Quantile regression is an important tool to determine the quality level of service, product, and operation systems via stochastic simulation. It is frequently known that the quantiles of the output distribution are monotonic functions of certain inputs to the simulation model. Because there is typically high variability in estimation of tail quantiles, it can be valuable to incorporate this information in quantile modeling. However, the existing literature on monotone quantile regression with multiple inputs is sparse. In this article, we propose a class of monotone regression models, which consists of functional analysis of variance (FANOVA) decomposition components modeled with Bernstein polynomial bases for estimating quantiles as a function of multiple inputs. The polynomial degrees of the bases for the model and the FANOVA components included in the model are selected by a greedy algorithm. Real examples demonstrate the advantages of incorporating the monotonicity assumption in quantile regression and the good performance of the proposed methodology for estimating quantiles. Supplementary materials for this article are available online.

KEY WORDS: Computer experiments; Monotonic metamodel; Quantile estimation.

1. INTRODUCTION

Prediction of the output of a computer model as a function of its inputs is an important problem in stochastic simulations of service, product, and operation systems. In these simulations, the outputs often have highly nonnormal distributions. When the output is nonnormal, estimation of the quantiles of the output rather than its mean and variance is more meaningful. Emulators are useful for reducing the cost of quantile estimation in stochastic simulation. However, compared to estimation of the mean, estimation of tail quantiles typically involves more uncertainty (a larger variance for the same sample size). Thus, it is of interest to use prior knowledge about the conditional quantile function to reduce estimation variability. Frequently, it is known that the quantiles are monotonic functions of the inputs. For instance, in open Jackson network queueing systems (Gross et al. 2013), the quantiles of the waiting time distribution are increasing functions of arrival rates into the system. As another example, quantiles of the solutions of stochastic differential equations (Xiu 2010) are often monotonic in the equation parameters. A specific example is the transient heat transfer problem considered in Cengel (2002), in which the temperature of a solid body immersed in a fluid is a solution of a differential equation. Various physical properties of the system (e.g., thermal conductivity) can be viewed as uncertain and assigned a probability distribution. Physical knowledge suggests that for a solid with lower temperature than its medium, the temperature in the solid should increase with time under all conditions. Thus, the quantiles of the temperature distribution should increase with time. In these examples, a model that is consistent with the monotonicity information will be advantageous due to three reasons. First, quantile estimation accuracy is expected to improve. As pointed out by a referee, for strictly monotone functions, any consistent estimator of the function that, when differentiated, yields a consistent derivative estimator will be monotonic for large sample sizes because the derivative will be strictly positive/negative. Consequently, the imposition of monotonicity will not improve performance asymptotically. However, significant improvements can often be seen when the sample size is small. Second, confidence intervals will not be unnecessarily wide because nonmonotonic predictions are ruled out. Third, the model is more interpretable and acceptable to the practitioner than one that is nonmonotonic. Due to these advantages, statistical methods for monotonic quantile estimation are of practical value. This article addresses the problem of monotonic quantile regression in stochastic simulation.

Linear regression models have long been used as a mean model in stochastic simulation experiments (Law 2014). Recently, Gaussian process (GP) emulators (Sacks et al. 1989; Currin et al. 1991) have been popularized for stochastic simulation by Jack P.C. Kleijnen and coauthors (Kleijnen 2009). Ankenman, Nelson, and Staum (2010) proposed a modification of the GP emulator, called stochastic kriging. The GP or stochastic kriging emulators are for estimating the mean. However, in many cases, interest centers on quantile estimation. This is the case for decision-making under uncertainty and nonparametric modeling of the output distribution. For example, in the design of a service system, a decision-maker may be interested in ensuring that the 0.9 quantile of the daily average waiting time of customers does not exceed a certain value. Note that maximization of quantiles of the utility is given a sound decision theory foundation by Rostek (2010). Quantile linear regression originated from the seminal work of Koenker and Bassett (1978), and there has been much further development on the subject.
(see Koenker 2005). Reich, Fuentes, and Dunson (2011) considered the problem of spatial modeling of quantiles. Recently, Plumlee and Tuò (2014) employed the GP model with a nugget to estimate quantiles in stochastic simulation.

Research on shape-restricted regression provides methods for building monotonic models. Chang et al. (2007), McKay Curtis and Ghosh (2011), and Wang and Ghosh (2012) used Bernstein polynomial models for monotonic mean regression with a single input. There are many other methods (e.g., Ramsay 1988; Mammen et al. 2001). Most monotone models can be used for both mean and quantile estimation. He and Ng (1999) discussed monotone quantile smoothing with a single input. Kim (2006) studied linear quantile models where the coefficients are monotonic functions of an input. There is less work on monotonic regression with multiple inputs. There are many other methods (e.g., Ramsay 1988; Ghosh 2011, and Wang and Ghosh 2012) used Bernstein regression with multiple inputs. Wang and Ghosh (2011) used multivariate Bernstein polynomials. Gluhovsky (2006) modified main effects and two-factor interactions with I-splines. Few articles explicitly consider monotone quantile regression with multiple inputs. Chernozhukov, Fernandez-Val, and Gallchon (2009) proposed a general method, called rearrangement, to monotize a nonmonotonic estimate of a function. Main-effects-only models (Koenker 2005) can be used but they are restrictive.

This article employs monotonic regression models obtained by modeling functional analysis of variance (FANOVA) components with Bernstein polynomial bases, called FANOVA Bernstein models, for quantile estimation in stochastic simulation. Stochastic simulation problems typically consist of multiple inputs and the output is often known to be monotonic functions of some of the inputs. Moreover, we can expect only a handful of inputs to have large effects on the output and a small subset of the large number of possible interactions between inputs to be significant. Thus, it is often of interest to build a monotonic quantile regression model with multiple inputs that includes only important main effects and interactions between inputs. We propose an approach based on identifying important FANOVA components by modeling each component using Bernstein polynomials. Monotonicity is guaranteed by imposing certain linear restrictions on the model coefficients. The degree of the Bernstein bases used and the FANOVA components are chosen by minimizing the Bayesian information criterion (BIC; Lian 2012).

The rest of the article is organized as follows: Section 2 reviews the univariate and multivariate Bernstein polynomials. Section 3 introduces the FANOVA Bernstein model and shows how monotonicity can be imposed. Section 4 discusses the monotone quantile regression problem. Section 5 proposes an algorithm for model selection, and discusses design and statistical inference issues. Two examples are given in Section 6. Section 7 concludes the article.

2. UNIVARIATE AND MULTIVARIATE BERNSTEIN POLYNOMIALS

The univariate Bernstein polynomials/bases of degree \( N \) on \([0, 1]\) are given by

\[
B(z, l, N) = \binom{N}{l} z^l (1 - z)^{N - l}, \quad l = 0, 1, \ldots, N. \tag{1}
\]

The univariate Bernstein model employed by Wang and Ghosh (2012) and others is

\[
f(z) = \sum_{l=0}^{N} \beta_l B(z, l, N). \tag{2}
\]

Because \( f'(z)/N = \sum_{l=0}^{N-1} (\beta_{l+1} - \beta_l) B(z, l, N - 1) \), the restrictions

\[
\beta_{l+1} - \beta_l \geq 0, \quad l = 0, 1, \ldots, N - 1 \tag{3}
\]

ensure that \( f'(z) \geq 0 \) for all \( z \in [0, 1] \).

For monotonic regression with \( d > 1 \) inputs on \([0, 1]^d\), Wang and Ghosh (2011) employed the multivariate Bernstein polynomials/bases \( B_d(z, l, N) = \prod_{j=1}^{d} B(z_j, l_j, N_j) \), for quantile estimation in stochastic simulation. Reich, Fuentes, and Dunson (2011) proposed a general method, called rearrangement, to monotonize a nonmonotonic estimate of a function. Main-effects-only models (Koenker 2005) can be used but they are restrictive.

