Sankhyā A : The Indian Journal of Statistics
https://doi.org/10.1007/s13171-023-00326-6

Quadratic Prediction of Time Series via Auto-Cumulants

Tucker S. McElroy
U.S. Census Bureau, Washington, DC, USA
Dhrubajyoti Ghosh and Soumendra Lahiri
Washington University in St. Louis, St Louis, USA

Abstract

Nonlinear prediction of time series can offer potential accuracy gains over linear methods when the process is nonlinear. As there are numerous examples of nonlinearity in time series data (e.g., finance, macroeconomics, image, and speech processing), there seems to be merit in developing a general theory and methodology. We explore the class of quadratic predictors, which directly generalize linear predictors, and show that they can be computed in terms of the second, third, and fourth auto-cumulant functions when the time series is stationary. The new formulas for quadratic predictors generalize the normal equations for linear prediction of stationary time series, and hence we obtain quadratic generalizations of the Yule-Walker equations; we explicitly quantify the prediction gains in quadratic over linear methods. We say a stochastic process is SECOND ORDER FORECASTABLE if quadratic prediction provides an advantage over linear prediction. One of the key results of the paper provides a characterization of second order forecastable processes in terms of the spectral and bi-spectral densities. We verify these conditions for some popular nonlinear time series models.

AMS (2000) subject classification. 62M10; 62M20.
Keywords and phrases. Bi-spectral density, Nonlinear prediction, Nonlinear processes, Quadratic prediction, Polyspectra.

1 Introduction

The problem of prediction can often be parsed as an attempt to find a good "estimator" of a target random variable Y given an available data random vector $\underline{X}$. Typically a joint distribution is posited, and the broader problem of prediction involves determining the conditional distribution $Y|\underline{X}$.

Dhrubajyoti Ghosh and Soumendra Lahiri both contributed equally to this work
Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13171-023-00326-6) contains supplementary material, which is available to authorized users.

Published online: 08 September 2023

The mean (when it exists) of this conditional distribution is the conditional expectation $\mathbb{E}[Y|\underline{X}]$, and is known to minimize the mean square error (MSE) between Y and all functions of $\underline{X}$. Because many prediction problems can be put into this context, much effort has been exerted in solving this general problem. When the joint distribution is Gaussian, the conditional expectation is a linear function of $\underline{X}$. This linear function is completely computable in terms of the first and second moments of the joint vector $(Y, \underline{X})$, as discussed in Brockwell and David (2016)(Chapter 2).

Even when the joint vector is non-Gaussian, a practitioner still might use the linear solution – knowing that this solution has the minimal MSE among all linear estimators – because it is simple to compute. Nevertheless, there can be a substantial predictive loss when non-Gaussian features are present in the data, such as asymmetry and excess kurtosis (cf. Maravall (1983), Brockett et al. (1988)). Nonlinearity in financial data has been documented in Hinich and Patterson (1985) and Abhyankar et al. (1997); Ramsey and Rothman (1996) discuss time irreversibility (i.e., nonlinearity) in macroeconomic data, whereas Harvey and Siddique (2000) discuss the implications of skewness in asset pricing. Oppenheim and Lim (1981) discuss the importance of obtaining phase information for applications to image and speech reconstruction. Mutschler (2018) provides auto-cumulant calculations for popular econometric models, motivated by known nonlinearities in consumption and interest rate data. Tjøstheim (1994) provides an overview of the benefits of nonlinear time series analysis.

One can formulate nonlinear prediction by generalizing the linear minimizer of prediction MSE to a nonlinear solution, leading to the search for a "universal predictor"(Kock and Teräsvirta (2011)). One approach to finding a universal predictor involves the Kolmogorov-Gabor (KG) polynomial, a finite-lag version of the universal predictor's truncated Volterra expansion (Volterra (2005), Shiryaev (1960), Brillinger (1970)), which is discussed in Tsay (1986) and Teräsvirta et al. (2010)). Recent work in econometrics utilizing the KG approach include Krolzig and Hendry (2001) and Castle and Hendry (2010). There is also substantial literature on bilinear processes (Rao (1981), Quinn (1982), Rao et al. (1983), Hinich and Patterson (1985), Liu and Brockwell (1988), Liu (1992), and Terdik (2012)) and the uses of bispectral analysis (Gabr (1988) and Rao and Gabr (2012)). In this paper, we formulate this universal predictor in terms of higher-order moments of a time series, which potentially could be estimated nonparametrically, avoiding the need to specify a model.

To fix ideas, let $\{X_t\}$ be a zero mean stationary time series with autocovariance function $\gamma(\cdot)$, and suppose that we observe a finite stretch $X_1, \ldots, X_T$ of the time series. The prediction target is a future value $Y = X_{T+L}$ of the $\{X_t\}$ process (for some given $L \geq 1$, the forecast lead) on the basis of the past P observations $\underline{X} = [X_{T-P+1}, \ldots, X_T]'$ (where $'$ denotes the transpose). In the case that $P = T$ we use all the available samples for prediction, but it may be advantageous to take $P < T$ when a large value of P necessitates the estimation of many parameters.

As a first step, one might consider predictors that are quadratic functions of $\underline{X}$, with the understanding that first and second moments will no longer be sufficient to describe the solution – as established in (4) below. Such a quadratic predictor could be computed with either parametric or nonparametric approaches: if a particular parametric model is supposed that specifies the needed auto-cumulants (but perhaps is agnostic about other auto-cumulants not needed to describe the solution), then one can simply plug into the formula once the model has been fitted. A nonparametric approach would forgo modeling, and instead proceed with a consistent estimation of the auto-cumulants. Adequate estimation of polyspectra has already been addressed in the statistics literature: see Brillinger (1965), Rosenblatt and Van Ness (1965), Lii and Rosenblatt (1990), Berg (2008), and Berg and Politis (2009). In the signal processing literature there is also much attention given to the subject: Nikias and Raghuveer (1987), Mendel (1991), and Nikias and Mendel (1993). See also Hinich and Wilson (1990), which investigates the cross bi-spectrum in spatial data.

Given that much is already known about polyspectral estimation and the properties of specific kinds of nonlinear models, our focus in this paper is instead on the properties of nonlinear predictors, elicited through the analysis of polyspectra and the development of recursive computational algorithms. With a deeper understanding of the properties of nonlinear predictors (of the multivariate polynomial type), we aim to facilitate the application of nonlinear prediction to time series data. We remark that the use of polynomials, and the quadratic basis in particular, is not canonical, but is merely a convenient, agnostic choice that is motivated by the Volterra expansion of the optimal predictor; this also indicates that taking cubic and quartic terms will further improve the MSE. One can utilize different bases but in order to do calculations one must either know the type of nonlinear process or have to resort to auto-cumulant computations. Our proposal is useful in contexts where the exact specification of the nonlinear process is not known, or there are computational difficulties associated with finding its optimal predictor.

We now briefly describe the specific findings and main contributions of the paper. For the mean corrected random variates $Y = X_{T+L}$ and $\underline{X} = [X_{T-P+1}, \ldots, X_T]'$, the class of quadratic predictors we consider here has the generic form:

$$g(\underline{X}) = a + \underline{b}' \underline{X} + \underline{X}' B \underline{X}$$

for some constant a, coefficient vector $\underline{b}$ for the linear part, and (symmetric) coefficient matrix B for the quadratic part. The zero mean condition on Y and unbiasedness considerations suggest taking $a = -\mathbb{E}[\underline{X}' B \underline{X}] = -\text{tr.}(B \Sigma_{\underline{X},\underline{X}})$, leaving $\underline{b}$ and B as the free parameters of the quadratic predictor; here $\Sigma_{\underline{X},\underline{X}} = \text{Var}[\underline{X}]$ is the covariance matrix of $\underline{X}$ with jkth entry $\gamma(j-k)$, and tr. is the trace operator. Using some suitable vectorization steps, we develop below a generalized version of the Yule-walker equations for the quadratic prediction problem and derive an explicit expression for the optimal choices of $\underline{b}$ and B under the squared error loss. We also derive necessary and sufficient conditions under which the quadratic prediction approach improves upon linear prediction, providing a complete characterization – see Theorem 2. We call a stochastic process *second order forecastable* if it satisfies these necessary and sufficient conditions. Thus, there is a benefit in using the quadratic prediction approach *only* for such second order forecastable processes.

We also give examples of some popular nonlinear time series models, such as ARCH and GARCH models (which are shown to have some strange behavior) and nonlinear Hermite processes that are second order forecastable. On computational aspects of the proposed methodology, we give an outline of an algorithm for computing the higher order auto-cumulants and associated polyspectra that yield the coefficients in the optimal quadratic predictor. Numerical results reported in the paper show nontrivial improvements over linear prediction, with relative gains in mean squared (prediction) error being as high as 79.76%; see Table 1 in Section 6 below. To summarize, the key contributions of the paper are to develop a new quadratic prediction methodology, provide results for the identification of second order forecastable processes (where the quadratic prediction method can offer improvements over classical linear prediction theory), and develop necessary computing tools to make the methodology applicable in practice.

The rest of the paper is organized as follows. In Section 2, we describe the quadratic prediction methodology. In Section 3, we quantify potential gains from using the quadratic prediction approach over the traditional linear prediction methodology. In Section 4 we provide a complete characterization of the class of second order forecastable processes under some regular-

ity conditions, and Section 5 gives some examples of such processes arising from nonlinear time series models. Computational aspects and results from a numerical study are reported in Section 6. Additional simulation results are given in the supplementary materials file. In Section 7 we provide two real data examples involving the Unemployment Rate and the Wolfer Sunspots data. The quadratic predictors give reductions of 16.5% and 29.2% in the mean squared errors in the two cases respectively. The Appendix contains proofs of the main results, as well as some technical details related to the derivation of the auto-cumulants of Hermite processes.

2 The Quadratic Yule-Walker Equations and the Best Quadratic Predictor

To state the formula for the quadratic predictor, we require the notion of auto-moment. With $\mathbb{Z}$ denoting the set of all integers and with $t_0 = 0$, for $r \geq 2$, let

$$\gamma_r(t_1, \ldots, t_{r-1}) = \mathbb{E} \prod_{j=0}^{r-1} X_{t_j}, \quad t_1, \ldots, t_{r-1} \in \mathbb{Z}$$

denote the rth order auto-moment function of the stationary (mean zero) process $\{X_t\}$. In particular, for $r = 2$, $\gamma_2(t) \equiv \gamma(t) = \mathbb{E}X_0X_t$ is the autocovariance function of $\{X_t\}$.

