The Institute of Statistical Mathematics
Research Memorandum No. 898

# A New Computational Approach to Density Estimation with Semidefinite Programming

Tadayoshi Fushiki * Shingo Horiuchi† Takashi Tsuchiya‡

November, 2003

**Abstract**

Density estimation is a classical and important problem in statistics. The aim of this paper is to develop a new computational approach to density estimation based on semidefinite programming (SDP), a new technology developed in optimization in the last decade. We express a density as the product of a nonnegative polynomial and a base density such as normal distribution, exponential distribution and uniform distribution. The difficult nonnegativity constraint imposed on the polynomial is expressed as a semidefinite constraint. Under the condition that the base density is specified, the maximum likelihood estimation of the coefficients of the polynomial is formulated as a variant of SDP which can be solved in polynomial-time with the recently developed interior-point methods. Since the base density typically contains just one or two parameters, if the likelihood function is easily maximized with respect to the polynomial part by SDP, then it is possible to compute the global maximum of the likelihood function by further maximizing the partially-maximized likelihood function with respect to the base density parameter. The primal-dual interior-point algorithms are used to solve the variant of SDP. The proposed model is flexible enough to express such properties as unimodality and symmetry which would be reasonably imposed on the density function. AIC (Akaike information criterion) is used to choose the best model. Through applications to several instances we demonstrate flexibility of the model and performance of the proposed procedure.

*The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, JAPAN. e-mail: fushiki@ism.ac.jp

†Access Network Service Systems Laboratories, NTT Corp., Makuhari, Chiba, 261-0023 Japan. e-mail: hor@ansl.ntt.co.jp

‡The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, JAPAN. e-mail: tsuchiya@sun312.ism.ac.jp . This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 15510144, 2003.

# 1 Introduction

Density estimation [11, 33, 34] is a classical and important problem in statistics. The aim of this paper is to develop a new computational parametric approach to density estimation. Roughly, there are two approaches to this problem; the nonparametric density estimation and the parametric density estimation. Kernel methods [34], maximum penalized likelihood methods [11, 13, 14, 40] and Bayesian methods [16, 36, 37, 38, 39] are among well-known nonparametric approaches. While the models behind these methods are flexible enough to express various types of distributions, they have several serious drawbacks in statistical inferences. Sometimes the models are too flexible and therefore we need some regularization or Bayesian framework to obtain a meaningful result. However, it is not easy to provide a simple and computationally tractable procedure to determine parameters such as bandwidth or smoothness weight. Local likelihood methods [9] and mixture distribution methods [18] are intermediate approaches between parametric and nonparametric approaches. While these approaches sometimes work well, they also have the same drawbacks as the nonparametric approaches. On the other hand, the parametric approach is not as popular as the methods mentioned above. This is partly because the approach fails to provide a flexible model due to the nonnegativity constraint imposed on the model function. An advantage of the parametric approach is that well-established tools for inference based on parametric statistical models can be directly applied. Another advantage is that the parametric density function can be utilized as a part in assembling more general statistical models such as regression models, time series models and point process models. We mention GALTHY [3, 4, 32] as an example of the parametric approach. In [6], parametric and nonparametric density estimation are studied from the viewpoint of MDL principle.

The common difficulty of these approaches is optimization of parameters. In many cases, determination of parameters in density estimation is formulated as a nonlinear constrained nonconvex, and sometimes even stochastic program. Therefore, it is difficult to develop a stable and robust method to solve it. Usually the obtained result is a local optimal solution and not the global one.

In this paper, we develop a new computational parametric approach to density estimation to overcome some of these difficulties. We restrict ourselves to one dimensional density estimation, though the idea can be extended to multi-dimensional case in a reasonable way. We model a density function as the product of a univariate nonnegative polynomial and a base density such as normal distribution, exponential distribution and uniform distribution. Thus, the support of the density function can be any of $(-\infty, \infty)$, $[0, \infty)$ and $[a, b]$, where $a, b \in \mathbf{R}$. The novelty of our approach lies in the use of semidefinite programming (SDP) [5, 7, 25, 43, 46], a new technology developed in the last decade in the field of optimization. SDP is the problem of minimizing a linear objective function over the intersection of an affine space and the space of symmetric positive semidefinite matrices. SDP is an extension of the classical linear programming, and has wide applications in control, combinatorial optimization, signal processing, communication systems design, optimal shape design etc. [7, 43, 46]. SDP can be solved efficiently in both theory and practice with the interior-point methods [5, 15, 17, 25, 26, 27]. We will show that the maximum likelihood estimate of our model can be computed efficiently and rigorously with the techniques of SDP.

A necessary and sufficient condition for a univariate polynomial to be nonnegative over the above-mentioned support sets can be expressed as semidefinite constraints. With this formula-

tion, when the base density is specified, the maximum likelihood estimation can be formulated as a variant of SDP, namely, maximizing the log determinant function under semidefinite constraint. The formulated optimization problem is not exactly a semidefinite program, but due to its special form, we can solve it also as efficiently as an ordinary semidefinite program with the interior-point method. Specifically, we use the primal-dual interior-point method to solve the variant of SDP [15, 17, 21, 26, 27, 41]. Akaike information criterion (AIC) [1, 2] is used to choose the best model. We demonstrate that the model is flexible enough and MAIC (minimum AIC) procedure gives reasonable estimates of the densities.

This paper is organized as follows. In Section 2, we explain our model and formulate the maximum likelihood estimation as a (variant of) semidefinite program. In Section 3, we briefly explain SDP and introduce interior-point methods to solve this problem. In Section 4, the performance of our method is demonstrated through application to several instances. In Section 5, we discuss possible applications of this approach to other areas of statistics taking up as an example the estimation of a nonstationary Poisson process. Section 6 is a concluding discussion.

# 2 Problem Formulation

Let $\{x_1, ..., x_N\}$ be data independently drawn from an unknown density $g(x)$ over the support $S \subseteq \mathbf{R}$. Our problem is to estimate $g(x)$ based on $\{x_1, ..., x_N\}$. If some prior information on $g(x)$ is available, we can use an appropriate statistical model to estimate $g(x)$. A flexible model is necessary to estimate $g(x)$ when we do not have enough information.

In this paper, we develop a computational approach to estimate $g(x)$ based on the following statistical model

$$f(x; \alpha, \beta) = p(x; \alpha)K(x; \beta), \tag{1}$$

where $p(x; \alpha)$ is a univariate polynomial with parameter $\alpha$ and $K(x; \beta)$ is a density function which is specified with parameter $\beta$ over the support $S$. The function $K(x; \beta)$ is referred to as a base density (function). We call $\alpha$ and $\beta$ a polynomial parameter and a base density parameter, respectively. The polynomial $p(x; \alpha)$ is nonnegative over $S$. In the following, we consider the models where the base density is normal distribution, exponential distribution and uniform distribution. These models will be referred to as "normal based model", "exponential based model" and "pure polynomial model", respectively.

Now, we associate a univariate polynomial with a matrix in the following way. Given $x \in \mathbf{R}$, we define $\boldsymbol{x}_d = (1, x, x^2, ..., x^{d-1}) \in \mathbf{R}^d$. In the following, we drop the subscript $d$ of $\boldsymbol{x}_d$ when there is no ambiguity. A polynomial of even degree $n(= 2d - 2)$ can be written as $\boldsymbol{x}^T Q \boldsymbol{x}$ by choosing an appropriate $Q \in \mathbf{R}^{d \times d}$. If a polynomial $q(x) = \sum_{i=0}^{n} q_i x^i$ is represented as $q(x) = \boldsymbol{x}^T Q \boldsymbol{x}$ with $Q \in \mathbf{R}^{d \times d}$, then we can recover the coefficient $q_k$ by

$$q_k = \mathrm{Tr}(E_k Q),$$

where $(i, j)$ element of $E_k$ is

$$(E_k)_{ij} = \begin{cases} 1 & \text{if } i + j - 2 = k \\ 0 & \text{otherwise} \end{cases}, \quad k = 1, ..., n.$$

Let $q'(x) = \sum_{i=0}^{n-1} q'_i x^i$ be the derivative of $q(x)$. The coefficient $q'_k$ is represented as

$$q'_k = (k+1)\mathrm{Tr}(E_{k+1}Q), \quad k = 1, ..., n-1.$$

The main theorem used in this paper together with SDP is the following.

**Theorem 2.1 ([24])** *Let $p(x)$ be a univariate polynomial of degree $n$. Then,*

(i) *$p(x) \geq 0$ over $S = (-\infty, \infty)$ iff $p(x) = \boldsymbol{x}_{(n/2+1)}^T Q \boldsymbol{x}_{(n/2+1)}$ holds for a symmetric positive semidefinite matrix $Q \in \mathbf{R}^{(n/2+1)\times(n/2+1)}$.*

(ii) *$p(x) \geq 0$ over $S = [a, \infty)$ iff*

(a)

$$p(x) = \boldsymbol{x}_{(n+1)/2}^T Q_1 \boldsymbol{x}_{(n+1)/2} + (x-a)\boldsymbol{x}_{(n-1)/2}^T Q_2 \boldsymbol{x}_{(n-1)/2}$$

*for symmetric positive semidefinite matrices*

$$Q_1 \in \mathbf{R}^{((n+1)/2)\times((n+1)/2)} \text{ and } Q_2 \in \mathbf{R}^{((n-1)/2)\times((n-1)/2)} \text{ (The case when } n \text{ is odd),}$$

(b)

$$p(x) = \boldsymbol{x}_{(n/2+1)}^T Q_1 \boldsymbol{x}_{(n/2+1)} + (x-a)\boldsymbol{x}_{n/2}^T Q_2 \boldsymbol{x}_{n/2}$$

*for symmetric positive semidefinite matrices*

$$Q_1 \in \mathbf{R}^{(n/2+1)\times(n/2+1)} \text{ and } Q_2 \in \mathbf{R}^{(n/2)\times(n/2)} \text{ (The case when } n \text{ is even).}$$

(iii) *$p(x) \geq 0$ over $S = [a, b]$ iff*

(a)

$$p(x) = (x-a)\boldsymbol{x}_{(n+1)/2}^T Q_1 \boldsymbol{x}_{(n+1)/2} + (b-x)\boldsymbol{x}_{(n+1)/2}^T Q_2 \boldsymbol{x}_{(n+1)/2} \tag{2}$$

*for symmetric positive semidefinite matrices*

$$Q_1,\ Q_2 \in \mathbf{R}^{((n+1)/2)\times((n+1)/2)} \text{ (The case when } n \text{ is odd),}$$

(b)

$$p(x) = \boldsymbol{x}_{(n/2+1)}^T Q_1 \boldsymbol{x}_{(n/2+1)} + (b-x)(x-a)\boldsymbol{x}_{n/2}^T Q_2 \boldsymbol{x}_{n/2} \tag{3}$$

*for symmetric positive semidefinite matrices*

$$Q_1 \in \mathbf{R}^{(n/2+1)\times(n/2+1)} \text{ and } Q_1 \in \mathbf{R}^{(n/2)\times(n/2)} \text{ (The case when } n \text{ is even).}$$

To explain our approach, we focus on the normal based model where $S = (-\infty, \infty)$ and $K(x;\beta)$ is a normal distribution with parameter $(\mu, \sigma)$. We assume that the parameter $\beta = (\mu, \sigma)$ is given. Under this condition, as will be explained below, the maximum likelihood estimation is formulated as a tractable convex program which can be solved easily with the technique of SDP. The readers will readily see that the exponential based model and the pure

polynomial model can be treated exactly in the same manner. In the following, $C \succeq (\succ) 0$ means that a matrix $C$ is symmetric positive semidefinite (definite).

We represent $p(x;\alpha)$ as $\boldsymbol{x}^T Q \boldsymbol{x}$ with some $Q \succeq 0$. The condition for $f(x;\alpha,\beta)$ to be a density function is written as

$$\int \boldsymbol{x}^T Q \boldsymbol{x} K(x;\beta) \mathrm{d}x = 1, \qquad Q \succeq 0.$$

It is easy to see that this condition is written as the following linear equality constraint with a semidefinite constraint

$$\mathrm{Tr}(M(\beta)Q) = 1, \quad Q \succeq 0,$$

where

$$M(\beta) = \int \boldsymbol{x}\boldsymbol{x}^T K(x;\beta) \mathrm{d}x.$$

Note that $M(\beta)$ is a matrix which can be obtained in closed form as a function of $\beta$ when $K$ is normal, exponential or uniform distribution. On the other hand, the log likelihood of the model (1) becomes

$$\begin{aligned}
\sum_{i=1}^N \log f(x_i;\alpha,\beta) &= \sum_{i=1}^N \{\log p(x_i;\alpha) + \log K(x_i;\beta)\} \\
&= \sum_{i=1}^N \log(\boldsymbol{x}^{(i)T} Q \boldsymbol{x}^{(i)}) + \sum_{i=1}^N \log K(x_i;\beta) \\
&= \sum_{i=1}^N \log \mathrm{Tr}(\boldsymbol{x}^{(i)} \boldsymbol{x}^{(i)T} Q) + \sum_{i=1}^N \log K(x_i;\beta),
\end{aligned}$$

where $\boldsymbol{x}^{(i)} = (1, x_i, x_i^2, ..., x_i^{d-1})^T$. Note that the term in log is linear in $Q$.

Therefore, the maximum likelihood estimation is formulated as follows:

$$\begin{array}{ll}
\max\limits_{Q,\beta} & \sum\limits_{i=1}^N \log\ \mathrm{Tr}(X^{(i)} Q) + \sum\limits_{i=1}^N \log K(x_i;\beta) \\
\text{s.t.} & \mathrm{Tr}(M(\beta)Q) = 1, \quad Q \succeq 0,
\end{array} \tag{4}$$

where $X^{(i)} = \boldsymbol{x}^{(i)} \boldsymbol{x}^{(i)T}$ for $i = 1, ..., N$. If we fix $\beta$ and regard $Q$ as the decision variable, this problem is a convex program closely related to SDP and can be solved efficiently both in theory and practice with the interior-point method [7, 25, 43, 46]. Let $g(\beta)$ be the optimal value of (4) when $\beta$ is fixed. Then we maximize $g(\beta)$ to obtain the maximum likelihood estimator. Since $\beta$ is typically with "one or two dimension" (e.g., location and scale parameter), maximization of $g$ can be done easily by grid search and nonlinear programming techniques [12, 28].

