MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Estimation of Exponential-Polynomial Distribution by Holonomic Gradient Descent

Jumpei HAYAKAWA and Akimichi TAKEMURA

METR 2014–10 March 2014

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

Estimation of exponential-polynomial distribution by holonomic gradient descent

Jumpei Hayakawa* and Akimichi Takemura*

March, 2014

Abstract

We study holonomic gradient decent for maximum likelihood estimation of exponential-polynomial distribution, whose density is the exponential function of a polynomial in the random variable. We first consider the case that the support of the distribution is the set of positive reals. We show that the maximum likelihood estimate (MLE) can be easily computed by the holonomic gradient descent, even though the normalizing constant of this family does not have a closed-form expression and discuss determination of the degree of the polynomial based on the score test statistic. Then we present extensions to the whole real line and to the bivariate distribution on the positive orthant.

Keywords and phrases: algebraic statistics, bivariate distribution, score test.

1 Introduction

Exponential distribution and the truncated normal distribution have been frequently used for positive continuous random variables (e.g., Chapter 19 and Section 13.10 of [7], [13]). Generalizing these two cases, in this paper we consider fitting a density function which is the exponential function of a polynomial in the random variable. For simplicity we first study the case of a positive random variable. For $x > 0$, consider the following density

$$f(x; \theta_1, \dots, \theta_d) = \frac{1}{A(\theta_1, \dots, \theta_d)} \exp(\theta_1 x + \cdots + \theta_d x^d), \qquad \theta_d < 0, \tag{1}$$

where

$$A(\theta_1, \dots, \theta_d) = \int_0^\infty \exp(\theta_1 x + \cdots + \theta_d x^d) \mathrm{d}x \tag{2}$$

is the normalizing constant of this density. In the following we write $A_d(\boldsymbol{\theta}) = A(\theta_1, \dots, \theta_d)$. We call (1) the *exponential-polynomial distribution of order d.* Although it is a natural generalization of the exponential ($d = 1$) and the truncated normal distribution ($d = 2$), it has been

*Graduate School of Information Science and Technology, University of Tokyo

rarely used in statistics. One reason is that $A_d(\boldsymbol{\theta})$ can not be written in a closed form. Another reason may be that the tail of the distribution is light because of the term $\theta_d x^d$, $\theta_d < 0$. However by having this term, we can allow arbitrary values of $\theta_1, \dots, \theta_{d-1}$ and have a flexible family of distributions.

Concerning the treatment of the normalizing constant, recently in [11] we proposed a new method, called the holonomic gradient decent (HGD), for evaluating the normalizing constant of the exponential family and for computing MLE. As in the subsequent works ([5], [12]), we show that HGD works well also for the case of exponential-polynomial distribution.

When we fit (1) to a given sample, the natural question we face is the determination of the order d of the model. The exponential-polynomial model has a special structure that the model of order $d-1$ with $\theta_d = 0$ and $\theta_{d-1} < 0$ is the boundary of the model of order d with $\theta_d < 0$. Hence we need to adapt standard model selection procedures to this non-regular case. We propose selection of d by a score test.

The organization of this paper is as follows. In Sections 2–4 we study exponential-polynomial distribution over the set of positive reals. In Section 2 we derive a differential equation satisfied by $A_d(\boldsymbol{\theta})$ and use the differential equation to compute MLE. In Section 3 we discuss how to determine the order d of the model by a score test. In Section 4 we present results of some numerical experiments. In Section 5 we extend the exponential-polynomial distribution to the whole real line and in Section 6 we study a bivariate exponential-polynomial distribution. We end the paper with some discussions on further extension of the model in Section 7.

2 Maximum likelihood estimation via holonomic gradient descent

Given a sample $\boldsymbol{x} = (x_1, \dots, x_n)$ of size n, $(1/n)$ times the log-likelihood function is written as

$$\bar{l}(\boldsymbol{\theta}; \boldsymbol{x}) = \theta_1 \bar{x} + \theta_2 \overline{x^2} + \cdots + \theta_d \overline{x^d} - \psi(\boldsymbol{\theta}), \qquad \psi(\boldsymbol{\theta}) = \log A_d(\boldsymbol{\theta}), \tag{3}$$

where $\overline{x^m} = \sum_{i=1}^n x_i^m / n$, $m = 1, \dots, d$. Let $\partial_m = \frac{\partial}{\partial \theta_m}$ denote the differentiation with respect to θ_m. In maximizing $\bar{l}$ with respect to $\boldsymbol{\theta}$, we want to compute its gradient

$$\nabla \bar{l} = \begin{bmatrix} \partial_1 \bar{l} \\ \vdots \\ \partial_d \bar{l} \end{bmatrix} = \begin{bmatrix} \bar{x} \\ \vdots \\ \overline{x^d} \end{bmatrix} - \begin{bmatrix} \partial_1 \psi \\ \vdots \\ \partial_d \psi \end{bmatrix}, \qquad \partial_m \psi(\boldsymbol{\theta}) = \frac{\partial_m A_d(\boldsymbol{\theta})}{A_d(\boldsymbol{\theta})}$$

and its Hessian matrix

$$H(\bar{l})(\boldsymbol{\theta}) = -H(\psi)(\boldsymbol{\theta}) = -\begin{bmatrix} \partial_1^2 \psi & \cdots & \partial_1 \partial_d \psi \\ \vdots & \dots & \vdots \\ \partial_d \partial_1 \psi & \cdots & \partial_d^2 \psi \end{bmatrix}, \qquad \partial_l \partial_m \psi(\boldsymbol{\theta}) = \frac{\partial_l \partial_m A_d(\boldsymbol{\theta})}{A_d(\boldsymbol{\theta})} - \frac{\partial_l A_d(\boldsymbol{\theta})}{A_d(\boldsymbol{\theta})} \frac{\partial_m A_d(\boldsymbol{\theta})}{A_d(\boldsymbol{\theta})}.$$

Note that $I(\boldsymbol{\theta}) = H(\psi)(\boldsymbol{\theta})$ is the Fisher information matrix for $\boldsymbol{\theta}$.

In (2) we can interchange the integration and the differentiation by elements of $\boldsymbol{\theta}$ as many time as needed. Hence derivatives of $A_d(\boldsymbol{\theta})$ can be evaluated by numerical integration. However it is cumbersome to perform numerical integration for the derivatives at every $\boldsymbol{\theta}$. The holonomic gradient decent allows us to compute $A_d(\boldsymbol{\theta})$ and its derivatives at any point by numerically solving a differential equation from those at an initial point $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. The fact that $A_d(\boldsymbol{\theta})$ is a holonomic function (cf. Section 1 and Appendix of [11], Chapter 6 of [6], [14]) guarantees the existence of a differential equation with polynomial coefficients satisfied by $A_d(\boldsymbol{\theta})$. Also, for our problem there is a convenient initial point (see (8) below), where $A_d(\boldsymbol{\theta})$ and its derivatives have a closed-form expression. Hence by using the holonomic gradient descent, we do not need any numerical integration for our problem.

Differentiating (2) by θ_1 we have

$$\partial_1 A_d(\boldsymbol{\theta}) = \int_0^\infty x \exp(\theta_1 x + \cdots + \theta_d x^d) \mathrm{d}x.$$

Repeating this i times we have

$$\partial_1^i A_d(\boldsymbol{\theta}) = \int_0^\infty x^i \exp(\theta_1 x + \cdots + \theta_d x^d) \mathrm{d}x. \tag{4}$$

However the right-hand side is also equal to $\partial_i A(\boldsymbol{\theta})$. Hence the following relation holds.

$$\partial_i A_d(\boldsymbol{\theta}) = \partial_1^i A_d(\boldsymbol{\theta}).$$

In general, for any higher-order mixed derivative $\partial_1^{j_1} \ldots \partial_d^{j_d} A(\boldsymbol{\theta})$ we have the relation

$$\partial_1^{j_1} \ldots \partial_d^{j_d} A_d(\boldsymbol{\theta}) = \partial_1^{j_1+2j_2+\cdots+dj_d} A_d(\boldsymbol{\theta}).$$

Hence all mixed derivatives reduce to the derivatives of $A_d(\boldsymbol{\theta})$ with respect to θ_1. It follows that for numerical purposes we only need to keep in memory the derivatives of $A_d(\boldsymbol{\theta})$ with respect to θ_1.

Now as a relation among the derivatives of $A_d(\boldsymbol{\theta})$ with respect to θ_1, we have the following theorem.

Theorem 2.1. *$A_d(\boldsymbol{\theta})$ satisfies the following differential equation*

$$(\theta_1 + 2\theta_2\partial_1 + 3\theta_3\partial_1^2 + \cdots + d\theta_d\partial_1^{d-1})A_d(\boldsymbol{\theta}) = -1. \tag{5}$$

Proof.

$$\begin{aligned}
-1 &= [\exp(\theta_1 x + \cdots + \theta_d x^d)]_0^\infty \\
&= \int_0^\infty \partial_x \exp(\theta_1 x + \cdots + \theta_d x^d)\, \mathrm{d}x \\
&= \int_0^\infty (\theta_1 + 2\theta_2 x + 3\theta_3 x^2 + \cdots + d\theta_d x^{d-1}) \exp(\theta_1 x + \cdots + \theta_d x^d) \mathrm{d}x \\
&= (\theta_1 + 2\theta_2\partial_1 + 3\theta_3\partial_1^2 + \cdots + d\theta_d\partial_1^{d-1})A_d(\boldsymbol{\theta}). \qquad \text{(by (4))}
\end{aligned}$$

□

By (5), $\partial_1^{d-1}A_d(\boldsymbol{\theta})$ is written in terms of lower-order derivatives as

$$\partial_1^{d-1}A_d(\boldsymbol{\theta}) = -\frac{1}{d\theta_d}(1 + \theta_1 + 2\theta_2\partial_1 + 3\theta_3\partial_1^2 + \cdots + (d-1)\theta_{d-1}\partial_1^{d-2})A_d(\boldsymbol{\theta}). \qquad (6)$$

Recursively differentiating this by θ_1 we see that all higher-order derivatives $\partial_1^m A_d(\boldsymbol{\theta})$, $m \geq d-1$, can be written in terms of the elements of a vector

$$F(\boldsymbol{\theta}) = [A_d(\boldsymbol{\theta}), \partial_1 A_d(\boldsymbol{\theta}), \ldots, \partial_1^{d-2}A_d(\boldsymbol{\theta})]^{\mathsf{T}},$$

where $^{\mathsf{T}}$ denotes the transpose of a vector or a matrix. If $F(\boldsymbol{\theta})$ can be evaluated at any point $\boldsymbol{\theta}$, we can compute MLE of the exponential-polynomial distribution.