A FANOVA decomposition of \( F : [0, 1]^d \rightarrow \mathbb{R} \) is given by

\[
F(z) = F_0 + \sum_{i=1}^{d} F_i(z_i) + \sum_{j=1}^{d} F_{i,j}(z_i, z_j) + \cdots + F_{1,\ldots,d}(z), \quad \text{where we call } F_0 \text{ the intercept, } F_{(i)} \text{ a main}
\]

3. FANOVA DECOMPOSITION AND MODELING

Theorem 1. Let \( L : [0, 1]^d \rightarrow \mathbb{R} \) be continuous. Then, \( f_N(z) = \sum_{l \in \mathcal{P}(N)} F(l/N) B_d(z, l, N) \), where \( I/N = (l_1/N_1, \ldots, l_d/N_d) \), converge uniformly to \( F \) as \( \min\{N_1, \ldots, N_d\} \rightarrow \infty \).

Let \( \beta = (\beta_i : I \in \mathcal{P}(N)) \) and \( \mathcal{S}_i(N) \) be the set of \( \beta_i \)'s that satisfy (5). If \( F \) is increasing in \( z_{k_1}, \ldots, z_{k_l} \), setting \( \beta_i = F(I/N) \) yields a \( \beta \) that is in \( \bigcap_{i=1}^{d} \mathcal{S}_i(N) \), that is, an \( f \) that is increasing in \( z_{k_1}, \ldots, z_{k_l} \). This observation and Theorem 1 imply that there exists \( \beta_N \in \bigcap_{i=1}^{d} \mathcal{S}_i(N) \) such that \( ||f_N - F||_{\infty} \rightarrow 0 \) for any sequence \( N_1, N_2, \ldots \) with \( \min\{N_1, \ldots, N_d\} \rightarrow \infty \), where \( N_i = (N_{i1}, \ldots, N_{id}) \). This justifies the use of constraints (5) to achieve monotonicity.

One important limitation of model (4) is that the number of bases increases very fast with the dimension \( d \). Thus, a large experiment is needed to fit the model when \( d \) is large.

The linear restrictions on the coefficients (5) are a sufficient but not necessary condition for monotonicity. Thus, coefficients that give a monotonic \( f \) may not satisfy (5) and it is not clear that (5) allows sufficient flexibility to model a large class of functions. A justification for imposing (5) is the following theorem, stated in page 115 of Gal and Anastassiou (2010).

Theorem 1. Let \( F : [0, 1]^d \rightarrow \mathbb{R} \) be continuous. Then, \( f_N(z) = \sum_{l \in \mathcal{P}(N)} F(l/N) B_d(z, l, N) \), where \( I/N = (l_1/N_1, \ldots, l_d/N_d) \), converge uniformly to \( F \) as \( \min\{N_1, \ldots, N_d\} \rightarrow \infty \).

Let \( \beta = (\beta_i : I \in \mathcal{P}(N)) \) and \( \mathcal{S}_i(N) \) be the set of \( \beta_i \)'s that satisfy (5). If \( F \) is increasing in \( z_{k_1}, \ldots, z_{k_l} \), setting \( \beta_i = F(I/N) \) yields a \( \beta \) that is in \( \bigcap_{i=1}^{d} \mathcal{S}_i(N) \), that is, an \( f \) that is increasing in \( z_{k_1}, \ldots, z_{k_l} \). This observation and Theorem 1 imply that there exists \( \beta_N \in \bigcap_{i=1}^{d} \mathcal{S}_i(N) \) such that \( ||f_N - F||_{\infty} \rightarrow 0 \) for any sequence \( N_1, N_2, \ldots \) with \( \min\{N_1, \ldots, N_d\} \rightarrow \infty \), where \( N_i = (N_{i1}, \ldots, N_{id}) \). This justifies the use of constraints (5) to achieve monotonicity.
effect, \( F_{(i,j)} \) a two-factor interaction, and similarly for the other components of the decomposition. Such decompositions can be employed to improve modeling of \( F \) because we can assume that a small subset of main effects and two-factor interactions are important. This assumption is supported by the effect sparsity and hierarchy principles (Wu and Hamada 2009). Efron and Stein (1981) pointed out that the FANOVA decomposition is unique if \( f \) has independent components and the conditions \( E_{1j} [F_{i1,\ldots,i_j}(z_{i_1}, \ldots, z_{i_j})] = 0, \ j = 1, \ldots, e \) are imposed, where the expectation is taken with respect to \( z_{i_j} \) while holding other inputs fixed. This decomposition is an important tool in global sensitivity analysis (Saltelli, Chan, and Scott 2000) and other inputs fixed. This decomposition is an important tool in global sensitivity analysis (Saltelli, Chan, and Scott 2000) and functional data analysis (Ramsay and Silverman 2005). This article adopts a different FANOVA decomposition obtained by imposing a different set of conditions that allow easier modeling with the multivariate Bernstein bases.

**Theorem 2.** Let \( F : [0, 1]^d \to \mathbb{R} \) be a function. Then, \( F \) has a unique FANOVA decomposition:

\[
F(z) = F_0 + \sum_{i=1}^{d} F_{(i)}(z_i) + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} F_{(i,j)}(z_i, z_j) + \cdots + F_{(1, \ldots, d)}(z) \tag{6}
\]

if \( F_{(1, \ldots, i, \ldots, j)}(x_1, \ldots, x_e) = 0 \forall (x_1, \ldots, x_e) \in \{ \xi \in [0, 1]^e : \prod_{e=1}^{d} \xi_e = 0 \}, \ e = 1, \ldots, d. \)

It is a simple matter to see that \( F_{(1, \ldots, i, \ldots, j)}(z_1, \ldots, z_d) \) can be approximated with a linear combination of multivariate Bernstein polynomials

\[
f_{(1, \ldots, i, \ldots, j)}(z_1, \ldots, z_d) = \sum_{(l_1, \ldots, l_e) \in \mathcal{Q}(N_1, \ldots, N_e)} \alpha_{(l_1, \ldots, l_e)} B_{e}(z_1, \ldots, z_d), \tag{7}
\]

where \( \mathcal{Q}(N_1, \ldots, N_e) = \prod_{e=1}^{d} \{1, \ldots, N_e\}. \) Note that the coefficients of the bases \( B_{e}(z_1, \ldots, z_d) \) are set to zero so that \( f_{(1, \ldots, i, \ldots, j)}(z_1, \ldots, z_d) = 0 \forall (z_1, \ldots, z_d) \in \{ (\xi_1, \ldots, \xi_e) \in [0, 1]^e : \prod_{e=1}^{d} \xi_e = 0 \}. \) Thus, it is very natural and convenient to model the FANOVA components in (6) with Bernstein bases.

Note that \( \sum_{l=0}^{N_1} B(z_1, L, N) = 1 \) so that an alternative set of bases with the same span, which we call the modified Bernstein bases is

\[
b(z, 0, N) = 1, \ b(z, l, N) = B(z, l, N), l = 1, \ldots, N. \tag{8}
\]

If we use the tensor product of the modified Bernstein bases (8) instead of the tensor product of the Bernstein bases (as in (4)) for function approximation, we obtain a model

\[
f(z) = \sum_{l \in \mathcal{L}(N)} \alpha_l b_d(z, l, N) \tag{9}
\]

that decomposes into the FANOVA components given in (7). This is because \( b(z_j, 0, N_j) = 1. \)

Sufficient conditions that ensure monotonicity of \( f(z) \) with respect to \( z_i \) can be obtained as follows: Since \( \{ B(z, 0, N), \ldots, B(z, N, N) \} \) and \( \{ b(z, 0, N), \ldots, b(z, N, N) \} \) have the same span, model (9) can be rewritten in terms of the bases \( \{ B(z_1, l_1, N_1) \cdots B(z_d, l_d, N_d) : l_1 = 0, \ldots, N_1, i = 1, \ldots, d \}, \) which gives model (4). By equating (9) and (4), we obtain (see Appendix A)

\[
\beta_{l_1, \ldots, l_d} = \sum_{e=1}^{d} \sum_{j_e \in \{0, \ldots, N_e\}} \alpha_{(l_1, \ldots, l_e)} \tag{10}
\]

From (10), we obtain the following theorem, which gives linear restrictions on the \( \alpha_{(l_1, \ldots, l_e)}'s \) that are sufficient to achieve monotonicity in an axial direction.

