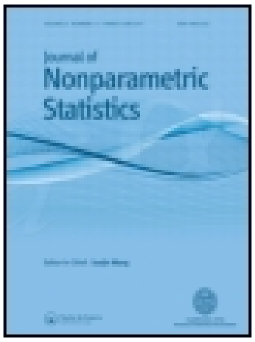

Taylor & Francis
Taylor & Francis Group

Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: http://www.tandfonline.com/loi/gnst20

Hermite expansion and estimation of monotonic transformations of Gaussian data

Ryan Janicki & Tucker S. McElroy

To cite this article: Ryan Janicki & Tucker S. McElroy (2016): Hermite expansion and estimation of monotonic transformations of Gaussian data, Journal of Nonparametric Statistics, DOI: 10.1080/10485252.2016.1139880

To link to this article: http://dx.doi.org/10.1080/10485252.2016.1139880

Published online: 27 Jan 2016.

Journal of Nonparametric Statistics, 2016
http://dx.doi.org/10.1080/10485252.2016.1139880

Taylor & Francis
Taylor & Francis Group

Hermite expansion and estimation of monotonic transformations of Gaussian data

Ryan Janicki* and Tucker S. McElroy

Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233-9100, USA

(*Received 16 June 2014; accepted 11 October 2015*)

This paper describes a semiparametric method for estimating a generic probability distribution using a basis expansion in L^2. We express the given distribution as a monotonic transformation of the Gaussian cumulative distribution function, expanded in a basis of Hermite polynomials. The coefficients in the basis expansion are functionals of the quantile function, and can be consistently estimated to give a smooth estimate of the transformation function. For situations in which the estimated function is not monotone, a projection approach is used to adjust the estimated transformation function to guarantee monotonicity. Two applications are presented which focus on the analysis of model residuals. The first is a data example which uses the residuals from the 2012 Small Area Income and Poverty Estimates model. The Hermite estimation method is applied to these residuals as a graphical method for detection of departures from normality and to construct credible intervals. The second example analyses residuals from time series models for the purpose of estimating the variance of the mean and median and comparing the results to the AR-sieve. This paper concludes with a set of numerical examples to illustrate the theoretical results.

Keywords: Hermite polynomials; small area; time series; AR-sieve; SAIPE; Hilbert space

AMS Subject Classification: 62G05; 62G07

1. Introduction

The Hermite polynomials have been a popular tool for approximation theory in fields such as probability, numerical analysis, and physics, due to the fact that this set of functions forms a complete orthonormal basis for functions with unbounded domains. This property allows for an expansion of any function in L^2 using the Hermite polynomials as basis functions.

In the statistical literature, orthogonal series have been used extensively for density estimation, notably by Schwartz (1967), who gave an expansion and simple moment-based estimators of the coefficients in the series for consistent density estimates in the integrated mean squared error sense. Madan and Milne (1994) and Jondeau and Rockinger (2001) used Hermite polynomial density estimates for perturbations of a standard Gaussian density to account for excess skewness and kurtosis for estimation of risk-neutral densities in their options pricing modes. Polynomial expansions are not restricted to the Hermite family; for example, Babu, Canty, and

*Corresponding author. Email: ryan.janicki@census.gov

Chaubey (2002) used Bernstein polynomials to obtain smooth cumulative distribution function and probability density function estimates for distributions with support on [0, 1].

The Hermite expansion has been used to understand the asymptotic behaviour of statistics – see Taqqu (1975). Recently, the Hermite expansion has been used to develop robust tests for normal distributional assumptions. Puuronen and Hyvärinen (2011) used Hermite polynomials for density approximation, and used estimated coefficients to measure the degree of non-Gaussianity in a sample, which is an important problem in the theory of independent component analysis (Hyvärinen and Oja 2000). Similar work is that of Bontemps and Meddahi (2005, 2012), who used Stein's lemma and the orthogonality of the Hermite polynomials to develop moment-based test statistics.

In this paper we represent the observed data as a monotonic transformation of standard Gaussian random variables, and expand this transformation function using a basis expansion. This is quite general, as the transformation function can be written as the composition of the quantile function and the Gaussian cumulative distribution function for any continuous population. The coefficients in the basis expansion of the transformation that we derive can be expressed as a functional of the theoretical quantile function. Using the empirical quantile function (eqf) as a plug-in estimator leads to an L-estimator, that is, a linear combination of order statistics. There is much literature on the properties and uses of L-estimators; see, for example, Mason and Shorack (1992) and David and Nagaraja (2003). In this paper, however, we focus on properties of the eqf for deriving consistent estimates for the Hermite coefficients in the expansion. This setup seems not to have been previously considered, and we find it to be a useful method for giving a semiparametric description of non-Gaussian distributions.

Closely related work is that of Menéndez, Ghosh, Künsch, and Tinner (2013), who modelled the errors in a nonparametric time series regression model as an unknown monotone transformation of an underlying latent Gaussian process, expanded in Hermite polynomials. In contrast to the work in this paper, Menéndez et al. (2013) use kernel smoothing to estimate the time trend function and to recover latent Gaussian process. They noted the important role of the first coefficient in the basis expansion in bandwith selection, and gave a consistent estimate of the coefficient for deriving a data-driven bandwidth in their kernel estimator of a time trend.

Alternatively, a parametric form of the marginal distribution could be specified. But the semiparametric form considered here is quite general, and by taking a larger truncation order in the series expansion, misspecification can be avoided, or at least have its impact diminished. Moreover, having identified the marginal distribution as a monotonic transformation of the Gaussian marginal allows for straightforward simulation of the data while allowing for atypical structures, in terms of skewness and kurtosis.

This paper is organised as follows: in Section 2.1, background on the properties of the Hermite polynomials and methodology for estimating a monotone transformation function based on an observed set of independent, identically distributed data is provided, and the statistical properties of this estimate are derived. Using empirical process theory, the asymptotic behaviour of the estimated coefficients is shown in Theorem 2.1, and consistent variance estimates of the estimated coefficients are given in Theorem 2.2. Section 2.2 extends the results to the case of stationary time series data. Section 3 describes a projection approach for modifying the nonparametric estimate of the transformation function to guarantee monotonicity for the case when either the estimate or the truncation of the transformation function is non-monotone, and the asymptotic behaviour of the projection estimate is given in Theorem 3.1. In Section 4.1, 2012 Small Area Income and Poverty Estimates (SAIPE) data is analysed using the Hermite estimation methodology, and in Section 4.2, application to time series models are investigated as an alternative to the AR-sieve. Numerical examples are given in Section 5 to illustrate the theoretical results of Sections 2 and 3. Concluding remarks are made in Section 6.

2. Methodology

2.1. *The i.i.d. case*

The normalised Hermite polynomials (see Samorodnitsky and Taqqu 1994) are defined as

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

for $k = 0, 1, 2, \ldots$. The collection of functions $\{H_k(x)\}_{k \geq 0}$ form a complete orthonormal basis for $L^2(d\Phi)$, where $\Phi(x)$ is the cumulative distribution function of the standard Gaussian distribution, so that

$$\int_{-\infty}^{\infty} H_k(x) H_l(x) \phi(x)\, \mathrm{d}x = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{if } k \neq l, \end{cases}$$

where $\phi(x) = (2\pi)^{-1/2} \mathrm{e}^{-x^2/2}$. Any function g satisfying $\int g^2(x)\phi(x)\, \mathrm{d}x < \infty$ therefore has the representation

$$g(x) = \sum_{k=0}^{\infty} J_k H_k(x), \tag{1}$$

where $J_k = \langle g, H_k \rangle$ are the Hermite coefficients, given as an inner product of g with the basis functions. This inner product is defined via $\langle f, h \rangle = \int_{-\infty}^{\infty} f(x) h(x) \phi(x)\, \mathrm{d}x$. Since the $\{H_k\}$ form a complete orthonormal system, the coefficients J_k tend to zero as $k \to \infty$.

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a population with cumulative distribution function F, which is assumed to be invertible, where each X_i is an unknown transformation of a standard Gaussian random variable. The transformation function is denoted by g, so that $X_i = g(Z_i)$, where the Z_i are i.i.d. $N(0, 1)$. For g to be identifiable, we assume that g is a strictly increasing function. For example, if $Z_i, i = 1, 2$ are standard normal random variables, and $Y_1 = g_I(Z_1) = -2 \log(1 - \Phi(Z_1))$ and $Y_2 = g_D(Z_2) = -2 \log(\Phi(Z_2))$, both Y_1 and Y_2 will have χ^2_2 distributions, but g_I is increasing while g_D is decreasing. As a second example, let $Y_1 = g_1(Z_1) = Z_1^2$ and $Y_2 = g_2(Z_2) = F_1^{-1}(\Phi(Z_2))$, where F_1^{-1} is the quantile function for the χ^2_1 distribution; both Y_1 and Y_2 will have χ^2_1 distributions, showing that the assumption of monotonicity is important.

Letting $Q(u) = F^{-1}(u) = \inf\{x : F(x) \geq u\}$ denote the quantile function of X_i, we easily see that $g = Q \circ \Phi$. This paper provides a method for estimating the unique monotonically increasing transformation function g.

If $\int g^2(x)\phi(x)\, \mathrm{d}x < \infty$ we can do a Hermite expansion on g (see Samorodnitsky and Taqqu 1994) so that g has the series representation (1). Let the normal quantile function be denoted Ξ, with derivative ξ (the normal quantile density function, Parzen 1979). It follows that

$$J_k = J_k(Q) = \int_{-\infty}^{\infty} Q(\Phi(x)) H_k(x) \phi(x)\, \mathrm{d}x = \int_0^1 Q(u) H_k(\Xi(u))\, \mathrm{d}u. \tag{2}$$

As noted above, the J_k tend to 0 as $k \to \infty$, so it is reasonable to expect that a model for g can be determined by truncating the infinite summation in Equation (1) at some level m; however it should be noted that the J_k do not necessarily tend to 0 monotonically. For example, if F is a distribution function that is continuous and symmetric around 0, then $J_{2k} = 0$ for all $k = 0, 1, 2, \ldots$, due to the fact that the Gaussian density and the Hermite polynomials H_{2k} are even functions, while $g(x) = Q(\Phi(x))$ is an odd function. That g is odd follows from the fact that $1 - \Phi(x) = F(Q(1 - \Phi(x))) = F(Q(\Phi(-x)))$ and $1 - \Phi(x) = 1 - F(Q(\Phi(x))) = F(-Q(\Phi(x)))$, due to the symmetry of F and Φ.

A nonparametric estimator of the quantile function Q is given by the inverse of the empirical distribution function (edf), which is equivalent to taking order statistics. Denote the edf by $\hat{F}(x) = (1/n)\sum_{i=1}^{n} I\{X_i \leq x\}$ and the inverse of the edf, that is, the eqf, by $\hat{Q}(u) = \hat{F}^{-1}(u) = X_{(k)}$, if $(k-1)/n < u \leq k/n$, $k = 1,\dots,n$, where $X_{(1)} < \cdots < X_{(n)}$ are the order statistics of the random sample $X_1,\dots,X_n$. Plugging the eqf in for Q in Equation (2) yields an estimate of J_k as follows:

$$\hat{J}_k = J_k(\hat{Q}) = \sum_{i=1}^{n} X_{(i)} \int_{\Xi[(i-1)/n]}^{\Xi[i/n]} H_k(x)\phi(x)\,\mathrm{d}x = \sum_{i=1}^{n} X_{(i)} \int_{(i-1)/n}^{i/n} H_k(\Xi(u))\,\mathrm{d}u.$$

This is a weighted sum of the order statistics, or an L-estimator, and the weights can be determined before the data analysis (they only depend on sample size and the normal cdf).

The asymptotic behaviour of this estimate can be described using empirical process theory and weak convergence results for the empirical quantile process. The main difficulty in proving asymptotic results for $\hat{J}_k$ is the divergence the tails of the empirical quantile process, that is, $P(\lim_{n\to\infty}\sup_{0<t<1}|\hat{Q}(t) - Q(t)| = \infty) = 1$, unless F has finite support (Csörgö and Horváth 1993). To overcome this undesirable tail behaviour, one can consider convergence of the quantile process on fixed compact sets, which holds under very weak conditions on the distribution function F (van der Vaart 1998), and then apply a bounded linear functional and the continuous mapping theorem. Alternatively, the processes

$$r_n^w(u) = \sqrt{n}w(u)(\hat{Q}(u) - Q(u))I\{1/(n+1) \leq u \leq n/(n+1)\}$$

for certain weight functions $w(u)$, can be used. Mason (1984) developed conditions on the weight function $w(u)$ and the quantile-density function $q(u)$ which guarantee weak convergence of $r_n^w(u)$ to the process $w(u)q(u)B(u)$. Alternative sets of conditions for which the empirical quantile process converges, as well as further discussion of the properties of the empirical quantile process can be found in Shorack and Wellner (1986) and Csörgö and Horváth (1993).

For $\epsilon > 0$, let $K_\epsilon(f) = \int_\epsilon^{1-\epsilon} f(u)H_k(\Xi(u))\,\mathrm{d}u$, and let $J_k^\epsilon = K_\epsilon(Q)$ and $\hat{J}_k^\epsilon = K_\epsilon(\hat{Q})$.

Theorem 2.1 *Let $X_1,\dots,X_n$ be independent, identically distributed random variables, with common distribution function F. Suppose F is continuously differentiable, with strictly positive derivative f. Then for any fixed $\epsilon > 0$,*

$$\sqrt{n}(\hat{\boldsymbol{J}}_m^\epsilon - \boldsymbol{J}_m^\epsilon) \overset{\mathcal{L}}{\Longrightarrow} \int_\epsilon^{1-\epsilon} q(u)B(u)\boldsymbol{H}_m(\Xi(u))\,\mathrm{d}u, \tag{3}$$

where $\hat{\boldsymbol{J}}_m^\epsilon = (\hat{J}_0^\epsilon,\dots,\hat{J}_m^\epsilon)^{\mathrm{T}}\boldsymbol{J}_m^\epsilon = (J_0^\epsilon,\dots,J_m^\epsilon)^{\mathrm{T}}$, and $\boldsymbol{H}_m(x) = (H_0(x),\dots,H_m(x))^{\mathrm{T}}$ is the vector of Hermite polynomials. The limit in Equation(3) *is normal with mean zero and covariance matrix consisting of terms*

$$V_{i,j}^\epsilon = \int_\epsilon^{1-\epsilon}\int_\epsilon^{1-\epsilon} q(u)q(v)H_i(\Xi(u))H_j(\Xi(v))(u \wedge v - uv)\,\mathrm{d}u\,\mathrm{d}v. \tag{4}$$

If, in addition, f is differentiable, and there exists a constant $\gamma > 0$ such that,

$$\sup_{0<t<1} t(1-t)\frac{|f'(Q(t))|}{f^2(Q(t))} \leq \gamma, \tag{5}$$

and for $k = 0,1,\dots,m$,

$$\int_0^1 \frac{H_k(\Xi(t))t^{1/2}(1-t)^{1/2}}{f(Q(t))}\,\mathrm{d}t < \infty, \tag{6}$$

then ϵ can be taken to be $\epsilon_n = 1/n$, and

$$\sqrt{n}(\hat{\boldsymbol{J}}_m^{\epsilon_n} - \boldsymbol{J}_m^{\epsilon_n}) \overset{\mathcal{L}}{\Longrightarrow} \int_0^1 q(u)B(u)\boldsymbol{J}_m(\Xi(u))\,\mathrm{d}u. \tag{7}$$

Proof The condition that F is continuously differentiable with positive density f is sufficient for weak convergence of the quantile process $\sqrt{n}(\hat{Q}(u) - Q(u))$ to $q(u)B(u)$ in $l^\infty[\epsilon, 1-\epsilon]$, for any fixed $\epsilon > 0$ (van der Vaart 1998, Corollary 21.5), where B is a standard Brownian bridge, and $q = 1/f(Q)$ is the derivative of Q, also known as the quantile-density function (Parzen 1979).

