Nonparametric curve estimation under monotonicity constraint

Rabi Bhattacharya[1,3] and Lizhen Lin[2]
[1] Department of Mathematics, University of Arizona, Tucson AZ 85721 , USA
[2] Department of Statistical Science, Duke University, Durham NC 27708, USA
[3] Corresponding author: Rabi Bhattacharya, email: rabi@math.arizona.edu

Abstract

A finite sample comparison is carried out for three recent nonparametric methodologies in estimating the monotone regression function F and its inverse F^{-1}. The methods are (1) the inverse kernel method DNP (Dette et al. (2005), Dette and Scheder (2010)), (2) the monotone spline (Kong and Eubank (2006)) and (3) the data adaptive method NAM (Bhattacharya and Lin (2010), (2011)), with roots in isotonic regression (Ayer et al. (1955), Bhattacharya and Kong (2007)). Comparative studies are carried out mainly in terms of length of the confidence interval in this article. Let there be m distinct values of the independent variable x among N observations y. The results show that if m is relatively small compared to N then generally the NAM performs best, while the DNP outperforms the other methods when m is $O(N)$ unless there is a substantial clustering of the values of the independent variable x.

Keywords: Finite sample comparative study, isotonic regression, monotone splines.

1 Introduction

Given a regression model $y = F(x) + \epsilon$ with $F(x)$ as the regression function and ϵ some independent error, it is often of interest in estimating $F(x)$ under some shape constraint such as the monotonicity constraint. For example, it is very reasonable to assume the biometric age-height charts to be monotonic in age over an appropriate range; econometric demand functions to be monotonic in price. Among other important examples are the estimation of the dose-response curves in bioassay and environmental studies which will be the main applications of the methodologies compared in our article. By constraining an estimate to be monotone one can avoid obtaining implausible estimates and produce estimates that are consistent with the prior knowledge and the biological background. There is a large literature in estimating the monotone does-response curve and its inverse using parametric models such as logistic models and probit models. However, it is in general hard to justify such functional forms and parametric modeling often leads to inconsistent estimates due to model misspecification. Thus a few nonparametric methodologies have been recently proposed. Three of the main nonparametric methods are (1) the inverse kernel method DNP (Dette et al. (2005), Dette and Scheder (2010)), (2) the monotone spline (Kong and Eubank (2006)) and (3) the data adaptive method NAM (Bhattacharya and Lin (2010), (2011)), with roots in isotonic regression (Ayer et al. (1955), Bhattacharya and Kong (2007)). These methods can be in general classified into three classes, methods that are kernel based, methods that exploit spline basis and methods that are based on isotonic regression estimates. Most of these methods bear optimal asymptotic properties. This article summarizes some results of an extensive finite sample comparison among these methods carried out in Bhattacharya and Lin (2013).

The article is organized as follows. Section 2 gives a brief description of the three nonparametric methods that are compared. Section 3 summarizes some of the results of the finite sample comparison in terms of simulated data.

2 Nonparametric methods

Consider the estimation of a monotone increasing regression function F on an interval $[a, b]$, $F' > 0$, based on observations (x_j, y_j), $j = 1, \cdots, N$, satisfying

$$y_j = F(x_j) + \epsilon_j \ (j = 1, \cdots, N), \tag{1}$$

where $a = x_1 \leq x_2 \leq \cdots \leq x_N = b$ are nonstochastic and ϵ_j $(j = 1, \cdots, N)$ are independent mean zero random variables; the distribution of ϵ_j may depend on x_j. Suppose there are m distinct values of x, say $a = z_1 < \cdots < z_m = b$, and n_i observations y_j for a given $x = z_i$, $n_1 + \cdots + n_m = N$, $m \to \infty$. In particular, one may allow $m = N$. Assume F has a continuous second derivative. In a dose-response study, $z_1, \ldots, z_m$ corresponding to m distinct dosages.