**Theorem 3.** Model (9) is increasing in \( z_k \) if

\[
\sum_{e=1}^{d} \sum_{j_e \in \{0, \ldots, N_e\}} \alpha_{(l_1, \ldots, l_e)} - \sum_{j_e \in \{0, \ldots, N_e\}} \alpha_{(l_1, \ldots, l_e)} = 0, \quad k = 0, \ldots, N_k - 1, l_e \in \{0, 1, \ldots, N_e\}, \ e \neq k. \tag{11}
\]

**Proof.** It follows from (10) that (9) with constraints (11) is equivalent to (4) with constraints (5).

Since it is often the case that only main effects and two-factor interactions are significant, we shall assume that \( F(z) = F_0 + \sum_{i=1}^{d} F_{(i)}(z_i) + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} F_{(i,j)}(z_i, z_j) \) in the remainder of the article. Proposition 1 demonstrates that the main effect \( F_k \) plays an important role in ensuring that \( F \) is increasing in \( z_k. \) It must be increasing. It must also offset all decreases in two-factor interactions involving input \( k, \) that is, the positive increment in \( F_k \) must be at least equal in magnitude to the sum of all negative increments in two-factor interactions involving input \( k. \)

**Proposition 1.** Suppose \( F(z) = F_0 + \sum_{i=1}^{d} F_{(i)}(z_i) + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} F_{(i,j)}(z_i, z_j) \) and \( F(z_1, \ldots, z_{k-1}, z_k + \Delta z_k, z_{k+1}, \ldots, z_d) - F(z) \geq 0 \forall z \in [0, 1]^d, \Delta z_k \in [0, 1 - z_k]. \)

Then,

\[
F_k(z_k + \Delta z_k) - F_k(z_k) \geq \sup_{z_k \in [0, 1]} [F_{(k,j)}(z_k, z_j) - F_{(k,j)}(z_k + \Delta z_k, z_j)] \tag{12}
\]

for all \( z_k \in [0, 1] \) and \( \Delta z_k \in [0, 1 - z_k], \) where \( z_{-k} = (z_1, \ldots, z_{k-1}, z_{k+1}, \ldots, z_d). \)

**Remark.** By abuse of notation, we let \( F_{(k,j)}(z_k, z_j) = F_{(i,j)}(z_k, z_j), j < k \) in this article.

The full FANOVA Bernstein model (with all main effects and interactions) is given by

\[
f_N(z) = f_0 + \sum_{i=1}^{d} f_{(i)}(z_i) + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} f_{(i,j)}(z_i, z_j). \tag{13}
\]
where \( f_0 = \alpha_0 \) is a constant,

\[
\begin{align*}
\tilde{f}_{i}(z_j) &= \sum_{l=1}^{N_i} \alpha_{l}^{(i)} B(z_j, l, N_i) + f_{i,j}(z_i, z_j) \\
&= \sum_{m=1}^{N_i} \sum_{l=1}^{N_i} \alpha_{l,m}^{(i,j)} B(z_i, l, N_i) B(z_j, m, N_j).
\end{align*}
\] (14)

Note that the Bernstein bases \( B(z_j, l, N_i) \) are used to model \( f_{i,j} \) and their products with \( B(z_j, m, N_j) \) are used as bases to model \( f_{i,j} \). The model (13) will be employed as the full model in the remainder of the article. Corollary 1 simplifies the result in Theorem 3 for model (13).

**Corollary 1.** Let \( f_N(z) \) be given by (13)–(14) and define

\[
\alpha_0^{(k,i)} = \alpha_0^{(k,i,0)} = 0 \quad \text{then,} \quad f_N(z) \text{ is increasing in } z_k
\]

\[
\alpha_{l+1}^{(k,i)} - \alpha_l^{(k,i)} + \sum_{j=1, j \neq k}^{d} \left[ \alpha_{l+1}^{(k,j)} - \alpha_l^{(k,j)} \right] \geq 0, l = 0, 1, \ldots, N_k - 1, m_j = 0, 1, \ldots, N_j. \hspace{1cm} (15)
\]

**Remark.** Suppose that \( \alpha_0^{(k,i)} = F_{l}(l/N_k) \) and \( \alpha_0^{(k,i,j)} = F_{l}(l/N_k) m_j/N_j \), then it can clearly be seen that (15) arise from imposing the monotonicity of \( F \) on the grid \( \{0, 1/N_k, \ldots, 1\} \). In particular, \( \alpha_{l+1}^{(k,i)} - \alpha_l^{(k,i)} + \sum_{j=1, j \neq k}^{d} \alpha_{l+1}^{(k,j)} - \alpha_l^{(k,j)} \geq 0 \), \( z_i, z_j \) is continuous (which implies that every \( F_{l}(l/N_k) \) and \( F_{l}(l/N_k) m_j/N_j \) is continuous). Suppose also that \( F \) is increasing in the inputs \( k_1, \ldots, k_L \). Let

\[
\tilde{f}_N(z) = f_0 + \sum_{i=1}^{d} \tilde{f}_{i}(z_i) + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} \tilde{f}_{i,j}(z_i, z_j), \hspace{1cm} (16)
\]

where \( f_0 = \alpha_0 \), \( \tilde{f}_{i}(z_i) = \sum_{l=1}^{N_i} F_{l}(l/N_k) B(z_i, l, N_i) \), and \( \tilde{f}_{i,j}(z_i, z_j) = \sum_{l=1}^{N_i} \sum_{m=1}^{N_i} F_{l}(l/N_k) m_j/N_j B(z_i, l, N_i) B(z_j, m_j/N_j) \). Then, the coefficients of \( \tilde{f}_N \) that is, the \( F_{l}(l/N_k) \)'s and \( F_{l}(l/N_k) m_j/N_j \)'s satisfy (15) for all \( k \in \{k_1, \ldots, k_L\} \), which implies that \( \tilde{f}_N \) is increasing in the inputs \( k_1, \ldots, k_L \). In addition, \( \tilde{f}_N \) converges uniformly to \( F \).

**Proof.** This follows from Theorem 1, Theorem 2, and Corollary 1.

### 4. MONOTONE QUANTILE REGRESSION MODEL

Let \( Y_z \) denote the stochastic simulation output given the vector of model inputs \( z \) and let \( f_z: z \rightarrow \mathbb{R} \) denote the function that maps \( z \) to the \( \tau \)th quantile. Thus,

\[
Y_z = f_z(z) + \varepsilon, \hspace{1cm} (17)
\]

where \( P(\varepsilon \leq 0 | z) = P(Y_z \leq f_z(z)) = \tau \). In this article, we shall consider the problem of estimating the conditional quantile function \( f_z \) in cases where \( f_z \) is known to be a monotonic function of one or more of the inputs. This problem is widely encountered but knowledge of monotonicity is often neglected in statistical modeling of \( f_z \).

Numerous physical relationships are monotonic and of the form \( \pi_0 = \prod_{j=1}^{p} \phi(\pi_j) \), where the \( \pi_j \)'s are either physical quantities or dimensionless parameters (Barenblatt 1996). Taking logarithm on both sides of the equations yields an additive relationship. Additive models have also been used in many other areas of statistical modeling (Hastie and Tibshirani 1990). These examples provide support for modeling monotonic relationships with only main effects and two-factor interactions, in addition to the effect hierarchy principle (Wu and Hamadah 2009). Moreover, the effect sparsity principle suggests that only a subset of main effects and two-factor interactions are important. Thus, we shall adopt a model selection approach that selects important main effects and two-factor interactions while enforcing monotonicity constraints.