In the following, we show that many properties of the density functions such as symmetry, monotonicity etc., can be expressed by adding linear equality constraints and/or semidefinite constraints to the above problem. Recall that we are dealing with the normal based model with the base density parameter $(\mu, \sigma)$. For simplicity, we also assume that $\mu = 0$.

If we consider a symmetric density function with respect to $x = 0$, we add several linear constraints $\mathrm{Tr}(E_i Q) = 0$ for all odd $i$ such that $1 \leq i \leq n$ because all coefficients of the odd degrees in the polynomial $p(x;\alpha)$ must be zero.

If we consider a density which is unimodal with the maximum at $x = \hat{x}$, we do the following. This condition is equivalent to the two conditions that $f(x)$ is monotone increasing in the interval $(-\infty, \hat{x}]$ and is monotone decreasing in the interval $[\hat{x}, \infty)$. The first monotone increasing

condition can be formulated as follows. Since $f'(x;\beta) = \{p'(x) - xp(x)/\sigma\}K(x;\beta)$, nonnegativity of $f'(x;\beta)$ in the interval $(-\infty, \hat{x}]$ is equivalent to the nonnegativity of $\{p'(x) - xp(x)/\sigma\}$ in the interval $(-\infty, \hat{x}]$. In view of the second statement of Theorem 2.1, we introduce symmetric positive semidefinite matrices $Q_1 \in \mathbf{R}^{(n/2+1)\times(n/2+1)}$ and $Q_2 \in \mathbf{R}^{(n/2+1)\times(n/2+1)}$ to represent

$$p'(x) - \frac{x}{\sigma}p(x) = \boldsymbol{x}^T Q_1 \boldsymbol{x} - (x - \hat{x})\boldsymbol{x}^T Q_2 \boldsymbol{x}.$$

Note that the degree of $p(x)$ is always even. The formulation is completed by writing down the conditions to associate $Q$ with $Q_1$ and $Q_2$. This amounts to the following $n$ linear equality constraints

$$(k+1)\mathrm{Tr}(E_{k+1}Q) - \frac{1}{\sigma}\mathrm{Tr}(E_{k-1}Q) = \mathrm{Tr}(E_k Q_1) - \mathrm{Tr}(E_{k-1}Q_2) + \hat{x}\mathrm{Tr}(E_k Q_2), \quad k = 1, ..., n,$$

where $E_l = 0$ for $l = -1$ and $l = n+1$. The other monotone decreasing condition on the interval $[\hat{x}, \infty)$ can be treated in a similar manner.

Thus, the approach is capable of handling various conditions about the density function in a flexible way just by adding semidefinite and linear constraints to (4). From the optimization point of view, the problem to be solved yet remains in the same tractable class. This point will be explained in more detail in the next section.

# 3 Semidefinite Programming and Interior-point Methods

In this section, we introduce SDP and the interior-point method for SDP, and explain how the interior-point method can be used in the maximum likelihood estimation formulated in the previous section. SDP [7, 25, 43, 46] is an extension of LP (linear programming) in the space of matrices, where a linear objective function is optimized over the intersection of an affine space and the cone of symmetric positive semidefinite matrices. SDP is tractable convex programming, and has a number of applications in combinatorial optimization, control theory, signal processing, structure design etc. [7, 43, 46]. A nice property about SDP is duality. As will be shown later, the dual problem of a semidefinite program becomes another semidefinite program, and under mild assumptions they have the same optimal value. The original problem is referred to as the primal problem in relation to the dual problem. The interior-point method solves SDP by generating a sequence in the interior of the feasible region. There are two types of the interior-point methods called the primal interior-point method and the primal-dual interior-point method. The first one generates iterates in the space of the primal problem while the other generates iterates in the both spaces of the primal and dual problems. We adopted the primal-dual method because it is more flexible and numerically stable. We first illustrate basic ideas of the interior-point methods with the primal method since it is more intuitive and easier to understand. Then we move on to introducing the primal-dual method.

**Remark on literatures:** There are many literatures on SDP. The paper [5] and the book [25] are fundamental works which brought many researchers' attention to this topic. We mention [43] as an earlier survey. The book [46] is a handbook of SDP in which various aspects of treated in detail. The book [7] is a recent textbook in which theory, algorithms and applications are

treated in depth from a unified point of view. Those who are interested in SDP software are recommended to see, for example, [35, 42, 47]. Benchmark of several SDP softwares are reported in [19, 20]. Though the original SDP treated in these literatures is somewhat different from the problem we deal with in this paper, they will give the readers some good idea about applicability of the interior-point method to our problem.

## 3.1 Semidefinite Programming and Primal Interior-point Methods

Let $A_{ij}$ ($i = 1, ..., m$ and $j = 1, ..., \bar{n}$) and $C_j$ ($j = 1, ..., \bar{n}$) be real symmetric matrices, where the size of the matrices $A_{ij}$ and $C_j$ are $n_j \times n_j$. A standard form of SDP is the following optimization problem with respect to $n_j \times n_j$ real symmetric matrix $X_j$, $j = 1, ..., \bar{n}$:

$$\begin{aligned} \text{(P)} \quad & \min_X \sum_{j=1}^{\bar{n}} \mathrm{Tr}(C_j X_j), \\ & \text{s.t.} \sum_{j=1}^{\bar{n}} \mathrm{Tr}(A_{ij} X_j) = b_i, \ i = 1, ..., m, \quad X_j \succeq 0, \quad j = 1, ..., \bar{n}. \end{aligned} \tag{5}$$

Here we denote $(X_1, ..., X_{\bar{n}})$ by $X$, and $X \succeq (\succ) 0$ means that each $X_j$ is symmetric positive semidefinite (definite). A feasible solution $X$ is called an interior feasible solution if $X \succ 0$ holds. Since the cone of positive semidefinite matrices is convex, SDP is a convex program. Even though the problem is highly nonlinear, it can be solved efficiently in both theoretical and practical sense with the interior-point method. The interior-point method is a polynomial-time method for SDP, and in reality, it can solve SDP involving matrices whose dimension is several thousands. To date, the interior-point method is the only practical method for SDP.

Now, let $\Omega$ be a subset of $\{1, ..., \bar{n}\}$, and consider the following problem where a convex function $-\sum_{j \in \Omega} \log \det X_j$ is added to the objective function in (5):

$$\begin{aligned} \text{($\widetilde{\text{P}}$)} \quad & \min_X \sum_{j} \mathrm{Tr}(C_j X_j) - \sum_{j \in \Omega} \log \det X_j, \\ & \text{s.t.} \sum_{j} \mathrm{Tr}(A_{ij} X_j) = b_i, \ i = 1, ..., m, \quad X_j \succeq 0, \quad j = 1, ..., \bar{n}, \end{aligned} \tag{6}$$

It is not difficult to see that the maximum likelihood estimation (4) can be cast into this problem as follows:

$$\begin{aligned} \text{(ML)} \quad & \min - \sum_{j=1}^{N} \log \det Y_j, \\ & \text{s.t.} \ Y_j - \mathrm{Tr}(X^{(j)} Q) = 0, \ i = 1, ..., N, \quad Y_j \succeq 0, \quad j = 1, ..., N, \\ & \qquad \mathrm{Tr}(MQ) = 1, \quad Q \succeq 0, \end{aligned}$$

where $Y_j, j = 1, ..., N$ are new variables of "one by one" matrix introduced to convert (4) to the form of (6). Thus, there are $\bar{n} = N + 1$ decision variables $Y_j (j = 1, ..., N)$ and $Q$ in this problem.

At a glance, the problem (6) looks more difficult than (5) because of the additional convex term in the objective function, however, due to its special structure, we can solve (6) as efficiently as (5) just by slightly modifying the interior-point method for (5) without loosing its

advantages [41, 44]. Due to its form of the objective function, the problem (6) has applications in statistics. For example, the maximum log likelihood estimate of the Gaussian graphical model for a given graph is formulated as (6), see, e.g., [29]. In [44], (6) is studied in detail from the viewpoint of applications and the primal interior-point method.

For the time being, we continue explanation of the interior-point method for (5) to illustrate its main idea. Since a main difficulty of SDP comes from its highly nonlinear shape of the feasible region (even though it is convex), it is important to provide a machinery to keep iterates away from the boundary of feasible region in order to develop an efficient iterative method. For this purpose, the interior-point method makes use of the logarithmic barrier function

$$-\sum_{j=1}^{\bar{n}} \log\det X_j.$$

The logarithmic barrier function is a convex function whose value diverges as $X$ approaches the boundary of the feasible region where one of $X_j$ becomes singular. Incorporating with this barrier function, let us consider the following optimization problem with a positive parameter $\nu$:

$$\begin{aligned} (\mathrm{P}_\nu) \quad & \min_X \sum_j \mathrm{Tr}(C_j X_j) - \nu \sum_j \log\det X_j, \\ & \text{s.t.} \sum_j \mathrm{Tr}(A_{ij} X_j) = b_i, \ i = 1, ..., m, \quad X_j \succeq 0, \quad j = 1, ..., \bar{n}, \end{aligned} \tag{7}$$

where $\nu$ is referred to as "barrier parameter." Since the log barrier function is strictly convex, $(\mathrm{P}_\nu)$ has a unique optimal solution. We denote by $\widehat{X}(\nu) = (\widehat{X}_1(\nu), ..., \widehat{X}_{\bar{n}}(\nu))$ the optimal solution of $(\mathrm{P}_\nu)$. By using the method of Lagrange multiplier, we see that $\widehat{X}(\nu)$ is a unique symmetric positive definite matrix $X$ satisfying the following system of nonlinear equations in unknown $X$ and (the Lagrangian multiplier) $y$:

$$\begin{aligned} & \nu X_j^{-1} = C_j - \sum_i A_{ij} y_i, \quad i = 1, ..., \bar{n}, \\ & \sum_j \mathrm{Tr}(A_{ij} X_j) = b_i, \ i = 1, ..., m, \ \ X_j \succeq 0, \ j = 1, ..., \bar{n}. \end{aligned} \tag{8}$$

The set

$$\mathcal{C}_{\mathrm{P}} \equiv \{\widehat{X}(\nu) : 0 < \nu < \infty\}$$

defines a smooth path which approaches the optimal solution of (P) as $\nu \to 0$. This path is called "the central trajectory of (5)." The main idea of the interior-point method is to solve the SDP with the following procedure to trace the central trajectory. Starting from a point close to the central trajectory $\mathcal{C}_{\mathrm{P}}$, we solve (5) by repeated application of the Newton method to (7), reducing $\nu$ gradually to zero.

A relevant part of this method is solution of (7) for each fixed $\nu$ where the Newton method is applied. The Newton method is basically a method for an unconstrained optimization problem, but the problem contains nontrivial constraints $X \succeq 0$. However, there is no difficulty in applying the Newton method here, because the problem is a minimization problem and the term $-\sum_j \log\det X_j$ diverges whenever as $X$ approach the boundary of the feasible region. Therefore, the Newton method is not bothered with the constraint $X \succeq 0$.

Another important issue here is initialization. We need an interior feasible solution to start. Usually this problem is resolved by so-called "two-phase method" or "Big-M method," which

are analogies of the techniques developed in classical linear programming. In the primal-dual method we introduce later, this point is resolved in a more elegant manner.

Now we extend the idea of interior-point method to solve (6). We consider the following problem with a positive parameter $\eta$:

$$\begin{aligned}(\widetilde{\mathrm{P}}_\eta)\quad &\min_X \sum_j \mathrm{Tr}(C_j X_j) - \sum_{j\in\Omega} \log\det X_j - \eta \sum_{j\notin\Omega} \log\det X_j, \\ &\text{s.t.} \sum_j \mathrm{Tr}(A_{ij} X_j) = b_i,\ i = 1, ..., m, \quad X_j \succeq 0, \quad j = 1, ..., \bar{n}.\end{aligned} \tag{9}$$

We denote by $\widetilde{X}(\eta)$ the optimal solution of (9), and define the central trajectory for (6) as

$$\mathcal{D}_{\mathrm{P}} \equiv \{\widetilde{X}(\eta) : 0 < \eta < \infty\}.$$

Note that $\widetilde{X}(\eta)$ approaches the optimal set of (6) as $\eta \to 0$.

Observe that the central trajectories $\mathcal{C}_{\mathrm{P}}$ and $\mathcal{D}_{\mathrm{P}}$ intersects at $\nu = 1$ and $\eta = 1$, i.e., $\widehat{X}(1) = \widetilde{X}(1)$. Therefore, we consider an interior-point method to solve (6) consisting of two stages. We first obtain a point $X^*$ close to $\widehat{X}(1)$ at Stage 1 with the ordinary interior-point method. In Stage 2, starting from $X^*$, a point close to the central trajectory $\mathcal{D}_{\mathrm{P}}$ for (6), we solve (9) with the Newton method repeatedly reducing $\eta$ gradually to zero. This idea is further incorporated with the primal-dual interior-point method in the next subsection.

## 3.2 Dual Problem, Primal-Dual Formulation, and Primal-Dual Interior-point Method

In this subsection, we introduce a dual problem and a primal-dual formulation of the optimization problems discussed in the previous section. First, the dual problem of (5) is defined as follows:

$$\begin{aligned}(\mathrm{D})\quad &\max_{y, Z_j} \sum_i b_i y_i, \\ &\text{s.t. } C_j - \sum_i A_{ij} y_i = Z_j,\ Z_j \succeq 0,\ j = 1, ..., \bar{n},\end{aligned} \tag{10}$$

where $Z_j$, $j = 1, ..., \bar{n}$, is $n_j \times n_j$ real symmetric matrix and $y = (y_1, ..., y_m)$ is $m$-dimensional real vector. We denote $(Z_1, ..., Z_{\bar{n}})$ by $Z$.