2.1 An adaptive nonparametric NAM

Given a set of weights w_i $(i = 1, \cdots, m)$, in the isotonic regression problem (1) the minimizer of

$$\sum_{i=1}^{m} (y_i - F(z_i))^2 w_i \tag{2}$$

over the class of all monotone nondecreasing F is given by (Ayer et al. (1955))

$$\widetilde{F}(z_i) = \max_{s \leq i} \min_{t \geq i} \frac{\sum_{s \leq q \leq t} y_q w_q}{\sum_{s \leq q \leq t} w_q}. \tag{3}$$

The estimate of the whole curve F is obtained by linear interpolation in $[\widetilde{F}(z_i)), \widetilde{F}(z_{i+1})]$, $i = 1, \cdots, m-1$. This also allows one to obtain an estimate $\widetilde{F}^{-1}$ of the inverse curve F^{-1}. Consider also the usual estimate $\widehat{F}(z_i)$ of $F(z_i)$ as the mean of those observations y_j for a given x value z_i . That is,

$$\widehat{F}(z_i) = S_i / n_i \tag{4}$$

where, S_i is the sum of those y_j with $x_j = z_i$. When the number of dosages m is large, one can divide the set of m sets $(z_1, \cdots, z_m,)$ into r adjacent nearly disjoint subgroups of approximately equal size $s(n)$ each, where r and $s(n)$ will be specified later. They satisfy the approximate equality

$$m \simeq rs(n). \tag{5}$$

For example, Group 1 comprises the z values $(z_1, z_{r+1}, z_{2r+1}, \cdots, z_{(s(n)-1)r+1)}, z_m)$, Group 2 is $(z_1, z_2, z_{r+2}, z_{2r+2}, \cdots, z_{(s(n)-1)r+2)}, z_m)$, $\cdots$, Group r is $(z_1, z_r, z_{2r}, \cdots, z_{(s(n)r)}, z_m)$. Now construct the linearly interpolated PAV estimate $\widetilde{F}_t$ of F as above, but using only the y_j's belonging to the t-th Group of z levels $(t = 1, \cdots, r)$. Then define the *NAM estimates* of F and F^{-1} as $\widetilde{F} = (1/r) \sum_{1 \leq t \leq r} \widetilde{F}_t$ and $\zeta = (1/r) \sum_{q \leq t \leq r} (\widetilde{F}_t)^{-1}$, respectively.

2.2 Kernel Based DNP Method

An important kernel-based method in estimating the effective dosage curve is the *DNP method* (following the terminology in Dette and Scheder (2010)). Let the response to x_i be y_i (0 or 1). The local linear estimator is obtained first by finding the solution to the following minimization problem: for a small $h > 0$, find the minimizer of

$$\min_{\beta_1, \beta_2} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right) (y_i - \beta_1 - \beta_2 (x_i - x))^2, \tag{6}$$

where $K(x)$ is a symmetric density on the real line $\mathbb{R}$ with a finite second moment, and h is the bandwidth. The estimator $\hat{\beta}_1(x)$ of β_1 is the estimator $\widehat{\widehat{F}}(x)$ of $F(x)$. The p-th quantile $ED_p = F^{-1}(p)$ is then estimated as

$$\widehat{ED}_p = \int_0^1 \int_{-\infty}^{p} \frac{1}{h_d} K_d \left(\frac{\widehat{\widehat{F}}(x) - u}{h_d} \right) du\, dx, \tag{7}$$

where h_d is small. Here K_d is a symmetric kernel with the same properties as K (e.g, $K_d = K$). But h and h_d are not of the same order, as shown below. Note that as $h_d \downarrow 0$, $\widehat{ED}_p$ converges to ED_p. To understand this, observe that $\frac{1}{h_d} K_d \left(\frac{\widehat{\widehat{F}}(x) - u}{h_d} \right) du$ converges to the Dirac measure $\delta_{\widehat{F}(x)}(du)$ as $h_d \downarrow 0$, so that the inner integral converges to the indicator function $1_{[\widehat{\widehat{F}}(x) \le p]}(x)$. The outer integral of this limit is the Lebesgue measure of the set $\{x : \widehat{\widehat{F}}(x) \le p\}$ which equals the length of this interval. This method of monotonization of a function $\widehat{\widehat{F}}$ is called *monotone or measure-preserving rearrangement* in Hardy et al. (1952). With an optimal choice of the bandwidth the estimate $p \to \widehat{ED}_p$ of F^{-1} attains asymptotically optimal error rates.

