國立交通大學

統計學研究所

碩士論文

單調無母數迴歸在廣義線性模型上之應用

Nonparametric Monotone Regression for Generalized Linear Models

研究生：文誠智

指導教授：洪志真 博士

中華民國九十六年六月

單調無母數迴歸在廣義線性模型上之應用

Nonparametric Monotone Regression for Generalized Linear Models

研究生：文誠智　　Student：Cheng-Chih Wen

指導教授：洪志真　　Advisor：Dr. Jyh-Jen Horng Shiau

國立交通大學

統計學研究所

碩士論文

A Thesis
Submitted to Institute of Statistics
College of Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics
June 2007
Hsinchu, Taiwan, Republic of China

中華民國九十六年六月

單調無母數迴歸在廣義線性模型上之研究

研究生：文誠智　　　　　　指導教授：洪志真 博士

國立交通大學統計學研究所

摘要

本篇文章裡，為解決WAT(Wafer Acceptance Test)-EC(Engineering Control)的問題，我們發展了單調無母數迴歸。藉由Gu(2002)，Zhang(2004)所提出的方法加以結合及修正，將反應變數拓展至整個指數族上，與此相關的演算法也會在本文中提出。我們利用Natural Cubic Splines的性質發展出有效率的計算法，並用模擬資料來探討其效率。當反應變數為Bernoulli或Poisson分部時，其模擬的結果都有不錯的表現。此外，在“真實函數”具有單調性的情形下，有單調限制估計量之ASE(Averages Square Error)與無單調限制並沒有明顯差異。然而，當無單調限制之估計量呈現出非單調時，則單調限制估計量在ASE上之表現會明顯優於前者。最後，我們將說明如何利用此方法來篩選EC中的WAT測試項目並且建立適當的管制上下限。

Nonparametric Monotone Regression for Generalized Linear Models

Student: Cheng-Chih Wen   Advisor: Dr. Jyh-Jen Horng Shiau

Institute of Statistic
National Chiao Tung University

Abstract

In this study, motivated by the WAT-EC problem, we develop a nonparametric monotone smoothing spline smoother for analyzing responses from exponential families by combining the methodologies provided in Gu (2002) and Zhang (2004) along with our modification. An algorithm with implementation details is provided. Computation is efficient because we utilize the characteristics of the natural cubic splines. The effectiveness of the proposed method is studied by simulation. The simulation results demonstrate that the proposed method performs well in the regression models with both the Bernoulli and Poisson responses. When the "true" function is monotonic, the proposed monotone estimator performs about the same as the unconstrained smoother in terms of the averaged squared error for the cases when the latter performs well. On the other hand, constrained smoother outperforms the unconstrained smoother when the unconstrained smoother produces non-monotone estimates. As an illustrative example, we demonstrate the proposed method can be used in screening WAT test items for more stringent engineering control and in setting appropriate control limits.

誌　謝

陳鄰安教授曾經說過：「數學可以因數學而數學，統計不能因統計而統計」，我用了兩年的碩士生涯才體會到這兩句話的涵義。不僅如此，除了在統計方面學到許多新知識，也認識很多好朋友。對於此篇論文的完成，我要感謝我的指導老師　洪志真教授仔細認真的批閱，以及教導我做學問應有的態度，並耐心回答我在論文上所遇到的問題，使我順利完成論文。也要感謝我的研究所同學們俊睿、益銘、花花、雪芳、永在和建威，給我鼓勵與歡笑，帶給我美好的碩士生回憶。

最後要感謝博士班的碩慧學姐，除了論文上的協助外，也時常不吝惜分享其人生經驗與價值觀，讓我了解自己欠缺的知識，並在我人生的規劃中給予適時的建議，受益良多。

此外還要感謝家人的支持，以及女朋友的體諒，使我可以毫無顧慮地專心完成論文。僅將此論文獻給我的老師、學姐、家人以及朋友們

文誠智　　　謹誌于

國立交通大學統計研究所

中華民國九十六年六月

Contents

1 Introduction

The Wafer Acceptance Test (WAT) in semiconductor manufacturing is aimed at monitoring whether the electric characteristics of devices, such as voltage, current, and resistance, are regular or not. In Fab, every wafer must go through WAT device testing. In addition, a more stringent control, hereinafter referred to as "engineering control (EC)", is further imposed on passing wafers. For implementing EC, engineers would need to select a set of critical WAT test items and determine more stringent "control" limits (than that of the regular WAT) for each of them. A wafer will be "held" up for further investigation when any of the EC items fails (i.e., exceeding the prescribed EC limits). However, without any objective assessment on how critical the WAT test items are toward the yield, engineers tend to select potential items as many as possible (sometimes even in hundreds) to perform the extra engineering control. Unfortunately, more than tolerable number of false alarms often occur with this practice. As a result, annoyed by the excess number of false alarms, engineers tend to ignore them, despite the extra efforts and costs spent in performing EC. Hence, to make EC more effective, it would be helpful if an assessment tool for screening EC test items is available for engineers and for evaluating the adequacy of the predetermined EC control limits as well.

Motivated by the above problem, as an assessment tool, we propose developing an EC performance curve for each WAT test item. The proposed EC performance curve of a test item aims at presenting the relationship between the EC passing rate and the circuit probe (C_p) yield of the wafers, where the EC passing rate is the probability that a randomly selected wafer passes this EC test item and the C_p yield is the proportion of the chips on the wafer that pass the functional test called the circuit probe test. The logic behind defining the EC performance curve as such is that an effective control of a critical EC test item would improve the process, which in turn leads to a higher C_p yield. Thus an EC test item that can discriminate the C_p yield to some extent would be worthwhile to perform the extra engineering control. As an illustrative example, Figure

1 displays the EC performance curves of three EC test items showing respectively the probabilities, denoted by $p(C_p)$ as a function of the C_p yield, of a wafer passing these EC test items. The dashed curve depicts that the passing probability is one or almost one for $C_p > .15$, indicating that most of wafers with C_p greater than .15 would pass this test easily. In other words, such an EC test item can only discriminate low-yield wafers. Conversely, the dotted curve discriminates only wafers with high yields. The solid curve discriminates better for the middle values of the C_p yield. Thus engineers can pick the EC test items based on the process under study or monitoring. For example, for a process with a fairly high C_p yield, an EC test with a performance curve similar to the dotted curve may be a good candidate for the engineering control.

We remark that a set of inappropriate control limits may make the performance curve of a critical-in-nature EC test item lose some or all of its discrimination power. Thus engineers may be able to adjust the control limits of an EC test to an appropriate level so that this particular EC test would have a desirable discrimination power for effective control.

Furthermore, although the WAT-EC tests and the circuit probe tests are quite different in nature, one tends to expect that the two test results of the same wafer should be somewhat positively associated. Thus it is reasonable to assume that the EC performance curves are monotonic.

An example of some similarity in nature as our WAT-EC problem is the item responses estimation problem discussed in Rossi, Wang, and Ramsay (2002). The data set consists of the responses of N examinees to n question items in a test. Assume that each item is answered either right or wrong. The authors proposed to estimate nonparametrically the probability that examinee j gets item i right from the discrete data and a covariate such as the IQ score for each of the N examinees via the EM Algorithm. Moreover, the discrimination power of the test items was also discussed in the paper.

To develop the EC performance curve for an EC test item, we adopt the nonparametric regression approach to estimating the functional relationship between the passing rate and the C_p yield. In the nonparametric regression approach, the only assumption on the regression function is smoothness and no functional form needs to be specified, which provides a great advantage of flexibility in function estimation and some convenience in modeling. The price we pay for adopting the nonparametric regression approach instead of the parametric regression approach is the slight inefficiency. However, this inefficiency only happens when the specified parametric regression model is adequate.

The statistical model we consider for the WAT-EC data is as follows. The independent variable (i.e., the covariate) of the regression is a random variable X representing the C_p yield of a wafer. The dependent variable Y is the corresponding pass/fail result of the WAT-EC test item for that wafer. Recall that $p(x)$ is the passing probability of the wafer with the C_p yield $X = x$. Then the dependent variable Y has a Bernoulli distribution with a passing probability $p(x)$. As mentioned before, we will estimate the EC performance curve $p(\cdot)$ by nonparametric *monotone* regression.

Although this study is motivated by the WAT-EC application, the nonparametric *monotone* regression estimation method developed in this work can be applied to applications with the dependent variable Y from the exponential family. For example, the number of particles on a wafer with a covariate affecting the number of particles may be modeled by a Poisson distribution in which the mean number of particles may be of interest and could be described as a monotone function of the covariate. Or the number of defects in a product item may be again modeled by a Poisson distribution and the covariate could be a process condition that is monotonically associated with the number of defects when the product item was manufactured.

The topic of monotone function smoothing has been discussed for quite a long time in the literature. One of the major techniques used in the monotone regression focuses on the first derivative of the function to be estimated. Under the assumption that the random errors follow the normal distribution, Ramsay (1998) proposed expressing the first derivative of a monotone function as the exponential of a smooth function and estimating the smooth exponent by a B-spline. Under the same Gaussian model for random errors, by adopting the penalized least squares approach, Zhang (2004) developed a simple method trying to obtain a monotone function estimate by forcing the estimated first derivative of the function to be non-negative or non-positive in computation. Based on the method constructed by Ramsay (1998), Wang (2000) extended the distribution to the exponential family and developed a two-step algorithm to implement a monotonic regression technique. For more research works on nonparametric monotone regression, see Ramsay (1998), Wang (2000), Zhang (2004), and the references cited therein.

In this study, we adopt a different approach from that in Wang (2000). Instead of the exponent approach by Ramsay (1998), We combine the penalized likelihood approach given in Gu (2002) for estimating the parameter function when responses are from exponential families with the Zhang's approach of forcing the estimated parameter function to be monotonic. With this approach, it is more natural that we use smoothing splines rather than B-splines. We remark that Zhang's method has a hard-to-see flaw and with this flaw the monotonicity of the estimated function cannot be guaranteed. We modify Zhang's method to ensure the monotonicity.

The rest of the paper is organized as follows. Section 2 reviews the three components we use in developing our estimation method, including smoothing splines as described in Green and Silverman (1994), Zhang's approach to forcing a smoothing spline estimate to be monotonic, and Gu's approach and algorithm for nonparametric regression when data are from exponential families. Section 3 describes the estimation method and the algorithm we propose in this paper. Section 4 examines the effectiveness of the proposed method by a simulation study for the cases of Bernoulli data and Poisson data. A comparison study is conducted to demonstrate the value of adding the monotone constraint. Section 5 returns to the motivated WAE-EC example and illustrates how to use the proposed method as an assessment tool. Section 6 concludes the paper with a brief summary and some discussions.

2 Literature Review

In the following subsections, we review the three components we use in developing the proposed method, including smoothing splines (in natural cubic splines), monotone smoothing splines (also in natural cubic splines), and smoothing splines for responses from exponential families.

2.1 Smoothing Splines - Natural Cubic Splines

Smoothing splines has been a very popular smoothing technique for decades. For computational purpose, we only review smoothing cubic splines as described in Green and Silverman (1994). For other aspects of smoothing splines, readers are referred to Wahba (1990), Eubank (1999), and Gu (2002) and the research works cited therein.

Consider the problem of fitting a curve from a set of noisy data $\{(x_1, Y_1), \ldots, (x_n, Y_n)\}$, where $x_i \in [a, b], i = 1, \ldots, n$. Let $S_2[a, b]$ denote the space of functions that are differentiable on $[a, b]$ with absolutely continuous first derivative and square integrable second derivative. Given any function g in $S_2[a, b]$, let the penalized sum of squares of g be

$$S(g) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - g(x_i)\}^2 + \lambda \int_a^b \{g''(x)\}^2 dx, \tag{1}$$

where $\lambda > 0$ is the smoothing parameter controlling the tradeoff between the closeness to data and the smoothness of the fitted curve. The smoothing spline estimate $\hat{g}$ is defined as the minimizer of the functional $S(g)$ over all $g \in S_2[a, b]$.

Suppose the real numbers $x_1, \ldots, x_n$ are given on interval $[a, b]$ such that $a \leq x_1 < x_2 < \ldots < x_n \leq b$. A function g defined on $[a, b]$ is called a cubic spline when the following two conditions are satisfied: (i) g is a cubic polynomial on each of the subintervals $(a, x_1), (x_1, x_2), \ldots, (x_n, b)$; (ii) the first and second derivatives of g are continuous at each knot x_i. If, in addition, the second and third derivatives of g are zero at a and b, then g is said to be a *natural cubic spline*. More specifically, a natural cubic spline is linear on the two boundary subintervals $[a, x_1]$ and $[x_n, b]$. It is well known that the smoothing spline estimate $\hat{g}$ is a natural cubic spline (de Boor, 2001;

Wahba, 1990).

Suppose that g is a natural cubic spline on interval $[a, b]$ with knots $x_1, \ldots, x_n$. Denote

$$g_i = g(x_i) \text{ and } \gamma_i = g''(x_i) \text{ for } i = 1, \ldots, n.$$

According to the definition of a natural cubic spline, the second derivative of g at x_1 and x_n are zero, that is, $\gamma_1 = \gamma_n = 0$. Let $\mathbf{g}$ be the vector $(g_1, \ldots, g_n)^T$ and $\boldsymbol{\gamma}$ be the vector $(\gamma_2, \ldots, \gamma_{n-1})^T$. They constructed the following two matrices Q and R for computation. Denote $h_i = x_{i+1} - x_i$ for $i = 1, \ldots, n-1$. Let Q be the $n \times (n-2)$ matrix with elements q_{ij} given by

$$q_{j-1,j} = h_{j-1}^{-1}, q_{j,j} = -h_{j-1}^{-1} - h_j^{-1}, q_{j+1,j} = h_j^{-1},$$

and $q_{ij} = 0$ for $|i - j| \geq 2$ for $i = 1, \ldots, n$ and $j = 2, \ldots, n-1$. Note that the top left element of Q is q_{12} and the bottom right element is $q_{n,n-1}$. R is the symmetric $(n-2) \times (n-2)$ band matrix with nonzero elements r_{ij} given by

$$\begin{aligned} r_{ii} &= \frac{1}{3}(h_{i-1} + h_i) \text{ for } 2 \leq i \leq n-1, \\ r_{i,i+1} &= r_{i+1,i} = \frac{1}{6}h_i \text{ for } 2 \leq i \leq n-2. \end{aligned}$$

The matrix R is strictly diagonal dominant, that is, $|r_{ii}| > \sum_{j \neq i} |r_{ij}|$ for each i. It follows that R is strictly positive definite. Define an $n \times n$ matrix K by

$$K = QR^{-1}Q^T.$$

Green and Silverman (1994) proved that the vectors $\mathbf{g}$ and $\boldsymbol{\gamma}$ can specify a natural cubic spline g if and only if the condition $Q^T\mathbf{g} = R\boldsymbol{\gamma}$ is satisfied. When this condition holds, the roughness penalty satisfies

$$\int_a^b \{g''(t)\}^2 dt = \boldsymbol{\gamma}^T R \boldsymbol{\gamma} = \mathbf{g}^T K \mathbf{g}.$$

Return to the curve fitting problem. Since the smoothing spline estimator $\hat{g}$ is a natural cubic spline, to minimize the penalized sum of squares functional (1), we only need to search over a finite-dimensional class of functions, i.e., the natural cubic splines with knots at the x_i's, instead of the infinite-dimensional space $S_2[a,b]$.