Under mild conditions, (5) and (10) have the optimal solutions with the same optimal value (the duality theorem) [7, 22, 25, 43]. Analogous to the case of (5), the central trajectory of (10) is defined as the set of the unique optimal solution of the following problem when parameter $\nu$ is changed:

$$\begin{aligned}(\mathrm{D}_\nu)\quad &\max_{y, Z_j} \sum_i b_i y_i + \nu \sum_j \log\det Z_j, \\ &\text{s.t. } C_j - \sum_i A_{ij} y_i = Z_j,\ Z_j \succeq 0,\ j = 1, ..., \bar{n}.\end{aligned} \tag{11}$$

We denote by $(\widehat{Z}(\nu), \hat{y}(\nu))$ the optimal solution of (11). Differentiation yields that $(\widehat{Z}(\nu), \hat{y}(\nu))$ is a unique optimal solution to the following system of nonlinear equations:

$$\begin{aligned}&\nu \sum_j A_{ij} Z_j^{-1} = b_i, \quad i = 1, ..., m, \\ &C_j - \sum_i A_{ij} y_j = Z_j,\ Z_j \succeq 0,\ j = 1, ..., \bar{n}.\end{aligned} \tag{12}$$

The set

$$\mathcal{C}_{\rm D} \equiv \{(\widehat{Z}(\nu), \hat{y}(\nu)) : 0 < \nu < \infty\}$$

defines a smooth path which approaches the optimal solution of (D) as $\nu \to 0$. This path is called "the central trajectory for (10)." Comparing (12) and (8), we see that $(\widehat{X}(\nu), \widehat{Z}(\nu), \hat{y}(\nu))$ is the unique optimal solution of the following bilinear system of equations:

$$\begin{array}{l} X_j Z_j = \nu I,\ j = 1, ..., \bar{n}, \\ \sum\limits_j \mathrm{Tr}(A_{ij} X_j) = b_i,\ i = 1, ..., m, \\ C_j - \sum\limits_i A_{ij} y_i = Z_j,\ j = 1, ..., \bar{n}, \\ X_j \succeq 0,\ j = 1, ..., \bar{n}, \quad Z_j \succeq 0,\ j = 1, ..., \bar{n}. \end{array} \tag{13}$$

Note that we also require $X_j = X_j^T$ and $Z_j = Z_j^T$ for each $X_j$ and $Z_j$, since "$\succeq$" means that a matrix is "symmetric" positive semidefinite. We define the central trajectory of (5) and (10) as

$$\mathcal{C} = \{\widehat{W}(\nu) :\ 0 < \nu < \infty\},$$

where $\widehat{W}(\nu) = (\widehat{X}(\nu), \widehat{Z}(\nu), \hat{y}(\nu))$. The primal-dual interior-point method solves (P) and (D) simultaneously by following the central trajectory $\mathcal{C}$ based on the formulation (13). Namely, we solve (13) repeatedly reducing $\nu$ gradually to zero.

Like in the primal method, a crucial part of the primal-dual method is the solution procedure of (13) for fixed $\nu$. There are several efficient iterative algorithms [15, 17, 21, 23, 26, 27] developed for this subproblem based on the Newton method for a system of nonlinear equations. We explain these methods in more detail in Appendix.

Now we introduce the dual counterpart of (6) as follows:

$$\begin{array}{ll} (\widetilde{\rm D}) & \max\limits_{y, Z_j} \sum\limits_i b_i y_i + \sum\limits_{j \in \Omega} \log\det Z_j + |\Omega|, \\ & \text{s.t. } C_j - \sum_i A_{ij} y_i = Z_j,\ Z_j \succeq 0,\ j = 1, ..., n. \end{array} \tag{14}$$

It is known that the optimal values of (14) and (6) coincides. In order to solve this problem, we consider the following optimization problem with parameter $\eta > 0$:

$$\begin{array}{ll} (\widetilde{\rm D}_\eta) & \max\limits_{y, Z_j} \sum\limits_i b_i y_i - \sum\limits_{j \in \Omega} \log\det Z_j, - \eta \sum\limits_{j \notin \Omega} \log\det Z_j \\ & \text{s.t. } C_j - \sum_i A_{ij} y_i = Z_j,\ Z_j \succeq 0,\ j = 1, ..., n. \end{array} \tag{15}$$

We denote by $(\widetilde{Z}(\eta), \tilde{y}(\eta))$ the unique optimal solution of this problem. We define the central trajectory for (14) as

$$\mathcal{D}_{\rm D} \equiv \{(\widetilde{Z}(\eta), \tilde{y}(\eta)) : 0 < \eta < \infty\}.$$

Note that $(\widetilde{Z}(\eta), \tilde{y}(\eta))$ approaches the optimal set of (9) as $\eta \to 0$. The set $\mathcal{D}_D$ of the solutions $\mathcal{D}_{\rm D}$ is referred to as the central trajectory for (14).

Analogous to (13), we have the following primal-dual formulation of $(\widetilde{X}(\eta), \widetilde{Z}(\eta), \tilde{y}(\eta))$:

$$\begin{array}{l} X_j Z_j = I,\ j \in \Omega \\ X_j Z_j = \eta I,\ j \notin \Omega \\ \sum\limits_j \mathrm{Tr}(A_{ij} X_j) - b_i = 0,\ i = 1, ..., m, \\ C_j - \sum\limits_i A_{ij} y_i - Z_j = 0,\ j = 1, ..., \bar{n}, \\ X_j \succeq 0\ j = 1, ..., \bar{n}, \quad Z_j \succeq 0,\ j = 1, ..., \bar{n}. \end{array} \tag{16}$$

We define

$$\mathcal{D} = \{(\widetilde{X}(\eta), \widetilde{Z}(\eta), \tilde{y}(\eta)) : \ 0 < \eta < \infty\}$$

as the primal-dual central trajectory of (6) and (14). The equation (16) can be solved efficiently with the same iterative methods for (13).

Now we are ready to describe a primal-dual interior-point method for (6) and (14). As in the case of the primal method, the primal-dual central trajectories $\mathcal{C}$ and $\mathcal{D}$ intersects at $\nu = \eta = 1$, i.e., $\widehat{W}(1) = \widetilde{W}(1)$. Therefore, we can solve (6) and (14) in two stages as follows. We first apply the ordinary primal-dual interior-point method for (5) and (10) to find a point $W^* = (X^*, Z^*, y^*)$ close to $\widehat{W}(1)$. Then starting from $W^*$, a point close to the central trajectory $\mathcal{D}$ for (6) and (14), we solve (6) by solving (16) approximately repeatedly reducing $\eta$ gradually to zero.

A remarkable advantage of the primal-dual method is its flexibility concerning initialization. In the primal formulation in the previous subsection, the method needs an initial feasible interior point. But obtaining such a point is already a nontrivial problem. In the primal-dual formulation, we can get around this difficulty, because the search directions can be computed for any $(X, Z, y)$ such that $X \succ 0$ and $Z \succ 0$. Generally such point does not necessarily satisfy linear equalities in (16), but we may let them be satisfied in the end of iterations, since they are linear. In that case, we approach the central trajectory from outside of feasible region.

Another important advantage of the primal-dual method is availability of an upperbound of the maximum value of the log likelihood function. This bound is obtained as the dual objective function value. Indeed, we observed in the numerical experiments conducted in the next section that feasibility of the dual problem (14) is satisfied more quickly and accurately in many cases, providing an upperbound on the maximum likelihood function value. This point is particularly important in a situation we need to compute the maximum likelihood estimate repeatedly for many models. With the dual information, we may truncate iteration in earlier stage if the dual information suggests the maximum likelihood will not be good.

We provided two versions of the primal-dual method in our implementation; (i) basic algorithm and (ii) predictor-corrector algorithm. The first one follows the central trajectory loosely. The method is simple and efficient, but often encounters difficulty for ill-conditioned problems as is reported later. The predictor-corrector algorithm follows the central trajectory more precisely. This method is slower but is robust and steady, suitable for ill-conditioned and difficult problems. We describe further details of these methods in Appendix. The primal-dual methods for (6) is studied in detail in [41], where somewhat different homotopy path leading to $\widetilde{W}(0)$ is introduced to follow. The two-stage algorithm we introduced here seems to work reasonably well for this statistical application so far.

# 4 Numerical Results

## 4.1 Outline

We conducted numerical experiments of our method with the following five models.

(i) Normal based model.

(ii) Exponential based model.

(iii) Pure polynomial model where the density function and its first derivative are assumed to be zero on the boundary of the support.

(iv) Normal based model where we require another additional condition that the estimated density function is unimodal.

(v) Exponential based model where we require another additional condition that the estimated density function is monotone decreasing.

Some of the datasets used in this experiment are generated by simulation from assumed distributions and others are taken from real datasets which have been often used for benchmark. The algorithms are coded in MATLAB and C, and all the numerical experiments are conducted under MATLAB 6.5 environment with the Windows OS. We used several platforms, but the typical one is like Pentium IV 2.4GHz with 1GB Memory. The code runs without trouble in a notebook computer equipped with a Pentium III 650MHz CPU and 256MB Memory.

As was explained before, the maximum likelihood estimation is computed in two steps. Namely, we optimize parameter $\alpha$ associated polynomials with SDP and at higher level we optimize parameter $\beta$ for the base density. We have $\beta = (\mu, \sigma)$ for (i), $\beta = \lambda$ for (ii), and $\beta = [a_{\min}, a_{\max}]$ for (iii), and in (iv), we have $\beta = (\mu, \sigma, \gamma)$, where $\gamma$ denotes the peak of distribution. Finally in (v), we have $\beta = \lambda$. Assuming that $\alpha$ is optimized with SDP, we just need to perform at most three (basically one or two) dimensional optimization problems to accomplish global optimization of the likelihood function.

According to the level of difficulty of SDP to be solved later, we employed

(a) Optimization by nonlinear optimization (for (i), (ii) and (v));

(b) Optimization by grid search (for (i), (ii) and (iv)):

(c) Optimization by manual (for (iii));

for the higher level optimization.

As was explained in the last section, we provided two versions of the interior-point methods; the basic algorithm and the predictor-corrector algorithm. Generally, we observed that SDPs for (i), (ii) and (v) are fairly easy while the ones for (iii) and (iv) are more difficult. The basic algorithm is robust and stable enough to solve (i), (ii) and (v) without trouble. On the other hand, the basic algorithm got into trouble when we tried to solve (iii) and (iv). In that case, we need to use a more sophisticated predictor-corrector algorithm. The typical number of iterations of the basic algorithm is around fifty, and the predictor-corrector algorithm is between one hundred to two hundred.

Concerning higher level optimization to determine $\beta$, we adopted nonlinear optimization and grid search procedure for (i), (ii), (iv) and (v). Thus, the whole procedure is fairly automated. As to (iii), optimization is done manually by trial and error, since the SDP routine requires more time and it is yet difficult to adjust appropriate stopping criteria. Further tune up of the code is a future subject of research. We compare the models with AIC [1, 2]. Here we note that the number of parameters should be reduced by one for one addition of one equality constraint. Therefore, if we define

$$\text{AIC} = -(\text{Log Likelihood}) + k,$$

where $k$ denotes the number of parameters, $k$ will become as follows for the five cases (i)$\sim$(v):

(i) Normal based model: $k = n + 2$
($\dim(\alpha) = n + 1,\ \dim(\beta) = 2$, (number of linear equalities) $= 1$).

(ii) Exponential based model: $k = n + 1$
($\dim(\alpha) = n + 1,\ \dim(\beta) = 1$, (number of linear equalities) $= 1$).

(iii) Pure polynomial model: $k = n - 2$
($\dim(\alpha) = n + 1,\ \dim(\beta) = 2$, (number of linear equalities) $= 5$).

(iv) Normal based model with unimodality: $k = n + 2$
($\dim(\alpha) = n + 1,\ \dim(\beta) = 3$, (number of linear equalities) $= 2$).

(v) Exponential based model with monotonicity: $k = n + 1$,
($\dim(\alpha) = n + 1,\ \dim(\beta) = 1$, (number of linear equalities) $= 1$).

Here $n$ is the degree of the polynomial $p(x; \alpha)$ in the model. We define AIC as one half of the usual definition of AIC. All the model contains the linear equality constraint that the integral of the estimated density over the support is one. The pure polynomial model (iii) contains additional equality constraints that the value of the density function and its derivative on the both end of its support is zero. In (iv) and (v), we did not give any "penalty term" on unimodality and monotonicity. In (iv), we introduce a new parameter to specify the peak of the density but we also have additional linear equality constraint that the derivative of the density is zero at the peak. Therefore, after all, the penalty term is the same as (i).

The data analyzed here is as follows.

(i) Normal based model

  (a) Simulated data 1 generated from a bimodal normal mixture distribution ($N = 100$).
  (b) Simulated data 2 generated from an asymmetric unimodal normal mixture distribution ($N = 250$).
  (c) Buffalo snowfall data ($N = 62$) [8, 30, 33].
  (d) Old faithful geyser data ($N = 107$) [45, 34].

(ii) Exponential based model

  (a) Simulated data 3 generated from a mixture distribution of an exponential distribution and a gamma distribution ($N = 200$).
  (b) Coal-mining disasters data ($N = 109$) [10].

(iii) Pure polynomial model

  (a) Old faithful geyser data
  (b) Galaxy data ($N = 82$) [31].

(iv) Normal based model with unimodality condition

  (a) The normal mixture distribution dataset treated in (i-b).

(v) Exponential based model with monotonicity condition

  (a) Coal-mining disasters data treated in (iii-b).

## 4.2 Normal based model

In this subsection, we show the results of the density estimation with the normal based model.

**[Simulated data 1: Bimodal normal mixture distribution]**

We generated 200 samples from a bimodal normal mixture distribution

$$\frac{0.3}{\sqrt{2\pi 0.5^2}} \exp\left(-\frac{(x+1)^2}{2\cdot 0.5^2}\right) + \frac{0.7}{\sqrt{2\pi 0.5^2}} \exp\left(-\frac{(x-1)^2}{2\cdot 0.5^2}\right).$$

Figures 1 (a)-(c) show the estimated densities from this data for $n = 2, 4, 6$. MAIC procedure picks up the model of degree 4 as the best model. It is seen that the estimated density (solid line) is close to the true density (broken line). In Figure 2, the change of AIC values is shown when the degree of the polynomial increases. This figure shows a typical behavior of AIC.