Next, recall that $Y = X_{T+L}$ and $\underline{X}' = [X_{T-P+1}, \ldots, X_T]$ have mean zero. The entire random vector consisting of $\underline{X}$ with Y as the final component is denoted $\underline{Z}$. If $\underline{Z}$ is Gaussian, $\mathbb{E}[Y|\underline{X}] = \underline{b}'\,\underline{X}$ with $\underline{b} = \mathrm{Var}[\underline{X}]^{-1}\mathrm{Cov}[Y, \underline{X}]$. Moreover, even when $\underline{Z}$ is non-Gaussian this same solution minimizes the linear prediction problem

$$\mathbb{E}[(Y - \underline{b}'\,\underline{X})^2] = \mathrm{Var}[Y] - 2\,\underline{b}'\,\mathrm{Cov}[\underline{X}, Y] + \underline{b}'\,\mathrm{Var}[\underline{X}]\,\underline{b},$$

which is verified by computing the gradient and Hessian with respect to $\underline{b}$. Whereas the generic nonlinear prediction problem minimizes the MSE difference between Y and all functions $g(\underline{X})$, the quadratic problem posits a predictor of the form

$$\begin{aligned} g(\underline{X}) &= \underline{b}'\,\underline{X} + \underline{X}'\,B\,\underline{X} - \mathbb{E}[\underline{X}'\,B\,\underline{X}] \\ &= \sum_{t=1}^{P} b_t\,X_{T+1-t} + \sum_{1 \leq s \leq u \leq P} B_{su}\,\left(X_{T+1-s}X_{T+1-u} - \gamma_2(u-s)\right), \quad (1) \end{aligned}$$

where $\underline{b} = (b_1, \ldots, b_P)' \in \mathbb{R}^P$ and B is a $P \times P$ ((weakly) upper-triangular) real matrix with entries B_{su}. The second term of (1) is a bilinear form, involving a two-dimensional array (i.e., the matrix B). The centering of this bilinear form is needed to ensure that the predictor $g(\underline{X})$ has mean zero, as otherwise a bias is introduced. It is important to specify that B is (weakly) upper-triangular, as we seek to identify the coefficients of B that minimize MSE, and these will not be identifiable unless we restrict the bilinear form. The entries of the linear form are not constrained. The optimal quadratic predictor is obtained by choosing the coefficients $\underline{b}$ and B such that $\mathbb{E}[(Y - g(\underline{X}))^2]$ is minimized. This can be done by taking partial derivatives of the quadratic form with respect to the free coefficients $\{b_t : 1 \leq t \leq P\}$ and $\{B_{su} : 1 \leq s \leq u \leq P\}$. Setting these partial derivatives equal to zero, one obtains the equations

$$\gamma_2(L+t-1) = \sum_{t'=1}^{P} b_{t'}\, \gamma_2(t-t') + \sum_{1 \leq s' \leq u' \leq P} B_{s'u'}\, \gamma_3(t-s', t-u')$$

$$\gamma_3(L+u-1, u-s) = \sum_{t'=1}^{P} b_{t'}\, \gamma_3(u-t', u-s)$$

$$+ \sum_{1 \leq s' \leq u' \leq P} B_{s'u'} \left(\gamma_4(u-s', u-u', u-s) - \gamma_2(u-s)\gamma_2(u'-s') \right). \quad (2)$$

These are the *generalized Yule-Walker equations* for the quadratic prediction problem. Note that in addition to the autocovariance function, it involves the third and fourth order auto-moments of the $\{X_t\}$ process.

The equations in (2) can be expressed in a compact form and solved for the coefficients by recasting them using suitable matrix notation. To that end, note that the bilinear form can be expressed in the following two alternative ways:

$$\underline{X}'\, B\, \underline{X} = \mathrm{tr.}\{\underline{X}\, \underline{X}'\, B\} = \mathrm{vec}[B]'\, \mathrm{vec}[\underline{X}\, \underline{X}'],$$

where recall that for an $m \times n$ matrix A, $\mathrm{vec}[A]$ is the $mn \times 1$ vector obtained by stacking the columns of A. The upper-triangular structure of B ensures that we can without loss of generality consider the (weak) vech (where only the elements on or above the diagonal are included in the column-wise vectorization of B) in lieu of vec in the above. Let $\underline{W} = \mathrm{vech}[\underline{X}\, \underline{X}']$, set $\beta' = [\underline{b}', \mathrm{vech}[B]']$, and let $\Sigma_{A,B} = \mathrm{Cov}[A, B]$ for any random vectors A and

B. Then, we have the following result on the best quadratic predictor (BQP) of Y.

Proposition 1 *Suppose that* $\Sigma_{\underline{X},\underline{X}}$ *and* $S \equiv \Sigma_{\underline{W},\underline{W}} - \Sigma_{\underline{W},\underline{X}}\,\Sigma^{-1}_{\underline{X},\underline{X}}\,\Sigma_{\underline{X},\underline{W}}$ *are invertible. Then, the BQP of* Y *is given by* $[\underline{X}', \underline{W}']\widehat{\beta}$ *where*

$$\widehat{\beta} = \begin{bmatrix} \Sigma^{-1}_{\underline{X},\underline{X}}\,\Sigma_{\underline{X},Y} - \Sigma^{-1}_{\underline{X},\underline{X}}\,\Sigma_{\underline{X},\underline{W}}\,S^{-1}\left(\Sigma_{\underline{W},Y} - \Sigma_{\underline{W},\underline{X}}\,\Sigma^{-1}_{\underline{X},\underline{X}}\,\Sigma_{\underline{X},Y}\right) \\ S^{-1}\left(\Sigma_{\underline{W},Y} - \Sigma_{\underline{W},\underline{X}}\,\Sigma^{-1}_{\underline{X},\underline{X}}\,\Sigma_{\underline{X},Y}\right) \end{bmatrix}.$$

Note that the optimal weights $\widehat{\beta}$ depend on the autocovariance function as well as the third and fourth order auto-moments of the $\{X_t\}$ process, as expected from the generalized Yule-Walker equations (2). In the next section, we explore the benefits of using the quadratic predictor over its linear counterpart, leading to the quadratic prediction principle that gives a criterion for establishing the superiority of the BQP over the best linear predictor (BLP).

Example: All-Pass Noise

Consider a nonlinear process $\{X_t\}$ given as an all-pass filter of i.i.d. noise: let $\{Z_t\}$ be i.i.d. with mean zero, variance μ_2, and non-zero third and fourth cumulants μ_3 and μ_4; let $\psi(z) = (1 - z/\phi)/(1 - \phi z)$ for $|\phi| < 1$, and define $X_t = \psi(B)Z_t$, where B is the backshift operator. Because $\psi(z)$ is a scaled all-pass filter (i.e., $\psi(z)\psi(1/z) = \phi^{-2}$), $\{X_t\}$ is a white noise of variance σ^2/ϕ^2. However, the third auto-moments are non-trivial: we can expand $\psi(z) = \sum_{\ell\geq 0}\psi_\ell z^\ell$ with coefficients $\psi_0 = 1$ and $\psi_\ell = \phi^{\ell-1}(\phi - 1/\phi)$ for $\ell \geq 1$, so that $\gamma_3(t_1, t_2) = \mu_3 \sum_{\ell\geq 0}\psi_\ell\psi_{\ell+t_1}\psi_{\ell+t_2}$. For instance, if $t_1, t_2 > 0$ then $\gamma_3(t_1, t_2) = \mu_3\phi^{t_1+t_2}(\phi - 1/\phi)^2\phi^{-2}(1 - \phi)/(1 - \phi^3)$. Setting $L = P = 1$, we can apply Proposition 1, obtaining a forecast $\widehat{X}_{t+1}$ of the form $bX_t + c(X_t^2 - \gamma_2(0))$. We need to compute the quantities

$$\begin{aligned} \gamma_3(0,0) &= \mu_3(1 + (\phi - 1/\phi)^3/(1 - \phi^3)) \\ \gamma_3(1,0) &= \mu_3(\phi - 1/\phi)(1 + \phi(\phi - 1/\phi)^2/(1 - \phi^3)) \\ \gamma_4(0,0,0) &= \mu_4(1 + (\phi - 1/\phi)^4/(1 - \phi^4)) + 3\gamma_2(0)^2; \end{aligned}$$

noting that $\gamma_2(1) = 0$, we find that $\widehat{\beta}' = S^{-1}[-\gamma_3(0,0)\gamma_3(1,0)/\gamma_2(0), \gamma_3(1,0)]$, where $S = \gamma_4(0,0,0) - \gamma_2(0)^2 - \gamma_3(0,0)^2/\gamma_2(0)$. In particular, the Yule-Walker estimator is zero (the best linear prediction is the mean), but for quadratic prediction the linear weight b is $-\gamma_3(0,0)\gamma_3(1,0)S^{-1}/\gamma_2(0)$, and the quadratic weight is $c = \gamma_3(1,0)/S$.

3 Improvement over the linear predictor

Clearly, the quadratic portion of the solution disappears entirely if and only if

$$\Sigma_{\underline{W},Y} - \Sigma_{\underline{W},\underline{X}}\, \Sigma_{\underline{X},\underline{X}}^{-1}\, \Sigma_{\underline{X},Y} = 0, \tag{3}$$

in which case $\widehat{\underline{b}}$ (the first component of $\widehat{\beta}$) also reduces to the linear solution $\Sigma_{\underline{X},\underline{X}}^{-1}\, \Sigma_{\underline{X},Y}$. An important observation here is that condition (3) involves the third auto-cumulant, but not the fourth (and higher) order auto-cumulants. The quantity on the left of (3) can also be viewed as $\Sigma_{\underline{W},\widehat{E}^{(1)}}$, where, with $\widehat{Y}^{(1)}$ denoting the BLP of $Y = X_{t+L}$, the random variable $\widehat{E}^{(1)} = Y - \widehat{Y}^{(1)}$ gives the error in the linear predictor. Hence it follows that there is no benefit to quadratic prediction if and only if the error in the linear predictor is uncorrelated with $\underline{W} = \text{vech}[\underline{X}\, \underline{X}']$. This is certainly the case for Gaussian $\underline{Z}$, where all third auto-moments are zero; even though S is invertible (it is now given by $\Sigma_{\underline{W},\underline{W}}$), the condition (3) is true. In general, the full expression for the BQP, $\widehat{Y}^{(2)}$, is

$$\widehat{Y}^{(2)} = \widehat{Y}^{(1)} + \Sigma_{\widehat{E}^{(1)},\underline{W}}\, S^{-1} \left[\underline{W} - \mathbb{E}[\underline{W}] - \Sigma_{\underline{W},\underline{X}}\, \Sigma_{\underline{X},\underline{X}}^{-1}\, \underline{X}\right]. \tag{4}$$

This expresses the quadratic estimator as the linear estimator plus a modification that is only present if (3) is violated. Also, this modification is based on the difference between $\underline{W}$ and its linear projection upon $\underline{X}$. It is easy to check that the minimum mean squared prediction error (attained by the BQP, $\widehat{Y}^{(2)}$) is $\Sigma_{Y,Y} - \Sigma_{Y,\underline{X}}\, \Sigma_{\underline{X},\underline{X}}^{-1}\, \Sigma_{\underline{X},Y} - \Sigma_{\widehat{E}^{(1)},\underline{W}}\, S^{-1}\, \Sigma_{\underline{W},\widehat{E}^{(1)}}$, where the first two terms correspond to the prediction error for a linear problem. In other words, the efficiency loss of using a linear estimator when a quadratic is warranted is the non-negative quantity

$$\Sigma_{\widehat{E}^{(1)},\underline{W}}\, S^{-1}\, \Sigma_{\underline{W},\widehat{E}^{(1)}}, \tag{5}$$

which is non-negligible when (3) is violated. We summarize this discussion in the following result.

Theorem 1 *Suppose that* $\Sigma_{\underline{X},\underline{X}}^{-1}$ *and* S^{-1} *exist. Then,*
(i) the optimal quadratic predictor of Y *under the squared error loss function is given by* (4).
(ii) The minimal quadratic prediction error is given by

$$\Sigma_{Y,Y} - \Sigma_{Y,\underline{X}}\, \Sigma_{\underline{X},\underline{X}}^{-1}\, \Sigma_{\underline{X},Y} - \Sigma_{\widehat{E}^{(1)},\underline{W}}\, S^{-1}\, \Sigma_{\underline{W},\widehat{E}^{(1)}}.$$

(iii) The quadratic predictor improves upon the linear predictor if and only if (3) *fails or equivalently,*

$$\Sigma_{\widehat{E}^{(1)},\underline{W}}\, S^{-1}\, \Sigma_{\underline{W},\widehat{E}^{(1)}} \neq 0.$$

Theorem 1 gives the formulae for the best quadratic predictor and the minimal prediction error of a quadratic predictor. It also precisely describes situations where quadratic prediction may improve upon linear prediction. Thus, Theorem 1 leads us to the following general *quadratic prediction principle*:
"There is no benefit to quadratic prediction if and only if the linear prediction error is orthogonal to quadratic functions of the data, i.e., when (3) *holds."*

Example: All-Pass Noise
Continuing our prior example of all-pass noise, we see that (4) is $\widehat{Y}^{(2)} = \gamma_3(1,0)S^{-1}(X_t^2 - \gamma_2(0) - \gamma_3(0,0)\gamma_2(0)^{-1}X_t)$, and the improvement (5) to MSE over linear prediction (which has MSE given by the process variance $\gamma_2(0)$) is $\gamma_3(1,0)^2/S$; this is zero if and only if $1+\phi(\phi-1/\phi)^2/(1-\phi^3) = 0$, which holds if and only if $\phi = 1, -1/2$. Since $|\phi| < 1$ by assumption, we see that there is improvement for quadratic prediction if $\phi \neq -1/2$.