Let g be the natural cubic spline formed by the vectors $\mathbf{g}$ and $\boldsymbol{\gamma}$, and matrices Q and R. Rewrite $S(g)$ in terms of these vectors and matrices as follows. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. Express the residual sum of squares about g as $(\mathbf{Y}-\mathbf{g})^T(\mathbf{Y}-\mathbf{g})$ and the roughness penalty term $\int\{g''(x)\}^2 dx$ as $\mathbf{g}^T K \mathbf{g}$ to obtain

$$\begin{aligned} S(g) &= \frac{1}{n}(\mathbf{Y}-\mathbf{g})^T(\mathbf{Y}-\mathbf{g}) + \lambda \mathbf{g}^T K \mathbf{g} \\ &= \frac{1}{n}\left\{\mathbf{g}^T(I+n\lambda K)\mathbf{g} - 2\mathbf{Y}^T\mathbf{g} + \mathbf{Y}^T\mathbf{Y}\right\}. \end{aligned}$$

Since K is non-negative definite, the matrix $I + n\lambda K$ is strictly positive definite. It therefore follows that $S(g)$ has a unique minimum, which can be expressed as

$$\mathbf{g} = (I + n\lambda K)^{-1}\mathbf{Y}. \tag{2}$$

Green and Silverman (1994) showed that the vector $\mathbf{g}$ can define the smoothing spline $\hat{g}$ uniquely. That is, over the space of all natural cubic splines with knots x_i, $S(g)$ has the unique minimum satisfying (2). Furthermore, the value of $g(x)$ at any point x can be specified by the vectors $\mathbf{g}$ and $\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ can be obtained by solving $Q^T\mathbf{g} = R\boldsymbol{\gamma}$. More specifically, on each subinterval $[x_i, x_{i+1}]$, $1 \le i \le n-1$, it can be shown that

$$\begin{aligned} g(x) \;=\; & \frac{(x-x_i)\,g_{i+1} + (x_{i+1}-x)\,g_i}{h_i} \\ & -\frac{1}{6}(x-x_i)(x_{i+1}-x)\left\{\left(1+\frac{x-x_i}{h_i}\right)\gamma_{i+1} + \left(1+\frac{x_{i+1}-x}{h_i}\right)\gamma_i\right\} \text{ for } x_i \le x \le x_{i+1}. \end{aligned}$$

If x is in the two boundary subintervals, by the fact that a natural cubic spline is linear on the boundary subintervals, we have

$$\begin{aligned} g(x) &= g_1 - (x_1 - x)g'(x_1) \text{ for } x \le x_1, \\ g(x) &= g_n + (x - x_n)g'(x_n) \text{ for } x \ge x_n, \end{aligned}$$

where $g'(x_1)$ and $g'(x_n)$ are derivatives of g at x_1 and x_n, respectively, which can be obtained by

$$\begin{aligned} g'(x_1) &= \frac{g_2 - g_1}{x_2 - x_1} - \frac{1}{6}(x_2 - x_1)\,\gamma_2, \\ g'(x_n) &= \frac{g_n - g_{n-1}}{x_n - x_{n-1}} + \frac{1}{6}(x_n - x_{n-1})\,\gamma_{n-1}. \end{aligned}$$

2.2 Monotone Smoothing Splines

Zhang (2004) proposed a simple and efficient monotone smoother based on smoothing spline estimation. The main idea is to impose a monotone constraint on the derivative of the estimated regression function.

Assume that data $\{(x_i, Y_i),\ i = 1, \ldots, n\}$ are sampled from the following nonparametric regression model

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $f(x)$ is an unknown smooth function with thrice continuous derivatives on interval $[a, b]$ and the random errors ϵ_i's are white noise with mean zero and standard deviation σ. For simplicity, assume that the design points x_i satisfy $a \leq x_1 < \ldots < x_n \leq b$.

A smooth estimator of f can be defined as the minimizer $\hat{f}$ of the following penalized least squares criterion:

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - f(x_i)\}^2 + \lambda \int \{f'''(x)\}^2 dx, \tag{3}$$

where again $\lambda > 0$ is the smoothing parameter.

To derive closed-form formulas for both $\hat{f}(x)$ and its derivative, write $f(x)$ in terms of $g(x) = f'(x)$ as

$$f(x) = f(x_1) + \int_{x_1}^{x} g(u)du. \tag{4}$$

Substitute (4) into (3) to obtain the following regularization criterion:

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - f(x_1) - \int_{x_1}^{x_i} g(x)dx\}^2 + \lambda \int \{g''(x)\}^2 dx. \tag{5}$$

Let $(\hat{f}(x_1), \hat{g})$ be the minimizer of (5).

By treating g as a natural cubic spline with knots x_i's for $i = 1, \ldots, n$, Zhang (2004) established the relationship between the function f and its derivative g using the method of Green and Silverman (1994) and gave the closed-form formulas for $\hat{f}$ and $\hat{g}$, respectively, as in the following.

Denote $f_i = f(x_i)$, $g_i = g(x_i)$, and $\gamma_i = g''(x_i)$ for $i = 1, \ldots, n$. Since g is a natural cubic spline, it follows that $\gamma_1 = \gamma_n = 0$. Let $\mathbf{f} = (f_1, \ldots, f_n)^T$, $\mathbf{g} = (g_1, \ldots, g_n)^T$, and

$\boldsymbol{\gamma} = (\gamma_2, \ldots, \gamma_{n-1})^T$. Let $h_i = x_{i+1} - x_i$, $i = 1, 2, \ldots, n-1$, and Q, R, K be the matrices as defined in Green and Silverman (1994, pages 12,13) (also defined in Subsection 2.1). According to their Theorem 2.1, it can be shown that g is a natural cubic spline with knots $x_1, \ldots, x_n$ if and only if $\boldsymbol{\gamma} = R^{-1}Q^T\mathbf{g}$. As mentioned before,

$$K = QR^{-1}Q^T \text{ and } \int_a^b g''(x)^2 dx = \mathbf{g}^T K \mathbf{g}.$$

Denote the n-dimensional column vector of zeros by $\mathbf{0}_n$. Let $C = (\mathbf{c}_1, \ldots, \mathbf{c}_n)^T$ and $D = (\mathbf{d}_1, \ldots, \mathbf{d}_n)^T$, where $\mathbf{c}_1 = \mathbf{0}_n$, $\mathbf{d}_1 = \mathbf{0}_{n-2}$, and, for $i = 2, 3, \ldots, n$,

$$\begin{aligned} \mathbf{c}_i &= (h_1, h_1 + h_2, \ldots, h_{i-2} + h_{i-1}, h_{i-1}, 0, \ldots, 0)^T, \\ \mathbf{d}_i &= (h_1^3 + h_2^3, \ldots, h_{i-2}^3 + h_{i-1}^3, h_{i-1}^3, 0, \ldots, 0)^T. \end{aligned}$$

Note that $\mathbf{c}_i$ and $\mathbf{d}_i$ are $n \times 1$ and $(n-2) \times 1$ vectors, respectively. According to Proposition 1 in Zhang (2004), the vector $\mathbf{f}$ can be expressed in terms of C, D, Q, R, and $\mathbf{g}$ as

$$\mathbf{f} = f(x_1) \cdot \mathbf{1}_n + \{\frac{1}{2}C - \frac{1}{24}DR^{-1}Q^T\}\mathbf{g}.$$

To find the estimator of the function f, denote

$$M = \frac{1}{2}C - \frac{1}{24}DR^{-1}Q^T \text{ , } \tilde{\mathbf{g}} = \begin{pmatrix} f_1 \\ \mathbf{g} \end{pmatrix} \text{ , } \tilde{M} = \begin{pmatrix} \mathbf{1}_n & M \end{pmatrix} \text{ , and } \tilde{K} = \begin{pmatrix} 1 & \mathbf{0}_n^T \\ \mathbf{0}_n & K \end{pmatrix} .$$

Then the minimizer of the regularization problem (5) is

$$\hat{\tilde{\mathbf{g}}} = \begin{pmatrix} \hat{f}_1 \\ \hat{\mathbf{g}} \end{pmatrix} = (\tilde{M}^T\tilde{M} + n\lambda\tilde{K})^{-1}\tilde{M}^T\mathbf{Y}, \tag{6}$$

where $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$. We then have the estimator $\hat{\mathbf{f}} = \tilde{M}\hat{\tilde{\mathbf{g}}}$.

We remark that Zhang (2004) has a typo that the matrix $\tilde{M}$ is defined as an $(n+1) \times (n+1)$ matrix

$$\tilde{M} = \begin{pmatrix} 1 & \mathbf{0}_n^T \\ \mathbf{1}_n & M \end{pmatrix},$$

which cannot be right since the vector $\mathbf{Y}$ in (6) is an $n \times 1$ vector.

Zhang (2004) developed a simple method with an attempt to obtain to a monotone estimate. Without loss of generality, assume that f is non-decreasing, hence g is non-negative. Then one would wish the estimator $\hat{g}$ to be non-negative as well. To achieve this goal, Zhang (2004) replaced $\hat{g}(x)$ by $\hat{g}_+(x) = max(\hat{g}(x), 0)$. He then estimated $\mathbf{f}$ by

$$\hat{\mathbf{f}} = \tilde{M} \begin{pmatrix} \hat{f}_1 \\ \hat{\mathbf{g}}_+ \end{pmatrix},$$

where $\hat{\mathbf{g}}_+ = (\hat{g}_+(x_1), \cdots, \hat{g}_+(x_n))^T$.

Unfortunately, the monotonicity of $\hat{\mathbf{f}}$ cannot be guaranteed. The reason is that the i-th element of $\hat{\mathbf{f}}$, denoted by $\hat{f}_i$, is not exactly the value of $\hat{f}(x_1) + \int_{x_1}^{x_i} \hat{g}_+(u)du$ as desired. Instead, it is $\hat{f}(x_1) + \int_{x_1}^{x_i} \tilde{g}_+(u)du$, where the function $\tilde{g}_+(\cdot)$ is the natural cubic spline interpolating $\{(x_i, \hat{g}_+(x_i), i = 1, ..., n\}$ and there is no guarantee that it would be nonnegative.

2.3 Regression with Responses from Exponential Families

For the response variable Y from an exponential family distribution along with a covariate x, consider the following conditional density function

$$f(y|x) = \exp\{\frac{y\eta(x) - q(\eta(x))}{\phi} + c(y, \phi)\}, \tag{7}$$

where q, and c are known functions, $\eta(x)$ is the parameter function to be estimated, and $\phi > 0$ is a known dispersion parameter independent of x. It is well known that $E[Y|x] = q'(\eta(x)) = \mu(x)$ and $Var[Y|x] = q''(\eta(x))\phi$.

Assume the responses Y_i corresponding to the covariate x_i, $i = 1, \cdots, n$, are i.i.d. from the exponential family distribution (7). Then the penalized log-likelihood functional can be expressed as

$$-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta(x_i) - q(\eta(x_i))\} + \frac{\lambda}{2}J(\eta), \tag{8}$$

where $J(\lambda)$ is the roughness penalty. Gu (2002) showed that the minimizer of (8) can be computed via the Newton iteration, which updates $\tilde{\eta}$ (the η obtained in the last iteration) by the minimizer of the following penalized weighted least squares functional

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\omega}_i\{\tilde{Y}_i - \eta(x_i)\}^2 + \lambda J(\eta),$$

where $\tilde{Y}_i = \tilde{\eta}(x_i) - \tilde{u}_i/\tilde{w}_i$ with $\tilde{u}_i = -Y_i + q'(\tilde{\eta}(x_i))$ and $\tilde{w}_i = q''(\tilde{\eta}(x_i))$. Note that $\tilde{\omega}_i > 0$ since $Var[Y|x] = q''(\eta(x))\phi > 0$ when $\eta(\cdot) = \tilde{\eta}(\cdot)$.

3 Proposed Method

3.1 Penalized Weighted Least Squares Function Estimator for Exponential Families

Consider the nonparametric regression for a generalized linear model specified within the exponential family. The data observed are i.i.d. samples (x_i, Y_i), for $i = 1, \ldots, n$, from an exponential family distribution. The covariate variable x_i is within some interval $[a, b]$ and the density of Y_i is

$$f(y|x) = \exp\{\frac{y\eta(x) - q(\eta(x))}{\phi} + c(y, \phi)\},$$

where q and c are known functions, and $\phi > 0$ is either known or considered as a nuisance parameter. The unknown parameter function $\eta(x)$ is the central aim of estimation. In our study, $\eta(x)$ needs to meet both smoothness and monotone conditions on interval $[a, b]$. Instead of maximizing the log-likelihood, we choose to minimize the following penalized log-likelihood functional

$$-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta(x_i) - q(\eta(x_i))\} + \frac{\lambda}{2}\int_a^b [D^{(m)}\eta(x)]^2 dx, \tag{9}$$

where λ is the smoothing parameter.

Gu (2002) gave a quadratic approximation of $-Y_i\eta(x_i) + q(\eta(x_i))$ at $\tilde{\eta}(x_i)$ as

$$\frac{1}{2}\tilde{\omega}_i\{\eta(x_i) - \tilde{\eta}(x_i) + \frac{\tilde{u}_i}{\tilde{w}_i}\}^2 + C_i,$$

where $\tilde{u}_i = -Y_i + q'(\tilde{\eta}(x_i))$, $\tilde{\omega}_i = q''(\tilde{\eta}(x_i))$, and C_i is not related to $\eta(x_i)$. Without imposing the constraint that $\eta(x)$ is monotonic, the minimizer of the penalized log likelihood functional (9) can be obtained by recursively finding the minimizer of the penalized weighted least squares functional

$$l = \frac{1}{n}\sum_{i=1}^{n}\tilde{\omega}_i\{\tilde{Y}_i - \eta(x_i)\}^2 + \lambda\int_a^b [D^{(m)}\eta(x)]^2 dx \tag{10}$$

via Newton iteration until convergence, where $\tilde{Y}_i = \tilde{\eta}(x_i) - \tilde{u}_i/\tilde{w}_i$.

3.2 Monotone Natural Spline Estimator

To impose the monotone constraint in the estimation of $\eta(x)$, we focus on the estimation of the first derivative of $\eta(x)$. Since $\eta(x)$ is non-increasing (or non-decreasing), the derivative $\eta'(x)$ is non-positive (or non-negative). There are reasons for dealing with $\eta'(x)$. Firstly, in practice, it is easier to impose non-positiveness (or non-negativeness) on $\eta'(x)$ than to impose monotonicity on $\eta(x)$. Secondly, we can derive a closed-form formula for $\eta'(x)$ easily by the property of natural cubic splines as given in Green and Silverman (1994).

