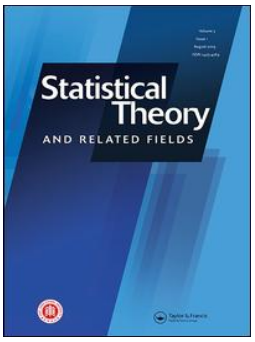

Taylor & Francis
Taylor & Francis Group

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Nonlinear prediction via Hermite transformation

Tucker McElroy & Srinjoy Das

To cite this article: Tucker McElroy & Srinjoy Das (2020): Nonlinear prediction via Hermite transformation, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2020.1856589

To link to this article: https://doi.org/10.1080/24754269.2020.1856589

Published online: 17 Dec 2020.

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2020.1856589

Nonlinear prediction via Hermite transformation

Tucker McElroy[a] and Srinjoy Das[b]

[a]U.S. Census Bureau, Washington, DC, USA; [b]University of California San Diego, La Jolla, CA, USA

ABSTRACT
General prediction formulas involving Hermite polynomials are developed for time series expressed as a transformation of a Gaussian process. The prediction gains over linear predictors are examined numerically, demonstrating the improvement of nonlinear prediction.

ARTICLE HISTORY
Received 25 October 2019
Revised 20 October 2020
Accepted 24 November 2020

KEYWORDS
Gaussian transformation; Hermite polynomial; nonlinear forecasting; time series

1. Introduction

The general prediction problem is to compute the best predictor of a random variable Y given a data vector $\underline{X}$, where a joint distribution is presumed to exist for Y and $\underline{X}$. If we define 'best' according to mean squared error (MSE) loss,[1] the best predictor (when the random variables are square integrable) is the conditional expectation $\mathbb{E}[Y \mid \underline{X}]$, which in the case of Gaussian variables is a linear function of $\underline{X}$. This linear function is completely computable in terms of first and second moments of the joint vector $(Y, \underline{X})$, as discussed in Brockwell and Davis (2013, Chapter 2). The problem can also be generalised to projection on infinite data sets, which arise in forecasting and signal extraction problems.

The theory for linear predictors is very well understood and is commonly applied to non-Gaussian data 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 (Brockett et al., 1988; Maravall, 1983). A common technique for handling such raw time series data is to apply a transformation that reduces asymmetry and kurtosis, thereby generating cumulants in the transformed space that more closely resemble Gaussian cumulants. Box-Cox transforms are an example of such functions, and are typically identified through exploratory analysis or via metadata; see discussion in McElroy (2016).

This paper provides exact formulas for non-linear prediction in scenarios where the non-Gaussian data process can be expressed as a univariate transformation of some Gaussian process. For some special cases, such as the log-normal distribution, exact formulas are already available for nonlinear predictors; here the general case is developed. The main result of the paper (Section 2) is an expansion of the conditional expectation in terms of Hermite coefficients of the transformation function – an idea that was utilised in Janicki and McElroy (2016) to model marginal quantiles. Here this technique is used to derive analytical expressions for predictors, along with the MSE; the solution is given as an explicit function of the Hermite coefficients of the transformation function, the various Hermite polynomials evaluated at the linear predictor, and further weights explicitly determined from the mean squared error of the linear predictor.

These results are general, in the sense that they can be applied to diverse contexts in statistics, such as linear models, spatial statistics, and multivariate analysis. But our applications are focussed on time series, and in particular on forecasting problems. In forecasting, the data vector $\underline{X} = [X_1, \ldots, X_T]'$ is a sample of size T from a time series $\{X_t\}$, and $Y = X_{T+1}$ represents the next unobserved value of the process. Backcasting involves setting $Y = X_0$, and missing value problems can similarly be addressed by letting $Y = X_t$ and $\underline{X} = [X_1, \ldots, X_{t-1}, X_{t+1}, \ldots, X_T]'$; see McElroy and McCracken (2017) for a recent treatment. These facets of the general methodology are developed in Section 2, and numerical comparisons are given in Section 3. The proofs are in an Appendix.

2. Nonlinear prediction

Our goal is to compute the minimal mean squared error (MSE) estimate of Y given data $\underline{X} = [X_1, \ldots, X_T]'$

CONTACT Tucker McElroy ✉ tucker.s.mcelroy@census.gov

[1] Other loss functions could of course be envisioned (mean absolute loss yields the conditional median, for example).

