IDENTIFICATION AND ESTIMATION OF CONTINUOUS-TIME RAINFALL-FLOW MODELS

Peter Young * **Hugues Garnier** **

* *Centre for Research on Environmental Systems and Statistics, Lancaster University, UK; and Integrated Catchment Assessment and Management Centre, School of Resources, Environment and Society, Australian National University, Canberra,* p.young@lancaster.ac.uk
** *Centre de Recherche en Automatique de Nancy, CRAN UMR 7039 CNRS-UHP-INPL, Université Henri Poincaré, Nancy 1, BP 239, F-54506 Vandœuvre-les-Nancy Cedex, France,* hugues.garnier@cran.uhp-nancy.fr

Abstract: The identification and estimation of rainfall-flow models is one of the most challenging problems in hydrology. This paper presents the results of direct continuous-time identification and estimation of a rainfall-flow model for the Canning, an ephemeral river in Western Australia, based on daily sampled data. It compares simplified and full implementations of the optimal instrumental variable algorithm used in the application and discusses the advantages of this direct method when compared with the alternative indirect approach to the problem based on discrete-time model estimation.

Keywords: Continuous-time models, hybrid models, instrumental variable, sampled data, parameter estimation, river catchments, rainfall-flow

1. INTRODUCTION

This paper [1] deals with the modelling of rainfall-flow dynamics in river systems. One important application for such models is in flood forecasting and warning, where recent research has shown the value of discrete (*e.g.* Young, 2002) and continuous-time (Young, 2004) transfer function models. As these references show, identification and estimation based on inductive Data-Based Mechanistic (DBM) modelling (*e.g.* Young, 1998 and the prior references therein) and State-Dependent Parameter (SDP) estimation (*e.g.* Young *et al.*, 2001) suggests that the rainfall flow model is of the 'Hammerstein' nonlinear form, with an input nonlinearity that converts the measured rainfall into 'effective rainfall': *i.e.* the rainfall that is effective in causing variations in flow. This effective rainfall then passes through a linear transfer function to yield the river flow. This paper presents the results of direct continuous-time (CT) identification and estimation of a rainfall-flow model for the Canning, an ephemeral river in Western Australia, based on daily sampled data. These results are obtained using the optimal **R**efined **I**nstrumental **V**ariable method of identification for **C**ontinuous-time systems (RIVC) (Young *et al.*, 2006) and they are compared with alternative 'indirect' approaches to the problem, where the CT model parameters are inferred from the parameters of discrete-time (DT) models. Such DT models are normally identified using conventional stochastic DT estimation

[1] This report is related to the paper presented at the 14th IFAC Symposium on System Identification (SYSID'2006), Newcastle (Australia), pp. 1276-1281, March 2006.

methods available for use in Matlab, such as the **R**efined **I**nstrumental **V**ariable (RIV) approach in the CAPTAIN Toolbox (Young, 2006*a*) and the **P**rediction **E**rror **M**ethod (PEM) in the Matlab System IDentification (SID) Toolbox.

2. OPTIMAL IV METHODS FOR CT MODELS

In recent years, there has been renewed interest in the problem of identifying continuous-time systems on the basis of discrete-time sampled data (see *e.g.*; Johansson, *et al.*, 1999; Bastogne *et al.*, 2001; Wang and Gawthrop, 2001; Garnier *et al.*, 2003; Garnier and Young, 2004; Moussaoui *et al.*, 2005).

In the RIVC algorithm (Young *et al.*, 2006) the relationship between the measured input and output is a continuous-time transfer function, while the noise is represented as a discrete-time AR or ARMA process. This RIVC algorithm is a direct development of the **S**implified **R**efined **I**nstrumental **V**ariable (SRIVC) method for continuous-time systems (Young and Jakeman, 1980) that has been used successfully for many years and has demonstrated the advantages of the stochastic formulation of the CT estimation problem over earlier deterministic methods.

The 'simplification' that characterises the name of the SRIVC method is the assumption, for the purposes of algorithmic development, that the additive noise is purely white in form. Although the inherent instrumental variable aspects of the resulting algorithm ensure that the parameter estimates are consistent and asymptotically unbiased in statistical terms, even if the noise happens to be coloured, the estimates are not statistically efficient (minimum variance) in this situation. This is because the special prefilters required in the estimation are not designed to account for the colour in the noise process. The hybrid RIVC estimation procedure follows logically from the RIV estimation method for DT models (Young, 1976; Young and Jakeman, 1979) by including concurrent DT noise model estimation and the use of this estimated noise model in the implementation of the prefilters.