Assume that x_i, $i = 1, \ldots, n$, are real numbers on interval $[a, b]$, which satisfy $a \leq x_1 < x_2 < \ldots < x_n \leq b$. Since we are more focused on the smoothness of $\eta'(x)$ than that of $\eta(x)$, it is natural to choose $m = 3$ in the roughness penalty functional of (10). For any $x \in [a, b]$,

$$\eta(x) = \eta(a) + \int_a^x g(u)du.$$

Substituting the above expression into (10), we have

$$l = \frac{1}{n}\sum_{i=1}^{n} \tilde{\omega}_i\{\tilde{Y}_i - \eta(a) - \int_a^{x_i} g(x)dx\}^2 + \lambda \int_a^b [g''(x)]^2 dx. \tag{11}$$

We remark that Zhang (2004) specified that $\hat{g}$, the minimizer of (5), is a natural cubic spline; however, according to Wahba (1990), $\hat{g}$ should be a piecewise quartic polynomial in $[x_1, x_n]$ and linear in the two boundary subintervals if the function space to search for the minimizer is $S_2[a, b]$.

Mainly for the computational purpose and also for simplicity, we shall restrict the function space to search for the minimizer of (11), denoted by $\hat{g}$, to be the class of natural cubic splines. That is, $\hat{g}$ is set to be a natural cubic spline. We remark that the class of natural cubic splines is a smaller space than $S_2[a, b]$ but it is rich enough for approximating any reasonably smooth function.

Using the same notation as in Green and Silverman (1994), let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$, $\boldsymbol{\gamma} = (\gamma_2, \ldots, \gamma_{n-1})^T$, $\mathbf{g} = (g_1, \ldots, g_n)^T$, where $\eta_i = \eta(x_i)$, $g_i = g(x_i)$, and $\gamma_i = g''(x_i)$ for $i = 1, \ldots, n$. If g is a natural cubic spline, $\gamma_1 = \gamma_n = 0$. We modify the method of

Zhang (2004) by expressing $\eta(x)$ as $\eta(x) = \eta(a)+\int_a^x g(u)du$ instead of $\eta(x_1)+\int_{x_1}^x g(u)du$. Then the matrices of C and D need to be modified as well. Denote $\eta(a)$ by η_a.

Let $h_i = x_{i+1}-x_i$, for $i = 1, \ldots, n-1$, $h_0 = x_1-a$, $k_1 = h_0\,(2+h_0/h_1)$, $k_2 = -{h_0}^2/h_1$, and $k_3 = -2h_0^2\,h_1$. Let the matrices Q, R, and K be defined as in Subsection 2.1. Modify the matrices C and D by replacing $\mathbf{c}_i$ with $\mathbf{c}_i + k_1 \cdot \mathbf{e}_1^n + k_2 \cdot \mathbf{e}_2^n$ and $\mathbf{d}_i$ with $\mathbf{d}_i + k_3 \cdot \mathbf{e}_1^n$, where $\mathbf{e}_i^k$ is the k-dimensional unit vector with the ith element being 1 and 0 elsewhere. In words, the matrix C is modified by adding k_1 to the first column and k_2 to the second column. Similarly, D is modified by adding k_3 to the first column. Let $M = \frac{1}{2}C - \frac{1}{24}DR^{-1}Q^T$ as before.

Proposition 1 gives the relationship between $\boldsymbol{\eta}$ and $\mathbf{g}$. All the proofs in this paper are given in the Appendix A.

Proposition 1 $\quad \boldsymbol{\eta} = \eta_a \cdot \mathbf{1}_n + M\,\mathbf{g}$.

The main reason that we consider $\eta(a)$ rather than $\eta(x_1)$ is that the matrix M constructed in this way is invertible while the matrix M in Zhang (2004) is not. Another advantage is that $\eta(a)$ can be specified by the vector $\mathbf{g}$, utilizing the fact that a natural cubic spline is linear on the boundary subinterval $[a, x_1]$, see the Appendix A.

Proposition 2 shows that the matrix M in Proposition 1 is invertible. In the following, when the dimension of a matrix or vector helps reading, we will add the dimension as the subscript in the notation. For example, $M_{n,n}$ denotes the n by n matrix M.

Proposition 2 $\quad M_{n,n}$ is invertible for all $n \geq 3$.

According to Proposition 1, the vector $\boldsymbol{\eta}$ can be specified by $\mathbf{g}$ and η_a. Given η_a, (11) can be written in matrix form as

$$\begin{aligned} l &= \frac{1}{n}(\tilde{\mathbf{Y}} - M\mathbf{g})^T\tilde{W}(\tilde{\mathbf{Y}} - M\mathbf{g}) + \lambda\mathbf{g}^T K\mathbf{g} \\ &= \frac{1}{n}\{\mathbf{g}^T(M^T\tilde{W}M + n\lambda K)\mathbf{g} - 2\tilde{\mathbf{Y}}^T\tilde{W}M\mathbf{g} + \tilde{\mathbf{Y}}^T\tilde{W}\tilde{\mathbf{Y}}\}, \end{aligned} \tag{12}$$

where $\tilde{\mathbf{Y}} = \{\tilde{Y}_1 - \eta_a, \ldots, \tilde{Y}_n - \eta_a\}^T$ and $\tilde{W} = diag(\tilde{\omega}_1, \ldots, \tilde{\omega}_n)$.

Since K is semi-positive definite, the matrix $M^T\tilde{W}M$ is strictly positive definite by Proposition 2 and $\tilde{\omega}_i > 0$ for $i = 1, \ldots, n$, it follows that (12) has a unique minimum at

$$\hat{\mathbf{g}} = (M^T\tilde{W}M + n\lambda K)^{-1} M^T\tilde{W}\tilde{\mathbf{Y}}. \tag{13}$$

Suppose the parameter function $\hat{\eta}$ is monotone increasing. Then $\hat{g}$ must be non-negative everywhere. To achieve this, we follow the same approach adopted by Zhang (2004) by setting $\hat{g}(x_i)$ to 0 when $\hat{g}(x_i) < 0$. Let $\hat{\mathbf{g}}_+ = (\hat{g}_+(x_1), \cdots, \hat{g}_+(x_n))^T$, where $\hat{g}_+(x_i) = max(\hat{g}(x_i), 0)$. We then use $\hat{\mathbf{g}}_+$ to construct vector $\hat{\boldsymbol{\eta}}$ instead of using $\hat{\mathbf{g}}$. That is,

$$\hat{\boldsymbol{\eta}} = \hat{\eta}_a \cdot \mathbf{1}_n + M\hat{\mathbf{g}}_+. \tag{14}$$

It is interesting to observe that $M\hat{\mathbf{g}}_+$ can be conceived as integrating the natural cubic spline that interpolates the points $\{(x_i, \hat{g}_+(x_i)), i = 1, \ldots, n\}$. Denote the interpolating natural cubic spline by $\tilde{g}_+$. Unfortunately, when $\hat{g}_+(x_k) = \hat{g}_+(x_{k+1}) = 0$ for some k, it is impossible to have $\hat{g}_+(x) \geq 0$ for all $x \in [x_k, x_{k+1}]$. Then the monotonicity of $\hat{\eta}$ is lost. To remedy this problem, we modify the estimate to ensure the monotonicity as follows.

For $x_i \leq x \leq x_{i+1}, i = 1, ..., n-1$, let

$$\hat{\eta}_{mon}(x) = \hat{\eta}(a) + \sum_{j=1}^{i} \tau_j \int_{x_{j-1}}^{x_j} \hat{g}_+(u)du + \tau_x \int_{x_i}^{x} \hat{g}_+(u)du,$$

where τ_i is an indicator function defined as 1 if $\hat{\eta}_i > \hat{\eta}_{i-1}$ and 0 otherwise and $\tau_x = 1$ when $\int_{x_i}^{x} \hat{g}_+(u)du > 0$ and 0 otherwise. Note that the value of $\int_{x_i}^{x} \hat{g}_+(u)du$ can be obtained as described in Subsection 2.1 since $\hat{g}_+$ is a natural cubic spline. It is obvious $\hat{\eta}_{mon}$ is monotonic.

For computational purpose, we shall express $\hat{\boldsymbol{\eta}}_{mon} = (\eta_{mon,1}, ..., \eta_{mon,n})^T$ in matrix form. Let $\tilde{M}$ be the $(n+1) \times (n+1)$ matrix given by

$$\tilde{M} = \begin{pmatrix} 1 & 0_n^T \\ 1_n & M \end{pmatrix}.$$

Let N be the $n \times (n+1)$ matrix with elements n_{ij}, $1 \leq i, j \leq n$, given by $n_{ii} = -1$, $n_{i,i+1} = 1$ for $1 \leq i \leq n$ and $n_{ij} = 0$ elsewhere. Let $S = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)$ be the $n \times n$ matrix given by $\mathbf{s}_i = (0_{i-1}^T, \tau_i \cdot 1_{n-i+1}^T)^T$ for $i = 1, \ldots, n$.

Proposition 3 $\hat{\boldsymbol{\eta}}_{mon} = \hat{\eta}_a \cdot \mathbf{1}_n + SN\tilde{M}\bar{\mathbf{g}}$, where $\bar{\mathbf{g}} = \begin{pmatrix} \hat{\eta}_a \\ \hat{\mathbf{g}}_+ \end{pmatrix}$.

Note that

$$N\tilde{M}\bar{g} = N\tilde{M}\begin{pmatrix}\hat{\eta}_a\\ \hat{\mathbf{g}}_+\end{pmatrix} = N\begin{pmatrix}\hat{\eta}_a\\ \hat{\eta}_1\\ \vdots\\ \hat{\eta}_n\end{pmatrix} = \begin{pmatrix}\hat{\eta}_1-\hat{\eta}_a\\ \hat{\eta}_2-\hat{\eta}_1\\ \vdots\\ \hat{\eta}_n-\hat{\eta}_{n-1}\end{pmatrix},$$

which computes the integral $\int_{x_i}^{x_{i+1}} \tilde{g}_+(u)du$ for each subinterval. The effect of S is to accumulate these integrals up to x_i but skip those integrals $\int_{x_i}^{x_{i+1}} g(u)du$ for which $\hat{\eta}_i \leq \hat{\eta}_{i-1}$.

3.3 Algorithm

We propose using the back-fitting approach to obtain estimators $\hat{\eta}_a$ and $\hat{\mathbf{g}}_+$. For back-fitting, see Hastie and Tibshirani (1990). Recall that the log-likelihood function for the exponential family is

$$l = -\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta(x_i) - q(\eta(x_i))\} + \frac{\lambda}{2}\int_a^b [g''(x)]^2 dx, \tag{15}$$

where $\eta(x_i) = \eta_a + M[i,.]\mathbf{g}$, if g is a natural cubic spline. Given $\mathbf{g}$, we can get a suitable value of η_a by solving the equation

$$\frac{\partial l}{\partial \eta_a} = -\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \frac{\partial q(\eta(x_i))}{\partial \eta_a}\} = 0.$$

Substitute the new η_a into (12) to obtain the new estimate of $\mathbf{g}$ by (13). Thus, repeat the iteration until the value of η_a converges. Since the equation $\partial l/\partial \eta_a = 0$ is non-linear, we propose using the Newton-Ralphson method to get the new estimate of $\eta(a)$. For details of the Newton-Ralphson method, see, for example, Burden and Faires (2001).

Back-fitting Algorithm

Step 1

Given the covariate $(x_1, \ldots, x_n)$, calculate the $n \times n$ matrix C, $n \times (n-2)$ matrices Q and D, $(n-2) \times (n-2)$ matrix R, then compute the matrices $K = QR^{-1}Q^T$ and $M = \frac{1}{2}C - \frac{1}{24}DR^{-1}Q^T$.

Step 2

Begin with iteration $k = 0$. Set $\hat{\mathbf{g}}^{(0)} = \left(\hat{g}_1^{(0)}, \ldots, \hat{g}_n^{(0)}\right)^T$ and $\hat{\eta}_a^{(0)}$ to some initial values. Set the tolerance level T.

Step 3

Given $\hat{\mathbf{g}}^{(k)}$, update $\hat{\eta}_a^{(k)}$ by $\hat{\eta}_a^{(k+1)}$, the minimizer of (15), with the Newton-Ralphson method.

Step 4

Construct the $n \times n$ diagonal matrix $\tilde{W}$ and $\tilde{\mathbf{Y}}$ with $\hat{\mathbf{g}}^{(k)}$ and $\hat{\eta}_a^{(k+1)}$. Then update $\hat{\mathbf{g}}^{(k)}$ by $\hat{\mathbf{g}}^{(k+1)} = \left(M^T\tilde{W}M + n\lambda K\right)^{-1} M^T\tilde{W}\tilde{\mathbf{Y}}$.

Step 5

If $\left|\frac{\hat{\eta}^{(k+1)}(a)-\hat{\eta}^{(k)}(a)}{\hat{\eta}^{(k)}(a)}\right| < T$, stop iterating and set vector $\hat{\boldsymbol{\eta}} = \hat{\eta}_a^{(k+1)} \cdot \mathbf{1}_n + SN\tilde{M}\bar{\mathbf{g}}_+$. Otherwise, set $\hat{\mathbf{g}}^{(k)} \leftarrow \hat{\mathbf{g}}_+^{(k+1)}$ and $\hat{\eta}_a^{(k)} \leftarrow \hat{\eta}_a^{(k+1)}$, return to **Step 3**.

4 Simulation Studies

4.1 Performance Evaluation

To evaluate the effectiveness of the proposed method, we apply the monotone regression smoother on some data generated from exponential family models, including Bernoulli and Poisson data as illustrative examples. The smoothing parameter λ is chosen by the Generalized Cross Validation (GCV) method proposed by Craven and Wahba (1979).

4.1.1 Bernoulli Data

n observations, $\{(x_i, Y_i), i = 1, \cdots, n\}$, are generated independently, in which x_i's are generated independently from interval $[0, 1]$ uniformly and Y_i is generated from the Bernoulli distribution with the probability function $P(Y_i = 1|x_i) = p(x_i)$, where $p(x)$ is a smooth monotone function for all $x \in [0, 1]$. The conditional density function of the Bernoulli response Y given the covariate x can be written as

$$f(y|x) = exp\{y\,\eta(x) - \log(1 + exp(\eta(x)))\},$$

where $\eta(x) = \log(p(x)/(1 - p(x)))$.

As an illustrative example, we choose n=200 and $p(x) = 1-(1-x^{4.5})^{2.5}$ for $x \in [0, 1]$. That is, $x_i \sim U(0, 1)$, $Y_i \sim Bernoulli(p(x_i))$, $i = 1, \ldots, 200$. The simulation results are displayed in Figure 2. The solid line is the target function $p(x)$, the dotted line is the estimated smooth curve under monotone constraint with smoothing parameter $\lambda = 0.00005$ chosen by GCV. It is observed that the estimated function is fairly close to the target function for this example.

In Figure 2(a), the dots are raw data $\{(x_i, Y_i)\}$, while the dots in Figure 2(b) are binned data. For binning data, the interval $[0, 1]$ is divided into 25 equally spaced subintervals. For each subinterval, we count the points and calculate the proportion of the 1's in that subinterval. The points plotted on Figure 2(b) are the proportion of $1's$ versus the midpoint of the corresponding subinterval. While it is hard to read from the raw data the information of $p(x)$, the binned data can follow the trend of the underlying target function $p(x)$ pretty well.

