A WAVELET TRANSFER MODEL FOR TIME-SERIES FORECASTING

Dr. DAMIEN FAY

Department of Mathematics, National University of Ireland, Galway,
Galway, Ireland
damien.fay@nuigalway.ie
http://www.maths.nuigalway.ie/~damien

Prof. JOHN RINGWOOD

Department of Electronic Engineering, National University of Ireland, Maynooth,
Maynooth,Ireland
john.ringwood@eeng.nuim.ie
http://www.eeng.may.ie/~jringwood

Received (19-09-2005)
Revised (-)

This paper is concerned with the case of an exogenous system in which a model is required to forecast a periodic output time-series using a causal input. A novel approach is developed in which the wavelet packet transform is taken of both the dependent time series *and* causal input. This results in two sets of basis dictionaries and requires two bases to be chosen. It is proposed that the best bases to choose are those which maximize the *mutual information*. Input selection is then implemented by eliminating those coefficients of the selected input basis with low mutual information. As an example, a model is constructed to forecast short-term electrical demand.

Keywords: wavelet packets; time-series; load forecasting.

1. Introduction.

Time series forecasting is concerned with forecasting a dependant time series, $y(k)$, with a set of causal variables, $U(k)$, by using a model, $f(\cdot)$, as:

$$y(k) = f(U(k)) + \varepsilon(k) \tag{1}$$

where $\varepsilon(k)$ is a residual term. However, estimation of $f(\cdot)$ is often a difficult task. This task may be aided by transforming the inputs and/or outputs into new domains *prior* to modeling as:

$$A\big(y(k)\big) = f'\big(B\big(U(k)\big)\big) + \varepsilon'(k) \tag{2}$$

where $A(\cdot)$ represents the output transform (or *output filtering*), $B(\cdot)$ represents the input transform (or *input pre-processing*), $\varepsilon'(k)$ is a residual term (note: $\varepsilon'(k) \neq \varepsilon(k)$ in general) and $f'(\cdot)$ denotes the new model. The purpose of $B(\cdot)$ is to eliminate non-causal inputs and reduce multicollinearity (cross-correlation) in the inputs [Ljung (1999)]. The purpose of $A(\cdot)$ is to transform the *dependent* time series, $y(k)$, into a time series that is more correlated to the input. In addition, the distribution of the residual term is altered which

can be advantageous, especially if the distribution of the original residual term, $\varepsilon(k)$, is non-Gaussian [Ljung (1999)].

Several types of transform have been applied in time series forecasting such as Principle Component Analysis (PCA) [Hiden *et al*. (1999)], Independent Component Analysis (ICA) [Roberts *et al*. (2004)], the Fourier Transform (FT) [Schoukens and Pintelon (1991)], the Wavelet Transform (WT) [Yao *et al*. (2000)] and the Wavelet Packet Transform (WPT) [Saito and Coifman (1997); Roberts *et al*. (2004); Milidiú *et al*. (1999); Nason and Sapatinas (2001)] among others. However, the WT and WPT would seem ideal for time series forecasting as unlike PCA, ICA and the FT, some time information is preserved in the transformed variables. In addition, the WPT allows an *adjustable* trade-off between time and frequency resolution in the transformed signal. The FT and WT have been used to transform *both* the input and output of a system prior to modeling [Schoukens and Pintelon (1991), Liu, (2005); Labat *et al*. (2000)]. However, the WPT has *not* been widely used for this purpose. The wavelet transfer model proposed in this paper is similar to that proposed by Ramsey and Lampart [1998]. However, as the focus of this paper is on time series *forecasting*, several unique problems arise such as the joint selection of $A(\cdot)$ and $B(\cdot)$ (Section 3.1) and input reduction (Section 3.2).