The hybrid form of the Box-Jenkins transfer function (BJTF) model is considered for two reasons. First, the theoretical and practical problems associated with the estimation of purely stochastic continuous-time ARMA models are avoided by formulating the problem in this manner. Second, as pointed out above, one of the main functions of the noise estimation is to improve the statistical efficiency of the parameter estimation by introducing appropriately defined prefilters into the estimation procedure. This can be achieved adequately on the basis of prefilters defined by reference to discrete-time AR or ARMA noise models.

This more sophisticated and statistically motivated RIVC method of CT identification and estimation is described and evaluated in Young *et al.* (2006) in order to demonstrate the advantages of the stochastic model formulation. This evaluation is based on comprehensive Monte Carlo Simulation (MCS) analysis. In the present paper, the practical advantages of this approach are demonstrated by the real example considered next in Section 3.

3. RAINFALL-FLOW MODELLING

The set of daily effective rainfall-flow data shown in Figure 1 is from the Canning, an ephemeral river in Western Australia (*i.e.* the river stops flowing during Summer). Note that the effective rainfall is considered here so that the dynamics are linear for the purposes of the present analysis. Full nonlinear modelling of another river system is considered in Young (2006*b*).

3.1 SRIVC Identification and Estimation

Previous discrete-time modelling of the data in Figure 1 has utilized the SRIV option of the RIV algorithm in the CAPTAIN Toolbox (*e.g.* Young *et al.*, 1997). Here, however, let us consider the application of the continuous-time SRIVC and RIVC algorithms. The bottom panel of Figure 1 compares the sampled output $\hat{y}_k$ of the SRIVC estimated continuous-time model (full line) with the measured flow. This model is identified in the second order, linear form:

$$\begin{aligned} \hat{y}(t) &= \frac{b_0 s^2 + b_1 s + b_2}{s^2 + a_1 s + a_2} u(t) \\ y_k &= \hat{y}_k + \xi_k \end{aligned} \tag{1}$$

where the subscript k in the second, observation equation denotes the sampled value of the associated variable on the k^{th} day, *i.e.* $y_k = y(k\Delta t)$, where Δt is the daily sampling interval; $\hat{y}_k$ is the sampled, deterministic output of the model; and ξ_k is the discrete-time noise process associated with y_k. This noise is identified by the AIC criterion (Akaike, 1974) as an AR(25) process and by the Schwarz (1978) criterion as ARMA(5,2).

The estimated parameters of the main TF model are as follows:

$$\begin{aligned} &\hat{a}_1 = 0.4428(0.0232); \quad \hat{a}_2 = 0.0208(0.0027); \\ &\hat{b}_0 = 0.0160(0.0012); \quad \hat{b}_1 = 0.0689(0.0018); \\ &\hat{b}_2 = 0.0057(0.0007) \end{aligned} \tag{2}$$

where the figures in parentheses are the estimated standard errors on the associated estimates. This