4.1.2 Poisson Data

The conditional density of the Poisson distribution can be expressed as

$$f(y|x) = \frac{\phi(x)^y \exp\left(-\phi(x)\right)}{y!} = \exp\{y\eta(x) - e^{\eta(x)} - log(y!)\},$$

where $\eta(x) = log\,\phi(x)$. Assume the target function is

$$\phi(x) = \log(x^2 + 1) \text{ for } 1 \leq x \leq 3.$$

200 covariates $\{x_i\}$ are generated independently from $U(1, 3)$, and the response Y_i follows Poisson($\phi(x_i)$), for $1 \leq i \leq 200$. Figure 3 shows the fitting results. In Figure 3, the solid line is the target function $\phi(x)$ and the dotted line is the estimated parameter function $\hat{\phi}(x)$ with smoothing parameter $\lambda_{GCV} = 1.5$ chosen by the GCV method. To

see how well the GCV estimate performs, we also show the "optimal" estimated curve (the dash-dot line) for which $\lambda_{opt} = 0.5$ about this example is the λ minimizing the averaged squared error (ASE) defined as

$$ASE(\hat{\phi}(x)) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\phi}(x_i) - \phi(x_i)\right)^2.$$

Note that $\hat{\phi}(\cdot)$ depends on λ.

Again, the dots in Figure 3(a) are raw data $\{(x_i, Y_i)\}$ while the dots in Figure 3(b) are the binned data. When comparing the two estimated curves in terms of ASE, we find the values of ASE are 0.01190 and 0.01189 for λ_{opt} and λ_{GCV}, respectively. There is no obvious difference between the curves corresponding to $\lambda = 1.5$ and $\lambda = 0.5$.

Consider another example in which the curvature of the mean function has more variation. Let

$$\phi(x) = \frac{1}{e} - e^{-x} + x - \frac{\sin(\pi x)}{\pi} \quad \text{for} \quad 1 \le x \le 3.$$

Similarly, generate 200 x_i's from $U(1, 3)$ randomly and Y_i's accordingly. $\lambda_{GCV} = 0.25$ while $\lambda_{opt} = 2.5 \times 10^{-5}$ by minimizing ASE criterion. The results are shown in Figure 4. The estimated curve with λ_{GCV} has obvious departures from the target function ϕ. On the other hand, the "optimal" curve with λ_{opt} captures the main trend of the target function. In addition, the ASE is 0.008 for λ_{opt} and is 0.031 for λ_{GCV}. Note that, as observed from Figure 4(b), the binned data are so noisy that it is hard to expect a data-driven method like GCV would perform well.

4.2 Monotone vs. Constraint-Free Smoothing Spline Estimator

We are interested in knowing whether adding the monotone constraint is value-added in estimation of monotone functions. In other words, if the underlying function is monotonic, would a regular (unconstrained) smoother performs poorer than or as well as a constrained smoother? Also, how often a regular smoother would produce a non-monotone estimate when the true function is monotonic? To answer these questions, we compare the proposed method with the method given in Gu (2002) by a simulation study.

Consider the Bernoulli example in Subsection 4.1.1. The true parameter function under study is of the form $p(x) = 1 - (1 - x^{\alpha})^{\beta}$, where α and β control the shape and the speed of going upward of the function. When $\alpha \geq 1$ and $\beta \geq 1$, $p(x)$ is increasing for all $x \in [0, 1]$. If β is much larger than α, the curve $p(x)$ climbs up to 1 rapidly. On the other hand, if β is much smaller than α, the curve reaches 1 slowly. Thus, three settings are studied: $(\alpha, \beta) = (1.98, 28), (6.28, 17.67)$, and $(6.9, 1.1)$ (in the order from fast to slow). Figure 1 shows these three functions. The effect of the sample size is also studied with $n = 50, 100$, and 200.

Let $\hat{p}_m(x)$ denote the estimate under the monotone constraint and $\hat{p}_g(x)$ denote the unconstrainted estimate developed by Gu (2002). Define $ASE(\hat{p}) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}(x_i) - p(x_i)\right)^2$ and calculate $ASE(\hat{p}_m)$ and $ASE(\hat{p}_g)$ for the same data set. For each setting of (α, β) and n, repeat the trial 10000 times. Table 1 displays the percentage of monotone $\hat{p}_g$ in 10000 trials. It is observed that Gu's estimate ($\hat{p}_g$) tends to produce more non-monotone estimates for smaller n. Table 2 shows the proportion of $ASE(\hat{p}_g) \geq ASE(\hat{p}_m)$. We see that, for all n under study, when the true function rises slowly, $\hat{p}_m$ performs better than $\hat{p}_g$, while $\hat{p}_g$ performs better than $\hat{p}_m$ for functions rising rapidly. And for all three target functions, the performance of $\hat{p}_m$ gets better when the sample size n gets larger. For illustration, Figure 5 shows some cases with monotone $\hat{p}_g$ and some with non-monotone $\hat{p}_g$ along with the corresponding $\hat{p}_m$.

For distribution comparison, boxplots of the 10000 $ASE(\hat{p}_g)$ and the corresponding 10000 $ASE(\hat{p}_m)$ are displayed in Figures 6-8. The sample quartiles of these estimates are given in Table 3. It seems there is no significant difference between the distributions of $ASE(\hat{p}_g)$ and $ASE(\hat{p}_m)$. In summary, $\hat{p}_m$ performs as well as $\hat{p}_g$ in terms of ASE, while ensuring monotonicity.

To see how bad (well) the unconstrained smoother can be when it produces a non-monotone (monotone) estimate, we generate 10000 cases of non-monotone (monotone) $\hat{p}_g$'s with $n = 100$. For the three target functions, Figures 9-11 compare the boxplots of

10000 $\hat{p}_g$ versus the corresponding 10000 $\hat{p}_m$ under the condition that $\hat{p}_g$ is monotonic (in panel (a)) or not monotonic (in panel (b)). It is clear to see from these figures that the constrained smoother outperforms the unconstrained one when $\hat{p}_g$ is not monotonic, while performing as well as $\hat{p}_g$ when $\hat{p}_g$ is monotonic. In summary, constraining monotonicity prevents the chance of poor estimation while attaining about the same performance for cases the unconstrained method performs well.

5 Examples

Return to the motivated WAT-EC example described above. In this section, we demonstrate the proposed method can be useful in setting appropriate EC limits to achieve better discrimination power and in the process of screening WAT test items for further engineering control.

For the first purpose, we generate responses Y_i's directly by checking if the measurements are within the control limits. Suppose that 200 wafers are taken from some lots and measured the voltage at some testkeys on each wafer. Consider the relationship between the C_p yield and the mean voltage $\mu(C_{p,i})$ for the i-th wafer satisfying

$$\mu(C_{p,i}) = 1 + 0.1\ \exp(-1.5\, C_{p,i}^3)\ T_i,$$

where $C_{p,i}$ is the yield of the i-th wafer generated randomly from the beta distribution $Beta(8,2)$ and T_i is a random variable with probability $P(T_i = 1) = P(T_i = -1) = 0.5$, for $i = 1\ldots, 200$. $Beta(8,2)$ is chosen to mimic the reality because it is skewed to the left. The function $\mu(C_p)$ indicates that the mean voltage approaches to the target voltage level 1 as C_p increases to 1. Suppose the distribution of voltage for the i-th wafer follows the normal distribution with mean $\mu(C_{p,i})$ and standard deviation 0.1. 9 measurements are generated for each wafer and the mean voltage $\bar{v}_i$ is computed. Figure 12 presents a histogram of these 200 mean voltages, for which the sample mean is 1.005 and sample standard deviation is 0.06. Suppose an engineer sets the upper (UCL) and lower control limits (UCL) at LCL=0.88 and UCL=1.12, respectively. Define $Y_i = 1$ if $\bar{v}_i$ is within the control limits and $Y_i = 0$ otherwise. With data $\{(C_{p,i}, Y_i)\}$, the passing rate for C_p yield is estimated and displayed in Figure 13(a). The passing rate curve seems to lack the discrimination power for $C_p \geq .6$ while showing high discrimination power for C_p in $(0.4, 0.5)$. Such an EC performance curve is not useful for most of the current processes in IC industries If this test item is critical in nature, the undesirable low discrimination power may be caused by the ad hoc choices of control limits. To demonstrate this, we reset the control limits to LCL=.93 and UCL=1.07. Then the EC performance curve displayed in Figure 13(b) indicates a fairly good discrimination power in the range of (.4, 1). A further reduction of the control limits to LCL=.97 and UCL=1.03 is too stringent, since the passing rate drops down to about 60% or below even for very high C_p.

Moreover, consider the case that there is no significant relationship between the C_p yield and voltage. Assume $C_{p,i} \sim Beta(8, 2)$ and mean voltage $\mu_i \sim Beta(10, 10) + 0.5$, for $i = 1, \ldots, 200$, and they are independent. Generate 200 $\bar{v}_i$ accordingly as described above. For these 200 $\bar{v}_i$, the sample mean is 0.99 and sample standard deviation is 0.10. Let UCL=1.1 and LCL=0.9. The corresponding passing rate presented in Figure 13(d) is flat for most of the C_p range. If we change the control limits, the resulting passing rate changes only in the level but has a similar pattern. A test items with such kind of EC performance curve indicates the test results may be not so much related to the C_p yield. Including this test item for EC may merely increase the number of false alarms. Consequently, this test item should not be chosen for engineering control.

6 Conclusions

In this study, motivated by the WAT-EC problem, we develop a nonparametric monotone smoothing spline smoother for analyzing responses from exponential families by combining the methodologies provided in Gu (2002) and Zhang (2004) along with our modification. An algorithm with implementation details is provided. Computation is efficient because we utilize the characteristics of the natural cubic splines. The simulation results demonstrate that the proposed method performs well in the regression models with both the Bernoulli and Poisson responses. When the “true” function is monotonic, the proposed monotone estimator performs about the same as the unconstrained smoother in terms of the averaged squared error for the cases when the latter performs well. On the other hand, constrained smoother outperforms the unconstrained smoother when the unconstrained smoother produces non-monotone estimates. Thus, the choice is obvious. If the function is monotonic in nature, then we should choose the method with the monotone constraint imposed.

As an illustrative example, we demonstrate the proposed method can be used in screening test items for engineering control and in setting appropriate control limits.

With the nonparametric regression in nature, the proposed method has the great advantage of model flexibility. Also since the method can be applied to all kinds of data as long as they follow the exponential family, it can find many potential applications in areas such as industries, medicine, education, social studies, and so on.

A Appendix: Proofs

Proposition 1.

$$\boldsymbol{\eta} = \eta_a \cdot \mathbf{1}_n + M\mathbf{g}.$$

Proof. Define $x_i - x_{i-1} = h_{i-1}$, for $i = 2, \ldots, n$, and let $h_0 = x_1 - a$. Since the function g is linear in subinterval $[a, x_1]$, it shows that

$$\int_a^{x_1} g(u)du = \frac{h_0(g_1 + g(a))}{2}. \tag{16}$$

By Green and Silverman (1994),

$$g_1' = \frac{g_2 - g_1}{x_2 - x_1} - \frac{1}{6}(x_2 - x_1)\gamma_2$$

and

$$g(x) = g_1 - (x_1 - x)g_1' \text{ for } x \le x_1.$$

Then $g(a)$ can be expressed in terms of g_1, g_2, and γ_2 as

$$g(a) = (1 + \frac{h_0}{h_1})g_1 - \frac{h_0}{h_1}g_2 + \frac{1}{6}h_0 h_1 \gamma_2.$$

Substituting the above expression into (16), we get

$$\int_a^{x_1} g(u)du = \frac{1}{2}\{h_0(2 + \frac{h_0}{h_1})g_1 - \frac{{h_0}^2}{h_1}g_2\} - \frac{1}{24}\{-2{h_0}^2 h_1 \gamma_2\}. \tag{17}$$

According to the proposition in Zhang (2004),

$$\int_{x_1}^{x_i} g(x)dx = \frac{1}{2}\mathbf{c}_i^T\mathbf{g} - \frac{1}{24}\mathbf{d}_i^T\boldsymbol{\gamma} \quad \text{for } 1 \le i \le n,$$

where $\mathbf{g}$ and $\mathbf{c}_i$ are $n \times 1$ vectors, $\boldsymbol{\gamma}$ and $\mathbf{d}_i$ are $(n-2) \times 1$ vectors. $\mathbf{c}_i$ and $\mathbf{d}_i$ are the same as that in Zhang (2004). Denoting $k_1 = (2 + h_0/h_1)h_0$, $k_2 = -{h_0}^2/h_1$, and $k_3 = -2{h_0}^2 h_1$, we have

$$\begin{aligned}
\eta_i &= \eta(a) + \int_a^{x_i} g(x)dx = \eta(a) + \int_a^{x_1} g(x)dx + \int_{x_1}^{x_i} g(x)dx \\
&= \eta(a) + \frac{1}{2}(k_1 g_1 + k_2 g_2 + \mathbf{c}_i^T\mathbf{g}) - \frac{1}{24}(\mathbf{d}_i^T\boldsymbol{\gamma} + k_3\gamma_2) \\
&= \eta(a) + \frac{1}{2}(\mathbf{c}_i + k_1 \cdot \mathbf{e}_1^n + k_2 \cdot \mathbf{e}_2^n)^T\mathbf{g} - \frac{1}{24}(\mathbf{d}_i + k_3 \cdot \mathbf{e}_1^{n-2})^T\boldsymbol{\gamma},
\end{aligned} \tag{18}$$

where $\mathbf{e}_i^k$ is the k-dimensional unit vector with the ith element being 1 and 0 elsewhere. Since $\boldsymbol{\gamma} = R^{-1}Q^T\mathbf{g}$, Proposition 1 then holds by (18). □

Proposition 2. *$M_{n,n}$ is invertible for all $n \geq 3$.*

Proof. Let $G_{n,n} = \frac{1}{2}C_{n,n}$ and $H_{n,n-2} = \frac{1}{24}D_{n,n-2}$. Then $M_{n,n} = G_{n,n} - H_{n,n-2}R^{-1}_{n-2,n-2}Q^T_{n-2,n}$. It suffices to show that $det(M_{n,n}) \neq 0$. Define the $(2n-2) \times (2n-2)$ matrix N_{2n-2} by

$$N_{2n-2} = \begin{bmatrix} G_{n,n} & H_{n,n-2} \\ Q^T_{n-2,n} & R_{n-2,n-2} \end{bmatrix}.$$

By the properties of block matrix decomposition,

$$\begin{bmatrix} G_{n,n} & H_{n,n-2} \\ Q^T_{n-2,n} & R_{n-2,n-2} \end{bmatrix} = \begin{bmatrix} I & HR^{-1} \\ O & I \end{bmatrix} \begin{bmatrix} M_{n,n} & O \\ O & R \end{bmatrix} \begin{bmatrix} I & O \\ R^{-1}Q^T & I \end{bmatrix}.$$

Since the matrix R is invertible, we then know that $det(M_{n,n}) \propto det(N_{2n-2})$. Thus, we only need to prove that $det(N_{2n-2}) \neq 0$ for all $n \geq 3$. More specifically, we simplify the matrix N by some elementary matrix operations into the matrix N'

$$det(N_{2n-2}) \propto det(N'_{2n-2}) = \begin{vmatrix} G'_{n,n} & H'_{n,n-2} \\ Q'^T_{n-2,n} & R'_{n-2,n-2} \end{vmatrix}, \tag{19}$$

where $G'_{n,n}$ is the matrix with entries $g_{i,j}$ given by $g_{i,i} = g_{i,i+1} = h^{-1}_{i-1}$ for $2 \leq i \leq n$, $g_{1,1} = k_1/2$, $g_{1,2} = k_2/2$, and $g_{i,j} = 0$ elsewhere; the matrix $H'_{n,n-2}$ has elements $\nu_{i,j}$, for $2 \leq i \leq (n-1)$, $\nu_{i,i-2} = \nu_{i,i-1} = \frac{h_{i-1}}{12}$, $\nu_{1,1} = \frac{-h_0^2 h_1}{12}$, $\nu_{2,1} = \frac{h_1}{12}$,$\nu_{n,n-2} = \frac{h_{n-1}}{12}$, and 0 elsewhere; $Q'^T_{n-2,n}$ is the $(n-2) \times n$ matrix with elements $q_{i,j}$, where $q_{i,i+1} = -2h_i^{-1} - 2h^{-1}_{i+1}$ for $1 \leq i \leq (n-2)$ and 0 elsewhere; the matrix $R'_{n-2,n-2}$ is a symmetric band matrix with entries $r_{i,i}$, $r_{i,i} = \frac{1}{4}(h_i + h_{i+1})$ for $1 \leq i \leq (n-2)$, $r_{i,i+1} = r_{i+1,i} = \frac{1}{12}h_{i+1}$ for $1 \leq i \leq (n-3)$, and 0 otherwise. When $n = 2$, N'_2 is a 2×2 matrix, where the first row is $(k_1/2, k_2/2)$ and the second row is (h_1^{-1}, h_1^{-1}). It is clearly that $det(N'_2) > 0$ since $k_1 > k_2$.

