Improving Partial Mutual Information Based Input Variable Selection for Data Driven Environmental and Water Resources Models

XUYUAN LI

B.E. (Hons)

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Civil, Environmental and Mining Engineering

Improving Partial Mutual Information Based Input Variable Selection for Data Driven Environmental and Water Resources Models

By:

Xuyuan Li, *B.E. (Hons)*

Supervised by:

Professor Holger R. Maier, *B.E. (Hons), Ph.D., MIEAust, CPEng (NPER)*
Professor of Integrated Water Systems Engineering, Associate Editor, Water Resources Research, Member of Editorial Board, Environmental Modelling and Software, School of Civil, Environmental & Mining Engineering,
The University of Adelaide

Dr. Aaron C. Zecchin, *Ph.D., B.E. (Hons), B.Sc. (Math & Comp. Sci.),*
Senior lecturer, School of Civil, Environmental & Mining Engineering,
The University of Adelaide

Thesis submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy

School of Civil, Environmental & Mining Engineering
Faculty of Engineering, Computer and Mathematical Sciences
The University of Adelaide
North Terrace, Adelaide, SA 5005, Australia
Phone: +61 8 8313 1575
Fax : +61 8 8303 4359
Email: xli@civeng.adelaide.edu.au, xliadelaide@gmail.com

TABLE OF CONTENT

LIST OF FIGURES

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES

NOMENCLATURE & ABBREVIATIONS

Symbols

$\hat{m}_{v_i}(X_{i^*}^j)$: residual estimate of v_i based on X_{i^*}

$(\boldsymbol{X}^j, y^j)$: observed pairs of input and output data

$\hat{\varphi}_4(g) = n^{-1}\sum_{i=1}^{n}\hat{L}^{(4)}(X^i; g)$: fourth order integrated squared density derivative

$F_{emp}(X_i^j)$: empirical CDF of the input variable estimated by a histogram

$F_{est}(X_i^j)$: estimated kernel based CDF of the input variable

$I_{X_i,y}$: mutual information

$I_{\boldsymbol{v}_i,u}$: partial mutual information

$\widehat{ISB}(h)$: estimation of the integrated squared bias

$K^{(n)}$: n^{th} derivative of kernel function K

$\widetilde{R(f'')}$: approximation of the integrated squared second derivative of f

$S_{x,i}^2$: sample variance of the input X_i

$S_{xy,i}$: covariance between input X_i and output y

S_y^2: sample variance of output y

X_{i^*} selected inputs

$X^{(j)}$: *j-th* data point formed by the interpolated and original data points

X_s: significant input set

$\hat{f}(\boldsymbol{X}, y)$: estimation of the joint probability density function between inputs $\boldsymbol{X}$ and output y

p_{t-n}: exogenous input with lag n

$\hat{y}$: estimation of the actual output y

$\bar{y}$: sample mean of the observations

$\boldsymbol{e_1}$: vector having 1 in the first entry and 0 elsewhere

ε_t: introduced error term

$\mu_2(K) = \int x^2K(x)dx$: second moment of K

$\mu_k(L) = \int u^kL(u)du$: *k-th* moment of L

$\mu_n(K)$: n^{th} moment of kernel function K

$\rho_{xy,i}$: correlation coefficient between input X_i and output y

σ_i: sample standard deviation of the X_i^j

$*$: convolution operation

h: kernel bandwidth

K: kernel function

$B(u; h_x)$: univariate boundary kernel with bandwidth h_x and variable $u = (X_i - X_i^j)/h_x$

$B(u, v; \boldsymbol{H})$: bivariate boundary kernel with bandwidth matrix $\boldsymbol{H}$ and vector (u, v) where $u = (X_i - X_i^j)/h_x$ and $v = (y - y^j)/h_y$

$E[y|\boldsymbol{X}]$: conditional expectation of output y given input $\boldsymbol{X}$

L: pilot kernel

$O(h)$: bias of density function

$P(t)$: lagged effective rainfall

$Q(t-1)$: lagged runoff

$R(K) = \int [K(x)]^2 dx$: integrated square of kernel function

a: left boundary of kernel density

c: right boundary of kernel density

e: number of effective inputs

$f(\boldsymbol{X}, y)$: joint probability density function between inputs $\boldsymbol{X}$and output y

g: pilot bandwidth

k: kurtosis

k: order of pilot kernel L

$m(x)$: regression function

m: number of inputs

n: number of observations

r: stage number of L

s: skewness

sup: supremum function

$\boldsymbol{H} = h_i^2 \begin{bmatrix} S_{x,i}^2 & S_{xy,i} \\ S_{xy,i} & S_y^2 \end{bmatrix}$: bivariate bandwidth matrix