4 Second Order Forecastable Processes

In this section, we present a characterization of time series models where the quadratic prediction improves on linear prediction. For definiteness, we shall restrict attention to the case where the goal is to predict $Y = X_{t+1}$ based on an infinite past $\{X_t, X_{t-1}, \ldots\}$; this is akin to taking the $L = 1$, $P = \infty$ case of the previous section.

We say a time series is a *second order forecastable process* if and only if the MSE of its one-step ahead BQP is less than the MSE of its one-step ahead BLP; this generalizes the classical set of linear processes, for which the one-step ahead BLP is the conditional expectation. Below, we develop some key results on linear projections of quadratic terms, and provide a characterization of second order forecastable processes.

A stationary process with moments of all orders can be described in terms of its polyspectra; this is the approach to a frequency domain analysis of time series advocated by Brillinger (1981). We proceed by describing these results and relating them to the familiar case of a linear process. Any strictly stationary time series $\{X_t\}$ with moments of all orders has auto-cumulant

functions κ of order $r+1$ (for $r \geq 1$) defined via

$$\kappa_{r+1}(\underline{h}) = \text{cum}[X_{t+h_1}, X_{t+h_2}, \ldots, X_{t+h_r}, X_t],$$

where $\underline{h} = [h_1, h_2, \ldots, h_r]'$ is a r-vector of lags. Note that the order of the auto-cumulant corresponds to the number of variables included $(r+1)$, not the number of lags (r). Strict stationarity – or more generally, stationarity of order $(r+1)$ – guarantees that κ_{r+1} is only a function of the lags, and hence t is immaterial.

Recall the standing assumption that $\mathbb{E}[X_t] = 0$, and also recall the definition of the auto-moment functions γ of order $(r+1)$:

$$\gamma_{r+1}(\underline{h}) = \mathbb{E}[X_{t+h_1} X_{t+h_2} \ldots X_{t+h_r} X_t].$$

For $r = 1, 2$, we have $\kappa_{r+1} = \gamma_{r+1}$, but for $r \geq 3$ the auto-cumulant and auto-moment functions are distinct.

For the discussion below, we fix $r \geq 1$ and assume that the order $(r+1)$ auto-cumulant function, denoted simply by $\kappa_{r+1}(\underline{h}) = \kappa(\underline{h})$ for ease of exposition, is absolutely summable over $\underline{h} \in \mathbb{Z}^r$. With such an assumption, the polyspectrum of order $(r+1)$ is well-defined. The corresponding polyspectral density of order $(r+1)$ is given (with $i = \sqrt{-1}$) by

$$f(\underline{\lambda}) = \sum_{\underline{h} \in \mathbb{Z}^r} \kappa(\underline{h}) \, \exp\{-i \, \underline{\lambda}' \underline{h}\},$$

where $\underline{\lambda} = [\lambda_1, \ldots, \lambda_r]'$ denotes a r-vector of frequencies. Brillinger (1965) provides an elegant discussion as to why it is preferable to consider the Fourier transform of auto-cumulants rather than that of auto-moments. When applying a linear filter $\Psi(B) = \sum_{j \in \mathbb{Z}} \psi_j B^j$ to such an $\{X_t\}$, yielding a new $\{Y_t\}$ defined by $Y_t = \Psi(B) X_t$, one can relate the polyspectra of the filter output to the polyspectra of the filter input. Let f_y and f_x denote polyspectra of order $(r+1)$ for the $\{Y_t\}$ and $\{X_t\}$ processes; then by Theorem 2.8.1 of Brillinger (1981),

$$f_y(\underline{\lambda}) = f_x(\underline{\lambda}) \, \Psi(e^{i \sum_{j=1}^r \lambda_j}) \prod_{j=1}^r \Psi(e^{-i\lambda_j}). \tag{6}$$

Recall that it follows from the Wold decomposition (McElroy and Politis (2020)) that a purely non-deterministic stationary time series $\{X_t\}$ can be expressed as $X_t = \Psi(B) \, Z_t$, where $\Psi(z)$ is a power series such that $\Psi(0) = 1$, and $\{Z_t\}$ is a white noise process, with its rth cumulant denoted by μ_r. When

$\{Z_t\}$ is i.i.d., we say the process $\{X_t\}$ is *linear*; this appellation is connected to the fact that the minimal MSE one-step ahead forecast function is linear in the past data. In such a case, the polyspectrum of order $(r+1)$ is given by

$$f(\underline{\lambda}) = \mu_{r+1}\, \Psi(e^{i\sum_{j=1}^{r}\lambda_j}) \prod_{j=1}^{r} \Psi(e^{-i\lambda_j}), \tag{7}$$

which follows from (6) and the fact that all polyspectra for an i.i.d. sequence are equal to the constant cumulant μ_{r+1}. If we relax the assumption that $\{Z_t\}$ is i.i.d., the above formula will no longer be valid; potentially the $\{Z_t\}$ has a non-constant polyspectra. Conversely, given a polyspectrum it is possible to factorize it under certain conditions; see Tekalp and Erdem (1989). For the case $r = 1$, the well-known spectral factorization theorem (McElroy and Politis (2020)) yields

$$f(\lambda) = \mu_2\, \Psi_2(e^{-i\lambda})\, \Psi_2(e^{i\lambda}), \tag{8}$$

where $\Psi_2(z)$ is a power series such that $\Psi_2(0) = 1$. This factorization is possible when the process is invertible, i.e., the spectral density is strictly positive.

Within the above context we now provide an equivalent characterization of second order forecastable processes. The endeavor to predict X_{t+1} in terms of both linear and quadratic functions of past data can be re-expressed as a linear function of both $\{X_{t-\ell}\}_{\ell\geq 0}$ and $\{X_{t-j}X_{t-k}\}_{j,k\geq 0}$. Using Lemma 2 of Bell (1984) the forecast only depends on the linear portion if and only if

$$\mathrm{Cov}[X_{t+1}, X_{t-j}X_{t-k} - \widehat{X_{t-j}X_{t-k}}] = 0 \tag{9}$$

for all $j, k \geq 0$, where $\widehat{X_{t-j}X_{t-k}}$ denotes the linear prediction of $X_{t-j}X_{t-k}$ on the basis of $\{X_{t-\ell}\}_{\ell\geq 0}$. In other words, if the above covariance is non-zero for some j and k, the process is second order forecastable – following the ideas discussed in Sections 2 and 3. To understand this condition better, we first derive $\widehat{X_{t-j}X_{t-k}}$; this is expressible as the mean $\gamma(k-j)$ plus some causal filter $\Pi^{(j,k)}(B)$ applied to X_t, i.e.,

$$\widehat{X_{t-j}X_{t-k}} = \gamma(k-j) + \Pi^{(j,k)}(B)X_t \equiv \gamma(k-j) + \sum_{h\geq 0} \pi_h^{(j,k)} X_{t-h}. \tag{10}$$

To state the formula for this filter, we need new notations. Let the order $(r+1)$ auto-cumulant generating function be denoted as

$$f_{r+1}(\underline{z}) = \sum_{\underline{h} \in \mathbb{Z}^r} \kappa(\underline{h})\, z_1^{h_1} \cdots z_r^{h_r},$$

which reduces to the polyspectral density $f_{r+1}(\underline{\lambda})$ when $z_j = e^{-i\lambda_j}$ for $j = 1, \ldots, r$. Next, for any function $g(z)$ that is analytic on an open annulus containing the unit circle of $\mathbb{C}$, let

$$\langle g(z) \rangle_z = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(e^{-i\lambda})\, d\lambda.$$

Also, for any Laurent series $\Psi(z)$ (see Ahlfors (1979)), let $[\Psi(z)]_r^s = \sum_{j=r}^s \psi_j z^j$ for integers $r \leq s$. The following result gives the linear projection of the quadratic term $X_{t-j}X_{t-k}$ for any $j, k \geq 0$.

Proposition 2 *Let $\{X_t\}$ be strictly stationary with third moments, and absolutely summable auto-cumulants of order 2 and 3. Suppose the spectral density is positive, so that the factorization (8) exists. Then for any $j, k \geq 0$, the power series in z defined by*

$$\Pi^{(j,k)}(z) = \frac{1}{\mu_2} \left[z^j \langle y^{k-j} f_3(zy^{-1}, y) \rangle_y / \Psi_2(z^{-1}) \right]_0^{\infty} \Psi_2(z)^{-1}$$

yields the filter $\Pi^{(j,k)}(B)$ that generates the optimal linear estimate of $X_{t-j}X_{t-k}$ via (10).

Using Proposition 2, we can now prove our main result about the characterization of the property of being a second order forecastable process. The main assumption is that there are no zeroes in the spectral density. This is not a substantially new restriction, because if the spectral density is not positive then linear forecasting is also impossible.

Theorem 2 *Let $\{X_t\}$ be strictly stationary with fourth moments, and absolutely summable auto-cumulants of order 2, 3, and 4. Suppose the spectral density is positive, so that the factorization (8) exists. Then $\{X_t\}$ is second order forecastable if and only if the expression*

$$\langle \langle z^{j+1} y^{k+1} f_3(z, y) / \Psi_2(z^{-1} y^{-1}) \rangle_y \rangle_z \tag{11}$$

is nonzero for some $j, k \geq 0$.

Remark 1 The condition (9) is also equivalent to

$$\mathrm{Cov}[X_{t+1} - \widehat{X}_{t+1}, X_{t-j}X_{t-k}] = 0$$

for all $j, k \geq 0$, and therefore by Theorem 2 condition (11) says that the linear forecast error is orthogonal to all quadratic functions of the past, i.e., the quadratic prediction principle holds. Note that equation (9) for a particular $j, k \geq 0$ is a special case of (3), and by considering all j, k we generalize the finite-sample treatment to the case of a semi-infinite sample.

Remark 2 As a consequence of Theorem 2, we shall say that any process $\{X_t\}$ satisfying those hypotheses is by definition a second order forecastable process if and only if (11) is nonzero for some $j, k \geq 0$. It is immediate that Gaussian processes and causal linear processes are not second order forecastable: in the former case, $f_3(z, y) \equiv 0$, and in the latter case, using (7),

$$f_3(z, y)/\Psi_2(z^{-1}y^{-1}) = \mu_3 \Psi_2(z)\Psi_2(y),$$

on the unit circle, so that (11) equals zero for all $j, k \geq 0$.

Example: All-Pass Noise

Because $f_2(z) = \mu_2/\phi^2$, we find that $\Psi_2(z) = 1$, and hence (11) is equal to $\gamma_3(-j-1, -k-1)$. Choosing $j = k = 1$, we find that $\gamma_3(-2, -2) = \mu_3((\phi^2 - 1) + \phi^2(\phi - 1/\phi)^3/(1 - \phi^3))$, which is $\phi^2 - 1$ times $\gamma_3(1, 0)$ already computed above – this is non-zero if $\phi \neq -1/2$, in which case Theorem 2 guarantees the process is second order forecastable. We can also compute the filter coefficients of $\Pi^{(j,k)}(z)$ from Proposition 2: for $h \geq 0$ we have

$$\pi_h^{(j,k)} = \langle z^{-h}\Pi^{(j,k)}(z)\rangle = \frac{\phi^2}{\mu_2}\langle z^{j-h}\langle y^{k-j}f(zy^{-1}, y)\rangle_y\rangle_z = \frac{\phi^2}{\mu_2}\gamma_3(h-j, h-k).$$

So the connection to second order forecastability is very clear in this case: having a non-zero third order auto-moment at negative lags ensures that $\mathrm{Cov}[X_{t+1}, X_{t-j}X_{t-k}]$ in (9) is non-zero for some $j, k \geq 0$, and since $\mathrm{Cov}[X_{t+1}, \widehat{X_{t-j}X_{t-k}}] = 0$ (because $\{X_t\}$ is a white noise) this guarantees that the quadratic predictor does not reduce to a linear predictor. Moreover, the quadratic filter coefficients are proportional to the third order auto-moments.