According to Lemma 1 stated below, for all $n \geq 3$, the determinant of N'_{2n-2} can be expressed in terms of $N'_{2(n-1)-2}$ by

$$det(N'_{2n-2}) = \frac{1}{4}det(N'_{2(n-1)-2}) + K_n,$$

where K_n is a positive number. Since $det(N'_2) > 0$, it follows clearly that $det(N'_{2n-2}) > 0$ for all $n \geq 3$, which implies that the matrix $M_{n,n}$ is invertible. □

For illustration, Appendix B gives N'_{2n-2} for $n = 4$ and 5 and expresses N'_{2*5-2} in terms of N'_{2*4-2}.

Lemma 1. *For the square matrix N'_{2n-2} defined by (19), for $n \geq 3$,*

$$det(N'_{2n-2}) = \frac{1}{4} det(N'_{2(n-1)-2}) + K_n, \tag{20}$$

for some $K_n > 0$.

Proof. We will show the equality (20) by induction.

Step 1:

When $n = 2$, it is clear that $det(N'_2) = h_0\ (h_0 + h_1)/{h_1}^2 > 0$.

When $n = 3$,

$$\begin{aligned} det(N'_4) &= \frac{h_0\,(3h_0\,(h_1+h_2)+h_1\,(2h_1+3h_2))}{12h_1^2h_2^2} \\ &= \frac{h_0\ (3\,h_0 + 3\,h_1)}{12\,{h_1}^2} + \frac{h_0\ \left(3\,h_0\,{h_1}^2 + 2\,{h_1}^3\right)}{12\,{h_1}^2\,{h_2}^2} + \frac{h_0\ \left(6\,h_0\,h_1 + 5\,{h_1}^2\right)}{12\,{h_1}^2\,h_2} \\ &= \frac{1}{4} det(N'_2) + K_3, \end{aligned}$$

where $K_3 = \frac{h_0\,\left(3\,h_0\,{h_1}^2+2\,{h_1}^3\right)}{12\,{h_1}^2\,{h_2}^2} + \frac{h_0\,\left(6\,h_0\,h_1+5\,{h_1}^2\right)}{12\,{h_1}^2\,h_2} > 0$. Thus (20) holds for $n = 3$.

When $n = 4$, we calculate the determinant of N'_{2*4-2}. It follows that

$$det(N'_6) = \frac{1}{4} det(N'_4) + K_4,$$

where

$$\begin{aligned} K_4 &= \frac{h_0\ \left(9\,h_0\,{h_1}^2\,{h_2}^2 + 6\,{h_1}^3\,{h_2}^2 + 12\,h_0\,h_1\,{h_2}^3 + 10\,{h_1}^2\,{h_2}^3 + 4\,h_0\,{h_2}^4 + 4\,h_1\,{h_2}^4\right)}{144\,{h_1}^2\,{h_2}^2\,{h_3}^2} \\ &+ \frac{h_0\ \left(18\,h_0\,{h_1}^2\,h_2 + 12\,{h_1}^3\,h_2 + 30\,h_0\,h_1\,{h_2}^2 + 25\,{h_1}^2\,{h_2}^2 + 12\,h_0\,{h_2}^3 + 12\,h_1\,{h_2}^3\right)}{144\,{h_1}^2\,{h_2}^2\,h_3}. \end{aligned}$$

It is clear that (20) holds since $h_0, h_1, h_2, h_3 > 0$.

Step 2:

We would like to prove (20) for general n by induction. For a fixed $k > 5$, assume that $det(N'_{2n-2}) > 0$ for $n = k-3$ and (20) holds for $n = k-2$ and $n = k-1$.

When $n = k$, consider the matrix form for N'_{2k-2}

$$N'_{2k-2} = \left[\begin{array}{cc} G'_{k,k} & H'_{k,k-2} \\ Q'^{T}_{k-2,k} & R'_{k-2,k-2} \end{array} \right].$$

By the structures of G',H',Q', and R', we can write the determinant of N'_{2k-2} in terms of $N'_{2(k-1)-2}$. More specifically, we interchange some column vectors of N'_{2k-2}, such that the i-th column vector becomes the $(i-1)$-th column vector for $i = k+1, \ldots, 2k-1$, and the k-th column vector becomes the $(2k-1)$-th column vector by some permutation operations. Do the same thing with the row vectors. We then obtain

$$det(N'_{2k-2}) = \left| \begin{array}{cc} N'_{2k-4} & U_{2k-4,2} \\ V_{2,2k-4} & T_{2,2} \end{array} \right| \qquad (21)$$

for some matrices $U_{2k-4,2}, V_{2,2k-4}$, and $T_{2,2}$. The matrix in (21) can be transformed into a lower triangle with some elementary matrix operations. More specifically, the block matrix in (21) can be written as

$$\left[\begin{array}{cc} N'_{2k-4} & U_{2k-4,2} \\ V_{2,2k-4} & T_{2,2} \end{array} \right] = \left[\begin{array}{cc} N'_{2k-4} & O_{2k-4,2} \\ V_{2,2k-4} & T'_{2,2} \end{array} \right] \cdot \left[\begin{array}{cc} I & -N'^{-1}_{2k-4} U_{2k-4,2} \\ O & I \end{array} \right].$$

We then obtain $det(N'_{2k-2}) = det(N'_{2k-4}) \cdot det(T'_{2,2})$. The matrix $T'_{2,2}$ is an upper triangle matrix, where $T'_{2,2}[1,1] = h^{-1}_{k-1}$ and

$$\begin{aligned} T'_{2,2}[2,2] &= \frac{h_{k-1} + h_{k-2}}{4} - \frac{h^2_{k-2}}{12} N'^{-1}_{2k-4}[2k-4, 2k-4] \\ &+ (2h^{-1}_{k-2} + 2h^{-1}_{k-1}) \left\{ \frac{h^2_{k-2}}{12} + \frac{h_{k-2}}{12} N'^{-1}_{2k-4}[k-1, 2k-4] \right\}. \qquad (22) \end{aligned}$$

By induction hypothesis, $det(N'_{2n-2}) = \frac{1}{4} det(N'_{2(n-1)-2}) + K_n$ holds for $n = k-1$ and $n = k-2$, and $det(N'_{2(k-3)-2}) > 0$, we get $N'^{-1}_{2k-4}[2k-4, 2k-4] < 4/h_{k-2}$ by Lemma 2, and $N'^{-1}_{2k-4}[k-1, 2k-4] > -h_{k-2}/3$ by Lemma 3. It follows that

$$\begin{aligned} T'_{2,2}[2,2] &> \frac{h_{k-1} + h_{k-2}}{4} - \frac{h^2_{k-2}}{12} \frac{4}{h_{k-2}} + (2h^{-1}_{k-2} + 2h^{-1}_{k-1}) \left\{ \frac{h^2_{k-2}}{12} - \frac{h_{k-2}}{12} \frac{h_{k-2}}{3} \right\} \\ &> \frac{h^2_{k-2}}{18} (2h^{-1}_{k-2} + 2h^{-1}_{k-1}) - \frac{h_{k-2}}{3} + \frac{h_{k-1} + h_{k-2}}{4} \\ &= \frac{h_{k-2}}{36} + \frac{h^2_{k-2}}{9h_{k-1}} + \frac{h_{k-1}}{4} > \frac{h_{k-1}}{4}. \end{aligned}$$

Then

$$det(N'_{2k-2}) = T'_{2,2}[2,2] \frac{1}{h_{k-1}} \, det(N'_{2k-4}) = \left(\frac{h_{k-1}}{4} + \left(T'_{2,2}[2,2] - \frac{h_{k-1}}{4} \right) \right) \frac{1}{h_{k-1}} \, det(N'_{2k-4}).$$

Define $K_k = \left(T'_{2,2}[2,2] - \frac{h_{k-1}}{4}\right)\frac{det(N'_{2k-4})}{h_{k-1}}$. Since $T'_{2,2}[2,2] > \frac{h_{k-1}}{4}$ and $det(N_{2k-4}) > 0$, then

$$det(N'_{2k-2}) = \frac{1}{4}det(N'_{2k-4}) + K_k > 0$$

as required. □

Lemma 2. *If* $det(N'_{2(k-1)-2}) = \frac{1}{4}det(N'_{2(k-2)-2}) + K_{k-1}$ *is satisfied, then*

$$N_{2k-4}^{'-1}[2k-4, 2k-4] < \frac{4}{h_{k-2}} \textit{ for } k \geq 4.$$

Proof. It is known that the elements of the inverse of N_{2k-4} can be expressed in terms of its cofactor matrix.

$$N_{2k-4}^{'-1}[2k-4, 2k-4] = \frac{det(cofactor(N'_{2k-4}[2k-4, 2k-4]))}{det(N'_{2k-4})}.$$

By the definition of N'_{2k-4}, it is easy to see that

$$det(cofactor(N'_{2k-4}[2k-4, 2k-4])) = \frac{1}{h_{k-2}}det(N'_{2(k-2)-2}).$$

Since $det(N'_{2(k-1)-2}) = \frac{1}{4}det(N'_{2(k-2)-2}) + K_{k-1}$ is satisfied by induction hypothesis, we obtain

$$N_{2k-4}^{'-1}[2k-4, 2k-4] = \frac{\frac{1}{h_{k-2}}det(N'_{2(k-2)-2})}{\frac{1}{4}det(N'_{2(k-2)-2}) + K_{k-1}} = \frac{4}{h_{k-2} + Y} < \frac{4}{h_{k-2}}$$

as required, since $Y = \frac{4*K_{k-1}}{\left(h_{k-2}\, det(N'_{2k-4})\right)} > 0$. □

Lemma 3. *If* $det(N'_{2n-2}) = \frac{1}{4}det(N'_{2(n-1)-2}) + K_n$ *is satisfied for* $n = k-1$, $n = k-2$, *and* $det(N'_{2n-2}) > 0$ *for* $n = k-3$, *then*

$$N_{2k-4}^{'-1}[k-1, 2k-4] > -\frac{h_{k-2}}{3} \textit{ for } k \geq 5.$$

Proof. Since $N'N'^{-1} = I$, it is obvious that $N'[k-1, \cdot]N^{'-1}[\cdot, 2k-4] = 0$, where $N'[k-1, \cdot]$ is the $(k-1)$-th row of N' and $N^{'-1}[\cdot, 2k-4]$ is the $(2k-4)$-th column of $N^{'-1}$. More specifically, it can be shown that

$$N_{2k-4}^{'-1}[k-2, 2k-4] + N_{2k-4}^{'-1}[k-1, 2k-4] = -\frac{h_{k-2}^2}{12}N_{2k-4}^{'-1}[2k-4, 2k-4].$$

Consider the cofactor of $N'_{2k-4}[k-2,2k-4]$. By the structure of N'_{2k-4}, we obtain

$$N'^{-1}_{2k-4}[k-2,2k-4] \propto det(cofator(N'_{2k-4}[k-2,2k-4])) = -\frac{h_{k-3}}{12h_{k-2}}\ det(A),$$

where A is a $(2k-6)\times(2k-6)$ matrix and all elements in A are the same as the elements in N'_{2k-6} except $A[2k-6,k-3]$ and $A[2k-6,2k-6]$, where $A[2k-6,k-3] = -2h_{k-2}^{-1} - 3h_{k-3}^{-1} < N'_{2k-6}[2k-6,k-3]$ and $A[2k-6,2k-6] = \frac{h_{k-3}}{4} + \frac{h_{k-2}}{6}$. Similarly, using elementary matrix operations, the $det(N'_{2k-6})$ and $det(A)$ can be formed by $N'_{2(k-3)-2}$. We obtain that

$$det(N'_{2k-6}) = \begin{vmatrix} N'_{2k-8} & O_{2k-8,2} \\ V'_{2,2k-8} & X_{2,2} \end{vmatrix} \text{ and } det(A) = \begin{vmatrix} N'_{2k-8} & O_{2k-8,2} \\ V'_{2,2k-8} & Y_{2,2} \end{vmatrix},$$

where both of X and Y are upper triangle matrices with $X[1,1] = Y[1,1] = h_{k-3}^{-1}$, and

$$\begin{aligned} X[2,2] &= \frac{h_{k-2}}{4} + \frac{h_{k-3}}{4} - c - N'_{2k-6}[2k-6,k-3]d, \\ Y[2,2] &= \frac{h_{k-2}}{6} + \frac{h_{k-3}}{4} - c - A[2k-6,k-3]d, \end{aligned}$$

where $c = \frac{h_{k-3}^2}{12} N'^{-1}_{2k-8}[2k-8,2k-8]$ and $d = \frac{h_{k-3}^2}{12} + \frac{h_{k-3}}{12} N'^{-1}_{2k-8}[k-4,2k-8]$. Since $det(N'_{2k-6}) = det(N'_{2k-8}) \cdot X[1,1] \cdot X[2,2] = \frac{1}{4} det(N'_{2k-8}) + K_{k-2} > 0$ is satisfied, we find that

$$K_{k-2} = h_{k-3}^{-1}\left(\frac{h_{k-3}}{4} - c - N'_{2k-6}[2k-6,k-3]d\right) > 0.$$

It follows that

$$\frac{h_{k-3}}{4} - c - A[2k-6,k-3]d > \frac{h_{k-3}}{4} - c - N'_{2k-6}[2k-6,k-3]d > 0.$$