The functional $K_\epsilon(f) = \int_\epsilon^{1-\epsilon} f(u)H_k(\Xi(u))\,\mathrm{d}u$ is bounded, and hence continuous, since for any $f \in l^\infty[\epsilon, 1-\epsilon]$,

$$\begin{aligned} |K_\epsilon(f)| &\le \int_\epsilon^{1-\epsilon} |f(u)||H_k(\Xi(u))|\,\mathrm{d}u \\ &\le \sup_{\epsilon \le u \le 1-\epsilon} |f(u)| \sup_{\epsilon \le u \le 1-\epsilon} |H_k(\Xi(u))|(1-2\epsilon) \equiv C(\epsilon, k)\|f\|_\infty. \end{aligned}$$

By the continuous mapping theorem, $K_\epsilon(\sqrt{n}(\hat{Q} - Q)) = \sqrt{n}(\hat{J}_k^\epsilon - J_k^\epsilon) \overset{\mathcal{L}}{\Longrightarrow} K_\epsilon(q(u)B(u)) = \int_\epsilon^{1-\epsilon} q(u)B(u)H_k(\Xi)\,\mathrm{d}u$. The limiting process in Equation (3) can be shown to be normal with mean zero and covariance terms $V_{i,j}$ in Equation (4) by applying Theorems 2.3 and 2.4 of Tanaka (1996).

For ϵ to be allowed to decrease with the sample size, more restrictive conditions on the density f are needed. Condition (5) is sufficient to guarantee the existence of a sequence of Brownian bridges $\{B_n(x)\}$ such that

$$\sup_{1/(n+1) \le u \le n/(n+1)} |\sqrt{n} f(Q(u))(\hat{Q}(u) - Q(u)) - B_n(u)| \xrightarrow{P} 0,$$

which implies weak convergence in $D[0, 1]$, the space of functions on $[0, 1]$ which are right continuous, and whose left-hand limits exist, of $\sqrt{n} f(Q(u))(\hat{Q}(u) - Q(u))I\{1/(n+1) \le u \le n/(n+1)\}$ to a standard Brownian bridge (see Csörgö 1983). The additional condition (6) is needed so that functionals of the quantile process defined on increasing domains can be taken, and allows for application of Theorem 6.4.1 of Csörgö and Horváth (1993), which gives $\int_{1/n}^{1-1/n} \sqrt{n} f(Q(u))(\hat{Q}(u) - Q(u))H_k(\Xi(u))/f(Q(u))\,\mathrm{d}u = \sqrt{n}(\hat{J}_k^{1/n} - J_k^{1/n}) \overset{\mathcal{L}}{\Longrightarrow} \int_0^1 q(u)B(u)H_k(\Xi(u))\,\mathrm{d}u$. ■

In practice, the choice of $\epsilon > 0$ effectively trims the sample; if ϵ is chosen to be less than $1/n$, then the entire sample will be used for estimation of J_k, while if ϵ is chosen to be greater than $1/n$, only $X_{(M)}, \ldots, X_{(n-M)}$, for some $M > 1$ will be used for inference. A larger ϵ will induce some bias in the estimate $\hat{J}_k^\epsilon$ of J_k, but will reduce the variance. For ϵ less than $1/n$, the difference in $\hat{J}_k^\epsilon$ and $\hat{J}_k$ appears through the differences in the coefficients $\int_\epsilon^{1/n} H_k(\Xi(u))\,\mathrm{d}u - \int_0^{1/n} H_k(\Xi(u))\,\mathrm{d}u$ and $\int_{(1-1/n)}^{1-\epsilon} H_k(\Xi(u))\,\mathrm{d}u - \int_{1-1/n}^1 H_k(\Xi(u))\,\mathrm{d}u$ for $X_{(1)}$ and $X_{(n)}$, respectively, so this bias can be effectively removed by choosing ϵ such that these differences are negligible.

In order to make use of formula (4), we must estimate the quantile-density q. There is much literature on the nonparametric and semiparametric estimation of this function (Parzen 1979; Jones 1992; Chen 1995; Chen and Parzen 1997) but here we adopt an approach that is coherent with our Hermite expansion of g in Equation (1). The approach is based on the observation that the quantile-density is integrated, and hence the variance expression can be written in terms of Q, which can be consistently estimated via $\hat{Q}$.

Theorem 2.2 *For $\epsilon > 0$, let $M = \min\{j : j/n > \epsilon\}$. If the conditions of Theorem 2.1 hold, the estimator*

$$\hat{V}^{\epsilon}_{i,j} = \sum_{l=M}^{n-M} \frac{l}{n}\frac{n-l}{n}(X_{(l+1)} - X_{(l)})^2 H_i\left(\Xi\left(\frac{l}{n}\right)\right) H_j\left(\Xi\left(\frac{l}{n}\right)\right)$$
$$+ \sum_{l=M}^{n-M-1}\left\{\frac{l}{n}H_j\left(\Xi\left(\frac{l}{n}\right)\right)(X_{(l+1)} - X_{(l)})\sum_{k=l+1}^{n-M}\frac{n-k}{n}H_i\left(\Xi\left(\frac{k}{n}\right)\right)(X_{(k+1)} - X_{(k)})\right\}$$
$$+ \sum_{l=M+1}^{n-M}\left\{\frac{n-l}{n}H_j\left(\Xi\left(\frac{l}{n}\right)\right)(X_{(l+1)} - X_{(l)})\sum_{k=M}^{l-1}\frac{k}{n}H_i\left(\Xi\left(\frac{k}{n}\right)\right)(X_{(k+1)} - X_{(k)})\right\} \quad (8)$$

is consistent for $V^{\epsilon}_{i,j}$ in Equation (4).

The proof is given in the appendix.

Remark 1 The matrix $\hat{V}^{\epsilon} = \{\hat{V}^{\epsilon}_{i,j}\}_{i,j=1}^{m}$ is positive definite, and the diagonal terms $\hat{V}^{\epsilon}_{i,i}$ simplify to

$$\hat{V}^{\epsilon}_{i,i} = \sum_{l=M}^{n-M} \frac{l}{n}\frac{n-l}{n}(X_{(l+1)} - X_{(l)})^2 H_i\left(\Xi\left(\frac{l}{n}\right)\right)^2$$
$$+ 2\sum_{l=M}^{n-M-1}\left\{\frac{l}{n}H_i\left(\Xi\left(\frac{l}{n}\right)\right)(X_{(l+1)} - X_{(l)})\sum_{k=l+1}^{n-M}\frac{n-k}{n}H_i\left(\Xi\left(\frac{k}{n}\right)\right)(X_{(k+1)} - X_{(k)})\right\}. \quad (9)$$

Equation (9) is a positive definite quadratic form in $H_i(\Xi(k/n))(X_{(k+1)} - X_{(k)})$, so that $\hat{V}^{\epsilon}_{i,i}$ will always be positive for any $\epsilon > 0$.

While obtaining estimates of J_k is of interest in its own right, we are principally interested in obtaining a workable approximation to g. The true function g can be estimated by using the first m estimated Hermite coefficients and the function $\hat{g}^{\epsilon}_m$, given by

$$\hat{g}^{\epsilon}_m(x) = \sum_{k=0}^{m} \hat{J}^{\epsilon}_k H_k(x).$$

The choice of m essentially dictates a ‘model’ for g, although we can take the viewpoint that $m = m(n)$ grows with the sample size, so that the method is semiparametric. The usual trade-offs between small and large choices of m then apply: small m induces a bias in our estimate of g, whereas a large m generates more stochastic terms in our estimate, and hence increases variance. This observation we demonstrate next.

Corollary 2.1 *Under the conditions of Theorem 2.1, for each x, the estimate $\hat{g}^{\epsilon}_m(x)$ of $g^{\epsilon}(x)$ has the following limiting behaviour:*

$$\sqrt{n}(\hat{g}^{\epsilon}_m(x) - g^{\epsilon}(x)) + R^{\epsilon}_m(x) = \sqrt{n}(\hat{g}^{\epsilon}_m(x) - g^{\epsilon}_m(x))$$
$$\overset{\mathcal{L}}{\Longrightarrow} \sum_{k=0}^{m}\int_{\epsilon}^{1-\epsilon} q(u)B(u)H_k(\Xi(u))\,\mathrm{d}u \cdot H_k(x) \quad (10)$$

as $n \to \infty$, where

$$R^{\epsilon}_m(x) = \sqrt{n}\sum_{k>m} J^{\epsilon}_k H_k(x),$$

so that

$$\sqrt{n}(\hat{g}^{\epsilon}_m(x) - g^{\epsilon}_m(x)) \overset{\mathcal{L}}{\Longrightarrow} N(0, V^{\epsilon}_m(x)). \tag{11}$$

The variance $V^{\epsilon}_m(x)$ is $\boldsymbol{H}^{\mathrm{T}}_m(x)\boldsymbol{V}^{\epsilon}\boldsymbol{H}_m(x)$, where $\boldsymbol{H}_m(x) = (H_0(x), \ldots, H_m(x))^{\mathrm{T}}$ is the vector of Hermite polynomials and $\boldsymbol{V}^{\epsilon}$ is the covariance matrix of the limiting Gaussian distribution in Corollary 2.1. *Conditions* (5) *and* (6) *must hold for $\epsilon = \epsilon_n = 1/n$.*

If, in addition, g is $j > 3$ times differentiable, and $\mathrm{d}^j/\mathrm{d}x^j(g(z)\phi(z))\mathrm{e}^{x^2/4}$ and all lower derivatives are in $L^1(\mathbb{R})$, then

$$\sqrt{n}\sum_{k=M+1}^{\infty} J_k H_k(x) = O\left(\frac{n^{1/2}}{m^{j/2-1}}\right). \tag{12}$$

Proof Convergence in Equation (11) follows immediately from Corollary 2.1 and the representation of the estimate of g in Equation (3). Equation (12) can be proved by adapting the proof of Theorem 1 of Boyd (1984). First, notice that

$$\frac{\mathrm{d}}{\mathrm{d}x}H_{n+1}(x) = \sqrt{n+1}H_n(x),$$

and that for all x and k (Abramowitz and Stegun 1965, p. 787),

$$|H_k(x)| \le \kappa \mathrm{e}^{x^2/4},$$

where $\kappa \approx 1.0864$. Using integration by parts,

$$J_k = \int_{-\infty}^{\infty} g(x)H_k(x)\phi(x)\,\mathrm{d}x = -\int_{-\infty}^{\infty}(g(x)\phi(x))'\frac{H_{k+1}(x)}{\sqrt{k+1}}\,\mathrm{d}x$$

so that

$$\begin{aligned}|J_k| &\le (k+1)^{-1/2}\int |(g(x)\phi(x))'|\mathrm{e}^{x^2/4}|H_{k+1}(x)\mathrm{e}^{-x^2/4}|\,\mathrm{d}x\\ &\le \kappa(k+1)^{-1/2}\int_{-\infty}^{\infty}|(g(x)\phi(x))'|\mathrm{e}^{x^2/4}\,\mathrm{d}x = \frac{C_1}{\sqrt{k+1}}.\end{aligned}$$

Integration by parts can be repeated j times, giving $|J_k| \le C_j(k+1)^{-j/2}$, where C_j is a constant independent of k and x. We therefore have

$$\sqrt{n}\left|\sum_{k>m} J_k H_k(x)\right| \le \sqrt{n}\kappa\mathrm{e}^{x^2/4}\sum_{k>m}|J_k| \le \sqrt{n}\kappa\mathrm{e}^{x^2/4}\sum_{k>m}\frac{1}{(m+1)^{j/2}} = O\left(\frac{n^{1/2}}{m^{j/2-1}}\right)$$

for $j > 3$. ■

Since $g^{\epsilon}(x) = g(x)$ for $x \in (\Xi(\epsilon), \Xi(1-\epsilon))$, $\sum_{k=0}^{\infty} J^{\epsilon}_k H_k(x) = \sum_{k=0}^{\infty} J_k H_k(x)$ on this interval. The rate of decay of the deterministic bias term $R^{\epsilon}_m(x)$ can thus be approximated by Equation (12). Equation (12) shows that this bias decreases as $m \to \infty$ and as the smoothness of the function g increases.

The above asymptotics are derived under an assumption of a fixed m. On the right-hand side of the convergence, the coefficient of $H_k(x)$ is a normal random variable given as an integral of a Brownian bridge. The variance of the right-hand side could be computed as a function of m; because the correlation between the terms is positive (due to the properties of Brownian Bridge), the variance increases as more terms are included. However, there seems to be no practical reason for computing this variance.

In order to determine m in practice, since we assume that g is a non-decreasing function, we should restrict our choice to $m = 2q + 1$ for some $q = 1, 2, \ldots$. One can then utilise Equation (3) to test the hypothesis that $J_{2q+1} = 0$ iteratively over increasing q. Once a $\hat{J}_{2q+1}$ is found to be not significantly different from zero, we may set $m = 2q - 1$. Since in general it will not be the case that all the $J_k = 0$ for all k greater than some m, we can take a semiparametric approach instead, and view m as increasing in n in order that the bias vanishes.

Remark 2 Noting that $F(x) = \Phi(g^{-1}(x))$, a density estimate of $f = F'$ is given by $\hat{f}(x) = \phi(\hat{g}_m^{-1}(x))(\hat{g}_m^{-1}(x))'$. This density estimate has a resemblance to the risk neutral density estimates of Madan and Milne (1994) and Jondeau and Rockinger (2001). The validity of $\hat{f}$ as a true density requires that the estimate $\hat{g}_m$ of g is monotone, which is not guaranteed by the previous methodology. An adjustment guaranteeing monotonicity is given in the next section.