2.3 Monotone B-Spline Smoothing

Given the regression model (1), the general smoothing spline problem is to find the estimate $\widehat{F}_h$ of the function F that minimizes the objective function (over the class of twice differentiable functions)

$$\mathcal{J} = \frac{1}{m} \sum_{j=1}^{m} (y_j - F(x_j))^2 w_j + h \int_{x_1}^{x_m} F''(x)^2 dx, \tag{8}$$

where w_j are positive weights and h is a smoothing parameter which controls the trade off between the smoothness of the curve and the fidelity to the data.

The existence and characterizations of the solution to (8) without any shape constraint on the regression function F are derived in Wahba (1990) (Also see Eubank (1999)). When F is *assumed to be monotone,* an approach by Kong and Eubank (2007) and Kelly and Rice (1990) with the use of the so-called B-spline basis $B_{j,4}$, represents a monotone F as a linear combination of the basis functions $B_{j,4}$, with coefficients β_j increasing with j. The existence of such F as a solution minimizing (8) can be easily shown.

For constructing confidence intervals using monotone spline estimates in quantal bioassay, Kong and Eubank (2007) proposed a form of parametric Bayesian inference for deriving confidence intervals. Our article constructs confidence intervals using the nonparametric bootstrap, which may be shown to be valid and which makes the procedure *fully nonparametric*. The optimal estimate of the smoothing parameter h is given by the GCV (generalized cross validation) algorithm (See Eubank (1999)).

3 Finite sample comparisons

In this subsection, 95% confidence intervals are constructed for the NAM, DNP, Spline estimates and MLE of the effective dosages ζ_p with 1000 samples of data simulated from some important parametric models. For each sample, NAM, DNP, Spline estimates and MLE of the effective dosage curve are obtained for 11 equidistant response levels in [0.05,0.85]. For each of the response level p, the lower confidence limits are given by the 2.5 percent quantile of the 1000 estimates and the upper confidence limits are given by the 97.5 percent quantile value of the 1000 estimates for each method.

The data are simulated from four parametric models which are Logistic model, Probit model, Beta model and Weibull model for the case $m = 5$, $n = 5, 10, 25$ and $m = 10$, $n = 5, 10, 25$. The comparisons are mainly carried out in terms of the length of the confidence interval for each method. The first three rows of the following tables record the lengths of the confidence intervals for the nonparametric estimates by NAM, DNP and Spline; the fourth

row records the difference $D1 =$ the length of the CI for DNP $-$ the length of CI for NAM; and the last row records the difference $D2 =$the length of the CI for Spline $-$ the length of CI for DNP. For lack of space only the results for the Weibull model are displayed here, more extensive results may be found in your Bhattacharya and Lin (2013). Some data examples are also given in Bhattacharya and Lin (2013).

As one can see from the results given in the following tables, for the most part, the NAM method yields the narrowest confidence intervals when $n = 10, 25$ for both of the case when $m = 5$ and $m = 10$ for for all four models. When $n = 5$, it seems that the DNP method works better for most of the cases. The NAM and DNP methods in general perform better than the monotone Spline method.

At the end of this subsection, some plots of the true confidence intervals for different estimates are given. The blue line in the middle is the true effective dosage curve, the red line with circles represents the confidence interval for the NAM estimate, the green line represents the confidence intervals for the MLE and the black line represents the confidence interval for the DNP estimate. For the graphs where the NAM confidence limits and the Spline (SP) confidence limits are compared, the blue lines represent the confidence intervals for the Spline estimates.