To check whether a given nonlinear process is second order forecastable, one needs the spectral factorization Ψ_2 (see McElroy (2018) for algorithms) and an expression for the bi-spectral density, so that (11) can be directly calculated. Hence, Theorem 2 is useful only when the auto-cumulants or

polyspecta are known. If we know the model for the nonlinear process, we may be able to compute the polyspectra and verify (11), but in such a case we might also be able to compute the conditional expectation, which is preferable. If we do not know the nonlinear model, or if it is difficult to compute its polyspectra (e.g., there are no analytical formulas available), then we cannot use Theorem 2; in this case, we can still do quadratic prediction based upon sample estimates of the auto-cumulants, and use Theorem 1 to assess whether there is a benefit over linear prediction. We present below a special case of Theorem 2 as a corollary.

Corollary 1 *Let* $\{X_t\}$ *be defined via* $X_t = \psi(B)Z_t$ *for an invertible power series* $\psi(z)$ *and a white noise sequence* $\{Z_t\}$ *with absolutely summable auto-cumulants of* g_2*,* g_3*, and* g_4*. Then* $\{X_t\}$ *is second order forecastable if and only if the expression*

$$\langle \langle z^{j+1} y^{k+1} \psi(z) \psi(y) g_3(z,y) \rangle_y \rangle_z \tag{12}$$

is nonzero for some $j, k \geq 0$*.*

Clearly, Corollary 1 generalizes the linear process case discussed in Remark 2. If $\{Z_t\}$ is an all-pass noise, it follows by similar arguments that the resulting $\{X_t\}$ is second order forecastable. As another example, suppose that $\{Z_t\}$ is a martingale difference sequence (so $\mathbb{E}[Z_t | Z_{t-1}, Z_{t-2}, \ldots] = 0$); then this is a white noise with third order auto-moment $\gamma_3(j,k) = 0$ unless (i) $j = k \geq 0$, or (ii) $k = 0$ and $j \leq 0$, or (iii) $j = 0$ and $k \leq 0$. Then $g_3(z,y) = \varphi(zy) + \vartheta(z^{-1}) + \upsilon(y^{-1})$ for power series φ, ϑ, and υ with corresponding coefficients $\gamma_3(j,j)$, $\gamma_3(-j,0)$, and $\gamma_3(0,-j)$ respectively; since (12) equals zero for all $j, k \geq 0$, causal filters of martingale difference sequences are not second order forecastable.

In general, a process that is second order forecastable may also be third order forecastable – this just says that the best cubic predictor has a lower MSE than the BQP. In fact, writing $\mathcal{F}_d$ for the class of dth order forecastable processes, we see that $\mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$, since trivially any BQP can be written as a cubic predictor where the third order terms are zero. In order to partition the space of nonlinear processes, we take the intersection of each $\mathcal{F}_d$ with the complement of all higher order classes, i.e., $\mathcal{A}_d$ is the set of *order d augurable processes*, defined as processes in $\mathcal{F}_d$ such that the best order k predictor, for all $k > d$, gives no reduction to the MSE. This is a stronger condition, so we refer to an augury rather than a forecast; it follows that $\mathcal{A}_d$ consists of all processes for which the Volterra expansion of the one-step ahead conditional

expectation truncates to order d. Although the augurable classes are more elegant since they form a partition, it is more difficult to check membership.

5 Illustrations of Second Order Forecastable Processes

We have already illustrated the paper's results on the all-pass noise process, but we now present further results for two classes of second order forecastable processes, each of which is simple to simulate and study.

5.1 Hermite Processes A class of nonlinear processes for which the auto-moments can be calculated fairly directly is the Hermite class. Let $\{Z_t\}$ be a mean zero, stationary Gaussian with autocovariance $c(h)$ such that $c(0) = 1$. The Hermite polynomials are defined for $k \geq 1$ ($H_0 \equiv 1$) as

$$H_k(x) = \frac{(-1)^k}{\sqrt{k!}}\, e^{x^2/2}\, \partial_x^k e^{-x^2/2}.$$

This definition differs from some authors (cf. Taqqu (1977)), where division by $\sqrt{k!}$ is omitted. For a sequence of coefficients $\{J_k\}_{k\geq 1}$ that are square summable, let $g(x) = \sum_{k=1}^{\infty} J_k\, H_k(x)$ and define a Hermite process via $X_t = g(Z_t)$. This is a zero mean nonlinear process; see Janicki and McElroy (2016) for more details. We describe a general method for computing the auto-moments in Appendix B. The derivation involves the Hermite generating function and some combinatorial concepts, and may be of independent interest; see Appendix B for details.

5.1.1 Exponential Hermite Process. Here we set $g(x) = e^x - \mu$ for $\mu = e^{c(0)/2}$, so that $\{X_t\}$ is a lognormal process. The auto-moments are:

$$\begin{aligned}
\gamma(h_1) &= \mu^2 \left(\exp\{c(h_1)\} - 1\right) \\
\gamma_3(h_1, h_2) &= \mu^3 \left(\exp\{c(h_1) + c(h_2) + c(h_1 - h_2)\}\right. \\
&\quad \left. - \exp\{c(h_1)\} - \exp\{c(h_2)\} - \exp\{c(h_1 - h_2)\} + 2\right) \\
\gamma_4(h_1, h_2, h_3) &= \mu^4 \left(\exp\{c(h_1) + c(h_2) + c(h_3) + c(h_1 - h_2) + c(h_1 - h_3)\right. \\
&\quad \left. + c(h_2 - h_3)\} - \exp\{c(h_1 - h_2) + c(h_1 - h_3) + c(h_2 - h_3)\}\right)
\end{aligned}$$

$$
\begin{aligned}
&- \exp\{c(h_2) + c(h_3) + c(h_2 - h_3)\} \\
&- \exp\{c(h_1) + c(h_3) + c(h_1 - h_3)\} \\
&- \exp\{c(h_1) + c(h_2) + c(h_1 - h_2)\} \\
&+ \exp\{c(h_2 - h_3)\} + \exp\{c(h_1 - h_3)\} + \exp\{c(h_1 - h_2)\} \\
&+ \exp\{c(h_3)\} + \exp\{c(h_2)\} + \exp\{c(h_1)\} - 3)\,,
\end{aligned}
$$

which are easily derived using the formula for the expectation of the lognormal distribution. For such a process, the minimal MSE predictor among all function classes is known to be the exponential of the sum of the linear estimator plus half its MSE; hence we know there is a benefit to nonlinear prediction, and the lognormal process will generally be a second order forecastable process.

Of course, in practice we may not know that our data follows such a lognormal process, or may have difficulty fitting the model; if we use nonparametric estimation of the auto-moments (see below) and utilize quadratic prediction, we can expect a benefit even when the exact model specification is unknown. For example, if $\{Z_t\}$ is an MA(q) then the third auto-moment function is zero whenever $|h_1|, |h_2| > q$, or $|h_1|, |h_1 - h_2| > q$, or $|h_2|, |h_1 - h_2| > q$. In the case $q = 1$, we obtain

$$
\begin{aligned}
&\gamma_3(0,0) = \mu^3\,(e^3 - 3e + 2) \\
&\gamma_3(\pm 1, 0) = \gamma_3(0, \pm 1) = \gamma_3(1,1) = \gamma_3(-1,-1) = \mu^3\,(e^{1+2c(1)} - 2e^{c(1)} - e + 2) \\
&\gamma_3(2,1) = \gamma_3(1,2) = \gamma_3(-2,-1) = \gamma_3(-1,-2) = \mu^3\,(e^{2c(1)} - 2e^{c(1)} + 1),
\end{aligned}
$$

all other values being zero. Hence the bi-spectrum is

$$
\begin{aligned}
f(z,y) = \mu^3 \Big(&(e^3 - 3e + 2) \\
&+ (e^{2c(1)} - 2e^{c(1)} + 1)\,(z^2 y + z^{-2} y^{-1} + z y^2 + z^{-2} y^{-2}) \\
&+ (e^{1+2c(1)} - 2e^{c(1)} - e + 2)\,(z + z^{-1} + y + y^{-1} + zy + z^{-1} y^{-1}) \Big).
\end{aligned}
$$

However, the autocovariance function corresponds to an MA(1) process, and hence $\Psi_2(z) = 1 - \theta z$ for some θ determined from $\gamma(0)$ and $\gamma(1)$ via spectral factorization. It follows that (11) equals $\sum_{\ell=0}^{\infty} \theta^\ell \gamma_3(\ell - j - 1, \ell - k - 1)$, which is nonzero in general.

5.1.2 Squared Hermite Process. Another case is given by assuming that only J_1 and J_2 are non-zero so that the process is expressed as:

$$X_t = J_1\, H_1(Z_t) + J_2\, H_2(Z_t) = J_1\, Z_t + J_2\, (Z_t^2 - 1)/\sqrt{2}.$$

Using the general method in Appendix B, the auto-moments are given in terms of J_1 and J_2 as follows:

$$\begin{aligned}
\gamma(h_1) &= J_1^2\, c(h_1) + J_2^2\, c(h_1)^2 \\
\gamma_3(h_1,h_2) &= \sqrt{2} J_1^2 J_2\ (c(h_1)c(h_2) + c(h_1)c(h_1-h_2) + c(h_2)c(h_1-h_2)) \\
&\quad + (\sqrt{2}J_2)^3\, c(h_1)\, c(h_2)\, c(h_1-h_2) \\
\gamma_4(h_1,h_2,h_3) &= J_2^4\ \left(c(h_3)^2\, c(h_1-h_2)^2 + c(h_2)^2\, c(h_1-h_3)^2 + c(h_1)^2\, c(h_2-h_3)^2\right) \\
&\quad + 4J_2^4\ (c(h_1)\, c(h_2)\, c(h_1-h_3)\, c(h_2-h_3) + c(h_2)\, c(h_3)\, c(h_1-h_2)\, c(h_1-h_3) \\
&\qquad + c(h_1)\, c(h_3)\, c(h_1-h_2)\, c(h_2-h_3)) \\
&\quad + J_1^4\ (c(h_3)\, c(h_1-h_2) + c(h_1)\, c(h_2-h_3) + c(h_2)\, c(h_1-h_3)) \\
&\quad + J_1^2\, J_2^2\ \left(c(h_3)\, c(h_1-h_2)^2 + c(h_1-h_2)\, c(h_3)^2 + c(h_1-h_3)\, c(h_2)^2\right. \\
&\qquad \left. + c(h_2-h_3)\, c(h_1)^2 + c(h_1)\, c(h_2-h_3)^2 + c(h_2)\, c(h_1-h_3)^2\right) \\
&\quad + 2\, J_1^2\, J_2^2\ (c(h_1)\, c(h_1-h_2)\, c(h_2-h_3) + c(h_1)\, c(h_2)\, c(h_1-h_3) \\
&\qquad + c(h_2)\, c(h_1-h_2)\, c(h_1-h_3) + c(h_1)\, c(h_1-h_3)\, c(h_2-h_3) \\
&\qquad + c(h_3)\, c(h_1-h_2)\, c(h_2-h_3) + c(h_2)\, c(h_1-h_3)\, c(h_2-h_3) \\
&\qquad + c(h_3)\, c(h_1-h_2)\, c(h_1-h_3) + c(h_1)\, c(h_3)\, c(h_2-h_3) \\
&\qquad + c(h_2)\, c(h_3)\, c(h_1-h_3) + c(h_2)\, c(h_3)\, c(h_1-h_2) \\
&\qquad + c(h_1)\, c(h_2)\, c(h_2-h_3) + c(h_1)\, c(h_3)\, c(h_1-h_2))\,.
\end{aligned}$$

So long as $J_2 \neq 0$, such squared Hermite processes can be second order forecastable. For example, if $\{Z_t\}$ is an MA(q), then $\gamma_3(h_1,h_2) = 0$ if $|h_1| > q$ or $|h_2| > q$ or $|h_1 - h_2| > q$. In the case that $q = 1$, we find that $\gamma_3(h_1,h_2)$ is given by

$$\begin{cases}
(\sqrt{2}J_2)^3 + 3\sqrt{2}J_1^2 J_2 & (h_1,h_2) = (0,0) \\
(\sqrt{2}J_2)^3\, c(1)^2 + \sqrt{2}J_1^2 J_2 (c(1)^2 + 2c(1)) & (h_1,h_2) = (\pm 1,0), (0,\pm 1), (1,1), (-1,-1) \\
\sqrt{2}J_1^2 J_2 c(1)^2 & (h_1,h_2) = (1,-1), (-1,1), (2,1), (-2,-1), (1,2), (-1,-2)
\end{cases}$$

Hence the bi-spectrum is non-trivial, whereas the autocovariance function corresponds to an MA(1) process. For this process to be second order forecastable it is necessary for $J_2 \neq 0$ and $c(1) \neq 0$ (since otherwise, the process

either reduces to a Gaussian process – which must be linear – or reduces to an i.i.d. process); an interesting particular case occurs with the choice $J_1^2 = -c(1)J_2^2$ (so $c(1) < 0$ is required), because then $\gamma_2(1) = 0$ and the process is a white noise. Then the third order auto-moment for $(h_1, h_2) \in \{(\pm 1, 0), (0, \pm 1), (1, 1), (-1, -1), (1, -1), (-1, 1), (2, 1), (-2, -1), (1, 2), (-1, -2)\}$ equals $\sqrt{2}J_1^6/J_2^3$; also (11) simply equals $\gamma_3(-j-1, -k-1)$, which is non-zero for several choices of $(j, k) \in \{(0, 0), (1, 0), (0, 1)\}$, and in these cases the process is second order forecastable.