**[Simulated data 2: Asymmetric unimodal normal mixture distribution]**

Here we generated a simulated dataset of 250 samples from a distribution proportional to

$$\exp\left(-\frac{x^2}{2}\right) + 5\exp\left(-\frac{(x-1)^2}{0.2}\right) + 3\exp\left(-\frac{(x-1)^2}{0.5}\right).$$

This is an asymmetric distribution which has a sharp peak around $x = 1$. The estimated density by MAIC procedure is shown in Figures 3. MAIC procedure chooses the model with degree 8. The values of log likelihood (LL) and AIC are LL $= -215.1$ and AIC $= 223.1$. The estimated density seems to have a bump on the left hand tail. Thus, we see that the estimation of the density function is more difficult on the left hand tail as long as we estimate the distribution just from the data. Later we will show how the estimation is stabilized if we assume the prior knowledge of unimodality of distribution.

**[Buffalo snowfall data]**

This data is the set of 63 values of annual snowfall in Buffalo, New York, USA from 1910 to 1972, in inches [8, 30, 33]. In Figures 4 (a)-(c), we show profiles of the distribution obtained with the maximum likelihood estimate when the degree of polynomial is decreased/increased. MAIC procedure chooses the model of degree 6 and seems to give a reasonable result.

**[Old faithful geyser data]**

This data contains duration times of 107 nearly consecutive eruption of the Old Faithful geyser in minutes [45, 34]. The estimated density is shown in Figure 5, where $n = 10$, LL $= -105.6$ and AIC $= 117.6$. It has longer tails in the both ends, reflecting the nature of the normal distribution. It seems that the tail of the distribution should be shorter. We also applied the pure polynomial model. The latter seems to give a better fit with shorter tail, as is shown later.

## 4.3 Exponential based model

In this subsection, we show the results of the density estimation with the exponential based model.

**[Simulated data 3: Mixture an exponential distribution and a gamma distribution]**

Here we generated a simulated data of 200 samples from a mixture distribution of an exponential distribution and a gamma distribution with shape parameter 4

$$0.2\left\{2\exp(-2x)\right\} + 0.8\left\{\frac{x^3}{3!}\exp(-x)\right\}.$$

MAIC procedure picks up the model with degree 2. As is seen in Figure 6, the estimated distribution obtained by MAIC procedure recovers fairly well the original distribution.

**[Coal-mining disasters data]**

Coal-mining disasters in Great Britain from 1875 to 1951 are reported in days [10]. See Figure 12 for the original sequence of disasters. Here we model it as a renewal process to estimate the interval density function with the exponential based model. MAIC procedure picks up the model with degree 4, where LL $= -699.0$ and AIC $= 704.0$. The estimated distribution is shown in Figure 7. It is seen that the distribution is considerably different from the exponential distribution. The estimated density seems to consists of three slopes. It is seen that there is a small bump around $x = 1200$. Later we will show how the estimated density will change if the density function is assumed to be monotonically decreasing.

## 4.4 Pure polynomial model

In this subsection, we show the profiles of the density functions estimated with the pure polynomial model.

**[Galaxy data]**

This data is obtained by measurements of the speed of galaxies in a region of the universe [31]. We applied the pure polynomial model to estimate the density from Galaxy Data, since the model with normal distribution base did not fit well. MAIC procedure chooses the model with degree 13. As is seen from Figure 8, the model seems to fit reasonably well to the data, suggesting that there are three clusters in distribution.

**[Old faithful geyser data]**

We applied the pure polynomial model to estimate the density function of the Old faithful geyser data. MAIC procedure chooses the model with degree 14, where LL $= -96.4$ and AIC $= 108.4$. The estimated density is shown in Figure 9. This model fits much better compared with the normal based model in terms of both likelihood and AIC. The model captures well the structure of distribution, suggesting that the distribution have a very short tail and the left peak is higher than the right one.

## 4.5 Normal based model with unimodality condition

As is demonstrated so far, our approach gives reasonable estimates for many cases. When we analyze real data, we sometimes have prior knowledge on the distribution such as symmetry or unimodality. It is difficult to incorporate such prior knowledge into model in nonparametric approaches. In the following, we pick up the simulated data 2 as an example, and compute the maximum likelihood estimate with the unimodality condition in our approach. Specifically, we observe how addition of the unimodality condition changes the estimated density function.

The distributions obtained by MAIC procedure adding an additional constraint of unimodality for the dataset is shown in Figure 10. LL and AIC are $-216.4$ and 224.5, respectively. We see that the bumps in the estimated density without unimodality condition disappeared and the shape of distribution looks closer to the original ones. LL and AIC get worse by about 1.5, where no "penalty" is added to AIC for the unimodality constraint.

### 4.6 Exponential based model with monotonicity condition

In Subsection 4.2, we estimated the interval density function of the coal-mining disasters data, and observed that the density function with the best AIC value has a bump around $x = 1200$. Here we estimate the density based on the same data with the exponential based model with monotonicity condition. In Figure 11, we show the best model we found, where the degree of polynomial is 4, LL $= -699.57$ and AIC $= 704.57$. We did not impose any "penalty" on monotonicity. These values of LL and AIC is almost the same as the ones we obtained in Subsection 4.2.

## 5 Other Applications

The approach proposed here can be applied to other areas of statistics such as point process, survival data analysis etc. In order to clarify this point, here we pick up estimation of the intensity function of a nonstationary Poisson process as an example. Suppose we have a nonstationary Poisson process whose intensity function is given by $\lambda(t)$, and let $t_1, ..., t_N = T$ be the sequence of the time when events were observed. We estimate $\lambda(t)$ as a nonnegative polynomial function on the interval $(0, T]$. The log likelihood is

$$\sum_{i=1}^{N} \log \lambda(t_i) - \int_0^T \lambda(t) \mathrm{d}t.$$

If $T$ is fixed, then we can apply exactly the same technique as density estimation developed in this paper. If we represent $\lambda(t)$ as (2) and/or (3) in Theorem 2.1, the term

$$\int_0^T \lambda(t) \mathrm{d}t$$

is represented as

$$\int_0^T \lambda(t) dt = \mathrm{Tr}(M_1 Q) + \mathrm{Tr}(M_2 Q_1),$$

where $M_1$ and $M_2$ are appropriate symmetric matrices. Therefore, the maximum likelihood estimation is formulated as the following problem:

$$\max \sum_{i=1}^{N} \log \Big(\mathrm{Tr}(X_1^{(i)} Q) + \mathrm{Tr}(X_2^{(i)} Q_1)\Big) - \Big(\mathrm{Tr}(M_1 Q) + \mathrm{Tr}(M_2 Q_1)\Big), \quad \text{s.t. } Q \succeq 0,\ Q_1 \succeq 0,$$

where $X_1^{(i)}$ and $X_2^{(i)}$, $(i = 1, ..., N)$ are matrices determined from the data. Thus, the problem just becomes (6) in this case. In Figure 12, we show the estimated intensity function $\lambda(t)$ for the coal-mining data with MAIC procedure. The procedure picks up the polynomial of degree

7, where LL = −690.8 and AIC = 698.8. In the previous section, we analyzed this data as a renewal process, and AIC of the estimated model is around 704.0 in the both of the cases where we require or not require monotonicity condition. Thus, we see that the nonstationary Poisson model seems to fit better in this case than the renewal model.

A similar technique can be applied to the analysis of other statistical problems such as estimation of a survival function for medical data etc. This is another interesting topic of further study.

# 6 Concluding Discussion

In this paper, we proposed a novel approach to the classical density estimation problem by means of semidefinite programming. We adapted standard interior-point methods for SDP to this problem, and demonstrated through various numerical experiments that the method gives reasonable estimate of the density function with MAIC procedure. We also showed that such properties as unimodality and monotonicity of the density function can be easily handled within this framework. There are several issues for further research.

The first issue is improvement of implementation. There are two aspects: (i) stability and robustness and (ii) speed and space. Concerning stability and robustness, we would say that our code works to some extent in a stable manner in the sense that it solves many of the SDP problems for optimizing the polynomial parameter $\alpha$, accomplishing enough level to be used in a grid search for the base density parameter $\beta$. But on the other hand, we admit that there still remain problems difficult to solve. Therefore, further tune up is necessary in particular if we synthesize it completely with a more sophisticated nonlinear optimization routine.

Next we discuss the issue of speed and space. In this study we put more emphasis on robustness of the algorithm. For this reason, the algorithm is a bit slower than expected in terms of the number of iterations. The code is not also fast enough yet in view of timing data. This is because our purpose of development of the current code written in MATLAB is to check feasibility of the idea and we did not pursuit efficiency. Therefore, there are several things to be done to make it faster.

From the algorithmic point of view, probably it would be possible to reduce the number of iterations by half or by one thirds if we incorporate with sophisticated implementation techniques like the Mehrotra predictor-corrector algorithm [35, 41]. Furthermore, it is possible to develop a code performing one iteration efficiently since our problem is a sparse problem with a special structure where the most of the blocks $X_i$ is one by one and there are only several blocks of relatively small semidefinite matrices. When we reasonably exploit these structures of this problem, the number of arithmetic operations required per iteration of the primal-dual method becomes $O(N^3)$, where $N$ is the number of data and the degree of the polynomial is assumed to be small. Memory requirement is proportional to $O(N^2)$. This suggests that our approach is not computationally expensive assuming that the number of iteration is typically up to fifty like other interior-point methods for SDP.

Still these factors may limit applicability of our approach for large problems. However, the current sophisticated implementation [35, 42, 47] solves the SDP problems involving the matrices whose dimension is several thousands. Therefore, we have a good chance to solve the density estimation problems where the number of data is up to several thousands.

Another remedy to deal with a large problem is to use a histogram density estimate and then smooth the histogram based on our method. Our method is easily extended to smooth a histogram. In this case, we are not bothered with choice of the number of the bins in constructing the histogram. The number of the bins, which corresponds to the number of samples in this paper, should be as large as possible within the range that the associated semidefinite program can be solved.

From the statistical point of view, the analysis of the properties of the model on the boundary would be important. There is a possibility that the estimated polynomial has a root on the real axis. This is a sort of irregular situation where the standard asymptotic theory for the maximum likelihood estimator does not hold. We "brutely" applied MAIC procedure even in this case, but it would be nice if a reasonable treatment of penalty term is developed. Note that this irregular condition can occur not only for nonnegativity constraint for the density $f(x)$ itself, but also for nonnegativity constraint for the derivative $f'(x)$ etc. Study of comparison of models with different supports based on AIC is also important in our context.

As was explained in the previous section, the techniques of this paper can be applied to other areas of statistics such as survival data analysis, point process etc. Development of applications to these areas is also an interesting further topic of research.

# References


[1] H. Akaike: Information theory and an extension of the maximum likelihood principle. In *Proceeding of the Second International Symposium on Information Theory*, eds., B. N. Petrov and F. Csàski, pp. 267-281, Akademiai Kiado, Budapest, 1973. (Also available as *Breakthroughs in Statistics*, Vol. 1, eds., S. Kotz and N. L. Johnson, pp. 610–624, 1992, Springer-Verlag.)

[2] H. Akaike: A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, Vol. 19 (1974), pp. 716-723.

[3] H. Akaike: On entropy maximization principle. *Applications of Statistics*, (Krishnaiah, P.R. ed.), pp. 27–41, North-Holland, 1977. *IEEE Trans. on Automatic Control*, Vol. 19 (1974), pp. 716-723.

[4] H. Akaike and E. Arahata: GALTHY, A probability density estimation. *Computer Science Monographs*, No. 9, The Institute of Statistical Mathematics, Tokyo, 1978.

[5] F. Alizadeh: Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM Journal on Optimization*, Vol. 5 (1995), pp. 13–51.

[6] A. R. Barron and T. M. Cover: Minimum complexity density estimation. *IEEE Transaction of Information Processing*, Vol. 37 (1991), pp. 1034-1054.

[7] A. Ben-Tal and A. Nemirovski: *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. *SIAM*, Philadelphia, PA, USA, 2001.

[8] J.-P. Carmichael: *The Autoregressive Method: A Method for Approximating and Estimating Positive Functions*. Ph. D. Thesis, SUNY, Buffalo, New York, 1976.

[9] J. B. Copas: Local likelihood based on kernel censoring. *Journal of Royal Statical Society Series B*, Vol. 57 (1995), pp. 221–235.

[10] D. R. Cox and P. A. W. Lewis: *The Statistical Analysis of Series of Events*. John Wiley, New York, 1966.

[11] P. P. B. Eggermont and V. N. LaRiccia: *Maximum Penalized Likelihood Estimation; Vol.1: Density Estimation.* Springer, 2001.

[12] R. Fletcher: *Practical Methods of Optimization* (2nd edition), John Wiley, 1989.

[13] I. J. Good and R. A. Gaskins: Nonparametric roughness penalties for probablity densities. *Biometrica*, Vol. 58 (1971), pp. 255–277.

[14] I. J. Good and R. A. Gaskins: Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and Meteorite Data. *Journal of American Statistical Association*, Vol. 75 (1980) pp. 42-56.

[15] C. Helmberg, F. Rendl, R. J. Vanderbei, and H. Wolkowicz: An interior-point method for semidefinite programming. *SIAM Journal on Optimization*, Vol. 6 (1996), pp.342–361.

[16] M. Ishiguro and Y. Sakamoto: A Bayesian approach to the probability density estimation. *Annals of the Institute of Statistical Mathematics.* Vol. 36 (1984), pp. 523–538.

[17] M. Kojima, S. Shindoh and S. Hara: Interior-point methods for the monotone semidefinite linear complementarity problem in symmetric matrices. *SIAM Journal on Optimization*, Vol. 7 (1997), pp. 86–125.