Figure 1: [Weibull]

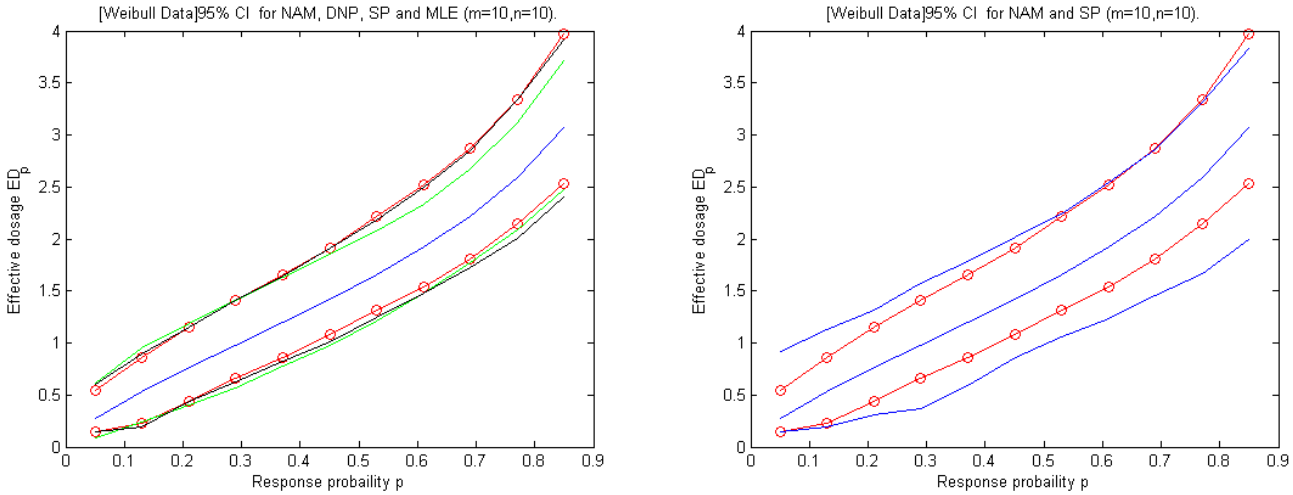

Figure 2: [Weibull]

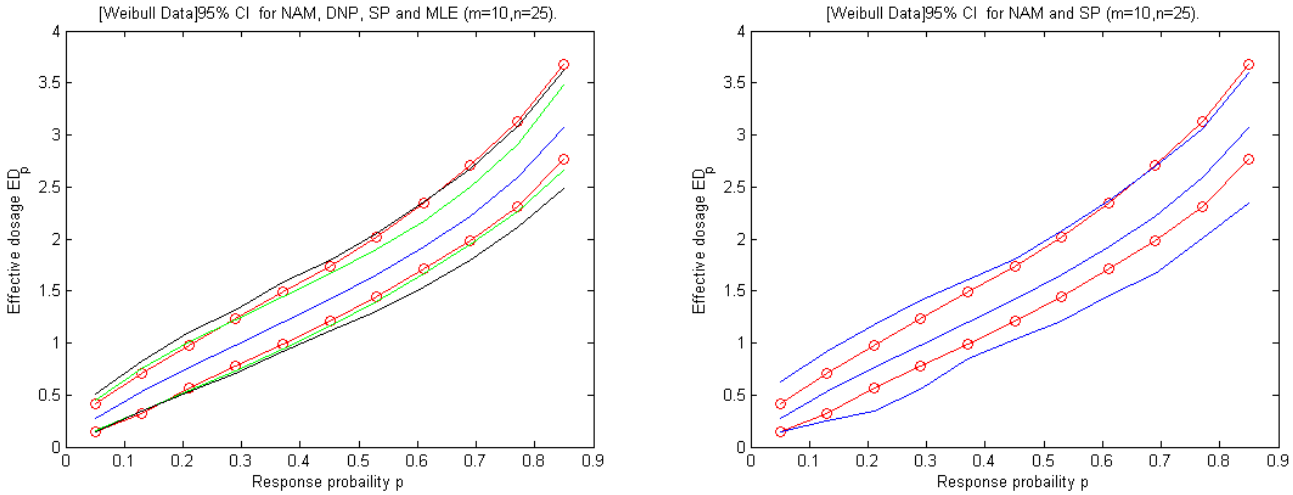

Table 1: [Weibull (m=5,n=5)] Lengths of Confidence Intervals for DNP, NAM(r=2) and Spline and Their Differences