Then $det(A) = det(N'_{2k-8}) \cdot Y[1,1] \cdot Y[2,2] > 0$ and $N'^{-1}_{2k-4}[k-2,2k-4] < 0$. Since $N'^{-1}_{2k-4}[2k-4,2k-4] < \frac{4}{h_{k-2}}$ is satisfied by Lemma 2, we obtain the the inequality $N'^{-1}_{2k-4}[k-1,2k-4] > -\frac{h_{k-2}}{3}$ as required. □

B Appendix: The Construction of N'_{2n-2} in Proposition 2

When $n = 4$, we find

$$N'_{2*4-2} = \begin{bmatrix} \frac{h_0\left(2+\frac{h_0}{h_1}\right)}{2} & \frac{-{h_0}^2}{2\,h_1} & 0 & 0 & \frac{-\left({h_0}^2\,h_1\right)}{12} & 0 \\ \frac{1}{h_1} & \frac{1}{h_1} & 0 & 0 & \frac{h_1}{12} & 0 \\ 0 & \frac{1}{h_2} & \frac{1}{h_2} & 0 & \frac{h_2}{12} & \frac{h_2}{12} \\ 0 & 0 & \frac{1}{h_3} & \frac{1}{h_3} & 0 & \frac{h_3}{12} \\ 0 & \frac{-2}{h_1}-\frac{2}{h_2} & 0 & 0 & \frac{h_1+h_2}{4} & \frac{h_2}{12} \\ 0 & 0 & \frac{-2}{h_2}-\frac{2}{h_3} & 0 & \frac{h_2}{12} & \frac{h_2+h_3}{4} \end{bmatrix}.$$

If $n = 5$, then

$$N'_{2*5-2} = \begin{bmatrix} \frac{h_0\left(2+\frac{h_0}{h_1}\right)}{2} & \frac{-{h_0}^2}{2\,h_1} & 0 & 0 & 0 & \frac{-\left({h_0}^2\,h_1\right)}{12} & 0 & 0 \\ \frac{1}{h_1} & \frac{1}{h_1} & 0 & 0 & 0 & \frac{h_1}{12} & 0 & 0 \\ 0 & \frac{1}{h_2} & \frac{1}{h_2} & 0 & 0 & \frac{h_2}{12} & \frac{h_2}{12} & 0 \\ 0 & 0 & \frac{1}{h_3} & \frac{1}{h_3} & 0 & 0 & \frac{h_3}{12} & \frac{h_3}{12} \\ 0 & 0 & 0 & \frac{1}{h_4} & \frac{1}{h_4} & 0 & 0 & \frac{h_4}{12} \\ 0 & \frac{-2}{h_1}-\frac{2}{h_2} & 0 & 0 & 0 & \frac{h_1+h_2}{4} & \frac{h_2}{12} & 0 \\ 0 & 0 & \frac{-2}{h_2}-\frac{2}{h_3} & 0 & 0 & \frac{h_2}{12} & \frac{h_2+h_3}{4} & \frac{h_3}{12} \\ 0 & 0 & 0 & \frac{-2}{h_3}-\frac{2}{h_4} & 0 & 0 & \frac{h_3}{12} & \frac{h_3+h_4}{4} \end{bmatrix}.$$

By some permutation operations, $det(N'_{2*5-2})$ can be expressed in terms of $det(N'_{2*4-2})$. That is,

$$\begin{aligned} det(N'_{2*5-2}) &= \begin{vmatrix} \frac{h_0\left(2+\frac{h_0}{h_1}\right)}{2} & \frac{-{h_0}^2}{2\,h_1} & 0 & 0 & \frac{-\left({h_0}^2\,h_1\right)}{12} & 0 & 0 & 0 \\ \frac{1}{h_1} & \frac{1}{h_1} & 0 & 0 & \frac{h_1}{12} & 0 & 0 & 0 \\ 0 & \frac{1}{h_2} & \frac{1}{h_2} & 0 & \frac{h_2}{12} & \frac{h_2}{12} & 0 & 0 \\ 0 & 0 & \frac{1}{h_3} & \frac{1}{h_3} & 0 & \frac{h_3}{12} & 0 & \frac{h_3}{12} \\ 0 & \frac{-2}{h_1}-\frac{2}{h_2} & 0 & 0 & \frac{h_1+h_2}{4} & \frac{h_2}{12} & 0 & 0 \\ 0 & 0 & \frac{-2}{h_2}-\frac{2}{h_3} & 0 & \frac{h_2}{12} & \frac{h_2+h_3}{4} & 0 & \frac{h_3}{12} \\ 0 & 0 & 0 & \frac{1}{h_4} & 0 & 0 & \frac{1}{h_4} & \frac{h_4}{12} \\ 0 & 0 & 0 & \frac{-2}{h_3}-\frac{2}{h_4} & 0 & \frac{h_3}{12} & 0 & \frac{h_3+h_4}{4} \end{vmatrix} \\ &= \begin{vmatrix} N'_{2*4-2} & U_{6,2} \\ V_{2,6} & T_{2,2} \end{vmatrix}. \end{aligned}$$

References

[1] Burden, R. L. and Faires, J. D. (2001). *Numerical Analysis,* Brooks/Cole, Australia.

[2] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerical Mathematik,* **31**, 377-403.

[3] de Boor, C. (2001). *A Practial Guide to Splines: with 32 figures,* Rev. ed., Springer, New York.

[4] Eubank, R. L.(1990). *Nonparametric Regression and Spline Smoothing,* 2nd ed., Marcel Dekker, New York.

[5] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models,* Chapman and Hall, London.

[6] Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach,* Chapman and Hall, London.

[7] Gu, C. (2002). *Smoothing Spline ANOVA Models,* Springer, New York.

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

[9] Rossi, N., Wang, X., and Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm, *Journal of Educational and Behavioral Statistics,* **27**, 291-317.

[10] Wahba, G. (1991). *Spline Models for Observational Data,* Philadelphia, Pennsylvania/SIAM.

[11] Wang, Z. (2000). An algorithm for generalized monotonic smoothing, *Journal of Applied Statistics,* **27**, 495-507.

[12] Zhang, J. T. (2004). A simple and efficient monotone smoother using smoothing splines, *Nonparametric Statistics,* **16(5)**, 779-796.

Table 1: The proportion of Monotony in 10000 repeats

		$n = 50$	$n = 100$	$n = 200$
$\hat{p}_g$	$\alpha = 1.98$ $\beta = 28$	77.89%	91.38%	96.2%
	$\alpha = 6.28$ $\beta = 17.67$	79.43%	93.46%	97.27%
	a=6.9 $\beta = 1.1$	85.54%	93.36%	94.67%

Table 2: Proportion of $\frac{ASE(\hat{p}_g)}{ASE(\hat{p}_m)} \geq 1$

	$n = 50$	$n = 100$	$n = 200$
$\alpha = 1.98$ $\beta = 28$	39.91%	43.76%	48.34%
$\alpha = 6.28$ $\beta = 17.67$	49.27%	52.5%	56.12%
$\alpha = 6.9$ $\beta = 1.1$	56.3%	61.45%	66.47%

Table 3: The distribution summary for $ASE(\hat{p}_g)$ and $ASE(\hat{p}_m)$, where Q1 is the first quartile , Q3 is the third quartile. The unit is 10^{-3}.

			Q1	Median	Mean	Q3
$\alpha = 1.98$ $\beta = 28$	$n = 50$	$\hat{p}_g$	2.3	4.9	7.0	9.3
		$\hat{p}_m$	2.3	4.9	6.8	9.1
	$n = 100$	$\hat{p}_g$	1.6	2.8	4	5.1
		$\hat{p}_m$	1.5	2.7	3.7	4.9
	$n = 200$	$\hat{p}_g$	1.0	1.7	2.2	2.8
		$\hat{p}_m$	0.83	1.5	2.0	2.6
$\alpha = 6.28$ $\beta = 17.67$	$n = 50$	$\hat{p}_g$	1.2	2.9	4.6	6.1
		$\hat{p}_m$	1.2	2.9	4.6	6.1
	$n = 100$	$\hat{p}_g$	0.71	1.7	2.7	3.6
		$\hat{p}_m$	0.7	1.6	2.5	3.4
	$n = 200$	$\hat{p}_g$	0.52	1.08	1.62	2.14
		$\hat{p}_m$	0.44	0.95	1.39	1.87
$\alpha = 6.9$ $\beta = 1.1$	$n = 50$	$\hat{p}_g$	1.4	3.1	4.8	6.3
		$\hat{p}_m$	1.8	3.2	5.6	6.9
	$n = 100$	$\hat{p}_g$	0.91	2.0	2.9	3.8
		$\hat{p}_m$	1.1	2.2	3.0	4.0
	$n = 200$	$\hat{p}_g$	0.54	1.1	1.6	2.1
		$\hat{p}_m$	0.59	1.2	1.6	1.2

Table 4: For sample size $n = 100$ and 10000 monotone $\hat{p}_g$, the Distribution summary of $ASE(\hat{p}_g)$ and $ASE(\hat{p}_m)$, where the Q1 is the first quartile , Q3 is the third quartile. The unit is 10^{-3}.

		Q1	Median	Mean	Q3
$\alpha = 1.98$	$\hat{p}_g$	0.7	1.6	2.5	3.3
$\beta = 28$	$\hat{p}_m$	0.8	1.7	2.6	3.5
$\alpha = 6.28$	$\hat{p}_g$	0.8	1.8	2.8	3.7
$\beta = 17.67$	$\hat{p}_m$	0.7	1.7	2.6	3.5
$\alpha = 6.9$	$\hat{p}_g$	1.3	2.5	3.4	4.4
$\beta = 1.1$	$\hat{p}_m$	1.1	2.2	3.1	4.1

Table 5: For sample size $n = 100$ and 10000 non-monotone $\hat{p}_g$, the distribution summary of $ASE(\hat{p}_g)$ and $ASE(\hat{p}_m)$, where the Q1 is the first quartile , Q3 is the third quartile. The unit is 10^{-3}.

		Q1	Median	Mean	Q3
$\alpha = 1.98$	$\hat{p}_g$	5.4	8.1	9.2	11.5
$\beta = 28$	$\hat{p}_m$	0.8	1.8	2.7	3.6
$\alpha = 6.28$	$\hat{p}_g$	5.0	7.4	8.4	10.5
$\beta = 17.67$	$\hat{p}_m$	0.8	1.8	2.8	3.7
$\alpha = 6.9$	$\hat{p}_g$	4.1	6.4	7.5	9.7
b=1.1	$\hat{p}_m$	1.1	2.3	3.3	4.3

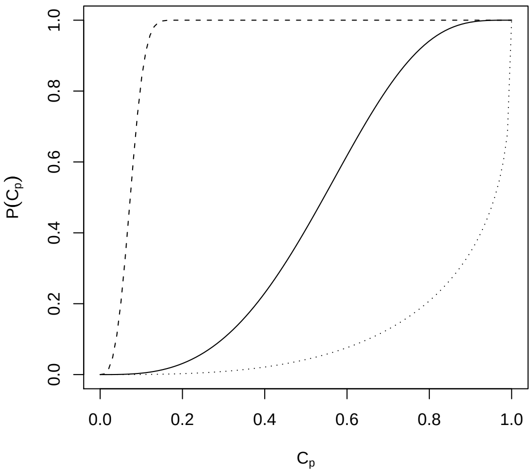
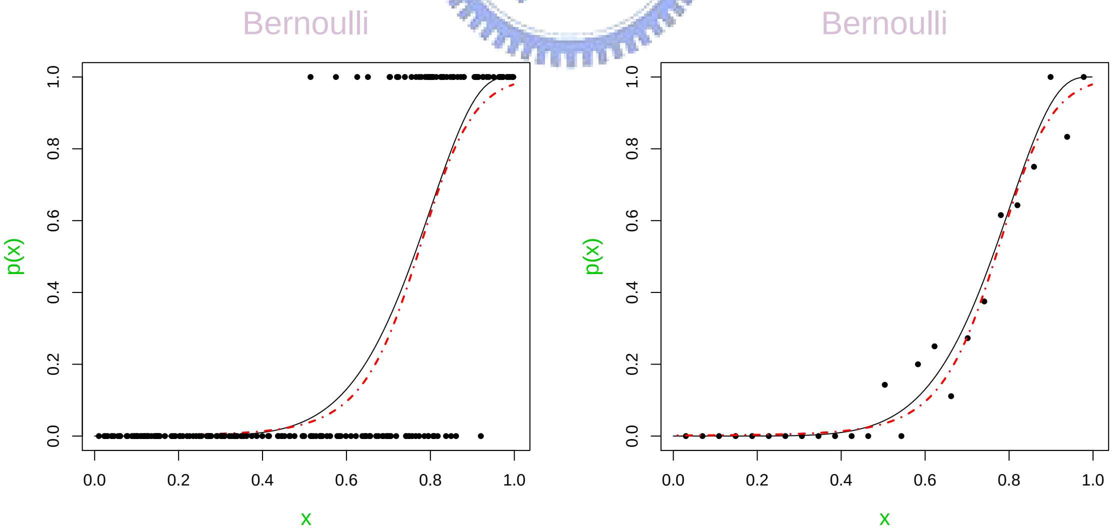


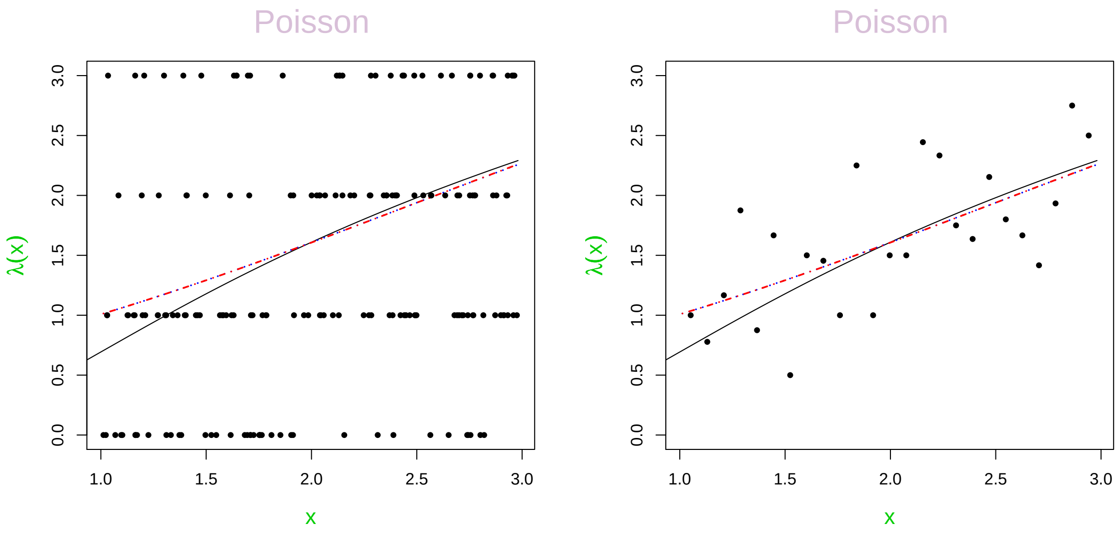
Figure 1: Examples of three EC performance curves.