Note that the directional derivative of $F(\boldsymbol{\theta})$ in the direction $\boldsymbol{h} = (h_1, \ldots, h_d)$ is written as

$$\frac{\partial}{\partial s}F(\boldsymbol{\theta} + s\boldsymbol{h}) = \sum_{j=1}^{d} h_j \partial_j F(\boldsymbol{\theta} + s\boldsymbol{h}) = \sum_{j=1}^{d} h_j \partial_1^j F(\boldsymbol{\theta} + s\boldsymbol{h}) = \sum_{j=1}^{d} h_j \begin{bmatrix} \partial_1^j A_d(\boldsymbol{\theta} + s\boldsymbol{h}) \\ \partial_1^{j+1} A_d(\boldsymbol{\theta} + s\boldsymbol{h}) \\ \vdots \\ \partial_1^{j+d-2} A_d(\boldsymbol{\theta} + s\boldsymbol{h}) \end{bmatrix}. \qquad (7)$$

When an appropriate initial point $\boldsymbol{\theta}_0$ and $F(\boldsymbol{\theta}_0)$ are given, (7) can be solved by standard solver for ordinary differential equation, such as the Runge-Kutta method.

As a convenient initial point consider $\boldsymbol{\theta}^0 = (0, 0, \ldots, 0, -c)$, $c > 0$. Then

$$\partial_1^m A_d(\boldsymbol{\theta}^0) = \int_0^\infty x^m \exp(-cx^d)\mathrm{d}x = \frac{1}{d}c^{-(1+m)/d}\Gamma(\frac{1+m}{d}), \qquad m \geq 0, \qquad (8)$$

which do not need numerical integration.

In summary, we have shown that the evaluation of $A_d(\boldsymbol{\theta})$ and the maximization of the likelihood function can be performed by using only a standard solver for an ordinary differential equation. As we see in Section 4 this method works quite well in practice.

3 Determination of the degree of the model

When we fit the exponential-polynomial distribution in (1) to a given sample, we need to determine the order d of the model. Suppose that we are fitting the model with order $d-1$ and wondering whether a model of order d fits better. One difficulty with (6) is that it becomes unstable as $\theta_d \to 0$, i.e., the differential equation (5) has a singularity at $\theta_d = 0$. Hence if the data really come from the model of order $d-1$, the estimation of the model of order d by our method tends to be unstable.

We can understand this problem by considering the parameter spaces of order $d-1$ and d. Let $\Omega_d = \{(\theta_1, \ldots, \theta_d) \mid \theta_d < 0\} \subset \mathbb{R}^d$ denote the parameter space of the model of order d. Ω_d is an open subset of $\mathbb{R}^d$. Now $\Omega_{d-1} = \{(\theta_1, \ldots, \theta_{d-1}, 0) \mid \theta_{d-1} < 0\}$ considered as a subset of $\mathbb{R}^d$ is on the boundary of Ω_d. See Figure 1. In (θ_{d-1}, θ_d)-plane, Ω_d is the lower half open plane and Ω_{d-1} the left half open θ_{d-1}-axis $\{(\theta_{d-1}, 0) \mid \theta_{d-1} < 0\}$. Since $A_d(\theta_1, \ldots, \theta_{d-1}, 0)$ is finite for

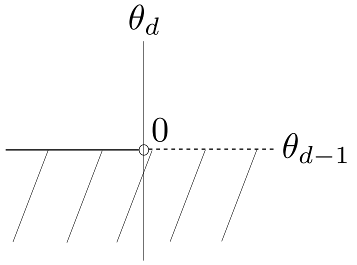

Figure 1: Model of order $d-1$ within the model of order d

$\theta_{d-1} < 0$, MLE may not exist in the open set Ω_d with positive probability. For each d, consider $\Omega_1, \ldots, \Omega_d$ as subsets of $\mathbb{R}^d$ and let $\bar{\Omega}_d = \Omega_1 \cup \cdots \cup \Omega_d$. Then $\psi_d(\boldsymbol{\theta}) = A_d(\boldsymbol{\theta})$ is strictly convex on $\bar{\Omega}_d$ and approaches $+\infty$ as $\boldsymbol{\theta}$ approaches the open boundary of $\bar{\Omega}_d$, such as the right half open θ_{d-1}-axis $\{(\theta_{d-1}, 0) \mid \theta_{d-1} > 0\}$ in Figure 1. Hence MLE always exists in $\bar{\Omega}_d$ but may not fall on Ω_d.

We now consider the hypothesis testing problem:

$$H_0 : \boldsymbol{\theta} \in \Omega_{d-1} \ \text{ v.s. } \ H_1 : \boldsymbol{\theta} \in \Omega_d. \tag{9}$$

If H_0 is true let $\boldsymbol{\theta}^* \in \Omega_{d-1}$ denote the true parameter vector and let $\hat{\boldsymbol{\theta}}_{d-1} = (\hat{\theta}_1, \ldots, \hat{\theta}_{d-1}, 0)$, $\hat{\theta}_{d-1} < 0$, denote the MLE under H_0. Then $\hat{\boldsymbol{\theta}}_{d-1}$ converges to $\boldsymbol{\theta}^*$ in probability.

The MLE $\hat{\boldsymbol{\theta}}_{d-1}$ under H_0 satisfies

$$\partial_j \bar{l}(\hat{\boldsymbol{\theta}}_{d-1}; \boldsymbol{x}) = 0, \qquad j = 1, \ldots, d-1.$$

Note that $\bar{l}(\hat{\boldsymbol{\theta}}_{d-1} + s\boldsymbol{h}; \boldsymbol{x})$ is strictly concave in $s \geq 0$ for any $\boldsymbol{h} = (h_1, \ldots, h_d)$, $h_d < 0$, i.e., on any half line emanating from $\hat{\boldsymbol{\theta}}_{d-1}$ into Ω_d. Hence on this half line, $\bar{l}(\hat{\boldsymbol{\theta}}_{d-1}; \boldsymbol{x})$ is maximized at $s = 0$ if and only if

$$0 \geq \frac{\partial}{\partial s} \bar{l}(\hat{\boldsymbol{\theta}}_{d-1} + s\boldsymbol{h}; \boldsymbol{x})|_{s=0} = \sum_{j=1}^{d} h_j \partial_j \bar{l}(\hat{\boldsymbol{\theta}}_{d-1}; \boldsymbol{x}) = h_d \partial_d \bar{l}(\hat{\boldsymbol{\theta}}_{d-1}; \boldsymbol{x}) \Leftrightarrow \ \partial_d \bar{l}(\hat{\boldsymbol{\theta}}_{d-1}; \boldsymbol{x}) \geq 0.$$

Note that the right-hand side does not depend on $\boldsymbol{h}$. Hence MLE does not exist on Ω_d and $\hat{\boldsymbol{\theta}}_{d-1}$ remains to be the MLE over $\bar{\Omega}_d$ if and only if $\partial_d \bar{l}(\hat{\boldsymbol{\theta}}_{d-1}; \boldsymbol{x}) \geq 0$.

Let the $d \times d$ Fisher information matrix $I(\boldsymbol{\theta}) = H(\psi)(\boldsymbol{\theta})$ be partitioned as

$$I(\boldsymbol{\theta}) = \begin{bmatrix} I_{d-1,d-1}(\boldsymbol{\theta}) & I_{d-1,d}(\boldsymbol{\theta}) \\ I_{d,d-1}(\boldsymbol{\theta}) & I_{dd}(\boldsymbol{\theta}) \end{bmatrix},$$

where I_{dd} is a scalar. Note that we put a comma between two subscripts when the subscripts are more complicated. Define

$$I_{dd\cdot 1,\ldots,d-1}(\boldsymbol{\theta}) = I_{dd}(\boldsymbol{\theta}) - I_{d,d-1}(\boldsymbol{\theta}) I_{d-1,d-1}(\boldsymbol{\theta})^{-1} I_{d-1,d}(\boldsymbol{\theta}).$$

In the standard case, where Ω_{d-1} is in the interior of Ω_d, the two-sided test based on

$$\frac{n(\partial_d \bar{l}(\hat{\boldsymbol{\theta}}_{d-1};\boldsymbol{x}))^2}{I_{dd\cdot 1,\ldots,d-1}(\hat{\boldsymbol{\theta}}_{d-1})}$$

is the score test for (9) (e.g., Section 7.7 of [10]). In our case Ω_{d-1} is the boundary of Ω_d and we reject H_0 if $\partial_d \bar{l}(\hat{\boldsymbol{\theta}}_{d-1};\boldsymbol{x})$ is negative and its absolute value is too large. However, from the form of the log-likelihood function in (3), the asymptotic null distribution $\partial_d \bar{l}(\hat{\boldsymbol{\theta}}_{d-1};\boldsymbol{x})$ is the same as in the standard case, i.e.,

$$\sqrt{n}\partial_d \bar{l}(\hat{\boldsymbol{\theta}}_{d-1};\boldsymbol{x}) \xrightarrow{d} \mathrm{N}(0, I_{dd\cdot 1,\ldots,d-1}(\boldsymbol{\theta}^*)) \qquad (n \to \infty).$$

Since $\hat{\boldsymbol{\theta}}_{d-1}$ converges to $\boldsymbol{\theta}^*$, we propose the following score test statistic

$$T_{d-1} = \frac{\sqrt{n}\partial_d \bar{l}(\hat{\boldsymbol{\theta}}_{d-1};\boldsymbol{x})}{\sqrt{I_{dd\cdot 1,\ldots,d-1}(\hat{\boldsymbol{\theta}}_{d-1})}}. \tag{10}$$

Let z_α denote the upper α quantile of N(0, 1). Given a significance level $\alpha < 1/2$, we can reject H_0 if $T_{d-1} \le -z_\alpha$, in view of the convergence in distribution

$$T_{d-1} \xrightarrow{d} \mathrm{N}(0,1) \qquad (n \to \infty). \tag{11}$$

4 Numerical experiments for the case of positive real line

We present results of some numerical experiments to show that MLE by HGD works well. We also check the asymptotic approximation in (11).