5.2 ARCH and GARCH Processes The class of ARCH and GARCH processes is extremely popular in modeling the log returns of stocks and indices in the financial sector. The market efficiency axiom indicates that any forecasts of such a process should have MSE equal to the variance, i.e., there is no benefit to be gained by prediction. It is easy to see that optimal linear forecasts equal zero (the mean), because the GARCH process is a white noise. (In contrast, the squared process has a non-trivial correlation, which allows the volatility to be forecasted with some success.) Moreover, the conditional expectation for the one-step ahead forecast is also zero, so the best predictor (in the MSE sense) equals the linear predictor, and there can not be any further advantages in using the quadratic prediction. In particular, (11) must be zero for all $j, k \geq 0$; we verify this below. However, for one-step behind backcasts, there is a potential advantage to the quadratic approach.

Conventionally, GARCH processes are defined in terms of driving inputs $\{Z_t\}$ that are symmetric, the first cases being studied having involved Gaussian distributions (Bollerslev (1986)). This was generalized to fat-tailed and asymmetric inputs – see Kotz et al. (2001), Trindade et al. (2010) and Kercheval and Liu (2011). These adaptations were driven by empirical considerations; here, we can show directly how kurtosis and asymmetry in the inputs impact the auto-cumulants and polyspectra of the GARCH process, and use them to verify (11). The GARCH(p,q) process is defined by

$$X_t = \sigma_t \, Z_t$$
$$\sigma_t^2 = a_0 + \sum_{j=1}^{p} a_j \, X_{t-j}^2 + \sum_{j=1}^{q} b_j \, \sigma_{t-j}^2,$$

where $Z_t \sim$ i.i.d.$(0, 1)$. We assume that all the a_j and b_j coefficients are non-negative, and that $\sum_j a_j + \sum_j b_j < 1$, which is sufficient to guarantee strict stationarity by Theorem 11.4.3 of McElroy and Politis (2020). Let $\omega(x) = \sum_{j=1}^{p} a_j x^j$ and $\theta(x) = 1 - \sum_{j=1}^{q} b_j x^j$. Set $\pi(z) = 1 - \omega(z)/\theta(z)$

and $\phi(z) = \theta(z) - \omega(z)$, so that $\pi(z) = \phi(z)/\theta(z)$. This power series will be written with a minus convention, i.e., $\pi(z) = 1 - \sum_{\ell \geq 1} \pi_\ell z^\ell$. Because the GARCH is a white noise, $f(\lambda) \equiv f_2(\lambda) = \mathbb{E}[X^2]$ for all λ, and

$$f_3(z,y) = \mathbb{E}[X^3] \left(\pi(zy)^{-1} + \pi(z^{-1})^{-1} + \pi(y^{-1})^{-1} - 2 \right).$$

(This is derived in Appendix A.) The expression for the fourth auto-cumulant function is omitted because it is extremely complicated, although the special case of the autocovariance for $\{X_t^2\}$ has a nice formula due to Bollerslev (1986). Because $\Psi_2(z) \equiv 1$ (i.e., the GARCH process is a white noise), we can deduce that condition (12) holds, and apply Corollary 1 to conclude that there is no benefit to quadratic forecasting over linear forecasting.

This result also holds for an ARMA-GARCH process, i.e., an ARMA process whose white noise innovations are a GARCH process. Letting the MA(∞) representation of the ARMA filter be denoted by $\psi(z)$, we see that the ARMA-GARCH process is of the type described in the hypothesis of Corollary 1, and we need to check condition (12). Hence we obtain, for all $j, k \geq 0$,

$$\langle\langle z^{j+1} y^{k+1} \psi(z)\psi(y) \left(\pi(zy)^{-1} + \pi(z^{-1})^{-1} + \pi(y^{-1})^{-1} - 2 \right) \rangle_y \rangle_z = 0.$$

The calculations in the GARCH case – and in particular, the expressions for the third order auto-cumulant – show that there is no correlation between X_t and past values of the squared process, and for this reason (together with the fact that the process is a white noise) there is no additional benefit to quadratic forecasting; however, if we instead examine one-step behind backcasts, now this correlation between the process and future values of the squared process leads to improvements in the quadratic predictors. Specifically, observe that backcasting is equivalent to forecasting the time-reversed GARCH process, for which we obtain (see Ch. 11, McElroy and Politis (2020)):

$$f_3(z,y) = \mathbb{E}[X^3] \left(\pi\big([zy]^{-1}\big)^{-1} + \pi(z)^{-1} + \pi(y)^{-1} - 2 \right), \text{ and}$$
$$\langle\langle z^{j+1} y^{k+1} f_3(z,y)/\Psi_2(z^{-1}y^{-1}) \rangle_y \big\rangle_z = \mathbb{E}[X^3]\, 1_{\{j=k\}}\, \widetilde{\pi}_{j+1},$$

where $\sum_{h \geq 0} \widetilde{\pi}_h x^h = \pi(x)^{-1}$. As a result, the BQP for the time reserved process will be better than its linear counterpart.

6 Computational Matters and Numerical Examples

We have implemented the methodology of this paper and applied it to various nonlinear processes, including the numerical examples reported below. In this section, we describe how the computations are done, and summarize the results. Recall that $\underline{W} = \text{vech}[\underline{X}\,\underline{X}']$. Construction of the matrix $\Sigma_{\underline{W},\underline{W}}$ proceeds by first building a larger 4-array out of the covariance of $\text{vec}[\underline{X}\,\underline{X}']$ with itself, allowing for redundancies. The indices i,j,k,ℓ for the four dimensions of the array each range from 1 to P. In R, applying the *matrix* operator to a 4-array constructs a block matrix, whereby j and ℓ are row and column block indices, and i and k are row and column indices within each block. For example, $3, 1, 4, 1$ represents the $3, 4$ entry in the upper left block of the matrix. In this ordering, the indices i and j correspond to various entries in the vector $\text{vec}[\underline{X}\,\underline{X}']$, conceived of as the transpose of $[\underline{X}'\,X_{T-P+1}, \underline{X}'\,X_{T-P+2}, \ldots, \underline{X}'\,X_T]$, or the collection of $\underline{X}'\,X_{T-P+j}$ with i giving the index within each $\underline{X}$. It follows that the i, j, k, ℓ entry of the array equals

$$\text{Cov}[X_{T-P+i}X_{T-P+j}, X_{T-P+k}X_{T-P+\ell}] = \gamma_4(i-\ell, j-\ell, k-\ell) - \gamma_2(i-j)\,\gamma_2(k-\ell).$$

(Note that auto-moment functions have many symmetries in their arguments, so there are many ways of writing the same quantity.) Once the entries of the 4-array have been filled in (inefficiently, by utilizing 4 nested loops over T elements), then certain row and column entries corresponding to the lower triangular entries of $\underline{X}\,\underline{X}'$ are omitted from the matrization of the array. In a similar fashion, we construct $\Sigma_{\underline{X},\underline{W}}$.

Hence the formulas for the BQP and its prediction error can be applied once the auto-moments are known. In practice, these can be obtained by fitting nonlinear models and plugging in the parameter estimates; alternatively, in cases where it is not practical to fit a nonlinear model, we can use nonparametric estimators. As described in Brillinger (1965), simple estimators of auto-moments and auto-cumulants can be constructed as follows: for a sample of size T and given a lag vector $\underline{h} = [h_1, h_2, \ldots, h_r]'$, define the index set $\mathcal{T}_{\underline{h}} = \cap_{\ell=0}^{r}\{1 - h_\ell, \ldots, T - h_\ell\}$ with $h_0 = 0$. Note that if any h_ℓ is greater than T or less than $1 - T$, then $\mathcal{T}_{\underline{h}} = \emptyset$. Then define the sample auto-moment of lag $\underline{h}$ as

$$\widehat{\gamma}_{r+1}(\underline{h}) = T^{-1} \sum_{t \in \mathcal{T}_{\underline{h}}} (X_{t+h_0} - \overline{X})(X_{t+h_1} - \overline{X}) \cdots (X_{t+h_r} - \overline{X}),$$

where $\overline{X} = T^{-1} \sum_{t=1}^{T} X_t$, and the sum is extended to be zero in cases where $\mathcal{T}_{\underline{h}} = \emptyset$. For any fixed $\underline{h}$, these estimators are asymptotically unbiased, with variance $O(T^{-1})$ assuming that all auto-cumulant functions are absolutely summable – see (2.6.1) of Brillinger (1981). However, values of $\underline{h}$ such that $\mathcal{T}_{\underline{h}}$ is small imply that there will be some bias in finite samples; note that $|h_j| < P$ for $1 \leq j \leq r$, so by restricting P to be much smaller than T we can ensure that bias is minimized. On the other hand, choosing a large P will make it difficult to estimate the filter coefficients consistently. Indeed, it is easy to show that the filter coefficients in the quadratic approach with P past variables can be estimated with accuracy $O_p(P^4 T^{-1/2})$, thereby imposing a natural upper bound on P.

In the simulations below, we set P to be a moderately large value to balance the trade off, following the choices considered in existing literature for other nonlinear prediction methods (cf. Fan and Yao (2008)).

For the simulation study, we consider the five different models as listed below:

- **Model IA:** $X_t = A\epsilon_t + B\epsilon_{t-1}^2 - B$ (with $\epsilon_t \sim^{\text{iid}}$ Unif(-1,1))
- **Model IB:** $X_t = A\epsilon_t + B\epsilon_{t-1}^2 - B$ (with $\epsilon_t \sim^{\text{iid}}$ Exp(1)-1)
- **Model IIA:** $X_t = J_1 H_1(Z_t) + J_2 H_2(Z_t)$ (with Z_t in (13) below and $\epsilon_t \sim^{\text{iid}}$ Gaussian(0,1))
- **Model IIB:** $X_t = J_1 H_1(Z_t) + J_2 H_2(Z_t)$ (with Z_t in (13) below and $\epsilon_t \sim^{\text{iid}}$ Exp(1)-1)
- **Model IIIA:** $X_t = \sum_{j \geq 0} \beta^j \prod_{n=0}^{j-1} e_{t-nk-\ell}\, e_{t-jk}$ (with $e_t \sim^{\text{iid}}$ Exp(1)-1)
- **Model IIIB:** $X_t = \sum_{j \geq 0} \beta^j \prod_{n=0}^{j-1} e_{t-nk-\ell}\, e_{t-jk}$ (with $e_t \sim^{\text{iid}}$ Gaussian(0,1)),

where in Models IIA and IIB, $\{Z_t\}$ is an AR(2) process given by

$$Z_t = 2\rho cos(\omega) Z_{t-1} - \rho^2 Z_{t-2} + \epsilon_t \tag{13}$$

with $\omega = \pi/4$ and ρ being a free parameter. Also, in Models IIIA and IIIB, k, ℓ are fixed integers. We considered three subcases of each of these models using different choices of parameters.