It is possible to model \( f_z \) using (4). However, this restricts the applicability of the method to low-dimension problems because the number of basis functions is \( M = \prod_{i=1}^{d} (N_i + 1) \), which can be very large. Note that we need a design of size at least \( M \) to model the correct model, which can be costly because the computer model simulations are time consuming. In contrast, even for the full FANOVA Bernstein model (13), the number of basis functions is \( N = 1 + \sum_{j=1}^{d} N_j + \sum_{j=1}^{d-1} \sum_{j=1}^{d} N_j N_j \). If \( N_1 = \cdots = N_d = N \), \( M = (N + 1)^d \) and \( N = 1 + dN + \left( \frac{d}{2} \right)N^2 \). For example, if \( N = 5 \) and \( d = 4 \), \( M = 1296 \) and \( N = 171 \). Due to the typically higher signal-to-noise ratio in stochastic simulation compared to physical experiments, the values of \( N_1, \ldots, N_d \) frequently need to be larger than one to model the true functions well with (4) or (13). Thus, the number of basis functions needed to model monotonic functions accurately using (4) with constraints (5) will often be too large. The number of constraints in (15) is \( N \prod_{i=1}^{d} (N_i + 1) \). If monotonicity in all \( d \) inputs is desired, we need \( \sum_{k=1}^{d} N_k \prod_{i=1}^{d} (N_i + 1) \) constraints. However, this number can be significantly reduced if only a subset of main effects and two-factor interactions in (13) are nonzero. If input \( k \) does not interact with inputs \( k_1, \ldots, k_L \), then we would only need \( N_k \prod_{i=1}^{d} (N_i + 1) \) constraints to enforce monotonicity in input \( k \). The approach in this article incorporates selection of main effects and two-factor interactions, which often significantly reduces the number of constraints needed to impose monotonicity.

We employ quantile regression models of the form (13) that allow for selection of main effects and two-factor interactions. The proposed FANOVA Bernstein model is given by

\[
\begin{align*}
f_z(z)(\delta, N) &= \alpha_0 + \sum_{i=1}^{d} \sum_{l=1}^{N_i} \alpha_{l}^{(i)} B(z_i, l, N_i) + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} \delta_{i,j} \\
&= \sum_{m=1}^{N_i} \sum_{l=1}^{N_i} \alpha_{l,m}^{(i,j)} B(z_j, l, N_i) B(z_j, m, N_j), \hspace{1cm} (18)
\end{align*}
\]

where \( \delta_{i,j} \) takes on the value 0 or 1. Note that if \( N_i = 0 \), then the main effect of input \( i \) is excluded from the model. In ad-
dition, because the same $N_i$ is used to model $f_{i(j)}$ and $f_{0(j)}$, the interaction $f_{i(j)}$ can be included in the model only when both $f_{i(j)}$ and $f_{0(j)}$ are included in the model, that is, $N_i > 0$ and $N_j > 0$. The value of $\delta_{i(j)}$ determine whether $f_{i(j)}$ is actually included. This incorporates the strong heredity principle, which states that a two-factor interaction can be in the model only if both of its parent main effects are active (Chipman 1996; Wu and Hamada 2009). Given $N$ and $\delta = (\delta_{1(2)}, \ldots, \delta_{d-1,d})$, $f_\tau$ is a linear model with coefficients $a_l^{(i)}$ corresponding to active main effects ($N_i > 0$) and $a_{l,mj}^{(i)}$ corresponding to active two-factor interactions ($N_i > 0$, $N_j > 0$, $\delta_{i(j)} = 1$). To force $f_\tau(z)|\{\delta, N\}$ to be increasing in $z_k$, we use (15). Note that if $N_k > 0$ and $\delta_{k,j} = 0$ for all $j \in J$, then (15) can be replaced with $a_{l+1}^{(k)} + \sum_{j \in J} |k| |j| f_{l+1,mj}^{(k,j)} - a_{l,mj}^{(k,j)} | \geq 0, \quad l = 0, 1, \ldots, N_k - 1, m_j = 0, 1, \ldots, N_j$.

For simplicity of notation, we order and rewrite the nonzero coefficients in (18) as $a_0, a_1, \ldots, a_M$. Let $\alpha = (a_0, a_1, \ldots, a_M)^T$, and $x(z) = (1, x_1(z), \ldots, x_M(z))$, where $x_i(z)$ denotes the value of the basis function corresponding to $\alpha_i$ at $z$. Given an experiment design $D = \{z^1, \ldots, z^n\} \subset [0, 1]^d$ and corresponding output data $Y = (Y_1, \ldots, Y_n)^T$, $\alpha$ can be estimated by minimizing the check loss

$$\min_{\alpha} \left[ \sum_{i=1}^n L_\tau(Y_i - x(z^i)^T \alpha) \right] \quad \text{s.t.} A \alpha \geq 0,$$

(19)

The problem (19) can be rewritten as a linear program

$$\min_{\alpha, v, w} \left[ \tau \mathbf{1}_n^T v + (1 - \tau) \mathbf{1}_n^T w \right] \quad \text{s.t.} A \alpha + v - w = Y, \quad v \geq 0, w \geq 0,$$

(20)

In (20), $X$ is a matrix with rows $x(z^1), \ldots, x(z^n)$ and $\mathbf{1}_n$ is an $n \times 1$ vector of 1’s.

We shall provide a justification for the approach given by (19) and (20). Suppose that there are $t$ distinct design points $z^1, \ldots, z^t$ each replicated $q$ times. As $q \to \infty$, the objective function in (19) normalized by $n = tq$ converges to $\sum_{i=1}^t \int E[L_\tau(Y_i - x(z^i)^T \alpha)]/t$, where $Y_i$ denotes the output at $z^i$. Thus, the integrated check loss $f_{[0,1]^d} E[L_\tau(Y_i - f_\tau(z)|\delta, N)]w(z)dz$ of $f_\tau(z)$, where $w$ is a density on $[0, 1]^d$, can be viewed as the population quantity being estimated by the objective function in (19). Let $\bar{F}(z) = f_\tau(z)|\{\delta = 1, N\}$, which is a model that includes all main effects and two-factor interactions when $N > 0$. Theorem 4 states that minimizing $f_{[0,1]^d} E[L_\tau(Y_i - \bar{F}(z)) - L_\tau(Y_i - F(z))]d\tau$ subject to (15), where $F$ is the true conditional quantile function, yields a quantile regression model $\bar{F}_N$ that achieves the minimum possible integrated check loss (which is achieved by $F$) as $\min\{N_1, \ldots, N_d\} \to \infty$.

Theorem 4. Let $F : [0, 1]^d \to \mathbb{R}$ be a continuous function that is increasing in $z_k$, $k \in \Omega \subset \{1, \ldots, d\}$ and $\omega$ be a density on $[0, 1]^d$. Let $\bar{Y}_i = F(z) + \varepsilon_i$, where $P(\varepsilon \leq 0) = P(Y_i \leq F(z)) = \tau_i \in (0, 1)$. Assume that $h_{\tau_i}(y)w(y)$ is the probability density function of $(Y_i, z_i)$. Denote by $\bar{F}_N(z)$ the function $\bar{F}(z) = f_\tau(z)|\{\delta = 1, N\}$ (see (18)) with $\alpha$ set to $\tilde{\alpha} = \arg\min_{\alpha} \int [0,1]^d E[L_\tau(Y_i - \bar{F}(z))]w(z)d\tau$ (15) holds for all $k \in \Omega, \alpha \in \mathcal{S}$, where $\mathcal{S}$ is a compact set that contains the coefficients of $\bar{F}_N$.