4.1 Performance of MLE by the holonomic gradient descent

The asymptotic distribution of MLE $\hat{\boldsymbol{\theta}}_d$ is

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_d - \boldsymbol{\theta}^*) \xrightarrow{d} \mathrm{N}_d(0, I(\boldsymbol{\theta}^*)^{-1}) \qquad (n \to \infty),$$

where $I(\boldsymbol{\theta}^*)$ is the Fisher information matrix at the true parameter $\boldsymbol{\theta}^*$. Write

$$p_i = \frac{\sqrt{n}(\hat{\theta}_i - \theta_i^*)}{\sqrt{I_{ii}^{-1}(\boldsymbol{\theta}^*)}}, \qquad i = 1, 2, \ldots, d, \tag{12}$$

where $I_{ii}^{-1}(\boldsymbol{\theta}^*)$ denotes the (i,i)-component of $I(\boldsymbol{\theta}^*)^{-1}$. Then

$$p_i \xrightarrow{d} \mathrm{N}(0,1) \qquad (n \to \infty). \tag{13}$$

Thus in our experiments we fix the true parameter $\boldsymbol{\theta}^*$, apply our method to simulated samples many times and we check the convergence of the empirical distribution of p_i to N(0, 1).

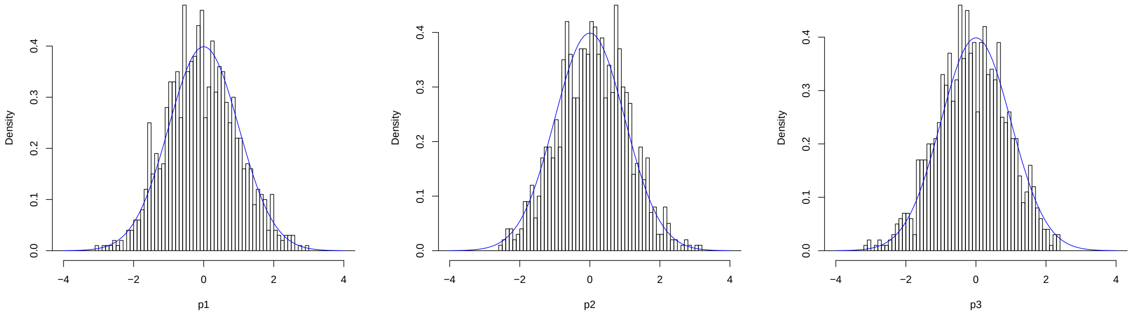

Figure 2: Histogram of p_i, $i = 1, 2, 3$ (from left to right) and density of N(0,1) for $d = 3$

We present simulation results for $d = 3$ in (1). We set $\boldsymbol{\theta}^* = (-1, 3, -2)$. In the experiment we used $n = 1000$ and iterated computing MLE 1000 times (i.e. the replication size is 1000). Computation of MLE quickly converged in each iteration. The histogram of p_i is given in Figure 2. The curved lines in these figures are the density function of $\mathrm{N}(0, 1)$. By comparing the histogram and the curved line we see that MLE by HGD works well.

4.2 Asymptotic approximation for score tests

We check the asymptotic approximation in (11) in the case of $d = 3, 4$. For $d = 3$ we set $\boldsymbol{\theta}^* = (3, -2, 0)$. The histograms of T_2 and T_3 are shown in Figure 3 (left to right). Again the asymptotic approximation works as expected.

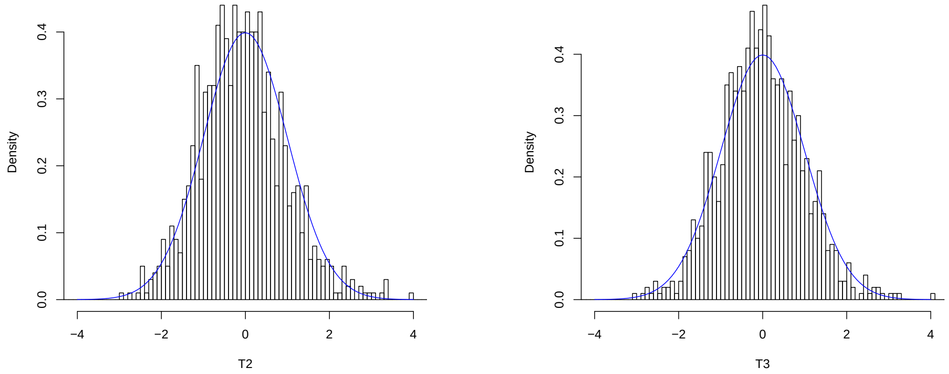

Figure 3: Histogram of T_{d-1} and the density of $\mathrm{N}(0, 1)$ for $d = 3, 4$

5 Exponential-polynomial distribution on the whole real line

In this section we extend the result of previous sections to the following density for the whole real line $\mathbb{R}^1$. Consider the density function

$$f(x;\theta_1,\dots,\theta_{2d}) = \frac{1}{A(\theta_1,\dots,\theta_{2d})}\exp(\theta_1 x + \cdots + \theta_{2d}x^{2d}), \qquad \theta_{2d} < 0, \tag{14}$$

where

$$A(\theta_1,\dots,\theta_{2d}) = \int_{-\infty}^{\infty} \exp(\theta_1 x + \cdots + \theta_{2d}x^{2d})\mathrm{d}x \tag{15}$$

is the normalizing constant of this density. In following we write $A_{2d}(\boldsymbol{\theta}) = A(\theta_1,\dots,\theta_{2d})$.

5.1 Maximum likelihood estimation for the whole line

The holonomic gradient decent is almost the same as in the previous sections. We have

$$\partial_1^i A_{2d}(\boldsymbol{\theta}) = \int_0^{\infty} x^i \exp(\theta_1 x + \cdots + \theta_{2d}x^{2d})\mathrm{d}x, \quad i = 1,2,\dots.$$

Also $\partial_i A_{2d}(\boldsymbol{\theta}) = \partial_1^i A_{2d}(\boldsymbol{\theta})$. In general $\partial_1^{j_1}\dots\partial_d^{j_d}A_{2d}(\boldsymbol{\theta}) = \partial_1^{j_1+2j_2+\dots+dj_d}A_{2d}(\boldsymbol{\theta})$. Hence all mixed derivatives reduce to the derivatives of $A_{2d}(\boldsymbol{\theta})$ with respect to θ_1. It follows that for numerical purposes we only need to keep in memory the derivatives of $A_{2d}(\boldsymbol{\theta})$ with respect to θ_1.

Now as a relation among the derivatives of $A_{2d}(\boldsymbol{\theta})$ with respect to θ_1 we have the following theorem.

Theorem 5.1. *$A_{2d}(\boldsymbol{\theta})$ satisfies the following differential equation*

$$(\theta_1 + 2\theta_2\partial_1 + 3\theta_3\partial_1^2 + \cdots + 2d\theta_d\partial_1^{2d-1})A_{2d}(\boldsymbol{\theta}) = 0. \tag{16}$$

Proof is omitted since it almost the same as the proof of Theorem 2.1, by noting

$$0 = [\exp(\theta_1 x + \cdots + \theta_{2d}x^{2d})]_{-\infty}^{\infty}.$$

By (16), $\partial_1^{2d-1}A_{2d}(\boldsymbol{\theta})$ is written in terms of lower-order derivatives as

$$\partial_1^{2d-1}A_{2d}(\boldsymbol{\theta}) = -\frac{1}{2d\theta_{2d}}(\theta_1 + 2\theta_2\partial_1 + 3\theta_3\partial_1^2 + \cdots + (2d-1)\theta_{2d-1}\partial_1^{2d-2})A_{2d}(\boldsymbol{\theta}).$$

Recursively differentiating this by θ_1 all higher-order derivatives $\partial_1^m A_{2d}(\boldsymbol{\theta})$, $m \ge 2d-1$, can be easily written in terms of $A_{2d}(\boldsymbol{\theta}), \partial_1 A_{2d}(\boldsymbol{\theta}), \dots, \partial_1^{2d-2}A_{2d}(\boldsymbol{\theta})$.

As a convenient initial point consider $\boldsymbol{\theta}^0 = (0,0,\dots,0,-c)$, $c > 0$. Then

$$\partial_1^m A_{2d}(\boldsymbol{\theta}^0) = \int_{-\infty}^{\infty} x^m \exp(-cx^{2d})\mathrm{d}x = \begin{cases} \frac{1}{d}c^{-(1+m)/2d}\Gamma(\frac{1+m}{2d}) & m = 0,2,4,\dots \\ 0 & m = 1,3,5,\dots \end{cases}$$

which do not need numerical integration.

5.2 Determination of the degree for the case of the whole line

For determining the order of the model we consider the testing problem

$$H_0 : \boldsymbol{\theta} \in \Omega_{2d-2} \text{ v.s. } H_1 : \boldsymbol{\theta} \in \Omega_{2d}.$$

The parameter space is illustrated in Figure 4, where Ω_{2d-2} corresponds to the origin.