NAM:	0.1787	0.2262	0.2414	0.2431	0.2598	0.2852	0.2871	0.2804	0.2926	0.3061	0.3232
DNP:	0.1375	0.2155	0.2508	0.2422	0.2522	0.2621	0.2702	0.2768	0.2899	0.3046	0.3127
Spline	0.2642	0.3008	0.3093	0.3624	0.3863	0.3905	0.3552	0.3202	0.3643	0.6655	0.7029
D1:	-0.0412	-0.0107	0.0093	-0.0009	-0.0076	-0.0230	-0.0168	-0.0036	-0.0028	-0.0015	-0.0105
D2:	0.1267	0.0853	0.0585	0.1202	0.1341	0.1284	0.0849	0.0434	0.0744	0.0891	0.0925

Table 2: [Weibull (m=5,n=10)] Lengths of Confidence Intervals for DNP, NAM(r=2) and Spline and Their Differences

NAM:	0.1123	0.1711	0.1883	0.1838	0.1848	0.2028	0.2022	0.1991	0.2246	0.2155	0.2132
DNP:	0.1238	0.1763	0.1854	0.2015	0.2019	0.2102	0.2163	0.2218	0.2233	0.2380	0.2506
Spline:	0.1238	0.1763	0.1854	0.2015	0.2019	0.2102	0.2163	0.2218	0.2233	0.2380	0.2506
D1:	0.0116	0.0052	-0.0029	0.0178	0.0171	0.0074	0.0141	0.0227	-0.0013	0.0225	0.0374
D2:	0.0545	0.0656	0.0853	0.0758	0.0744	0.0580	0.0375	0.0291	0.0336	0.0637	0.0720

Table 3: [Weibull (m=5,n=25)] Lengths of Confidence Intervals for DNP, NAM(r=2) and Spline and Their Differences

NAM:	0.0814	0.1013	0.1120	0.1274	0.1257	0.1246	0.1333	0.1392	0.1414	0.1393	0.1530
DNP:	0.0967	0.1360	0.1449	0.1682	0.1575	0.1471	0.1685	0.1890	0.1714	0.1754	0.1914
Spline	0.1264	0.1646	0.2063	0.2143	0.1739	0.1593	0.1779	0.1853	0.1824	0.1837	0.2117
D1:	0.0153	0.0347	0.0329	0.0408	0.0317	0.0226	0.0352	0.0498	0.0300	0.0361	0.0384
D2:	0.0297	0.0286	0.0613	0.0461	0.0164	0.0121	0.0094	-0.0037	0.0110	0.0083	0.0203

Table 4: [Weibull (m=10,n=5)] Lengths of Confidence Intervals for DNP, NAM(r=3) and Spline and Their Differences

NAM:	0.6268	0.8851	0.9274	0.9700	1.0744	1.1611	1.2431	1.4632	1.5249	1.6640	1.9524
DNP:	0.5311	0.8156	0.9539	0.9713	1.0605	1.1368	1.1874	1.2617	1.3447	1.5111	1.8156
Spline	0.8892	1.2904	1.3509	1.4651	1.5616	1.5528	1.7029	1.7617	1.9285	2.0212	2.6157
D1:	-0.0957	-0.0695	0.0264	0.0013	-0.0139	-0.0244	-0.0556	-0.2015	-0.1802	-0.1528	-0.1368
D2:	0.3581	0.4748	0.3970	0.4937	0.5011	0.4160	0.5155	0.5000	0.5838	0.5101	0.8001

Table 5: [Weibull (m=10,n=10)] Lengths of Confidence Intervals for DNP, NAM(r=3) and Spline and Their Differences