2. The Wavelet Packet Transform.

The WPT is implemented by successively filtering an input, $y(k)$ with specifically designed high pass, H, and low pass, G, filters forming a WPT tree (Figure 1). This is followed by a down-sampling by two*. As H and G form *perfect reconstruction filters*, the original data can be reconstructed from the down-sampled coefficients. With successive filtering, the level of frequency resolution increases at the expense of time resolution. As the option exists to filter each branch independently an adjustable time-frequency resolution trade-off is possible (three alternative trees or *packets* are shown in Figure 1) [for an excellent textbook on wavelets see Percival and Walden (1999)].

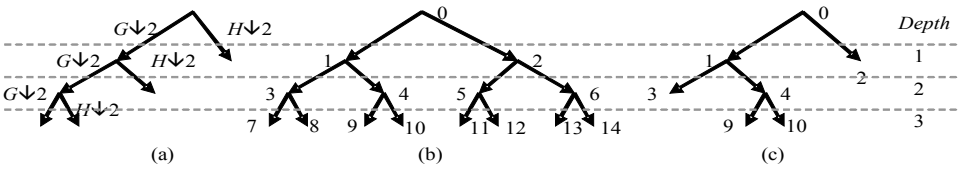


Figure 1. Diagram of the WPT to a depth of three. (a) packet {7,8,4,2} (the wavelet transform), (b) the complete wavelet packet tree and labeled nodes (c) example of another wavelet packet {3,9,10,2}.

3. The Wavelet Transfer Model.

The wavelet transfer model first pre-filters the input and output using the wavelet packet transform. Input selection is then applied and a non-linear model is used to relate the transformed input to the output as:

$$AY(k) = f(S^{\circ}B^{\circ}U(k)) + \varepsilon'(k) \tag{3}$$

* i.e. removing every second element of the filtered signal, denoted ↓2.

where A is a (WPT) basis transform of the output, $Y(k)=[y(k)\ y(k\text{-}1)\ \ldots\ y(k\text{-}s)]$, B is a (WPT) basis transform of the input, $U(k)$, S represents the *shrinkage* operator which reduces the dimensionality of the input (see Section 3.2), f is a non-linear function, $\varepsilon'(k)$ is a vector of (filtered) error terms[†], s is the period of the data and ^{o} denotes *after*.

3.1 *Packet selection technique.*

Define:

$$\mathbf{D}_1 = \{A_i\}_{i=1}^{N_1} \qquad \text{and} \qquad \mathbf{D}_2 = \{B_j\}_{j=1}^{N_2} \tag{4}$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are wavelet packet *dictionaries* of all possible WPT transforms of $Y(k)$ and $U(k)$, respectively. A_i and B_j are the elements of those dictionaries and N_1 and N_2 their respective lengths. The aim of packet selection is to choose an element of $\mathbf{D}_1$ and $\mathbf{D}_2$ *jointly*. It is proposed here to use the *Mutual Information* (MI, defined below) between the transformed input and output to determine the optimal transform:

$$(A, B) = \arg\max_{i,j} I(A_i Y(k); B_j U(k)) \tag{5}$$

where A and B are the bases to be chosen and $I(U;Y)$ is the MI defined as:

$$I(U;Y) = \int_Y \int_U f_{U,Y}(u,y) \log \frac{f_{U,Y}(u,y)}{f_U(u) f_Y(y)} du dy \tag{6}$$

where $f_U(u)$ and $f_Y(y)$ are the (multi-variate) probability distributions of U and Y respectively. $f_{U,Y}(u,y)$ is the joint PDF between U and Y. Saito *et al*. [2002] proposed a Local Discriminant Basis (LDB) algorithm for calculating the MI for a *classification* problem. However, estimating $f_{U,Y}(u,y)$ for multi-variate continuous data is a difficult task [Darbellay (1999)]. An approximation of the MI may be made by means of multi-variate Gaussian kernels as [Nilsson *et al*. (2002)]:

$$I(U;Y) \approx \sum_{j=1}^{M} \int_Y \int_U \alpha_j G_{j_{U,Y}}(u,y) \log \frac{G_{j_{U,Y}}(u,y)}{G_{j_U}(u) G_{j_Y}(y)} du dy \quad , \quad \sum_{j=1}^{M} \alpha_j = 1 \tag{7}$$