Table 1 gives MSEs for 5 different models for the linear and quadratic predictions, and also the MSE obtained by using a nonparametric approach

proposed in Chapter 10 of Fan and Yao (2008). The sample size is taken to be $T = 100$ and the past window length is taken to be $P = 20$. More tables are given in the Supplement for different choices of sample size (T) and past window length (P).

From Table 1, we find that the quadratic prediction engenders significant improvement in almost in all of the cases over the competing approaches. For the nonparametric method of Fan and Yao (2008), the improvement obtained by the quadratic approach can be as high as 92%, although there are cases where the nonparametric approach is superior (cf. Model IB, case 3). On the other hand, the quadratic prediction always provided improvements over the linear prediction, with the amount of improvement ranging from modest (e.g., 1.09% for Model IB, case 3) to substantial (79.76% for Model IIIB, case 2). Hence, while quadratic prediction may not always provide substantially better results than linear prediction (say, when the process is Gaussian), it performs better than its linear counterpart in presence of nonlinearity. On a cautionary note, we also point out that when the sample size is not large, the improvement in MSE (5) of using the BQP over the BQL could be offset by additional error due to parameter estimation uncertainty, in view of the fact that quadratic filters require the estimation of more filter coefficients.

7 Data Analysis

We consider two real data applications. The first example treats the case of forecasting the Wolfer sunspots, and the second example examines nonlinear forecasts of unemployment data.

7.1 Wolfer Sunspots The time series of Wolfer sunspots has been studied extensively. We consider a monthly vintage starting in 1749. Examination of the series shows large cyclical movements due to the known solar behavior, and the oscillations have an asymmetric shape. As the period is roughly 11 years, or 132 time units, longer-term forecasts should use auto-cumulants containing this many lags. We instead examine the one-step ahead forecast, which should not be greatly impacted by the solar oscillation. With $P = 30$ as the size of the predictor set used in the quadratic prediction problem, and with the whole sample used to estimate the auto-moments non-parametrically, we computed the one-step ahead prediction MSE resulting from both the linear and quadratic estimators: We found that the quadratic approach leads to a **29.2 % reduction** in MSE.

7.2 Unemployment Rate We also examined unemployment rate data from the Bureau of Labor Statistics. This series is the monthly Seasonally Adjusted Unemployment Rate (16 years and over, series id LNS14000000),

Table 1: MSE Comparison for $T = 100$ and $P = 20$

Model	Parameters	Quad Pred	Fan & Yao	Linear
IA	$A = -0.235,\ B = 0.376$	0.15	0.22 (31.81)	0.32 (53.12)
	$A = -0.350,\ B = 0.100$	0.14	0.18 (22.22)	0.24 (41.67)
	$A = 0.350,\ B = -0.100$	0.18	0.21 (14.29)	0.28 (35.71)
IB	$A = -0.235,\ B = 0.376$	0.59	1.28 (53.91)	1.12 (47.32)
	$A = -0.350,\ B = 0.100$	0.20	0.32 (37.50)	0.42 (52.38)
	$A = 0.350,\ B = -0.100$	0.27	0.16 (−68.75)	0.27 (1.09)
IIA	$\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$	0.77	1.22 (36.89)	1.32 (41.67)
	$\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$	0.43	0.85 (49.41)	0.73 (41.09)
	$\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$	1.14	15.80 (92.78)	1.58 (27.84)
IIB	$\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$	2.50	8.65 (71.09)	7.19 (65.23)
	$\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$	1.04	3.97 (73.81)	1.38 (24.63)
	$\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$	3.42	14.86 (76.98)	7.98 (57.14)
IIIA	$k = 1,\ l = 2,\ \beta = 0.3$	1.05	3.25 (67.69)	3.89 (73.01)
	$k = 2,\ l = 5,\ \beta = 0.3$	1.15	1.39 (17.27)	3.14 (63.37)
	$k = 5,\ l = 2,\ \beta = 0.3$	0.67	1.23 (32.52)	2.12 (60.85)
IIIB	$k = 1,\ l = 2,\ \beta = 0.3$	0.92	1.26 (26.98)	2.02 (54.45)
	$k = 2,\ l = 5,\ \beta = 0.3$	0.52	1.06 (50.94)	2.57 (79.76)
	$k = 5,\ l = 2,\ \beta = 0.3$	0.83	1.56 (46.79)	2.47 (66.39)

The values in the parentheses represent relative percentage improvements in MSE when Quadratic Prediction is used compared to the respective competing methods

covering the period January 1948 through July 2019, of the Labor Force Statistics from the Current Population Survey. This was downloaded from the Bureau of Labor Statistics on 4:30 PM, August 8, 2019 (https://data.bls.gov/timeseries/LNS14000000). The series is of considerable interest to economists and policy-makers, and is fairly smooth with occasional bursts of activity. A crude autoregressive fit indicates an AR(13) may be adequate from the standpoint of linear time series modeling, and hence we will use $P = 13$ as the size of the predictor set (though with the entire sample used to nonparametrically estimate the auto-moments). With this choice of P, the one-step ahead prediction MSE is **16.5 % reduced** by the quadratic predictor as compared to the linear one.

Funding Information Research partially supported by NSF grant number DMS 1811998.

Data availibility All Datasets used in this paper are publicly available, as indicated in the paper.

Declarations This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the U.S. Census Bureau.

Conflicts of interest The authors do not have any personal financial interests related to the subject matters discussed in this manuscript.

References

Abhyankar, A., Copeland, L. S., and Wong, W. (1997). Uncovering nonlinear structure in real-time stock-market indexes: the s&p 500, the dax, the nikkei 225, and the ftse-100. *Journal of Business & Economic Statistics*, 15(1):1–14.

Ahlfors, L. (1979). *Complex Analysis*. McGraw-Hill, New York.

Bell, W. (1984). Signal extraction for nonstationary time series. *The Annals of Statistics*, 12(2):646–664.

Berg, A. (2008). Multivariate lag-windows and group representations. *Journal of Multivariate Analysis*, 99(10):2479–2496.

Berg, A. and Politis, D. N. (2009). Higher-order accurate polyspectral estimation with flat-top lag-windows. *Annals of the Institute of Statistical Mathematics*, 61(2):477–498.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Brillinger, D. R. (1965). An introduction to polyspectra. *The Annals of mathematical statistics*, pages 1351–1374.

Brillinger, D. R. (1970). The identification of polynomial systems by means of higher order spectra. *Journal of Sound and Vibration*, 12(3):301–313.

Brillinger, D. R. (1981). *Time series: data analysis and theory*, volume 36. Siam.

Brockett, P. L., Hinich, M. J., and Patterson, D. (1988). Bispectral-based tests for the detection of gaussianity and linearity in time series. *Journal of the American Statistical Association*, 83(403):657–664.

Brockwell, P. J. and Davis, R. A. (2016). *Introduction to time series and forecasting*. springer.

Castle, J. L. and Hendry, D. F. (2010). A low-dimension portmanteau test for non-linearity. *Journal of Econometrics*, 158(2):231–245.

Fan, J. and Yao, Q. (2008). *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media.

Gabr, M. (1988). On the third-order moment structure and bispectral analysis of some bilinear time series. *Journal of Time Series Analysis*, 9(1):11–20.

Harvey, C. R. and Siddique, A. (2000). Conditional skewness in asset pricing tests. *The Journal of Finance*, 55(3):1263–1295.

Hinich, M. J. and Patterson, D. M. (1985). Identification of the coefficients in a non-linear: time series of the quadratic type. *Journal of Econometrics*, 30(1-2):269–288.

Hinich, M. J. and Wilson, G. R. (1990). Detection of non-gaussian signals in non-gaussian noise using the bispectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(7):1126–1131.

Janicki, R. and McElroy, T. S. (2016). Hermite expansion and estimation of monotonic transformations of gaussian data. *Journal of Nonparametric Statistics*, 28(1):207–234.

Kercheval, A. N. and Liu, Y. (2011). Risk forecasting with garch, skewed t distributions, and multiple timescales. *Handbook of Modeling High-Frequency Data in Finance*, 4:163.

Kock, A. B. and Teräsvirta, T. (2011). Forecasting with nonlinear time series models. *Oxford handbook of economic forecasting*, pages 61–87.

Kotz, S., Kozubowski, T. J., and Podgórski, K. (2001). Asymmetric multivariate laplace distribution. In *The Laplace distribution and generalizations*, pages 239–272. Springer.

Krolzig, H.-M. and Hendry, D. F. (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*, 25(6-7):831–866.

Lii, K. and Rosenblatt, M. (1990). Asymptotic normality of cumulant spectral estimates. *Journal of Theoretical Probability*, 3(2):367–385.

Liu, J. (1992). On stationarity and asymptotic inference of bilinear time series models. *Statistica Sinica*, pages 479–494.

Liu, J. and Brockwell, P. J. (1988). On the general bilinear time series model. *Journal of Applied Probability*, 25(3):553–564.

Maravall, A. (1983). An application of nonlinear time series forecasting. *Journal of Business & Economic Statistics*, 1(1):66–74.

McElroy, T. (2018). Recursive computation for block-nested covariance matrices. *Journal of Time Series Analysis*, 39(3):299–312.

McElroy, T. S. and Politis, D. N. (2020). *Time Series: A First Course with Bootstrap Starter*. CRC Press.

Mendel, J. M. (1991). Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications. *Proceedings of the IEEE*, 79(3):278–305.

Mutschler, W. (2018). Higher-order statistics for dsge models. *Econometrics and statistics*, 6:44–56.

Nikias, C. L. and Mendel, J. M. (1993). Signal processing with higher-order spectra. *IEEE Signal processing magazine*, 10(3):10–37.

Nikias, C. L. and Raghuveer, M. R. (1987). Bispectrum estimation: A digital signal processing framework. *Proceedings of the IEEE*, 75(7):869–891.

Oppenheim, A. V. and Lim, J. S. (1981). The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541.

Quinn, B. (1982). Stationarity and invertibility of simple bilinear models. *Stochastic Processes and their Applications*, 12(2):225–230.

Ramsey, J. B. and Rothman, P. (1996). Time irreversibility and business cycle asymmetry. *Journal of Money, Credit and Banking*, 28(1):1–21.

Rao, M. B., Rao, T. S., and Walker, A. (1983). On the existence of strictly stationary solutions to bilinear equations. *J. Time Series Anal*, 4:95–l.

Rao, T. S. (1981). On the theory of bilinear time series models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(2):244–255.

Rao, T. S. and Gabr, M. (2012). *An introduction to bispectral analysis and bilinear time series models*, volume 24. Springer Science & Business Media.

Rosenblatt, M. and Van Ness, J. W. (1965). Estimation of the bispectrum. *The Annals of Mathematical Statistics*, 36(4):1120–1136.

Shiryaev, A. N. (1960). Some problems in the spectral theory of higher-order moments. i. *Theory of Probability & Its Applications*, 5(3):265–284.

Taqqu, M. S. (1977). Law of the iterated logarithm for sums of non-linear functions of gaussian variables that exhibit a long range dependence. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 40:203–238.

Tekalp, A. M. and Erdem, A. T. (1989). Higher-order spectrum factorization in one and two dimensions with applications in signal modeling and nonminimum phase system identification. *IEEE Transactions on Acoustics Speech and Signal Processing*, 37(10):1537–1549.

Teräsvirta, T., Tjøstheim, D., Granger, C. W. J., et al. (2010). *Modelling nonlinear economic time series*. Oxford University Press Oxford.

Terdik, G. (2012). *Bilinear stochastic models and related problems of nonlinear time series analysis: a frequency domain approach*, volume 142. Springer Science & Business Media.