$\boldsymbol{X} = [X_1 \ldots X_m]^T$: input variables

$\boldsymbol{h} = [h_1 \quad \cdots \quad h_n]^T$: kernel bandwidth vector

Abbreviation

ACF: auto-correlation function

AIC: Akaike information criterion

AMISE: asymptotic mean integrated squared error

ANNs: artificial neural networks

BCV: biased cross validation

BCVDPI: a combination of BCV and DPI

BE: backward elimination (pruning)

BJ: Box-Jenkins

BK: boundary kernel

BP: back-propagation algorithm

CE: coefficient of efficiency

CK: conventional kernel

CSR: correct selection rate

CT: computational time

DELSA: distributed evaluation of local sensitivity analysis

DPI: 2-stage direct plug-in

EAR4: exogenous auto-regressive time series model (with time order up to 4)

EMISE: exact mean integrated squared error

ES: exhaustive search

ETC: empirical translation correction

EVT1: extreme value type I distribution

EXP: exponential distribution

FS: forward selection

GAMMA: gamma distribution

GRNN: general regression neural network

GRR: Gaussian reference rule

GSS: golden section search

HS: heuristic search

ICAIVS: hybrid independent component analysis and input variable selection filter

IIS: tree-based iterative input variable selection

IoAd: index of agreement

IVS: input variable selection

KDEs: kernel density estimates

K-S: Kolmogorov-Smirnov statistic

KT: kernel transformation

L1UL: Lock 1 upper river level

LBE: local bandwidth (enlarging)

LBR: local bandwidth (reducing)

LHOP: local high order polynomial

LLM: local linear method

LLP: local linear polynomial

LOGN: log-normal distribution

LOGPT3: log-Pearson type III distribution

LOS: Loxton river salinity

LQP: local quadratic polynomial

LSCV: least squared cross validation

MAE: mean absolute error

MAS: Mannum river salinity

MBS: Murray Bridge river salinity

MCE: modified coefficient of efficiency

MI: mutual information

MIoAd: modified index of agreement

MLPANNs: multi-layer perceptron artificial neural networks

MLPs: multi-layer perceptrons

MOS: Morgan river salinity

MPI: modified persistence index

MVC: multi-variable calibration

MVCA: multi-variable calibration with mean absolute error as the objective function

MVCS: multi-variable calibration with squared error as the objective function

NL: nonlinear input-output function

NORM: normal distribution

NS: normal scale

OM: optimisation method

PA: pseudo-data approach

PACF: partial auto-correlation function

PC: partial correlation

PCA: principal component analysis

PDF: probability density function

PI: persistence index

PMI: partial mutual information

PNNs: probabilistic neural networks

PSO: particle swarm optimisation

PT3: Pearson type III distribution

RBFs: radial basis functions

RC: reflection correction

RE: residual estimation

RMSE: root mean squared error

RNNs: recurrent neural networks

RVSDEM: recursive variable selection embedded in dynamic emulation models

SCV: smoothed cross validation

SOM-GAGRNN: self-organising map genetic algorithm general regression neural network

SVC: single variable calibration

SVCA: single variable calibration with mean absolute error as the objective function

SVCS: single variable calibration with squared error as the objective function

SVO: single variable optimisation

SVR: single variable regression

TEAR10: threshold exogenous auto-regressive time series model (with time order up to 10)

WAS: Waikerie river salinity

ABSTRACT

Artificial neural networks (ANNs), as one of the most commonly used data driven models for environmental and water resources problems, have been applied successfully and extensively over the last two decades and are still gaining in popularity. Consideration of the methods used in the steps in the development of ANNs, which consist of data collection, data processing, input variable selection, data division, calibration and validation, are vitally important, as ANN model development is based on data, rather than understanding of the underlying physical processes.

Among these methods, input variable selection (IVS) plays a significant role, as the performance of the developed model can be compromised if inputs having a pronounced relationship with the modelled output are omitted. In contrast, calibration becomes extremely challenging and modelling validation, as well as knowledge extraction, are problematic if redundant or superfluous inputs are included. Given the facts explained above, various techniques have been developed for the sake of more accurate IVS.

Partial mutual information (PMI) is one of the most promising approaches to IVS, as it has a number of desirable properties, such as the ability to account for input relevance, the ability to cater to both linear and non-linear input-output relationships and the ability to check the redundancy of selected inputs. PMI is a stepwise input selection algorithm, which only selects one variable per iteration, as part of which the strength of the relationship between each potential input and the output is quantified using mutual information (MI) and input redundancy is accounted for by removing the influence of already selected inputs. This is achieved by developing models between the selected input and the output and assessing the strength of the relationship (in terms of MI) between the remaining potential inputs and the residuals of these models during the next iteration, which is referred to as PMI.