[18] G. McLachlan and D. Peel: *Finite Mixture Models.* John-Wiley, 2000.

[19] Hans Mittelmann: An independent benchmarking of SDP and SOCP solvers. *Mathematical Programming*, Vol. 95 (2003), pp. 407–430.

[20] Hans Mittelmann: Benchmarks for optimization software. http://plato.asu.edu/bench.html

[21] R. D. C. Monteiro: Primal-dual path-following algorithms for semidefinite programming. *SIAM Journal on Optimization*, Vol. 7, No. 3 (1997), pp. 663-678.

[22] R. D. C. Monteiro and M. J. Todd: *Path Following Methods*. In *Handbook of Semidefinite Programming; Theory, Algorithms, and Applications* (eds. H. Wolkowicz, R. Saigal and L. Vandenberghe), Kluwer Academic Publishers, 2000, pp. 267-306.

[23] R. D. C. Monteiro and T. Tsuchiya: polynomial convergence of a new family of primal-dual algorithms for semidefinite programming. *SIAM Journal on Optimization*, Vol. 9 (1999), pp. 551–577.

[24] Yu. Nesterov: Squared functional systems and optimization problems. In *High performance optimization* (eds. H. Frenk, K. Roos, T. Terlaky and S. Zhang), pp. 405–440, Kluwer, Dordrecht, 2000.

[25] Yu. Nesterov and A. Nemirovskii: *Interior-point Methods for Convex Programming.* SIAM Publisher, Philadelphia, 1994.

[26] Y. E. Nesterov and M. Todd: Self-scaled barriers and interior-point methods for convex programming. *Mathematics of Operations Research*, Vol. 22, No. 1 (1997).

[27] Y. E. Nesterov and M. Todd: Primal-dual interior-point methods for self-scaled cones. *SIAM Journal on Optimization*, Vol. 8 (1998), pp. 324–364.

[28] J. Nocedal and S. J. Wright: *Numerical Optimization*, Springer, 1999.

[29] A. Ohara and T. Tsuchiya: Graphical modelling with interior point methods and combinatorial optimization techniques: an analysis on relationship between lawyers and judges in Japan, *Research Memorandum* No. 774, The Institute of Statistical Mathematics, Tokyo, October, 2000.

[30] E. Parzen: Nonparametric statistical data modeling, *Journal of American Statistical Association*, Vol. 74 (1979), pp. 105-131.

[31] K. Roeder: Density estimation with confidence sets exemplified by superclusters and voids in the galaxies, *Journal of American Statistical Society*, Vol. 85 (1990), No. 11, pp. 617-624.

[32] Y. Sakamoto: *Model Analysis of Categorical Data* (in Japanese), Kyoritsu Shuppan, Tokyo, 1985. (English Transration by the author is available as *Categorical Data Analysis by AIC*, Kluwer Publisher, 1991.)

[33] D. W. Scott: *Multivariate Density Estimation.* John Wiley and Sons, USA, 1992,

[34] B. W. Silvermann: *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.

[35] Jos F. Sturm: Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. Interior point methods. *Optimization Methods and Software*, Vol. 11/12 (1999), pp. 625–653.

[36] K. Tanabe: Bayes model and ABIC (in Japanese). *Communications of the Operations Research Society of Japan*, Vol. 30 (1985), pp. 178-183.

[37] K. Tanabe and M. Sagae: A Bayes method for nonparametric univariate and bivariate probability density estimation. Unpublished manuscript, 1984.

[38] K. Tanabe and M. Sagae: An empirical Bayes method for nonparametric density estimation. *Cooperative Research Report of the Institute of Statistical Mathematics*, Vol. 118 (1999), pp. 11-37 (An abbreviated version of [37]).

[39] K. Tanabe, M. Sagae and S. Ueda: BNDE FORTRAN subroutines for Bayesian nonparametric univariate and bivariate density estimators. Research Report, Department of Information Sciences, Science University of Tokyo, 1984. (A revised version with the same title and the authors is available as *Computer Science Monograph*, Vol. 24, the Institute of Statistical Mathematics, Tokyo, 1988.)

[40] R. A. Tapia and J. R. Thompson: *Nonparametric Probability Density Estimation*, Johns Hopkins University Press, Baltimore, USA, 1978.

[41] Kim-Chuan Toh: Primal-dual path-following algorithms for determinant maximization problems with linear matrix inequalities. *Computational Optimization and Applications*, Vol. 14 (1999), pp. 309–330.

[42] R. H. Tütüncu, K. C. Toh, and M. J. Todd: Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical Programming*, Vol. 95 (2003), pp. 189–217.

[43] L. Vandenberghe and S. Boyd: Semidefinite programming. *SIAM Review*, Vol. 38 (1996), pp. 49–95.

[44] L. Vandenberghe, S. Boyd, and S-P. Wu: Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, Vol. 19 (1998), pp. 499–533.

[45] S. Weisberg: *Applied Linear Regression*. John Wiley, New York, 1985.

[46] H. Wolkowicz, R. Saigal and L. Vandenberghe: *Handbook of Semidefinite Programming; Theory, Algorithms, and Applications*. Kluwer Academic Publishers, 2000.

[47] M. Yamashita, K. Fujisawa and M. Kojima: Implementation and evaluation of SDPA 6.0 (SemiDefinite Programming Algorithm 6.0), *Optimization Methods and Software*, Vol.18 (2003), pp. 491-505.

# Appendix

# A Variants of the Newton Method to Solve (13) and (16)

In this section, we outline the variants of the Newton method for (13) and (16). Since (13) and (16) are systems of bilinear and linear equations, we may think of directly applying the Newton method to these systems. But this idea does not work because the $X$-part and $Z$-part of the Newton direction generally does not lie in the space of symmetric matrices. This difficulty is remedied by replacing the equations $X_jZ_j = \zeta I$, $X_j = X_j^T$ and $Z_j = Z_j^T$ in (13) and (16) (N. B. the latter two conditions on symmetry of $X_j$ and $Z_j$ do not appear explicitly in these systems) with another equivalent system of bilinear equations, say, $\Phi_\zeta^{(j)}(X_j, Z_j) = 0$ for $j = 1, ..., \bar{n}$, and then applying the Newton method to the modified system of equations. A typical example of $\Phi_\zeta^{(j)}$ is

$$\Phi_\zeta^{(j)}(X_j, Z_j) = \frac{1}{2}(X_jZ_j + Z_jX_j) - \zeta I.$$

It is known that $\Phi_\zeta^{(j)}(X_j, Z_j) = 0$ iff $X_j$ and $Z_j$ are symmetric matrices satisfying $X_j Z_j = \zeta I$. Then we apply the Newton method to the system of linear and bilinear equations

$$\begin{array}{l} \Phi_\nu^{(j)}(X_j, Z_j) = 0, \ j = 1, ..., \bar{n} \\ \sum\limits_j \mathrm{Tr}(A_{ij} X_j) = b_i, \ i = 1, ..., m, \\ C_j - \sum\limits_i A_{ij} y_i = Z_j, \ j = 1, ..., \bar{n}, \\ X_j \succeq 0 \ j = 1, ..., \bar{n}, \quad Z_j \succeq 0, \ j = 1, ..., \bar{n}. \end{array} \tag{17}$$

to solve (13) and

$$\begin{array}{l} \Phi_1^{(j)}(X_j, Z_j) = 0, \ j \in \Omega \\ \Phi_\eta^{(j)}(X_j, Z_j) = 0, \ j \in \{1, ..., \bar{n}\} \backslash \Omega \\ \sum\limits_j \mathrm{Tr}(A_{ij} X_j) = b_i, \ i = 1, ..., m, \\ C_j - \sum\limits_i A_{ij} y_i = Z_j, \ j = 1, ..., \bar{n}, \\ X_j \succeq 0 \ j = 1, ..., \bar{n}, \quad Z_j \succeq 0, \ j = 1, ..., \bar{n}. \end{array} \tag{18}$$

to solve (16). We do not go into further details of this issue but just mention that there are two popular search directions employed to solve (13) called HRVW/KSH/M direction [15, 17, 21] and NT direction [26, 27] which are generated as the Newton directions for certain reasonable choices of $\Phi_\nu^{(j)}(X_j, Z_j)$, adding another twist to $\Phi_\nu$ introduced above. These directions are known to be efficient both in theory and practice [22, 35, 42, 47]. We used HRVW/KSH/M direction in our implementation.

# B Basic Algorithm and Predictor-Corrector Algorithm

In this appendix, we describe in more detail the two versions of the primal-dual methods we implemented, namely, the basic algorithm and the predictor-corrector algorithm. First we introduce a few relevant quantities necessary to explain the methods.

For $W = (X, Z, y)$ such that $X \succ 0$ and $Z \succ 0$, we define

$$\mu(X, Z) \equiv \frac{\sum_{j=1}^{\bar{n}} \mathrm{Tr}(X_j Z_j)}{\sum_{j=1}^{\bar{n}} n_j}.$$

Suppose that $X$ and $Z$ satisfy the centrality condition $X_j Z_j = \nu I$ for all $j = 1, ..., \bar{n}$ for some $\nu > 0$. It is easy to verify that

$$\mu(X, Z) = \nu.$$

Therefore, we associate any point $W = (X, Z, y)$ such that $X \succ 0$ and $Z \succ 0$ to a point on the central trajectory $\mathcal{C}$ with the barrier parameter $\nu = \mu(X, Z)$.

We also introduce analogous concepts for the central trajectory $\mathcal{D}$. Namely, we introduce the duality gap as

$$\tilde{\mu}(X, Z) \equiv \frac{\sum_{j \notin \Omega} \mathrm{Tr}(X_j Z_j)}{\sum_{j \notin \Omega} n_j}.$$

[**Basic Algorithm**]

As was mentioned before, the algorithm consists of two stages. In the first stage it finds a point $W^*$ in the neighborhood of $\widehat{W}(1)$. Then in the second stage, the algorithm generates a sequence approaching an optimal solution of (6).

[**Stage 1**]

**(Step 0)** Let $W^0 = (X^0, Z^0, y^0)$ be an initial solution satisfying $X^0 \succ 0$ and $Z^0 \succ 0$. set $k = 1$;

**(Step 1)** If $W^k$ is sufficiently close to $\widehat{W}(1)$, then go to **Stage 2**.

**(Step 2)** Compute $\nu^{k+1}$. (See the updating scheme of $\nu$ described below.)

**(Step 3)** Set $\nu := \nu^{k+1}$, and compute the Newton direction $\Delta W = (\Delta X, \Delta Z, \Delta y)$ to (17).

**(Step 4)** If $X^k + \Delta X \succ 0$ and $Z^k + \Delta Z \succ 0$ holds, then $W^{k+1} = W^k + \Delta W$. Otherwise take a step with a fixed fraction $\theta \in (0, 1)$ of the way to the boundary of semidefinite cones, i.e., compute the maximum step $t^*$ such that $X^k + t^*\Delta X \succeq 0$ and $Z^k + t^*\Delta Z \succeq 0$ holds, and set $W^{k+1} = W^k + \theta t^* \Delta W$.

**(Step 5)** Set $k := k + 1$, and return to **(Step 1)**.

[**Stage 2**]

**(Step 0)** Set $\eta^k = 1$.

**(Step 1)** If $\eta^k$ is sufficiently small, then return $X^k$ as an optimal solution to (6)

**(Step 2)** Compute $\eta^{k+1}$. (See the updating scheme of $\eta$ described below.)

**(Step 3)** Set $\eta := \eta^{k+1}$, and compute the Newton direction $\Delta W = (\Delta X, \Delta Z, \Delta y)$ to (18).

**(Step 4)** If $X^k + \Delta X \succ 0$ and $Z^k + \Delta Z \succ 0$ holds, then $W^{k+1} = W^k + \Delta W$. Otherwise take a step with a fixed fraction $\theta \in (0, 1)$ of the way to the boundary of semidefinite cones, i.e., compute the maximum step $t^*$ such that $X^k + t^*\Delta X \succ 0$ and $Z^k + t^*\Delta Z \succ 0$ holds, and set $W^{k+1} = W^k + \theta t^* \Delta W$.

**(Step 5)** Set $k := k + 1$, and return to **(Step 1)**.

We set initial value as $X^0 = Z^0 = \sqrt{\nu^0} I$ for $\nu^0 > 0$ sufficiently large. Updating scheme of $\nu$ and $\eta$ of Stage 1 (Step 1) and Stage 2 (Step 2) is an important ingredient of the method. We took the following strategy:

**Updating Scheme of $\nu$ and $\eta$ in Basic Algorithm**

- **Stage 1 (Step 1)**
  1. If $\mu^k \geq 10$, then set $\nu^{k+1} = 0.5\mu(X^k, Z^k)$.
  2. If $1 \leq \mu(X^k, Z^k) \leq 10$, then set $\nu^{k+1} = 1$.

- **Stage 2 (Step 2)**

  1. $\eta^{k+1} = 0.2\tilde{\mu}(X^k, Z^k)$.

**[Predictor-Corrector Algorithm]**

In the predictor-corrector algorithm, we trace the central trajectory more precisely. This strategy is very important to solve difficult problems. For this purpose, we introduce a neighborhood of the central trajectory $\mathcal{C}$

$$\mathcal{N}(\beta) \equiv \left\{(X, Z, y) : \sqrt{\sum_j \|X_j^{1/2} Z_j X_j^{1/2} - \mu(X, Z) I\|_F^2} \leq \beta \mu(X, Z),\ X \succeq 0,\ Z \succeq 0\right\}, \quad (19)$$

where $\beta \in [0, 1)$ is the parameter to determine the area of the neighborhood. Note that whether a point $(X, Z, y)$ is in $\mathcal{N}(\beta)$ or not depends only on $X$ and $Z$.