Then,

$$\lim_{\min\{N_1, \ldots, N_d\} \to \infty} \int [0,1]^d E\left[ L_\tau(Y_i - \bar{F}_N(z)) - L_\tau(Y_i - F(z)) \right] w(z)d\tau = 0.$$

5. MODEL SELECTION, STATISTICAL INFERENCE, AND DESIGN

For practical purposes, we need a data-driven procedure to determine the values of $N_1, \ldots, N_d$ and the $\delta_{i(j)}$’s. By increasing $N$, model (18) can approximate functions of increasingly complex shapes. On the other hand, setting $N_j = 0$, model (18) will be independent of $z_j$ and by setting $\delta_{k,j} = 0, \forall j \neq k$, it will be additive in input $k$. Thus, by selecting appropriate $N$ and $\delta$, (18) can be used to model a wide variety of functions with moderately large number of inputs.

We use the BIC (Lian 2012) for comparing models with different $N$ and $\delta$. The BIC is

$$\text{BIC} = \ln \left[ \sum_{i=1}^n L_\tau(Y_i - x(z^i)^T \alpha) \right] + |\tilde{\alpha}| \ln n/(2n),$$

(21)

where $\tilde{\alpha}$ is the optimal solution to (20) and $|\tilde{\alpha}|$ is the number of elements in $\tilde{\alpha}$. The BIC has been proven to be asymptotically consistent for model selection under certain conditions. We set $\text{BIC} = \infty$ for all models with model matrix $X$ that is not of full column rank. Without any monotonicity constraints, the model coefficients are not identifiable and we can reduce the number of columns of $X$ (which is equal to $|\tilde{\alpha}|$) without changing $\sum_{i=1}^n L_\tau(Y_i - x(z^i)^T \alpha)$. The situation is more complicated when there are monotonicity constraints. However, if there exists an optimal solution $\tilde{\alpha}$ to (20) such that $|\tilde{\alpha}| > 0$, then there are multiple optimal solutions to (20). Thus, it makes sense to rule out models that give an optimal solution to (20) and $|\tilde{\alpha}| > 0$, which is a model that includes all main effects and two-factor interactions when $N > 0$.

Because the number of models is huge, we use a heuristic optimization approach to search for the best model. We suppose that $N_i \leq N_{\max}, i = 1, \ldots, d$, where $N_{\max}$ is an upper bound specified by the modeler. The model search algorithm is given in Figure 1.

TECHNOMETRICS, MAY 2016, VOL. 58, NO. 2
For each bootstrap replication, we sample finds the best model. When the \((i, j)\) interaction is added, the algorithm also finds the best \(N_i\) conditioned on \(N_k = N_k^*\), \(k \neq i\) and the best \(N_j\) conditioned on \(N_k = N_k^*\), \(k \neq j\). The reason for this step is that if \(N_i^*\) and \(N_j^*\) are both large, then adding the \((i, j)\) interaction into the model will increase the number of terms by \(N_i^* N_j^*\), which is huge. This may give a model with far too many terms (perhaps even more than the number of experiment runs) and a high BIC value. As such, the \((i, j)\) interaction may not be able to enter the model even though it is significant. By reducing \(N_i\) or \(N_j\), the \((i, j)\) interaction can be allowed to enter the model. The best option involving two-factor interactions is removed from and added to the model in each bootstrap sample. A best option involving two-factor interactions is adopted each time Step 3 is performed. A best option can involve removing interaction \((i^*, j^*)\), adding interaction \((i^*, j^*)\) and changing \(N_i\), or adding interaction \((i^*, j^*)\) and changing \(N_j\). This process is repeated until no improvements in the BIC can be made, which leads to termination of the algorithm. Note that existing model selection methods cannot be used for the problem in this article. The LASSO and other penalties (Wu and Liu 2009) can be used to select model terms but not the polynomial degrees \(N\). Moreover, they do not impose the effect heredity principle, which have been found improve model selection when the number of model terms far exceeds the number of distinct experiment runs (Chipman 1996), as in the problems considered in this article.

To quantify model and parameter estimation uncertainty, we employ a bootstrap procedure. In stochastic simulations, it is common and sensible to employ replicated designs, where each distinct input setting is replicated \(q\) times. Let the ordered observations at a distinct design point \(z^i\) be denoted by \(Y_{(1)}, \ldots, Y_{(q)}\). Then, we estimate the quantile function \(\xi(\tau) = f_1(z^i)\) with the piecewise linear function that connects the points \((0, Y_{(1)} - 0.5), (Y_{(2)} - Y_{(1)})/(0.5/q), Y_{(1)}, \ldots, (q - 0.5)/q, Y_{(q)}\)), \((1, Y_{(q)} + 0.5Y_{(q)} - Y_{(q-1)})\)). The first point is obtained by continuing the linear trend between \((0.5/q, Y_{(1)})\) and \((1.5/q, Y_{(2)})\) and similarly for the last point. For each bootstrap replication, we sample \(q\) uniform random numbers on \([0, 1]\) and compute the value of the piecewise linear quantile function at the \(q\) numbers to obtain \(q\) random samples of the response. This is called a smoothed bootstrap (Silverman and Young 1987). The design is fixed and not resampled. Using each bootstrap sample of the output as data, we find the best model using the model search algorithm in Figure 1. The collection of best models found with the procedure gives an indication of model uncertainty. In addition, a \(100(1 - \theta)\%\) confidence interval for a quantile is constructed from the \(1 - \theta/2\) and \(1 - \theta/2\) sample quantiles of the estimates given by the best model in each bootstrap sample.

Optimal designs for the FANOVA Bernstein model is a complex problem that we shall leave for further research. In the examples, we employ the cosine maximin Latin hypercube design (cosine MLHD) proposed by Dette and Pepelyshev (2010) and Tan (2015). Both Dette and Pepelyshev (2010) and Tan (2015) demonstrated that the cosine MLHD improves performance of GP models. Furthermore, Tan (2015) showed that cosine MLHDs are good for fitting polynomial models. Since (18) is a polynomial model, cosine MLHDs should be a good choice for fitting the model. We suggest that the number of replicates \(q\) for the design be chosen so that the expected number of observations above or below the quantile to be estimated is at least one, that is, \(q \min[\tau, 1 - \tau] \geq 1\). This will allow reasonably reliable estimation of the quantile at each distinct design setting using only the data for that setting.

When estimating multiple conditional quantile functions, the problem of quantile crossing can occur, that is, an estimated \(\tau_1\) quantile can be larger than an estimated \(\tau_2\) quantile at a given \(z\) even though \(\tau_1 < \tau_2\). This problem can be overcome by rearranging the quantile estimates, as proposed by Chernozhukov, Fernandez-Val, and Galichon (2009).

6. EXAMPLES

This section presents two examples in which quantiles of an M/M/1 queueing system and a real model are estimated. Appendix B gives a numerical example based on a more complex queueing system. In each example, we compare the performance of the FANOVA Bernstein model to the stationary GP model with Gaussian correlation function and a nugget term employed by Plumlee and Tu (2014), which we call the quantile GP model. The quantile GP parameters are estimated by the method of maximum likelihood (Currin et al. 1991),
using the sample quantiles at distinct design points as data. Since the quantile GP is nonmonotonic, we also consider rearranging the quantile GP predictions and credible limits (Chernozhukov, Fernandez-Val, and Galichon 2009). This procedure is attractive because Chernozhukov, Fernandez-Val, and Galichon (2009) proved that it does not worsen and can improve prediction accuracy and coverage. Computational implementation of rearrangement requires monotonizing the function estimate on a grid. Fine grids are needed to approximate the theoretical rearrangement estimator well. This is too computationally intensive when there are more than two inputs. Thus, we apply the rearrangement procedure on a coarse grid \([0, 0.1, \ldots, 1]^d\) for Examples 2 and 3, which have \(d = 5\) and \(d = 4\), respectively. The average rearrangement estimator suggested by Chernozhukov, Fernandez-Val, and Galichon (2009) is employed, which requires rearranging with respect to all \(d!\) ordering of the inputs. The grid partitions the input space \([0, 1]^d\) into cells, which are the smallest hypercubes with grid points as corners. To predict at a point not on the grid, we use model (4) with \(N = (1, \ldots, 1)\) fitted by interpolating the rearranged predictions at the corners of the cell containing the point. To fit the model, the inputs are rescaled so that the cell is \([0, 1]^d\). This gives a continuous monotonic predictor. Note that Chernozhukov, Fernandez-Val, and Galichon (2009) did not provide any suggestion on the choice of grid size and monotonic interpolator.

