Identification of Symmetric Noncausal Processes ⋆

Qiugang Lu [a] Philip D. Loewen [c] R. Bhushan Gopaluni [a] Michael G. Forbes [b]
Johan U. Backström [b] Guy A. Dumont [d] Michael S. Davies [d]

[a] *Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4*

[b] *Honeywell Process Solutions, North Vancouver, BC, Canada, V7J 3S4*

[c] *Department of Mathematics, University of British Columbia, Vancouver, BC, Canada, V6T 1Z3*

[d] *Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada, V6T 1Z3*

Abstract

We propose a maximum likelihood estimation approach for the identification of symmetric noncausal models. Such models are used to represent the cross-directional dynamic response of many industrial processes that are generally modeled with a high-dimensional gain matrix. Reducing the number of parameters in a noncausal model can significantly reduce the uncertainty associated with parameter estimates. We adapt the maximum likelihood method to treat symmetric noncausal models by showing that every symmetric noncausal process admits a spectrally equivalent causal model. It is then proved that the maximum likelihood estimate of this causal model converges to that of the original noncausal model. We present an iterative identification algorithm to efficiently estimate the parameters in noncausal models. Finally, we show that the parameter covariance estimate obtained from the causal model also converges to that of the noncausal model, which lays a foundation for optimal input design in noncausal processes. Several numerical examples illustrate the effectiveness of the proposed algorithm.

Key words: Noncausal model, Maximum likelihood estimation, Optimal input design, Paper machine

1 Introduction

Over the last few decades system identification of causal models has received extensive attention and a number of classical methods such as the prediction-error method, maximum likelihood estimation and subspace identification have been available in the literature [1,2,3]. Identification of noncausal models has not attracted the same amount of research focus in the system identification community, possibly due to the rarity of noncausal processes in the process industry. However, in applications where the independent variable indexes space rather than time, noncausal behavior is both physically realizable and relevant. An example is the cross-directional (CD) process of a paper machine, where a bumped actuator generates responses on both sides (cf. Fig. 1). If we treat this cross direction as an axis that is analogous to the conventional time axis, then the actuator response on both sides would correspond to 'past' and 'future'—essentially a noncausal behavior. Performing noncausal identification, preferably by adapting currently accessible techniques for causal models, forms the motivation of this work.

In the areas of image processing and astronomical data processing there have been efforts devoted to noncausal identification of autoregressive models [4,5,6]. A primary issue is the identifiability of noncausal models when the white noise is Gaussian. Specifically, given a noncausal AR process [1] $\{X_x\}$ of order p driven by independent and identically distributed white noise w_x, it is possible to find a purely causal (or a purely noncausal) AR model of order p that fits the spectrum of the original process $\{X_x\}$ and is driven by some other white noise $\hat{w}_x$ [7]. Since a Gaussian process is completely characterized by its second order properties, a noncausal system driven by Gaussian disturbances is not identifiable [4]. If a transfer function, $G(q)$, is purely causal

⋆ Corresponding author Bhushan Gopaluni. Tel. +01-604 827 5668.

Email addresses: qglu@chbe.ubc.ca (Qiugang Lu), loew@math.ubc.ca (Philip D. Loewen), bhushan.gopaluni@ubc.ca (R. Bhushan Gopaluni), michael.forbes@honeywell.com (Michael G. Forbes), johan.backstrom@honeywell.com (Johan U. Backström), guyd@ece.ubc.ca (Guy A. Dumont), miked@ece.ubc.ca (Michael S. Davies).

[1] We will use the variable 'x' (not 't') in this paper to highlight the fact that the processes we consider are noncausal in space (not time).

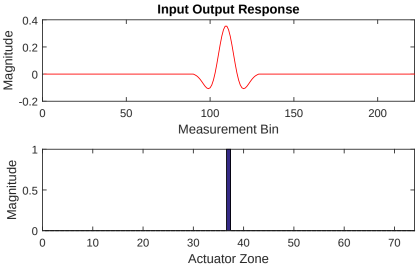


Fig. 1. Typical noncausal CD response in a paper machine to a step change in an actuator.

or purely anticausal, then $|G(e^{i\omega})|$ can be uniquely determined given the order of $G(q)$. On the other hand, if $G(q)$ is noncausal, then it is not possible to uniquely determine $G(q)$ from the second order properties of the process even if the correct order is known. This is due to the fact that

$$|e^{i\omega} - z_0| = |e^{-i\omega} - \bar{z}_0|. \tag{1}$$

In other words, if z_0 is a pole/zero of $G(q)$, then $G(q)$ and a transfer function with the conjugate inverse of z_0 as its pole/zero will have the same spectrum. The phase information is lost due to the above equality. Traditionally, this ambiguity is overcome by restricting the model search to the set of causal models when the process is driven by Gaussian noise.

Due to the non-identifiability of noncausal processes driven by Gaussian noise, most research on identification of noncausal processes is focused on identification of models driven by non-Gaussian noise [8]. Symmetric noncausal impulse response identification of ARMA models was considered in [5]. In that paper, the idea was to find a spectrally equivalent causal model and extract the noncausal model based on assumptions on the set of possible models. A maximum-likelihood method for estimating a noncausal ARMA model and the Cramer-Rao lower bound are derived in [9]. There is also literature on identification of noncausal ARMA models using higher order spectral analysis [10] to deal with non-Gaussian noise. Recent advances on noncausal identification are more focused on non-Gaussian univariate and multivariate autoregressive processes [11,12]. In the realm of system identification, the noncausal model is more often utilized as a tool to address the nonlinearity in the feedback or identification of unstable systems in closed-loop [13,14,15]. Moreover, noncausal models have a close connection with causal unstable models, as pointed out in [16,17]. Therefore, the available techniques for identifying causal unstable models, e.g., noncausal filtering in [17] and modifying noise model with an all-pass filter to make the predictor stable in [16], are applicable in theory to identify noncausal models. In principle, the noncausal filtering technique in [17] developed for unstable causal models can be implemented to handle the identification of noncausal models. However, with this method the user has to provide the gradient for the solvers, which may require extensive efforts in computing the gradient for noncausal models. For the method in [16], when converting a noncausal model into a causal unstable model, a new sequence of driving noise is produced which usually has a much higher variance than the original noise. This can significantly reduce the signal-to-noise ratio (SNR), and thus one has to pay extra attention when implementing this approach. In the current work, we present a novel iterative identification scheme for unbiased estimation of symmetric noncausal ARMAX models subject to Gaussian noise, and this method can avoid the above issues. We handle the identifiability issue by searching for model parameters in a set of causal models that are spectrally equivalent to the original noncausal models.

As mentioned above, our work is motivated by the control of paper machines. We are interested in controlling paper quality variables such as moisture, basis weight and caliper across the width of the emerging paper sheet, known as the Cross Direction (CD). This requires for a good model of the response of the CD process to actuator moves. A typical noncausal CD response is shown in Fig. 1. Current industrial practice is to measure such responses empirically, using bump tests on the actuators. In this work, we develop a general framework for noncausal process identification with CD modeling of paper machines as the target application.

System identification for sheet and film processes has gained considerable interest in the literature. The state-of-the-art on this topic has been covered thoroughly in [18] and [19]. The mainstream work on identification treats the CD process as a large-scale multi-input-multi-output (MIMO) system by discretizing the process model along the cross direction [20,21]. This is equivalent to using an FIR model for the CD response. As a consequence, the estimated model parameters have large uncertainty due to the large number of free parameters [19]. However, to the best knowledge of the authors, reports on treating the CD process as a low-order parametric noncausal model still remain scarce. In this paper, our central goal is to establish a framework for identifying noncausal ARMAX models with currently available results on causal system identification. This framework can easily be extended to situations with more complex model structures, such as Box-Jenkins models. Specifically, in this work we use a low order noncausal model in the cross direction to minimize the variance of parameter estimates. The method proposed in this paper provides an asymptotically efficient estimate of the CD model using a maximum likelihood approach. In [20] also, a maximum likelihood approach was used, however, our work is different due to the utilization of a *parsimonious noncausal* model instead of an FIR type model. In [18], a method for designing optimal inputs (for the actuator profile) has been developed. The optimal input is designed by minimizing a measure of the confidence region around a steady-state process model, using least squares. However, for robust control, a good description of the uncertainty in the frequency domain is useful. There is extensive literature on input design for robust control [22]. Unfortunately, all of these methods deal only with causal models. The other objective of this paper is to present an optimal input design algorithm for noncausal models based on the covariance formulation from our noncausal identification method. We have presented some preliminary results in [23]. Based on our previous results, in this work, we propose a framework for identification of symmetric noncausal processes. It includes rigorous proofs for all theorems, convergence analysis of our iterative identification algorithm and an optimal input design algorithm for noncausal models. The focus of this paper is on open-loop noncausal identification, however, the proposed results can easily be extended to closed-loop identification.

This paper is organized as follows: In Section 2, we prepare the stage for a rigorous description of the problem by presenting the assumptions and definitions used in this work and deriving an important result showing the spectral equivalence of any symmetric noncausal model to a causal model. In Section 3, we prove that the maximum likelihood estimate of the noncausal process and its causal-equivalent model asymptotically converge to the same value. In Section 4, we propose a new identification algorithm for symmetric noncausal processes. In Section 5, we prove that the covariance estimates of the original noncausal process and the causal-equivalent model asymptotically converge to the same value. In that section, we also briefly outline a method for designing inputs. In Section 6, a few simulation examples are presented. The paper concludes in Section 7.

2 Preliminaries

2.1 Assumptions & Notation

Suppose that the process measurements y_x are related to inputs u_x and noise e_x according to the following "true system"

$$\mathcal{S}: \quad A(q,\theta_0)y_x = B(q,\theta_0)u_x + C(q,\theta_0)e_x, \tag{2}$$

where A, B and C are polynomials in q, the forward shift operator with respect to x. The analysis presented in this paper requires all the polynomial coefficients to be real.

2.1.1 Notation

We index discrete stochastic processes with the independent variable x. N denotes the size of a collected data set. We use $\{s_x\}$ to represent a sequence of values for a signal s, where s_x stands for the signal s evaluated at x in space. Usually the range of x will be obvious from the context. By default we assume $x \in \{1, \cdots, N\}$. $y_x \in \mathbb{R}, u_x \in \mathbb{R}, e_x \in \mathbb{R}$ represent output (CD profile), input (actuator profile) and Gaussian white noise respectively. Noise variance is represented by σ^2. Causal-equivalent signal of s_x is represented by $\widetilde{s}_x$. For brevity we represent the data set $\{y_x\}$ by $\mathbf{y}$ and $\{\widetilde{y}_x\}$ by $\widetilde{\mathbf{y}}$. θ_0 is the true parameter vector containing the coefficients of the polynomials in (2). We replace θ_0 with θ to emphasize that the parameter is unknown wherever necessary. θ^i represents the parameter estimate after iteration i. n represents the number of unknown scalar parameters in θ. E is reserved for expectation over the probability space $(\mathcal{X}, \mathcal{F}, P)$, where $\mathcal{X}$ is the event space, $\mathcal{F}$ is a σ-algebra on $\mathcal{X}$ and P is a complete measure defined on $\mathcal{F}$. Hence, any random variable, s_x, should be written as $s_x(\xi)$ for some $\xi \in \mathcal{X}$. However, for the sake of brevity we ignore the argument, ξ, whenever it is obvious. The spectrum of a signal s_x is denoted by $\Phi_s(\omega)$, where ω is the frequency. $\|.\|$ is used to denote the vector 2-norm in Euclidean space or the Frobenius norm for matrices.

2.1.2 Assumptions

(A1) The true parameter vector θ_0 lies in a compact and convex subset Ω of $\mathbb{R}^n$.

(A2) For any signal s_x, $E[s_x] = m_s(x)$, $|m_s(x)| \le C_s, \forall x$.

(A3) For any signal s_x, $E[s_x s_{x+\tau}] = R_s(x, x+\tau)$, $|R_s(x, x+\tau)| \le C_s, \forall\, x, \tau, \lim_{N\to\infty} \frac{1}{N}\sum_{x=1}^{N} R_s(x, x-\tau) = R_s(\tau), \quad \forall\ \tau$.

(A4) For each $\theta \in \Omega$, the model polynomials $A(q,\theta)$, $B(q,\theta)$, $C(q,\theta)$ have no zeros on the unit circle and all zeros are stable.

(A5) For each $\theta \in \Omega$, each polynomials $T = A, B, C$ admits a factorization of the form $T(q,\theta) = T^{+}(q,\theta)T^{-}(q,\theta)$, involving a strictly causal polynomial, $T^{-}(q,\theta) = \sum_{i=0}^{n_T} t_i q^{-i}$, and a strictly anticausal polynomial, with $T^{+}(q,\theta) = T^{-}(q^{-1},\theta)$ on account of symmetry. Here $t_i = a_i, b_i, c_i$ when $T = A, B, C$, respectively.

(A6) $a_0 = c_0 = 1$.

(A7) e_x, $\widetilde{e}_x$ and their respective derivatives with respect to the parameter vector θ can be represented using families of uniformly stable filters (see definition below) acting on the known data $\{y_x\}$ and $\{u_x\}$.

Note that assumption (A5) implies that the model (2) possesses a symmetric response. Further, since each of the polynomials generically denoted T has real coefficients, the values of T^{-} and T^{+} at $q = e^{i\omega}$ form a complex-conjugate pair, so $T(e^{i\omega},\theta) = \left|T^{-}(e^{i\omega},\theta)\right|^2$ is real and positive for all ω, θ. As we will show, the symmetry property makes it possible to find a causal-equivalent representation for this class of noncausal models. Note that most CD responses in paper machines can be modeled using (2) [24].