Remark 3 Since $Q(u) = g(\Xi(u))$, Corollary 2.1 can be used to obtain an estimate of the quantile function. Making the substitution $x = \Xi(u)$ in Equation (10) we have an approximation for the quantile function. The variance in Equation (11) can be estimated using Theorem 2.2 which gives a confidence bound for the estimated quantile function.

Remark 4 Suppose that the probability density function of the data in original scale is $p(x;\theta) = \phi(g^{-1}(x);\theta)/\dot{g}(g^{-1}(x))$, where θ is the vector of model parameters. We can do a Hermite expansion on g^{-1} as well, using the same mathematics, and obtain

$$g^{-1}(z) = \sum_{k=0}^{\infty} \frac{G_k}{k!} H_k(z)$$

and

$$G_k = \langle g^{-1}, H_k \rangle = \int_{-\infty}^{\infty} \Xi(F(x)) H_k(x)\, \mathrm{d}x = \int_0^1 \Xi(u) H_k(Q(u))\, \mathrm{d}u.$$

Note that if we tried to devise estimates of G_k, as we did earlier with the Hermite expansion of g, we would obtain a nonlinear functional of the eqf, which is not convenient. However, this formulation is more convenient for Bayesian analysis, since we only need put priors on the G_k coefficients, as well as a prior on the model order (or truncation order of the sum). In this way we can evaluate $p(x; \{G_k\}, \theta)$ quite easily.

2.2. *Dependent sequences*

The results of the previous subsection require that the observations are independent and identically distributed. An extension can be made to accommodate dependence using analogous results from empirical process theory for dependent sequences. This requires stronger conditions on the distribution function F and on the degree of dependence. Let $\{X_n, n \geq 1\}$ be a sequence of square integrable random variables defined on a probability space $(\Omega, \mathcal{F}, P)$, and let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two

σ-algebras contained in $\mathcal{F}$. Define

$$\alpha(\mathcal{F}_1, \mathcal{F}_2) = \sup_{A \in \mathcal{F}_1, B \in \mathcal{F}_2} |P(A \cap B) - P(A)P(B)|$$

as a measure of dependence between $\mathcal{F}_1$ and $\mathcal{F}_2$. Let $\mathcal{F}_n^m = \sigma(X_i, n \le i \le m)$ be the σ-algebras generated by the random variables $X_n, \dots, X_m$, and

$$\alpha(n) = \sup_{k \ge 1} \alpha(\mathcal{F}_1^k, \mathcal{F}_{n+k}^{\infty}).$$

The sequence $\{X_n, n \ge 1\}$ is said to be α-mixing if $\alpha(n) \to 0$ as $n \to \infty$. Csörgö and Yu (1996) prove several theorems with different sets of conditions on the mixing rates and on the distribution function F for weak convergence of the quantile process, one of which is stated next.

Theorem 2.3 (Csörgö and Yu (1996)) *Let $\{X_n, n \ge 1\}$ be a stationary sequence of random variables with common continuous distribution function F. Assume that*

(1) *F is twice differentiable on (a, b), where*

$$a = \sup\{x : F(x) = 0\}, \; b = \inf\{x : F(x) = 1\}, \; -\infty \le a < b \le \infty;$$

(2) *$F'(x) = f(x) > 0$ on (a, b);*
(3) *for some $\gamma > 0$ we have*

$$\sup_{a<x<b} F(x)\{1 - F(x)\} \frac{|f'(x)|}{f^2(x)} = \sup_{0<t<1} t(1-t) \frac{|f'(Q(t))|}{f^2(Q(t))} \le \gamma;$$

(4) *either $A \wedge B > 0$, where $A = \lim_{x \searrow a} f(x) < \infty$ and $B = \lim_{x \nearrow b} f(x) < \infty$, or if $A = 0$ (resp. $B = 0$), then f is non-decreasing (resp. non-increasing) on an interval to the right of a (resp. to the left of b);*
(5) *$\{F(X_n), n \ge 1\}$ is a stationary α-mixing sequence of uniform $[0, 1]$ random variables with*

$$\alpha(n) = O(n^{-\theta-\eta}) \quad \text{for some } \theta \ge 1 + \sqrt{2} \text{ and } \eta > 0.$$

Then

$$\sqrt{n} f(Q(t))(\hat{Q}(t) - Q(t)) \overset{\mathcal{L}}{\Longrightarrow} B^*(t) \text{ in } D[0, 1], \tag{13}$$

where $D[0, 1]$ is the space of functions which are right continuous and whose left-hand limits exist. $B^*(t)$ is a mean-zero continuous Gaussian process defined on $[0, 1]$ with $B^*(0) = B^*(1) = 0$ and covariance function

$$E(B^*(s)B^*(t)) = s \wedge t - st + \sum_{k=2}^{\infty} \{\mathrm{Cov}(I\{F(X_1) \le s\}, I\{F(X_k) \le t\}) + \mathrm{Cov}(I\{F(X_k) \le s\}, I\{F(X_1) \le t\})\}.$$

Remark 5 Because $F(X_t) \in A$ if and only if $Z_t \in \Phi^{-1}(A)$ (because g is assumed to be invertible), it follows that $\{F(X_n), n \ge 1\}$ is α-mixing if and only if $\{Z_n, n \ge 1\}$ is α-mixing, and $\alpha(n)$ is the same for both processes. It is known that a stationary Gaussian process is α-mixing if and only if it is completely regular (Ibragimov and Rozanov 1978), which precludes long-range dependent processes, but includes processes with short-range dependence such as autoregressive and moving average time series. In particular, if $\{Z_n, n \ge 1\}$ is an autoregressive or moving average process, then condition (5) of Theorem 2.3 is satisfied.

Assuming the conditions of Theorem 2.3, an analogous result to Corollary 2.1 can be obtained by applying the continuous linear functional $H(f) = \int_{\epsilon}^{1-\epsilon} q(u)f(u)H_k(\Xi(u))\,\mathrm{d}u$ to Equation (13), that is,

$$\sqrt{n}(\hat{J}_k^{\epsilon} - J_k^{\epsilon}) \overset{\mathcal{L}}{\Longrightarrow} \int_{\epsilon}^{1-\epsilon} q(u)B^*(u)H_k(\Xi(u))\,\mathrm{d}u, \tag{14}$$

and

$$\sqrt{n}(\hat{g}_m^{\epsilon}(x) - g_m^{\epsilon}(x)) \overset{\mathcal{L}}{\Longrightarrow} \sum_{k=0}^{m} \int_{\epsilon}^{1-\epsilon} q(u)B^*(u)H_k(\Xi(u))\,\mathrm{d}u \cdot H_k(x) \tag{15}$$

as $n \to \infty$, where $B^*(\cdot)$ is the Gaussian process defined in Theorem 2.3. The limiting random variables in Equations (14) and (15) are mean-zero Gaussian random variables. Unfortunately, the dependence in the sequence $\{X_n, n \geq 1\}$ makes the covariance of these limiting Gaussian random variables far more complicated than in the i.i.d. case, so that the methods used to derive a consistent variance estimator in Theorem 2.2 can not be directly applied.

While an analytical expression for the variance of $\hat{J}_k^{\epsilon}$ is not easily obtainable, subsampling can be used to estimate the variance of the left-hand side of Equation (14). Politis and Romano (1994) gives the details of subsampling for dependent data, which can be summarised as follows: choose $1 \leq b \leq n$ such that $b/n \to 0$ and $b \to \infty$, and let $\mathcal{B}_i = (X_i, \ldots, X_{i+b-1})$ for $i = 1, \ldots, N = n - b + 1$. The subsampling estimator of the distribution of $\sqrt{n}(\hat{J}_k^{\epsilon} - J_k^{\epsilon})$ based on the overlapping blocks $\mathcal{B}_i$ is

$$\hat{L}_n(x) = \frac{1}{N}\sum_{i=1}^{N} I\left\{\sqrt{b}(\hat{J}_{i,k}^{\epsilon} - \hat{J}_k^{\epsilon}) \leq x\right\}, \quad x \in \mathbb{R},$$

where $\hat{J}_{i,k}^{\epsilon}$ is the estimator of J_k^{ϵ}, calculated from the block of data $\mathcal{B}_i$. The subsampling estimator $\hat{L}_n(x)$ can be used to obtain an estimator of the variance of $\hat{J}_k^{\epsilon}$, which is given by

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{J}_k^{\epsilon}) = \frac{b}{n}\left[\frac{1}{N}\sum_{i=1}^{N}(\hat{J}_{i,k}^{\epsilon})^2 - \left(\frac{1}{N}\sum_{i=1}^{N}\hat{J}_{i,k}^{\epsilon}\right)^2\right].$$

Let $L(x)$ be the distribution function for the limiting Gaussian random variable in Equation (14). Under the conditions of Theorem 2.3, the subsampling estimator $\hat{L}_n(x)$ converges uniformly to $L(x)$ in probability as $n \to \infty$ (Politis and Romano 1994).

The results of this subsection hold for fixed ϵ. We conjecture that the results hold, as in the i.i.d. case, with $\epsilon = \epsilon_n = 1/n$. However, the theory for quantile processes of dependent sequences is not nearly as well-developed as it is for the i.i.d. case, and the results for integrated quantile processes used in the proof of Theorem 2.1 do not apply when the data is correlated.

3. Non-monotone estimates

Throughout, we have assumed that the function g in Equation (1) is a non-decreasing function. However, even though g is monotone, it is not necessarily true that the truncated series

$$g_{2q+1}(x) = \sum_{k=0}^{2q+1} J_k(Q)H_k(x) \equiv \sum_{j=0}^{2q+1} a_j x^j \tag{16}$$

is monotone in x. Let $\boldsymbol{a}_{2q+1} = (a_1, \ldots, a_{2q+1})^{\mathrm{T}}$ and $\boldsymbol{J}_{2q+1} = (J_1(Q), \ldots, J_{2q+1}(Q))^{\mathrm{T}}$; the vectors $\boldsymbol{a}_{2q+1}$ and $\boldsymbol{J}_{2q+1}$ are related through the linear transformation $\boldsymbol{a}_{2q+1} = \boldsymbol{A}_{2q+1}\boldsymbol{J}_{2q+1}$, where

$\boldsymbol{A}_{2q+1}$ is a $(2q+1) \times (2q+1)$ non-singular upper triangular matrix with 1s on the main diagonal. For example, if $q = 1$, $\boldsymbol{a}_3 = (J_0 - J_2/\sqrt{2}, J_1 - 3J_3/\sqrt{6}, J_2/\sqrt{2}, J_3/\sqrt{6})^{\mathrm{T}}$ and for $q = 2, \boldsymbol{a}_5 = (J_0 - J_2/\sqrt{2} + 3J_4/\sqrt{24}, J_1 - 3J_3/\sqrt{6} + 15J_5/\sqrt{120}, J_5/\sqrt{2} - 6J_4/\sqrt{24}, J_3/\sqrt{6} - 10J_5/\sqrt{120}, J_4\sqrt{24}, J_5/\sqrt{120})$.

As described in Section 2, let

$$\hat{g}_{2q+1}(x) = \sum_{k=0}^{2q+1} J_k(\hat{Q}_n)H_k(x) \equiv \sum_{j=0}^{2q+1} \hat{a}_j x^j \tag{17}$$

be the polynomial estimate of degree $2q+1$ of $g(x)$ based on a sample $X_1, \dots, X_n$ from a population F, where $X_j = g(Z_j)$ and the Z_j are i.i.d. $N(0,1)$ (dropping the superscript ϵ for ease of presentation). Since $\hat{\boldsymbol{J}}_{2q+1}$ is a consistent estimator of $\boldsymbol{J}_{2q+1}$, if g_{2q+1} is not monotone, $\hat{g}_{2q+1}$ will not necessarily be monotone, even for large values of n. Furthermore, even if g_{2q+1} is non-decreasing, it is possible that for small and moderate values of n, the estimate $\hat{g}_{2q+1}$ may not be monotone, which is a problem if we want to work with the inverse function g^{-1}.

For cases in which the estimated polynomial $\hat{g}_{2q+1}(x)$ is not monotone, it is desirable to find adjusted estimates $\tilde{\boldsymbol{J}}_{2q+1} = (\tilde{J}_1, \dots, \tilde{J}_{2q+1})^{\mathrm{T}}$ (the constant term J_0 plays no role in the monotonicity of the polynomial) which are 'close' to the original estimates $\hat{\boldsymbol{J}}_{2q+1}$, such that the polynomial based on $\tilde{\boldsymbol{J}}_{2q+1}$,

$$\tilde{g}_{2q+1}(x) = \sum_{k=1}^{2q+1} \tilde{J}_k H_k(x) \equiv \sum_{k=1}^{2q+1} \tilde{a}_k x^k, \tag{18}$$

is monotone. The question is then how to calculate $\tilde{\boldsymbol{J}}_{2q+1}$ and to understand the behaviour of these adjusted estimates. The discussion that follows describes a projection approach to calculating a set of coefficients $\tilde{\boldsymbol{J}}_{2q+1}$ which guarantee a non-decreasing estimate of g.

The polynomials of degree $2q+1$ can be identified by the coefficients. Let $\Theta_{2q+1} \subset \mathbb{R}^{2q+1}$ be the closed convex cone consisting of the coefficients of non-decreasing polynomials. For any vector $\boldsymbol{a} = (a_1, \dots, a_{2q+1})^{\mathrm{T}} \in \mathbb{R}^{2q+1}$ let $\|\boldsymbol{a}\|_2 = (\sum_{k=1}^{2q+1} a_k^2)^{1/2}$. We can obtain the coefficients of a non-decreasing polynomial by finding the projection, $\tilde{\boldsymbol{a}}_{2q+1}$ of $\hat{\boldsymbol{a}}_{2q+1}$ onto Θ_{2q+1}, that is, the point $\tilde{\boldsymbol{a}}_{2q+1}$ such that $\|\tilde{\boldsymbol{a}}_{2q+1} - \hat{\boldsymbol{a}}_{2q+1}\|_2 \leq \|\boldsymbol{a} - \hat{\boldsymbol{a}}_{2q+1}\|_2$ for all $\boldsymbol{a} \in \Theta_{2q+1}$. Since Θ_{2q+1} is a non-empty closed convex subset of a Hilbert space, the point $\tilde{\boldsymbol{a}}_{2q+1}$ exists and is unique. The coefficients of the Hermite polynomials can then be found via the inverse transformation $\tilde{\boldsymbol{J}}_{2q+1} = \boldsymbol{A}_{2q+1}^{-1}\tilde{\boldsymbol{a}}_{2q+1}$. Similarly, if g_{2q+1} is not monotone, the first $2q+1$ coefficients, $\boldsymbol{a}_{2q+1}$ in (16), can be projected onto Θ_{2q+1} to obtain $\boldsymbol{a}^*_{2q+1}$, which can then be transformed to $\boldsymbol{J}^*_{2q+1} = \boldsymbol{A}_{2q+1}^{-1}\boldsymbol{a}^*_{2q+1}$. We call $\tilde{\boldsymbol{J}}_{2q+1}$ the projected estimates, and, following the terminology from the misspecified model literature (Sawa 1978), $\boldsymbol{J}^*_{2q+1}$ the pseudo-true values. Clearly, if g_{2q+1} is non-decreasing, then $\boldsymbol{J}^*_{2q+1} = \boldsymbol{J}_{2q+1}$.