Tjøstheim, D. (1994). Non-linear time series: a selective review. *Scandinavian Journal of Statistics*, pages 97–130.

Trindade, A. A., Zhu, Y., and Andrews, B. (2010). Time series models with asymmetric laplace innovations. *Journal of Statistical Computation and Simulation*, 80(12):1317–1333.

Tsay, R. S. (1986). *Nonlinearity tests for time series. Biometrika*, 73(2):461– 466.

Volterra, V. (2005). *Theory of functionals and of integral and integrodifferential equations*. Courier Corporation.

Appendix A: Proofs

A.1 Proof of Proposition 1 Write $\mathcal{X}' = [\underline{X}', \text{vech}[\underline{X}\,\underline{X}']']$ for the complete data vector, and set $\beta' = [\underline{b}', \text{vech}[B]']$. Then, we can express a generic predictor $g(\underline{X}) = \underline{b}'\underline{X} + \underline{X}'B\underline{X} - \mathbb{E}\underline{X}'B\underline{X}$ as $g(\underline{X}) = \beta'\{\mathcal{X} - \mathbb{E}[\mathcal{X}]\}$. Next recall that $\Sigma_{A,B} = \text{Cov}[A, B]$ for random vectors A and B. Hence, the quadratic MSE can be expressed as

$$\mathcal{Q}(\beta) \equiv \mathbb{E}[(Y - g(\underline{X}))^2] = \text{Var}[Y] - 2\,[\Sigma_{Y,\underline{X}},\ \Sigma_{Y,\underline{W}}]\,\beta + \beta'\,M\,\beta,$$

where

$$M = \begin{bmatrix} \Sigma_{\underline{X},\underline{X}} & \Sigma_{\underline{X},\underline{W}} \\ \Sigma_{\underline{W},\underline{X}} & \Sigma_{\underline{W},\underline{W}} \end{bmatrix}.$$

Now setting $\frac{\partial \mathcal{Q}(\beta)}{\partial \beta} = 0$, we get the matrix equivalent of the generalized Yule-Walker equations in (2): $M\beta = [\Sigma_{Y,\underline{X}},\ \Sigma_{Y,\underline{W}}]$. Note that when both $\Sigma_{\underline{X},\underline{X}}$ and the Schur complement S are invertible, the matrix M is invertible, and it can be directly checked that

$$M^{-1} = \begin{bmatrix} \Sigma^{-1}_{\underline{X},\underline{X}} + \Sigma^{-1}_{\underline{X},\underline{X}}\Sigma_{\underline{X},\underline{W}}S^{-1}\Sigma_{\underline{W},\underline{X}}\Sigma^{-1}_{\underline{X},\underline{X}} & -\Sigma^{-1}_{\underline{X},\underline{X}}\Sigma_{\underline{X},\underline{W}}S^{-1} \\ -S^{-1}\Sigma_{\underline{W},\underline{X}}\Sigma^{-1}_{\underline{X},\underline{X}} & S^{-1} \end{bmatrix}.$$

The proposition now follows by computing $[\Sigma_{Y,\underline{X}},\ \Sigma_{Y,\underline{W}}]M^{-1}$, and taking the transpose. □

A.2 Proof of Proposition 2 Without loss of generality suppose that $\mathbb{E}[X_t] = 0$. Take $j, k \geq 0$ as fixed integers throughout the proof. First we note that the linear estimator $\widehat{X_{t-j}X_{t-k}}$ has to take the form (10) by basic linear projection theory (Brockwell and David (2016)), given that $\mathbb{E}[X_{t-j}X_{t-k}] = \gamma(k-j)$. The error process for the optimal linear estimator needs to be uncorrelated with $X_{t-\ell}$ for all $\ell \geq 0$, which yields the equations

$$\gamma_3(\ell - j, \ell - k) = \sum_{h \geq 0} \pi_h^{(j,k)}\,\gamma_2(\ell - h), \quad \text{which holds if and only if}$$

$$\langle\langle z^{j-\ell}y^{k-\ell}\,f_3(z,y)\rangle_y\rangle_z = \langle \Pi^{(j,k)}(z)\,z^{-\ell}f(z)\rangle_z, \quad \text{which holds if and only if}$$

$$\langle z^{j-\ell}\langle y^{k-j}\,f_3(zy^{-1}, y)\rangle_y\rangle_z = \mu_2\,\langle \Pi^{(j,k)}(z)\,z^{-\ell}\Psi_2(z)\Psi_2(z^{-1})\rangle_z.$$

The last line is obtained by using the change of variable $z \mapsto zy^{-1}$ (this amounts to shifting the frequency λ in $z = e^{-i\lambda}$, so there is no impact from the chain rule), and applying the spectral factorization (8). Consider summing this last equation against β_ℓ for $\ell \geq 0$, where we define these

coefficients for any desired $h \geq 0$ via $\beta_\ell = \widetilde{\psi}^{(2)}_{\ell-h}$ for $\ell \geq h$, and zero otherwise. Here $\widetilde{\psi}^{(2)}_k$ is the kth coefficient of $\Psi_2(z)^{-1}$. This definition means that $\sum_{\ell \geq 0} \beta_\ell z^\ell = \Psi_2(z)^{-1} z^h$, and we can provide this construction for any $h \geq 0$. The application of these coefficients yields a new system of equations, which hold for all $h \geq 0$:

$$\langle z^{j-h} \langle y^{k-j} f_3(zy^{-1}, y)\rangle_y / \Psi_2(z^{-1})\rangle_z = \mu_2 \langle \Pi^{(j,k)}(z)\, z^{-h} \Psi_2(z)\rangle_z.$$

Note that $\Pi^{(j,k)}(z)\, \Psi_2(z)$ is a power series (i.e., it corresponds to some causal filter), and hence the right hand side of the above equation is just μ_2 times the h^{th} coefficient of this power series. It follows that there can be no anti-causal portion of the left-hand side of the equation, i.e.,

$$\Big[z^j \langle y^{k-j} f_3(zy^{-1}, y)\rangle_y / \Psi_2(z^{-1}) \Big]_{-\infty}^{-1} = 0, \quad \text{and}$$
$$\Big[z^j \langle y^{k-j} f_3(zy^{-1}, y)\rangle_y / \Psi_2\big(z^{-1}\big) \Big]_{0}^{\infty} = \mu_2\, \Pi^{(j,k)}(z) \Psi_2(z).$$

Note that $\mu_2 \neq 0$, so dividing by $\mu_2 \Psi_2(z)$ yields the stated formula. □

A.3 Proof of Theorem 2 The theorem claims that (9) holds (for all $j, k \geq 0$) if and only if (11) equals zero (for all $j, k \geq 0$). Fixing arbitrary $j, k \geq 0$, (9) holds if and only if

$$\gamma_3(-j-1, -k-1) = \sum_{h \geq 0} \pi_h^{(j,k)}\, \gamma_2(-1-h).$$

Utilizing the result of Proposition 2, (9) holds if and only if

$$\begin{aligned}
\langle \langle z^{j+1} y^{k+1} f_3(z, y)\rangle_y \rangle_z &= \langle \Pi^{(j,k)}(z)\, z f(z)\rangle_z \\
&= \langle z\Psi_2(z^{-1}) \Big[z^j \langle y^{k-j} f_3(zy^{-1}, y)\rangle_y / \Psi_2(z^{-1}) \Big]_0^{\infty} \rangle_z \\
&= \langle z\Psi_2(z^{-1}) z^j \langle y^{k-j} f_3(zy^{-1}, y)\rangle_y / \Psi_2(z^{-1})\rangle_z \\
&\quad - \langle z\Psi_2(z^{-1}) \Big[z^j \langle y^{k-j} f_3(zy^{-1}, y)\rangle_y / \Psi_2(z^{-1}) \Big]_{-\infty}^{-1} \rangle_z .
\end{aligned}$$

To obtain the last equality, we have used the fact that a Laurent series $\Theta(z)$ can be written as $[\Theta(z)]_0^\infty = \Theta(z) - [\Theta(z)]_{-\infty}^{-1}$. Next, note that

$$\langle z\Psi_2(z^{-1}) z^j \langle y^{k-j} f_3(zy^{-1}, y)\rangle_y / \Psi_2(z^{-1})\rangle_z = \langle \langle (zy^{-1})^{j+1} y^{k+1} f_3(zy^{-1}, y)\rangle_y \rangle_z .$$

Therefore, (9) holds if and only if

$$0 = \langle z\Psi_2(z^{-1})\Big[z^j\langle y^{k-j}\, f_3(zy^{-1},y)\rangle_y/\Psi_2(z^{-1})\Big]_{-\infty}^{-1}\rangle_z$$
$$= \langle \Psi_2(z^{-1})\Big[z^{j+1}\langle y^{k-j}\, f_3(zy^{-1},y)\rangle_y/\Psi_2(z^{-1})\Big]_{-\infty}^{0}\rangle_z,$$

which uses the fact that for any Laurent series $\Theta(z)$, $z[\Theta(z)]_{-\infty}^{-1} = [z\Theta(z)]_{-\infty}^{0}$. The final expression is the integral of the product of two power series, and hence the product of their index-zero coefficients must be zero; because $\Psi_2(0) = 1$, (9) holds if and only if

$$0 = \langle z^{j+1}\langle y^{k-j}\, f_3(zy^{-1},y)\rangle_y/\Psi_2(z^{-1})\rangle_z.$$

Now applying the transformation $z \mapsto zy$, we obtain (11) equals zero for all $j, k \geq 0$. □

A.4 Proof of Corollary 1 Because $\{Z_t\}$ is a white noise, $g_2(z) \propto 1$, and $f_2(z) \propto \psi(z)\psi(z^{-1})$. By invertibility of $\psi(z)$, the spectral density is positive and we find that $\Psi_2(z) = \psi(z)$. Also, by (6) we see that $f_3(z,y) = \psi(z)\psi(y)\psi(z^{-1}y^{-1})g_3(z,y)$, and hence (11) becomes (12). □

A.5 Derivation of the GARCH Bi-spectrum This derivation is taken from the solution to Exercise 11.27 of McElroy and Politis (2020). Since σ_t only depends on past values $X_{t-1}, X_{t-2}, \ldots$, for $k_1 \neq k_2 \neq 0$ the third auto-cumulant $\mathbb{E}[X_t\, X_{t+k_1}\, X_{t+k_2}] = 0$; this is because one of the three indices t, $t+k_1$, and $t+k_2$ is largest – say it is t. Then $X_t = Z_t\,\sigma_t$ and Z_t is independent of σ_t and X_{t+k_1} and X_{t+k_2}, so that

$$\mathbb{E}[X_t\, X_{t+k_1}\, X_{t+k_2}] = \mathbb{E}[Z_t]\,\mathbb{E}[\sigma_t\, X_{t+k_1}\, X_{t+k_2}] = 0.$$

Also, if $k_1 = k_2 = 0$ then clearly the third auto-cumulant equals $\mathbb{E}[X^3] = \mu_3$. If $k_1 = k_2 \neq 0$, we obtain

$$\mathbb{E}[X_0\, X_{k_1}^2] = \mathbb{E}[Z_{k_1}^2]\,\mathbb{E}\left[X_0\left(a_0/\theta(1) + \sum_{j=1}^{\infty}\pi_j\, X_{k_1-j}^2\right)\right] = \sum_{j=1}^{\infty}\pi_j\,\mathbb{E}[X_0\, X_{k_1-j}^2]$$

when $k_1 > 0$, where we have used Corollary 11.4.5 of McElroy and Politis (2020) to express $X_{k_1}^2$ in terms of past squared values of the process. Otherwise, if $k_1 < 0$ the auto-cumulant is zero, because Z_0 is independent of the other variables in that case. Let $\nu_k = \mathbb{E}[X_0\, X_k^2]$, which is zero if

$k < 0$. Then we have shown $\nu_k = \sum_{j=1}^{\infty} \pi_j \, \nu_{k-j}$, or $\pi(B) \, \nu_k = 0$; with $\nu_0 = \mathbb{E}[X^3]$, this homogeneous Ordinary Differencing Equation can be rewritten as $\pi(z) \, \nu(z) = \mathbb{E}[X^3]$, and therefore $\nu(z) = \mathbb{E}[X^3]/\pi(z)$.