Analogously, a neighborhood $\widetilde{\mathcal{N}}(\beta)$ of $\mathcal{D}$ is defined as

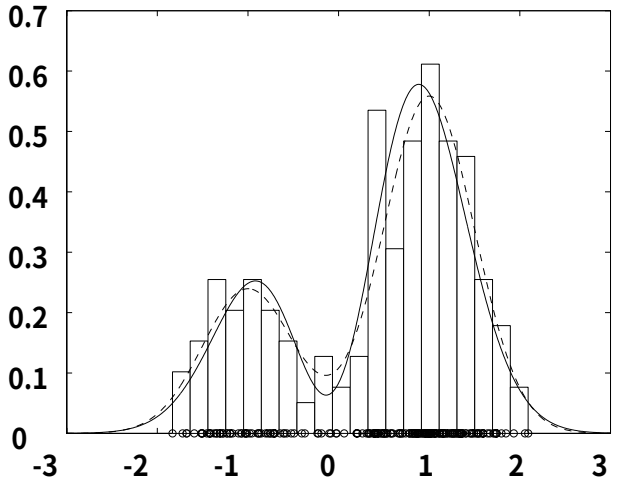
$$\begin{aligned}\widetilde{\mathcal{N}}(\beta) \ \equiv \ \Bigg\{&(X, Z, y) : \sqrt{\sum_{j\in\Omega} \|X_j^{1/2} Z_j X_j^{1/2} - I\|_F^2 + \sum_{j\notin\Omega} \|X_j^{1/2} Z_j X_j^{1/2} - \tilde{\mu}(X, Z) I\|_F^2} \\ &\leq \beta\tilde{\mu}(X, Z),\ X \succeq 0,\ Z \succeq 0\Bigg\}.\end{aligned}$$

Now, we are ready to describe the predictor-corrector algorithm. The predictor-corrector algorithm follows the central trajectory more closely. For this reason, the algorithm is more robust and stable than the basic Algorithm, capable of handling ill-conditioned problems which the basic algorithm cannot solve. Below we outline the method.

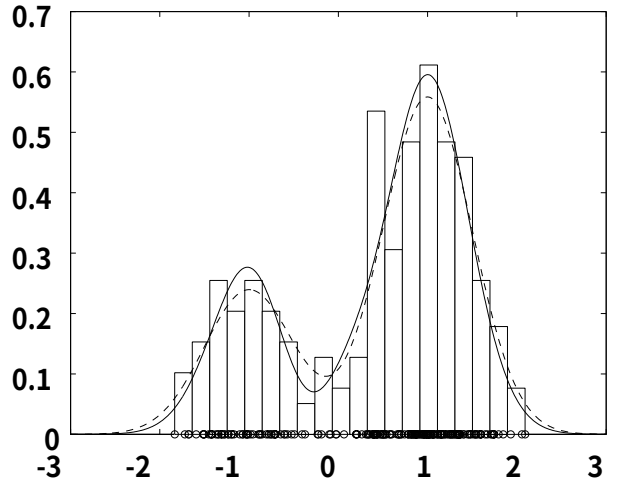
The method consists of two stages: Stage 1 starts from an initial point $(X^0, Z^0, y^0) = (\nu^0 I, \nu^0 I, 0)$ for $\nu^0 > 0$ and finds a point $W^k$, say, in a sufficiently small neighborhood of $\widehat{W}(1)$. Starting from $W^k$, Stage 2 generates a sequence approaching $\widetilde{W}(0)$.

Two neighborhoods $\mathcal{N}(\beta)$ and $\mathcal{N}(2\beta)$ are provided in Stage 1 to guide the iterates smoothly to $\widetilde{W}(1)$. One iteration of the predictor-corrector algorithm consists of the predictor step and corrector step. In the beginning of the predictor step the iterate is assumed to stay in smaller neighborhood $\mathcal{N}(\beta)$. The Newton direction for (17) with $\nu = 0$ is computed, and the largest step is taken in the direction of the Newton direction toward the boundary of $\mathcal{N}(2\beta)$. Then corrector steps is performed to bring the iterate back again to $\mathcal{N}(\beta)$. This step is designed so that the progress made in the predictor step is not lost. Repeating this procedure, the predictor-corrector algorithm generates a sequence approaching to $\widehat{W}(1)$.

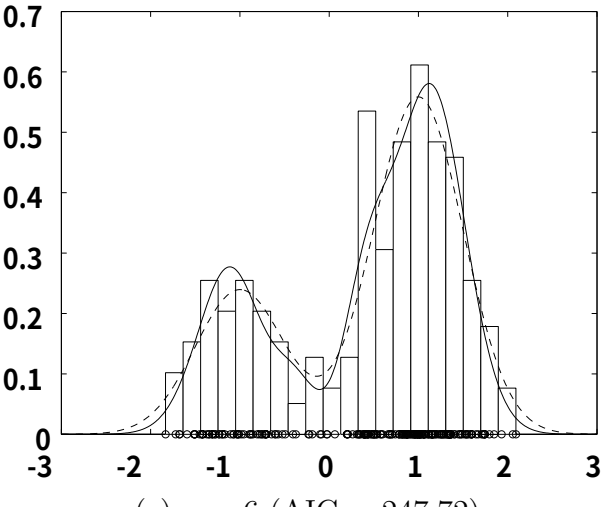
Stage 2 goes in a similar way. We use different neighborhoods $\widetilde{\mathcal{N}}(\beta)$ and $\widetilde{\mathcal{N}}(2\beta)$ for guide. In the predictor step the iterate is assumed to stay in smaller neighborhood $\widetilde{\mathcal{N}}(\beta)$. The Newton direction for (18) with $\eta = 0$ is computed, and the largest step is taken in the direction of the Newton direction toward the boundary of $\widetilde{\mathcal{N}}(2\beta)$. Then corrector steps is performed to bring the iterate back again to $\widetilde{\mathcal{N}}(\beta)$. This steps are designed so that the progress made in the predictor step is not lost. Repeating this procedure, the predictor-corrector algorithm generates a sequence approaching $\widetilde{W}(0)$.




(a) $n = 2$ (AIC = 248.19)




(b) $n = 4$ (AIC = 247.26)




(c) $n = 6$ (AIC = 247.72)

Figure 1: Estimated density function from the simulated data 1 with different degrees of polynomials (normal based model). Each data point is shown by a circle; The bars are the histogram density estimation (the number of the bins is 20); The solid line is the estimated density; The broken line is the true density.

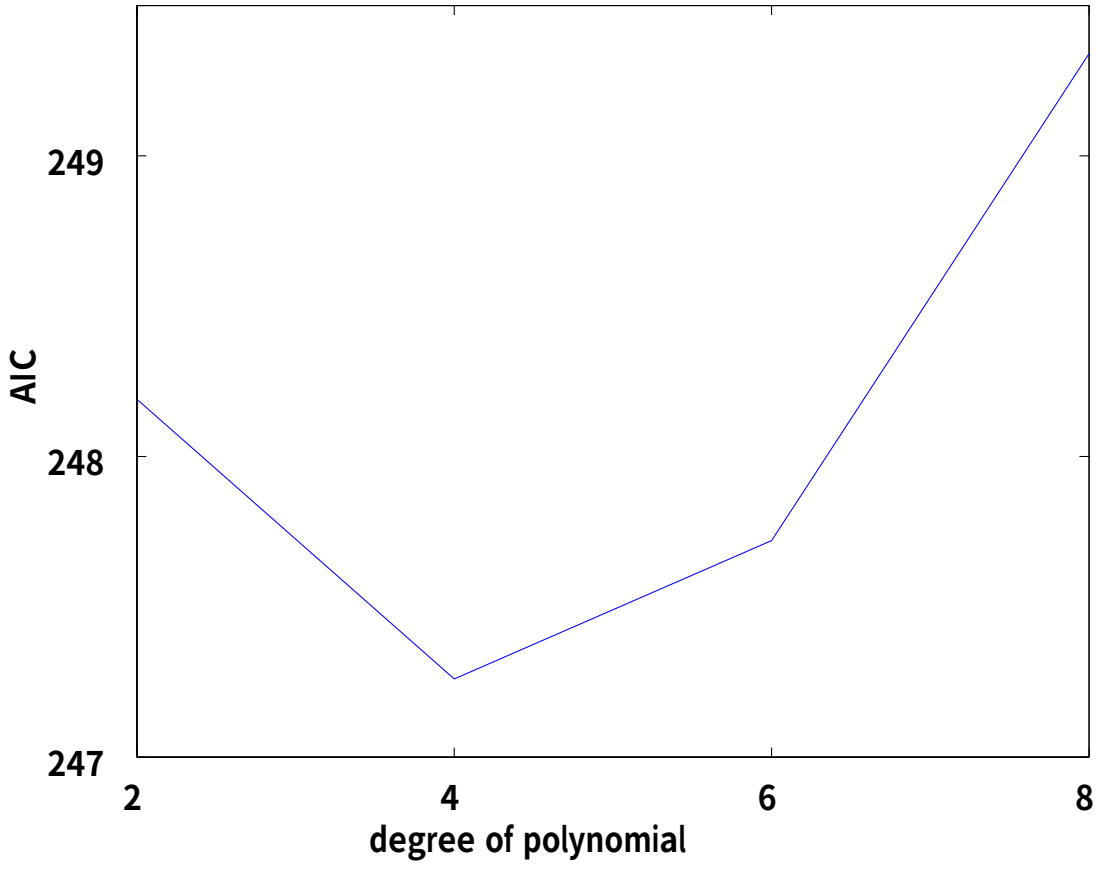



Figure 2: AIC of the estimated density function from the simulated data 1.

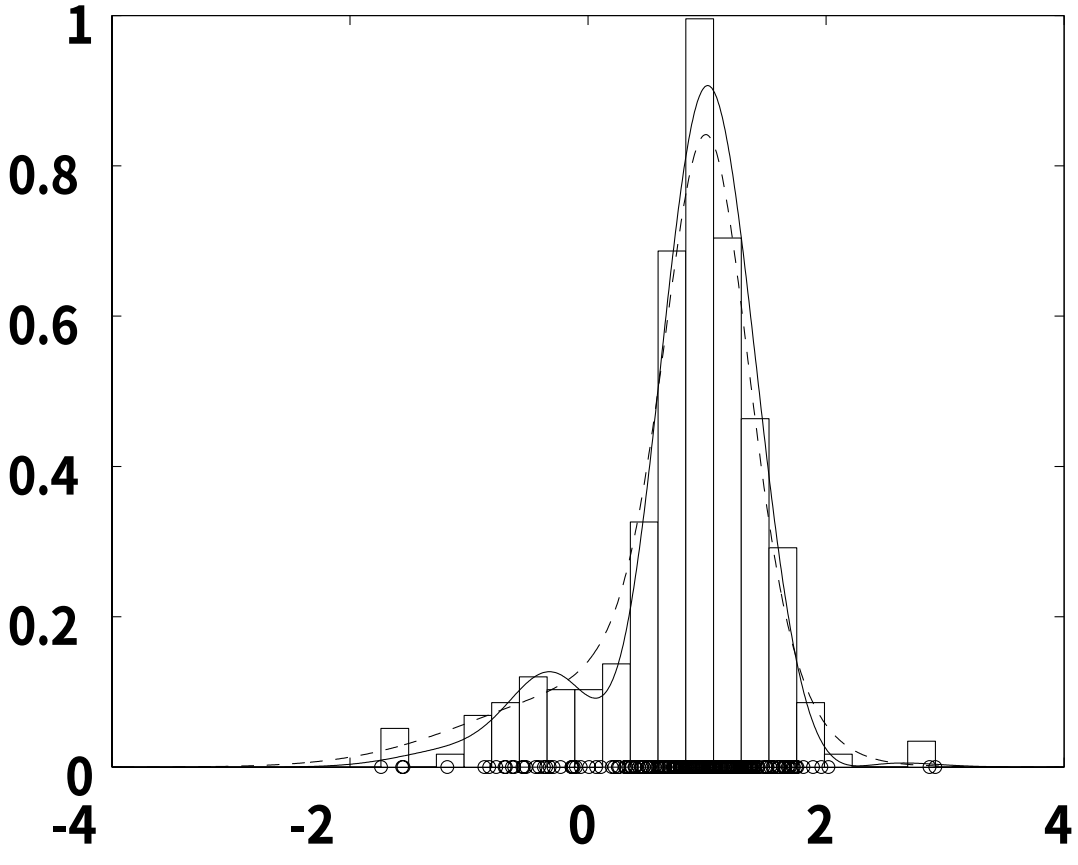
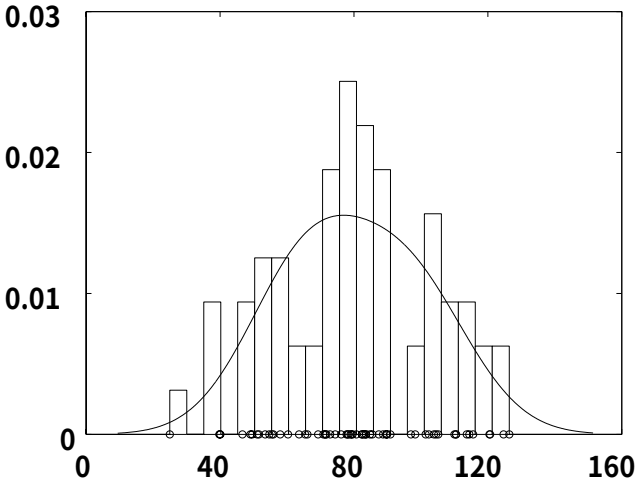
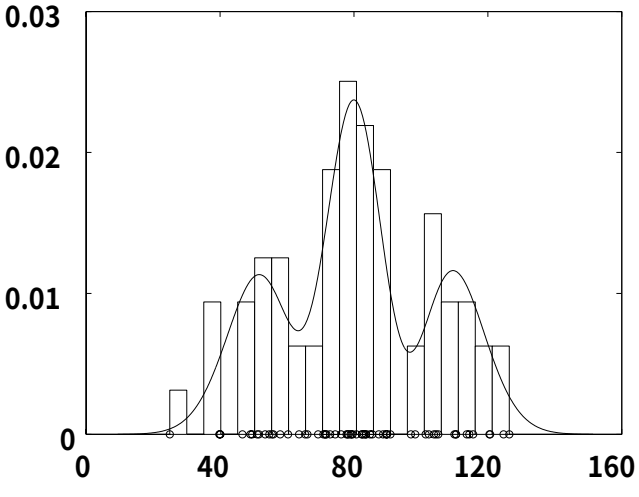


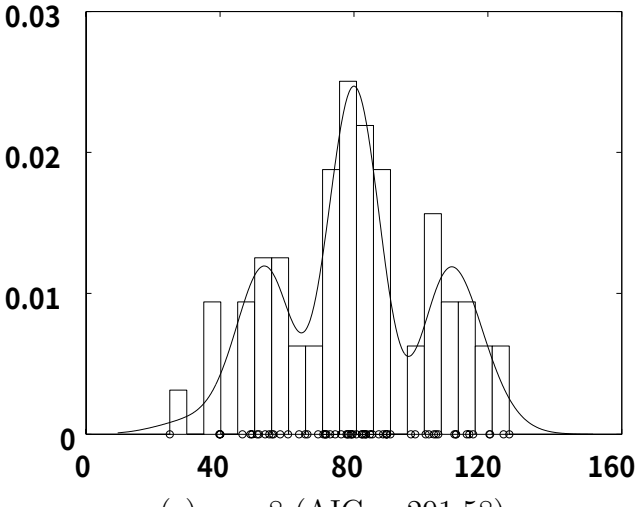

Figure 3: Estimated density function from the simulated data 2 (normal based model). Each data point is shown by a circle; The bars are the histogram density estimation (the number of the bins is 20); The solid line is the estimated density; The broken line is the true density.