The polynomial $\tilde{g}_{2q+1}$ based on $\tilde{\boldsymbol{J}}_{2q+1}$ in Equation (18) is non-decreasing, so the derivative $\tilde{g}'_{2q+1}(x)$ must be non-negative for all x, with positive leading coefficient. Hence all roots of the derivative must be complex conjugate pairs, say $u_l \pm iv_l$, so that the derivative can be written

$$\begin{aligned} \tilde{g}'_{2q+1}(x) &= \sum_{k=0}^{2q}(k+1)a_{k+1}x^k = (2q+1)a_{2q+1}\prod_{l=1}^{q}(x - [u_l + iv_l])(x - [u_l - iv_l]) \\ &= (2q+1)a_{2q+1}\prod_{l=1}^{q}(x^2 - 2u_l x + u_l^2 + v_l^2). \end{aligned} \tag{19}$$

The Hermite coefficients $\tilde{\boldsymbol{J}}_{2q+1}$ in Equation (18) can be related to the parameters $\boldsymbol{u} = (u_1, \dots, u_q)^{\mathrm{T}}$ and $\boldsymbol{v} = (v_1, \dots, v_q)^{\mathrm{T}}$ by expanding Equation (19) and matching coefficients to the derivative of Equation (18); let $M_k(\boldsymbol{\theta})$ denote this mapping, where $\boldsymbol{\theta}^{\mathrm{T}} = (\boldsymbol{u}^{\mathrm{T}}, \boldsymbol{v}^{\mathrm{T}}, a)$ and a is the leading coefficient.

Denote the squared-error loss function by

$$L = L(\boldsymbol{\theta}; Q) = \sum_{k=1}^{2q+1} \{M_k(\boldsymbol{\theta}) - J_k(Q)\}^2. \tag{20}$$

The coefficients $\tilde{\boldsymbol{J}}_{2q+1}$ can be found by minimising $\|\boldsymbol{a} - \hat{\boldsymbol{a}}_{2q+1}\|_2$ over Θ_{2q+1}, or equivalently, by obtaining the parameter $\hat{\boldsymbol{\theta}}$ which minimises the loss function $L(\boldsymbol{\theta}; \hat{Q})$ and setting $\tilde{J}_k = M_k(\hat{\boldsymbol{\theta}})$. Similarly, let $\boldsymbol{\theta}^*$ be the minimiser of $L(\boldsymbol{\theta}; Q)$ so that $J_k^* = M_k(\boldsymbol{\theta}^*)$. Since the projected estimate $\tilde{\boldsymbol{J}}_{2q+1}$ and the pseudo-true value $\boldsymbol{J}^*_{2q+1}$ are unique, the sets of roots, combined with the leading coefficients $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$, must also be unique. For example, if $q = 1$, the reparameterisation is $M_3(\boldsymbol{\theta}) = \tilde{J}_3$, $M_2(\boldsymbol{\theta}) = -6\tilde{J}_3 u/\sqrt{6}$, and $M_1(\boldsymbol{\theta}) = 3\tilde{J}_3/\sqrt{6} + 3\tilde{J}_3(u^2 + v^2)/\sqrt{6}$, and $\boldsymbol{\theta}^*$ can be found by numerically minimising Equation (20). If $\hat{g}_3$ is not monotone, it is easy to show that the value v^* in the minimiser $\boldsymbol{\theta}^*$ must be zero, so that the projected polynomial, $\tilde{g}_3$, will have a single real root of multiplicity 2.

This problem is similar to the problem of parameter estimation in misspecified models (Huber 1967; White 1982), in that misspecification occurs due to the fact an insufficient number of terms has been included for accurate estimation of g, and the polynomial that is ultimately estimated is non-monotone. Using the theory of misspecified models, we expect that $\tilde{\boldsymbol{J}}_{2q+1}$ is an estimator of $\boldsymbol{J}^*_{2q+1}$:

THEOREM 3.1 *Suppose the function $g_{2q+1}(x)$ in Equation* (16) *is not monotone. If*

$$\sqrt{n}(\hat{\boldsymbol{J}}_{2q+1} - \boldsymbol{J}_{2q+1}) \overset{\mathcal{L}}{\Longrightarrow} N_{2q+1}(\mathbf{O}, \boldsymbol{V}),$$

the projected estimate $\tilde{\boldsymbol{J}}_{2q+1}$ is a consistent estimator of $\boldsymbol{J}^*_{2q+1}$. If, in addition, the Hessian matrix $\nabla_{\boldsymbol{\theta}} \nabla^{\mathrm{T}}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; Q)$ is non-singular with continuous inverse in a neighbourhood of $\boldsymbol{\theta}^*$, $\sqrt{n}(\tilde{\boldsymbol{J}}_{2q+1} - \boldsymbol{J}^*_{2q+1})$ is asymptotically Gaussian.

Proof Let Π denote the projection mapping onto Θ_{2q+1} and let $\|\cdot\|_{op}$ be the operator norm on the space of $(2q+1) \times (2q+1)$ matrices induced by the vector norm $\|\cdot\|_2$. Then

$$\begin{aligned} \|\tilde{\boldsymbol{J}}_{2q+1} - \boldsymbol{J}^*_{2q+1}\|_2 &= \|(\boldsymbol{A}^{-1}_{2q+1} \circ \Pi \circ \boldsymbol{A}^{-1}_{2q+1}) \circ (\hat{\boldsymbol{J}}_{2q+1} - \boldsymbol{J}_{2q+1})\|_2 \\ &\le \|\boldsymbol{A}^{-1}_{2q+1}\|_{op} \|\boldsymbol{A}_{2q+1}\|_{op} \Big\| \hat{\boldsymbol{J}}_{2q+1} - \boldsymbol{J}_{2q+1}\|_2. \end{aligned}$$

Since the operator norms of $\boldsymbol{A}_{2q+1}$ and $\boldsymbol{A}^{-1}_{2q+1}$ are bounded and $\hat{\boldsymbol{J}}_{2q+1}$ is a consistent estimator of $\boldsymbol{J}_{2q+1}$, it follows that $\tilde{\boldsymbol{J}}_{2q+1}$ is a consistent estimator of $\boldsymbol{J}^*_{2q+1}$. The projected estimate $\tilde{\boldsymbol{J}}_{2q+1}$ and the pseudo-true value $\boldsymbol{J}^*_{2q+1}$ are unique, so there must be unique sets of roots, $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$, which minimise $L(\boldsymbol{\theta}; \hat{Q})$ and $L(\boldsymbol{\theta}; Q)$, respectively. Since $\tilde{\boldsymbol{J}}_{2q+1}$ is consistent for $\boldsymbol{J}^*_{2q+1}$, it follows that $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}^*$. Furthermore, since $\sqrt{n}(\hat{\boldsymbol{J}}_{2q+1} - \boldsymbol{J}_{2q+1})$ converges weakly to a Gaussian random variable, $\sqrt{n}(\tilde{\boldsymbol{J}}_{2q+1} - \boldsymbol{J}^*_{2q+1}) = \boldsymbol{O}_p(1)$.

Since $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}^*$, and each $M_k(\boldsymbol{\theta})$ is a polynomial function of $\boldsymbol{\theta}$, a Taylor series expansion of $M_k(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}^*$ is given by

$$M_k(\hat{\boldsymbol{\theta}}) - M_k(\boldsymbol{\theta}^*) = \nabla^{\mathrm{T}}{}_{\boldsymbol{\theta}} M_k(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + o_p(1).$$

Note that $\hat{\boldsymbol{\theta}}$ minimises $L(\boldsymbol{\theta}; \hat{Q})$ and $\boldsymbol{\theta}^*$ minimises $L(\boldsymbol{\theta}; Q)$, so $\nabla_{\boldsymbol{\theta}} L(\hat{\boldsymbol{\theta}}, \hat{Q}) = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*, Q) = 0$. Hence

$$\nabla_{\boldsymbol{\theta}} L(\hat{\boldsymbol{\theta}}; Q) = \nabla_{\boldsymbol{\theta}} L(\hat{\boldsymbol{\theta}}; Q) - \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*; Q) = \nabla_{\boldsymbol{\theta}} \nabla^{\mathrm{T}}{}_{\boldsymbol{\theta}} L(\boldsymbol{\xi}; Q)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \tag{21}$$

for some $\boldsymbol{\xi}$ between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$,

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; Q) = 2 \sum_{k=1}^{2q+1} (M_k(\boldsymbol{\theta}) - J_k(Q)) \nabla_{\boldsymbol{\theta}} M_k(\boldsymbol{\theta}),$$

and

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} L(\hat{\boldsymbol{\theta}}; Q) &= \nabla_{\boldsymbol{\theta}} L(\hat{\boldsymbol{\theta}}; Q) - \nabla_{\boldsymbol{\theta}} L(\hat{\boldsymbol{\theta}}; \hat{Q}) \\ &= 2 \sum_{k=1}^{2q+1} (M_k(\hat{\boldsymbol{\theta}}) - J_k(Q)) \nabla_{\boldsymbol{\theta}} M_k(\hat{\boldsymbol{\theta}}) - 2 \sum_{k=1}^{2q+1} (M_k(\hat{\boldsymbol{\theta}}) - J_k(\hat{Q})) \nabla_{\boldsymbol{\theta}} M_k(\hat{\boldsymbol{\theta}}) \\ &= 2 \sum_{k=1}^{2q+1} \nabla_{\boldsymbol{\theta}} M_k(\hat{\boldsymbol{\theta}})(J_k(\hat{Q}) - J_k(Q)). \end{aligned} \tag{22}$$

Since $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^*$, $\sqrt{n}(J_k(\hat{Q}) - J_k(Q)) = O_p(1)$, and $\nabla_{\boldsymbol{\theta}} M_k(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$, by the continuous mapping theorem,

$$\begin{aligned} &2\sqrt{n} \sum_{k=1}^{2q+1} \nabla_{\boldsymbol{\theta}} M_k(\hat{\boldsymbol{\theta}})(J_k(\hat{Q}) - J_k(Q)) = 2\sqrt{n} \sum_{k=1}^{2q+1} \nabla_{\boldsymbol{\theta}} M_k(\boldsymbol{\theta}^*)(J_k(\hat{Q}) - J_k(Q)) + \boldsymbol{o}_p(1) \\ &\equiv 2\sqrt{n}\boldsymbol{D}(\hat{\boldsymbol{J}}_{2q+1} - \boldsymbol{J}_{2q+1}) + \boldsymbol{o}_p(1) \overset{\mathcal{L}}{\Longrightarrow} N_{2q+1}(\boldsymbol{0}, \boldsymbol{W}), \end{aligned} \tag{23}$$

where $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{\theta}^*) = (\nabla_{\boldsymbol{\theta}} M_1(\boldsymbol{\theta}^*), \ldots, \nabla_{\boldsymbol{\theta}} M_{2q+1}(\boldsymbol{\theta}^*))^{\mathrm{T}}$ and $\boldsymbol{W} = \boldsymbol{W}(\boldsymbol{\theta}^*) = 4\boldsymbol{D}\boldsymbol{V}\boldsymbol{D}^{\mathrm{T}}$. Combining Equations (21)–(23) gives

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) &= \sqrt{n}(\nabla_{\boldsymbol{\theta}} \nabla^{\mathrm{T}}{}_{\boldsymbol{\theta}} L(\boldsymbol{\xi}; Q))^{-1} \nabla_{\boldsymbol{\theta}} L(\hat{\boldsymbol{\theta}}; Q) \\ &= 2\sqrt{n}(\nabla_{\boldsymbol{\theta}} \nabla^{\mathrm{T}}{}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*; Q))^{-1} \boldsymbol{D}(\hat{\boldsymbol{J}}_{2q+1} - \boldsymbol{J}_{2q+1}) + \boldsymbol{o}_p(1) \\ &\equiv 2\sqrt{n}\boldsymbol{H}\boldsymbol{D}(\hat{\boldsymbol{J}}_{2q+1} - \boldsymbol{J}_{2q+1}) + \boldsymbol{o}_p(1), \end{aligned}$$

so that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \overset{\mathcal{L}}{\Longrightarrow} N_{2q+1}(\boldsymbol{0}, \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}).$$

Let $\boldsymbol{M}(\boldsymbol{\theta}) = (M_1(\boldsymbol{\theta}), \ldots, M_{2q+1}(\boldsymbol{\theta}))^{\mathrm{T}}$. By the multivariate delta method,

$$\sqrt{n}(\tilde{\boldsymbol{J}}_{2q+1} - \boldsymbol{J}^*_{2q+1}) = \sqrt{n}(\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta}^*)) \overset{\mathcal{L}}{\Longrightarrow} N(0, (\nabla_{\boldsymbol{\theta}} \boldsymbol{M}(\boldsymbol{\theta}^*))^{\mathrm{T}} \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}} (\nabla_{\boldsymbol{\theta}} \boldsymbol{M}(\boldsymbol{\theta}^*))).$$

■

Remark 6 A Lagrange multiplier approach could be used instead when working with a third degree polynomial; a third degree polynomial is monotone if and only if

$\hat{J}_2^2 + 3\hat{J}_3^2 - 6\hat{J}_1\hat{J}_3/\sqrt{6} \leq 0$. Hence, the Hermite coefficients $\tilde{\boldsymbol{J}}_3$ can be found through use of the Lagrangian $\Lambda(\boldsymbol{J}, \lambda) = \|\boldsymbol{J}_3 - \hat{\boldsymbol{J}}_3\|^2 + \lambda(\hat{J}_2^2 + 3\hat{J}_3^2 - 6\hat{J}_1\hat{J}_3/\sqrt{6})$. The asymptotic behaviour of $\tilde{\boldsymbol{J}}_3$ can then be found by following methods similar to those in Silvey (1959). However, for higher degree polynomials, this method is not tractable, as there is no obvious way to describe monotonicity of the polynomial through a set of functional constraints, so that a Lagrangian cannot be written.

Remark 7 Due to the orthogonality of the Hermite functions,

$$\int_{-\infty}^{\infty} (\hat{g}_m(x) - \tilde{g}_m(x))^2 \phi(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \left(\sum_{k=0}^{m} \hat{J}_k H_k(x) - \sum_{k=0}^{m} \tilde{J}_k H_k(x) \right)^2 \phi(x)\,\mathrm{d}x$$

$$= \sum_{k,l=0}^{m} (\hat{J}_k - \tilde{J}_k)(\hat{J}_l - \tilde{J}_l) \int_{-\infty}^{\infty} H_k(x) H_l(x) \phi(x)\,\mathrm{d}x = \sum_{k=0}^{m} (\hat{J}_k - \tilde{J}_k)^2.$$

Hence, obtaining adjusted coefficients which minimise the squared-error loss function (20) subject to a monotonicity constraint is equivalent to obtaining an adjusted estimate of the transformation function by minimising an integrated squared-error loss over the space of polynomials, subject to a monotonicity constraint.