2.2 Spectral Equivalence of Causal and Noncausal Models

In order to prove certain convergence results in a later section, we adapt a key causal definition in Ljung [1, p. 27] as follows.

Definition 1 *Let $G(q,\theta) = \sum_{k=-\infty}^{\infty} g_k(\theta)q^{-k}$ be a transfer function depending on a parameter θ. Given a parameter set Ω, call G uniformly stable on Ω when $\sum_{k=-\infty}^{\infty} \sup_{\theta\in\Omega} \left|g_k(\theta)\right| < +\infty$.*

We now present a lemma on the uniform stability of sum and product of uniformly stable transfer functions belonging to the same model set.

Lemma 1 *If the given noncausal filters $G(q,\theta)$ and $H(q,\theta)$ depending on $\theta \in \Omega$ are both uniformly stable on Ω, then the following filters are also uniformly stable on Ω:*

(i) $G(q,\theta)+H(q,\theta)$,
(ii) $G(q,\theta)H(q,\theta)$.

Proof. The sequences of impulse-response coefficients for $G+H$ and GH are, respectively, the sum and the convolution of the individual coefficient sequences for G and H. Both these operations produce an absolutely summable result when applied to a pair of absolutely summable inputs. The result follows. ■

A signal that satisfies assumptions A2 and A3 is said to be quasi-stationary—see Ljung [1, p. 34]. The following result shows that a signal filtered by a stable noncausal model is quasi-stationary.

Theorem 1 *Let $\{w_x\}$ be a quasi-stationary process and let $G(q)$ be a symmetric stable noncausal model of the form $G(q)=\sum_{k=-\infty}^{\infty} g_k q^{-k}$, where $g_k = g_{(-k)}$. Then the filtered signal $s_x = G(q)w_x = G^+(q)G^-(q)w_x$, is also quasi-stationary. Moreover, its spectrum is related to the spectrum of w_x by $\Phi_s(\omega) = |G^-(e^{i\omega})|^4\Phi_w(\omega) = |G^+(e^{i\omega})|^4\Phi_w(\omega)$.*

Proof. The proof of quasi-stationarity of s_x follows the lines of Theorem 2.2 in Ljung [1] and hence it is not repeated here. From the same theorem in Ljung [1], we have

$$\begin{aligned}\Phi_s(\omega) &= [G^+(e^{i\omega})G^-(e^{i\omega})]\Phi_w[G^+(e^{i\omega})G^-(e^{i\omega})]^* \\ &= |G^-(e^{i\omega})|^4\Phi_w(\omega). \end{aligned} \tag{3}$$

The final equation follows due to the symmetry in $G(q)$. ■

Corollary 1 *For the model in (2), the output spectrum is given by $\Phi_y(\omega) = \frac{|B(e^{i\omega},\theta_0)|^2}{|A(e^{i\omega},\theta_0)|^2}\Phi_u + \frac{|C(e^{i\omega},\theta_0)|^2}{|A(e^{i\omega},\theta_0)|^2}\Phi_e$, under the assumption that input u_x is not correlated with the noise e_x.*

This theorem is useful in studying the frequency domain properties of estimated models.

Traditionally, causal models have been identified by minimizing prediction errors (prediction-error methods). In the case of noncausal models, it is not possible to define a prediction error. However, there has been some work reported on using a "two-sided prediction error" [25,26]. The idea in these methods is to use a weighted average of the "past" and "future" values to find the current prediction error. Most of the reported work along these lines focuses on ARMA models. In this paper, our aim is to make identification of noncausal models accessible to the wealth of methods available for identification of causal models. We first show that a causal equivalent of the noncausal sequence $\{y_x\}$ can be generated by a causal model and a possibly different realization of noise. By equivalent time series, we mean a time series with the same spectrum as $\{y_x\}$. We then show that the maximum likelihood estimates produced using data generated from the causal model and the original noncausal model are the same in the probabilistic sense.

In order to estimate the causal-equivalent time series, we need to know the true model itself. Hence, the algorithm proposed in this paper takes an iterative approach. We make an initial guess for the model parameters and use that to estimate the causal-equivalent time series, which in turn is used in identifying a causal-equivalent model. The idea behind the proposed method is shown in Fig. 2.

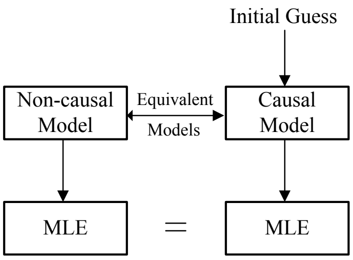


Fig. 2. The main idea of this paper showing equivalence of noncausal model and a corresponding causal model along with the equivalence of their respective maximum likelihood estimates (MLE).

The following result shows that there exists a causal-equivalent model.

Proposition 1 *Consider the noncausal ARMAX model*

$$A(q)y_x = B(q)u_x + C(q)e_x, \tag{4}$$

where

$$\begin{aligned}A(q) &= A^+(q)A^-(q), \text{ such that } A^+(q^{-1}) = A^-(q),\\ B(q) &= B^+(q)B^-(q), \text{ such that } B^+(q^{-1}) = B^-(q),\\ C(q) &= C^+(q)C^-(q), \text{ such that } C^+(q^{-1}) = C^-(q),\end{aligned}$$

and e_x is Gaussian white noise with variance σ^2. Assume that there are no zeros of $A(q), B(q), C(q)$ on the unit circle. Then there exist causal polynomials $\widetilde{A}(q)$, $\widetilde{B}(q)$, $\widetilde{C}(q)$, a Gaussian white noise sequence $\widetilde{e}_x$, and a sequence $\{\widetilde{y}_x\}$ with same spectral characteristics as $\{y_x\}$, satisfying the causal invertible equation

$$\widetilde{A}(q)\widetilde{y}_x = \widetilde{B}(q)u_x + \widetilde{C}(q)\widetilde{e}_x. \tag{5}$$

Moreover, if u_x and e_x are independent, then u_x and $\widetilde{e}_x$ are also independent.

Proof. After enumerating the anti-causal zeros of the respective polynomials A, B, C by α_j, β_j, γ_j, we define the causal, minimum-phase polynomials

$$\begin{aligned}\widetilde{A}(q) &= A(q)\prod_{1\le j\le n_a}\left(\frac{q^{-1}-\alpha_j}{q-\alpha_j}\right) := A(q)\pi_A(q),\\ \widetilde{B}(q) &= B(q)\prod_{1\le j\le n_b}\left(\frac{q^{-1}-\beta_j}{q-\beta_j}\right) := B(q)\pi_B(q),\\ \widetilde{C}(q) &= C(q)\prod_{1\le j\le n_c}\left(\frac{q^{-1}-\gamma_j}{q-\gamma_j}\right) := C(q)\pi_C(q).\end{aligned}$$

Each of π_A, π_B, π_C is an all-pass filter. Consider

$$\widetilde{y}_x = \frac{\pi_B}{\pi_A} y_x. \tag{6}$$

From the definition of π_B and π_A, we know that $\Phi_{\widetilde{y}}(\omega) = \Phi_y(\omega), \forall\omega$. Multiplying both sides of (4) by $\frac{\pi_B}{\pi_A}$, after some manipulations, we arrive at $\widetilde{y}_x = \frac{\widetilde{B}}{\widetilde{A}} u_x + \frac{\widetilde{C}}{\widetilde{A}} \widetilde{e}_x$, where $\widetilde{e}_x = \frac{\pi_B}{\pi_C} e_x$. Note that $\widetilde{e}_x$ is a Gaussian white noise sequence with variance σ^2. We can see that there exist a Gaussian white noise sequence, $\{\widetilde{e}_x\}$, and a sequence $\{\widetilde{y}_x\}$, defined by (6), such that the original noncausal model (4) admits a causal-equivalent representation (5) in the sense that the two have identical output spectra. ■

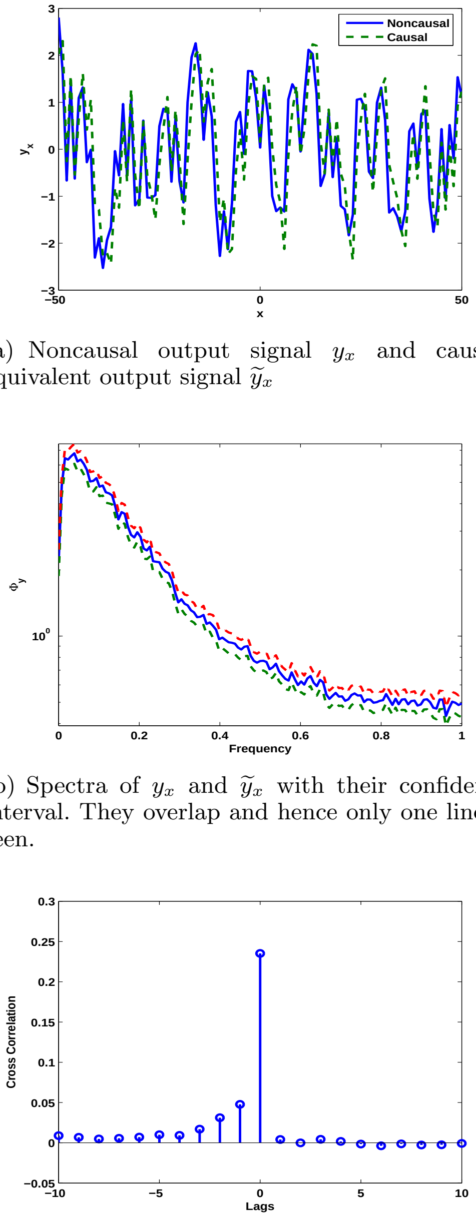
The above proposition associates a spectrally equivalent causal system with any symmetric noncausal model. We propose to estimate the parameters of the noncausal system using maximum likelihood estimates from the causal-equivalent model. The text below provides both theoretical and empirical justification for this approach.

Example 1 Consider this noncausal ARX model:

$$(1 - \theta_1 q^{-1})(1 - \theta_1 q) y_x = (1 - \theta_2 q^{-1})(1 - \theta_2 q) u_x + e_x,$$

with reference parameter values $\theta_1^0 = 0.5$, $\theta_2^0 = 0.2$. The reference model is excited with a white signal (u_x) of unit variance; the noise variance is 0.09. In order to minimize variance errors, we use a very large data set with $N = 1280000$ data points. We first identify the above model with the collected noncausal data using ordinary least squares, as is common with causal ARX models. The results, $\theta_1 = 0.5880$ and $\theta_2 = 0.3143$, are grossly inaccurate. This is to be expected, because the "current" noise is correlated with both "past" and "future" outputs due to noncausality (please refer to [1], p. 205 for a discussion on reasons for bias [2]). Using the same data set but with the causal-equivalent input and output data, the estimated model parameters are much more accurate: $\theta_1 = 0.4998$ and $\theta_2 = 0.1996$.

Example 2 Consider the noncausal ARMAX model (2) with coefficients $A(q) = (1 - 0.5q^{-1})(1 - 0.5q)$, $B(q) = (1 - 0.2q^{-1})(1 - 0.2q)$, $C(q) = (1 - 0.4q^{-1})(1 - 0.4q)$. Then we have $\widetilde{A}(q) = (1 - 0.5q^{-1})(1 - 0.5q^{-1})$, $\widetilde{B}(q) = (1 - 0.2q^{-1})(1 - 0.2q^{-1})$, $\widetilde{C}(q) = (1 - 0.4q^{-1})(1 - 0.4q^{-1})$. This model is simulated using Gaussian white noise of unit variance for e_x and a random binary input in the Nyquist frequency range of 0 to 1 for u_x. The data length $N = 100000$. A plot showing the original noncausal output and the causal-equivalent output is shown in Fig. 3(a). Their spectra are shown in Fig. 3(b). It is clear that even though the two stochastic processes are different they have the same spectrum and moreover, as shown in Fig. 3(c), there is no correlation between $\widetilde{y}_x$ and future $\widetilde{e}_x$.



(a) Noncausal output signal y_x and causal-equivalent output signal $\widetilde{y}_x$

(b) Spectra of y_x and $\widetilde{y}_x$ with their confidence interval. They overlap and hence only one line is seen.

(c) Cross correlation between $\widetilde{y}_x$ and $\widetilde{e}_x$ at different lags

Fig. 3. Noncausal and causal-equivalent output signal

2.3 Open-loop Noncausal Filtering Approach

For the open-loop noncausal ARX system (4), a more intuitive identification approach can be used when the input signal u_x is deterministic. Specifically, system (4) can be equivalently written as

$$y_x = s_x + \frac{(C^-(q))^2}{(A^-(q))^2} \bar{e}_x, \tag{7}$$

where $s_x = \frac{B^+(q)B^-(q)}{A^+(q)A^-(q)} u_x$, $\bar{e}_x = \frac{C^+(q)A^-(q)}{C^-(q)A^+(q)} e_x$. Note that $\bar{e}_x$ is also Gaussian and white with same spectrum as e_x, but

[2] It is important to note that the amount of bias depends on the pole/zero locations of the process and noise transfer functions

perhaps a different realization. Since s_x is deterministic, the predictor form of (7) is

$$\hat{y}_x = \frac{(A^-(q))^2}{(C^-(q))^2} s_x + \frac{(C^-(q))^2 - (A^-(q))^2}{(C^-(q))^2} y_x. \tag{8}$$

The prediction error for the noncausal filtering approach arising from (7)-(8) is

$$\bar{\varepsilon}_x = \frac{(A^-(q))^2}{(C^-(q))^2} (y_x - s_x), \tag{9}$$

whereas the prediction error $\widetilde{\varepsilon}_x$ from (5) is an all-pass version of $\bar{\varepsilon}_x$, i.e., $\bar{\varepsilon}_x = \frac{\pi_B}{\pi_A}\widetilde{\varepsilon}_x$. Thus, asymptotically in sample size, for open-loop identification, the noncausal filtering approach shown here yields identical estimates (also with the same asymptotic properties) as the method proposed in previous section.