(a)

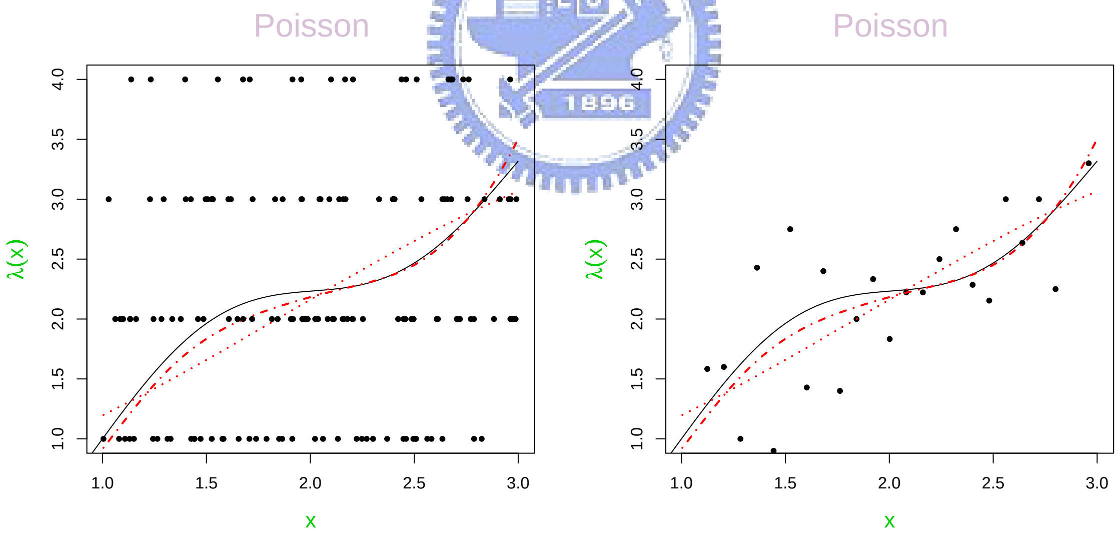
(b)

Figure 2: Bernoulli data. The solid line is the target function $p(x) = 1 - (1 - x^{4.5})^{2.5}$. The dotted line is the estimated curve with monotone constraint. Dots in the left panel represent samples (x_i, y_i), $i = 1, \ldots, 200$. Dots in the right panel are binned data.



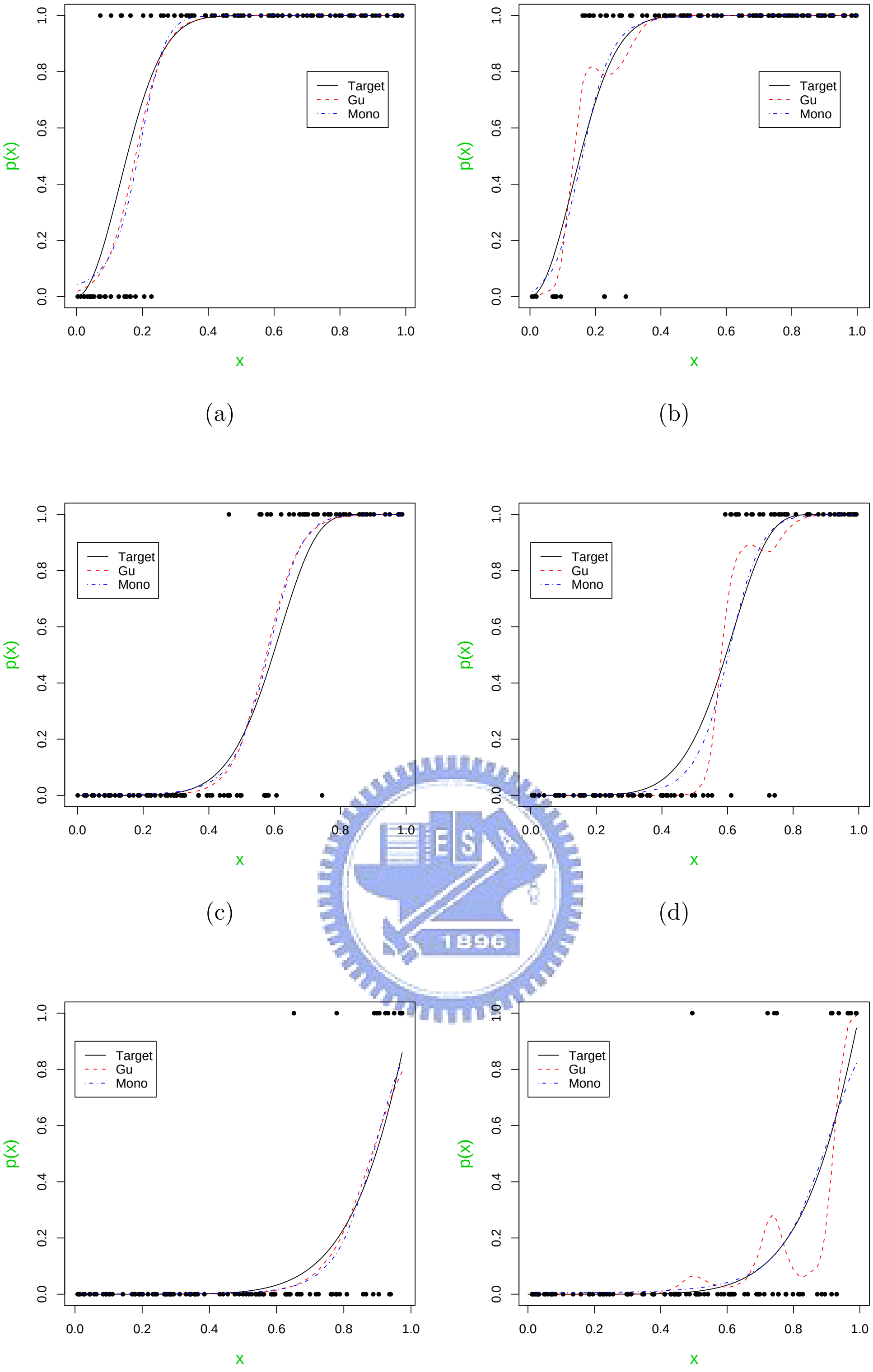
(a) (b)

Figure 3: Poisson data. The solid line is the target function $\phi(x) = \log(x^2+1)$. Dotted line is the estimated curve by GCV method. Dash-dot line is the estimated curve with smoothing parameter 1.5. Dots in the left panel are points (x_i, y_i), $i = 1, \ldots, 200$. Dots in the right panel are binned data.



(a) (b)

Figure 4: Poisson data. The solid line is the target function $\phi(x) = \frac{1}{e} - e^{-x} + x - \frac{\sin(\pi\, x)}{\pi}$. The dotted line is the estimated curve by GCV method. Dash-dot line is the estimated curve with smoothing parameter 2.5×10^{-5}. Dots in the left panel are points (x_i, y_i), $i = 1, \ldots, 200$. Dots in the right panel are binned data.



(a)

(b)

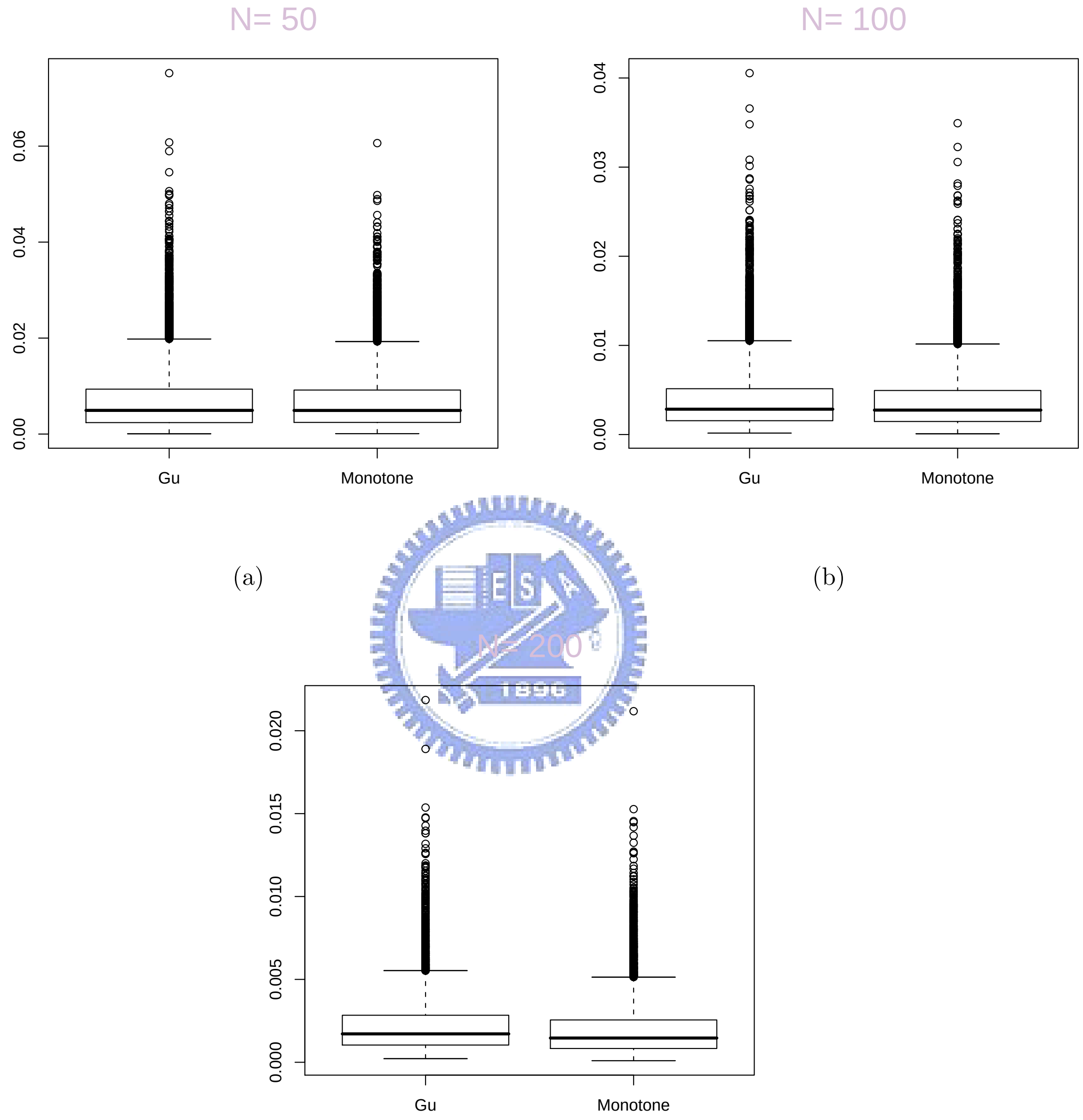
(c)

(d)

(e)

(f)

Figure 5: Illustrative examples. The solid line is the target function $p(x) = 1 - (1 - x^{\alpha})^{\beta}$. For panels from top to bottom, (α, β) are (1.98,28), (6.28,17.67), and (6.9,1.1), respectively. The dashed line is the estimated curve $\hat{p}_g$ by Gu(2002). The dash-dot line is the estimated curve $\hat{p}_m$ with monotone constraint. The dots are data points. The left three panels shows the cases with monotone $\hat{p}_g$, while $\hat{p}_g$ in the right panels are not monotonic. The sample size $n = 100$.

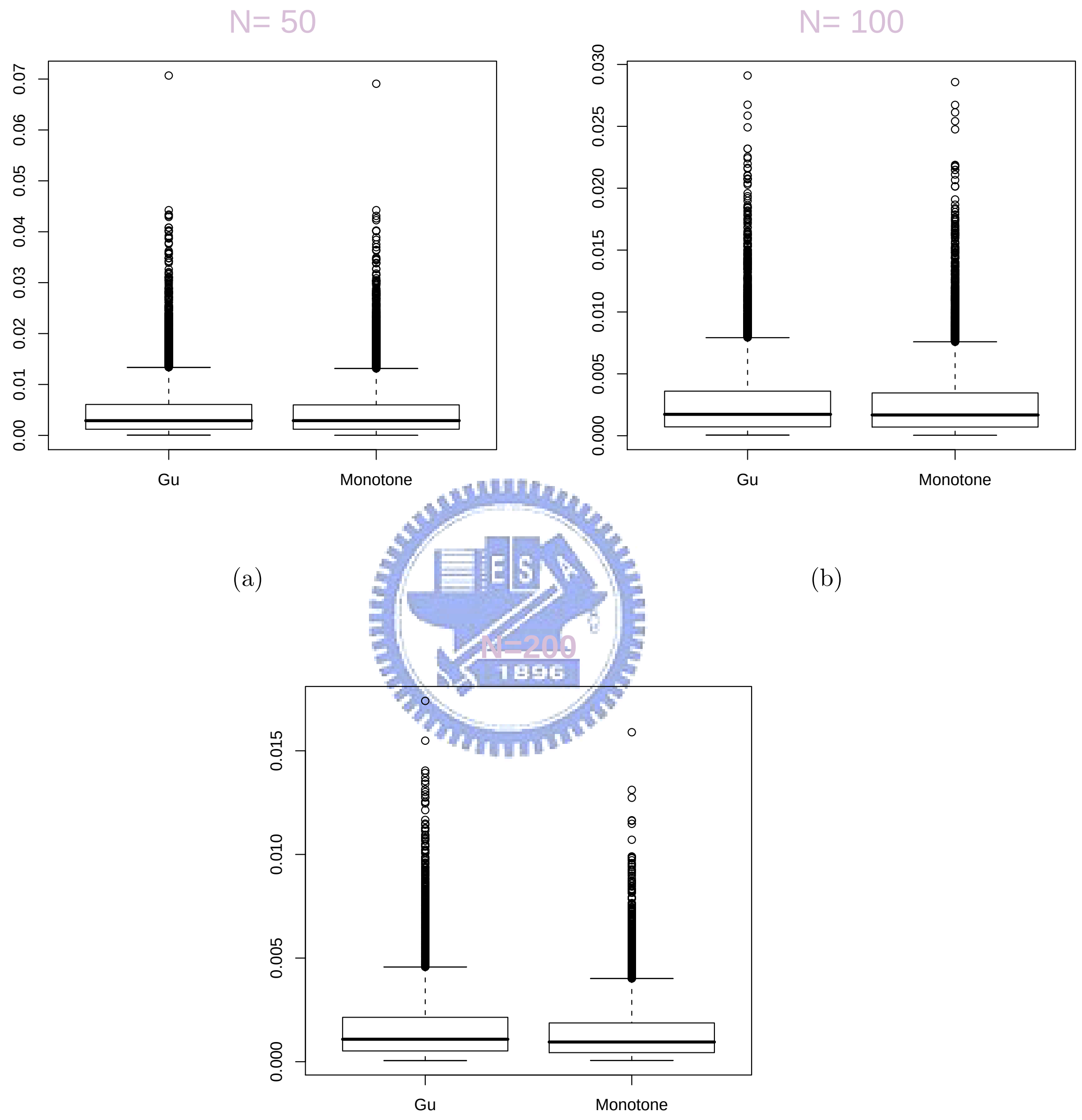


(a)

(b)

(c)

Figure 6: Box-plots of 10000 $ASE(\hat{p}_g)$ and 10000 $ASE(\hat{p}_m)$. $p(x) = 1 - (1 - x^{1.98})^{28}$. Panels (a), (b), and (c) are for sample size n=50, 100, and 200, respectively. For each panel, "Gu" indicates $\hat{p}_g$, and "monotone" indicates $\hat{p}_m$.