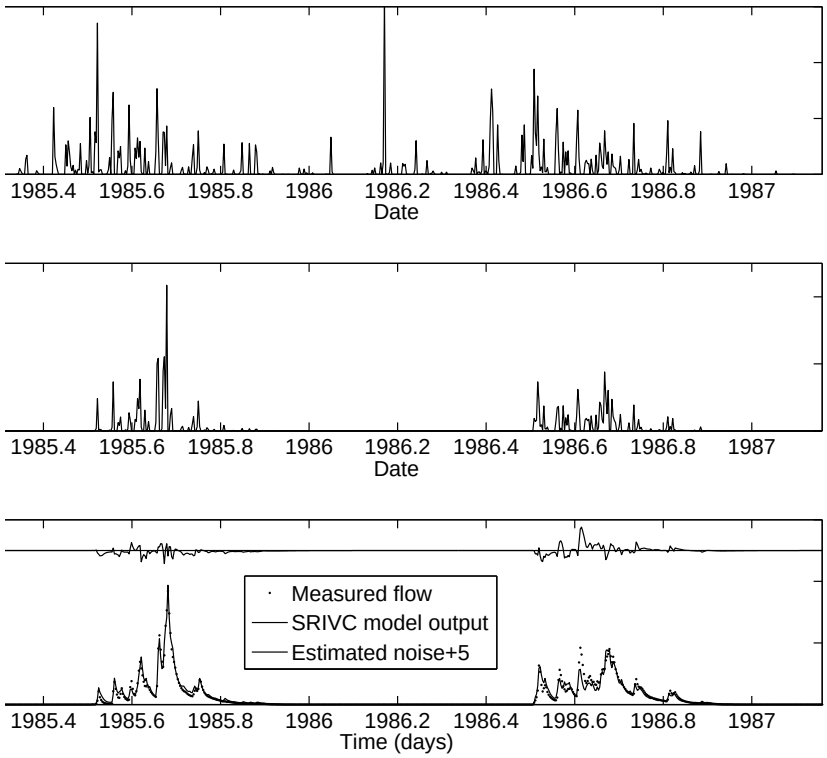

Fig. 1. Measured rainfall (top panel), effective rainfall (middle panel) and flow (lower panel) data from the River Canning, Western Australia. The lower panel also shows the output $\hat{y}(t)$ of the SRIVC estimated continuous-time model (full line) and the associated noise (error, $y_k - \hat{y}_k$) series (+5).

model has coefficients of determination, based on its simulated deterministic output $\hat{y}(t)$, of $R_T^2 = 0.958$; (*i.e.* 95.8% of the measured flow variance is explained by $\hat{y}(t)$); and a standard coefficient of determination based on the one-step-ahead prediction errors of $R^2 = 0.983$.

Frequency response plots of the estimated AR(25) and ARMA(5,2) noise models are shown in Figure 2. Here, the ARMA(5,2) model was estimated in two ways, as indicated on the plot: namely, using the PEM algorithm in the Matlab SID Toolbox (dashed line); and using a recent modification of the method described in Young (1985), shown as a full line. This latter algorithm is based on high order AR modelling coupled with subsequent ARMA estimation based on a combination of IV and least squares estimation (ARMA-IV) (Young, 2006*c*). The residual variances of the two ARMA models are 0.00353 for PEM and somewhat less at 0.00339 for ARMA-IV, while the residual variance of the AR(25) model is 0.00292. Note that the main spectral characteristics of these various noise models are similar and all appear at higher frequencies, well outside the bandwidth of the estimated model. The autocorrelation function of both the AR(25) and ARMA(5,2) noise model residuals shows that they are not significantly autocorrelated but their variance changes rather radically over time (*i.e.* they are 'heteroscedastic': see Young, 2002), with no noise at all over the Summer periods when there is no flow, and high variance when the largest flows occur over the Winter periods. Of course, the fact that the noise has this non-standard property should alert us to possible problems and question the application of AR/ARMA modelling: this is considered later.

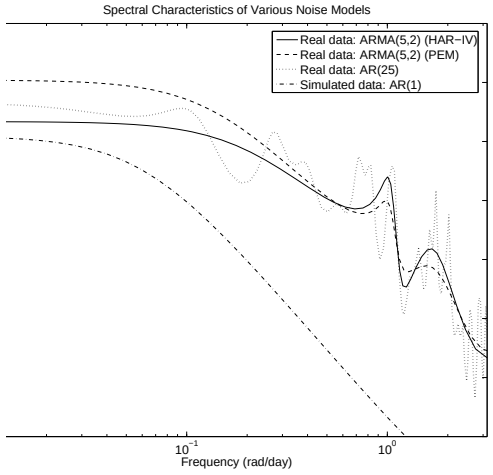

Fig. 2. Spectral characteristics of various noise models (see text). The lower dashdot line is discussed later in Section 3.3.

3.2 Derived Parameter and Indirect Estimation

In conformity with the Data-Based Mechanistic (DBM) modelling philosophy (see *e.g.* Young

(1998) and the prior references therein), it is important to interpret the model (1) with the parameter estimates of (2) in a physically meaningful manner. Following previous research on rainfall-flow modelling (*e.g.* Young, 2002, 2004) this is possible if the TF model is decomposed into a parallel connection of two, first order transfer functions that can be associated with the surface and groundwater characteristics of the river catchment. In the present case, the respective time constants (residence times) $\hat{T}_1$ and $\hat{T}_2$ of the two transfer functions are defined as follows:

$$\begin{aligned} &\text{Surface processes} : \hat{T}_1 = 2.57 \; days \\ &\text{Groundwater processes} : \hat{T}_2 = 18.71 \; days \end{aligned} \tag{3}$$

In other words, and not surprisingly, the rainfall affects the flow via the surface processes much more quickly than it affects the flow through the groundwater system.