It is possible to fit a monotonic model with main effects and two-factor interactions that are modeled with linear splines. However, linear splines are not differentiable and can be poor approximations of differentiable computer models widely seen in practice. Linear spline methods proposed by Koenker, Ng, and Portnoy (1994) and Koenker and Mizera (2004) are implemented in the quantreg package in R. The quantreg package only has the capability to fit a model with \textit{prespecified} main effects and two-factor interactions. Automatic selection of significant main effects and two-factor interactions is needed when there are many FANOVA components. Moreover, the package can only fit linear spline models that have monotonic main effects. It does not provide the capability to fit monotonic linear spline models with two-factor interactions.

6.1 Example 1: M/M/1 Queue

This example illustrates some advantages of the proposed method over the quantile GP model, rearranged quantile GP model, and a cubic Bernstein model (2) without the monotonicity constraint (fitted by check loss minimization) for a transient M/M/1 queuing system. The first arrival occurs at time 0, and the response \(Y\) is the average time the first 10 customers spend in the system. The mean time between arrivals is fixed at one and the mean service time \(z_1\) can take on any value in \([0, 1]\). We shall model quantiles of \(Y\) as a function of \(z_1\) using a design with four distinct points \(0.02, 0.34, 0.66, 0.98\). Ten independent replicates are run at each point. The degree \(N_1\) of the FANOVA Bernstein model is restricted to \([0, 1, 2, 3]\) and 95% confidence intervals for the model are constructed with the proposed bootstrap method. Credible limits for the quantile GP model are obtained from the posterior GP given in Currin et al. (1991).

In Figure 2(a)–2(d), we plot estimation results for the 0.9 quantile obtained with two sets of data. Figure 2(a) and 2(c) is constructed with the same data (similarly for Figure 2(b) and 2(d)). An accurate estimate (based on 50,000 samples) of the 0.9 quantile is plotted as a solid line with crosses in each figure. The experiment data are plotted as circles. Point estimates and 95% confidence limits (500 bootstrap samples) for the FANOVA
Bernstein model are plotted as dashed lines. The cubic Bernstein model is plotted as solid lines. Point estimates and 95% credible limits for the quantile GP are plotted as dotted lines in Figure 2(a) and 2(b), while those for the rearranged quantile GP are plotted as dash-dot lines in Figure 2(c) and 2(d). Figure 2(a) and 2(c) shows that credible intervals obtained from the quantile GP or rearranged quantile GP models may not be reasonable because they do not take into account the increase in variance of \( Y \) as \( z_1 \) increases. The quantile GP model uses a constant variance nugget term. Consequently, the credible intervals for the model have nearly the same width for all \( z_1 \). In contrast, the FANOVA Bernstein model provides reasonable confidence intervals that increase in width as \( z_1 \) increases.

Figure 2(b) and 2(d) demonstrates that the monotonicity constraints make estimation of quantiles robust against outliers. The figure illustrates the case where there is one outlier at \( z_1 = 0.66 \). This outlier seriously affects the cubic Bernstein and quantile GP models, which rise close to the outlier and give nonmonotonic predictions. The quantile GP also gives interval predictions that are too narrow because the estimated nugget variance is near zero. Rearranging the quantile GP yields no visible improvements in predictions but creates sharp bends in the prediction curves, which is physically implausible in this example. In contrast, the FANOVA Bernstein model provides accurate monotonic predictions that are only slightly affected by the outlier. The confidence intervals are wide due to the outlier, which is a reasonable behavior.

Simulation results presented in Appendix C show that the FANOVA Bernstein model provides more accurate predictions and better prediction interval coverage than the quantile GP and rearranged quantile GP models.

6.2 Example 2: Temperature of a Solid Sphere Immersed in Fluid

In this example, we model quantiles in a heat transfer problem described by Cengel (2002). The sphere is immersed in a fluid with a higher temperature at time 0. Heat transfer takes place between the sphere and the fluid via convection and inside the sphere via conduction. The temperature \( Y \) in the sphere depends on the nine inputs \( u_1, \ldots, u_9 \) shown in Table 1. The table also gives the units for each of the 10 variables (output and inputs). The relationship between \( Y \) and \( u_1, \ldots, u_9 \) is the solution of a partial differential equation. It is given explicitly by

\[
Y = u_3 + u_4 \sum_{i=1}^{\infty} \frac{4 \sin \eta_i - \eta_i \cos \eta_i}{2 \eta_i - \sin(2\eta_i)} e^{-\eta_i z_1} \sin(\eta_i u_1) \eta_i u_1, \\
1 - \eta_i \cot \eta_i = \pi_1, \quad \eta_i \in ((i - 1)\pi, i\pi), \\
\pi_1 = \frac{u_5 u_7 / u_6, \pi_2 = u_6 u_2 / (u_3 u_0 u_5^2)},
\]

where \( \pi_1 \) is the Biot number and \( \pi_2 \) is the Fourier number. The \( \eta_i \)'s are the solutions of the equation \( 1 - \eta_i \cot \eta_i = \pi_1 \), where \( \eta_i \in ((i - 1)\pi, i\pi) \). Our computations approximate the infinite series in (22) with the sum of the first four terms only, which is quite accurate. The above model is useful for manufacturing process design (see chap. 8 of Jaularia and Torrance 2003).

From physical considerations, we know that \( Y \) is increasing in \( u_1, \ldots, u_5 \) whenever \( u_4 < 0 \), that is, whenever the temperature of the medium is higher than the initial temperature of the sphere. The temperature at a point nearer the surface of the sphere is higher than the temperature at a point farther from the surface. Thus, \( Y \) is increasing in \( u_1 \). The temperature \( Y \) should clearly increase with time \( u_2 \) since heat flows into the sphere. If the temperature of the medium \( u_3 \) gets higher, the temperature at any point in the sphere should get higher. Moreover, if the initial temperature of the sphere is closer to the temperature of the medium \( (u_4) \) increases), the temperature at any point in the sphere should be higher. Finally, as the convective heat transfer coefficient \( u_5 \) increases, heat transfer into the sphere increases. Thus, \( Y \) is increasing in \( u_5 \). Note that the arguments above hold for each \( u_i \) conditional on \( u_j, j \neq i \).