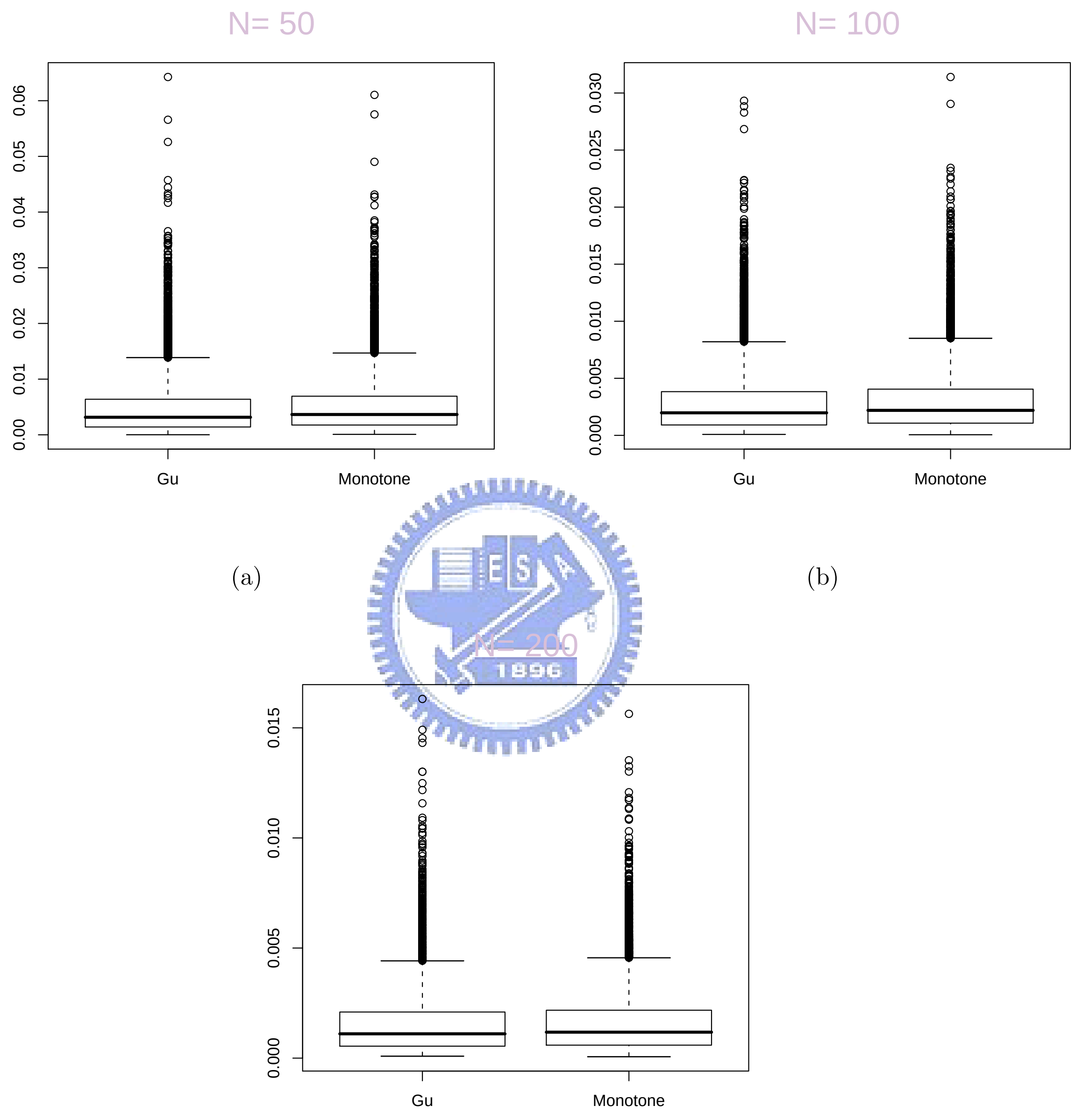


(a)

(b)

(c)

Figure 7: Box-plots of 10000 $ASE(\hat{p}_g)$ and 10000 $ASE(\hat{p}_m)$. $p(x) = 1-(1-x^{6.28})^{17.67}$. "Gu" is $\hat{p_g}$, and "monotone" is $\hat{p_m}$.

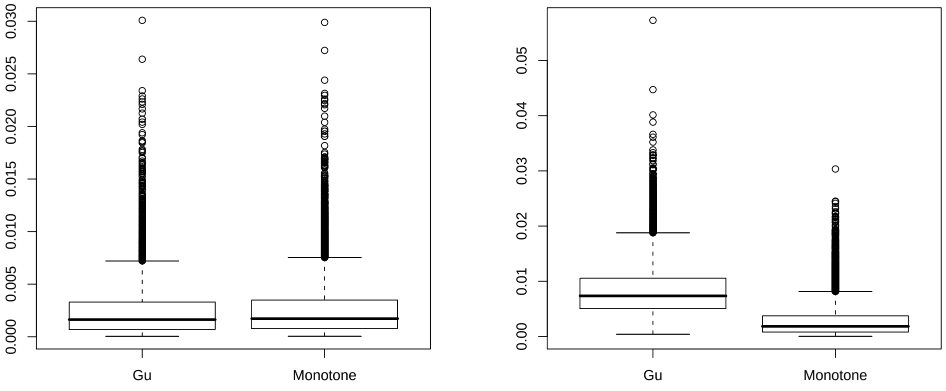


(a)

(b)

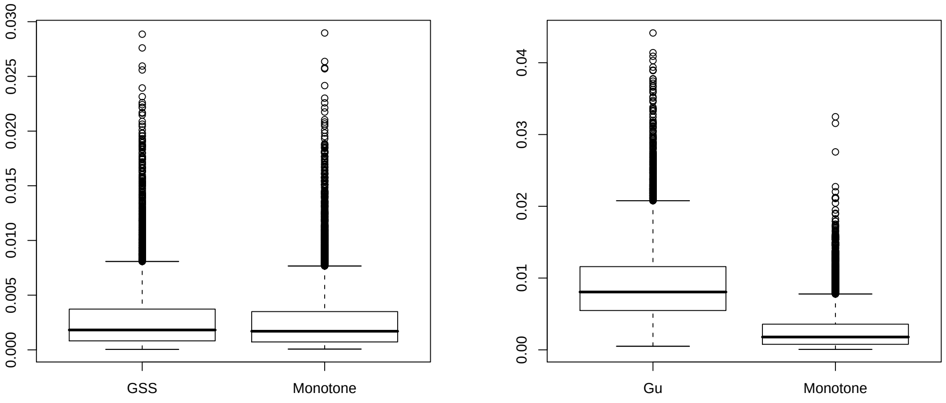
(c)

Figure 8: Box-plots of 10000 $ASE(\hat{p}_g)$ and 10000 $ASE(\hat{p}_m)$. $p(x) = 1 - (1 - x^{6.9})^{1.1}$. "Gu" is $\hat{p_g}$, and "monotone" is $\hat{p_m}$.



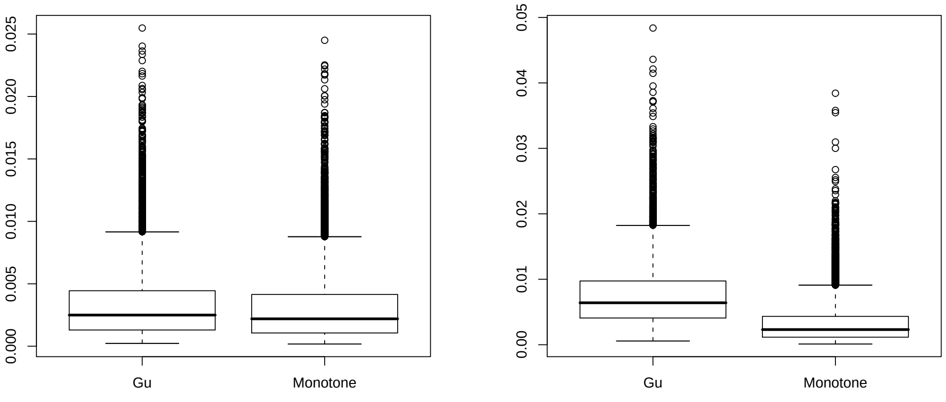
(a) (b)

Figure 9: Box-plots of 10000 $ASE(\hat{p_g})$ and 10000 $ASE(\hat{p_m})$. $p(x) = 1-(1-x^{1.98})^{28}$ and $n = 100$. ”Gu” is $\hat{p}_g$, and ”Monotone” is $\hat{p}_m$. Panel (a) is for 10000 monotone $\hat{p}'_g s$ while panel (b) is for 10000 non-monotone $\hat{p}'_g s$ shows in panel (b)



(a) (b)

Figure 10: Box-plots of 10000 $ASE(\hat{p_g})$ and 10000 $ASE(\hat{p_m})$. $p(x) = 1-(1-x^{6.28})^{17.67}$. “Gu” is $\hat{p}_g$, and“Monotone” is $\hat{p}_m$. Panel (a) is for 10000 monotone $\hat{p}'_g s$ while panel (b) is for 10000 non-monotone $\hat{p}'_g s$ shows in panel (b)



(a)

(b)

Figure 11: Box-plots of 10000 $ASE(\hat{p_g})$ and 10000 $ASE(\hat{p_m})$. $p(x) = 1 - (1 - x^{6.9})^{1.1}$. “Gu” is $\hat{p}_g$, and “Monotone” is $\hat{p}_m$. Panel (a) is for 10000 monotone $\hat{p}_g's$ while panel (b) is for 10000 non-monotone $\hat{p}_g's$ shows in panel (b)

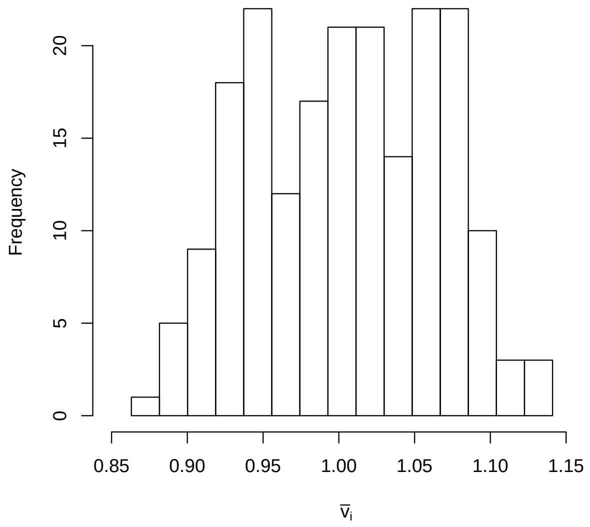
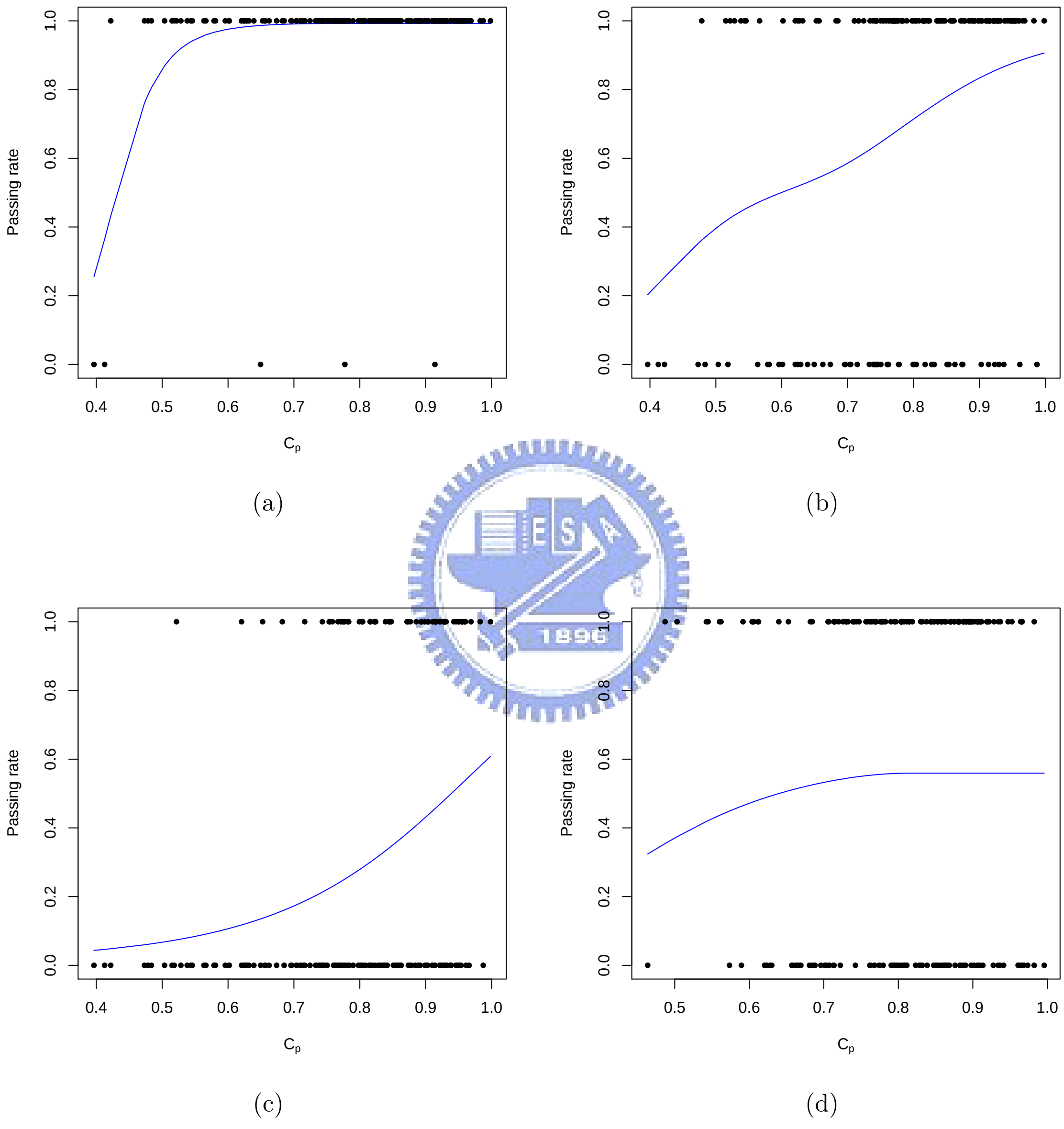


Figure 12: Histogram of 200 mean voltages . The sample mean is 1.005 and standard deviation is 0.06.



(a)

(b)

(c)

(d)

Figure 13: Estimated passing rates(i.e., EC performance curves) under three control limits. (LCL,UCL) of mean voltage for panels (a), (b), and (c) are $(0.88, 1.12)$, $(0.93, 1.07)$, $(0.97, 1.03)$, respectively. Panel (d) is an EC performance curve when mean voltage is independent of the C_p yield.