(finite-sample case) or $\underline{X} = \{X_s : s \leq T\}$ (semi-infinite sample case). This estimate is the conditional expectation $\mathbb{E}[Y \,|\, \underline{X}]$, denoted $\widehat{Y}$ for short. We presume that there exists an invertible function f such that $Z_t = f(X_t)$ yields a Gaussian process $\{Z_t\}$. Moreover, $f(Y)$ is a Gaussian random variable whose joint distribution with the process $\{Z_t\}$ is known. In the case of forecasting, backcasting, or missing value imputation, $f(Y) = Z_t$ for some t.

In the following development, it is important to impose that $f(Y)$ be standard normal, although this would rarely be the case if f is obtained by exploratory analysis. (For example, if $Y = X_{T+1}$ and $f(x) = \log(x)$ by exploratory analysis, it would rarely be the case that $\log(X_{T+1})$ would have unit variance.) Let $Z_\star$ denote the standardisation of $f(Y)$, i.e., $Z_\star = (f(Y) - \mathbb{E}[f(Y)])/\sqrt{\mathrm{Var}[f(Y)]}$. We will define g as the inverse map from $Z_\star$ to Y, so that

$$g(x) = f^{-1}(x\sqrt{\mathrm{Var}[f(Y)]} + \mathbb{E}[f(Y)]).$$

The mapping $Y = g(Z_\star)$ allows us to obtain a Hermite expansion of g; if the marginal distribution of $Z_\star$ were non-Gaussian, we could instead have recourse to the Appell polynomials (Varma, 1951).

The main result of the paper is an expression for $\mathbb{E}[Y \,|\, \underline{X}]$ in terms of $\mathbb{E}[Z_\star \,|\, \underline{Z}]$, with $\underline{Z} = [Z_1, \ldots, Z_T]'$. This is of interest because such a Gaussian conditional expectation has a well-known linear formula; see Chapter 4 of McElroy and Politis (2020). In applications, one might transform the data by applying f, then model the Gaussian process, and finally compute $\mathbb{E}[Z_\star \,|\, \underline{Z}]$ by plugging into the linear formulas. Then the main formula of this paper can be used to obtain $\mathbb{E}[Y \,|\, \underline{X}]$ by inserting $\mathbb{E}[Z_\star \,|\, \underline{Z}]$ and its MSE, denoted by V, which would also be available in applications. In particular,

$$V = \mathbb{E}\left[\left(Z_\star - \mathbb{E}[Z_\star \,|\, \underline{Z}]\right)^2 \right] \qquad (1)$$

and is given by formulas in McElroy and Politis (2020); also, see (11) below. We define the space $\mathbb{L}_2(\mathrm{d}\Phi)$ (where Φ is the standard normal cumulative distribution function) as all functions that are square integrable with respect to the measure $\mathrm{d}\Phi$. An inner product on this space is defined via $\langle f, h \rangle = \int_{-\infty}^{\infty} f(x) h(x) \phi(x)\, \mathrm{d}x$, where ϕ is the standard normal probability density function. Hence, we can also write $\langle f, g \rangle = \mathbb{E}[f(W)\ g(W)]$ for $W \sim \mathcal{N}(0,1)$. It follows that we can do a Hermite expansion on g (see Janicki & McElroy, 2016 for background):

$$g(x) = \sum_{k=0}^{\infty} J_k H_k(x) \qquad (2)$$

with H_k the normalised Hermite polynomials (Roman, 1984), and $J_k = \langle g, H_k \rangle$ the Hermite coefficients. The Hermite polynomials are defined via

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

for $k \geq 0$, and form a complete orthonormal system. (Hence, the coefficients J_k tend to zero as $k \to \infty$.) Plugging $Z_\star$ into (2) yields $Y = \sum_{k=0}^{\infty} J_k H_k(Z_\star)$, and applying the conditional expectation operator (which is linear) yields

$$\widehat{Y} = \mathbb{E}[Y \,|\, \underline{X}] = \sum_{k=0}^{\infty} J_k \mathbb{E}[H_k(Z_\star) \,|\, \underline{X}]. \qquad (3)$$

Evidently, the nonlinear predictor can be computed in terms of conditional expectations of $H_k(Z_\star)$, and the formula is summarised in our main theorem below. The proof relies upon the Hermite generating function

$$h(x,t) = \exp\{xt - t^2/2\} = \sum_{k=0}^{\infty} \frac{t^k}{\sqrt{k!}} H_k(x). \qquad (4)$$

(The second equality follows from the definition of the Hermite polynomials, and is discussed in Roman (1984).) From (4) we see that

$$\partial_t^k h(x,t)|_{t=0} = \sqrt{k!} H_k(x). \qquad (5)$$

Therefore, $J_k = \mathbb{E}[g(W)\ \ H_k(W)] = \frac{1}{\sqrt{k!}} \partial_t^k \mathbb{E}[g(W) h(W,t)]|_{t=0}$. We next discuss the optimal predictor.