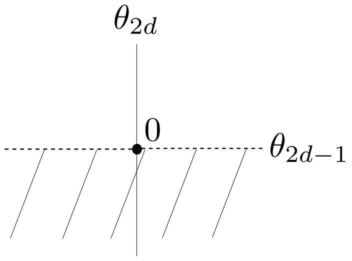

Figure 4: Model of order $2d - 2$ within the model of order $2d$

Here we need to do more careful analysis than in Section 3. The difficulty in this case is that $A_{2d}(\boldsymbol{\theta})$ in (15) is infinite for $\theta_{2d-1} \neq 0, \theta_{2d} = 0$:

$$A(\theta_1, \ldots, \theta_{2d}, \theta_{2d-1}, 0) = \infty, \qquad \forall \theta_{2d-1} \neq 0.$$

Hence we can not take the partial derivative of $A_{2d}(\boldsymbol{\theta})$ with respect to θ_{2d-1} at $(\theta_1, \ldots, \theta_{2d-2}, 0, 0)$. However $\partial_{2d-1} A(\theta_1, \ldots, \theta_{2d})$ and $\partial_{2d} A(\theta_1, \ldots, \theta_{2d})$ exist, as long as $\theta_{2d} < 0$. Also if $\theta_{2d-2} < 0$, as $(\theta_{2d-1}, \theta_{2d}) \to (0, 0)$ in such a way that $|\theta_{2d-1}/\theta_{2d}|$ is bounded, by the dominated convergence theorem we have

$$\begin{aligned} \lim_{\substack{(\theta_{2d-1},\theta_{2d})\to(0,0) \\ |\theta_{2d-1}/\theta_{2d}|:\text{bounded}}} \partial_{2d-1} A(\theta_1, \ldots, \theta_{2d}) &= \int_{-\infty}^{\infty} x^{2d-1} \exp(\theta_1 x + \cdots + \theta_{2d-2} x^{2d-2}) \mathrm{d}x \\ &= A(\boldsymbol{\theta}_{2d-2}) E_{\boldsymbol{\theta}_{2d-2}}(X^{2d-1}), \end{aligned}$$

and

$$\lim_{\substack{(\theta_{2d-1},\theta_{2d})\to(0,0) \\ |\theta_{2d-1}/\theta_{2d}|:\text{bounded}}} \partial_{2d} A(\theta_1, \ldots, \theta_{2d}) = A(\boldsymbol{\theta}_{2d-2}) E_{\boldsymbol{\theta}_{2d-2}}(X^{2d}),$$

where $\boldsymbol{\theta}_{2d-2} = (\theta_1, \ldots, \theta_{2d-2}, 0, 0)$, $\theta_{2d-2} < 0$ and $E_{\boldsymbol{\theta}_{2d-2}}$ denotes the expected value under $\boldsymbol{\theta}_{2d-2}$. Let $\hat{\boldsymbol{\theta}}_{2d-2}$ denote MLE under H_0.

We now redefine the $(2d) \times (2d)$ Fisher information matrix $I(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_{2d-2}$ as

$$\tilde{I}(\boldsymbol{\theta}_{2d-2}) = \begin{bmatrix} \tilde{I}_{2d-2,2d-2}(\boldsymbol{\theta}_{2d-2}) & \tilde{I}_{2d-2,2d}(\boldsymbol{\theta}_{2d-2}) \\ \tilde{I}_{2d,2d-2}(\boldsymbol{\theta}_{2d-2}) & \tilde{I}_{2d,2d}(\boldsymbol{\theta}_{2d-2}) \end{bmatrix} = \lim_{\substack{(\theta_{2d-1},\theta_{2d})\to(0,0) \\ |\theta_{2d-1}/\theta_{2d}|:\text{bounded}}} \begin{bmatrix} I_{2d-2,2d-2}(\boldsymbol{\theta}) & I_{2d-2,2d}(\boldsymbol{\theta}) \\ I_{2d,2d-2}(\boldsymbol{\theta}) & I_{2d,2d}(\boldsymbol{\theta}) \end{bmatrix},$$

where $I_{2d,2d}$ is a 2×2 matrix. Define

$$\tilde{I}_{2d,2d\cdot 1,\ldots,2d-2}(\boldsymbol{\theta}) = \tilde{I}_{2d,2d}(\boldsymbol{\theta}) - \tilde{I}_{2d,2d-2}(\boldsymbol{\theta}) \tilde{I}_{2d-2,2d-2}(\boldsymbol{\theta})^{-1} \tilde{I}_{2d-2,2d}(\boldsymbol{\theta}).$$

For testing H_0 we again propose to use a score statistic

$$T_{2d-2} = n[\partial_{2d-1}\bar{l}(\hat{\boldsymbol{\theta}}_{2d-2};\boldsymbol{x}), \partial_{2d}\bar{l}(\hat{\boldsymbol{\theta}}_{2d-2};\boldsymbol{x})]\ \tilde{I}_{2d,2d\cdot1,\dots,2d-2}(\hat{\boldsymbol{\theta}}_{2d-2})^{-1}\begin{bmatrix}\partial_{2d-1}\bar{l}(\hat{\boldsymbol{\theta}}_{2d-2};\boldsymbol{x})\\ \partial_{2d}\bar{l}(\hat{\boldsymbol{\theta}}_{2d-2};\boldsymbol{x})\end{bmatrix}. \tag{17}$$

We reject H_0 if

$$T_{2d-2} \geq \chi^2_2(\alpha), \tag{18}$$

where $\chi^2_2(\alpha)$ is the upper α-quantile of the χ^2 distribution with two degrees of freedom. Numerical performance of this test is confirmed in the next subsection.

5.3 Numerical experiments for the whole line

For checking the asymptotic distribution of the MLE, we compare the empirical distribution of p_i in (12) with N$(0,1)$ for $2d = 4$ and $\boldsymbol{\theta}^* = (1, 4, -2, -3)$. For checking (18) we compare the empirical distribution of T_{2d-2} of (17) with the χ^2 distribution with 2 degrees of freedom. For $2d - 2 = 2$ we choose $\boldsymbol{\theta}^* = (2, -1, 0, 0)$.

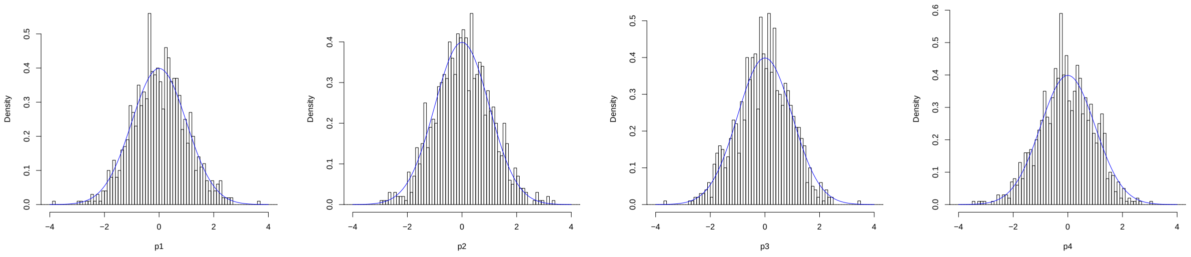

Figure 5: Histogram p_i and the density of N$(0, 1)$ for $2d = 4$

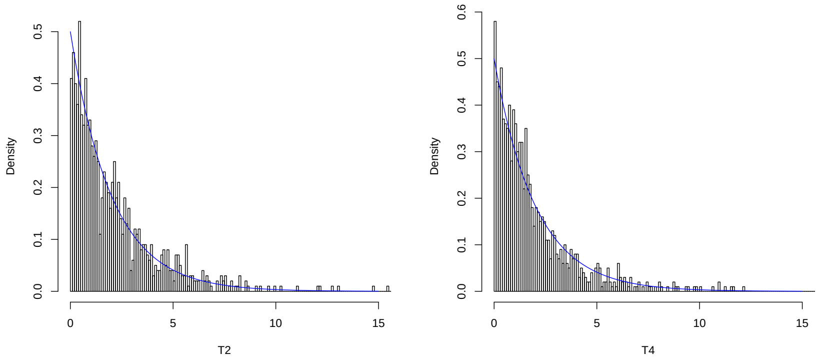

Figure 6: Histogram of T_{2d-2} of (17) and the density of $\chi^2(2)$ for $2d = 4, 6$.

Figure 5 shows for $2d = 4$ the histogram of p_i and the density of N(0, 1). We see that they agree with each other. Figure 6 shows for $2d = 4, 6$ the histogram of T_{2d-2} of (17) and the density of the chi-square distribution with 2 d.f. We again see a good agreement.

6 Bivariate exponential-polynomial distribution on the positive orthant

In this section we develop holonomic gradient descent for bivariate exponential-polynomial distribution on the positive orthant. The differential equations needed for HGD are more difficult to derive than in the univariate case. Also the problem of singularity of the system of differential equations arises in the bivariate case.