The SRIVC results in (2) can be compared with the results obtained by indirect identification and estimation. For example, using the SRIV option of the RIV algorithm in CAPTAIN and converting the estimated [2 3 0] model to continuous-time, using the `d2cm` algorithm in Matlab ('zoh' option), yields the following estimates for the continuous time model parameters:

$$\begin{aligned} \hat{a}_1 = 0.4425; \quad \hat{a}_2 = 0.0209; \\ \hat{b}_0 = 0.0159; \quad \hat{b}_1 = 0.0688; \quad \hat{b}_2 = 0.0057 \end{aligned} \tag{4}$$

The time constants associated with this model are 2.57 and 18.6 days and it has an $R_T^2 = 0.958$.

It is interesting to compare these results with those produced by the PEM, OE, BJ and IV4 algorithms in the SID Toolbox. The PEM algorithm produces very poor results if the noise is modelled as an ARMA(5,2) or AR(25) process with a coefficient of determination of $R_T^2 = 0.917$ and of $R_T^2 = 0.903$ respectively. However, it yields a model that is closer to the SRIV estimated model if the noise model is not included (simple TF model, effectively OE estimation), with a marginally smaller coefficient of determination $R_T^2 = 0.953$ and estimated time constants of 3.13 and 14.77 days [2]. Exactly the same results are obtained from the OE and BJ algorithms, as would be expected. However, IV4 produces very poor results with a negative real pole in the discrete-time model and an $R_T^2 = 0.933$. Interestingly, if the SRIV estimates are used as starting values for PEM, the coefficients of determination is increased to almost the same as the SRIV estimated model, at $R_T^2 = 0.957$, and the estimated time constants are then 2.49 and 15.7 days. Of course, PEM estimation is clearly redundant in this situation if it needs to be started by the SRIV estimates in order to ensure convergence to a similarly acceptable model. Note that for the results shown in (4), no standard errors are given because these are not available after the `d2cm` transformation and would need to be evaluated in some manner: *e.g.* by Monte Carlo Simulation (MCS) analysis based on the estimated parametric error covariance matrix provided by the discrete-time estimation algorithms.

The uncertainty on the estimates of any parameters derived from the directly estimated CT model, such as the time constants $\hat{T}_1$ and $\hat{T}_2$, is also not available and so needs to be estimated by MCS analysis. For example, Figure 3 presents the results of such MCS analysis using a simulation model based on the SRIVC estimated parameters and a normally distributed white noise chosen to provide a signal/noise ratio similar to that on the real data. These results are based on 5000 realizations and are presented in the form of the normalised histograms (empirical distributions) of the two time constants. The full line in these plots shows the best fitting normal distribution curve, based on the first two moments (mean and variance) of the realisations in each case. The values of these statistics are given above each subplot. Not surprisingly, given that the information content of the data in this regard is naturally less than that for the short time constant, the long time constant estimate is not so well defined and is somewhat skewed towards higher values. Note that the PEM-based indirect estimates lie at the very edge of these distributions.

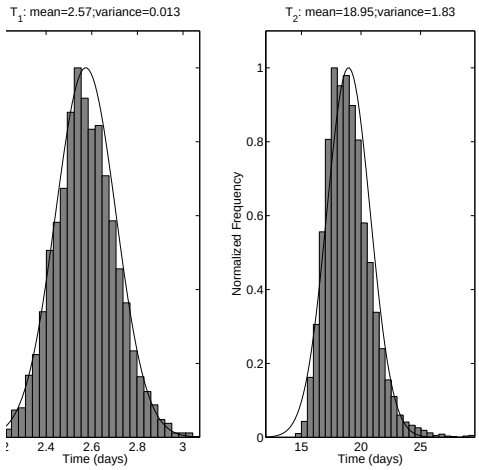

Fig. 3. Normalized histograms (empirical distributions) of the two time constants based on MCS analysis: T_1 in the left panel and T_2 in the right panel.