Theorem 2.1: *Suppose that $Z_\star$ is standard normal, and $Y = g(Z_\star)$ with g given by* (2), *and $\widehat{Z}_\star = \mathbb{E}[Z_\star \,|\, \underline{Z}]$ is the linear prediction in the transformed space, with MSE given by V in* (1). *Then the MSE optimal nonlinear predictor $\widehat{Y}$ of Y given the data $\underline{X}$ is*

$$\mathbb{E}[Y \,|\, \underline{X}] = \sum_{k=0}^{\infty} \frac{J_k}{\sqrt{k!}} \sum_{\ell=0}^{k} \binom{k}{\ell} \sqrt{\ell!} H_\ell(\widehat{Z}_\star) \kappa_{k-\ell}, \qquad (6)$$

where κ_ℓ is the ℓth moment of a Gaussian variable of variance V, i.e., κ_ℓ equals 1 if $\ell = 0$, and equals zero if ℓ is odd and equals $\sqrt{V}^{\ell} (\ell - 1)!!$ if ℓ is even. With $\varepsilon = Y - \widehat{Y}$, the optimal prediction MSE is

$$\mathbb{E}[\varepsilon^2] = \sum_{k=1}^{\infty} J_k^2 \left(1 - (1 - V)^k\right). \qquad (7)$$

The optimal predictor (6) can be explicitly computed to any desired level of accuracy, truncating the first summation over k to a desirable level. Note that this result applies to finite-samples, semi-infinite samples, and bi-infinite samples, allowing us to apply (6) to forecasting (from an infinite past) and missing value problems.

Remark 2.1: Since $\sum_{k=0}^{\infty} J_k^2 = \langle g, g \rangle$, we see that the Hermite coefficients are square summable if $g \in \mathbb{L}_2(\mathrm{d}\Phi)$, and the MSE is well-defined, being bounded above by $\sum_{k=1}^{\infty} J_k^2 V^k$. (Note that $V \leq 1$ because $Z_\star$ has variance one.) Because $\mathrm{Var}[\widehat{Y}^2] = \mathrm{Var}[Y^2] - \mathbb{E}[\varepsilon^2]$, we see that $\widehat{Y}$ is square integrable, and hence the formula (6) converges.

Remark 2.2: In applications, we typically will know $f(Y) \mid \underline{Z}$ rather than $Z_\star \mid \underline{Z}$, but the latter can be obtained from the former by applying the normalising transform described above, i.e.,

$$\mathbb{E}[Z_\star \mid \underline{Z}] = (\mathbb{E}[f(Y) \mid \underline{Z}] - \mathbb{E}[f(Y)])/\sqrt{\mathrm{Var}[f(Y)]}$$
$$V = \mathbb{E}[(f(Y) - \mathbb{E}[f(Y) \mid \underline{Z}])^2]/\mathrm{Var}[f(Y)].$$

Remark 2.3: To apply Theorem 2.1 we must obtain the Hermite coefficients J_k. Either g is known (obtained by exploratory analysis) or estimated (see Janicki & McElroy, 2016), and then the Hermite coefficients can be obtained via Monte Carlo simulation:

$$J_k = \langle g, H_k \rangle = \mathbb{E}[g(W)\, H_k(W)] \approx M^{-1} \sum_{m=1}^{M} g(W_m) H_k(W_m)$$

for $W_1, \ldots, W_m$ i.i.d. standard normal. Another approach to computation involves the generating function:

$$\begin{aligned} J_k &= \frac{1}{\sqrt{k!}} \partial_t^k \mathbb{E}[g(W) \quad h(W,t)] \Big|_{t=0} \\ &= \frac{1}{\sqrt{2\pi\, k!}} \partial_t^k \int g(w) \exp\{wt - t^2/2\} \\ &\quad \times \exp\{-w^2/2\}\, \mathrm{d}w \Big|_{t=0} \\ &= \frac{1}{\sqrt{2\pi\, k!}} \partial_t^k \int g(w) \exp\{-(w-t)^2/2\}\, \mathrm{d}w \Big|_{t=0} \\ &= \frac{1}{\sqrt{2\pi\, k!}} \partial_t^k \int g(w+t) \exp\{-w^2/2\}\, \mathrm{d}w \Big|_{t=0} \\ &= \frac{1}{\sqrt{k!}} \partial_t^k \mathbb{E}[g(W+t)] \Big|_{t=0} = \frac{1}{\sqrt{k!}} \mathbb{E}[g^{(k)}(W)]. \end{aligned}$$