4. Applications

4.1. *SAIPE*

The U.S. Census Bureau's SAIPE programme provides current estimates of income and poverty within school districts, counties, and states for the age groups 0–4, 5–17, 18–64, and 65 and older. The published estimates of the proportion in poverty within each state are model-based, and employ both the direct estimate of the proportion in poverty from the American Community Survey (ACS) and regression predictions of poverty based on administrative records and census data. The model used is the Fay–Herriot model (Fay and Herriot 1979),

$$\begin{aligned} y_i &= Y_i + e_i, \\ Y_i &= \boldsymbol{x}^{\mathrm{T}}{}_i \boldsymbol{\beta} + u_i, \end{aligned} \tag{24}$$

$i = 1, \ldots, n$, where the e_i are independent $N(0, V_i)$ with known variances V_i, and the u_i are i.i.d. $N(0, \sigma^2)$, independent of the e_i. The state model has $n = 51$ small areas for the 50 states and the District of Columbia, and the y_i are the direct survey estimates from the ACS one-year estimates of the true proportion in poverty Y_i. The covariates $\boldsymbol{x}_i$ in the regression part of the model include the tax return poverty rate, the tax non-filer rate, the Supplemental Nutrition Assistance Program (SNAP) participation rate, Supplemental Security Income (SSI) recipiency rate, and the residuals from a regression of the Census 2000 poverty ratios on the previous four covariates.

For known σ^2, the best linear unbiased predictor of (BLUP) of Y_i is given by

$$\tilde{Y}_i = (1 - h_i) y_i + h_i \boldsymbol{x}^{\mathrm{T}}{}_i \tilde{\boldsymbol{\beta}},$$

where $h_i = V_i/(\sigma^2 + V_i)$, $\boldsymbol{\Sigma} = \mathrm{diag}(V_1 + \sigma^2, \ldots, V_{51} + \sigma^2)$, and

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\sigma^2) = (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{y}.$$

The BLUP $\tilde{Y}_i$ is thus a weighted average of the regression estimate $\boldsymbol{x}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}$ and the direct survey estimate y_i.

The variance component σ^2 can be estimated by numerically computing the restricted maximum likelihood (REML) estimator (Harville 1977), that is, the value $\hat{\sigma}^2$ which maximises the restricted likelihood

$$L(\sigma^2;\boldsymbol{y}) = |\boldsymbol{\Sigma}|^{-1/2}|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}|^{-1/2}\mathrm{e}^{-1/2}(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}).$$

The empirical best linear unbiased predictor (EBLUP) of Y_i is

$$\hat{Y}_i = (1-\hat{h}_i)y_i + \hat{h}_i\boldsymbol{x}^{\mathrm{T}}{}_i\hat{\boldsymbol{\beta}},$$

where $\hat{h}_i = V_i/(\hat{\sigma}^2 + V_i)$ and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}^2)$.

The empirical Bayes interval for Y_i

$$I_i^C(\alpha) = \hat{Y}_i \pm z_{\alpha/2}V_i^{1/2}(1-\hat{h}_i)^{1/2}$$

has the property that

$$\mathbb{P}(Y_i \in I_i^C(\alpha)) = 1-\alpha+O(n^{-1}),$$

where $\mathbb{P}$ is the probability distribution induced by the joint distribution of Equation (24) (Chatterjee, Lahiri, and Li 2008). The validity of the above interval depends on the assumed normality of the error terms e_i and v_i. If the error terms have departures from normality, a score other than $z_{\alpha/2}$ should be used.

For evaluating the validity of the distributional assumptions in linear mixed effects models such as Equation (24), Lange and Ryan (1989) suggested using weighted QQ plots based on the EBLUP of the area-specific random effects u_i as a graphical method for detecting departures from normality. The BLUP of u_i is given by

$$\tilde{u}_i = \frac{\sigma^2}{\sigma^2+V_i}(y_i - \boldsymbol{x}^{\mathrm{T}}{}_i\hat{\boldsymbol{\beta}}), \tag{25}$$

and the variance of $\tilde{u}_i$ is

$$\left(\frac{\sigma^2}{\sigma^2+V_i}\right)^2(\sigma^2 + V_i - \boldsymbol{x}^{\mathrm{T}}{}_i(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i). \tag{26}$$

The EBLUP of u_i, denoted by $\hat{u}_i$, is $\tilde{u}_i$ in Equation (25), with σ^2 replaced by its REML estimate, $\hat{\sigma}^2$. Similarly, the variance of $\hat{u}_i$ can be estimated by replacing σ^2 with $\hat{\sigma}^2$ in Equation (26). Under the model (24), the set of standardised EBLUPs, $\hat{u}_i/\sqrt{\hat{Var}(\hat{u}_i)}$, are approximately i.i.d. $N(0,1)$ random variables.

The reasonableness of the normality assumption for the SAIPE data set can be evaluated by using the methodology of Section 2. Treating the standardised residuals as a set of i.i.d. random variables, and writing $g(Z_i) = \hat{u}_i/\sqrt{\hat{Var}(\hat{u}_i)}$, where the Z_i are i.i.d. standard Gaussian random variables, we can estimate the transformation function g by estimating the Hermite coefficients. The ϵ used in calculating the coefficients was 10^{-7}, chosen to be small compared to $1/n$, the inverse of the sample sizes, so that all the data was used and any bias induced with the choice of ϵ in estimation of the first and last coefficients would be small. If a plot of this estimate differs from a straight line, the assumption of normality should be questioned and different model assumptions may need to be made.

Figure 1 shows the third and fifth degree Hermite polynomial estimates of the transformation function g for the standardised residuals under two different models using 2012 data. The first is the estimated transformation function of the standardised residuals of the estimates of the

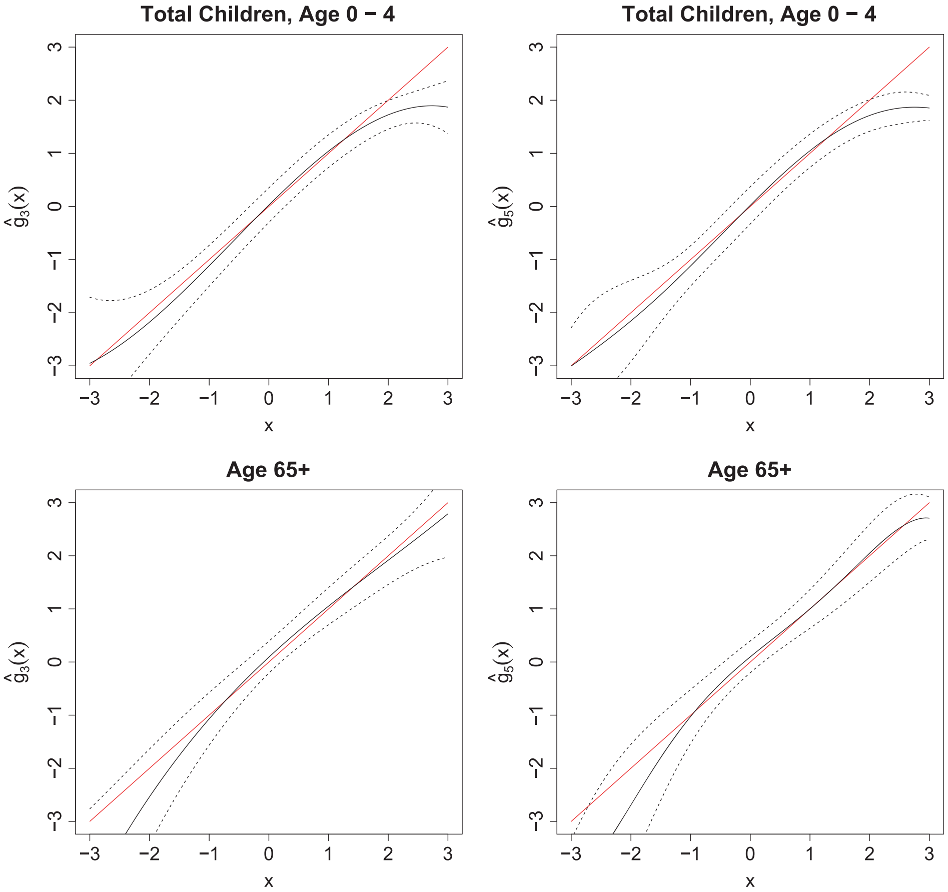

Figure 1. Third and fifth degree Hermite polynomial estimates of the transformation function of the standardised EBLUPs for age groups 0–4 and 65 + . The estimated functions are solid black lines and the dashed lines give pointwise 95% confidence intervals. The diagonal line corresponding to a Gaussian distribution is shown in red.

proportion in poverty of children aged 0–4. The second set of plots is for the proportion in poverty in the 65 and older age group. The solid black line is the estimated function, and the solid red line shows the straight line corresponding to a normal distribution. The dashed lines give a pointwise 95% confidence interval based on Corollary 2.1 and the estimated covariance in Theorem 2.2. Comparing the plot of the estimated transformation function along with the confidence intervals to the diagonal line can be used as a visual diagnostic for departures from normality.

The plot of the estimates $\hat{g}_3$ and $\hat{g}_5$ in Figure 1 for the 65 and over age group diverge from the diagonal line for the negative values, due mainly to a few negative standardised residuals that are large in magnitude. However, the confidence intervals for each of these estimated functions contain the diagonal, suggesting no strong evidence against the normality of the errors in the model for proportion in poverty in the age 65 and older group. Compare this to the plots for the 0–4 age group. The estimated function is close to the diagonal line for negative values of x, but is beneath the diagonal for positive values of x, suggesting skew in standardised residuals. Also, the confidence intervals do not contain the diagonal for larger values of x. For comparison, the sample skewness of the standardised EBLUPs is -0.333 for the 0–4 age group and -0.677 for the 65 + age group. The Kolmogorov–Smirnov, Shapiro–Wilk, and Anderson–Darling tests for normality give p-values of 0.901, 0.466, and 0.485, respectively, for the 0–4 age group, and 0.876,0.163, and 0.248, respectively, for the 65 + age group.

Constructing an estimate of g also allows for the calculation of a confidence interval that does not depend on the assumption of normality of the error terms in the Fay–Herriot model (24). The distribution function of the residuals, F, satisfies the relationship $F = \Phi(g^{-1})$. Critical values x_1 and x_2 can be chosen such that

$$\begin{aligned} 1-\alpha &= P(x_1 \le X \le x_2) = P(Z \le g^{-1}(x_2)) - P(Z \le g^{-1}(x_1)) \\ &= \Phi(g^{-1}(x_2)) - \Phi(g^{-1}(x_1)). \end{aligned} \tag{27}$$

The cutoff points x_1 and x_2 can be approximated by first using the sample to estimate g^{-1} with $\hat{g}_m^{-1}$ (or if the estimate $\hat{g}_m$ is not monotone, the adjusted estimate $\tilde{g}_m$ from Theorem 3.1), and choosing points x_1 and x_2 to satisfy $\Phi(\hat{g}_m^{-1}(x_2)) - \Phi(\hat{g}_m^{-1}(x_1)) = 1-\alpha$. For the age group 0 – 4, the values of x_1 and x_2 which satisfy Equation (27) for $\alpha = 0.05$ and minimise the distance $x_2 - x_1$ are $x_2 = 1.86$ and $x_1 = -2.07$.

4.2. *AR-sieve bootstrap*

Suppose we observe a sample from a stationary time series $\{Y_t\}$, and we want to estimate the sampling distribution of a statistic $T_n = T_n(Y_1, Y_2, \dots, Y_n)$. One could proceed by applying an AR-sieve (Bühlmann 1997) (or the MA sieve of McMurry and Politis (2010)) to the data to produce approximately uncorrelated residuals $X_1, X_2, \dots, X_n$. Suppose the residuals satisfy $X_t = g(Z_t)$ for i.i.d. Gaussian variables Z_t; we can construct $\hat{g}_m$ using the methodology of Section 2, that is, estimate the Hermite coefficients in ascending order, only stopping when they are no longer significantly different from zero, and obtaining the corresponding estimate of g.

Next we describe a procedure analogous to the parametric bootstrap (Shao and Tu 1995). Given an estimate of g, a pseudo-sample can be obtained by drawing n i.i.d. standard normal variables, and letting

$$X_t^{(J)} = \hat{g}_m(Z_t^{(J)})$$

for $t = 1, 2, \dots, n$ and $J \ge 1$. Then each sample can be re-correlated by reversing the sieve, in this manner producing pseudo-samples $Y_1^{(J)}, Y_2^{(J)}, \dots, Y_n^{(J)}$. Then the statistic can be calculated via $T_n^{(J)} = T_n(Y_1^{(J)}, Y_2^{(J)}, \dots, Y_n^{(J)})$, and the collection of these evaluations over various J is $T_n^{(1)}, T_n^{(2)}, \dots$. These quantities then form an estimate of the sampling distribution of T_n. This procedure is essentially a modification of the AR-sieve, but instead of sampling from the empirical distribution of the model residuals, we sample from a smoothed distribution by estimating a transformation function.

There are some cautions about this approach. First, it is only feasible for stationary time series. Second, an appropriate AR-sieve must be identified – which requires some crude time series modelling – and the series must be de-meaned. This only removes second-order dependencies in the time series. The method is only useful for non-Gaussian time series – else one can proceed with a simpler approach – which cannot be reduced to independence by a de-correlation. Note that the identification of an AR-sieve is tantamount to modelling the time series, since in either case one attempts to find a 'whitening transformation' (McElroy and Holan 2009). Third, one applies the quantile methodology of Section 3 by assuming the data to be independent, or weakly dependent, as in Section 2.2. Of course, $\hat{g}$ is a statistical estimate of g, so further error enters in our generation of the pseudo-samples.

Given these provisos, how useful is the method in practice? The method is likely to be reasonably effective with weakly correlated time series with highly non-standard marginal distributions

– for example, with much skewness and/or kurtosis. For example, one might employ the technique for heavy-tailed data (but so long as the mean exists) for which the bootstrap works badly (Athreya 1987). The method resembles a parametric bootstrap, where the parent distribution is nonparametrically estimated. In order to evaluate the technique, we present results from numerical experiments.