We standardize \( u_1, \ldots, u_5 \) and work with the standardized input \( z = (z_1, \ldots, z_9) \in [0, 1]^5 \), where \( z_3 = (u_3 - 250)/(270 - 250) \) and similarly for the other \( z_i \)'s. The values of \( u_6, \ldots, u_9 \) are randomly sampled from their ranges in each computer model run. Thus, the value of \( Y \) returned by the computer model is stochastic. We shall model the 0.1, 0.5, and 0.9 quantiles of \( Y \). The difference between the 0.9 and 0.1 quantiles is useful for quantifying the spread of the distribution of \( Y \) and the quantity \( (f_{0.9} - f_{0.5}) - (f_{0.5} - f_{0.1}) \) gives a measure of the skewness. The relationship between \( f \) and \( z_2 \) (time) has an asymptote, as shown in Figure D1 in Appendix D. This is due to the temperature achieving its steady state value. This feature is manifested in

| Table 1. Description, units, minimum, and maximum values of variables in Example 2. The description column gives a physical description of each input and the output; the units column gives the physical unit for each variable; the minimum and maximum columns give the minimum and maximum values of each input for the experimental design |
|---|---|---|---|
| \( Y \) | Temperature | K | – | – |
| \( u_1 \) | Distance from center/ radius of sphere | – | 0 | 1 |
| \( u_2 \) | Time | s | 0 | 2000 |
| \( u_3 \) | Temperature of medium (fluid) | K | 250 | 270 |
| \( u_4 \) | Initial sphere temperature \( - \) temperature of medium | K | –100 | –30 |
| \( u_5 \) | Convective heat transfer coefficient | kgs\(^{-1}\)K\(^{-1}\) | 180 | 210 |
| \( u_6 \) | Thermal conductivity | kgs\(^{-1}\)K\(^{-1}\) | 30 | 100 |
| \( u_7 \) | Radius of sphere | m | 0.05 | 0.15 |
| \( u_8 \) | Specific heat | m\(^3\)K\(^{-2}\) | 300 | 500 |
| \( u_9 \) | Density | kgm\(^{-3}\) | 7000 | 9000 |

TECHNOMETRICS, MAY 2016, VOL. 58, NO. 2
For the multivariate Bernstein model, we set many physical systems. The spread of the distribution of $Y$ also changes with $z_2$ (see Figure D1).

We perform 50 simulations to compare the performance of the proposed approach to the multivariate Bernstein, quantile GP, and rearranged quantile GP models. The design employed is a 40-run cosine MLHD replicated 10 times. Each cosine MLHD is generated by first generating an MLHD $U$ and transforming the MLHD to obtain a cosine MLHD $D = [\cos(U \pi) + 1]/2$. Note that $U$ is obtained by generating a 2500 Latin hypercube designs and selecting the one that performs the best with respect to the maximin criterion. Thus, $D$ is different for each simulation.

In each simulation, the design $D$ is generated and the output from the simulator is obtained. Then, quantile regression is performed using the multivariate Bernstein model (4) with the monotonicity constraints (5), the FANova Bernstein model, the quantile GP model, and the rearranged quantile GP model. The value of $N_{\text{max}}$ for the FANOVA Bernstein model is fixed at seven.

For the multivariate Bernstein model, we set $N = (1, 1, 1, 1, 1)$. Note that this model requires $2^5 = 32$ distinct design points so that its coefficients are identifiable. If each component of $N$ is increased to two, we would need $3^5 = 243$ distinct design points. This demonstrates a disadvantage of the multivariate Bernstein model. Two quantile estimation performance measures are computed in each simulation run: the square error and absolute error averaged over a set of validation data points, which are the mean squared error (MSE) and mean absolute error (MAE), respectively. The validation data consist of 1000 replicates of a 500-run MLHD design. The square error at each of the 500 data points is $f_\tau - f_\hat{\tau})^2$, where $f_\tau$ is the sample quantile of the 1000 replicates computed using Matlab’s `prctile` function, and $f_\hat{\tau}$ is the quantile estimate obtained using one of the four alternative models above (similarly for the absolute error). In Appendix D, we plot the point and interval predictions given by the FANOVA Bernstein and rearranged quantile GP model for some simulated data and discuss model selection results.

Table 2 presents the simulation mean and standard deviation of the MSE and MAE for the FANOVA Bernstein model and the mean difference in MSE/MAE between a competing model and the FANOVA Bernstein model. We see that the FANOVA Bernstein model is substantially better than the multivariate Bernstein model in all cases. Table 2 shows that in the simulations, the FANOVA Bernstein model performs better than the quantile GP and rearranged quantile GP models for $\tau = 0.5$ and $\tau = 0.9$ but performs worse when $\tau = 0.1$. Since the same data are used to fit the four models in each simulation, the fraction of simulations in which the difference in MSE/MAE is positive can be used to test whether the FANOVA Bernstein model is significantly better. These fractions are given in Table 2. Since

| $\tau = 0.9$ | Multivariate Bernstein − FANOVA Bernstein | Quantile GP − FANOVA Bernstein | Rearranged Quantile GP − FANOVA Bernstein |
|--------------|------------------------------------------|---------------------------------|---------------------------------------------|
| **FANOVA Bernstein** | **Mean** | **Mean** | **Mean** | **Mean** | **Mean** | **Mean** | **Mean** |
| MSE | 1.883 | 115.629 | 5.878 | 4.224 |
| Std | 1.132 | 1.000 | 0.980 | 1.000 |
| MAE | 0.859 | 6.272 | 0.953 | 0.757 |
| Std | 0.228 | 1.000 | 1.000 | 1.000 |

| $\tau = 0.1$ | Multivariate Bernstein − FANOVA Bernstein | Quantile GP − FANOVA Bernstein | Rearranged Quantile GP − FANOVA Bernstein |
|--------------|------------------------------------------|---------------------------------|---------------------------------------------|
| **FANOVA Bernstein** | **Mean** | **Mean** | **Mean** | **Mean** | **Mean** | **Mean** | **Mean** |
| MSE | 7.022 | 53.604 | −1.179 | −1.630 |
| Std | 2.175 | 1.000 | 0.220 | 0.160 |
| MAE | 1.962 | 4.373 | −0.141 | −0.217 |
| Std | 0.305 | 1.000 | 0.260 | 0.180 |

| $\tau = 0.5$ | Multivariate Bernstein − FANOVA Bernstein | Quantile GP − FANOVA Bernstein | Rearranged Quantile GP − FANOVA Bernstein |
|--------------|------------------------------------------|---------------------------------|---------------------------------------------|
| **FANOVA Bernstein** | **Mean** | **Mean** | **Mean** | **Mean** | **Mean** | **Mean** | **Mean** |
| MSE | 3.817 | 91.399 | 3.972 | 2.427 |
| Std | 1.872 | 1.000 | 0.880 | 0.800 |
| MAE | 1.396 | 6.365 | 0.616 | 0.425 |
| Std | 0.305 | 1.000 | 0.900 | 0.840 |
Table 3. Mean and standard deviation of coverage of 95% confidence/credible intervals given by FANOVA Bernstein, quantile GP, and rearranged quantile GP models.

| | FANOVA Bernstein | Quantile GP | Rearranged quantile GP |
|---|---|---|---|
| τ = 0.9 | Mean | 0.91 | 0.82 | 0.85 |
| | Std | 0.09 | 0.10 | 0.07 |
| τ = 0.1 | Mean | 0.91 | 0.83 | 0.85 |
| | Std | 0.06 | 0.10 | 0.09 |
| τ = 0.5 | Mean | 0.87 | 0.80 | 0.84 |
| | Std | 0.07 | 0.11 | 0.08 |

\( P(B / 50 \geq 0.68) = 0.0077 \), where \( B \) is a binomial random variable with 50 trials and success probability 0.5. We say that the FANOVA Bernstein model is significantly better if the observed fraction is larger than or equal to 0.68. We see that the FANOVA Bernstein model is significantly better (with respect to both MSE and MAE) than the quantile and rearranged quantile GP models for estimating the 0.5 and 0.9 quantile but significantly worse for the 0.1 quantile.

We assess the coverage of 95% confidence/credible intervals of the FANOVA Bernstein, quantile GP, and rearranged quantile GP models. The coverage of the intervals for a model is the fraction that contains the sample quantiles of the validation data. The 95% confidence intervals for the FANOVA Bernstein model are constructed with 100 bootstrap samples. Table 3 gives the mean and standard deviation of the coverages. We see that the FANOVA Bernstein model gives coverage closer to the nominal coverage than the quantile GP or rearranged quantile GP models.