The last expression denotes the k-fold derivative of the function. In cases where g is explicitly known, this can be an easier approach to getting the Hermite coefficients.

In the following examples we suppose that $\widehat{Z}_\star = \mathbb{E}[Z_\star \mid \underline{Z}]$ and $V = \mathbb{E}[(\widehat{Z}_\star - Z_\star)^2]$ are known and available to the practitioner; we present various cases of transforms f, and apply the results of Theorem 2.1.

Example 2.1 (Gaussian): A simple affine transformation $g(x) = \sigma\, x + \mu$ ensures that Y is still Gaussian, with mean μ and variance σ^2. In this case $J_0 = \mu$ and $J_1 = \sigma$, and we more simply have $\mathbb{E}[Y \mid \underline{X}] = \sigma\, \widehat{Z}_\star + \mu$.

Example 2.2 (Lognormal): Suppose $g(x) = \mathrm{e}^x$; applying the generating function method of Remark 2.3, we obtain

$$J_k = \frac{1}{\sqrt{k!}} \mathbb{E}[\exp\{W\}] = \frac{\mathrm{e}^{1/2}}{\sqrt{k!}}$$

for all $k \geq 0$. This can be utilised in (3), together with (5), to yield

$$\begin{aligned} \mathbb{E}[Y \mid \underline{X}] &= \mathrm{e}^{1/2} \sum_{k=0}^{\infty} \frac{1}{k!} \partial_t^k \left(h(\widehat{Z}_\star, t)\, \mathrm{e}^{Vt^2/2} \right) \Big|_{t=0} \\ &= \mathrm{e}^{1/2} \left(h(\widehat{Z}_\star, t)\, \mathrm{e}^{Vt^2/2} \right) |_{t=1} \\ &= \mathrm{e}^{1/2} \exp\{\widehat{Z}_\star + (V-1)/2\} \\ &= \exp\{\widehat{Z}_\star + V/2\}, \end{aligned}$$

which corresponds to the result of McElroy (2010). Applying Theorem 2.1 to compute the optimal MSE, we see that $\mathbb{E}[(\widehat{Y} - Y)^2] = \mathrm{e}^2(1 - \mathrm{e}^{-V})$.

Example 2.3 (Uniform): For $\{Z_t\}$ with standard normal marginal, set $g = \Phi$, so that $\{X_t\}$ has a marginal distribution that is uniform on $(0, 1)$. Then by Remark 2.3, we find that $g^{(k)}(x) = \phi^{(k-1)}(x)$, and hence

$$\begin{aligned} J_k &= \frac{1}{\sqrt{k!}}, \mathbb{E}[\phi^{(k-1)}(W)] \\ &= k^{-1/2}(-1)^{k-1} \mathbb{E}[H_{k-1}(W)\, \phi(W)], \end{aligned}$$

where $W \sim \mathcal{N}(0, 1)$.

Example 2.4 (Logistic): Consider a logistic transform given by $g(x) = \mathrm{e}^x/(1 + \mathrm{e}^x)$. The first few derivatives of g are

$$\begin{aligned} g^{(1)}(x) &= \mathrm{e}^x(1 + \mathrm{e}^x)^{-2} \\ g^{(2)}(x) &= (\mathrm{e}^x - \mathrm{e}^{2x})(1 + \mathrm{e}^x)^{-3} \\ g^{(3)}(x) &= (\mathrm{e}^x - 4\,\mathrm{e}^{2x} + \mathrm{e}^{3x})(1 + \mathrm{e}^x)^{-4}. \end{aligned}$$

By Monte Carlo, we obtain $J_0 = 0.500, J_1 = 0.207, J_2 = 0$, and $J_3 = -.025$.