The models we consider are

M1:

$$X_t = \sum_{i=1}^{48} \phi_i X_{t-i} + \varepsilon_i,$$

where $\phi_j = (-1)^{j+1} 7.5/(j+1)^3$ for $j = 1, \ldots, 48$, and ε_i are i.i.d. $t_3/\sqrt{3}$, and M2:

$$X_t = 0.8X_{t-1} - 0.5\varepsilon_{t-1} + \varepsilon_t,$$

where ε_t are i.i.d. $t_3/\sqrt{3}$ random variables.

Models M1 and M2 are from Bühlmann (1997), with the exception that Bühlmann (1997) used $N(0, 1)$ errors in model M1 and a mixture of $N(0, 1)$ and $N(0, 100)$ errors in M2.

For the simulation studies, sample sizes of $n = 64$ or 512 were used, with 500 bootstrap replications. For each example, 250 simulations were run. As with the SAIPE example in the previous subsection, a value of $\epsilon = 10^{-7}$ was chosen, so that all data would be used in the estimation of the coefficients, and any bias in the first and last coefficient estimates would be minimal. For each model, two cases were considered: when the statistic of interest is the mean $T_n = n^{-1} \sum_{i=1}^{n} X_i$ or the median $T_n = \mathrm{med}\{X_1, \ldots, X_n\}$, and a bootstrap is needed to estimate the variance of T_n. The true value of $\sigma_n^2 = n * \mathrm{var}(T_n)$ was found by simulating each model 10,000 times. All computational work in this paper was done using R (R Development Core Team 2011).

Table 1 presents results for $T_n = n^{-1} \sum_{i=1}^{n} X_i$. In this table, $(\sigma_n^2)^* = n * \mathrm{var}^*(T_n^*)$, the bootstrap variance estimate of the mean, and $\mathrm{RMSE} = \mathrm{MSE}((\sigma_n^2)^*)/\sigma_n^4$ is the relative mean square error (MSE) of the bootstrap variance estimate. Four bootstrap methods are compared: the AR-Sieve (AR) of Bühlmann (1997), a bootstrap using third degree Hermite polynomials (H3), a bootstrap using fifth degree Hermite polynomials (H5), and a bootstrap for which the degree of the Hermite polynomial used is data driven.

We see in Table 1 higher accuracy of variance estimation using Hermite polynomials to approximate the distribution of the residuals compared to sampling from the empirical distribution in terms of RMSE, particularly for smaller sample sizes for both models. There is not a major difference between a third degree polynomial compared to a fifth degree polynomial or a data-driven polynomial degree choice, although we generally see a reduction in bias with a higher-degree polynomial at the cost of increased variance.

Table 2 presents the same set of results as does Table 1, with the exception that here we are interested in estimating the variance of the median. As can be seen in Table 2, there is little advantage to smoothing the distribution of the model residuals via polynomial expansion – the AR sieve was slightly better in model M1 for $n = 64$, while the Hermite bootstraps were slightly better in M2 for $n = 64$, and there was almost no difference using either procedure for either model for $n = 512$. A simpler bootstrap for the distribution of the median is most likely more useful since the median is more robust to outliers or heavy tails than the mean. However, when the statistic of interest is the mean, and there is the possibility of extreme observations due to heavy tails which can greatly influence the statistic, smoothing the distribution of the residuals may give better performance, as the interpolation due to the estimated transformation function allows one to sample deeper in the tails of the distribution in the bootstrap procedure. We note that in other numerical examples where the error terms are Gaussian, there was little or no benefit to

Table 1. A comparison of the AR sieve bootstrap with a bootstrap using the Hermite estimation procedure.

Model		Method	σ_n^2	$E[(\sigma_n^2)^*]$	$\mathrm{SD}[(\sigma_n^2)^*]$	RMSE
M1	$n = 64$	AR	13.47	10.43	11.25	0.75
		H3		9.64	9.43	0.57
		H5		9.84	9.71	0.59
		HN		9.70	9.52	0.58
	$n = 512$	AR	12.02	13.64	5.86	0.25
		H3		12.91	4.73	0.16
		H5		13.10	5.03	0.18
		HN		13.04	4.97	0.18
M2	$n = 64$	AR	6.02	3.72	4.78	0.77
		H3		3.39	3.93	0.61
		H5		3.44	4.02	0.63
		HN		3.40	3.93	0.61
	$n = 512$	AR	5.56	5.19	2.62	0.23
		H3		4.93	2.09	0.15
		H5		5.02	2.22	0.17
		HN		4.99	2.21	0.17

Notes: H3 and H5 use polynomials of order 3 and 5, respectively, while HN indicates that the order of the polynomial to be used was estimated. The statistic of interest is $T_n = \sum_{i=1}^{n} X_i/n$, and $\sigma_n^2 = n * \mathrm{var}(T_n)$. The last three columns show the mean, standard error, and relative MSE of the bootstrap estimate of σ_n^2.

Table 2. A comparison of the AR sieve bootstrap with a bootstrap using the Hermite estimation procedure.

Model		Method	σ_n^2	$E[(\sigma_n^2)^*]$	$\mathrm{SD}[(\sigma_n^2)^*]$	RMSE
M1	$n = 64$	AR	9.46	8.70	7.72	0.67
		H3		9.33	8.04	0.72
		H5		8.97	7.94	0.70
		HN		9.28	8.08	0.73
	$n = 512$	AR	9.69	10.59	3.32	0.13
		H3		10.67	3.42	0.13
		H5		10.56	3.33	0.13
		HN		10.65	3.36	0.13
M2	$n = 64$	AR	4.13	2.81	2.97	0.62
		H3		2.97	2.86	0.56
		H5		2.67	2.78	0.58
		HN		2.91	2.87	0.57
	$n = 512$	AR	4.07	3.73	1.40	0.12
		H3		3.65	1.44	0.13
		H5		3.57	1.41	0.13
		HN		3.62	1.43	0.13

Notes: H3 and H5 use polynomials of order 3 and 5, respectively, while HN indicates that the order of the polynomial to be used was estimated. The statistic of interest is $T_n = \mathrm{med}\{X_1, \ldots, X_n\}$, and $\sigma_n^2 = n * \mathrm{var}(T_n)$. The last three columns show the mean, standard error, and relative MSE of the bootstrap estimate of σ_n^2.

smoothing the distribution of the model residuals compared to using a bootstrap procedure with samples drawn from the edf.

Table 3 shows the number of times out of 500 simulations a polynomial order was used in procedure (HN), in which the order of the polynomial approximation was data driven, and chosen by sequentially testing the significance of each coefficient. Generally a low-order polynomial of degree 3 or 5 was sufficient to approximate the transformation function, a result which seems consistent with other simulation studies, including those shown in Section 5. However, as the

Table 3. For procedure HN in Tables 1 and 2, the number of times a polynomial of each order was selected under models M1 and M2 with sample sizes of $n = 64$ and $n = 512$.

		Polynomial order			
Model	n	3	5	7	9
M1	64	178	60	10	2
	512	164	72	9	5
M2	64	195	48	7	0
	512	163	78	4	5

observed sample size increases, higher-degree polynomials are more frequently used for accurate approximation.

5. Numerical examples

In this section, numerical examples are presented to investigate the finite sample properties of the estimators and to verify the theoretical results of the previous sections. For each example, 100,000 simulations were performed, and the bias and MSE of the estimators $\hat{J}_k$ for $k = 0, \ldots, 5$ are presented. Also, for each simulation and each k, the 95% confidence interval, $\hat{J}_k \pm 1.96\sqrt{\hat{V}_k/n}$ is calculated, where $\hat{V}_k$ is as in Equation (8), and the proportion of times this interval covers the true value J_k, as well as the average lower and upper endpoints of the confidence intervals is reported. Because the Beta distribution has compact support, the estimated coefficients and variance estimates can be computed with ϵ set to 0.

Table 4 shows the results of a simulation study with samples X_j, $j = 1, \ldots, n$ from the Beta$(5, 1)$ distribution for different values of n. The transforming function for the Beta$(5, 1)$ distribution is $g(x) = \Phi(x)^{1/5}$. Since this function g contracts a standard Gaussian random variable from one with range over the entire real line to a random variable with range concentrated on $(0, 1)$, the estimates $\hat{J}_k$ converge quickly as there is little bias or MSE with an observed sample of only $n = 100$. However, a larger sample size is needed to get confidence intervals coverage rates close to 95%.

The Beta$(5, 1)$ distribution is an example of a distribution for which the transforming function $g(x) = \Phi(x)^{1/5}$ is monotonically increasing, but for which the first 4 coefficients, $J_0 = 0.8333, J_1 = 0.1334, J_2 = -0.0445, J_3 = -0.0045$ lead to a non-monotone polynomial. From these true values J_1, J_2, J_3, the pseudo-true values can be found numerically using the method described in Section 3 and are $J_1^* \approx 0.1338, J_2^* \approx -0.0416, J_3^* \approx 0.0056$. The asymptotic covariance matrix for the projected estimates $\tilde{J}_1, \tilde{J}_2$ and $\tilde{J}_3$ can be found using Theorem 3.1, the pseudo-true values J_1^*, J_2^* and J_3^* and the asymptotic covariance matrix for $\hat{J}_1, \hat{J}_2$ and $\hat{J}_3$

$$\boldsymbol{V}_3 = \begin{bmatrix} 0.01286 & -0.00549 & -0.00129 \\ & 0.00705 & -0.00298 \\ & & 0.00446 \end{bmatrix}, \tag{28}$$

and is

$$\boldsymbol{V}_3^* = \begin{bmatrix} 0.0131 & -0.0048 & 0.0010 \\ & 0.0071 & -0.0021 \\ & & 0.0006 \end{bmatrix}. \tag{29}$$

Table 4. Performance of the estimates $\hat{J}_k$ when $g(x) = \Phi(x)^{1/5}$.

n	k	Bias	AV	Average CI		% Coverage
100	0	0.0001	0.0197	(0.8060,	0.8607)	0.9442
	1	−0.0019	0.0126	(0.1104,	0.1525)	0.9121
	2	0.0027	0.0067	(−0.0555,	−0.0280)	0.8478
	3	−0.0033	0.0038	(−0.0181,	0.0024)	0.8452
	4	0.0031	0.0020	(0.0013,	0.0203)	0.9135
	5	−0.0020	0.0014	(−0.0108,	0.0067)	0.9505
500	0	0.0000	0.0200	(0.8210,	0.8456)	0.9480
	1	−0.0004	0.0129	(0.1232,	0.1428)	0.9410
	2	0.0006	0.0070	(−0.0508,	−0.0368)	0.9205
	3	−0.0009	0.0044	(−0.0106,	−0.0002)	0.8937
	4	0.0010	0.0027	(0.0046,	0.0129)	0.8977
	5	−0.0010	0.0017	(−0.0048,	0.0028)	0.9318
1000	0	0.0000	0.0198	(0.8246,	0.8420)	0.9492
	1	−0.0002	0.0129	(0.1262,	0.1401)	0.9460
	2	0.0003	0.0071	(−0.0492,	−0.0391)	0.9334
	3	−0.0005	0.0045	(−0.0089,	−0.0012)	0.9124
	4	0.0006	0.0029	(0.0053,	0.0113)	0.9036
	5	−0.0006	0.0019	(−0.0033,	0.0020)	0.9234
5000	0	0.0000	0.0200	(0.8294,	0.8372)	0.9490
	1	0.0000	0.0129	(0.1302,	0.1365)	0.9487
	2	0.0001	0.0071	(−0.0467,	−0.0421)	0.9453
	3	−0.0001	0.0045	(−0.0065,	−0.0028)	0.9383
	4	0.0002	0.0030	(0.0064,	0.0093)	0.9275
	5	−0.0002	0.0021	(−0.0014,	0.0010)	0.9247
10,000	0	0.0000	0.0199	(0.8306,	0.8361)	0.9491
	1	0.0000	0.0129	(0.1311,	0.1356)	0.9502
	2	0.0000	0.0070	(−0.0461,	−0.0428)	0.9487
	3	−0.0001	0.0045	(−0.0059,	−0.0033)	0.9432
	4	0.0001	0.0031	(0.0067,	0.0088)	0.9358
	5	−0.0001	0.0022	(−0.0010,	0.0007)	0.9275

Notes: The column AV shows the variance of $\sqrt{n}(\hat{J}_k - J_k)$ over the 100,000 simulations (compare to (28)in Section 5). The final columns show the average lower and upper endpoints of the 95% confidence intervals, and the proportion of times the intervals covered the true values. The MSE of the estimated coefficients are nearly zero for each sample size and each k, so are not shown.

Table 5. Estimated coefficients of the projected transformation function.

		$n = 100$	500	1000	5000	10,000
$k = 1$	Bias	−0.0019	0.0000	−0.0002	0.0000	0.0000
	AV	0.0128	0.0132	0.0131	0.0132	0.0131
$k = 2$	Bias	0.0031	0.0001	0.0004	0.0001	0.0000
	AV	0.0064	0.0070	0.0070	0.0071	0.0070
$k = 3$	Bias	−0.0006	0.0000	−0.0001	0.0000	0.0000
	AV	0.0004	0.0004	0.0005	0.0005	0.0005

Notes: The column AV shows the variance of $\sqrt{n}(\tilde{J}_k - J_k^*)$ over the 100,000 simulations (compare to Equation (29)in Section 5). The MSE for each of the estimates is nearly zero so is not shown.

Table 5 shows the bias and MSE of the projected estimates $\tilde{J}_k$, as well as the variance of $\sqrt{n}(\tilde{J}_k - J_k^*)$ (AV). Since the estimates $\hat{J}_k$ are accurate for small values of n, and the values of V_3^* are not large, it is not surprising that the projected estimates converge rapidly to the pseudo-true values.

The second example uses the logistic distribution, which like the standard normal distribution has mean zero and is symmetric around zero, but has heavier tails than the normal distribution.