NAM:	0.3905	0.6339	0.7121	0.7498	0.7925	0.8316	0.8952	0.9822	1.0641	1.2010	1.4386
DNP:	0.4510	0.6948	0.7144	0.7759	0.8158	0.8971	0.9377	1.0168	1.1194	1.3422	1.5087
Spline:	0.7749	0.9332	1.0090	1.2055	1.1872	1.1504	1.1786	1.3059	1.3999	1.6432	1.8403
D1:	0.0606	0.0609	0.0023	0.0261	0.0233	0.0655	0.0425	0.0346	0.0553	0.1411	0.0701
D2:	0.3239	0.2384	0.2946	0.4296	0.3714	0.2533	0.2409	0.2891	0.2805	0.3010	0.3317

Table 6: [Weibull (m=10,n=25)] Lengths of Confidence Intervals for DNP, NAM(r=3) and Spline and Their Differences

NAM:	0.2672	0.3900	0.4080	0.4489	0.4933	0.5273	0.5798	0.6388	0.7249	0.8204	0.9159
DNP:	0.3644	0.4817	0.5750	0.6135	0.6636	0.6875	0.7515	0.8263	0.8802	0.9672	1.1384
Spline:	0.4724	0.6704	0.8248	0.8549	0.7569	0.7691	0.8664	0.9039	1.0354	1.0535	1.2535
D1:	0.0973	0.0917	0.1670	0.1646	0.1703	0.1602	0.1716	0.1874	0.1552	0.1468	0.2225
D2:	0.1080	0.1887	0.2499	0.2415	0.0932	0.0815	0.1150	0.0776	0.1552	0.0862	0.1150

References

[1] AYER, M., BRUNK, H.D., EWING, G.M., REID, W.T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann.Math.Statist.* **26** 641-647.

[2] BHATTACHARYA, R. and KONG, M. (2007). Consistency and asymptotic normality of the estimated effective dose in bioassay. *J. Statist. Plan. Inf.* **137** 643-658.

[3] BHATTACHARYA, R. and LIN, L. (2010). An adaptive nonparametric method in benchmark analysis for bioassay and environmental studies. *Statist. Probab. Letters*. **80**, 1947-1953.

[4] BHATTACHARYA, R. and LIN, L. (2011). Nonparametric benchmark analysis in risk assessment: a comparative study by simulation and data analysis. *Sankhya, The Indian Journal of Statistics Ser.B* **73**, Issue 1(2011), 144-163.

[5] BHATTACHARYA, R. and LIN, L. (2013). Recent progress in the nonparametric estimation of monotone curves -with applications to bioassay and environmental risk assessment (with Bhattacharya, R.) . *Comput. Statist. Data Anal.*, **63**, pp. 63–80.

[6] DETTE, H. and SCHEDER, R. (2010). A finite sample comparison of nonparametric estimates of the effective dose in quantal bioassay. *J. Statist. Comput. Simulation.* **80 (5)** 527–544.

[7] DETTE,H., NEUMEYER, N. and PLIZ, K.F. (2005). A note on nonparametric estimation of the effective dose in quantal bioassay. *J.Amer.Statist.Assoc.* **100** 503-510.

[8] EFRON, B. and TIBSHIRANI,R.J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, London.

[9] EUBANK, R.L. (1999). *Nonparametric Regression and Spline Smoothing.* 2nd Edition. Marcel Dekker, New York.

[10] HARDY, G. H.; LITTLEWOOD, J. E.; AND PÓLYA, G. *Inequalities.* Cambridge, England: Cambridge University Press, 1952.

[11] KELLY, C, and RICE, J. (1990).Monotone smoothing with application to dose response curves and the assessment of synergism. BIOMETRICS **46**, 1071–1085.

[12] Kong, M. and Eubank, R.L. (2006). Monotone smoothing with application to dose-response curve. *Comm. Statist. Simulation Comput*. **35**, no. 4, 991-1004,

[13] MÜLLER,H.G. and SCHMITT,T. (1988). Kernel and probit estimation in quantal bioassay. *J.Amer.Statist.Assoc* **83**(403) pp, 750-759.

[14] Utreras, F.I. (1985). Smoothing noisy data under monotonicity constraints: existence, characterization and convergence rates. *Numer. Math.* **74** , 611–625.

[15] WAHBA, G. (1990). *Spline Models for Observational Data.* CBMS-NSF Series, **59**. SIAM, Philadelphia, PA.