Example 2.5 (Square): With $g(x) = x^2$, we have $J_0 = 1, J_1 = 0, J_2 = \sqrt{2}$, and $J_k = 0$ for $k > 2$. Then the optimal nonlinear predictor is $\mathbb{E}[Y \mid \underline{X}] = \widehat{Z}_\star^2 + V$. It is simple to check that the error is

$$\epsilon = (Z_\star - \widehat{Z}_\star)(Z_\star + \widehat{Z}_\star) - V,$$

which has mean zero and is independent of all functions of the data. The prediction MSE is $4V - 2V^2$.

3. Comparing linear and nonlinear prediction

It is of interest to understand how much benefit nonlinear prediction provides. Clearly, if g is affine (Example 2.1) then the minimal MSE is equal to $\sigma^2 V$, the same as the linear prediction, but we can expect gains to the degree that g differs from the affine case.

Remark 3.1: A related nonlinear predictor that is sometimes used in applications is defined via $\widetilde{Y} = g(\widehat{Z})$, but unfortunately this estimator can be biased. Following the same arguments used in the proof of Theorem 2.1,

$$\begin{aligned} Y - \widetilde{Y} &= \sum_{k=0}^{\infty} \frac{J_k}{\sqrt{k!}} \partial_t^k \left(h(Z,t) - h(\widehat{Z}_\star, t) \right) \Big|_{t=0} \\ &= \sum_{k=0}^{\infty} \frac{J_k}{\sqrt{k!}} \left(\exp\{\widehat{Z}_\star t - t^2/2\} \right. \\ &\quad \times \left. \left[\exp\{(Z - \widehat{Z}_\star t\} - 1 \right] \right) \Big|_{t=0}, \end{aligned}$$

so that the expectation of the quantity in parentheses is

$$\mathbb{E}[\exp\{\widehat{Z}_\star t - t^2/2\}] \left[\exp\{V t^2/2\} - 1 \right] > 0.$$

Hence there is no guarantee that the bias is zero.

More properly, a comparison can be made to the best linear predictor. When $\{Z_t\}$ is a strictly stationary time series, then $\{X_t\}$ is as well, and we can in some cases determine the best linear estimator's MSE for comparison. Note that in this special case, the mean and variance of each variable Z_t is the same, and hence without any loss of generality we may assume that $\{Z_t\}$ is standardised, i.e., each Z_t is standard normal. Because $X_t = g(Z_t)$, it follows from Taniguchi and Kakizawa (2000, p. 319) that $\mathbb{E}[X_t] = J_0$ and

$$\gamma_X(h) = \sum_{k=0}^{\infty} J_k^2 \gamma_Z(h)^k - J_0^2 = \sum_{k=1}^{\infty} J_k^2 \gamma_Z(h)^k, \qquad (8)$$

where $\{\gamma_X(h)\}$ and $\{\gamma_Z(h)\}$ are the autocovariance sequences of $\{X_t\}$ and $\{Z_t\}$, respectively. So in principle we can understand the second order structure of $\{X_t\}$ in terms of the Hermite coefficients and the original autocovariances.

Suppose further that we are interested in one-step ahead forecasting from a sample of size T. The best linear predictor is obtained by solving the Yule-Walker equations in $\{\gamma_X(h)\}$, and the MSE of such is given by

$$\gamma_X(0) - [\gamma_X(1), \ldots, \gamma_X(T)] \Gamma_X^{-1} [\gamma_X(1), \ldots, \gamma_X(T)]', \qquad (9)$$

where Γ_X is the T-dimensional Toeplitz covariance matrix with jkth entry $\gamma_X(j-k)$. We know that such an MSE must be greater than the minimal MSE provided in Theorem 2.1, with equality occuring only in the case that a linear estimator is globally optimal (e.g., the time series is linear, or is Gaussian). If instead we are forecasting from an infinite past, then the MSE of the linear predictor is the innovation variance σ^2 given by Kolmogorov's formula:

$$\sigma^2 = \exp\left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \log f_X(\lambda)\, \mathrm{d}\lambda \right\}. \qquad (10)$$

Here, $f_X(\lambda) = \sum_{h=-\infty}^{\infty} \gamma_X(h)\, \mathrm{e}^{-i\lambda h}$ is the spectral density of $\{X_t\}$. Therefore, for either a finite sample or for an infinite past, we can determine the linear predictor MSE for a stationary process $\{X_t\}$ by first computing $\gamma_X(h)$ from $\gamma_Z(h)$ via (8), followed by application of (9) or (10) as each case requires. As for the best (nonlinear) predictor, its MSE is given by (7) of Theorem 2.1, where

$$\begin{aligned} V = \gamma_Z(0) - [\gamma_Z(1), \ldots, \gamma_Z(T)] \Gamma_Z^{-1} \\ \times [\gamma_Z(1), \ldots, \gamma_Z(T)]' \end{aligned} \qquad (11)$$

is the analogue of (9) for the $\{Z_t\}$ process.

