IVS framework datasets

Greer B. Humprhey, Stefano Galelli*, Holger R. Maier
Andrea Castelletti, Graeme C. Dandy, Matthew S. Gibbs

1 AR

Equations for the two linear AR models are given by:

$$AR1 : x_t = 0.9x_{t-1} + 0.866\epsilon_t \tag{1}$$

$$AR9 : x_t = 0.3x_{t-1} - 0.6x_{t-4} - 0.5x_{t-9} + \epsilon_t \tag{2}$$

where ϵ is random Gaussian noise with zero mean and unit standard deviation. These models were used for testing the PMI IVS algorithm in Sharma (2000). For each model, x_t was arbitrarily initialised and a total of $N + 500$ data points were generated, where N was set equal to 500 and 70 (see Humprhey et al. (2014)). The first 500 points were discarded to reduce any effects from the arbitrary initialisation. A total of 15 candidate inputs, $x_{t-1}, x_{t-2}, \ldots, x_{t-15}$, were generated using each of the models. Only one and three of these 15 inputs are relevant for modelling the AR1 and AR9 outputs, respectively.

2 TAR

The nonlinear TAR models are given as follows:

$$TAR1 : x_t = \begin{cases} -0.9x_{t-3} + 0.1\epsilon_t & \text{if } x_{t-3} \leq 0 \\ 0.4x_{t-3} + 0.1\epsilon_t & \text{if } x_{t-3} > 0 \end{cases} \tag{3}$$

$$TAR2 : x_t = \begin{cases} -0.5x_{t-6} + 0.5x_{t-10} + 0.1\epsilon_t & \text{if } x_{t-6} \leq 0 \\ 0.8x_{t-10} + 0.1\epsilon_t & \text{if } x_{t-6} > 0 \end{cases} \tag{4}$$

As with the AR models, these models were also used in Sharma (2000). Again, x_t was arbitrarily initialised and a total of $N + 500$ data points were generated, with the first 500 points discarded to reduce initialisation effects. For both models, N was set equal to 500.

*Corresponding author. Tel: +65 6499 4786, E-mail address: stefano_galelli@sutd.edu.sg

3 NL

The NL family of datasets were generated using the following equation from Bowden et al. (2005a):

$$y = x_2^2 + \cos(x_6) + 0.35 \sin(x_9) + s\epsilon \tag{5}$$

where s is a scaling factor used to alter the level of noise in the output. A total of 15 candidate inputs, $x_1, x_2, \ldots, x_{15}$, were randomly sampled from $x \sim N(0, \Sigma)$, where Σ is a predefined covariance matrix. Similar to the AR models, $N + 500$ data points were generated, where N was set equal to 500 and 70 (see Humprhey et al. (2014)). The first 500 points were discarded to reduce any random initialisation effects. For the 'NL_500' and 'NL_70' datasets (see Humprhey et al. (2014)), s was set equal to zero, such that the output y contains no noise. Furthermore, Σ was set equal to the identity matrix; thus, the 15 candidate inputs were sampled independently of one another (i.e. there was no collinearity between the candidate input variables). For the 'NL2' dataset, on the other hand, s was set equal to one, resulting in a high level of noise in the output (according to the definitions of noise levels given above). In addition, for this dataset, the covariance matrix Σ was defined such that 25 pairs of inputs were highly correlated (correlation > 0.7) and the 15 candidate inputs were generated accordingly. Thus, the resulting degrees of collinearity for this dataset was equal to 25.

4 Bank

The 'Bank' family of datasets are a slight variation of a subset of the original Bank regression benchmark datasets provided on the DELVE repository (Rasmussen et al., 1996). These datasets were generated from a simplistic simulator of daily banking activity, which returns the rejection rate (rej) or fraction of customers that are turned away from banks because all the open tellers have full queues.