Although PMI IVS has already been applied successfully to a number of studies in hydrological and water resource modelling, present implementations predominantly depend on the assumption that the data used to develop the model follow a Gaussian distribution. This assumption has the potential to affect two steps in the PMI process, including the estimation of MI/PMI and the estimation of the residuals. In terms of MI/PMI estimation, this requires kernel density estimates of the modelling data to be obtained for the estimation of marginal and joint probability density functions (PDFs), which rely on estimates of kernel bandwidths (or smoothing parameters) and in most studies, the Gaussian reference rule is used for this purpose, which only results in optimal bandwidth estimates if the modelling data follow a Gaussian distribution. However, this is unlikely to be the case when dealing with water resources and other environmental data. In terms of residual estimation (RE), this has generally been done using general regression neural networks (GRNNs), which also require estimates of kernel bandwidths to be obtained and therefore suffers from the same issues as MI/PMI estimation.

The purpose of this thesis is to assess the impact the assumption that the data follow a Gaussian distribution has on the performance of PMI IVS and the efficacy of potential methods for overcoming any problems associated with this assumption. In order to achieve this, a large number of numerical tests are conducted on synthetic data with different degrees of normality and non-linearity, investigating the effectiveness of a range of options for (i) bandwidth estimation (caused by making Gaussian assumptions for non-Gaussian circumstances when adopting kernel based estimations in both MI/PMI and RE) and (ii) for dealing with boundary issues (caused by using a symmetrical kernel for bounded/unsymmetrical data when implementing kernel based estimations in both MI/PMI and RE), as well as methods for RE that do not require kernel density estimates. The results from these tests are used to develop preliminary guidelines for the selection of the most appropriate bandwidth and the most effective treatment of the boundary issue, which are then validated for two water resources case studies with different data properties and problem linearity, including forecasting of river salinity in the River Murray, Australia, and rainfall-runoff modelling in the Kentucky River, USA.

The major research contributions are presented in three journal publications. The motivations underlying these publications include: 1) the development and testing of rigorous and novel analytical procedures for assessing if, and to what degree, the performances of residual and MI estimates are affected by bandwidth selection and boundary issues; 2) clear explanation of the inaccurate performance of conventional PMI IVS under the influence of bandwidth selection and boundary issues; 3) the development of effective preliminary guidelines based upon synthetic studies dealing with both bandwidth selection and boundary issues under different scenarios categorised by data normality and problem linearity; 4) the development of more robust and reliable PMI IVS software for realistic environmental and water resource problems. Overall, the research outcomes suggest that the performance of PMI IVS is significantly influenced by bandwidth selection and boundary issues and can be effectively improved by following the proposed empirical guidelines, although the findings of this work could be tested more broadly, including for data sets with a wider range of attributes, such as different degrees of noise, collinearity and interdependency, as well as incomplete information.

STATEMENT OF ORIGINALITY

I **Xuyuan Li** hereby certify that this work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

A list of works contained within this thesis is given in Section 5.3.

Signature__________________________ Date___________________________

ACKNOWLEDGEMENTS

I would take this opportunity to gratefully acknowledge my principal supervisor Professor Holger R. Maier and co-supervisor Dr. Aaron C. Zecchin, who have supervised me with invaluable support, encouragement, and patience. I really appreciate the critical and rigorous attitude, timely and effective feedback, great foresight, and comfortable research environment provided by both of my supervisors. I have enjoyed studying under their supervision and learnt a lot from them during my candidature. Without their excellent supervision, the present thesis would hardly be possible.

I would like to sincerely thank Prof. A. Sharma, Dr. G.J. Bowden, Dr. R.J. May and Dr. G.B. Humphrey who brought me into the present research topic and kindly provided their suggestions and research materials.

I would express my appreciation to my fellow postgraduates within the School for sharing their wisdom, experiences, successes, and lessons with me. It was my pleasure to meet Dr. Xun Sun, Dr. Wenyan Wu, Dr. Feifei Zheng, Dr. Liang Huang, Dr. Jeffrey P. Newman, Dr. Christopher Stokes, and Dr. Tao Zhang and have had a number of in-depth academic discussions with them. Special thanks are also given to School Editor Barbara Brougham, Computer Technician Dr. Stephen Carr, all School Administrators, and the research scholarship (AGRS) provided by the University of Adelaide.

Last but not least, I would like to give an immense gratitude to my parents, Mr Xuelong Li and Ms Yongjing Liu, for their altruistic care and support from beginning to end (*ab ovo usque ad mala*).