We provide an illustration in the case of an MA(1) process with various values of θ, and sample size $T = 100$. The innovation variance is set equal to $(1+\theta^2)^{-1}$ so that $\gamma_Z(0) = 1$, as required by the above discussion. For the MA(1) process with $T = 100$, the value of V given by (11) is the same (up to the fourth decimal place) as the innovation variance $(1+\theta^2)^{-1}$. For transformations, we study $g(x) = x^2$, $g(x) = \mathrm{e}^x$, and the logistic. Observe that from (8) the process $\{X_t\}$ will be m-dependent if $\{Z_t\}$ is (although the converse need not be true). Hence, if we obtained a sample from $\{X_t\}$ it would likely be identified with an MA(q) model, and the parameter estimates (e.g., obtained using a Whittle likelihood, which is valid for non-Gaussian processes so long as the cumulants are summable) would likely converge to those corresponding to the spectral factorisation of f_X. Thus, our illustration provides an accurate rendition of the prediction MSE one would obtain in the case of linear or nonlinear predictors, only with the impact of parameter estimation error completely removed (Tables 1–3).

Table 1. MSE for linear and non-linear predictors applied to a squared MA(1) process of parameter θ.

θ	0	.2	.4	.6	.8
V	1.000	0.9615	0.8621	0.7353	0.6098
Linear	2.000	1.9973	1.9713	1.9211	1.8795
Non-linear	2.000	1.9970	1.9620	1.8599	1.6954

Table 2. MSE for linear and non-linear predictors applied to an exponential MA(1) process of parameter θ.

θ	0	.2	.4	.6	.8
V	1.000	0.9615	0.8621	0.7353	0.6098
Linear	4.6708	4.5985	4.3851	4.1192	3.9269
Non-linear	4.6708	4.5642	4.2688	3.8470	3.3732

Table 3. MSE for linear and non-linear predictors applied to a logistic MA(1) process of parameter θ.

θ	0	.2	.4	.6	.8
V	1.000	0.9615	0.8621	0.7353	0.6098
Linear	0.0433	0.0417	0.0375	0.0323	0.0274
Non-linear	0.0433	0.0417	0.0374	0.0320	0.0266

In each case, the degree of benefit to nonlinear prediction increases with θ, as to be expected; however, there are large discrepancies between the three functions. When $\theta = .8$, the logistic transformation offers only a 3% improvement with nonlinear prediction, indicating that linear prediction is almost just as good as the conditional expectation. With $g(x) = x^2$ the analogous improvement is 11%, and is 16% for $g(x) = \mathrm{e}^x$, indicating some real benefit to nonlinear prediction.

We end with a remark on how a confidence interval can be constructed using simulations. Let the formula given by (6) be denoted $h(\widehat{Z}_\star)$. Then the optimal prediction error can be written

$$\varepsilon = g(\widehat{Z}_\star + \delta) - h(\widehat{Z}_\star),$$

where $\delta = Z_\star - \widehat{Z}_\star$ is the (linear) prediction error for the Gaussian variable, and is uncorrelated with (and hence independent of) $\widehat{Z}_\star$. Also $\delta \sim \mathcal{N}(0, V)$. Therefore, in cases where it is easy to simulate $\widehat{Z}_\star$ (e.g., suppose we are forecasting from a Gaussian ARMA process) we can independently draw δ and compute ε for repeated Monte Carlo draws, thereby obtaining a confidence interval for Y.

Disclaimer

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.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Tucker McElroy is Senior Time Series Mathematical Statistician at the U.S. Census Bureau.

Srinjoy Das is a Postdoctoral Scholar in the Mathematics department at the University of California, San Diego.

References

Brockett, P. L., Hinich, M. J., & 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. https://doi.org/10.1080/01621459.1988.10478645

Brockwell, P. J., & Davis, R. A. (2013). *Time series: Theory and methods*. Springer Science & Business Media.