Within the Bank simulator, customers come from three different residential areas (a1, a2, a3), each having a defined centre location (a1cx, a1cy; a2cx, a2cy; a3cx, a3cy) and population size (a1pop, a2pop, a3pop). There are three different banks that customers may choose from (b1, b2, b3), which also have defined locations (b1x, b1y; b2x, b2y; b3x, b3y). Customers are generated from the different residential areas according to a Poission distribution with an intensity given by the area's population. They then choose their preferred bank depending on the distance from their home, where the location of their home is randomly generated from a Gaussian distribution with covariance matrix specified according to the residential area to which the customer belongs. The choice of bank follows the Boltzmann distribution, which also depends on a fictitious temperature (temp) parameter. Each bank has a number of open tellers with queues of varying lengths and these tellers may open and close according to demand. The tellers have various efficiencies,

while the customers have tasks of varying complexity and various levels of patience and may change queue if their patience expires. A maximum queue length (mxql) is specified and customers are rejected from the bank if the queues at all open tellers are already of this length.

There are two versions of the Bank model provided on the DELVE repository: one with an 8-dimensional input vector and the other with a 32-dimensional input vector. In fact, both versions depend on the same 32 input variables; however, the 8-dimensional version provides no information about the values of the remaining 24 inputs. Thus, this represents the special case where there is incomplete information about the target data. Additionally, there are 4 instances of each model version:

- fairly linear with moderate noise (fm);
- fairly linear with high noise (fh);
- nonlinear with moderate noise (nm); and
- nonlinear with high noise (nh).

For the purposes of the IVS evaluation framework, only the four instances of the 8-dimensional input version of the model are used. The 8 input variables for which values are provided include a1cx, a1cy, b2x, b2y, a2pop, a3pop, temp and mxql. To fill the candidate input pool, the remaining 24 input variables from the corresponding 32-dimensional input versions were used. These remaining inputs are considered to be irrelevant because, although these variables were used to generate the target data, the particular random realisations of the variables provided in the 32-dimensional versions were not. As such, each of the Bank datasets included in the proposed framework contains 32 candidate inputs (i.e. 8 relevant and 24 irrelevant inputs).

The original Bank datasets each include 8192 samples. From these, 400 samples were randomly selected to make up each of the dataset replicates; thus there is some minor overlap between the 30 replicates. For further details, see Humprhey et al. (2014).

5 Friedman

The 'Friedman' family of datasets were generated by the following function, suggested by Friedman (1991):

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + s\epsilon \tag{6}$$

Again, s is a scaling factor used to alter the level of noise in the output. A total of P input variables $x_1, x_2, \ldots, x_P$ were sampled from $x \sim U(0, 1)$. To enable an

evaluation of an IVS algorithm's performance given different proportions of extraneous inputs, P was set equal to both 10 and 50 (only the first five inputs are relevant). Furthermore, both moderate (m) and high (h) noise levels were added to the generated output to enable sensitivities of IVS algorithms to different noise levels to be investigated (see Humprhey et al. (2014)).

In the case that $P = 10$, datasets were generated with both zero and 25 degrees of collinearity (corresponding to 'c0' and 'c25', respectively), in order to measure the robustness of an algorithm to input collinearity. The 10 candidate inputs were generated such that they exhibited a pre-specified linear correlation according to the method described by Schumann (2009) for generating correlated uniform variates. Similar to the 'NL2' dataset, the covariance matrix was defined such that 25 pairs of inputs were highly correlated (correlation > 0.7). When the degree of collinearity was equal to zero, the candidate inputs were simply sampled independently from $x \sim U(0, 1)$. Accounting for the different dimensionalities, noise levels and degrees of collinearity, a total of six Friedman datasets were generated (with 30 replicates of each), as shown in Humprhey et al. (2014). In each case, the length of the dataset was $N = 250$.

The Friedman regression function (sometimes known as 'add10') given by eq. (6) has been used in numerous studies for benchmarking IVS (e.g. Chang (2013); Doquire and Verleysen (2012); Haleem and Abdel-Aty (2011); Loh (2012); Yang and Ong (2011); Tikka (2009); Fernando et al. (2009); Doksum et al. (2008); May et al. (008a)) and other regression methods. Additionally, there are instances of the Friedman model available on the DELVE (Rasmussen et al., 1996) and WEKA (Hall et al., 2009) repositories. However, many of the aforementioned studies have generated new realisations of the dataset, rather than using those provided on the data repositories. Moreover, the length and dimensionality of the generated datasets were generally not varied, nor was the level of noise in the output or the degree of collinearity amongst the inputs. The Friedman datasets provided on the WEKA platform by Amasyali and Ersoy (2009), however, include 80 different realisations of the model with different dimensionalities, different sample sizes and different degrees of collinearity. Yet, despite this desired variation in dataset properties, the candidate inputs in these datasets are not uniformly distributed on the range $[0, 1]$, as originally proposed by Friedman (1991), due to the method used to induce collinearity. This affects the way in which the individual inputs influence the behaviour of the output (primarily inputs x_1 and x_2). Furthermore, replicates of these 80 realisations have not been provided.