Let

$$
\begin{aligned}
h(\boldsymbol{\theta}, x, y) &= \exp\left(\sum_{0 \le i+j \le d} \theta_{ij} x^i y^j \right) \\
&= \exp(\theta_{10}x + \theta_{01}y + \theta_{20}x^2 + \theta_{11}xy + \theta_{02}y^2 + \cdots + \theta_{d0}x^d + \cdots + \theta_{0d}y^d)
\end{aligned}
$$

and consider the density function

$$
f(x, y; \boldsymbol{\theta}) = \frac{1}{A(\boldsymbol{\theta})} h(\boldsymbol{\theta}, x, y),
$$

where

$$
A(\boldsymbol{\theta}) = \int_0^\infty \int_0^\infty h(\boldsymbol{\theta}, x, y) \mathrm{d}x \mathrm{d}y
$$

is the normalizing constant. We call this distribution a bivariate exponential-polynomial distribution of degree d. Here the parameter vector $\boldsymbol{\theta}$ belongs to the parameter space

$$
\boldsymbol{\Theta} = \{\boldsymbol{\theta} \mid A(\boldsymbol{\theta}) < \infty\}. \tag{19}
$$

We consider the structure of $\boldsymbol{\Theta}$ below in Section 6.3. We note that if $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, then $h(\boldsymbol{\theta}, x, y)$ satisfies

$$
\begin{aligned}
h(\boldsymbol{\theta}, x, y) &\to 0 \qquad (x \to \infty), \\
h(\boldsymbol{\theta}, x, y) &\to 0 \qquad (y \to \infty).
\end{aligned} \tag{20}
$$

Given the sample $z = \{(x_i, y_i)\}_{i=1}^n$, $(1/n)$ times the log-likelihood function is written as

$$
\begin{aligned}
\bar{l}(\boldsymbol{\theta}, z) &= \sum_{1 \le i+j \le d} \theta_{ij} \overline{x^i y^j} - \log A(\boldsymbol{\theta}) \\
&= \theta_{10}\bar{x} + \theta_{01}\bar{y} + \cdots + \theta_{st}\overline{x^s y^t} + \cdots + \theta_{d0}\overline{x^d} + \cdots + \theta_{0d}\overline{y^d} - \log A(\boldsymbol{\theta}).
\end{aligned} \tag{21}
$$

From (21) the gradient vectors is given as

$$\nabla \bar{l}(\boldsymbol{\theta}, z) = \begin{bmatrix} \overline{x} \\ \overline{y} \\ \vdots \\ \overline{x^s y^t} \\ \vdots \\ \overline{x^d} \\ \overline{x^{d-1}y} \\ \vdots \\ \overline{y^d} \end{bmatrix} - \frac{1}{A(\boldsymbol{\theta})} \begin{bmatrix} \partial_{10}A(\boldsymbol{\theta}) \\ \partial_{01}A(\boldsymbol{\theta}) \\ \vdots \\ \partial_{st}A(\boldsymbol{\theta}) \\ \vdots \\ \partial_{d0}A(\boldsymbol{\theta}) \\ \partial_{d-1,1}A(\boldsymbol{\theta}) \\ \vdots \\ \partial_{0d}A(\boldsymbol{\theta}) \end{bmatrix}, \tag{22}$$

where $\partial_{ij} = \partial / \partial\theta_{ij}$. As in the univariate case we would like to avoid numerical integration for $\partial_{ij}A(\boldsymbol{\theta})$, $0 \le i + j \le d$, in every step of iteration for obtaining MLE.

6.1 Maximum likelihood estimation for the bivariate case

We first derive differential equations satisfied by $A(\boldsymbol{\theta})$. Since there are terms like xy, we need to obtain different types of differential equations, which were not needed in the univariate case.

Let

$$A_x(\boldsymbol{\theta}) = \int_0^\infty h(\boldsymbol{\theta}, x, 0)\mathrm{d}x = \int_0^\infty \exp(\theta_{10}x + \theta_{20}x^2 + \cdots + \theta_{d0}x^d)\mathrm{d}x, \tag{23}$$

$$A_y(\boldsymbol{\theta}) = \int_0^\infty h(\boldsymbol{\theta}, 0, y)\mathrm{d}y = \int_0^\infty \exp(\theta_{01}y + \theta_{02}y^2 + \cdots + \theta_{0d}y^d)\mathrm{d}y. \tag{24}$$

The values of (23), (24) and their derivatives with respect to θ_{ij} can be obtained easily from our results for the univariate case. Hence in the following derivation we treat them as known or already evaluated.

We differentiate $A(\boldsymbol{\theta})$ by θ_{01} or θ_{10}. Then

$$\partial_{10}A(\boldsymbol{\theta}) = \int_0^\infty \int_0^\infty xh(\boldsymbol{\theta}, x, y)\mathrm{d}x\mathrm{d}y, \quad \partial_{01}A(\boldsymbol{\theta}) = \int_0^\infty \int_0^\infty yh(\boldsymbol{\theta}, x, y)\mathrm{d}x\mathrm{d}y$$

and we have

$$\partial_{10}^s \partial_{01}^t A(\boldsymbol{\theta}) = \int_0^\infty \int_0^\infty x^s y^t h(\boldsymbol{\theta}, x, y)\mathrm{d}x\mathrm{d}y. \tag{25}$$

On the other hand,

$$\partial_{st}A(\boldsymbol{\theta}) = \int_0^\infty \int_0^\infty x^s y^t h(\boldsymbol{\theta}, x, y)\mathrm{d}x\mathrm{d}y. \tag{26}$$

From (25), (26) we have

$$(\partial_{st} - \partial_{10}^s \partial_{01}^t)A(\boldsymbol{\theta}) = 0.$$

Furthermore corresponding to Theorem 2.1, we have the following theorem.

Theorem 6.1. *$A(\boldsymbol{\theta})$ satisfies the following differential equations.*

$$\left(\sum_{1\le i+j\le d, 1\le i} i\theta_{ij}\partial_{10}^{i-1}\partial_{01}^{j}\right)A(\boldsymbol{\theta}) = -A_y(\boldsymbol{\theta}), \tag{27}$$

$$\left(\sum_{1\le i+j\le d, 1\le j} j\theta_{ij}\partial_{10}^{i}\partial_{01}^{j-1}\right)A(\boldsymbol{\theta}) = -A_x(\boldsymbol{\theta}). \tag{28}$$

Proof. By symmetry we only show (27). We have

$$\begin{aligned}\int_0^\infty\int_0^\infty \partial_x h(\boldsymbol{\theta},x,y)\mathrm{d}x\mathrm{d}y &= \int_0^\infty\left\{\int_0^\infty \partial_x h(\boldsymbol{\theta},x,y)\mathrm{d}x\right\}\mathrm{d}y\\ &= \int_0^\infty [h(\boldsymbol{\theta},x,y)]_{x=0}^{x=\infty}\,\mathrm{d}y = -\int_0^\infty h(\boldsymbol{\theta},0,y)\mathrm{d}y \qquad \text{(by (20))}\\ &= -A_y(\boldsymbol{\theta}). \end{aligned}\tag{29}$$

On the other hand,

$$\begin{aligned}\int_0^\infty\int_0^\infty \partial_x h(\boldsymbol{\theta},x,y)\mathrm{d}x\mathrm{d}y &= \int_0^\infty\int_0^\infty \partial_x \exp\big(\sum_{0\le i+j\le d}\theta_{ij}x^i y^j\big)\mathrm{d}x\mathrm{d}y\\ &= \int_0^\infty\int_0^\infty \left(\sum_{1\le i+j\le d,1\le i} i\theta_{ij}x^{i-1}y^j\right)\exp\big(\sum_{0\le i+j\le d}\theta_{ij}x^iy^j\big)\mathrm{d}x\mathrm{d}y\\ &= \left(\sum_{1\le i+j\le d,1\le i} i\theta_{ij}\partial_{10}^{i-1}\partial_{01}^{j}\right)\int_0^\infty\int_0^\infty h(\boldsymbol{\theta},x,y)\mathrm{d}x\mathrm{d}y \qquad \text{(by (25))}\\ &= \left(\sum_{1\le i+j\le d,1\le i} i\theta_{ij}\partial_{10}^{i-1}\partial_{01}^{j}\right)A(\boldsymbol{\theta}). \end{aligned}\tag{30}$$

(27) follows from (29) and (30). □

In the univariate case, the important fact was that higher-order derivatives of $A_d(\boldsymbol{\theta})$ are written as rational function combinations of lower-order derivatives of $A_d(\boldsymbol{\theta})$. In (27), (28), the highest order of derivatives in $A_d(\boldsymbol{\theta})$ is $d-1$ and there are d derivatives of order $d-1$:

$$\partial_{10}^{d-1}, \partial_{10}^{d-2}\partial_{01}, \cdots, \partial_{10}\partial_{01}^{d-2}, \partial_{01}^{d-1}.$$

If we want to evaluate these d derivatives of order $d-1$ by solving a system of equations, then we do not have enough equations for $d \ge 3$, because there are only two equations in Theorem 6.1. We need to have more differential equations.

To obtain more equations, we operate the following set

$$O_q = \{\partial_{10}^q, \partial_{10}^{q-1}\partial_{01}, \partial_{10}^{q-2}\partial_{01}^2, \cdots, \partial_{01}^q\}$$

of $q + 1$ differential operators of the same order q to (27) and (28). In order to determine q, we count the number of differential equations obtained after operating O_q.

The highest order of derivatives after operating O_q to (27), (28) is $q + d - 1$ and there are the following $q + d$ derivatives

$$\partial_{10}^{q+d-1}, \partial_{10}^{q+d-2}\partial_{01}, \cdots, \partial_{10}\partial_{01}^{q+d-2}, \partial_{01}^{q+d-1}.$$

On the other hand there are $2(q+1)$ differential equations after operating O_q to (27), (28). Hence we have the right number of equations if we take $q + d = 2(q + 1)$ or

$$q = d - 2.$$

In view of

$$\partial_{10}A_y(\boldsymbol{\theta}) = 0, \qquad \partial_{01}A_x(\boldsymbol{\theta}) = 0,$$

when we operate

$$O_{d-2} = \{\partial_{10}^{d-2}, \partial_{10}^{d-3}\partial_{01}, \partial_{10}^{d-4}\partial_{01}^{2}, \cdots, \partial_{01}^{d-2}\}$$

to (27), (28), we have the following system of differential equations.

$$\begin{bmatrix} \partial_{10}^{d-2} & 0 \\ \partial_{10}^{d-3}\partial_{01} & \vdots \\ \vdots & \vdots \\ \partial_{10}\partial_{01}^{d-3} & \vdots \\ \partial_{01}^{d-2} & 0 \\ 0 & \partial_{10}^{d-2} \\ \vdots & \partial_{10}^{d-3}\partial_{01} \\ \vdots & \vdots \\ \vdots & \partial_{10}\partial_{01}^{d-3} \\ 0 & \partial_{01}^{d-2} \end{bmatrix} \begin{bmatrix} \left(\sum\limits_{1\le i+j\le d, 1\le i} i\theta_{ij}\partial_{10}^{i-1}\partial_{01}^{j}\right)A(\boldsymbol{\theta}) \\ \left(\sum\limits_{1\le i+j\le d, 1\le j} j\theta_{ij}\partial_{10}^{i}\partial_{01}^{j-1}\right)A(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \partial_{10}^{d-2} & 0 \\ \partial_{10}^{d-3}\partial_{01} & \vdots \\ \vdots & \vdots \\ \partial_{10}\partial_{01}^{d-3} & \vdots \\ \partial_{01}^{d-2} & 0 \\ 0 & \partial_{10}^{d-2} \\ \vdots & \partial_{10}^{d-3}\partial_{01} \\ \vdots & \vdots \\ \vdots & \partial_{10}\partial_{01}^{d-3} \\ 0 & \partial_{01}^{d-2} \end{bmatrix} \begin{bmatrix} -A_y(\boldsymbol{\theta}) \\ -A_x(\boldsymbol{\theta}) \end{bmatrix} = - \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \partial_{01}^{d-2}A_y(\boldsymbol{\theta}) \\ \partial_{10}^{d-2}A_x(\boldsymbol{\theta}) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{31}$$