Table 6. Performance of the estimates $\hat{J}_k$ when $g(x) = \log(\Phi(x)) - \log(1 - \Phi(x))$.

n	k	Bias	$\sqrt{\text{MSE}}$	Average CI		% Coverage
100	0	0.0006	0.1814	(−0.3517,	0.3529)	0.9468
	1	−0.0327	0.1573	(1.4902,	2.0572)	0.9012
	2	−0.0003	0.1375	(−0.2116,	0.2110)	0.8779
	3	−0.1028	0.1470	(−0.0728,	0.1942)	0.5366
	4	0.0001	0.0655	(−0.1257,	0.1259)	0.9342
	5	−0.0898	0.1032	(−0.2336,	0.0376)	0.7942
500	0	−0.0002	0.0815	(−0.1590,	0.1585)	0.9479
	1	−0.0071	0.0707	(1.6644,	1.9342)	0.9362
	2	−0.0001	0.0662	(−0.1179,	0.1178)	0.9247
	3	−0.0321	0.0667	(0.0426,	0.2203)	0.7531
	4	0.0000	0.0454	(−0.0605,	0.0606)	0.8369
	5	−0.0492	0.0578	(−0.1149,	0.0001)	0.5918
1000	0	−0.0003	0.0574	(−0.1126,	0.1121)	0.9492
	1	−0.0038	0.0500	(1.7062,	1.8989)	0.9429
	2	0.0001	0.0473	(−0.0873,	0.0875)	0.9356
	3	−0.0188	0.0475	(0.0739,	0.2155)	0.8167
	4	0.0000	0.0360	(−0.0497,	0.0497)	0.8482
	5	−0.0342	0.0428	(−0.0821,	−0.0027)	0.5526
5000	0	−0.0001	0.0257	(−0.0503,	0.0502)	0.9491
	1	−0.0009	0.0223	(1.7620,	1.8490)	0.9493
	2	0.0000	0.0215	(−0.0412,	0.0413)	0.9439
	3	−0.0052	0.0213	(0.1210,	0.1956)	0.9010
	4	0.0001	0.0187	(−0.0302,	0.0303)	0.8987
	5	−0.0130	0.0203	(−0.0431,	0.0008)	0.6542
10,000	0	0.0000	0.0181	(−0.0356,	0.0355)	0.9505
	1	−0.0003	0.0157	(1.7752,	1.8369)	0.9492
	2	0.0000	0.0153	(−0.0295,	0.0295)	0.9464
	3	−0.0029	0.0150	(0.1332,	0.1880)	0.9204
	4	0.0000	0.0137	(−0.0233,	0.0234)	0.9122
	5	−0.0082	0.0145	(−0.0341,	0.0014)	0.7228
50,000	0	0.0000	0.0081	(−0.0159,	0.0159)	0.9502
	1	−0.0001	0.0071	(1.7924,	1.8201)	0.9500
	2	0.0000	0.0068	(−0.0133,	0.0133)	0.9500
	3	−0.0008	0.0067	(0.1499,	0.1756)	0.9411
	4	0.0000	0.0065	(−0.0118,	0.0119)	0.9337
	5	−0.0026	0.0066	(−0.0209,	−0.0007)	0.8410

Notes: The final columns show the average lower and upper endpoints of the 95% confidence intervals, and the proportion of times the intervals covered the true values.

The transforming function is $g(x) = \log(\Phi(x)) - \log(1 - \Phi(x))$, and the true values of the coefficients are $J_0 = J_2 = J_4 = 0, J_1 \approx 1.806, J_3 \approx 0.164$, and $J_5 \approx -0.008$. As was done with the examples in Section 4, ϵ was set to 10^{-7} so as to be small compared to $1/n$, so that all observed data would be used in the estimation of the coefficients and their variances. Table 6 shows the performance of the estimates $\hat{J}_k$. The convergence is much slower than the example using the Beta distribution, with much larger bias and MSE, particularly for $\hat{J}_5$. Also, a much larger sample size is needed for accurate coverage of the confidence intervals.

While the estimates $\hat{J}_k$, $\hat{V}_k$, and $\tilde{J}_k$ are of interest in their own right, the primary goal is to obtain an estimate of the function g. For each example and sample size, after each simulation, the estimate $\hat{g}_k(x)$ of g was calculated. Figures 2 and 3 plot, for different sample sizes, the true function g in red against plots of the mean and 5th and 95th percentiles of $\hat{g}_k$ for each value of x. A reasonable approximation of g in each of the presented examples is given by either $\hat{g}_3$ or $\hat{g}_5$ for small values of $n = 100$ or 500. For samples of $n = 1000$ or 5000, the estimated functions agree nearly exactly with g in the range $|x| < 3$, with some variation in the tails $|x| > 3$.

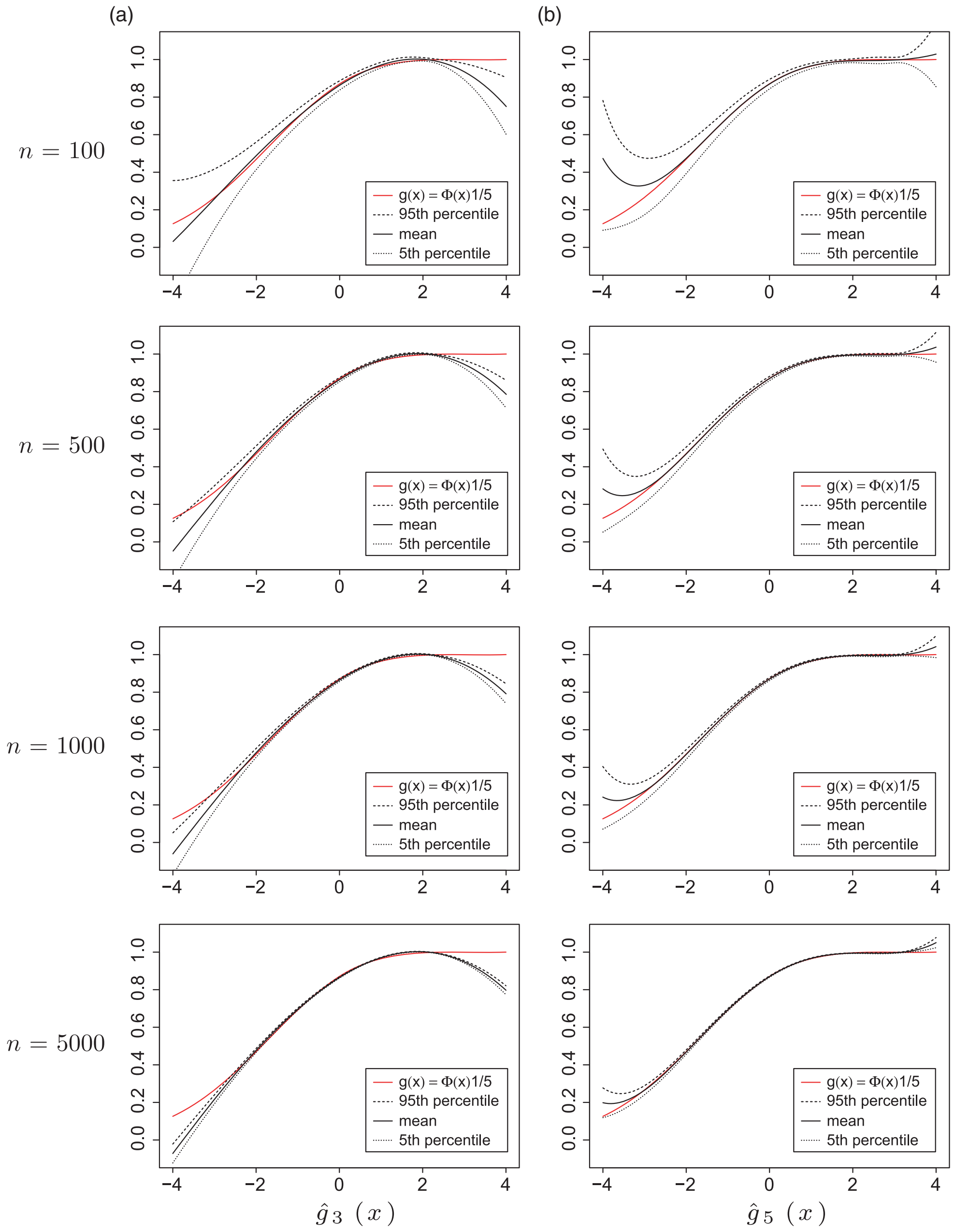

Figure 2. Summary plots of the estimates of the transformation function leading to a Beta distribution. The left column uses a third degree Hermite polynomial estimate and the right column uses a fifth Hermite degree polynomial estimate. The true transformation function is shown in red, the pointwise mean of the 100,000 estimated functions is the solid black line, and the 5th and 95th percentiles are given as dashed lines.

6. Concluding remarks

This paper presents a method for estimating continuous, monotone transformations of Gaussian random variables. The setup is very general, since any random variable X with a continuous distribution function F will be equal in distribution to $g(Z) = F^{-1}(\Phi(Z))$ for a standard Gaussian

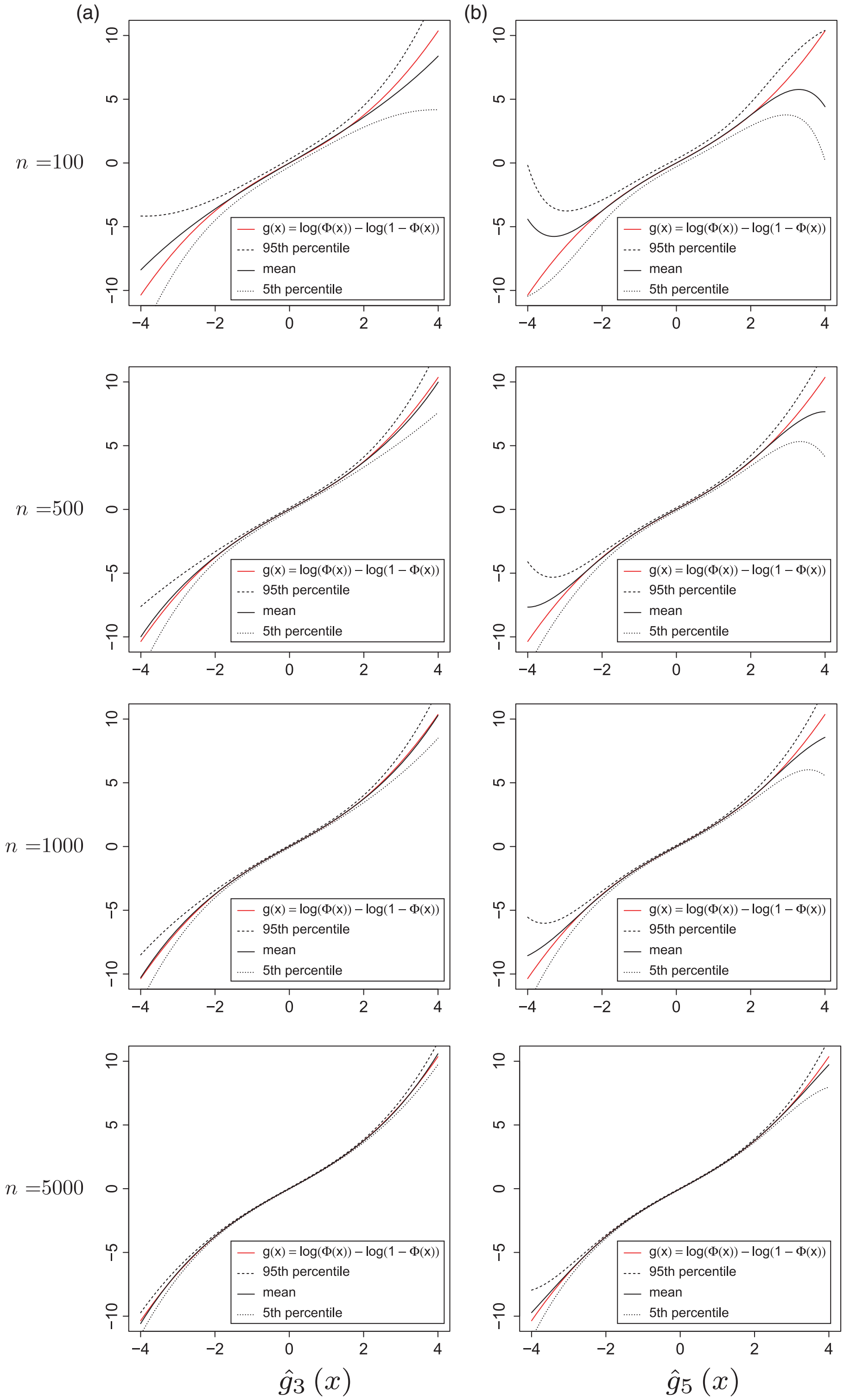

Figure 3. Summary plots of the estimates of the transformation function leading to a logistic distribution. The left column uses a third degree polynomial estimate and the right column uses a fifth degree polynomial estimate. The true transformation function is shown in red, the pointwise mean of the 100,000 estimated functions is the solid black line, and the 5th and 95th percentiles are given as dashed lines.

random variable Z. The use of Hermite polynomials allows for a basis expansion in $L^2(d\Phi)$ of the transformation function g, and after a polynomial degree is chosen, the coefficients can be consistently estimated. We have also proposed a method for estimating the degree of the Hermite polynomial to use, which sequentially checks each estimated coefficient for difference from zero. While this method is somewhat heuristic, it is intuitive and easy to implement. For situations in which the estimated transformation function is not monotone, we have shown that projection methods can be used to adjust the estimated coefficients in a minimal way to produce a monotone estimate. Furthermore, having a monotone estimate of g immediately gives a smooth estimator of the distribution, density and quantile functions.

To be useful in practice, the observations need to be identically distributed. The examples presented in this paper focus on the analysis of model residuals, which should approximately satisfy this requirement for correctly specified models. The proposed procedure also allows for a graphical method for detecting departures from normality, which is an important problem in the theory of independent component analysis (Hyvärinen and Oja 2000).

This paper only considers the analysis of univariate data. Multivariate extensions of the methods presented in this paper are challenging for two reasons. First, a multivariate quantile process must be defined, and a limit process found. In contrast to the univariate setting, there is not a universally accepted definition of a quantile for multivariate data. Einmahl and Mason (1992) discuss some possibilities for multivariate quantile processes, in terms of classes of sets, and prove functional central limit theorems for the associated processes. The second challenge is in selecting a vector version of the Hermite polynomials which, like the multivariate quantile process, is not uniquely defined. Holmquist (1996) gives a description of a class of vector Hermite polynomials, which are orthogonal with respect to the multivariate normal density. These polynomials depend on the covariance matrix, and can be defined as differentials of the multivariate normal density. However, exact expressions for these polynomials are complicated. For a particular multivariate quantile process it may be possible to find an associated set of multivariate Hermite polynomials, and to extend the results of this paper to the analysis of multivariate data sets. Due to these difficulties, it was suggested by a referee that for multivariate analysis, it may make sense to consider instead more flexible parametric models, such as mixtures of multivariate skew distributions (Lee and McLachlan 2013, 2014).

The choice of Hermite polynomials is arbitrary. One could use Appell polynomials instead, which generalise away from the Gaussian distribution – see Taniguchi and Kakizawa (2000). Although this may have advantages for certain situations, part of the appeal of the Hermite approach is the easy simulation of the Gaussian distribution. If using an Appell polynomial instead, a distribution that is easy to simulate – such as an exponential or student t – should be selected.

Acknowledgements

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. The authors wish to thank the referees and the associate editor for their constructive comments and suggestions which greatly improved the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Abramowitz, M., and Stegun, I.A. (eds.) (1965), *Handbook of Mathematical Functions*, New York: Dover.

Athreya, K.B. (1987), 'Bootstrap of the Mean in the Infinite Variance Case', *The Annals of Statistics*, 15, 724–731.