7. CONCLUSIONS

This article proposes a class of FANOVA Bernstein models for quantile estimation in stochastic simulation. The models are obtained by modeling main effects and two-factor interactions in a FANOVA decomposition with Bernstein bases. The model coefficients are estimated with a linear program as in quantile regression. Monotonicity with respect to the inputs is imposed by linear constraints on the coefficients. Two-factor interactions and the degree of the polynomial bases for each input are selected by an algorithm that uses the BIC as model selection criterion. A smoothed bootstrap method is employed to perform statistical inference. Three examples that involve estimating the quantiles of real stochastic simulation models illustrate that quantile point estimates obtained with the proposed model are robust to outliers in the response and can be more accurate than the quantile GP or rearranged quantile GP models.

A few areas require further research. First, a faster procedure for constructing confidence intervals is desirable. The bootstrap procedure for obtaining confidence intervals is time-consuming due to the need to repeat the model selection step. If the model selection step is not repeated, the confidence intervals that are obtained can suffer from undercoverage. Second, a rigorous method is needed to choose the design points and the number of replicates.

SUPPLEMENTARY MATERIALS

Appendices.pdf: This file contains Appendices A–D. Appendix A contains proofs of mathematical results. Appendix B gives a numerical example in which quantiles of a queueing simulation model with four inputs are modeled. Appendix C gives simulation results for Example 1. Appendix D gives illustrations and results for Example 2.

ACKNOWLEDGMENTS

The author thanks the associate editor and two referees for helpful comments. This research was supported by City University of Hong Kong Start-Up Grant 7200364 and Early Career Scheme (ECS) project 21201414 funded by the Research Grants Council of Hong Kong.

[Received August 2014. Revised February 2015.]

REFERENCES

Ankenman, B., Nelson, B. L., and Staum, J. (2010), “Stochastic Kriging for Simulation Metamodeling,” Operations Research, 58, 371–382. [180]
Barenblatt, G. I. (1996), Scaling, Self-Similarity, and Intermediate Asymptotics: Dimensional Analysis and Intermediate Asymptotics (Vol. 14), Cambridge: Cambridge University Press. [183]
Cengel, Y. A. (2002), Heat Transfer: A Practical Approach, New York: McGraw-Hill. [180,187]
Chang, I. S., Chien, L. C., Hsiung, C. A., Wen, C. C., and Wu, Y. J. (2007), “Shape Restricted Regression with Random Bernstein Polynomials,” IMS Lecture Notes-Monograph Series, 54, 187–202. [181]
Chernozhukov, V., Fernandez-Val, I., and Galichon, A. (2009), “Improving Point and Interval Estimators of Monotone Functions by Rearrangement,” Biometrika, 96, 559–575. [181,185]
Chipman, H. (1996), “Bayesian Variable Selection With Related Predictors,” The Canadian Journal of Statistics, 24, 17–36. [181,185]
Curtin, C., Mitchell, T., Morris, M., and Ylivieska, D. (1991), “Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments,” Journal of the American Statistical Association, 86, 953–963. [180,185,186]
Dette, H., and Pepelyshev, A. (2010), “Generalized Latin Hypercube Design for Computer Experiments,” Technometrics, 52, 421–429. [185]
Efron, B., and Stein, C. (1981), “The Jackknife Estimate of Variance,” The Annals of Statistics, 9, 586–596. [182]
Gal, S. and Anastassiou, G. A. (2010), Shape-Preserving Approximation by Real and Complex Polynomials, New York: Springer. [181]
Gluhovsky, I. (2006), “Smooth Isotonic Additive Interaction Models with Application to Computer System Architecture Design,” Technometrics, 48, 176–192. [181]
Gross, D., Shortle, J. F., Thompson, J. M., and Harris, C. M. (2013), Fundamentals of Queueing Theory (4th ed.), New York: Wiley. [180]
Hastie, T. J., and Tibshirani, R. J. (1990), Generalized Additive Models (Vol. 43), Boca Raton, FL: CRC Press. [183]
He, X., and Ng, P. (1999), “COBS: Qualitatively Constrained Smoothing via Linear Programming,” Computational Statistics, 14, 315–338. [181]
Jaluria, Y., and Torrance, K. E. (2003), “Shape Restricted Regression with Random Bernstein Polynomials,” IMS Lecture Notes-Monograph Series, 54, 187–202. [181]
Kim, M. (2006), “Quantile Regression with Shape-Constrained Varying Coefficients,” Sankhya, 68, 369–391. [181]
Kleijnen, J. P. (2009), “Kriging Metamodeling in Simulation: A Review,” European Journal of Operational Research, 192, 707–716. [180]
Koenker, R. (2005), Quantile Regression (Vol. 38), Cambridge: Cambridge University Press. [181,184]
Koenker, R., and Bassett Jr, G. (1978), “Regression Quantiles,” Econometrica: Journal of the Econometric Society, 66, 3–50. [180,184]
Koenker, R., and Mizera, I. (2004), “Penalized Triograms: Total Variation Regularization for Bivariate Smoothing,” Journal of the Royal Statistical Society, Series B, 66, 145–163. [186]
Koenker, R., Ng, P., and Portnoy, S. (1994), “Quantile Smoothing Splines,” Biometrika, 81, 673–680. [186]
Law, A. (2014), Simulation Modeling and Analysis (5th ed.), New York: McGraw-Hill. [180]
Lian, H. (2012), “A Note on the Consistency of Schwarz's Criterion in Linear Quantile Regression with the SCAD Penalty,” Statistics & Probability Letters, 82, 1224–1228. [181,184]
Mammen, E., Marron, J. S., Turlach, B. A., and Wand, M. P. (2001), “A General Projection Framework for Constrained Smoothing,” Statistical Science, 16, 232–248. [181]

McKay Curtis, S., and Ghosh, S. K. (2011), “A Variable Selection Approach to Monotonic Regression with Bernstein Polynomials,” Journal of Applied Statistics, 38, 961–976. [181]

Plumlee, M., and Tuò, R. (2014). “Building Accurate Emulators for Stochastic Simulations via Quantile Kriging,” Technometrics, 56, 466–473. [181,185]

Ramsay, J. O. (1988), “Monotone Regression Splines in Action,” Statistical Science, 3, 425–441. [181]

Ramsay, J. O., and Silverman, B. W. (2005), Functional Data Analysis (2nd ed.), New York: Springer. [182]

Reich, B. J., Fuentes, M., and Dunson, D. B. (2011), “Bayesian Spatial Quantile Regression,” Journal of the American Statistical Association, 106, 6–20. [181]

Rostek, M. (2010). “Quantile Maximization in Decision Theory,” The Review of Economic Studies, 77, 339–371. [180]

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), “Design and Analysis of Computer Experiments,” Statistical Science, 4, 409–423. [180]

Saltelli, A., Chan, K., and Scott, M. (eds.) (2000), Sensitivity Analysis, New York: Wiley. [182]

Silverman, B. W., and Young, G. A. (1987), “The Bootstrap: To Smooth or not to Smooth?,” Biometrika, 74, 469–479. [185]

Tan, M. H. Y. (2015), “Stochastic Polynomial Interpolation for Uncertainty Quantification with Computer Experiments,” Technometrics, 57, 457–467. [185]

Wang, J., and Ghosh, S. K. (2011), “Shape Restricted Nonparametric Regression Based on Multivariate Bernstein Polynomials,” Technical Report no. 2640, Department of Statistics, NC State University. [181]

——— (2012), “Shape Restricted Nonparametric Regression With Bernstein Polynomials,” Computational Statistics & Data Analysis, 56, 2729–2741. [181]

Wu, C. F. J., and Hamada, M. S. (2009), Experiments: Planning, Analysis, and Optimization (2nd ed.), New York: Wiley. [182,183]

Wu, Y., and Liu, Y. (2009), “Variable Selection in Quantile Regression,” Statistica Sinica, 19, 801–817. [185]

Xiu, D. (2010), Numerical Methods for Stochastic Computations: A Spectral Method Approach, New Jersey: Princeton University Press. [180]