Janicki, R., & McElroy, T. (2016). Hermite expansion and estimation of monotonic transformations of Gaussian data. *Journal of Nonparametric Statistics*, *28*(1), 207–234. https://doi.org/10.1080/10485252.2016.1139880

Maravall, A. (1983). An application of nonlinear time series forecasting. *Journal of Business & Economic Statistics*, *1*(1), 66–74. https://doi.org/10.1080/07350015.1983.10509325

McElroy, T. (2010). A nonlinear algorithm for seasonal adjustment in multiplicative component decompositions. *Studies in Nonlinear Dynamics and Econometrics*, *14*(4). Article 6. https://doi.org/10.2202/1558-3708.1756

McElroy, T. (2016). On the measurement and treatment of extremes in time series. *Extremes*, *19*(3), 467–490. https://doi.org/10.1007/s10687-016-0254-4

McElroy, T., & McCracken, M. (2017). Multi-step ahead forecasting of vector time series. *Econometric Reviews*, *36*(5), 495–513. https://doi.org/10.1080/07474938.2014.977088

McElroy, T., & Politis, D. (2020). *Time series: A first course with bootstrap starter*. Chapman Hall.

Roman, S. (1984). *The umbral calculus*. Academic Press.

Taniguchi, M., & Kakizawa, Y. (2000). *Asymptotic theory of statistical inference for time series*. Springer.

Varma, R. S. (1951). On Appell polynomials. *Proceedings of the American Mathematical Society*, *2*(4), 593–596. https://doi.org/10.1090/S0002-9939-1951-0042547-5

Appendix

Proof of Theorem 2.1: From (3) and (5) we obtain

$$\mathbb{E}[H_k(Z_\star) \,|\, \underline{X}] = \frac{1}{\sqrt{k!}} \partial_t^k \mathbb{E}[h(Z_\star, t) \,|\, \underline{X}] \Big|_{t=0}.$$

We can write $Z_\star \,|\, \underline{X} \sim \mathcal{N}(\widehat{Z}_\star, V)$, and using the property that $Z_\star - \widehat{Z}_\star$ is independent of all functions of the data, we obtain

$$\begin{aligned} \mathbb{E}[h(Z_\star, t) \,|\, \underline{X}] &= \mathbb{E}\left[\exp\{Z_\star t - \widehat{Z}_\star t\} \exp\{\widehat{Z}_\star t - t^2/2\} \,|\, \underline{X}\right] \\ &= \mathbb{E}[\exp\{Z_\star t - \widehat{Z}_\star t\} \,|\, \underline{X}] \exp\{\widehat{Z}_\star t - t^2/2\} \\ &= \exp\{V\, t^2/2\} h(\widehat{Z}_\star, t). \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}[H_k(Z_\star) \,|\, \underline{X}] &= \frac{1}{\sqrt{k!}} \partial_t^k \left(h(\widehat{Z}_\star, t)\, \mathrm{e}^{V\, t^2/2} \right) \Big|_{t=0} \\ &= \frac{1}{\sqrt{k!}} \sum_{\ell=0}^{k} \binom{k}{\ell} \partial_t^\ell h(\widehat{Z}_\star, t) \partial_t^{k-\ell}\, \mathrm{e}^{V\, t^2/2}|_{t=0} \\ &= \frac{1}{\sqrt{k!}} \sum_{\ell=0}^{k} \binom{k}{\ell} \sqrt{\ell!} H_\ell(\widehat{Z}_\star) \kappa_{k-\ell}. \end{aligned}$$

The prediction error is

$$\begin{aligned} \varepsilon = Y - \mathbb{E}[Y \,|\, \underline{X}] &= \sum_{k=0}^{\infty} J_k \left(H_k(Z_\star) - \mathbb{E}[H_k(Z_\star) \,|\, \underline{X}] \right) \\ &= \sum_{k=1}^{\infty} \frac{J_k}{\sqrt{k!}} \partial_t^k \left(h(Z_\star, t) - \mathbb{E}[h(Z_\star, t) \,|\, \underline{X}] \right) \Big|_{t=0} \\ &= \sum_{k=0}^{\infty} \frac{J_k}{\sqrt{k!}} \partial_t^k \left(h(\widehat{Z}_\star, t) \right. \\ &\quad \cdot \left. \left[\exp\{(Z_\star - \widehat{Z}_\star)t\} - \exp\{V\, t^2/2\} \right] \right) \Big|_{t=0}. \end{aligned}$$