3.3 Full RIVC Estimation

As far as full RIVC estimation is concerned, there is one problem with the present example: namely,

[2] All ARMAX estimation was poor with maximum $R_T^2 = 0.929$

the nature of the naturally occurring noise. As pointed out previously, this noise is heteroscedastic (see bottom panel, Figure 1). This means that some of the assumptions about the noise process required for the full RIVC estimation are violated because the AR(25) and ARMA(5,2) noise models are rather poor representations of the real noise. As a result, application of the full RIVC model yields reasonable estimates of the parameter values because of the instrumental variable mechanism, which is much less vulnerable to the nature of the noise. On the other hand, the results are very similar to those obtained with the simpler SRIVC algorithm. The main reason for this is that the AR(25) and ARMA(5,2) models are dominated by higher frequency characteristics (see Figure 2), with no significant roots close to the unit circle in the complex z domain. As a result, there is only a small improvement in statistical efficiency resulting from the introduction of the noise model and related enhancement of the adaptive prefilters. As can often be the case, however, the SRIVC parameter and covariance matrix estimates are quite good and would be satisfactory for most practical applications. Indeed, the results obtained here and in other studies we have carried out, suggest that SRIVC provides an excellent default algorithm for day-to-day applications of this kind. Of course, as illustrated in other simulation studies reported by Young *et al.* (2006), the advantages of full RIVC estimation can be significant in the case of more severely coloured noise, particularly when the AR model has roots near to the unit circle. For example, the results shown in Table 1 were obtained by MCS analysis using a simulation model based on the SRIVC estimated parameters but with the noise process simulated as a first order AR process with denominator polynomial $C(z^{-1}) = 1 - 0.96z^{-1}$ and a normally distributed white noise input chosen to provide a signal/noise ratio similar to that on the real data. The spectral characteristics of this noise are shown by the dash-dot line in Figure 2 (the gain has been adjusted a little for greater clarity).

The MCS results in Table 1 now reveal a significant reduction in the standard errors on the RIVC parameter estimates when compared with the SRIVC estimates. Note, however, that because of the inherent instrumental variable and prefiltering aspects of the SRIVC algorithm, the parameter estimates are statistically consistent and have relatively low variance. This demonstrates, once again, the robust characteristics of this simpler algorithm.

4. CONCLUSIONS

This paper has described the application of optimal, continuous-time identification and estimation methods to the problem of modelling rainfall-flow dynamics in an ephemeral river system. With the help of numerical simulations, the paper has also demonstrated a number of interesting properties of the SRIVC and RIVC algorithms used in this application. In particular, the results reveal that, in this case, little advantage is obtained from the use of the full RIVC method in comparison with the simpler SRIVC method because of the particular 'heteroscedastic' and high frequency noise characteristics. However, subsequent MCS analysis, based on the stochastic simulation of the SRIVC estimated rainfall-flow model with additive low frequency AR noise replacing the natural noise, have showed the advantages of the new RIVC approach.

It is felt that the results presented in this paper, based on a stochastic formulation of the CT transfer function estimation problem, provide a theoretically elegant and practically useful approach to the modelling of stochastic linear systems from discrete-time sampled data. It is an approach that has a number of advantages in scientific terms, since it provides a differential equation model that conforms with models used in most scientific research, where conservation equations (mass, energy etc.) are normally formulated in terms of differential equations. It is also a model defined by a unique set of parameter values that are not dependent on the sampling interval, so eliminating the need for conversion from discrete to continuous-time that is an essential element of indirect approaches to estimation based on discrete-time model estimation.

The SRIVC and RIVC algorithms used in this paper are both available in the CONTSID (http://www.cran.uhp-nancy.fr/contsid/) and CAPTAIN (http://www.es.lancs.ac.uk/cres/captain/) Toolboxes for Matlab. CONTSID concentrates on all of the major identification tools for continuous-time systems including SRIVC/RIVC; while CAPTAIN is a more general toolbox for stationary, nonstationary and nonlinear time series analysis, having tools for discrete (SRIV/RIV) and continuous-time (SRIVC/RIVC) TF identification, as well as algorithms for time variable and state dependent (nonlinear model) parameter estimation, nonstationary signal extraction and adaptive forecasting.