Although the noncausal filtering approach in this section seems more straightforward, it is not applicable when u_x is stochastic or generated in closed-loop. In contrast, our method in previous sections can be easily extended to closed-loop case. Moreover, from the implementation perspective, for the noncausal filtering approach, the computation of the prediction error and its gradient with respect to parameters still involve extensive noncausal parts. This requires a special treatment. In particular, the transient effects due to noncausal filtering may not be negligible when the sample size is small. In contrast, our approach relies only on causal identification methods that are readily available in current system identification solvers. For simplicity, in the following sections, we focus on exploring the statistical properties of our approach for open-loop data.

3 Maximum Likelihood Estimation

In this section we demonstrate the asymptotic convergence of a prediction-error type objective function to the log-likelihood function for both causal and noncausal models. We further show that the log-likelihood function of a noncausal model converges asymptotically to that of its causal-equivalent model. Moreover, the maximum likelihood estimates of a noncausal model and its causal-equivalent model coincide if there is a unique maximum associated with the log-likelihood functions. This observation allows us to identify a noncausal model by identifying its causal-equivalent model. The expression for the asymptotic log-likelihood function of stable, causal models was derived in [1] and the expression for nongaussian, noncausal autoregressive models was derived in [9]. We generalize the result in [9] to uniformly stable noncausal ARMAX models.

3.1 Likelihood Function

We consider the following standard objective function in prediction-error methods

$$V^N(\theta) = -\frac{1}{N}\sum_{x=1}^{N} \frac{1}{2} e_x^2(\theta), \tag{10}$$

where $e_x(\theta) = \frac{A(q,\theta)}{C(q,\theta)}\left(y_x - \frac{B(q,\theta)}{A(q,\theta)} u_x\right)$. From (2), we can see that there is a one-to-one correspondence between y_x and e_x. Maximum likelihood estimation theory [2], ensures that maximizing the log-likelihood function of the output data is equivalent to maximizing the log-likelihood function of the noise, e_x. Recall that the (averaged) log-likelihood function of the data given inputs can be expressed as follows (see Appendix A):

$$\mathcal{L}^N = \frac{1}{N}\sum_{x=1}^{N} \log f_e(e_x|u_1, \cdots, u_N, \theta),$$

where $f_y(\cdot)$ and $f_e(\cdot)$ denote the density functions of y_x and e_x respectively. Noting the availability of similar results for the causal situation [1,2], we now show that the objective function in (10) is uniformly close to the true log-likelihood function for both causal and noncausal models, provided that they are uniformly stable and the noise is Gaussian.

Proposition 2 *Let $\mathcal{L}_c^N(\widetilde{\mathbf{y}}, \theta)$ and $\mathcal{L}^N(\mathbf{y}, \theta)$ denote the average log-likelihood functions of the data sets $\{\widetilde{y}_1, \widetilde{y}_2, \cdots, \widetilde{y}_N\}$ and $\{y_1, y_2, \cdots, y_N\}$ generated respectively by the causal and noncausal models in* (5) *and* (4)*. Let $K_V = \log \sigma\sqrt{2\pi}$. If the filters generating both the causal and noncausal data sets are uniformly stable, then, as $N \to \infty$,*

$$\sup_{\theta\in\Omega} \left| \frac{1}{\sigma^2} V_c^N(\theta) - \mathcal{L}_c^N(\widetilde{\mathbf{y}}, \theta) - K_V \right| \to 0, \quad \textit{w.p.1},$$
$$\sup_{\theta\in\Omega} \left| \frac{1}{\sigma^2} V^N(\theta) - \mathcal{L}^N(\mathbf{y}, \theta) - K_V \right| \to 0, \quad \textit{w.p.1}.$$

Here V^N is as defined in (10) *and V_c^N is defined analogously for the model in* (5)*.*

Proof. We generalize the proof in [9] to uniformly stable noncausal ARMAX models. See Appendix A for details. ■

The above proposition suggests using the sum of squared errors, $\widetilde{e}_x(\theta)$, as our objective function for finding the maximum likelihood estimate. As we show in the next section, the uniform convergence established above is the key to convergence of the corresponding parameter estimates.

3.2 Convergence of Causal and Noncausal Parameter Estimates

We now prove our main theoretical result. It shows that the likelihood function of the original noncausal model is uniformly close to the likelihood function of the causal-equivalent model, with the gap closing as the data set grows. This result makes use of the following extension of Theorem 2B.1 in Ljung [1], which states that a signal generated by a uniformly stable family of systems is 'uniformly ergodic'.

Theorem 2 *Consider two uniformly stable families of (possibly noncausal) filters $G(q,\theta)$ and $H(q,\theta)$, $\theta \in \Omega$. Let u_x be a bounded signal, i.e., $\sup_x |u_x| < +\infty$, and define signals $s_x(\theta) = G(q,\theta)u_x + H(q,\theta)e_x$ for each θ, where e_x is Gaussian white noise with variance σ^2. Then the following statement holds with probability 1: as $N \to \infty$,*

$$\sup_{\theta\in\Omega} \left\| \frac{1}{N} \sum_{x=1}^{N} \left[s_x(\theta)s_x(\theta)^T - \mathbb{E}s_x(\theta)s_x(\theta)^T \right] \right\| \to 0.$$

Proof. The proof is very similar to that of Theorem 2B.1 in Ljung [1] and hence omitted. The only difference is in making use of the noncausal Definition 1 in this paper. ■

The above theorem will be used many times in the rest of this paper. It shows that for a signal generated by any set of uniformly stable filters, the sample covariance converges uniformly to the ensemble covariance as the data length tends to infinity.

Proposition 3 *Assume that the following filters are uniformly stable on Ω:*

$$\frac{B(q,\theta)}{C(q,\theta)}, \quad \frac{A(q,\theta)}{C(q,\theta)}, \quad \frac{\pi_B(q,\theta)}{\pi_C(q,\theta)}. \tag{11}$$

Define $\mathcal{L}(\theta) = \sigma^2(\theta) - \sigma\sqrt{2\pi}$, where

$$\sigma^2(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(\left| \frac{A(e^{i\omega},\theta)B(e^{i\omega},\theta_0)}{C(e^{i\omega},\theta)A(e^{i\omega},\theta_0)} - \frac{B(e^{i\omega},\theta)}{C(e^{i\omega},\theta)} \right|^2 \Phi_u + \left| \frac{A(e^{i\omega},\theta)C(e^{i\omega},\theta_0)}{C(e^{i\omega},\theta)A(e^{i\omega},\theta_0)} \right|^2 \Phi_e \right) d\omega. \tag{12}$$

Then as $N \to \infty$,

$$\sup_{\theta\in\Omega} \left| \mathcal{L}^N(\mathbf{y},\theta) - \mathcal{L}(\theta) \right| \overset{w.p.1}{\to} 0, \tag{13}$$

$$\sup_{\theta\in\Omega} \left| \mathcal{L}_c^N(\widetilde{\mathbf{y}},\theta) - \mathcal{L}(\theta) \right| \overset{w.p.1}{\to} 0, \tag{14}$$

$$\sup_{\theta\in\Omega} \left| \mathcal{L}_c^N(\widetilde{\mathbf{y}},\theta) - \mathcal{L}^N(\mathbf{y},\theta) \right| \overset{w.p.1}{\to} 0. \tag{15}$$

Proof. We begin with (15). Recall that $\theta_0 \in \Omega$ by assumption. For any given $\theta \in \Omega$, the error $e_x(\theta)$ satisfies

$$e_x(\theta) = \left(\frac{A(q,\theta)B(q,\theta_0)}{C(q,\theta)A(q,\theta_0)} - \frac{B(q,\theta)}{C(q,\theta)} \right) u_x + \left(\frac{A(q,\theta)C(q,\theta_0)}{C(q,\theta)A(q,\theta_0)} \right) e_x(\theta_0). \tag{16}$$

Similarly, the error for the causal-equivalent model is

$$\widetilde{e}_x(\theta) = \frac{\pi_B(q,\theta)}{\pi_C(q,\theta)} \left(\frac{A(q,\theta)B(q,\theta_0)}{C(q,\theta)A(q,\theta_0)} - \frac{B(q,\theta)}{C(q,\theta)} \right) u_x + \frac{\pi_B(q,\theta)}{\pi_C(q,\theta)} \left(\frac{A(q,\theta)C(q,\theta_0)}{C(q,\theta)A(q,\theta_0)} \right) e_x(\theta_0). \tag{17}$$

Now consider the function $\sigma^2(\theta)$ defined in (12). Combining Parseval's theorem, Theorem 2, and Lemma 1, we have, as $N \to \infty$,

$$\sup_{\theta\in\Omega} \left| -\frac{1}{N} \sum_{x=1}^{N} e_x^2(\theta) + \sigma^2(\theta) \right| \overset{w.p.1}{\to} 0, \tag{18}$$

$$\sup_{\theta\in\Omega} \left| -\frac{1}{N} \sum_{t=1}^{N} \widetilde{e}_x^2(\theta) + \sigma^2(\theta) \right| \overset{w.p.1}{\to} 0. \tag{19}$$

Since (for each $\theta \in \Omega$ and $\omega \in \mathbb{R}$) $|A(e^{i\omega})| = |\widetilde{A}(e^{i\omega})|$, $|B(e^{i\omega})| = |\widetilde{B}(e^{i\omega})|$ and $|C(e^{i\omega})| = |\widetilde{C}(e^{i\omega})|$, the variances of $e_x(\theta)$ and $\widetilde{e}_x(\theta)$ uniformly converge to σ^2. From the triangle inequality, for any given N, we have

$$\sup_{\theta\in\Omega} \left| V_c^N(\theta) - V^N(\theta) \right| \le \sup_{\theta\in\Omega} \left| V_c^N(\theta) + \frac{1}{2}\sigma^2(\theta) \right| + \sup_{\theta\in\Omega} \left| V^N(\theta) + \frac{1}{2}\sigma^2(\theta) \right|. \tag{20}$$

From (18) and (19), it follows that the probability space $\mathscr{X}$ contains two subsets Γ_c and Γ (possibly empty) such that $\mathbb{P}(\Gamma) = \mathbb{P}(\Gamma_c) = 0$ and

$$\sup_{\theta\in\Omega} \left| -\frac{1}{N} \sum_{x=1}^{N} e_x^2(\xi,\theta) + \sigma^2(\theta) \right| \to 0 \; \forall \; \xi \notin \Gamma,$$

$$\sup_{\theta\in\Omega} \left| -\frac{1}{N} \sum_{x=1}^{N} \widetilde{e}_x^2(\xi,\theta) + \sigma^2(\theta) \right| \to 0 \; \forall \xi \notin \Gamma_c.$$

(Note that $\widetilde{\mathbf{y}}$ and $\mathbf{y}$ also depend on the random parameter $\xi \in \mathscr{X}$ through the noise term.) For each ξ in $\mathscr{X} - (\Gamma \cup \Gamma_c)$, sending $N \to \infty$ in (20) and using Proposition 2 gives

$$\lim_{N\to\infty} \sup_{\theta\in\Omega} \left| \mathcal{L}_c^N(\widetilde{\mathbf{y}}(\xi),\theta) - \mathcal{L}^N(\mathbf{y}(\xi),\theta) \right| = 0. \tag{21}$$

Of course $\mathbb{P}(\Gamma_c \cup \Gamma) \le \mathbb{P}(\Gamma_c) + \mathbb{P}(\Gamma) = 0$, so (15) follows.