(a) $n = 4$ (AIC = 291.88)




(b) $n = 6$ (AIC = 289.99)




(c) $n = 8$ (AIC = 291.58)

Figure 4: Estimated density function from the Buffalo snowfall data with different degrees of polynomials (normal based model). Each data point is shown by a circle; The bars are the histogram density estimation (the number of the bins is 20); The solid line is the estimated density.

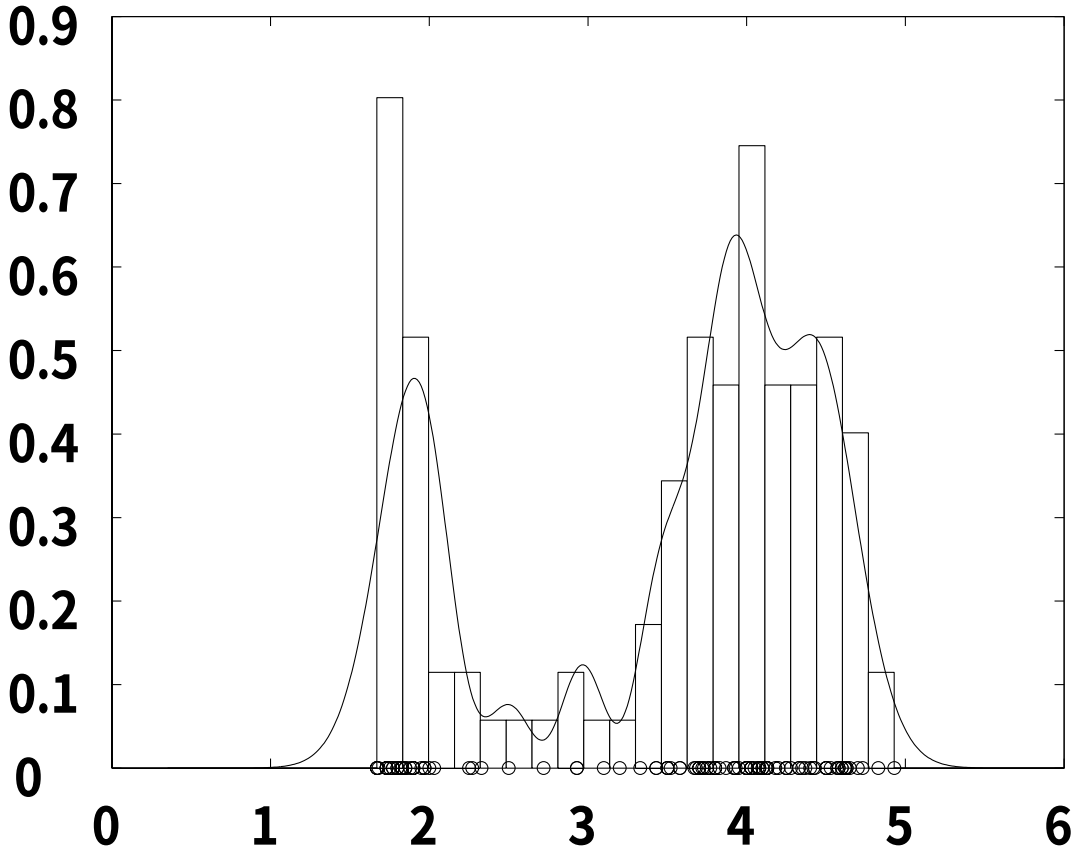



Figure 5: Estimated density function from the Old faithful geyser data (normal based model). Each data point is shown by a circle; The bars are the histogram density estimation (the number of the bins is 20); The solid line is the estimated density.

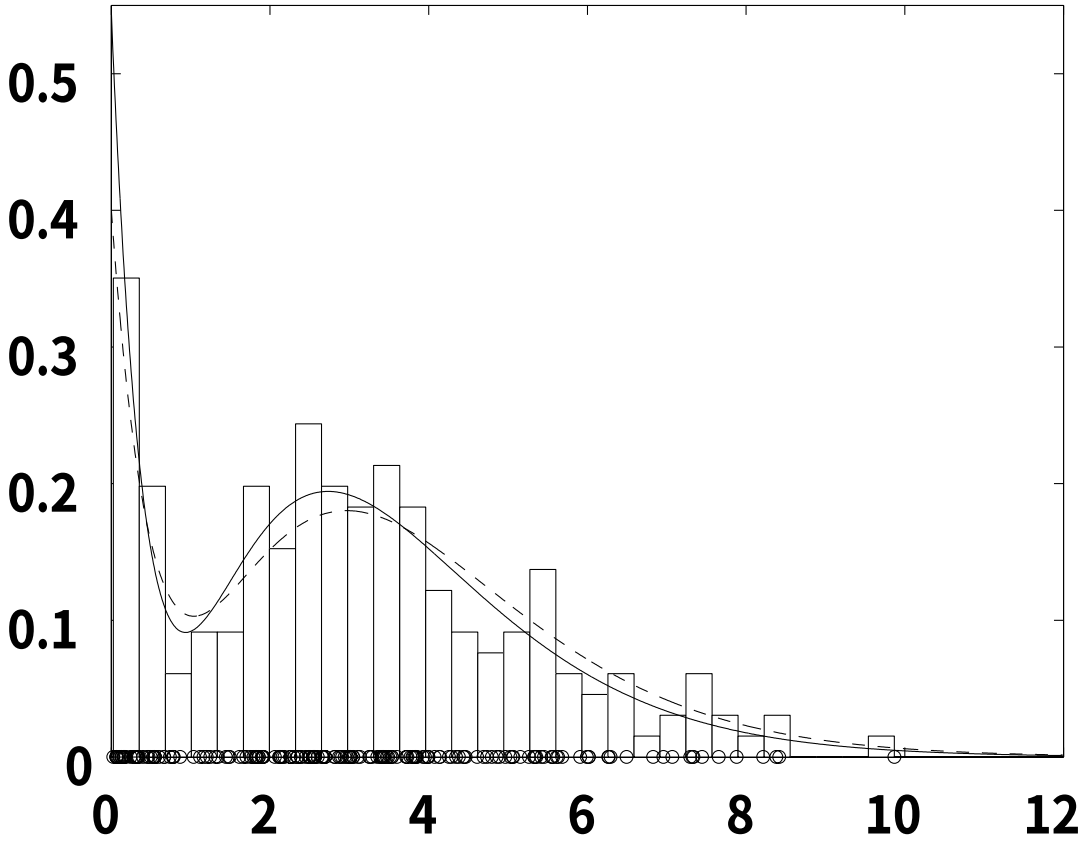



Figure 6: Estimated density function from the simulated data 3 (exponential based model). Each data point is shown by a circle; The bars are the histogram density estimation (the number of the bins is 20); The solid line is the estimated density; The broken line is the true density.

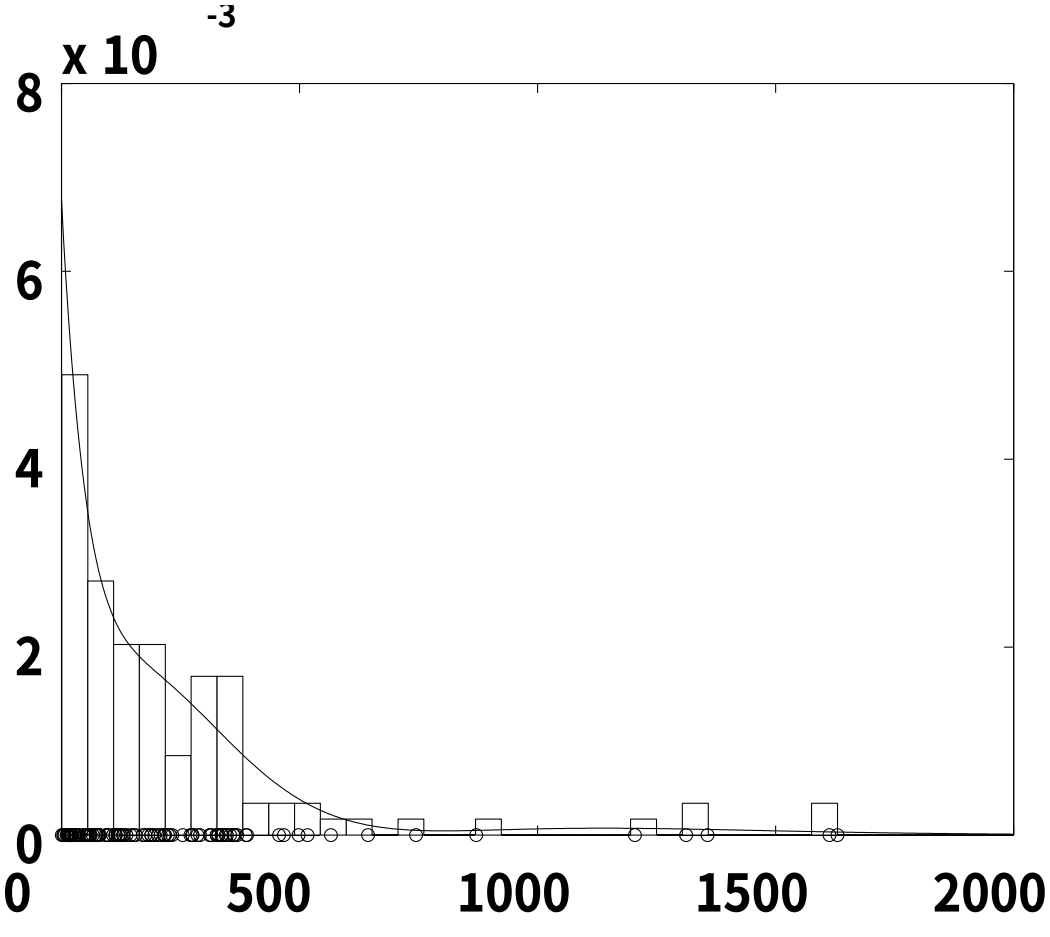



Figure 7: Estimated density function from the coal-mining disasters data (exponential based model). Each data point is shown by a circle; The bars are the histogram density estimation (the number of the bins is 20); The solid line is the estimated density.

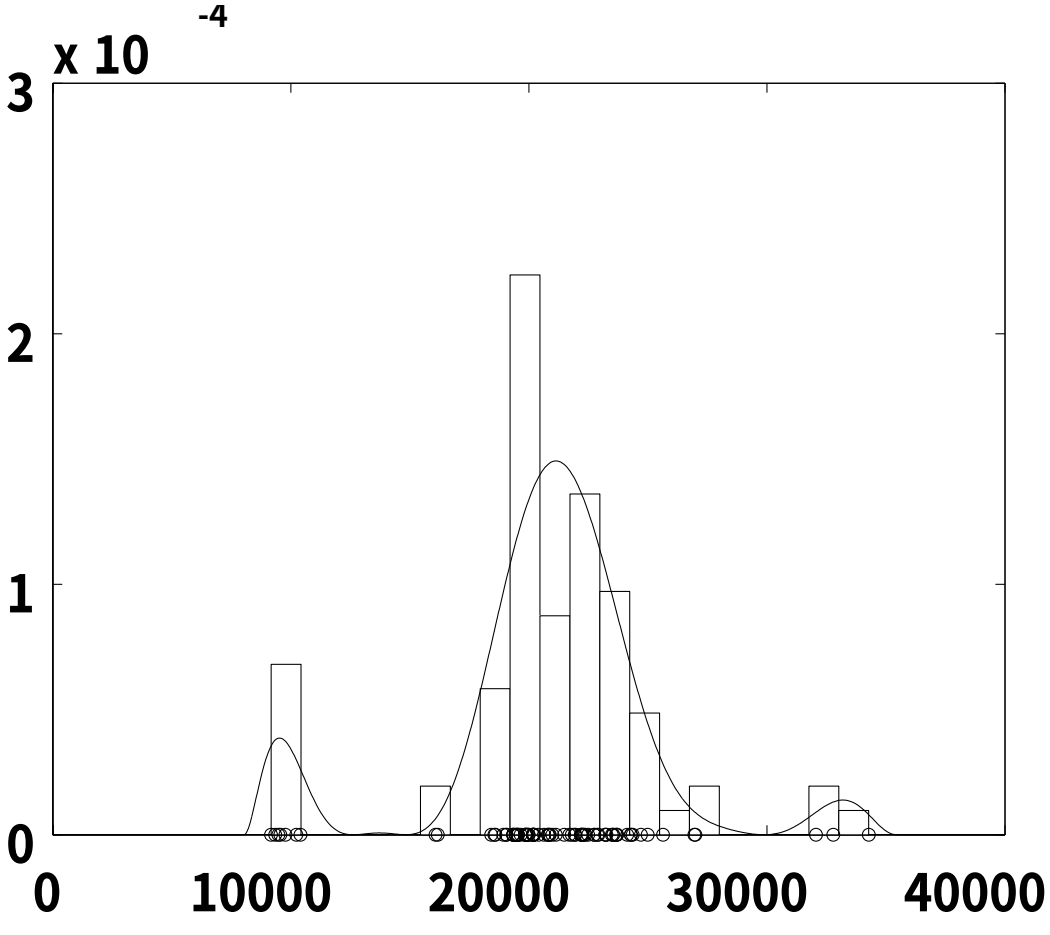



Figure 8: Estimated density function from the Galaxy data (pure polynomial model). Each data point is shown by a circle; The bars are the histogram density estimation (the number of the bins is 20); The solid line is the estimated density.