We transform (31) to a system of differential equations to solve for the derivatives of the highest order

$$\partial_{10}^{2d-3}, \partial_{10}^{2d-4}\partial_{01}, \cdots, \partial_{10}\partial_{01}^{2d-4}, \partial_{01}^{2d-3}.$$

For any pair of non-negative integers (s, t) satisfying $s + t = d - 2$ let

$$\begin{aligned} \phi(s,t) &= s\partial_{10}^{s-1}\partial_{01}^{t} + \theta_{10}\partial_{10}^{s}\partial_{01}^{t} + \sum_{2\le i+j\le d-1, 1\le i} i\theta_{ij}\partial_{10}^{s+i-1}\partial_{01}^{j+t}, \\ \psi(s,t) &= t\partial_{10}^{s}\partial_{01}^{t-1} + \theta_{01}\partial_{10}^{s}\partial_{01}^{t} + \sum_{2\le i+j\le d-1, 1\le j} j\theta_{ij}\partial_{10}^{s+i}\partial_{01}^{t+j-1}. \end{aligned} \tag{32}$$

Then (31) is transformed to

$$\begin{cases} \sum\limits_{i+j=d,1\leq i} i\theta_{ij}\partial_{10}^{s+i-1}\partial_{01}^{t+j}A(\boldsymbol{\theta}) = -\partial_{10}^{s}\partial_{01}^{t}A_y(\boldsymbol{\theta}) - \phi(s,t)A(\boldsymbol{\theta}), \\ \sum\limits_{i+j=d,1\leq j} j\theta_{ij}\partial_{10}^{s+i}\partial_{01}^{t+j-1}A(\boldsymbol{\theta}) = -\partial_{10}^{s}\partial_{01}^{t}A_x(\boldsymbol{\theta}) - \psi(s,t)A(\boldsymbol{\theta}). \end{cases} \tag{33}$$

In matrix form (33) is expressed as

$$P(\boldsymbol{\theta})\begin{bmatrix} \partial_{10}^{2d-3}A(\boldsymbol{\theta}) \\ \partial_{10}^{2d-4}\partial_{01}A(\boldsymbol{\theta}) \\ \vdots \\ \partial_{10}\partial_{01}^{2d-4}A(\boldsymbol{\theta}) \\ \partial_{01}^{2d-3}A(\boldsymbol{\theta}) \end{bmatrix} = Q(\boldsymbol{\theta}),$$

where

$$P(\boldsymbol{\theta}) = \begin{bmatrix} d\theta_{d0} & \cdots & \cdots & 2\theta_{2,d-2} & \theta_{1,d-1} & & & \\ & d\theta_{d0} & \cdots & \cdots & 2\theta_{2,d-2} & \theta_{1,d-1} & & \\ & & & \cdots & \cdots & & & \\ & & & d\theta_{d0} & \cdots & \cdots & 2\theta_{2,d-2} & \theta_{1,d-1} \\ \theta_{d-1,1} & 2\theta_{d-2,2} & \cdots & \cdots & d\theta_{0d} & & & \\ & \theta_{d-1,1} & 2\theta_{d-2,2} & \cdots & \cdots & d\theta_{0d} & & \\ & & & \cdots & \cdots & & & \\ & & & \theta_{d-1,1} & 2\theta_{d-2,2} & \cdots & \cdots & d\theta_{0d} \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \\ a \end{matrix}}\right\} (d-1)\text{ rows} \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \\ a \end{matrix}}\right\} (d-1)\text{ rows} \end{matrix}, \tag{34}$$

$$Q(\boldsymbol{\theta}) = -\begin{bmatrix} \phi(d-2,0)A(\boldsymbol{\theta}) \\ \phi(d-3,1)A(\boldsymbol{\theta}) \\ \vdots \\ \phi(1,d-3)A(\boldsymbol{\theta}) \\ \partial_{01}^{d-2}A_y(\boldsymbol{\theta}) + \phi(0,d-2)A(\boldsymbol{\theta}) \\ \partial_{10}^{d-2}A_x(\boldsymbol{\theta}) + \psi(d-2,0)A(\boldsymbol{\theta}) \\ \psi(d-3,1)A(\boldsymbol{\theta}) \\ \vdots \\ \psi(1,d-3)A(\boldsymbol{\theta}) \\ \psi(0,d-2)A(\boldsymbol{\theta}) \end{bmatrix}. \tag{35}$$

In $P(\boldsymbol{\theta})$ empty elements in the matrix are zeros. We give further consideration of $P(\boldsymbol{\theta})$ in the next section.

If $\det P(\boldsymbol{\theta}) \neq 0$,

$$\begin{bmatrix} \partial_{10}^{2d-3}A(\boldsymbol{\theta}) \\ \partial_{10}^{2d-4}\partial_{01}A(\boldsymbol{\theta}) \\ \vdots \\ \partial_{10}\partial_{01}^{2d-4}A(\boldsymbol{\theta}) \\ \partial_{01}^{2d-3}A(\boldsymbol{\theta}) \end{bmatrix} = P^{-1}(\boldsymbol{\theta})Q(\boldsymbol{\theta}). \tag{36}$$

Hence from (32), (35), (36) we see that $\partial_{10}^{2d-3}A(\boldsymbol{\theta}), \partial_{10}^{2d-4}\partial_{01}A(\boldsymbol{\theta}), \cdots, \partial_{10}\partial_{01}^{2d-4}A(\boldsymbol{\theta}), \partial_{01}^{2d-3}A(\boldsymbol{\theta})$, are written as rational function combinations of elements of the vector

$$F(\boldsymbol{\theta}) = [A(\boldsymbol{\theta}), \partial_{10}A(\boldsymbol{\theta}), \partial_{01}A(\boldsymbol{\theta}), \cdots, \partial_{10}^{2d-4}A(\boldsymbol{\theta}), \partial_{10}^{2d-5}\partial_{01}A(\boldsymbol{\theta}), \cdots, \partial_{10}\partial_{01}^{2d-5}A(\boldsymbol{\theta}), \partial_{01}^{2d-4}A(\boldsymbol{\theta})]^\top.$$

If we can evaluate $F(\boldsymbol{\theta})$ at any $\boldsymbol{\theta}$, then by (22) we can obtain the maximum likelihood estimate of the bivariate exponential-polynomial distribution. As in the univariate case, if the initial values of $F(\boldsymbol{\theta}_0)$ can be evaluated at $\boldsymbol{\theta}_0$, then the value of $F(\boldsymbol{\theta})$ at any other point $\boldsymbol{\theta}$ can be obtained by solving the differential equation.

In the univariate case, the origin $\theta_d = 0$ was the only singular point of the differential equation (5). In the bivariate case the set $\{\boldsymbol{\theta} | \det P(\boldsymbol{\theta}) = 0\}$ is the set of singularities of (31). This singularity causes difficulty for HGD and in the next section we investigate $\det P(\boldsymbol{\theta})$.

6.2 Evaluation of the determinant of the Pfaffian system

We prove that $\det P(\boldsymbol{\theta})$ in (34) is given by the discriminant of a polynomial equation. We use the basic results on determinantal expression for resultants and discriminants (cf. Chapter 12 of [4], Section 3.3 of [1]). Let two polynomials $f(x)$, $g(x)$ be denoted as

$$f(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0 = a_m \prod_{i=1}^{m}(x - \alpha_i), \tag{37}$$

$$g(x) = b_n x^n + b_{m-1}x^{m-1} + \cdots + b_0 = b_n \prod_{i=1}^{n}(x - \beta_i).$$

The resultant $R(f, g)$ is defined as

$$R(f,g) = a_m^n b_n^m \prod_{i=1}^{m}\prod_{j=1}^{n}(\alpha_i - \beta_j).$$

Then the determinantal expression of $R(f, g)$ is given as follows ((1.12) of Chapter 12 of [4], Lemma 3.3.4 of [1]).

$$R(f,g) = \det \begin{bmatrix} a_m & a_{m-1} & \cdots & \cdots & \cdots & a_0 & & & \\ & a_n & a_{m-1} & \cdots & \cdots & \cdots & a_0 & & \\ & & & & \cdots & \cdots & \cdots & & \\ & & & a_m & a_{m-1} & \cdots & \cdots & \cdots & a_0 \\ b_n & b_{n-1} & \cdots & b_0 & & & & & \\ & b_n & b_{n-1} & \cdots & b_0 & & & & \\ & & & & \cdots & & & & \\ & & & & & \cdots & & & \\ & & & & & & \cdots & & \\ & & & & & b_n & b_{n-1} & \cdots & b_0 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a\\ a\\ a\\ a \end{matrix}}\right\} n \text{ rows} \\ \left.\vphantom{\begin{matrix} a\\ a\\ a\\ a\\ a\\ a \end{matrix}}\right\} m \text{ rows} \end{matrix}$$

We also consider the discriminant. For $f(x)$ in (37) the discriminant for the equation $f(x) = 0$ is given by

$$D = (-1)^{m(m-1)/2} a_m^{2(m-1)} \prod_{1 \le i < j \le m} (\alpha_i - \alpha_j)^2.$$

Let

$$p(x;\boldsymbol{\theta}) = \theta_{d0}x^d + \theta_{d-1,1}x^{d-1} + \theta_{d-2,2}x^{d-2} + \cdots + \theta_{1,d-1}x + \theta_{0d}. \tag{38}$$