6 Salinity

The 'Salinity' datasets are based on a real water resources dataset used for forecasting salinity in the River Murray at Murray Bridge, South Australia, 14 days

in advance. This case study has previously been used in numerous studies related to the development and testing of statistical modelling methods (Maier and Dandy, 1996, 1998a,b; Bowden et al., 2002, 2003; Kingston et al., 2005, 2008), including IVS methods (Bowden et al., 2005b; Fernando et al., 2009). Therefore, it was considered to provide a good benchmark dataset for the purposes of IVS evaluation. However, as the "true" inputs that are relevant for modelling these data are unknown, an artificial neural network was used to generate a synthetic output time series closely resembling the natural data, but with known inputs. In the IVS study conducted by Fernando et al. (2009), the inputs found to be most relevant for forecasting Murray Bridge salinity with a 14 day lead time (MBS_{t+13}) included salinities at Mannum (MAS_{t-1}) and Waikerie (WAS_{t-1}), as well as flow at lock 7 ($L7F_{t-1}$), all at a time lag of 1 day. These inputs make sense from a physical point of view, since during times of low flow ($\approx$ 6500 ML/day), the travel time between Mannum and Murray Bridge is approximately 14 days, while, during times of intermediate flow (16,000–17,000 ML/day), the travel time between Waikerie and Murray Bridge is approximately 14 days (Maier and Dandy, 1996). The third input variable, $L7F_{t-1}$, has been found to improve the accuracy in forecasting high values of salinity, since such salinities are associated with very low flow values (Fernando et al., 2009). The resulting ANN model used to generate the synthetic salinity data took the following form:

$$MBS_{t+13} = f(MAS_{t-1}, WAS_{t-1}, L7F_{t-1}) + s\epsilon \tag{7}$$

where $f()$ represents a single hidden layer, four hidden node, fully connected multilayer perceptron (MLP).

The available daily dataset spans the period from December 1996 to March 1998 and includes salinities at five locations, flows at three locations and river levels at eight locations at and upstream from Murray Bridge (a total of 16 input variables). The candidate input pool was filled with time lagged values of each of these 16 variables, $\left\{x_{j,t-1}, x_{j,t-2}, \ldots, x_{j,t-L}; j = 1, \ldots, 16\right\}$, where L is the maximum time lag included. To allow the investigation of IVS algorithm performance given different proportions of irrelevant candidate inputs, L was set equal to both 5 and 10, producing datasets with 80 and 160 potential inputs, respectively. Accounting for the appropriate lags of the input and output variables resulted in datasets comprised of 4120 and 4115 samples for the 5 and 10 lag datasets, respectively. Furthermore, the noise scaling factor, s, in eq. (7) was varied such that low (l), moderate (m) and high (h) noise levels were added to the generated output to enable sensitivities of IVS algorithms to different noise levels to be investigated. This resulted in a total of six Salinity datasets, as shown in Humprhey et al. (2014).

As the candidate inputs for the Salinity model are real observed data, in order to generate the 30 replicates of the Salinity datasets, only the random noise component, ϵ, in eq. (7) was randomly regenerated. Thus, for the low noise datasets

there is very little variation in the 30 output time series. However, datasets were randomly perturbed to remove any time structure in the input and output variables.