where $G_{j_{U,Y}}(u,y)$, $G_{j_U}(u)$, $G_{j_Y}(y)$ are multi-variate Gaussian distributions for the j^{th} kernel, M denotes the number of modes in the approximated distributions and α_j is the j^{th} weight associated with each kernel to ensure that the total probability equals one. The optimum mean and covariance matrices for the kernels may be estimated using the Expectation Maximization (EM) algorithm [Dempster *et. al*. (1977)]. Given a Gaussian kernel the expression for the approximate MI then reduces to:

$$I(U;Y) \approx \sum_{j=1}^{M} \alpha_j \left| \hat{C}_{j_{UY}} \right| \Big/ \left| \hat{C}_{j_U} \right\| \hat{C}_{j_Y} \Big| \tag{8}$$

[†] Note that $f(\cdot)$ makes a *forecast* of the transformed output, $\hat{Y}'(k)$, and not of $Y(k)$. Typically $f(\cdot)$ will be trained to minimize some cost function (e.g. the Mean Squared Error, MSE) of the forecast errors. However, in this case $f(\cdot)$ minimizes the cost function with respect to $\varepsilon'(k)$ and not $\varepsilon(k)$. This is sometimes advantageous [Ljung, (1991)] as it may remove disturbances at high and low frequencies that are not wanted during modeling.

where $|\cdot|$ denotes the determinant, $\hat{C}_{jUY}$, $\hat{C}_{jY}$ and $\hat{C}_{jU}$ are the sample cross and auto-covariance matrices of the j^{th} kernel.

3.2 *Input selection.*

Input selection requires reduction in the dimension of $BU(k)$. Typically, a threshold is used in which wavelet coefficients with mutual information (or entropy in the univariate case) below the threshold are eliminated [Percival and Walden (2000)]. However, the purpose here is to reduce the dimension of the input space to a specific size. Given A and B (calculated in Section 3.1), input selection is implemented by retaining those variables that *individually* have the highest mutual information with the output as:

$$U'' = \{Au_{j_m} : m = 1,\ldots,N_{\dim} / \underset{j_m \in \{i / j_{m-1},\ldots,j_1\}}{\arg\max} \hat{I}(Au_{j_m}; BY) \tag{9}$$

where U'' is the reduced input set of dimension N_{dim}, Au_l is the l^{th} element of AU and j_m are the indices of the retained elements. $\hat{I}(Au_l; BY)$ is estimated as in Eq. (8).

4. Example Application: Hourly Electricity Demand Forecasting.

Hourly electrical demand is a time-series driven by human activity which is influenced by weather; temperature and humidity being the dominant causal variables. The data spans the years 1986-2000, only Mondays to Fridays and only the months January to March. In addition, this data has been *de-trended*. The data has been split into three different groups for analysis; training set (400×24 points), validation set (170×24 points) and test set (170×24 points). Finally, note that this data is periodic with a period of 24 (hours) and that full details of the above can be found in [Fay *et al*., (2003)]. In Figure 2 a rise in temperature from indices 87:100 and a corresponding fall in the detrended demand at indices 95:100, are indicated. This example suggests that a low frequency component in the temperature (i.e. the average temperature between indices 87:100) is causing a corresponding change in the dependant variable but at a later time and for a shorter period. Thus, the wavelet transfer model would seem ideal in identifying these time-frequency correlations between the input and output.