This polynomial will also appear in the next section in the investigation of the parameter space $\boldsymbol{\Theta}$ in (19). The discriminant $D(\boldsymbol{\theta})$ of the polynomial equation $p(x,\boldsymbol{\theta}) = 0$ is given as ((1.29) of Chapter 12 of [4], Definition 3.3.3 of [1])

$$D(\boldsymbol{\theta}) = \frac{1}{\theta_{d0}} R(p, p'), \tag{39}$$

where

$$R(p,p') = \det \left[\begin{array}{ccccccccc}
\theta_{d0} & \theta_{d-1,1} & \cdots & \cdots & \cdots & \theta_{0d} & & & \\
 & \theta_{d0} & \theta_{d-1,1} & \cdots & \cdots & \cdots & \theta_{0d} & & \\
 & & & & \cdots & \cdots & \cdots & & \\
 & & & \theta_{d0} & \theta_{d-1,1} & \cdots & \cdots & \cdots & \theta_{0d} \\
d\theta_{d0} & \cdots & \cdots & 2\theta_{2,d-2} & \theta_{1,d-1} & & & & \\
 & d\theta_{d0} & \cdots & \cdots & 2\theta_{2,d-2} & \theta_{1,d-1} & & & \\
 & & & \cdots & \cdots & & & & \\
 & & & & \cdots & \cdots & & & \\
 & & & & d\theta_{d0} & \cdots & \cdots & 2\theta_{2,d-2} & \theta_{1,d-1}
\end{array}\right]
\begin{matrix}
\left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} (d-1) \text{ rows} \\
\left.\begin{matrix} \\ \\ \\ \\ \\ \end{matrix}\right\} d \text{ rows}
\end{matrix}.$$

Using (39) we give the following theorem on the relation of $\det P(\boldsymbol{\theta})$ in (34) and the discriminant $D(\boldsymbol{\theta})$ of polynomial equation $p(x;\boldsymbol{\theta}) = 0$ in (38).

Theorem 6.2.

$$\det P(\boldsymbol{\theta}) = d^{d-2} D(\boldsymbol{\theta}). \tag{40}$$

Proof. Define a $(2d-1) \times (2d-1)$ matrix S as

$$S = \begin{bmatrix} \theta_{d0} & 0 & -\theta_{d-2,2} & -2\theta_{d-3,3} & \cdots & -(d-1)\theta_{0d} & 0 & \cdots & 0 \\ \mathbf{0} & & & & P(\boldsymbol{\theta}) & & & & \end{bmatrix},$$

where $\mathbf{0}$ is a column vector of zeros of size $2d-2$. Expanding the determinant with respect to the fist column we have

$$\det S = \theta_{d0} \det P(\boldsymbol{\theta}). \tag{41}$$

On the other hand we add the i-row to the $(i+d)$-th rows $(1 \le i \le d-1)$ and then add the

$(d+1)$-st row multiplied by $d-1$ to the first row. Then we obtain

$$\det S \;= d^{d-2} \det \begin{bmatrix} d\theta_{d0} & \cdots & \cdots & 2\theta_{2,d-2} & \theta_{1,d-1} & & & & \\ & d\theta_{d0} & \cdots & \cdots & 2\theta_{2,d-2} & \theta_{1,d-1} & & & \\ & & & \cdots & \cdots & & & & \\ & & & & \cdots & \cdots & & & \\ & & & & d\theta_{d0} & \cdots & \cdots & 2\theta_{2,d-2} & \theta_{1,d-1} \\ \theta_{d0} & \theta_{d-1,1} & \cdots & \cdots & \cdots & \theta_{0d} & & & \\ & \theta_{d0} & \theta_{d-1,1} & \cdots & \cdots & \cdots & \theta_{0d} & & \\ & & & & \cdots & \cdots & \cdots & & \\ & & & \theta_{d0} & \theta_{d-1,1} & \cdots & \cdots & \cdots & \theta_{0d} \end{bmatrix}.$$

By interchanging rows

$$\begin{aligned} \det S \;&= d^{d-2}(-1)^{d(d-1)}R(p,p') \\ &= \theta_{d0}d^{d-2}\det D(\boldsymbol{\theta}). \qquad \text{(by (39))} \end{aligned} \tag{42}$$

From (41), (42) we have

$$\det P(\boldsymbol{\theta}) = d^{d-2}D(\boldsymbol{\theta}).$$

□

6.3 Structure of the parameter space for the bivariate case

In this section we investigate the parameter space $\boldsymbol{\Theta}$. By the transformation

$$x = r\cos\omega, \quad y = r\sin\omega,$$

define

$$H(\boldsymbol{\theta},\omega) = \int_0^\infty \tilde{h}(\boldsymbol{\theta},r,\omega)\mathrm{d}r, \quad \tilde{h}(\boldsymbol{\theta},r,\omega) = rh(\boldsymbol{\theta},r\cos\omega,r\sin\omega).$$

Since $\tilde{h}$ is non-negative, by Fubini's theorem $A(\boldsymbol{\theta})$ is written as

$$A(\boldsymbol{\theta}) = \int_0^{\pi/2} H(\boldsymbol{\theta},\omega)\mathrm{d}\omega.$$

It is easily seen that $H(\boldsymbol{\theta},\omega)$ is continuous in the compact interval $\omega \in [0,\pi/2]$ and $A(\boldsymbol{\theta}) = \infty$ if and only if $H(\boldsymbol{\theta},\omega) = \infty$ for some $\omega \in [0,\pi/2]$. Note that

$$H(\boldsymbol{\theta},\omega) < \infty \quad \Leftrightarrow \quad h(\boldsymbol{\theta}, y/\tan\omega, y) \to 0 \quad (y \to \infty).$$

The coefficient of the highest degree term in y of $h(\boldsymbol{\theta},ay,y)$ is $p(a;\boldsymbol{\theta})$, where $p(x;\boldsymbol{\theta})$ is given in (38). If $p(a;\boldsymbol{\theta}) < 0$ for all $a \geq 0$, then $\boldsymbol{\theta} \in \Theta$. Hence if we define

$$\boldsymbol{\Theta}' = \{\boldsymbol{\theta} \mid \forall a \geq 0, p(a;\boldsymbol{\theta}) < 0\}, \tag{43}$$

then $\mathbf{\Theta}' \subset \mathbf{\Theta}$. Note that for $\boldsymbol{\theta} \in \mathbf{\Theta} \setminus \mathbf{\Theta}'$ there exists $a \geq 0$ such that $p(a; \boldsymbol{\theta}) = 0$, i.e., the term of order d in y vanishes on the ray $\{(ay, y), y \geq 0\}$. In this sense $\boldsymbol{\theta} \in \mathbf{\Theta} \setminus \mathbf{\Theta}'$ may be considered as a model of order $d - 1$. We call $\mathbf{\Theta}'$ in (43) the parameter space of a *proper* order-d model.

The above consideration gives insight into the structure of $\mathbf{\Theta}$, but it is still difficult to decide whether $\boldsymbol{\theta} \in \mathbf{\Theta}'$ for a given $\boldsymbol{\theta}$. We propose the following easier method for determination. Note that $\theta_{d0} < 0, \theta_{0d} < 0$ is a trivial restriction and we assume this. Now clearly we have

$$p(x; \boldsymbol{\theta}) < 0 \quad (\forall x \geq 0) \;\Leftrightarrow\; \begin{cases} \theta_{d0} < 0, \\ p(x; \boldsymbol{\theta}) \text{ does not have a positive root.} \end{cases}$$

Following the argument in [2], we now move $\boldsymbol{\theta}$ from an initial point in $\mathbf{\Theta}'$, keeping $\theta_{d0} < 0, \theta_{0d} < 0$, and consider when $p(x; \boldsymbol{\theta})$ is no longer negative for some $x > 0$, i.e., when $p(x; \boldsymbol{\theta}) = 0$ has a positive root. There are two cases.

1. A real root moves from the negative real line to the positive real line.
2. A complex root moves to the positive real line.

The fist case corresponds to $\theta_{0d} > 0$, but this does not happen by our assumption. Complex roots for a polynomial with real coefficients appear in conjugate pairs and in the second case we have a multiple root on the positive real line. Hence under the assumption $\theta_{d0} < 0, \theta_{0d} < 0$, a positive root appears if and only if the discriminant $D(\boldsymbol{\theta})$ of $p(x; \boldsymbol{\theta}) = 0$ becomes 0 and the root becomes positive. Note that $D(\boldsymbol{\theta}) = 0$ may also happen because of negative or complex multiple roots.

Based on this observation consider the complement of the hypersurface $\{\boldsymbol{\theta} \mid D(\boldsymbol{\theta}) = 0\}$ in $\{\boldsymbol{\theta} \mid \theta_{d0} < 0, \theta_{0d} < 0\}$:

$$\mathbf{\Theta}'' = \{\boldsymbol{\theta} \mid \theta_{d0} < 0, \theta_{0d} < 0\} \setminus \{\boldsymbol{\theta} \mid D(\boldsymbol{\theta}) = 0\}$$

$\mathbf{\Theta}''$ consists of disjoint open connected components ("chambers"), which we denote by $\mathbf{\Theta}''_i, i \in I$. Then $\mathbf{\Theta}''$ is partitioned as

$$\mathbf{\Theta}'' = \bigcup_{i \in I} \mathbf{\Theta}''_i.$$

Note that the number of positive roots of $p(x; \boldsymbol{\theta})$ is constant in each chamber $\mathbf{\Theta}''_i$. Hence if $\mathbf{\Theta}''_i \cap \mathbf{\Theta}' \neq \emptyset$, then $\mathbf{\Theta}''_i \subset \mathbf{\Theta}'$, namely each $\mathbf{\Theta}''_i$ is either a subset of $\mathbf{\Theta}'$ or disjoint from $\mathbf{\Theta}'$. Define

$$I^* = \{i \in I \mid \mathbf{\Theta}''_i \cap \mathbf{\Theta}' \neq \emptyset\} = \{i \in I \mid \mathbf{\Theta}''_i \subset \mathbf{\Theta}'\}.$$

Since the hypersurface $\{\boldsymbol{\theta} \mid D(\boldsymbol{\theta}) = 0\}$ has measure zero, we have the following theorem concerning $\mathbf{\Theta}'$ in (43).

