. We have used PI as an approach for predictor identification and subsequently used Partial Weights (PW) as a means of assigning the relative contribution of the identified predictors. We have shown that the application of this logic with optimum number of predictors to a range of linear, non-linear and dynamic synthetic datasets leads to outcomes that are better than or more or less comparable with the alternatives used currently.

While the approach presented is cast in a system specification context, where the response is directly assumed to be a function of the plausible predictors considered, extensions of the argument when the predictors need to be collectively transformed to a different variable space (as in principal component regression) are trivial to formulate. Further work will focus on extending the methods presented to take into account issues related to input uncertainty [*Chowdhury and Sharma*, 2007] and multiple responses, typical of the problems often dealt with in hydro-climatological applications.

REFERENCES

Chowdhury, S. and A. Sharma (2007), Mitigating Parameter Bias in Hydrological Modelling due to Uncertainty in Covariates, Journal of Hydrology 340(doi:10.1016/j.jhydrol.2007.04.010): 197-204.

Lall, U., and A. Sharma 91996), A nearest neighbor bootstrap for time series resampling, Water. Resour. Res., 32, 679-693.

Sharma, A., U. Lall and D.G. Tarboton, (1998), Kernel bandwidth selection for a first order nonparametric streamflow simulation model, Stochastic Hydrology and Hydraulics 12(1): 33-52.

Sharma, A., (2000a), Seasonal to interannual rainfall ensemble forecasts for improved water supply management: 1. A strategy for system predictor identification, J. Hydrol., 239, 232-239.

Sharma, A. (2000b), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3 - A nonparametric probabilistic forecast model." Journal of Hydrology 239(1-4): 249-258.

Sharma, A., K. C. Luk, I. Cordery and U. Lall (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 2 - Predictor identification of quarterly rainfall using ocean-atmosphere information, Journal of Hydrology 239(1-4): 240-248.

Sharma, A., and R. Mehrotra (2014), An information theoretic alternative to model a natural system using observational information alone, Water Resources Research, 49, doi:10.1002/2013WR013845.

Tong, H. (1990), Nonlinear Time Series Analysis: A Dynamical Systems Perspective, Academic, San Diego, California, USA.