Turning to (13), we use Appendix A and Proposition 2 to get

$$\sup_{\theta\in\Omega}\left|V^N(\theta)/\sigma^2 - K_V - \mathcal{L}^N(\mathbf{y},\theta)\right| \overset{w.p.1}{\to} 0. \tag{22}$$

Then, from the triangle inequality,

$$\begin{aligned}
&\sup_{\theta\in\Omega}\left|-\sigma^2(\theta)/\sigma^2 - K_V - \mathcal{L}^N(\mathbf{y},\theta)\right| \le \\
&\quad \sup_{\theta\in\Omega}\left|-\sigma^2(\theta)/\sigma^2 - V^N(\theta)/\sigma^2\right| \\
&\quad + \sup_{\theta\in\Omega}\left|V^N(\theta)/\sigma^2 - K_V - \mathcal{L}^N(\mathbf{y},\theta)\right|.
\end{aligned}$$

Thanks to (22) and (18), the argument from this inequality to (13) is very similar to the one from (20) to (21). Likewise for (14). ■

Corollary 2 *Under the assumptions in Proposition 3,*

$$\left\|\max_{\theta\in\Omega}\mathcal{L}_c^N(\widetilde{\mathbf{y}},\theta) - \max_{\theta\in\Omega}\mathcal{L}^N(\mathbf{y},\theta)\right\| \overset{w.p.1}{\to} 0, \tag{23}$$

$$\left\|\max_{\theta\in\Omega}\mathcal{L}^N(\mathbf{y},\theta) - \max_{\theta\in\Omega}\mathcal{L}(\theta)\right\| \overset{w.p.1}{\to} 0, \tag{24}$$

$$\left\|\max_{\theta\in\Omega}\mathcal{L}_c^N(\widetilde{\mathbf{y}},\theta) - \max_{\theta\in\Omega}\mathcal{L}(\theta)\right\| \overset{w.p.1}{\to} 0. \tag{25}$$

Proof. All three statements follow from Proposition 3 and the following inequality, valid for any real-valued functions f and g on Ω:

$$\left|\sup_{\theta\in\Omega} f - \sup_{\theta\in\Omega} g\right| \le \sup_{\theta\in\Omega}\left|f(\theta) - g(\theta)\right|. \tag{26}$$

To justify (26), suppose for simplicity that f and g attain their suprema, i.e., some $\theta_f, \theta_g \in \Omega$ obey $f(\theta_f) = \sup_{\theta\in\Omega} f$, $g(\theta_g) = \sup_{\theta\in\Omega} g$. Then clearly $f(\theta_g) \le f(\theta_f)$, i.e., $-f(\theta_f) \le -f(\theta_g)$. Adding $g(\theta_g)$ to both sides gives

$$g(\theta_g) - f(\theta_f) \le g(\theta_g) - f(\theta_g) \le \left|g(\theta_g) - f(\theta_g)\right|. \tag{27}$$

Repeating this argument with the roles of f and g reversed establishes

$$f(\theta_f) - g(\theta_g) \le \left|f(\theta_f) - g(\theta_f)\right|. \tag{28}$$

Together, inequalities (27) and (28) confirm our claim:

$$\begin{aligned}
\left|\sup_{\theta\in\Omega} f - \sup_{\theta\in\Omega} g\right| &\le \max\left\{\left|f(\theta_f) - g(\theta_f)\right|, \left|f(\theta_g) - g(\theta_g)\right|\right\} \\
&\le \sup_{\theta\in\Omega}\left|f(\theta) - g(\theta)\right|.
\end{aligned}$$

Now consider (23). For each fixed $\xi \in \mathcal{X}$ such that $\sup_{\theta\in\Omega}\left|\mathcal{L}_c^N(\widetilde{\mathbf{y}},\theta) - \mathcal{L}^N(\mathbf{y},\theta)\right| \to 0$ as $N \to \infty$, defining $f(\theta) = \mathcal{L}_c^N(\widetilde{\mathbf{y}},\theta)$ and $g(\theta) = \mathcal{L}^N(\mathbf{y},\theta)$ and applying (26) implies $\left|\sup_{\theta\in\Omega}\mathcal{L}_c^N(\widetilde{\mathbf{y}},\theta) - \sup_{\theta\in\Omega}\mathcal{L}^N(\mathbf{y},\theta)\right| \to 0$ as $N \to \infty$. According to Proposition 3, the set of ξ to which this reasoning applies has probability 1. Similar arguments establish (24) and (25). ■

The results above show that the log-likelihood function of the causal-equivalent model uniformly approximates that for the original noncausal model as the sample count tends to infinity. Further, the maximum values of the causal and noncausal log-likelihood functions are asymptotically equal, given that both causal and noncausal models are uniformly stable. It remains to prove that the set of maximizing parameters for each of these two functions are the same, so that we can use the causal-equivalent model for identification. This requires some care, because (as shown in Example 3 below) uniform convergence of a sequence of functions is not enough to guarantee convergence of their maximizing inputs.

Proposition 4 *Consider the following sets of maximizing parameters:*

$$\begin{aligned}
\mathcal{D}_c^N &= \operatorname*{argmax}_{\theta\in\Omega}\mathcal{L}_c^N(\widetilde{\mathbf{y}}_N,\theta), \\
\mathcal{D}^N &= \operatorname*{argmax}_{\theta\in\Omega}\mathcal{L}^N(\mathbf{y}_N,\theta), \\
\mathcal{D} &= \operatorname*{argmax}_{\theta\in\Omega}\mathcal{L}(\theta),
\end{aligned}$$

where $\mathcal{L}_c^N$, $\mathcal{L}^N$, and $\mathcal{L}$ are defined in Propositions 2 and 3 above. Then, with probability 1, we have

(i) Each of the sets $\mathcal{D}_c^N$, $\mathcal{D}^N$, and $\mathcal{D}$ is closed and nonempty.

(ii) $\emptyset \ne \limsup_{N\to\infty}\ \mathcal{D}_c^N \subseteq \mathcal{D}$.

(iii) $\emptyset \ne \limsup_{N\to\infty}\ \mathcal{D}^N \subseteq \mathcal{D}$.

(iv) In particular, if $\mathcal{D} = \{\theta_0\}$ is a singleton, then one has both $\theta_c^N \to \theta_0$ and $\theta^N \to \theta_0$ for any sequences $\theta_c^N \in \mathcal{D}_c^N$ and $\theta^N \in \mathcal{D}^N$.

Note that the "lim sup" operation in the statement above applies to a sequence of sets. For a generic sequence of sets $D^N \subseteq \mathbb{R}^n$, we have $\theta \in \limsup_{N\to\infty} D^N$ if and only if $\theta = \lim_{N\to\infty}\theta^N$ for some sequence $\{\theta^N\}$ with the property that $\theta^N \in D^N$ for infinitely many N (See Resnick [27], p. 6).

Remark. The conclusions of Proposition 4 can be improved by replacing the sets $\mathcal{D}^N, \mathcal{D}_c^N$ with the respective enlargements defined as follows. Let $r_N \ge 0$ be any nonnegative sequence of scalars obeying $r_N \to 0$ as $N \to \infty$, and $\widehat{\mathcal{D}}^N = \left\{\theta + r_N u :\ \theta \in \mathcal{D}^N,\ |u| \le 1\right\} \cap \Omega$, $\widehat{\mathcal{D}}_c^N = \left\{\theta + r_N u :\ \theta \in \mathcal{D}_c^N,\ |u| \le 1\right\} \cap \Omega$. This is relevant because iterative schemes designed to find points in $\mathcal{D}_c^N$ typically return inexact results, which can be described using sets like $\widehat{\mathcal{D}}_c^N$.

Proof. The stated conclusions follow from the basic properties of the log-likelihood functions involved here and the uniform convergence properties established in the cited propositions. For any fixed realization ξ in $\mathcal{X}$ of the random processes considered here (outside some set with zero probability), we reason as follows. All three variants of $\mathcal{L}(\theta)$ are continuous, and any continuous function is guaranteed to attain a maximum value over any compact set; the set of maximizers must be closed. This establishes (i). Further, all three variants of $\mathcal{L}(\theta)$ are continuously differentiable functions of the parameter vector θ, whose gradients are uniformly bounded on the set Ω as shown in Section 5. Thus the deterministic arguments detailed in Appendix B establish (ii) and (iii). As for (iv), when $\mathcal{D} = \{\theta_0\}$ is a singleton set, the compactness of Ω guarantees that every sequence $\{\theta^N\}$ obeying $\theta^N \in \mathcal{D}^N$ for each N must have a convergent subsequence; from (iii), that subsequence must converge to θ_0. Since this reasoning applies to arbitrary sequences $\theta^N \in \mathcal{D}^N$, it is impossible for any such sequence to fail to converge to θ_0. The situation for $\mathcal{D}_c^N$ is analogous. ■

Remark 1 The above result generalizes Theorem 8.2 in Ljung [1] by allowing non-singleton sets for $\mathcal{D}^N$. The results in *(ii)* and *(iii)* show that the maximum-likelihood parameter estimates from the causal-equivalent model and the original noncausal model asymptotically approach true maximizers of the asymptotic log-likelihood function. If there is a unique maximizer in the limit, then the causal and non-casual data will provide the same asymptotic estimates. (Example 3 suggests that the causal and noncausal estimates could differ when $\mathcal{D}$ contains more than one element—a situation we consider unlikely.)

From now on, we focus on maximum likelihood estimation of the causal-equivalent model. Variance expressions for estimates of causal models are derived in Ljung [1] (p. 291). In Section 5, we provide variance expressions for the causal-equivalent model with its special structure and also show that if $\limsup_{N\to\infty} \bar{\mathcal{D}}_c^N \cap \limsup_{N\to\infty} \bar{\mathcal{D}}^N \neq \emptyset$, then the variance of the common maxima of causal and noncausal models will be asymptotically identical.

Example 3 Uniform proximity between the causal and noncausal models may not be enough to guarantee that the corresponding maximum likelihood parameter estimates are close together. Here is a deterministic analytical model that clarifies the issue. We use a one-dimensional parameter set $\Omega = [-1.5, 1.5]$ and define $f_N(\theta) = -(\theta^2 - 1)^2 - \frac{\theta}{N}$, $g_N(\theta) = -(\theta^2 - 1)^2 + \frac{\theta}{N}$. Both sequences $\{f_N\}$, $\{g_N\}$ converge uniformly on Ω to the same limit function $h(\theta) = -(\theta^2 - 1)^2$. The discrepancies are caused by linear perturbations that shrink as N increases: these perturbations make $f_N > h_N > g_N$ when $\theta < 0$, but reverse these inequalities when $\theta > 0$. As illustrated in Fig. 4, the maximizers of $\{f_N\}$ form a sequence θ_N^f converging to $\theta^f = -1$, whereas the maximizers of $\{g_N\}$ form a sequence θ_N^g converging to $\theta^g = +1$. Of course the maximum values converge to 0 (which is the maximum value of h), and both parameter sequences mentioned above converge to inputs that maximize h. The existence of distinct global maximizers for the limit function h allows these two outcomes to hold even though $\left|\theta_N^f - \theta_N^g\right| \geq 2$ for all N.

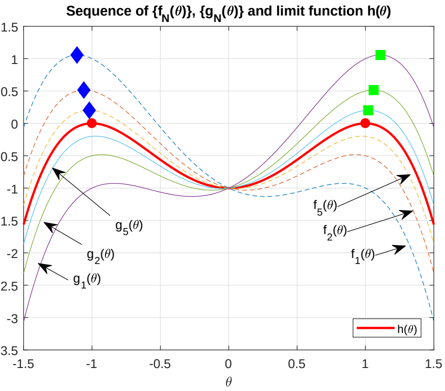


Fig. 4. Sample functions f_N, g_N and h in Example 3.

For the identification problem considered in this paper, we have not imposed hypotheses to guarantee that the limiting log-likelihood function $\mathcal{L}(\theta)$ defined in Proposition 3 achieves its maximum at a unique point on Ω. When it does (which we expect to be typical in practice), conclusion (iv) of Proposition 4 applies.

4 Identification Algorithm

From the previous section one can see that if the log-likelihood function has a unique maximizer, then the maximum likelihood estimate of the causal-equivalent model converges to that of the original noncausal model. This result lays a theoretical foundation for using the causal-equivalent data $\{\widetilde{y}_x, u_x\}$ and corresponding causal identification techniques as a way to identify a symmetric noncausal model. However, as shown in (6), acquiring the causal-equivalent data depends on an unknown all-pass filter $\frac{\pi_B}{\pi_A}$. We propose an iterative approach to deal with this. To start, one must have input-output measurements u_x and y_x for $x \in \{1, 2, \ldots, N\}$ and an approximation, θ, for the true parameter vector θ_0. Often an initial guess can be obtained from physical insight into the process. For instance, if we know an approximate time constant, then it can be used as a first guess for that parameter. Iteration proceeds as follows:

(1) Use the current approximation θ to compute $\widetilde{y}_x(\theta) = (\pi_B(\theta)/\pi_A(\theta))y_x$.
(2) Use the data $\widetilde{y}_x$ obtained from step 1 with u_x to identify the structured model in (5).
(3) Extract an updated approximation for θ from the identified model.
(4) If Steps 2 and 3 produce a sufficiently small change in the original vector θ, declare success and stop; otherwise, return to Step 1.

The steps above are supported by standard software. Our numerical experiments used Matlab's System Identification Toolbox, specifically, the function `idgrey` to encode the structured linear model and the function pem to perform the optimization.

To formalize the iterations with a view to convergence analysis, let $F(q,\theta) = \frac{\pi_B(q,\theta)}{\pi_A(q,\theta)}$ denote the filter applied in step (1). Note that $F(q,\theta)$ is all-pass for each θ. Then define a version of the average prediction-error as a function of inputs in $\Omega \times \Omega$:

$$J^N(\theta,\theta') = \frac{1}{N}\sum_{x=1}^{N}\left[\frac{\widetilde{A}(q,\theta)}{\widetilde{C}(q,\theta)}\left(\widetilde{y}_x(\theta') - \frac{\widetilde{B}(q,\theta)}{\widetilde{A}(q,\theta)}u_x\right)\right]^2,$$

where $\widetilde{y}_x(\theta') = F(q,\theta')y_x$. When the current parameter estimate in the iteration is $\theta = \theta_n$, step (1) of the algorithm produces $\widetilde{y}_x(\theta_n)$ and then steps (2)–(3) define $\theta_{n+1} = \operatorname{argmin}_{\theta\in\Omega} J^N(\theta,\theta_n)$. In particular,

$$J^N(\theta_{n+1},\theta_n) \leq J^N(\theta,\theta_n), \forall\theta\in\Omega. \tag{29}$$

The following proposition addresses the convergence of the algorithm above in the large-N limit. We define $J(\theta,\theta') = \lim\limits_{N\to\infty} J^N(\theta,\theta')$.