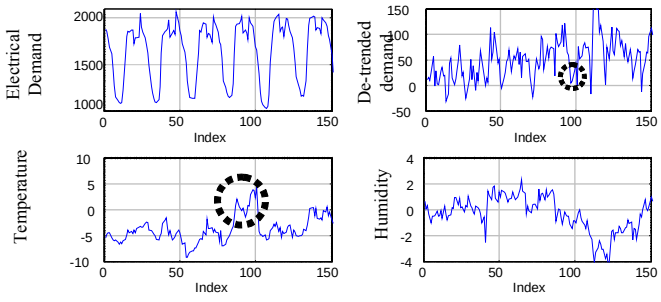


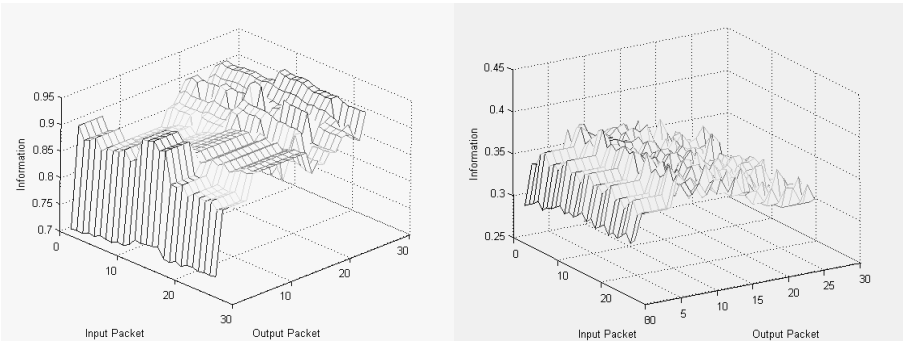
Figure 2. Graph of original and de-trended electrical demand, temperature and humidity.

For the purposes of this paper the output time-series is the *de-trended* demand, $y(k)$, and there are ***two*** inputs, temperature and humidity, denoted $u^t(k)$ and $u^h(k)$ respectively. The WPT to a depth of four is taken of $y(k)$, $u^t(k)$ and $u^h(k)$ using Daubechie's '*D4'* wavelet

[Percival and Walden (2000)], giving three dictionaries $\boldsymbol{D}_1$, $\boldsymbol{D}_2^1$ and $\boldsymbol{D}_2^2$. $Y(k)$, $U^l(k)$ and $U^h(k)$ are constructed as:

$$Y(k) = [y(k) \ldots y(k\text{-}24)] \quad U^l(k) = [u^l(k) \ldots u^l(k\text{-}72)] \quad U^h(k) = [u^h(k) \ldots u^h(k\text{-}72)] \quad k=24,48,\ldots \qquad (10)$$

Note that $U^l(k)$ and $U^h(k)$ contain weather data up to a lag of 3 days (72 hours). After three days it is considered that the weather has no effect on the demand [Fay *et al*., 2003]. In addition, note that as the data is periodic, it is sufficient to take every 24th value of k [§]. The input selection reduces the number of input variables to seven[**]. Figure 3 shows the mutual information between the inputs and outputs for different packet transforms, calculated using M =1 (This is equivalent to using the correlation).



(a) (b)

Figure 3. Graph of mutual information between de-trended input and (a) temperature (b) humidity.

Table 1, below, summarizes the optimal packets chosen for the input-outputs in Figure 3. As can be seen, the transformed temperature has higher mutual information with the transformed de-trended load and so this is chosen as the transform to be applied.

Table 1. Packet transforms that share the maximum mutual information with de-trended load.

Variable	Input Packet number	Input packet nodes	Output packet number	Output packet nodes.	Mutual Information
Temperature	26	{3,9,10,2}	12	{7,8,9,10,11,12,6}	0.9455
Humidity	7	{3,4,5,13,14}	9	{3,4,2}	0.3901

The next stage is to model $AY(k)$ with $BU(k)$ using a feed-forward neural network. The network used is similar to that described in [Fay *et al*. (2003)] (the inputs differ) and so it is not described here. For comparison the Wavelet Transfer Model (WTM) is compared to a Transfer Model (TM) in which the WPT is not applied, i.e. A=1, B=1 (note: input selection is still applied). Table 2 summarizes the results.