Next, considering the cases that $k_1 = 0 \neq k_2$ and $k_2 = 0 \neq k_1$, we obtain

$$\kappa_3(k_1, k_2) = \mathbb{E}[X_0^2 \, X_{k_2}] = \mathbb{E}[X_0 \, X_{-k_2}^2] = \nu_{-k_2}$$
$$\kappa_3(k_1, k_2) = \mathbb{E}[X_0^2 \, X_{k_1}] = \mathbb{E}[X_0 \, X_{-k_1}^2] = \nu_{-k_1}$$

respectively, where $k_2, k_1 < 0$. Together these expressions yield the third auto-cumulant function, and the bi-spectral density is

$$\begin{aligned} f_3(z, y) &= \mathbb{E}[X^3] + \sum_{k>0} \nu_k \, (zy)^k + \sum_{k<0} \nu_{-k} \, z^k + \sum_{k<0} \nu_{-k} \, y^k \\ &= \mathbb{E}[X^3] + (\nu(zy) - \mathbb{E}[X^3]) + (\nu(z^{-1}) - \mathbb{E}[X^3]) + (\nu(y^{-1}) - \mathbb{E}[X^3]) \\ &= \mathbb{E}[X^3] \, \left(\pi(zy)^{-1} + \pi(z^{-1})^{-1} + \pi(y^{-1})^{-1} - 2 \right). \end{aligned}$$

Appendix B: Computing Auto-moments of Hermite Processes

Define the generating function $h(x, t) = \exp\{xt - t^2/2\}$, which is related to the Hermite polynomials via

$$h(x, t) = \sum_{k=0}^{\infty} \frac{t^k}{\sqrt{k!}} \, H_k(x).$$

It is known that the autocovariance function is given by $\gamma(h) = \sum_{\ell \geq 0} J_\ell^2 \, c(h)^\ell$ assuming that $\sum_{\ell \geq 0} J_\ell^2 < \infty$ (see Taqqu (1977)), and we generalize this below. It can be shown that the formula for the order $(r+1)$ auto-moment is

$$\gamma_{r+1}(\underline{h}) = \sum_{\underline{\ell} \geq 0} J_{\ell_0} \cdot J_{\ell_1} \cdots J_{\ell_r} \prod_{j=0}^{r} (\ell_j!)^{-1/2} \, \frac{\partial^{\ell_j}}{\partial s_j^{\ell_j}} \mathbb{E} \left[\exp \left\{ \sum_{i=0}^{r} s_i Z_{t+h_i} - \sum_{i=0}^{r} s_i^2/2 \right\} \right] |_{s_j = 0},$$

where $h_0 = 0$, and $\underline{\ell} = [\ell_0, \ell_1, \ldots, \ell_r]'$. (Here we assume that decay conditions on J_ℓ hold, sufficient to guarantee that the above expansion for $\gamma_{r+1}(\underline{h})$ is valid; Taqqu (1977) has a related treatment and corresponding conditions.) The expectation can be expanded as follows: let $K = \{(m, n) : 0 \leq m <$

$n \le r\}$. Then it follows from the formula for the mean of the lognormal distribution that

$$\mathbb{E}\left[\exp\left\{\sum_{i=0}^{r} s_i Z_{t+h_i} - \sum_{i=0}^{r} s_i^2/2\right\}\right] = \exp\left\{\sum_{(m,n)\in K} s_m s_n\, c(h_m - h_n)\right\},$$

and by differentiation the auto-moments can be determined. We now consider the cases of the third and fourth auto-moments. For $r = 2$ we compute

$$\begin{aligned}
&\frac{\partial^{\ell_0}}{\partial s_0^{\ell_0}}\frac{\partial^{\ell_1}}{\partial s_1^{\ell_1}}\frac{\partial^{\ell_2}}{\partial s_2^{\ell_2}} \exp\{s_0 s_1\, c(h_1) + s_0 s_2\, c(h_2) + s_1 s_2\, c(h_1 - h_2)\}\Big|_{s_0=s_1=s_2=0} \\
&= \frac{\partial^{\ell_0}}{\partial s_0^{\ell_0}}\frac{\partial^{\ell_1}}{\partial s_1^{\ell_1}}\frac{\partial^{\ell_2}}{\partial s_2^{\ell_2}} \sum_{n_0,n_1,n_2\ge 0} \Big[\frac{s_0^{n_0+n_1} s_1^{n_0+n_2} s_2^{n_1+n_2}}{n_0!n_1!n_2!} \\
&\qquad\qquad c(h_1)^{n_0}\, c(h_2)^{n_1}\, c(h_1-h_2)^{n_2}\Big]\Big|_{s_0=s_1=s_2=0},
\end{aligned}$$

which is nonzero only if $\ell_0 = n_0 + n_1$, $\ell_1 = n_0 + n_2$, and $\ell_2 = n_1 + n_2$.

We can succinctly write these conditions: let $\mathbb{N}_+ = \mathbb{N} \cup \{0\}$, and denote a vector of indices by $\underline{n} = [n_0, n_1, n_2]'$. Then defining

$$A_3 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \qquad \mathcal{N}_{\underline{\ell}} = \{n_0, n_1, n_2 \in \mathbb{N}_+ : \underline{\ell} = A_3\, \underline{n}\},$$

it follows that

$$\gamma_3(h_1,h_2) = \sum_{\underline{\ell}\in\mathbb{N}_+{}^3} J_{\ell_0}\cdot J_{\ell_1}\cdot J_{\ell_2} \sum_{\underline{n}\in\mathcal{N}_{\underline{\ell}}} \frac{\sqrt{\ell_0!\ell_1!\ell_2!}}{n_0!n_1!n_2!}\, c(h_1)^{n_0}\, c(h_2)^{n_1}\, c(h_1-h_2)^{n_2}.$$

For $r = 3$ we have

$$\begin{aligned}
&\frac{\partial^{\ell_0}}{\partial s_0^{\ell_0}}\frac{\partial^{\ell_1}}{\partial s_1^{\ell_1}}\frac{\partial^{\ell_2}}{\partial s_2^{\ell_2}}\frac{\partial^{\ell_3}}{\partial s_3^{\ell_3}} \exp\{s_0 s_1\, c(h_1) + s_0 s_2\, c(h_2) + s_0 s_3\, c(h_3) + \\
&\qquad s_1 s_2\, c(h_1-h_2) + s_1 s_3\, c(h_1-h_3) + s_2 s_3\, c(h_2-h_3)\}\,|_{s_0=s_1=s_2=s_3=0} \\
&= \frac{\partial^{\ell_0}}{\partial s_0^{\ell_0}}\frac{\partial^{\ell_1}}{\partial s_1^{\ell_1}}\frac{\partial^{\ell_2}}{\partial s_2^{\ell_2}}\frac{\partial^{\ell_3}}{\partial s_3^{\ell_3}} \sum_{n_0,n_1,n_2,n_3,n_4,n_5\ge 0} \frac{s_0^{n_0+n_1+n_2} s_1^{n_0+n_3+n_4} s_2^{n_1+n_3+n_5} s_3^{n_2+n_4+n_5}}{n_0!n_1!n_2!n_3!n_4!n_5!} \\
&\qquad c(h_1)^{n_0}\, c(h_2)^{n_1}\, c(h_3)^{n_2}\, c(h_1-h_2)^{n_3}\, c(h_1-h_3)^{n_4}\, c(h_2-h_3)^{n_5}|_{s_0=s_1=s_2=s_3=0},
\end{aligned}$$

which is nonzero only if $\ell_0 = n_0+n_1+n_2$, $\ell_1 = n_0+n_3+n_4$, $\ell_2 = n_1+n_3+n_5$, and $\ell_3 = n_2 + n_4 + n_5$. In this case we define

$$A_4 = \begin{bmatrix} 1\,1\,1\,0\,0\,0 \\ 1\,0\,0\,1\,1\,0 \\ 0\,1\,0\,1\,0\,1 \\ 0\,0\,1\,0\,1\,1 \end{bmatrix} \qquad \mathcal{N}_{\underline{\ell}} = \{n_0, n_1, n_2, n_3, n_4, n_5 \in \mathbb{N}_+ : \underline{\ell} = A_4\,\underline{n}\},$$

and obtain $\gamma_4(h_1, h_2, h_3)$ is given by

$$\sum_{\underline{\ell}\in\mathbb{N}_+^4} J_{\ell_0} \cdot J_{\ell_1} \cdot J_{\ell_2} \cdot J_{\ell_3} \sum_{\underline{n}\in\mathcal{N}_{\underline{\ell}}} \frac{\sqrt{\ell_0!\ell_1!\ell_2!\ell_3!}}{n_0!n_1!n_2!n_3!n_4!n_5!}$$
$$c(h_1)^{n_0}\, c(h_2)^{n_1}\, c(h_3)^{n_2} c(h_1-h_2)^{n_3}\, c(h_1-h_3)^{n_4}\, c(h_2-h_3)^{n_5}.$$

We can now apply these formulas to the case of a quadratic Hermite process. For $r = 2$ (and restricting ourselves to $\ell_0, \ell_1, \ell_2 \leq 2$), the only nonzero terms occur when either $\ell_0 = \ell_1 = \ell_2 = 2$ (which implies $n_0 = n_1 = n_2 = 1$), or for $\underline{\ell}' = [1, 1, 2]$ (corresponding to $\underline{n}' = [0, 1, 1]$), $\underline{\ell}' = [1, 2, 1]$ (corresponding to $\underline{n}' = [1, 0, 1]$), or $\underline{\ell}' = [2, 1, 1]$ (corresponding to $\underline{n}' = [1, 1, 0]$); then we obtain the stated formula. For $r = 3$, there are sixteen configurations for the sum over the ℓ indices, as each can take the value one or two. It turns out that eight of these configurations can yield solutions in terms of $n_0, \ldots, n_5$, which in turn are each constrained to be zero or one. These configurations are described in Table 2. By carefully organizing the results, we obtain the stated expression for the fourth-order auto-moment.

Table 2: Configurations of ℓ and n indices

$\underline{\ell}$	$\mathcal{N}_{\underline{\ell}}$
1 1 1 1	[0 0 1 1 0 0], [1 0 0 0 0 1], [0 1 0 0 1 0]
1 1 2 2	[0 0 1 1 0 1], [0 1 0 0 1 1], [1 0 0 0 0 2]
1 2 1 2	[0 0 1 1 1 0], [1 0 0 0 1 1], [0 1 0 0 2 0]
1 2 2 1	[0 0 1 2 0 0], [1 0 0 1 0 1], [0 1 0 1 1 0]
2 1 1 2	[0 0 2 1 0 0], [1 0 1 0 0 1], [0 1 1 0 1 0]
2 1 2 1	[0 1 1 1 0 0], [1 1 0 0 0 1], [0 2 0 0 1 0]
2 2 1 1	[1 0 1 1 0 0], [2 0 0 0 0 1], [1 1 0 0 1 0]
2 2 2 2	[0 0 2 2 0 0], [1 0 1 1 0 1], [2 0 0 0 0 2], [1 1 0 0 1 1], [0 1 1 1 1 0], [0 2 0 0 2 0]

Left column gives values for ℓ_0, ℓ_1, ℓ_2, and ℓ_3, and right column gives corresponding possible values for n_0, n_1, n_2, n_3, n_4, and n_5

Tucker S. McElroy
Research and Methodology
Directorate, U.S. Census Bureau,
4600 Silver Hill Road, 20233
Washington, DC, USA
E-mail: tucker.s.mcelroy@census.gov

Dhrubajyoti Ghosh and Soumendra Lahiri
Department of Mathematics and Statistics, Washington University in St. Louis, 1 Brookings Drive, 63130
Missouri St Louis, USA
E-mail: d.ghosh@wustl.edu
E-mail: s.lahiri@wustl.edu

Paper received: 20 December 2022