Proposition 5 *In the large-N limit, each iteration of the identification algorithm above decreases the prediction error. More precisely, as $N\to\infty$, we have*

$$J(\theta_{n+1},\theta_{n+1}) \leq J(\theta_n,\theta_n) \quad \textit{for}\; n = 1,2,\ldots. \tag{30}$$

Proof. As detailed above (see (29)), if $\theta = \theta_n$ in step (1) of the algorithm, then steps (2)–(3) produce θ_{n+1} for which (asymptotically in N) $J(\theta_{n+1},\theta_n) \leq J(\theta,\theta_n), \quad \forall \;\theta\in\Omega$. In particular,

$$J(\theta_{n+1},\theta_n) \leq J(\theta_n,\theta_n). \tag{31}$$

The next step in the iteration holds the identified causal model fixed and re-filters the observations with the most recent parameter estimate, θ_{n+1}. This updates the prediction error objective by effectively substituting $\theta' = \theta_{n+1}$ in $J(\theta_{n+1},\theta')$. To assess this, we rewrite [3]

$$J^N(\theta,\theta') = \frac{1}{N}\sum_{x=1}^{N}\left[\frac{\pi_A A(\theta)}{\pi_C C(\theta)}\left(\widetilde{y}_x(\theta') - \frac{F(\theta)B(\theta)}{A(\theta)}u_x\right)\right]^2 \tag{32}$$

From the true system model (2), we have

$$\widetilde{y}_x(\theta') = F(\theta')\frac{B(\theta_0)}{A(\theta_0)}u_x + F(\theta')\frac{C(\theta_0)}{A(\theta_0)}e_x. \tag{33}$$

Substituting (33) in (32) and using Parseval's theorem in the limit $N\to\infty$, we get

$$J(\theta,\theta') = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\left|\frac{A(\theta)}{C(\theta)}\right|^2\left|\frac{B(\theta_0)}{A(\theta_0)} - \frac{F(\theta)}{F(\theta')}\frac{B(\theta)}{A(\theta)}\right|^2\Phi_u \right.$$
$$\left. + \left|\frac{A(\theta)C(\theta_0)}{C(\theta)A(\theta_0)}\right|^2\Phi_e\right)d\omega. \tag{34}$$

We have used the independence between $\{u_x\}$ and $\{e_x\}$ and the fact that F is always an all-pass filter to simplify the expression.

We now apply (34) with $\theta = \theta_{n+1}$:

$$J(\theta_{n+1},\theta') = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\left|\frac{A(\theta_{n+1})}{C(\theta_{n+1})}\right|^2\left|\frac{B(\theta_0)}{A(\theta_0)} - \frac{F(\theta_{n+1})}{F(\theta')}\right.\right.$$
$$\left.\left.\cdot\frac{B(\theta_{n+1})}{A(\theta_{n+1})}\right|^2\Phi_u + \left|\frac{A(\theta_{n+1})C(\theta_0)}{C(\theta_{n+1})A(\theta_0)}\right|^2\Phi_e\right)d\omega. \tag{35}$$

Now both $B(\theta)$ and $A(\theta)$ are noncausal and symmetric, so their values when $q = e^{i\omega}$ are real and positive. We therefore let

$$r(\theta,\omega) = \frac{B(\theta,\omega)}{A(\theta,\omega)}, \tag{36}$$

noting that $r(\theta,\omega) > 0$ always. In addition, $F(\theta)$ is an all pass filter and therefore

$$F(\theta,\omega) = e^{i\phi(\theta,\omega)}, \;\; \forall\;\omega, \tag{37}$$

for some function ϕ. Using (36) and (37) and dropping the frequency variable for simplicity,

$$\begin{aligned}
&\left|\frac{B(\theta_0)}{A(\theta_0)} - \frac{F(\theta_{n+1})}{F(\theta')}\frac{B(\theta_{n+1})}{A(\theta_{n+1})}\right|^2\\
&= \left|r(\theta_0) - r(\theta_{n+1})e^{i[\phi(\theta_{n+1})-\phi(\theta_n)]}\right|^2\\
&= (r(\theta_0) - r(\theta_{n+1})\cos([\phi(\theta_{n+1}) - \phi(\theta')])^2\\
&\qquad +(r(\theta_{n+1})\sin([\phi(\theta_{n+1}) - \phi(\theta')]))^2\\
&= r^2(\theta_0) + r^2(\theta_{n+1})\\
&\qquad -2r(\theta_0)r(\theta_{n+1})\cos([\phi(\theta_{n+1}) - \phi(\theta')]).
\end{aligned} \tag{38}$$

The cosine value that minimizes the expression in (38) is 1, and this is achieved when $\phi(\theta') - \phi(\theta_{n+1})$ is an integer multiple of 2π. In particular, the choice $\theta' = \theta_{n+1}$ is a minimizer, giving

$$\theta_{n+1} \in \operatorname*{argmin}_{\theta'\in\Omega} J(\theta_{n+1},\theta'). \tag{39}$$

In combination with (31), this implies the desired result:

$$J(\theta_{n+1},\theta_{n+1}) \leq J(\theta_{n+1},\theta_n) \leq J(\theta_n,\theta_n). \tag{40}$$

[3] For simplicity, we omit the arguments q in the time domain and $e^{i\omega}$ in the frequency domain.

■

The numerical experiments described in Section 6 confirm that the descent property established above is typical in situations where the number of measurements is sufficient, and, more generally, whenever the signal-to-noise ratio is sufficiently large.

5 Input Design

Input design for causal systems is a well established area of research. In general, input design methods shape the uncertainty of the identified model in a particular fashion that suits the user. The early work on experiment design was based on asymptotic variance expressions first derived in [28]. These variance expressions are asymptotic in both model order and sample size. These methods have also been successfully implemented in practice [29]. However, recent work has shown that variance expressions based on asymptotic model order are not accurate [30,31,32]. In [31], it was shown that if the uncertainty is large, asymptotic expressions lead to inaccurate results. However, if a particular condition is satisfied by the model class then the asymptotic results provide reasonable accuracy despite large uncertainty. In view of the results showing unreliability of asymptotic variance expressions, a number of results based on convex optimization techniques have appeared in the literature [33,22]. A more recent technique uses graph theory to solve the nonconvex problem of closed-loop input design in the presence of input and output probabilistic bounds as well as non-linear feedback [34]. For a good summary of these methods see [35,36] and the references therein.

In this paper, we do not intend to develop a new method for input design. We would like to utilize the existing methods for the noncausal process. In order to achieve that, we need to show asymptotic convergence of the covariance estimate of the causal-equivalent model to that of the original noncausal model. This in turn makes it possible to use the covariance estimates of the causal equivalent model for input design.

5.1 Convergence of Covariance Estimates

The Cramer-Rao lower bound provides a lower bound on the variance of an unbiased estimate $\hat{\theta}$ of θ_0 and is given by Ljung [1, p. 214]: $\mathrm{cov}(\hat{\theta}) \geq M_0^{-1}$, where M_0 is the Fisher Information Matrix [4]

$$M_0 = \mathbb{E}\left\{\nabla\mathcal{L}_c(\widetilde{\mathbf{y}},\theta_0)^T\nabla\mathcal{L}_c(\widetilde{\mathbf{y}},\theta_0)\right\}. \tag{41}$$

Maximum likelihood estimates achieve the Cramer-Rao lower bound. Hence, in this section we show that the Cramer-Rao lower bounds of the causal and noncausal models are asymptotically equal.

For inspiration, recall Proposition 2, which established (with probability 1) two limiting relations, namely, the original and causal-equivalent versions of $\mathcal{L}^N(\theta) \to \sigma^2(\theta) + \sigma\sqrt{2\pi}$ as $N \to \infty$. This convergence is uniform with respect to $\theta \in \Omega$. Now we need the following improvement of these relations involving derivatives.

Lemma 2 *Let us assume that $\mathcal{L}^N(\theta)$ and $\mathcal{L}^N(\theta)$ are differentiable for all N with respective derivatives $\nabla\mathcal{L}^N(\theta)$ and $\nabla\mathcal{L}_c^N(\theta)$ that are continuous. Also assume that the derivatives converge uniformly then,*

$$\sup_{\theta\in\Omega}\left\|\nabla\mathcal{L}^N(\theta) - \nabla\sigma^2(\theta)\right\| \overset{w.p.1}{\to} 0, \tag{42}$$

$$\sup_{\theta\in\Omega}\left\|\nabla\mathcal{L}_c^N(\theta) - \nabla\sigma^2(\theta)\right\| \overset{w.p.1}{\to} 0, \tag{43}$$

and

$$\sup_{\theta\in\Omega}\left\|\nabla\mathcal{L}^N(\theta) - \nabla\mathcal{L}_c^N(\theta)\right\| \overset{w.p.1}{\to} 0. \tag{44}$$

Proof. See Appendix C. ■

Lemma 3 *Let $F^N, G^N : \Omega \to \mathbb{R}^n$ be sequences of functions satisfying $\sup_{\theta\in\Omega}\left|F^N(\theta) - G^N(\theta)\right| \to 0$ as $N \to \infty$. In order to deduce that*

$$\sup_{\theta\in\Omega}\left|m^N(F^N(\theta)) - m^N(G^N(\theta))\right| \to 0. \tag{45}$$

for some given sequence of continuously differentiable functions $m^N : \mathbb{R}^n \to \mathbb{R}$, it suffices to show that

(a) there exists $R > 0$ such that $\left|F^N(\theta)\right| \leq R$ for all $N \in \mathbb{N}$ and all $\theta \in \Omega$, and
(b) there exists a continuous function $\phi\colon \mathbb{R}^n \to \mathbb{R}$ such that $\sup_N\left|\nabla m^N(p)\right| \leq \phi(p), p \in \mathbb{R}^n$, where ∇ is the gradient w.r.t. p in this lemma.

Proof. See Appendix D. ■

Theorem 3 *Let us suppose that the assumptions in Propositions 1, 2, 4 and Lemma 3 are satisfied and that the maximum likelihood estimates of the causal and noncausal models are obtained as shown in Proposition 3. Furthermore, assume that the maximizer of the likelihood function is unique, as in Proposition 4(iv). Then for the covariances of the estimated parameter vectors obtained using causal and noncausal likelihood functions, given by*

$$\begin{aligned} R_c^N &= \left[\mathbb{E}\left\{\nabla\mathcal{L}_c^N(\widetilde{\mathbf{y}},\theta_0)^T\nabla\mathcal{L}_c^N(\widetilde{\mathbf{y}},\theta_0)\right\}\right]^{-1}, \\ R^N &= \left[\mathbb{E}\left\{\nabla\mathcal{L}^N(\mathbf{y},\theta_0)^T\nabla\mathcal{L}^N(\mathbf{y},\theta_0)\right\}\right]^{-1}, \end{aligned} \tag{46}$$

we have

$$|R_c^N - R^N| \overset{w.p.1}{\to} 0. \tag{47}$$

[4] Note that ∇ denotes the gradient with respect to parameter vector θ unless otherwise specified.

Proof. Using Lemma 3 and the conditions (a) and (b) in conjunction with Lemma 2, the result in this theorem follows. The condition (a) in Lemma 3 is valid if $\nabla\mathcal{L}_c^N(\widetilde{\mathbf{y}},\theta)$ and $\nabla\mathcal{L}^N(\mathbf{y},\theta)$ are bounded by some constants R_c and R respectively. This is usually true for stable systems in the neighborhood of the true parameter vector θ_0. The condition in (b) is satisfied if $\nabla R_c^N(\theta)$ and $\nabla R^N(\theta)$ are bounded by continuous function of θ independent of N. ■

Remark 2 This is an important result that is useful in designing experiments. It states that the covariance estimates of the causal and noncausal models converge to the same value asymptotically. Hence, inputs for identification of the original noncausal model can be designed by minimizing the parameter covariance of the causal-equivalent model. In the following paragraphs we adapt this approach and design inputs.

5.2 Input design

It is well known that for the (causal) prediction error method, when the noise is Gaussian and the criterion is quadratic, the covariance of parameter estimates coincides with the Cramer-Rao lower bound as $N \to \infty$ [1, p. 287]. From Theorem 3 it follows that, asymptotically, the Cramer-Rao lower bounds for the causal model (4) and the noncausal model (5) are equivalent. Thus, the input design of noncausal models can be instead conducted based on the asymptotic parameter covariance of causal-equivalent models.