Table 2. A comparison of the wavelet transfer model and a model without the WPT.

Model	MSE Training Set	MSE Validation Set	MSE Novelty Set
WTM	3586	5236	6815
TM	3929	5876	7569

[§] i.e. the data is arranged by day, see Eq. (10).
[**] This number is chosen subjectively with experience.

5. Conclusions.

The WPT-based model has been shown to have merit for the task of electrical demand forecasting. Some minor drawbacks include the restrictive assumption that the input and output are drawn from a multi-variate Gaussian distribution, which may not be a good approximation of the actual distribution. In addition as mutual information is not additive and choosing the optimal packet bases can be computationally expensive.

6. Acknowledgments

The authors would like to thank the Irish national grid operator (EirGrid Plc).

7. References

Darbellay, G.A., (1999). An estimator of the mutual information based on a criterion for independence, *Computational Statistics and Data Analysis*, **32**: 1-17

Dempster, A.P., Laird, N.M., Rubin, D.B., (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B,* **39 (1)**: 1-38

Fay, D., Ringwood, J.V., Condon, M., Kelly, M., (2003). 24-hour electrical load data—a sequential or partitioned time series? *Neurocomputing*, **55**: 469 – 498

Hiden, H.G., Willis, M.J., Tham, M.T., Montague, G.A., (1999). Non-linear principal components analysis using genetic programming, *Computers and Chemical Engineering*, **23**: 413-425

Labat, D., Ababou, R., Mangin, A., (2000). Rainfall-runoff relations for karstic springs. Part II: continuous wavelet and discrete orthogonal multiresolution analyses*, Journal of Hydrology*, **238**: 149-178

Liu, L.T., Hsu, H.T., Grafarend, E.W, (2005). Wavelet coherence analysis of length-of-day variations and El Niño-southern oscillation, *Journal of Geodynamics*, **39**: 267–275

Ljung, L. (1999). *System Identification: Theory for the User*, (2nd ed), Prentice Hall, N.J.

Mallat, S.G., (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**: 674-693

Milidiú, R.L., Machado, R.J., Rentería, R.P., (1999). Time series forecasting through wavelets transformation and a mixture of expert models, *Neurocomputing*, **28**: 145-156

Nason, G.P., Sapatinas, T., (2001). Wavelet Packet Transfer Function Modelling of Nonstationary Time Series, *Statistics and Computing*, **12**: 45-56

Nilsson, M., Gustafsson, H., Andersen, S.V., Kleijn, W.B., (2002). Gaussian mixture model based mutual information estimation between frequency bands in speech**,** *IEEE International Conference on Acoustics, Speech, and Signal Processing*, **1**:525-528

Percival, D.B., Walden, A.T., (2000). *Wavelet Methods for Time Series Analysis*, Cambridge Univ. Press, Cambridge

Ramsey, J.B., Lampart, C. (1998). "The decomposition of economic relationships by time scale using wavelets: expenditure and income", *Studies in Nonlinear Dynamics and Econometrics*, **3**: 23–42

Roberts S., Roussos, E., Choudrey, R., (2004). Hierarchy, priors and wavelets: structure and signal modelling using ICA, *Signal Processing*, **84**: 283 – 297

Saito, N., Coifman, R.R., (1997). Extraction of geological information from acoustic well-logging waveforms using time-frequency wavelets, *Geophysics*, **62(6)**: 1921-1930

Saito, N., Coifman, R.R., Geshwind, F.B., Warner, F., (2002). Discriminant feature extraction using empirical probability density estimation and a local basis library, *Pattern Recognition*, **35**: 2841–2852

Schoukens, J., Pintelon, R., (1991). *Identification of Linear Systems: a Practical Guideline to Accurate Modeling*, Pergamon Press, London

Yao, S.J., Song, Y.H., Zhang, L.Z., Cheng, X.Y. (2000). Wavelet transform and neural networks for short-term electrical load, *Energy Conversion and Management*, **41**: 1975-1988