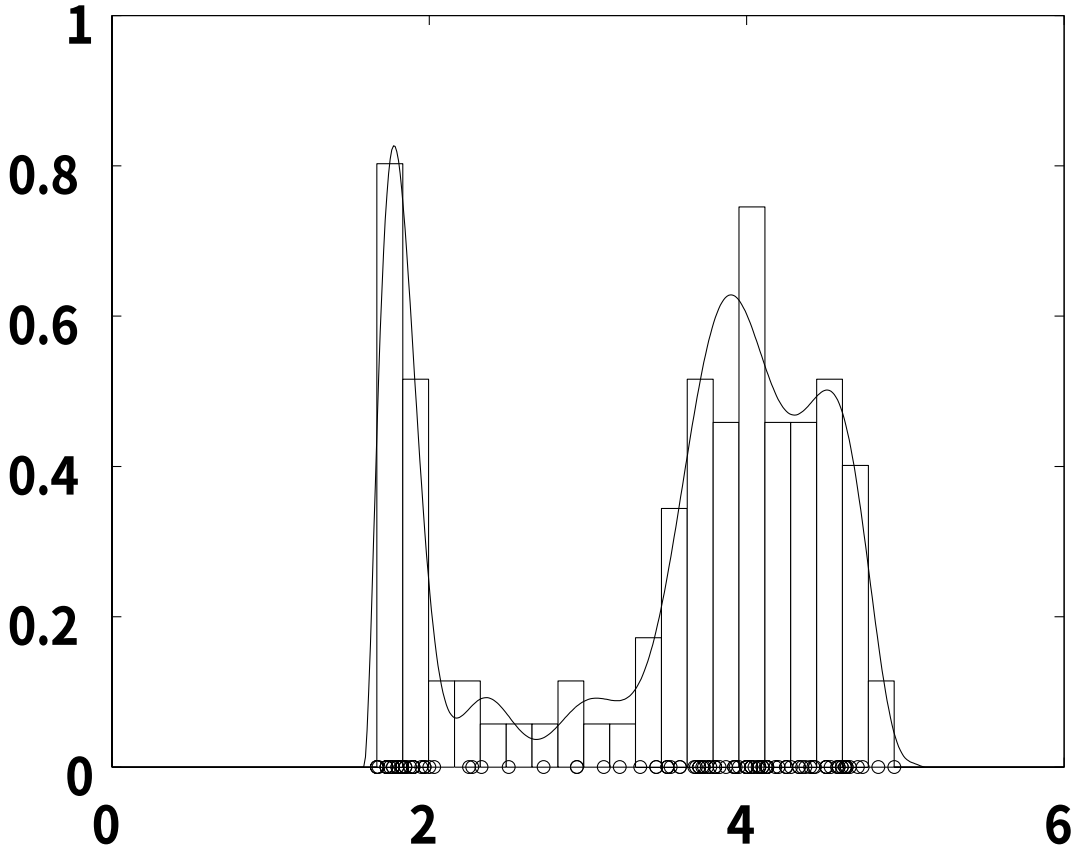



Figure 9: Estimated density function from the Old faithful geyser data (pure polynomial model). Each data point is shown by a circle; The bars are the histogram density estimation (the number of the bins is 20); The solid line is the estimated density.

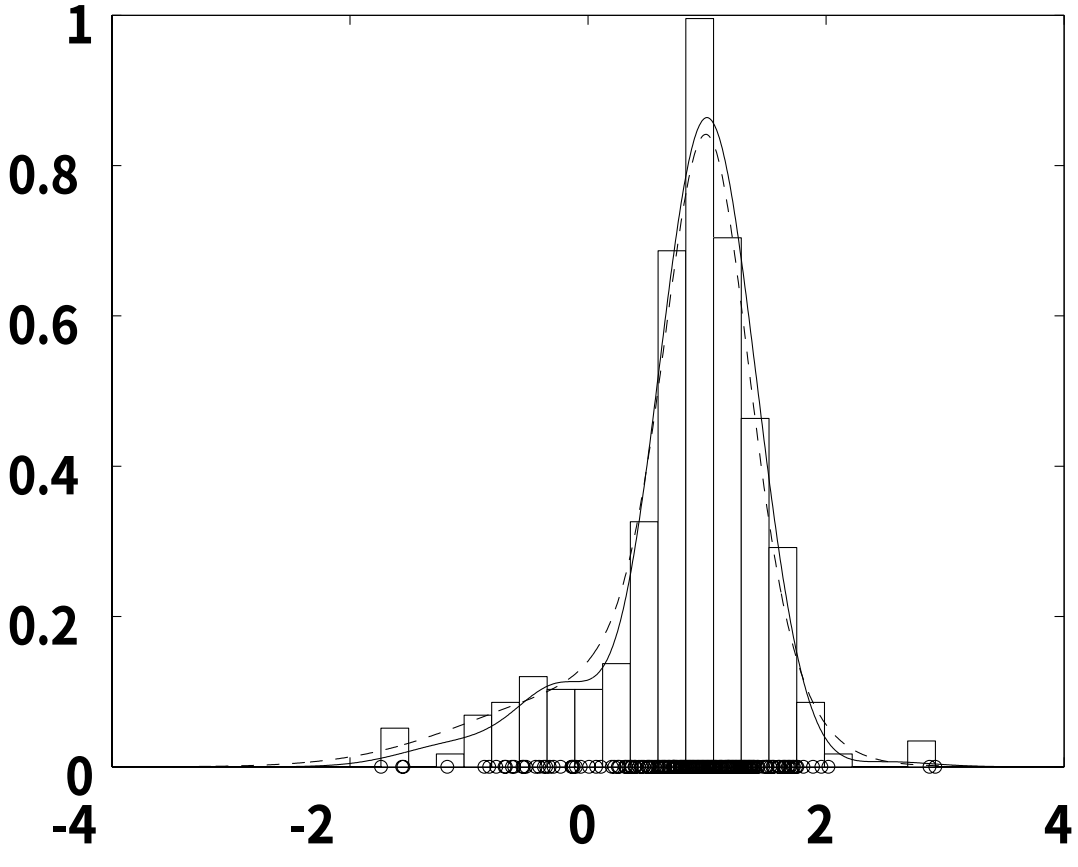



Figure 10: Estimated density function from the simulated data 2 (normal based model with unimodality constraint). Each data point is shown by a circle; The bars are the histogram density estimation (the number of the bins is 20); The solid line is the estimated density; The broken line is the true density.

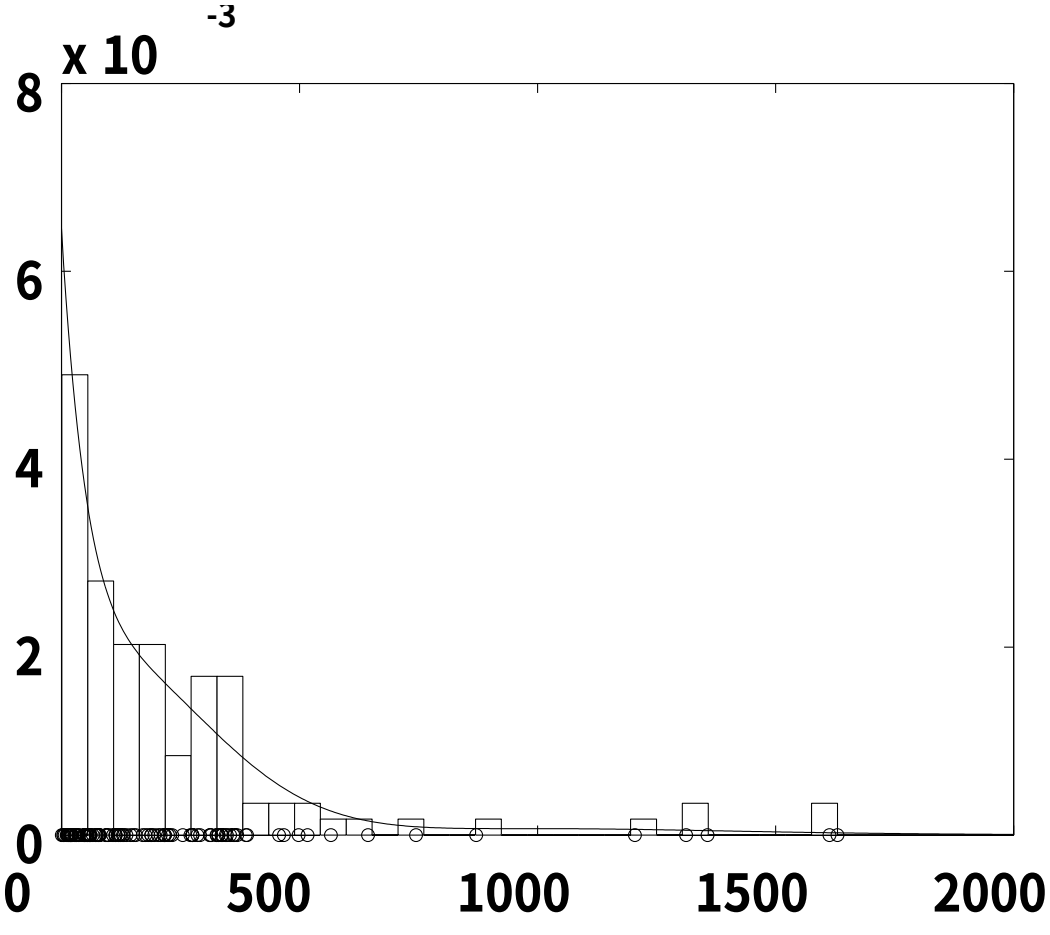



Figure 11: Estimated density function from the coal-mining disasters data (exponential based model with monotonicity). Each data point is shown by a circle; The bars are the histogram density estimation (the number of the bins is 20); The solid line is the estimated density.

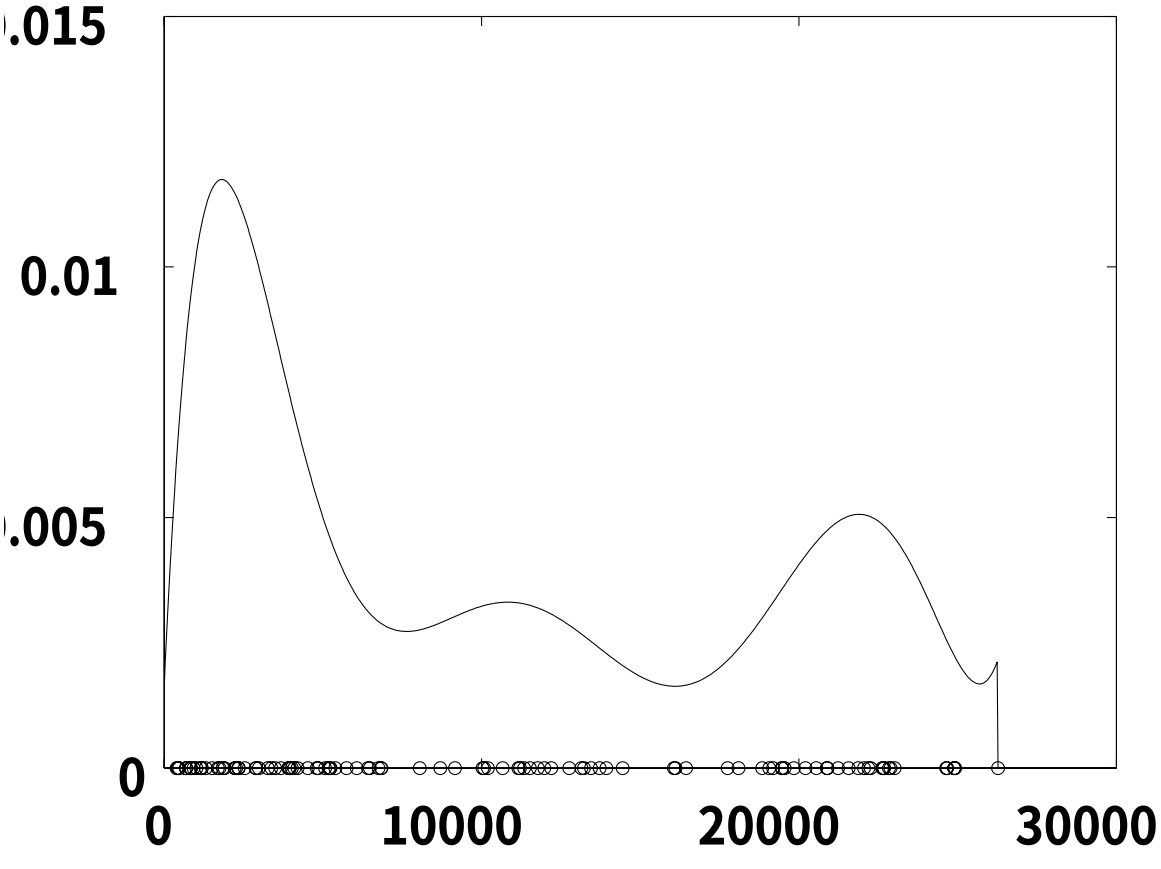



Figure 12: Estimated intensity function from the coal-mining disasters data. Each disaster is shown by a circle.