For the structured causal model (5), the one-step-ahead prediction at x is $\hat{y}(x,\theta) = \frac{\widetilde{B}}{\widetilde{C}}u_x + \left(1 - \frac{\widetilde{A}}{\widetilde{C}}\right)y_x$. Following the result in Ljung [1, p. 282], the asymptotic parameter covariance obeys

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \to \mathcal{N}(0, P_\theta), \quad \text{as } N \to \infty, \tag{48}$$

where $\psi(x,\theta) = \nabla\hat{y}(x,\theta)$ helps define

$$P_\theta = \sigma^2(\theta_0)E\left[\psi(x,\theta_0)\psi(x,\theta_0)^T\right]^{-1}.$$

Now assuming open loop conditions, the inverse of the covariance matrix can be expressed in the frequency domain as a linear function of the input spectrum, $\Phi_u(\omega)$ [1] (p. 291),

$$\begin{aligned} P_\theta^{-1} = & \frac{1}{2\pi\sigma^2}\int_{-\pi}^{\pi}\mathcal{F}_u(e^{i\omega},\theta_0)\Phi_u(\omega)\mathcal{F}_u^*(e^{i\omega},\theta_0)d\omega \\ & +\frac{1}{2\pi}\int_{-\pi}^{\pi}\mathcal{F}_e(e^{i\omega},\theta_0)\mathcal{F}_e^*(e^{i\omega},\theta_0)d\omega, \end{aligned} \tag{49}$$

where $\mathcal{F}_u(\theta_0) = \frac{\widetilde{A}(q,\theta_0)}{\widetilde{C}(q,\theta_0)}\nabla\left(\frac{\widetilde{B}(q,\theta)}{\widetilde{A}(q,\theta)}\right)\Big|_{\theta=\theta_0}$, $\mathcal{F}_e(\theta_0) = \frac{\widetilde{A}(q,\theta_0)}{\widetilde{C}(q,\theta_0)}\nabla\left(\frac{\widetilde{C}(q,\theta)}{\widetilde{A}(q,\theta)}\right)\Big|_{\theta=\theta_0}$. With a finite-dimensional parameterization approach as shown in [22], the spectrum of input signal is expressed as

$$\Phi_u(\omega) = \Psi(e^{i\omega}) + \Psi^*(e^{i\omega}), \tag{50}$$

where $\Psi(e^{i\omega}) = \sum_{k=0}^{M-1} c_k e^{-i\omega k}$, M is the number of parameters and c_k, $k = 1,\ldots,M$, are real decision variables in the input design. A necessary condition for the parameterized function $\Phi_u(\omega)$ to be a spectrum is that $\Phi_u(\omega) \geq 0,\ \forall\omega$. This infinite dimensional constraint can be converted to a finite dimensional and convex form by the KYP lemma, see [22]. In this work we intend to find an optimal input design problem to minimize the covariance of the transfer function estimate, i.e.,

$$\begin{aligned} \min_{c_k, k=1,\ldots,M} \quad & \gamma \\ \text{s.t.} \quad & \Phi_u(\omega) \geq 0, \\ & |F_u(e^{i\omega})|^2 cov[\hat{G}(e^{i\omega})/G_0(e^{i\omega})] \leq \gamma,\ \forall\omega, \\ & \frac{1}{2\pi}\int_{-\pi}^{\pi}|W_u(e^{i\omega})|^2\Phi_u(\omega)d\omega \leq \alpha, \\ & \frac{1}{2\pi}\int_{-\pi}^{\pi}|W_y(e^{i\omega})|^2\Phi_y(\omega)d\omega \leq \beta, \end{aligned} \tag{51}$$

where $F_u(e^{i\omega})$ is a frequency-wise weighting function and $cov[\hat{G}(e^{i\omega})]$ is the covariance of the transfer function estimate under the proposed noncausal model identification method. The parameters α and β specify upper bounds on the admissible input and output signal powers, respectively. With the above parameterization of the input spectrum in (50), the preceding input design can be recast as a convex optimization problem with linear matrix inequality (LMI) constraints. Specifically, the parameter covariance P_θ^{-1} can be expressed as

$$\begin{aligned} P_\theta^{-1} &= \sum_{k=0}^{M-1} c_k B_k^P(\theta_0) + R_0(\theta_0), \quad \text{where} \\ B_k^P(\theta_0) &= \frac{1}{2\pi\sigma^2}\int_{-\pi}^{\pi}\left[\mathcal{F}_u(e^{i\omega},\theta_0)(e^{i\omega k} + e^{-i\omega k})\right. \\ & \qquad \left. \cdot\mathcal{F}_u^*(e^{i\omega},\theta_0)\right]d\omega, \\ R_0(\theta_0) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathcal{F}_e(e^{i\omega},\theta_0)\mathcal{F}_e^*(e^{i\omega},\theta_0)d\omega. \end{aligned}$$

The positivity constraint on the power spectrum in (51) can be transformed into an LMI using the KYP lemma:

$$\begin{bmatrix} Q - A^TQA & -A^TQB \\ -B^TQA & -B^TQB \end{bmatrix} + \begin{bmatrix} 0 & C^T \\ C & D + D^T \end{bmatrix} \geq 0,$$

where $Q = Q^T$, and $\{A, B, C, D\}$ is a controllable state-space realization of $\Psi(e^{i\omega}) = \sum_{k=0}^{M-1} c_k e^{-i\omega k}$. Analogously, we can also formulate the other three conditions as LMIs in terms of the parameters c_k. The above optimization-based input design can be extended to cases with complicated quality constraints and objective functions. In this paper we only make use of the above formulations for an illustrative purpose.

6 Numerical Example

6.1 Identification

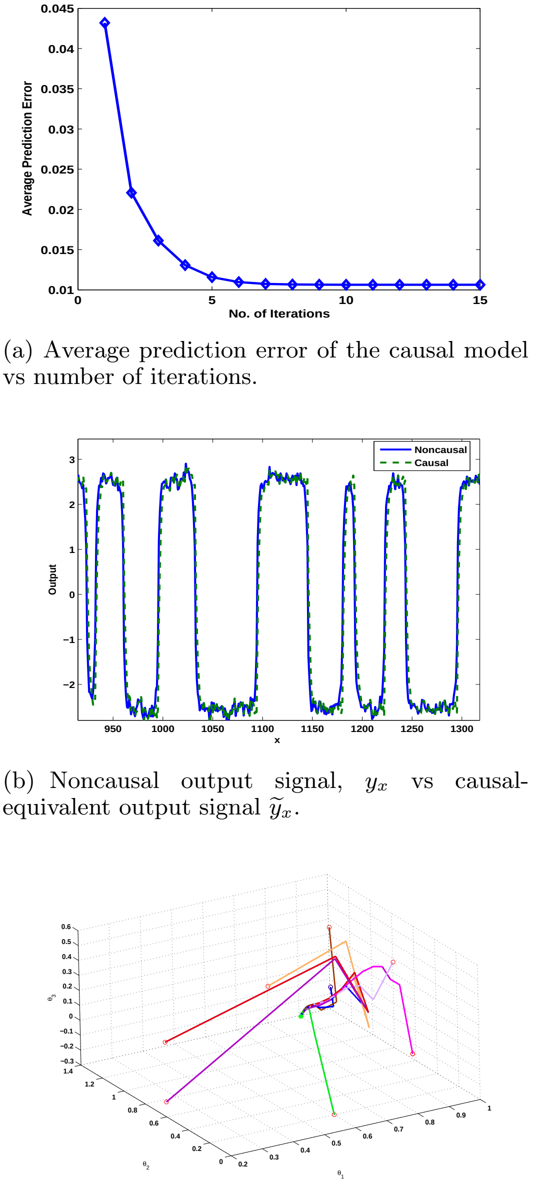
Example 4 Consider the following noncausal model

$$(1-0.5q^{-1})(1-0.5q)y_x = (1-0.2q^{-1})(1-0.2q)u_x \\ +(1-0.4q^{-1})(1-0.4q)e_x, \qquad (52)$$

where e_x is Gaussian white noise with variance 0.01. The causal-equivalent model is given by

$$(1-\theta_1 q^{-1})^2\widetilde{y}_x = (1-\theta_2 q^{-1})^2 u_x + (1-\theta_3 q^{-1})^2 \widetilde{e}_x, \qquad (53)$$

where $\widetilde{y}_x = \frac{1-\theta_2 q^{-1}}{1-\theta_1 q^{-1}}\frac{1-\theta_1 q}{1-\theta_2 q}y_x$. We use the function `idinput` in Matlab to generate a sequence of random binary signal as u_x. To examine the convergence of our iterative identification algorithm, we generate a large data set with size $N = 5000$ to minimize variance errors. Starting with an initial guess of parameters θ_1 and θ_2, we obtain an estimate of $\{\widetilde{y}_x\}$. Note that we do not need an initial guess for the noise model. This is an extra advantage of our approach since in practice engineers have a good idea of typical process model parameters while it is not easy to acquire any *a priori* information regarding the true noise model. We then minimize the prediction errors of the causal-equivalent model and obtain a new estimate for the model parameters. With the updated model parameters we re-estimate $\{\widetilde{y}_x\}$ and repeat the identification exercise. These iterations are continued until the parameter estimate converges. In this example we choose $[0.8; 0.4; 0.6]$ as an initial guess of the parameter vector. The average prediction error for the estimated causal model decreases monotonically as iteration proceeds, as shown in Fig. 5(a). The estimated parameters using the proposed method are shown in the first row of Table 1 with an estimated variance of $\widetilde{e}_x$ of 0.0106. In principle, an increase in the data size can reduce the effect of noise on the actual parameter estimates [1, p. 295]. The original noncausal signal and its filtered causal-equivalent signal are shown in Fig. 5(b). The phase shift due to all-pass filter is visible.



(a) Average prediction error of the causal model vs number of iterations.

(b) Noncausal output signal, y_x vs causal-equivalent output signal $\widetilde{y}_x$.

(c) The parameter trajectories during the iterations for different initial guesses. Initial guess represented by 'o' and the converged estimate by '*'.

Fig. 5. Simulation results of Example 4

In order to test the performance of the algorithm, we tried a number of initial guesses. A plot showing the trajectories of the parameter estimates with the iterations is shown in Fig. 5(c). For all initial guesses in the neighborhood of the true parameters, the estimates converged within a few iterations, typically about $15-30$ for this example. Table 1 also demonstrates the identification results using noncausal MLE. Notice that this approach is similar to the one used in [17] for unstable causal models. One can see this approach can also yield accurate parameter estimates. However, the user has to provide the gradient vector of the predictor of the underlying noncausal model. Our method can avoid this issue since it can be implemented using currently available tools for causal system identification. To further show the advantage of the proposed noncausal identification method relative to that in [16], we first convert (52) into a causal unstable model with a new noise $\widetilde{e}_x$ sequence and perform the prediction error identification based on the gradient shown in [16]. Note that with this method, the variance of new noise increases to $\sigma^2_{\widetilde{e}} = \sigma^2_e/a^2_{n_a}$ (a_{n_a} is the last coefficient of polynomial $A(q)$), thus reducing the SNR in the data. Table 1 shows the identification results based on our method and that in [16] for (52). One can see that our method successfully identifies the true parameters, whereas the method in [16] yields an imprecise estimate of θ_1. This is due to the reduced SNR explained above. Increasing the level of excitation signal would improve the parameter estimates, as shown in the last row of Table 1. In practice, for stable models, a_{n_a} is usually small (much smaller that the value in this example) and thus we

Table 1
Estimated parameters with our approach, the method based on [16] and the noncausal MLE similar to [17]

	θ_1	θ_2	θ_3
True parameters	0.5	0.2	0.4
Our method with $\sigma_u^2 = 1$	0.4989	0.1996	0.4003
Method based on [17], $\sigma_u^2 = 1$	0.4825	0.1829	0.4027
Method based on [16], $\sigma_u^2 = 1$	0.6241	0.2165	0.3979
Method based on [16], $\sigma_u^2 = 4$	0.5128	0.2011	0.3980

expect our method to perform better than that from [16] in identifying practical noncausal models.

To further compare the proposed method with the noncausal MLE method based on [17], we perform 100 Monte-Carlo simulations with $N = 200$. The resulting estimates have mean and standard deviation values of $(0.4993 \pm 0.0083), (0.1989 \pm 0.0106), (0.4005 \pm 0.0231)$, respectively. The corresponding results from the noncausal MLE method are $(0.5040 \pm 0.0099), (0.2062 \pm 0.0129), (0.4259 \pm 0.0307)$. Evidently both methods give unbiased estimates, although the parameter covariance differs. A possible reason for the slightly larger uncertainty of parameter estimates associated with the noncausal MLE method is the transient effect at the beginning and at the end of the predicted output sequence, especially when the sample size is small, as in this test. Note that for the noncausal MLE method the noncausual filtering occurs in the computation of prediction errors and gradients at each step. This issue is less severe for our proposed method as our algorithm relies mostly on causal identification. Moreover, the noncausal MLE method involves a complex noncausal gradient computation, which is another disadvantage from the implementation perspective.

Example 5 We now consider a realistic CD model in a paper machine. This typical model is taken from [24] [5],

$$(1 - 0.3465q^{-1} + 0.3025q^{-2})(1 - 0.3465q + 0.3025q^2)y_x \\ = u_x + (1 - 0.3q^{-1})(1 - 0.3q)e_x \qquad (54)$$

The estimated causal-equivalent model is of the form

$$(1 - \theta_1 q^{-1} + \theta_2 q^{-2})^2 y_x = u_x + (1 - \theta_3 q^{-1})^2 e_x.$$

The response of this model is similar to the one shown in Fig. 1. As in the previous example, we have used a random binary signal with unit variance as u_x. Typical paper machines have significant noise (and/or disturbances) in the measurements. Hence to reflect reality, we have used a noise covariance of 0.3. The number of data points available from a paper machine depends on the number of scanner measurements across the width of the paper. We have also limited the data set to $N = 1000$. Clearly, with a smaller data set and higher noise covariance, the uncertainty of the estimated model increases. A 3D plot of the estimated parameters for 200 Monte Carlo simulations is shown in Fig. 6(a). The Nyquist and Bode plots of the estimated models from the Monte Carlo simulations are shown in Figures 6(b) and 6(c), respectively. As shown in the figure, the variance in the frequency domain is large around the corner frequency [6]. This is typically the case with the data from paper machines. Hence, a good input designed to minimize this uncertainty and to find the "best" nominal model is essential for CD model identification of paper machines.