Babu, G.J., Canty, A.J., and Chaubey, Y.P. (2002), 'Application of Bernstein Polynomials for Smooth Estimation of a Distribution and Density Function', *Journal of Statistical Planning and Inference*, 11, 377–392.

Bontemps, C., and Meddahi, N. (2012), 'Testing Distributional Assumptions: A GMM Approach', *The Journal of Applied Econometrics*, 27, 978–1012.

Bontemps, C., and Meddahi, N (2005), 'Testing Normality: A GMM Approach', *The Journal of Econometrics*, 124, 149–186.

Boyd, J.P. (1984), 'Asymptotic Coefficients of Hermite Function Series', *Journal of Computational Physics*, 54, 382–410.

Bühlmann, P. (1997), 'Sieve Bootstrap for Time Series', *Bernoulli*, 3, 123–148.

Chatterjee, S., Lahiri, P., and Li, H. (2008), 'Parametric Bootstrap Approximation to the Distribution of EBLUP and Related Prediction Intervals in Linear Mixed Models', *The Annals of Statistics*, 36, 1221–1245.

Chen, C. (1995), 'Uniform Consistency of Generalized Kernel Estimators of Quantile Density', *The Annals of Statistics*, 23, 2285–2291.

Chen, C., and Parzen, E. (1997), 'Unified Estimators of Smooth Quantile and Quantile Density Functions', *The Journal of Statistical Planning and Inference*, 59, 291–307.

Csörgö M. (1983), *Quantile Processes with Statistical Applications*, Philadelphia: Society for Industrial and Applied Mathematics.

Csörgö M., and Horváth, L. (1993), *Weighted Approximations in Probability and Statistics*, New York: John Wiley & Sons.

Csörgö M., and Yu, H. (1996), 'Weak Approximations for Quantile Processes of Stationary Sequences', *Canadian Journal of Statistics*, 24, 403–430.

David, H.A., and Nagaraja, H.N. (2003), *Order Statistics* (3rd ed.), Hoboken, NJ: New York.

Einmahl, J.H., and Mason, D.M. (1992), 'Generalized Quantile Processes', *The Annals of Statistics*, 20, 1062–1078.

Fay, R.E., and Herriot, R.E. (1979), 'Estimates of Income for Small Places: An Application of James Stein Procedures to Census Data', *Journal of the American Statistical Association*, 74, 269–277.

Harville, D.A. (1977), 'Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems', *Journal of the American Statistical Association*, 72, 320–338.

Holmquist, B. (1996), 'The d-Variate Vector Hermite Polynomials of order k', *Linear Algebra and its Applications*, 238, 155–190.

Huber, P.J. (1967), 'The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions', *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 221–233.

Hyvärinen, A., and Oja, E. (2000), 'Independent Component Analysis: Algorithms and Applications', *Neural Networks*, 13, 411–430.

Ibragimov, I.A., and Rozanov, I.U.A. (1978), *Gaussian Random Processes*, New York: Springer-Verlag.

Jondeau, E., and Rockinger, M. (2001), 'Gram-Charlier Densities', *Journal of Economic Dynamics and Control*, 25, 1457–1483.

Jones, M.C. (1992), 'Estimating Densities, Quantiles, Quantile Densities and Density Quantiles', *The Annals of Mathematical Statistics*, 44, 721–727.

Lange, N., and Ryan, L. (1989), 'Assessing Normality in Random Effects Models', *Annals of Statistics*, 17, 624–642.

Lee, S., and McLachlan, G.J. (2014), 'Finite Mixtures of Multivariate Skew t-distributions: Some Recent and New Results', *Annals of Statistics.*, 24, 181–202.

Lee, S.X., and McLachlan, G.J. (2013), 'Model-based Clustering and Classification with Non-Normal Mixture Distributions (with discussion)', *Statistical Methods and Applications*, 22, 427–454.

Madan, D.B., and Milne, F. (1994), 'Contingent Claims Valued and Hedged by Pricing and Investing in a Basis', *Mathematical Finance*, 4, 223–245.

Mason, D.M. (1984), 'Weak Convergence of the Weighted Empirical Quantile Process in $L^2[0, 1]$', *Annals of Probability*, 12, 243–255.

Mason, D.M., and Shorack, G.R. (1992), 'Necessary and Sufficient Conditions for Asymptotic Normality of L-Statistics', *Annals of Probability*, 20, 1779–1804.

McElroy, T., and Holan, S. (2009), 'Spectral Domain Diagnostics for Testing Model Proximity and Disparity in Time Series Data', *Statistical Methodology*, 6, 1–20.

McMurry T. L., and Politis, D.N. (2010), 'Banded and Tapered Estimates of Autocovariance Matrices and the Linear Process Bootstrap', *Journal of Time Series Analysis*, 31, 471–482.

Menéndez, P., Ghosh, S., Künsch, H.R., and Tinner, W. (2013), 'On Trend Estimation Under Monotone Gaussian Subordination with Long-Memory: Application to Fossil Pollen Series', *Journal of Nonparametric Statistics*, 25, 765–785.

Parzen, E. (1979), 'Nonparametric Statistical Data Modeling', *Journal of the American Statistical Association*, 74, 105–121.

Politis, D.N., and Romano, J.P. (1994), 'Large Sample Confidence Intervals Based on Subsamples Under Minimal Assumptions', *Annals of Statistics*, 22, 2031–2050.

Puuronen, J., and Hyvärinen, A. (2011), "Hermite Polynomials and Measures of Non-gaussianity," in *Proceedings of the 21st International Conference on Artificial Neural Networks – Volume Part II*, Espoo, Finland, ICANN'11, Berlin, Heidelberg: Springer-Verlag, pp. 205–212.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0.

Samorodnitsky, G., and Taqqu, M.S. (1994), *Stable Non-Gaussian Random Processes: Stochastic Models With Infinite Variances*, New York: Chapman & Hall.

Sawa, T. (1978), ‘Information Criterion for Discriminating Among Alternative Regression Models’, *Econometrica*, 46, 1273–1291.

Schwartz, S.C. (1967), ‘Estimation of Probability Density by an Orthogonal Series’, *The Annals of Mathematical Statistics*, 38, 1261–1265.

Shao, J., and Tu, D. (1995), *The Jackknife and Bootstrap*, New York: Springer-Verlag.

Shorack, G.R., and Wellner, J.A. (1986), *Empirical Processes With Applications To Statistics*, New York: John Wiley & Sons.

Silvey, S.D (1959), ‘The Lagrangian Multiplier Test’, *The Annals of Mathematical Statistics*, 30, 389–407.

Tanaka, K. (1996), *Time Series Analysis: Nonstationary and Noninvertible Distribution Theory*, New York: John Wiley & Sons.

Taniguchi, M., and Kakizawa, Y. (2000), *Asymptotic Theory of Statistical Inference For Time Series*, New York: Springer.

Taqqu, M.S. (1975), ‘Weak Convergence to Fractional Brownian Motion and to the Rosenblatt Process’, *Zeitschrift fur. Wahrscheinlichkeitstheorie und verwandte Gebiete*, 31, 287–302.

van der Vaart, A.W. (1998), *Asymptotic Statistics*, New York: Cambridge University Press.

White, H. (1982), ‘Maximum Likelihood Estimation of Misspecified Models’, *Econometrica*, 50, 1–25.

Appendix. Estimation of the variance of $\hat{J}_k$

Let $h_k(u) = H_k(\Xi(u))$. From the covariance $V^{\epsilon}_{i,j}$ in Equation (4), we obtain

$$\begin{aligned}
V^{\epsilon}_{i,j} = {} & \int_{\epsilon}^{1-\epsilon} \int_{\epsilon}^{v} uq(u)h_i(u)\,\mathrm{d}u\, q(v)h_j(v)(1-v)\,\mathrm{d}v \\
& + \int_{\epsilon}^{1-\epsilon} \int_{v}^{1-\epsilon} (1-u)q(u)h_i(u)\,\mathrm{d}u q(v)h_j(v)v\,\mathrm{d}v \\
= {} & \epsilon Q(1-\epsilon)h_i(1-\epsilon) \int_{\epsilon}^{1-\epsilon} vq(v)h_j(v)\,\mathrm{d}v - \int_{\epsilon}^{1-\epsilon} vq(v)h_j(v)U_i(Q,v)\,\mathrm{d}v \\
& - \epsilon Q(\epsilon)h_i(\epsilon) \int_{\epsilon}^{1-\epsilon} (1-v)q(v)h_j(v)\,\mathrm{d}v - \int_{\epsilon}^{1-\epsilon} (1-v)q(v)h_j(v)L_i(Q,v)\,\mathrm{d}v
\end{aligned}$$

via integration by parts, where

$$\begin{aligned}
L_i(Q,v) &= \int_{\epsilon}^{v} Q(u)\,\mathrm{d}[uh_i(u)] \quad \text{and} \\
U_i(Q,v) &= \int_{v}^{1-\epsilon} Q(u)\,\mathrm{d}[(1-u)h_i(u)].
\end{aligned}$$

Using the fact that $L_i(Q,\epsilon) = 0 = U_i(Q, 1-\epsilon)$ and integration by parts again, we obtain

$$V^{\epsilon}_{i,j} = R(Q,\epsilon) + \int_{\epsilon}^{1-\epsilon} Q(v)\,\mathrm{d}[h_j(v)(vU_i(Q,v) + (1-v)L_i(Q,v))]\,\mathrm{d}v,$$

where

$$\begin{aligned}
R(Q,\epsilon) = {} & -\epsilon^2 Q(\epsilon)Q(1-\epsilon)(h_i(1-\epsilon)h_j(\epsilon) + h_i(\epsilon)h_j(1-\epsilon)) \\
& + \epsilon(1-\epsilon)(Q(\epsilon)^2 h_i(\epsilon)h_j(\epsilon) + Q(1-\epsilon)^2 h_i(1-\epsilon)h_j(1-\epsilon)) \\
& - \epsilon Q(1-\epsilon)(h_i(1-\epsilon)L_j(Q, 1-\epsilon) + h_j(1-\epsilon)L_i(Q, 1-\epsilon)) \\
& + \epsilon Q(\epsilon)(h_i(\epsilon)U_j(Q,\epsilon) + h_j(\epsilon)U_i(Q,\epsilon)).
\end{aligned}$$

Substituting $\hat{Q}$ for Q into this expression and simplifying the integrals appropriately gives, for $\epsilon < v < 1 - \epsilon$,

$$\begin{aligned}
U_i(\hat{Q}, v) &= X_{(n-M+1)}(\epsilon h_i(1-\epsilon) - (1 - \max(1 - M/n, v))h_i(\max(1 - M/n, v))) \\
&\quad + \sum_{l=[vn]+1}^{n-M} X_{(l)} \left\{ \left(1 - \frac{l}{n}\right) h_i\left(\frac{l}{n}\right) - \left(1 - \frac{l-1}{n}\right) h_i\left(\frac{l-1}{n}\right) \right\} \\
&\quad + X_{(M)}((1 - M/n)h_i(M/n) - (1 - \min(\epsilon, v))h_i(\min(v, \epsilon)))I\{v \le M/n\}, \\
L_i(\hat{Q}, v) &= X_{(M)}(\min(v, M/n)h_i(\min(v, M/n)) - \epsilon h_i(\epsilon)) \\
&\quad + \sum_{l=M+1}^{[vn]} X_{(l)} \left\{ \left(\frac{l}{n}\right) h_i\left(\frac{l}{n}\right) - \left(\frac{l-1}{n}\right) h_i\left(\frac{l-1}{n}\right) \right\} \\
&\quad + X_{(n-M+1)}(\min(v, 1-\epsilon)h_i(\min(v, 1-\epsilon)) \\
&\quad - (n-M)h_i((n-M)/n)/n)I\{v \ge 1 - M/n\}, \\
R(\hat{Q}, \epsilon) &= -\epsilon^2 X_{(M)} X_{(n-M+1)}(h_i(1-\epsilon)h_j(\epsilon) + h_i(\epsilon)h_j(1-\epsilon)) \\
&\quad + \epsilon(1-\epsilon)(X_{(M)}^2 h_i(\epsilon)h_j(\epsilon) + X_{(n-M+1)}^2 h_i(1-\epsilon)h_j(1-\epsilon)) \\
&\quad - \epsilon X_{(n-M+1)}(h_i(1-\epsilon)L_j(\hat{Q}, 1-\epsilon) + h_j(1-\epsilon)L_i(\hat{Q}, 1-\epsilon)) \\
&\quad + \epsilon X_{(M)}(h_i(\epsilon)U_j(\hat{Q}, \epsilon) + h_j(\epsilon)U_i(\hat{Q}, \epsilon)),
\end{aligned}$$

and

$$\begin{aligned}
&\int_{\epsilon}^{1-\epsilon} \hat{Q}(v)\, \mathrm{d}[h_j(v)(vU_i(\hat{Q}, v) + (1-v)L_i(\hat{Q}, v))]\, \mathrm{d}v \\
&= -\epsilon h_j(\epsilon)X_{(M)}U_i(\hat{Q}, \epsilon) + \epsilon h_j(1-\epsilon)X_{(n-M+1)}L_i(\hat{Q}, 1-\epsilon) \\
&\quad - \sum_{l=M}^{n-M} (X_{(l+1)} - X_{(l)}) h_j\left(\frac{l}{n}\right) \left(\frac{l}{n} U_i\left(\hat{Q}, \frac{l}{n}\right) + \left(1 - \frac{l}{n}\right) L_i\left(\hat{Q}, \frac{l}{n}\right)\right).
\end{aligned}$$

Combining the above terms and simplifying the expressions gives

$$\begin{aligned}
\hat{V}_{i,j}^{\epsilon} &= \sum_{l=M}^{n-M} \frac{l}{n} \frac{n-l}{n} (X_{(l+1)} - X_{(l)})^2 H_i\left(\Xi\left(\frac{l}{n}\right)\right) H_j\left(\Xi\left(\frac{l}{n}\right)\right) \\
&\quad + \sum_{l=M}^{n-M-1} \left\{ \frac{l}{n} H_j\left(\Xi\left(\frac{l}{n}\right)\right) (X_{(l+1)} - X_{(l)}) \sum_{k=l+1}^{n-M} \frac{n-k}{n} H_i\left(\Xi\left(\frac{k}{n}\right)\right) (X_{(k+1)} - X_{(k)}) \right\} \\
&\quad + \sum_{l=M+1}^{n-M} \left\{ \frac{n-l}{n} H_j\left(\Xi\left(\frac{l}{n}\right)\right) (X_{(l+1)} - X_{(l)}) \sum_{k=M}^{l-1} \frac{k}{n} H_i\left(\Xi\left(\frac{k}{n}\right)\right) (X_{(k+1)} - X_{(k)}) \right\}.
\end{aligned}$$

From the weak convergence of the quantile process and the continuous mapping theorem, we can deduce that $\hat{V}_k^{\epsilon} - V_k^{\epsilon} = O_P(n^{-1/2})$.