7 Kentucky

Similar to the Salinity datasets, the 'Kentucky' datasets are derived from natural rainfall-runoff data available for the Kentucky River basin, with output data synthetically generated by an ANN model. The original Kentucky dataset is described in Jain et al. (2004); Jain and Srinivasulu (2006) and has been utilised for assessing and comparing statistical modelling methods in various studies (Srinivasulu and Jain, 2009; Wu et al., 2012; Bowden et al., 2012). The dataset consists of 26 years (1960-1972 and 1977-1989) of average daily streamflow data (cfs) from the Kentucky River at LD10 and average daily rainfall (mm) from five rain gauges within the Kentucky River catchment. Jain and Srinivasulu (2006) transformed the five rainfall series into a single effective rainfall (Er) variable and used these data at time steps t, t-1, and t-2 (Er_t, Er_{t-1}, Er_{t-2}) together with flow values at time steps t-1 and t-2 (Q_{t-1}, Q_{t-2}) to model the flow at time t (Q_t). These inputs were selected based on correlation analyses between Er and Q values on the present flow value. However, in developing the ANN model used for generating the synthetic flow series, it was found that input Er_{t-2} only provided a very minor contribution in predicting the output flows and that a model with better generalisation ability could be obtained when this input was excluded from the model. Thus, the resulting ANN model used to generate the synthetic flow data took the following form:

$$Q_t = f(Q_{t-1}, Q_{t-2}, Er_t, Er_{t-1}) + \epsilon \tag{8}$$

where $f(\cdot)$ represents a single hidden layer, four hidden node, fully connected MLP. The candidate input pool was filled with time lagged values of Q and Er, as well as the current value of Er: $\{Er_t, Er_{t-1}, Er_{t-2}, \ldots, Er_{t-L}, Q_{t-1}, Q_{t-2}, \ldots, Q_{t-L}\}$, where L was set equal to 10. Accounting for the appropriate lags of the input and output variables resulted in a dataset comprised of 4739 samples and 21 potential inputs (see Humprhey et al. (2014)).

Similar to the Salinity datasets, in order to generate the 30 replicates of the Kentucky dataset, only the random noise component, ϵ, in eq. (8) was randomly regenerated. Again, the datasets were randomly perturbed to remove any time structure in the input and output variables.

8 Miller

The 'Miller' datasets are based on the function provided in Miller (1984, 2002) where:

$$y = x_1 - x_2 \tag{9}$$

and y is orthogonal to x_1 and almost orthogonal to x_2. This model gives an example of a situation where the separate variables x_1 and x_2 have little or no predictive value on their own, but jointly, they perfectly describe the output. Consequently, IVS methods which only consider the relevance of a single variable at a time are unlikely to select the correct input subset.

For this framework, 200 samples of each variable were generated, where all variables were integer valued, as follows:

$$\begin{aligned} x_1 &= \begin{cases} -10 & \text{if } r < 0.5, \\ 10 & \text{otherwise.} \end{cases} \\ x_2 &= x_1 + i \end{aligned} \tag{10}$$

where $r \sim U(0, 1)$ and i was generated from a discrete uniform distribution on the range $[-3, 3]$. As a result, y is also integer valued on the range $[-3, 3]$. Although variables x_1 and x_2 are not correlated with y, they are highly correlated with one another (correlation > 0.9). This example illustrates that very high variable correlation does not necessarily infer redundancy.

Similar to Miller (1984, 2002), a third candidate input variable, x_3, was generated such that it was linearly correlated with y (correlation ≈ 0.6), with values between -4 and 4. Given that the correlation between x_3 and y is significantly greater than the correlation between either x_1 or x_2 and y, IVS methods which only consider the relevance of a single variable at a time are likely to select a suboptimal input subset for this dataset.

References

Amasyali, M. F. and Ersoy, O. K. (2009). A study of meta learning for regression. Ece technical reports.

Bowden, G. J., Dandy, G. C., and Maier, H. R. (2003). Data transformation for neural network models in water resources applications. *Journal of Hydroinformatics*, 5(4):245–258.

Bowden, G. J., Maier, H. R., and Dandy, G. C. (2002). Optimal division of data for neural network models in water resources applications. *Water Resources Research*, 38(2):2–1.

Bowden, G. J., Maier, H. R., and Dandy, G. C. (2005a). Input determination for neural network models in water resources applications. Part 1. Background and methodology. *Journal of Hydrology*, 301(1-4):75–92.

Bowden, G. J., Maier, H. R., and Dandy, G. C. (2005b). Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river. *Journal of Hydrology*, 301(1-4):93–107.

Bowden, G. J., Maier, H. R., and Dandy, G. C. (2012). Real-time deployment of artificial neural network forecasting models: Understanding the range of applicability. *Water Resources Research*, 48(10):W10549.