6.2 Input Design

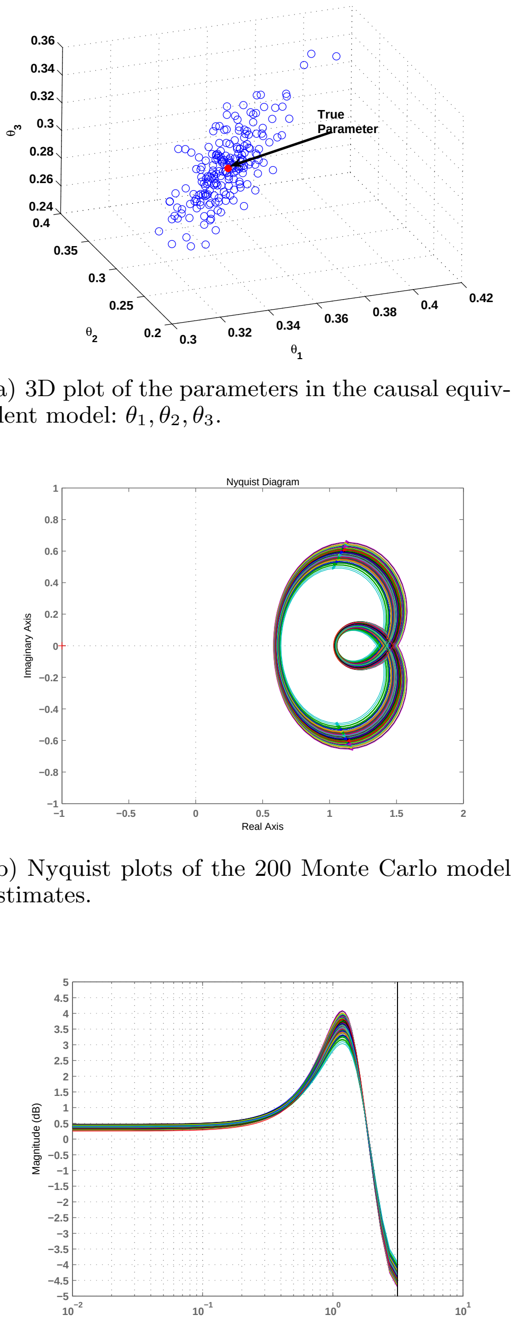
Example 6 In this example, we design an optimal input for the model in (54) based on (51). Let us suppose that we would like to minimize the covariance of the transfer function estimates in identifying (54), particularly the covariance around the corner frequency region. To this end, we use the true process model as a weighting function $F_u(e^{i\omega})$ in (51). In order to make a fair comparison between the optimal input and the random binary input in Example 5, we do not consider the output power constraint here and choose $W_u = 1$ and $\alpha = 1$ in (51) such that these two inputs have the same variance. The constraint on the positivity of the input power spectrum is formulated as an LMI. With the above setup, the spectrum of the optimal input is shown in Fig. 7(a) (dashed-dotted line). All the other simulation conditions are similar to those in Example 5. In Example 5, the variance in the low frequency range is small and that around the corner frequency is large. However, in this example, the Bode plot of models estimated from Monte Carlo simulations shows higher variance in the low frequency region and lower variance around the corner frequency as shown in Fig. 7(b). Thus we can expect better control performance if the controller is designed based on the process model estimate under the optimal input signal.

7 Conclusion

In general, identification of noncausal processes with Gaussian noise is rather difficult. In this work, we present a maximum likelihood approach to identify symmetric noncausal models. It is shown that a symmetric noncausal model admits a causal-equivalent model in the sense of equivalent output spectrum. We further show that the log-likelihood function as well as its maximizer of a noncausal model converges asymptotically to those of its causal-equivalent model, respectively, given that the likelihood functions have a unique maximum. We further propose an iterative identification approach to identify the causal-equivalent model and such algorithm is proved to be convergent. In addition, we show that the parameter covariance matrices of the causal and noncausal models converge to the same value asymptotically, which lays a foundation for the input design of using the parameter covariance of the causal-equivalent model. Several examples are presented to demonstrate and

[5] Please note that there is no noise model in [24]. We have added a noise model to make it more realistic.

[6] Please note that the y-axis is in decibels

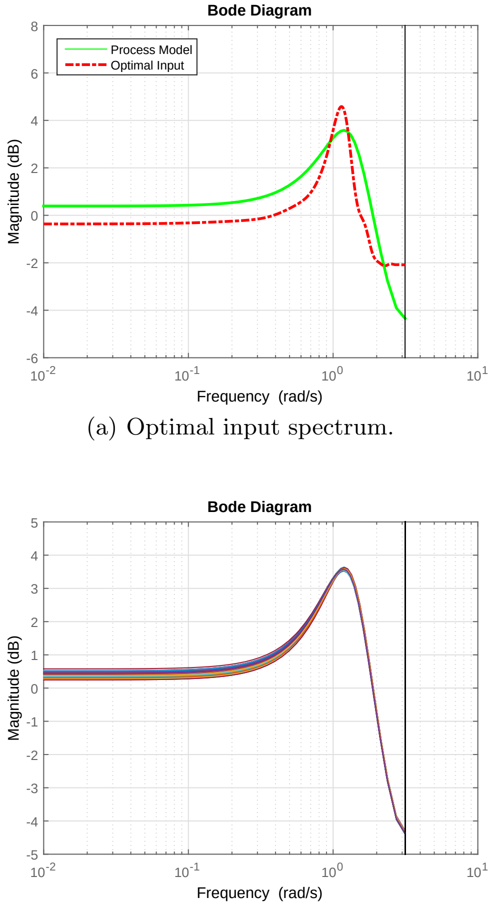


(a) 3D plot of the parameters in the causal equivalent model: $\theta_1, \theta_2, \theta_3$.

(b) Nyquist plots of the 200 Monte Carlo model estimates.

(c) Bode plots of the 200 Monte Carlo model estimates.

Fig. 6. Simulation results of Example 5



(a) Optimal input spectrum.

(b) Bode plots of the 200 Monte Carlo model estimates.

Fig. 7. Simulation results of Example 6

verify the related results. However, similar to Steiglitz-McBride method, due to the prefiltering of the data in the iteration identification algorithm proposed in Section 4, consistency and asymptotic efficiency of parameter estimates may not be ensured when implementing this algorithm.

A summary of the advantages and disadvantages, both theoretical properties and implementation issues, of the noncausal identification techniques mentioned in this article are as follows. For the noncausal MLE in [17], its merits lie in that it yields unbiased estimate and no multiple optimizations are needed for estimating parameters. However, it requires users to provide (noncausal) gradients, and has larger parameter uncertainties when sample size is small than our method due to the transient effects from the noncausal filtering in computing predictor errors and gradients. For the method in [16], it does not involve iterative identifications. However, a new noise sequence having larger variance is produced and this reduces the SNR, leading to worse parameter estimates than our method. It also requires users to provide gradients. For the method in Section 2.3, it is straightforward to understand. However, it is only suitable for open-loop identification or deterministic input, and computation of the gradient involves extensive noncausal filtering. In terms of our method, it can be easily implemented with current system identification toolboxes, and users do not need to provide gradient information. Moreover, it can be easily extended to closed-loop identification and is independent of SNR. However, it is computationally more expensive due to the iterative identification algorithm.

Acknowledgements

The authors greatly appreciate the insightful feedback and valuable comments from the three reviewers. In particular, the open-loop noncausal filtering approach presented in Section 2.3 is proposed by an anonymous reviewer. The analytical Example 3 is also suggested by the same reviewer. We are grateful to this particular reviewer and his/her comments significantly improved the quality of the paper. We also sincerely thank the Editor and the Associated Editor for their help and guidance throughout the entire review process.

Appendix A Proof of Proposition 2

We generalize the approach in [9]. To fix notation, we rearrange (4) as

$$\begin{aligned} e_x &= \frac{A(q,\theta)}{C(q,\theta)}\left[y_x - \frac{B(q,\theta)}{A(q,\theta)}u_x\right] \\ &= \sum_{k=-\infty}^{\infty} g_k(\theta)u_{x-k} + \sum_{k=-\infty}^{\infty} h_k(\theta)y_{x-k} \\ &= G(q,\theta)u_x + H(q,\theta)y_x, \end{aligned}$$

using the impulse response coefficients associated with $G = -B/C$ and $H = A/C$ in the usual way: $G(q,\theta) = \sum_{k=-\infty}^{\infty} g_k(\theta)q^{-k}$, $H(q,\theta) = \sum_{k=-\infty}^{\infty} h_k(\theta)q^{-k}$. Now suppose that we have the input-output data $\{y_1,\cdots,y_N\}$ and $\{u_1,\cdots,u_N\}$. For each x in $\{1,\ldots,N\}$, we define $G_x(q,\theta) = \sum_{k=x-N}^{x-1} g_k(\theta)q^{-k}$, $\widetilde{G}_x = G - G_x, H_x(q,\theta) = \sum_{k=x-N}^{x-1} g_k(\theta)q^{-k}$, $\widetilde{H}_x = H - H_x$. This notation splits the noncausal transfer functions into two parts: $G_x(q,\theta)$ and $H_x(q,\theta)$ correspond to known values, while $\widetilde{G}_x(q,\theta)$ and $\widetilde{H}_x(q,\theta)$ correspond to unknown past and future values. Our system takes the compact form

$$e_x = G_x(q,\theta)u_x + H_x(q,\theta)y_x + \widetilde{G}_x(q,\theta)u_x + \widetilde{H}_x(q,\theta)y_x.$$

Now the average log-likelihood function of the data given inputs is

$$\begin{aligned} \mathcal{L}^N &= \frac{1}{N}\log f_y(y_1,\cdots,y_N|u_1,\cdots,u_N,\theta) \\ &= \frac{1}{N}\log f_e(e_1,\cdots,e_N|u_1,\cdots,u_N,\theta) \\ &= \frac{1}{N}\sum_{x=1}^{N}\log f_e(e_x|u_1,\cdots,u_N,\theta), \end{aligned} \tag{55}$$

where $f_y(\cdot)$ and $f_e(\cdot)$ denote the density functions of y_x and e_x respectively [7], and the second equality follows [8] from Lemma 5.1 in Ljung [1]. The values $e_1,\cdots,e_N$ involved here are not all known, so we cannot maximize $\mathcal{L}^N$ with respect to θ to find a maximum likelihood estimate. Instead, we introduce the following approximation for e_x:

$$\widetilde{e}_x = \sum_{k=x-N}^{k=x-1} g_k(\theta)u_{x-k} + \sum_{k=x-N}^{k=x-1} h_k(\theta)y_{x-k}. \tag{56}$$

Then we maximize the approximate log-likelihood function

$$\widetilde{\mathcal{L}}^N := \frac{1}{N}\sum_{x=1}^{N}\log f_{\widetilde{e}}(\widetilde{e}_x|u_1,\cdots,u_N,\theta).$$

To complete the proof, we show that, with probability 1, $|e_x - \widetilde{e}_x| \to 0$ uniformly. This ensures that $|\widetilde{\mathcal{L}}^N - \mathcal{L}^N| \to 0$ with probability one as the data length increases.

Let $\Delta e_x = |e_x - \widetilde{e}_x|$ denote the difference of interest. With the notation above,

$$\begin{aligned} \Delta e_x &= \left|\widetilde{G}_x(q,\theta)u_x + \widetilde{H}_x(q,\theta)y_x\right| \\ &= \left|\sum_{k\notin[x-1,x-N]} g_k u_{x-k} + \sum_{k\notin[x-1,x-N]} h_k y_{x-k}\right|. \end{aligned}$$

Let us now define the following sum,

$$S_x(\varepsilon) = \sum_{k=1}^{x}\mathbb{P}(\Delta e_k > \varepsilon). \tag{57}$$

Note that $x \in [1,N]$ and therefore $x \le N$. Now as $x \to \infty$ and $N \to \infty$ simultaneously, the interval $[x-1,x-N]$ gets bigger, and g_k and h_k such that $k \notin [x-1,x-N]$ get smaller (since g_k and h_k are absolutely convergent). Considering that u_x are deterministic and bounded, and e_x are Gaussian, it follows that $\limsup_{x\to\infty} \ E\left[(\Delta e_x)^2\right] = 0$. Now using Theorem 3.39 in Rudin [37] we have

$$\sum_{x=1}^{\infty} E\left[(\Delta e_x)^2\right] < \infty. \tag{58}$$

Chebyshev's inequality [1, p. 542] implies

$$\mathbb{P}(\Delta e_x > \varepsilon) \le \frac{1}{\varepsilon^2}E\left[(\Delta e_x)^2\right]. \tag{59}$$

From (57),(58) and (59) it follows that $\lim_{x\to\infty} S_x(\varepsilon) < \infty$. Using the Borel-Cantelli lemma [1, p. 542], we conclude that as $x \to \infty$, $\Delta e_x \overset{w.p.1}{\to} 0$. Therefore, from (57), it follows that $\sup_{\theta\in\Omega}|e_x - \widetilde{e}_x| \overset{w.p.1}{\to} 0$. Since e_x is Gaussian, $f_e(.)$ is a smooth function and therefore we conclude that $\sup_{\theta\in\Omega}|\mathcal{L}^N - \widetilde{\mathcal{L}}^N| \overset{w.p.1}{\to} 0$. Noting that

[7] With a slight abuse of notation, we use the same name for the density function of the sequence $\{e_1,\cdots,e_N\}$ and the individual random variables e_x.

[8] Although the cited Lemma is derived for causal systems, a similar proof holds for noncausal systems.

$f_e(e_x) = 1/(\sigma\sqrt{2\pi})e^{-e_x^2/2\sigma^2}$, we conclude that

$$\sup_{\theta\in\Omega}\left|V^N(\theta) - \sigma^2\mathcal{L}^N(\theta)\right| \overset{w.p.1}{\to} \log\sigma\sqrt{2\pi}. \tag{60}$$

The analogous result for causal systems is well-known and hence the proof is not repeated here.