REFERENCES

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Bastogne, T., H. Garnier and P. Sibille (2001). A PMF-based subspace method for continuous-

Parameters	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$	$\hat{a}_1$	$\hat{a}_2$	$\hat{c}_1$
True values	**0.0160**	**0.0689**	**0.0057**	**0.4428**	**0.0208**	**−0.96**
SRIVC (SR)	0.0162	0.0696	0.0069	0.4719	0.0245	/
stand. error	0.0008	0.0012	0.0005	0.0162	0.0018	/
SRIVC (MCS)	0.0160	0.0689	0.0058	0.4451	0.0217	/
stand. error	0.0008	0.0023	0.0020	0.0450	0.0076	/
RIVC (SR)	0.0164	0.0695	0.0067	0.4651	0.0237	-0.948
stand. error	0.0002	0.0006	0.0009	0.0184	0.0038	0.010
RIVC (MCS)	0.0160	0.0689	0.0058	0.4450	0.0212	-0.955
stand. error	0.0002	0.0007	0.0011	0.0206	0.0045	0.011

Table 1. SRIVC and full RIVC parameter estimates in the case of the simulated effective rainfall-flow system. SR denotes single run results and MCS denotes results from MCS analysis based on 100 random realizations.

time model identification. Application to a multivariable winding process. *International Journal of Control* **74**(2), 118–132.

Garnier, H. and P.C. Young (2004). Time-domain approaches for continuous-time model identification from sampled data. In: *Invited tutorial paper for the American Control Conference (ACC'2004)*. Boston, MA (USA).

Garnier, H., M. Mensler and A. Richard (2003). Continuous-time model identification from sampled data. Implementation issues and performance evaluation. *International Journal of Control* **76**(13), 1337–1357.

Johansson, R., M. Verhaegen and C.T. Chou (1999). Stochastic theory of continuous-time state-space identification. *IEEE Transactions on Signal Processing* **47**(1), 41–50.

Moussaoui, S., D. Brie and A. Richard (2005). Regularization aspects in continuous-time model identification. *Automatica* **41**(2), 197–208.

Wang, L. and P. Gawthrop (2001). On the estimation of continuous-time transfer functions. *International Journal of Control* **74**(9), 889–904.

Young, P. C., A. J. Jakeman and D. Post (1997). Recent advances in the data-based modelling and analysis of hydrological systems. *Water Science and Technology* **36**, 99–116.

Young, P.C. (1976). Some observations on instrumental variable methods of time-series analysis. *International Journal of Control* **23**, 593–612.

Young, P.C. (1985). *Handbook in Statistics: Time Series in the Time Domain*. Chap. Recursive identification, estimation and control, pp. 213–255. E.J. Hannan and P.R. Krishnaiah and M.M. Rao ed.. North Holland.

Young, P.C. (1998). Data-based mechanistic modeling of environmental, ecological, economic and engineering systems. *Journal of Modelling & Software* **13**, 105–122.

Young, P.C. (2002). Advances in real-time flood forecasting". *Philosophical Trans. Royal Society, Physical and Engineering Sciences* **360**(9), 1433–1450.

Young, P.C. (2004). Identification and estimation of continuous-time hydrological models from discrete-time data. In: *International Conference on Hydrology: Science and Practice for the 21st century*. London (U.K.).

Young, P.C. (2006*a*). The Captain toolbox for Matlab. In: *Proceedings of the 14th IFAC Symposium on System Identification (SYSID'2006)*. Newcastle (Australia).

Young, P.C. (2006*b*). Data-based mechanistic modelling and river flow forcasting. In: *Proceedings of the 14th IFAC Symposium on System Identification (SYSID'2006)*. Newcastle (Australia).

Young, P.C. (2006*c*). An instrumental variable approach to ARMA model identification and estimation. In: *Proceedings of the 14th IFAC Symposium on System Identification (SYSID'2006)*. Newcastle (Australia).

Young, P.C. and A.J. Jakeman (1979). Refined instrumental variable methods of time-series analysis: Part I, SISO systems. *International Journal of Control* **29**, 1–30.

Young, P.C. and A.J. Jakeman (1980). Refined instrumental variable methods of time-series analysis: Part III, extensions. *International Journal of Control* **31**, 741–764.

Young, P.C., H. Garnier and M. Gilson (2006). An optimal instrumental variable approach for identifying hybrid continuous-time Box-Jenkins model. In: *Proceedings of the 14th IFAC Symposium on System Identification (SYSID'2006)*. Newcastle (Australia).

Young, P.C., P. McKenna and J. Bruun (2001). Identification of nonlinear stochastic systems by state dependent parameter estimation. *International Journal of Control* **74**, 1837–1857.