Note that $Z_\star - \widehat{Z}_\star$ is orthogonal to all linear functions of the data; because this error is Gaussian, it is moreover independent of all functions of the data $\underline{X}$. It follows that $\mathbb{E}[\varepsilon] = 0$, because $\mathbb{E}[\exp\{(Z_\star - \widehat{Z}_\star)t\}] = \exp\{V\,t^2/2\}$. Moreover, for any function $\ell(\underline{X})$ of the data,

$$\begin{aligned}\mathbb{E}[\epsilon\;\ell(\underline{X})] = \sum_{k=0}^{\infty} \frac{J_k}{\sqrt{k!}} \partial_t^k \Big(& h(\widehat{Z}_\star, t)\ell(\underline{X}) \\ & \cdot \mathbb{E}\Big[\exp\{(Z_\star - \widehat{Z}_\star)t\} \\ & - \exp\{V\,t^2/2\}\Big]\Big)\Big|_{t=0} = 0.\end{aligned}$$

This verifies optimality. To compute the MSE, first observe that

$$\begin{aligned}\varepsilon^2 = \sum_{j,k=1}^{\infty} \frac{J_j\,J_k}{\sqrt{j!}\sqrt{k!}} \partial_s^j \partial_t^k \Big(& \exp\{\widehat{Z}_\star(s+t) - (s^2+t^2)/2\} \\ & \cdot \Big[\exp\{(Z_\star - \widehat{Z}_\star)(s+t)\} \\ & - \exp\{V\,t^2/2 + (Z_\star - \widehat{Z}_\star)s\} \\ & - \exp\{V\,s^2/2 + (Z_\star - \widehat{Z}_\star)t\} \\ & + \exp\{V(s^2+t^2)/2\}\Big]\Big)|_{s,t=0}.\end{aligned}$$

Note that $\mathbb{E}[\widehat{Z}_\star] = \mathbb{E}[Z_\star] = 0$, because $Z_\star$ is standard normal. Moreover, due to orthogonality, $Z_\star = (Z_\star - \widehat{Z}_\star) + \widehat{Z}_\star$ with the two summands orthogonal, and hence $1 = \mathbb{E}[Z_\star^2] = V + \mathbb{E}[\widehat{Z}_\star{}^2]$. Using these facts and again using the independence property, we take the expectation of ε^2 and obtain

$$\begin{aligned}\mathbb{E}[\varepsilon^2] &= \sum_{j,k=1}^{\infty} \frac{J_j\,J_k}{\sqrt{j!}\sqrt{k!}} \partial_s^j \partial_t^k \Big(\exp\{\mathbb{E}[\widehat{Z}_\star](s+t) \\ &\quad + \mathbb{E}[\widehat{Z}_\star{}^2](s+t)^2/2 - (s^2+t^2)/2\} \\ &\quad \cdot \Big(\exp\{V(s+t)^2/2\} - \exp\{V(s^2+t^2)/2\}\Big)\Big)|_{s,t=0} \\ &= \sum_{j,k=1}^{\infty} \frac{J_j\,J_k}{\sqrt{j!}\sqrt{k!}} \partial_s^j \partial_t^k \\ &\quad \times \Big(\exp\{(1-V)(s+t)^2/2 - (s^2+t^2)/2\} \\ &\quad \cdot \Big(\exp\{V(s+t)^2/2\} - \exp\{V(s^2+t^2)/2\}\Big)\Big)|_{s,t=0} \\ &= \sum_{j,k=1}^{\infty} \frac{J_j\,J_k}{\sqrt{j!}\sqrt{k!}} \partial_s^j \partial_t^k \Big(\exp\{st\}\cdot\Big(1 - \exp\{-V\,st\}\Big)\Big)\Big|_{s,t=0}.\end{aligned}$$

Now, it is straight-forward to show that for any constant A,

$$\partial_s^j \partial_t^k \exp\{A\,st\}|_{s,t=0} = \begin{cases} A^k\,k! & \text{if } j = k \\ 0 & \text{else.} \end{cases}$$

Applying this with $A = 1$ and $A = -V$ yields

$$\begin{aligned}\mathbb{E}[\varepsilon^2] &= \sum_{k=1}^{\infty} \frac{J_k^2}{k!} \left(1 - (1-V)^k\right) k! \\ &= \sum_{k=1}^{\infty} J_k^2 \left(1 - (1-V)^k\right).\end{aligned}$$

■