Appendix B Proof of Proposition 4

Proceeding in general deterministic notation, we consider how the set of maximizing inputs can evolve for a uniformly convergent sequence of functions. The ingredients are a compact convex set Ω in $\mathbb{R}^n$, functions $f^N, f\colon \Omega \to \mathbb{R}$ ($N \in \mathbb{N}$), and sets

$$D^N = \arg\max_{\Omega} f^N = \left\{\theta \in \Omega : \ f^N(\theta) \geq f^N(\theta') \ \forall\theta' \in \Omega\right\},$$
$$D = \arg\max_{\Omega} f = \left\{\theta \in \Omega : \ f(\theta) \geq f(\theta') \ \forall\theta' \in \Omega\right\}.$$

For $r \geq 0$, we use the notation $\mathbb{B} = \left\{\theta \in \mathbb{R}^n : \ \left|\theta\right| \leq 1\right\}$, $D + r\mathbb{B} = \left\{d + ru : \ d \in D, \ u \in \mathbb{B}\right\}$.

Lemma 4 *Assume that each of the functions f^N, f is continuously differentiable on the compact convex set Ω. If*

$$\sup_{\theta\in\Omega}\left|f^N(\theta) - f(\theta)\right| \to 0 \textit{ as } N \to \infty, \textit{ and}$$
$$\sup_{\theta\in\Omega}\left|\nabla f^N(\theta) - \nabla f(\theta)\right| \to 0 \textit{ as } N \to \infty,$$

and r_N is any sequence with nonnegative values obeying $r_N \to 0$ as $N \to \infty$, then

$$\limsup_{N\to\infty}\left[\left(D^N + r_N\mathbb{B}\right) \cap \Omega\right] \subseteq D.$$

Proof. Let M_0 denote the maximum value of $\left|\nabla f\right|$ over Ω. This is a finite real number because Ω is a compact set and $\left|\nabla f(\theta)\right|$ depends continuously on $\theta \in \Omega$. Then let $M = M_0 + 1$. For all N sufficiently large, our hypotheses imply that $\sup_\Omega \left|\nabla f^N\right| \leq M$, and consequently that $\left|f^N(\theta_2) - f^N(\theta_1)\right| \leq M\left|\theta_2 - \theta_1\right|, \forall\theta_1, \theta_2 \in \Omega$. Now let any sequence of indices $N_k \uparrow \infty$ be given, and consider an arbitrary convergent sequence $w_{N_k} \in \Omega \cap \left(D^{N_k} + r_{N_k}\mathbb{B}\right)$. Let $w = \lim_k w_{N_k}$. By definition, $w_{N_k} \in \Omega$ and $w_{N_k} = \widehat{\theta}_{N_k} + r_{N_k}u_{N_k}$ for some $\widehat{\theta}_{N_k} \in D^{N_k}$ and $u_{N_k} \in \mathbb{B}$. Now the compactness of Ω guarantees that along a suitable subsequence (we do not relabel), $\widehat{\theta}_{N_k}$ converges to some $\widehat{\theta}$ in Ω. Choosing any θ in Ω, we fix k and consider the following quantity: $f(w_{N_k}) - f^{N_k}(\theta) = \left(f(w_{N_k}) - f^{N_k}(\widehat{\theta}_{N_k})\right) + \left(f^{N_k}(\widehat{\theta}_{N_k}) - f^{N_k}(\theta)\right)$. By definition of $\widehat{\theta}_{N_k}$, the rightmost difference shown here is nonnegative. Therefore

$$\begin{aligned} f(w_{N_k}) - f^{N_k}(\theta) &\geq f(w_{N_k}) - f^{N_k}(\widehat{\theta}_{N_k}) \\ &= \left(f(w_{N_k}) - f^{N_k}(w_{N_k})\right) + \left(f^{N_k}(w_{N_k}) - f^{N_k}(\widehat{\theta}_{N_k})\right). \end{aligned}$$

As $k \to \infty$ (through the subsequence named earlier), we have $\left|f(w_{N_k}) - f^{N_k}(w_{N_k})\right| \to 0$ by the hypothesis of uniform convergence of the given sequence of functions. Further, as shown above,

$$\left|f^{N_k}(w_{N_k}) - f^{N_k}(\widehat{\theta}_{N_k})\right| \leq M\left|w_{N_k} - \widehat{\theta}_{N_k}\right| \leq Mr_{N_k} \to 0.$$

Thus the inequality above supports the limiting statement that $f(w) - f(\theta) \geq 0$. But $\theta \in \Omega$ is arbitrary, so this shows that $w \in D$, as required. ■

Appendix C Proof of Lemma 2

The results in (42) and (43) follow directly from Theorem 7.17 in Rudin [37] by writing it for multivariable functions. In order to prove (44), let us start with the triangle inequality

$$\begin{aligned} &\left\|\nabla\mathcal{L}^N(\theta) - \nabla\mathcal{L}_c^N(\theta)\right\| \leq \\ &\quad \left\|\nabla\mathcal{L}^N(\theta) - \nabla\sigma^2(\theta)\right\| + \left\|\nabla\mathcal{L}_c^N(\theta) - \nabla\sigma^2(\theta)\right\|. \end{aligned} \tag{61}$$

Therefore,

$$\begin{aligned} &\sup_{\theta\in\Omega}\left\|\nabla\mathcal{L}^N(\theta) - \nabla\mathcal{L}_c^N(\theta)\right\| \leq \\ &\sup_{\theta\in\Omega}\left\|\nabla\mathcal{L}^N(\theta) - \nabla\sigma^2(\theta)\right\| + \sup_{\theta\in\Omega}\left\|\nabla\mathcal{L}_c^N(\theta) - \nabla\sigma^2(\theta)\right\|. \end{aligned} \tag{62}$$

Using (62) in combination with (42) and (43) in the limit $N \to \infty$, we have

$$\sup_{\theta\in\Omega}\left\|\nabla\mathcal{L}^N(\theta) - \nabla\mathcal{L}_c^N(\theta)\right\| \overset{w.p.1}{\to} 0. \tag{63}$$

Appendix D Proof of Lemma 3

Combining the assumption of uniform proximity with condition (a), we deduce that there exists $R_1 \geq R$ for which $\max\left\{\left|F^N(\theta)\right|, \left|G^N(\theta)\right|\right\} \leq R_1, \forall N \in \mathbb{N}, \ \theta \in \Omega$. The continuous function ϕ in assumption (b) must attain its maximum on the closed ball of radius R_1 centered at the origin of $\mathbb{R}^n$, so there exists a constant K such that $\sup_N\left|\nabla m^N(p)\right| \leq K$, whenever $\left|p\right| \leq R_1$. Consequently, for each $\theta \in \Omega$ and $N \in \mathbb{N}$, we have

$$\begin{aligned} \left|m^N(F^N(\theta)) - m^N(G^N(\theta))\right| &\leq K\left|F^N(\theta) - G^N(\theta)\right| \\ &\leq K\sup_{\theta\in\Omega}\left|F^N(\theta) - G^N(\theta)\right|. \end{aligned}$$

The right side converges to 0 by hypothesis, so the result follows. ■

References

[1] L. Ljung. *System Identification: Theory for the User.* Prentice-Hall, New Jersey, 1999.

[2] T. Söderström and P. Stoica. *System Identification.* Prentice-Hall, Inc., New Jersey, 1988.

[3] Rik Pintelon and Johan Schoukens. *System Identification: A Frequency Domain Approach.* John Wiley & Sons, 2012.

[4] Rongning Wu and Richard A Davis. Least absolute deviation estimation for general autoregressive moving average time-series models. *Journal of Time Series Analysis*, 31(2):98–112, 2010.

[5] J.K. Tugnait. Modeling and identification of symmetric noncausal impulse responces. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(5):1171–1181, 1986.

[6] B. Andrews, M. Calder, and R.A. Davis. Maximum likelihood estimation for α-stable autoregressive processes. *The Annals of Statistics*, 37(4):1946–1982, 2009.

[7] P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting.* Springer-Verlag, New York, 2002.

[8] R. Davis and L. Song. Noncausal vector AR processes with application to economic time series. *Columbia University: Working Paper*, pages 1–34, 2012.

[9] O. Shalvi and E. Weinstein. Maximum likelihood and lower bounds in system identification with non-Gaussian inputs. *IEEE Transactions on Information Theory*, 40(2):328–339, 1994.

[10] M. Bakrim and D. Aboutajdine. Higher-order statistics based blind estimation of non-gaussian bidimensional moving average models. *Signal Processing*, 86(10):3031–3042, 2006.

[11] Markku Lanne and Jani Luoto. Noncausal bayesian vector autoregression. *Journal of Applied Econometrics*, 31(7):1392–1406, 2016.

[12] Christian Gourieroux and Joann Jasiak. Filtering, prediction and simulation methods for noncausal processes. *Journal of Time Series Analysis*, 37(4):405–430, 2016.

[13] Urban Forssell and Lennart Ljung. A projection method for closed-loop identification. *IEEE Transactions on Automatic Control*, 45(11):2101–2106, 2000.

[14] Khaled F Aljanaideh and Dennis S Bernstein. Closed-loop identification of unstable systems using noncausal FIR models. *International Journal of Control*, pages 1–18, 2016.

[15] Maciej Niedźwiecki and Szymon Gackowski. New approach to noncausal identification of nonstationary stochastic FIR systems subject to both smooth and abrupt parameter changes. *IEEE Transactions on Automatic Control*, 58(7):1847–1853, 2013.

[16] Urban Forssell and Lennart Ljung. Identification of unstable systems using output error and Box-Jenkins model structures. *IEEE Transactions on Automatic Control*, 45(1):137–141, 2000.

[17] Håkan Hjalmarsson and Urban Forssell. *Maximum likelihood estimation of models with unstable dynamics and non-minimum phase noise zeros.* Linköping University Electronic Press, 1998.

[18] A.P. Featherstone, J. VanAntwerp, and R.D. Braatz. *Identification and Control of Sheet and Film Processes.* Springer Science & Business Media, 2000.

[19] D.M. Gorinevsky and C. Gheorghe. Identification tool for cross-directional processes. *IEEE Transactions on Control Systems Technology*, 11(5):629–640, 2003.

[20] L.G. Bergh and J.F. MacGregor. Spatial control of sheet and film forming processes. *The Canadian Journal of Chemical Engineering*, 65(1):148–155, 1987.

[21] M.E. Ammar and G.A. Dumont. Automatic tuning of robust constrained cross-direction controllers. *International Journal of Adaptive Control and Signal Processing*, 30(1):1550–1567, 2016.

[22] H. Jansson and H. Hjalmarsson. Input design via LMIs admitting frequency-wise model specifications in confidence regions. *IEEE Transactions on Automatic Control*, 50(10):1534–1549, 2005.

[23] R Bhushan Gopaluni, Philip D Loewen, Mohammed Ammar, Guy A Dumont, and Michael S Davies. Identification of symmetric noncausal processes: Cross-directional response modelling of paper machines. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 6744–6749, 2006.

[24] D. Gorinevsky and G. Stein. Structured uncertainty analysis of spatially distributed paper machine process control. In *Proceedings of the 2001 American Control Conference*, volume 3, pages 2225–2230, 2001.

[25] J.J. Hsue and A.E. Yagle. Similarities and differences between one-sided and two-sided linear prediction. *IEEE Transactions on Signal Processing,*, 43(1):345–349, 1995.

[26] Abdellah Kacha, Francis Grenez, and Khier Benmahammed. Time–frequency analysis and instantaneous frequency estimation using two-sided linear prediction. *Signal Processing*, 85(3):491–503, 2005.

[27] Sidney I Resnick. *A probability path.* Springer Science & Business Media, 2013.

[28] L. Ljung. Asymptotic variance expressions for identified black-box transfer function models. *IEEE Transactions on Automatic Control*, 30(9):834–844, 1985.

[29] Y. Zhu. Multivariable process identification for MPC: the asymptotic method and its applications. *Journal of Process Control*, 8(2):101–115, 1998.

[30] Håkan Hjalmarsson and Jonas Martensson. A geometric approach to variance analysis in system identification. *IEEE Transactions on Automatic Control*, 56(5):983–997, 2011.

[31] S. Garatti, M.C. Campi, and S. Bittanti. Assessing the quality of identified models through the asymptotic theory – when is the result reliable? *Automatica*, 40(8):1319–1332, 2004.

[32] Marco C Campi and Erik Weyer. Non-asymptotic confidence sets for the parameters of linear transfer functions. *IEEE Transactions on Automatic Control*, 55(12):2708–2720, 2010.

[33] M. Gevers, X. Bombois, R. Hildebrand, and G. Solari. Optimal experiment design for open and closed-loop system identification. *Communications in Information and Systems*, 11(3):197–224, 2011.

[34] Patricio E Valenzuela, Cristian R Rojas, and Håkan Hjalmarsson. A graph theoretical approach to input design for identification of nonlinear dynamical models. *Automatica*, 51:233–242, 2015.

[35] Roland Hildebrand, Michel Gevers, and Gabriel Elías Solari. Closed-loop optimal experiment design: solution via moment extension. *IEEE Transactions on Automatic Control*, 60(7):1731–1744, 2015.

[36] Christian A Larsson, Afrooz Ebadat, Cristian R Rojas, Xavier Bombois, and Håkan Hjalmarsson. An application-oriented approach to dual control with excitation for closed-loop identification. *European Journal of Control*, 29:1–16, 2016.

[37] W. Rudin. *Principles of Mathematical Analysis.* McGraw-Hill Publishing Co., 1976.