Chang, Y. (2013). Variable selection via regression trees in the presence of irrelevant variables. *Communications in Statistics - Simulation and Computation*, 42(8):1703–1726.

Doksum, K., Tang, S., and Tsui, K.-W. (2008). Nonparametric variable selection: The EARTH algorithm. *Journal of the American Statistical Association*, 103(484):1609–1620.

Doquire, G. and Verleysen, M. (2012). Feature selection with missing data using mutual information estimators. *Neurocomputing*, 90(0):3–11.

Fernando, T. M. K. G., Maier, H. R., and Dandy, G. C. (2009). Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach. *Journal of Hydrology*, 367(3–4):165–176.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Haleem, K. and Abdel-Aty, M. (2011). Application of GLASSO in variable selection and crash prediction at unsignalized intersections. *Journal of Transportation Engineering*, 138(7):949–960.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

Humprhey, G., Galelli, S., Maier, H., Castelletti, A., Dandy, G., and Gibbs, M. (2014). An evaluation framework for input variable selection algorithms for environmental data-driven models. *Environmental Modelling & Software*, -:–. Submitted.

Jain, A. and Srinivasulu, S. (2006). Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques. *Journal of Hydrology*, 317(3–4):291–306.

Jain, A., Sudheer, K. P., and Srinivasulu, S. (2004). Identification of physical processes inherent in artificial neural network rainfall runoff models. *Hydrological Processes*, 18(3):571–581.

Kingston, G. B., Lambert, M. F., and Maier, H. R. (2005). Bayesian training of artificial neural networks used for water resources modeling. *Water Resources Research*, 41(12):W12409.

Kingston, G. B., Maier, H. R., and Lambert, M. F. (2008). Bayesian model selection applied to artificial neural networks used for water resources modeling. *Water Resources Research*, 44(4):W04419.

Loh, W.-Y. (2012). *Variable selection for classification and regression in large p, small n problems*, chapter 10, pages 135–159. Lecture Notes in Statistics. Springer New York.

Maier, H. R. and Dandy, G. (1996). The use of artificial neural networks for the prediction of water quality parameters. *Water Resources Research*, 32(4):1013–1022.

Maier, H. R. and Dandy, G. C. (1998a). The effect of internal parameters and geometry on the performance of back-propagation neural networks: An empirical study. *Environmental Modelling and Software*, 13(2):193–209.

Maier, H. R. and Dandy, G. C. (1998b). Understanding the behaviour and optimising the performance of back-propagation neural networks: an empirical study. *Environmental Modelling and Software*, 13(2):179–191.

May, R. J., Maier, H. R., Dandy, G. C., and Fernando, T. M. K. G. (2008a). Nonlinear variable selection for artificial neural networks using partial mutual information. *Environmental Modelling & Software*, 23(10–11):1312–1326.

Miller, A. J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, 147(3):389–425.

Miller, A. J. (2002). *Subset Selection in Regression*. Monographs on Statistics and Applied Probability. Chapman & Hall / CRC, 2nd edition.

Rasmussen, C. E., Neal, R. M., Hinton, G. E., van Camp, D., Revow, M., Ghahramani, Z., Kustra, R., and Tibshirani, R. (1996). Data for evaluating learning in valid experiments (delve).

Schumann, E. (2009). Generating correlated uniform variates.

Sharma, A. (2000). Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - a strategy for system predictor identification. *Journal of Hydrology*, 239(1-4):232–239.

Srinivasulu, S. and Jain, A. (2009). River flow prediction using an integrated approach. *Journal of Hydrologic Engineering*, 14(1):75–83.

Tikka, J. (2009). Simultaneous input variable and basis function selection for rbf networks. *Neurocomputing*, 72(10–12):2649–2658.

Wu, W., May, R., Dandy, G. C., and Maier, H. R. (2012). A method for comparing data splitting approaches for developing hydrological ann models. In *Proceedings of the 6th Biennial Meeting of the International Environmental Modelling and Software Society, Leipzig, D*.

Yang, J.-B. and Ong, C.-J. (2011). Feature selection using probabilistic prediction of support vector regression. *Neural Networks, IEEE Transactions on*, 22(6):954–962.