Theorem 6.3. *Except for a set of measure zero*

$$\mathbf{\Theta}' = \bigcup_{i \in I^*} \mathbf{\Theta}''_i. \tag{44}$$

Although it is difficult to completely characterize the boundaries of $\mathbf{\Theta}_i''$'s for general d, if the boundary between $\mathbf{\Theta}_i, i \in I^*$, and $\mathbf{\Theta}_j$, $j \in I^*$, corresponds to negative or complex multiple roots, then the boundary also belongs to $\mathbf{\Theta}'$.

We illustrate the partition (44) for the case of $d = 3$. For any $c_1, c_2 > 0$, we have $p(x;\boldsymbol{\theta}) < 0, \forall x > 0$ if and only if $c_1 p(c_2 x;\boldsymbol{\theta}) < 0, \forall x > 0$. This implies that we can assume $\theta_{03} = \theta_{30} = -1$ without loss of generality in considering the partition (44). In this case the discriminant is written as

$$D(\boldsymbol{\theta}) = \theta_{12}^2\theta_{21}^2 - 4\theta_{12}^3 - 4\theta_{21}^3 + 18\theta_{12}\theta_{21} - 27.$$

On the $(\theta_{12}, \theta_{21})$-plane, $D(\boldsymbol{\theta}) = 0$ consists of two curves as illustrated in Figure 7. In Figure 7, chamber A corresponds to two positive roots and one negative root, chamber B corresponds to two complex roots and one negative root, and chamber C corresponds to three negative roots. Hence the partition in (44) is $B \cup C$. The boundary between B and C also belongs to $\mathbf{\Theta}'$.

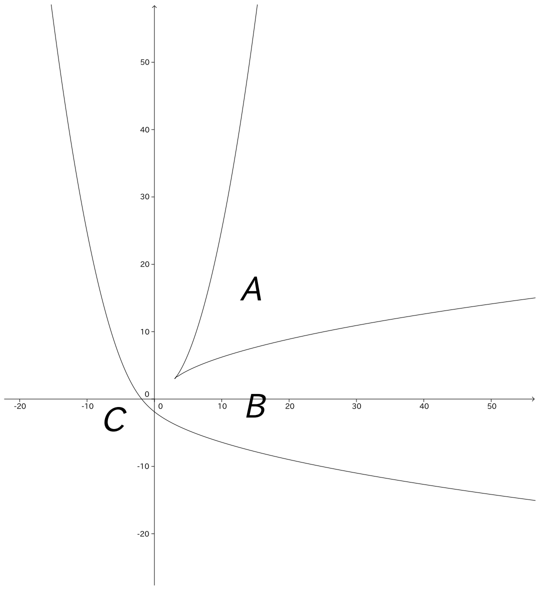

Figure 7: Partition of Theorem 6.3 for $d = 3$

For maximum likelihood estimation we need to take an initial point in each chamber $\mathbf{\Theta}_i''$, $i \in I^*$, of Theorem 6.3 and perform the numerical integration only for those initial points. It is difficult to give a simple initial point $\mathbf{\Theta}_i''$ for all $i \in I^*$. For some $i \in I^*$, the following simple initial point $\boldsymbol{\theta}_0$ is available. For $c_1 > 0, c_2 > 0$ define $\boldsymbol{\theta}_0$ by

$$\theta_{d0} = -c_1,\ \theta_{0d} = -c_2, \quad \theta_{ij} = 0 \ \text{ for } (i, j) \neq (0, d), (d, 0).$$

Then

$$p(x;\boldsymbol{\theta}_0) = -c_1 x^d - c_2 = 0, \quad p'(x;\boldsymbol{\theta}_0) = -dc_1 x^{d-1} = 0$$

do not have a common root and $R(p, p') \neq 0$. Hence $D(\boldsymbol{\theta}_0) \neq 0$ and $\det P(\boldsymbol{\theta}) \neq 0$ by (40). Furthermore clearly $p(x;\boldsymbol{\theta}_0)$ is negative for $x > 0$. Hence $\boldsymbol{\theta}_0 \in \mathbf{\Theta}_i''$, $i \in I^*$. For this $\boldsymbol{\theta}_0$ the normalizing constant and its derivatives are easily evaluated as

$$\begin{aligned}\partial_{ij}A(\boldsymbol{\theta}_0) &= \int_0^\infty \int_0^\infty x^i y^j \exp(-c_1 x^d - c_2 y^d)\mathrm{d}x\mathrm{d}y \\ &= \int_0^\infty x^i \exp(-c_1 x^d)\mathrm{d}x \int_0^\infty y^j \exp(-c_2 y^d)\mathrm{d}y \\ &= \frac{1}{d} c_1^{-(i+1)/d}\Gamma\left(\frac{i+1}{d}\right)\frac{1}{d} c_2^{-(j+1)/d}\Gamma\left(\frac{j+1}{d}\right).\end{aligned}$$

Although we do not show numerical results for the bivariate case, for $d = 2$ the computation of the normalizing constant and MLE is fast and the asymptotic distribution of MLE has been checked. For $d = 3$, the computation of the normalizing constant is fast, but the computation of MLE is somewhat heavy at current implementation in MATLAB. This seems to be due to high dimensionality (9 parameters) of the model for $d = 3$.

7 Some discussions

In this paper we discussed the maximum likelihood estimation of the exponential-polynomial distribution. Here we discuss some possible extensions of the distribution and topics for further research.

In the exponential-polynomial distribution we have a polynomial as the exponent of the exponential function. We can add another polynomial to the exponential-polynomial distribution, if this polynomial is non-negative over the sample space. Recall that the problem concerning non-negative polynomials was also essential for understanding the structure of the parameter space for the bivariate exponential-polynomial distribution in Section 6.3. Let

$$p(x; \boldsymbol{\eta}) = \eta_0 + \eta_1 x + \cdots + \eta_h x^h$$

be a polynomial in x. Consider the following density on the positive real line:

$$\begin{aligned} f(x; \boldsymbol{\eta}, \boldsymbol{\theta}) &= \frac{1}{\tilde{A}(\boldsymbol{\eta}, \boldsymbol{\theta})} p(x; \boldsymbol{\eta}) \exp(\theta_1 x + \cdots + \theta_d x^d), \\ \tilde{A}(\boldsymbol{\eta}, \boldsymbol{\theta}) &= \int_0^\infty p(x; \boldsymbol{\eta}) \exp(\theta_1 x + \cdots + \theta_d x^d)\mathrm{d}x. \end{aligned}$$

The normalizing constant $\tilde{A}(\boldsymbol{\eta}, \boldsymbol{\theta})$ can be evaluated as

$$\tilde{A}(\boldsymbol{\eta}, \boldsymbol{\theta}) = \sum_{i=0}^h \eta_i \int_0^\infty x^i \exp(\theta_1 x + \cdots + \theta_d x^d) = \sum_{i=0}^h \eta_i \partial_1^i A_d(\boldsymbol{\theta}),$$

where $A_d(\boldsymbol{\theta})$ is given in (2). Hence from the view point of holonomic gradient descent this generalization can be easily handled. However, in the estimation of this density we need to guarantee that $p(x; \hat{\boldsymbol{\eta}})$ is a non-negative polynomial for $x \geq 0$. This problem was considered in Fushiki et al. ([3]). They showed that the maximum likelihood estimation under the restriction of non-negativity of $p(x; \hat{\eta})$ can be performed with the technique of semidefinite programming. We can also use the parameterization of non-negative polynomials given in Proposition 3.3 of [9]. See also Section 9, Chapter V of [8].

For the univariate case we derived score tests for determining the order d of the model. The difficulty in model selection is the fact that the model of order $d - 1$ is on the boundary of the model of order d. In this paper we did not discuss the problem of model selection for the bivariate case, because the boundary is much more difficult compared to the univariate case, as discussed in Section 6.3. Also in the bivariate case, as the model of order d we included all monomials $x^d, x^{d-1}y, \dots, y^d$ of order d. However we may omit some monomials among these $d + 1$ monomials. The structure of the boundary of the model seems to depend on the choice of monomials of order d. Model selection procedures for the bivariate case is left to a future study.

Acknowledgment The authors are grateful to Satoshi Kuriki for very useful suggestions.

References

[1] H. Cohen. *A Course in Computational Algebraic Number Theory*, volume 138 of *Graduate Texts in Mathematics*. Springer-Verlag, Berlin, 1993.

[2] N. Fukuma and T. Mori. On positivity of linear combinations of polynomials. *IEEJ Transactions on Electronics, Information and Systems*, 113(10):798–803, 1993. (In Japanese).

[3] T. Fushiki, S. Horiuchi, and T. Tsuchiya. A maximum likelihood approach to density estimation with semidefinite programming. *Neural Comput.*, 18(11):2777–2812, 2006.

[4] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Modern Birkhäuser Classics. Birkhäuser Boston Inc., Boston, MA, 2008. Reprint of the 1994 edition.

[5] H. Hashiguchi, Y. Numata, N. Takayama, and A. Takemura. Holonomic gradient method for the distribution function of the largest root of a Wishart matrix. *Journal of Multivariate Analysis*, 117:296–312, 2013.

[6] T. Hibi, editor. *Grobner Bases: Statistics and Software Systems*. Springer, Tokyo, Japan, 2013.

[7] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions. Vol. 1*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1994. A Wiley-Interscience Publication.

[8] S. Karlin and W. J. Studden. *Tchebycheff Systems: With Applications in Analysis and Statistics*. Pure and Applied Mathematics, Vol. XV. John Wiley & Sons Inc., 1966.

[9] N. Kato and S. Kuriki. Likelihood ratio tests for positivity in polynomial regressions. *J. Multivariate Anal.*, 115:334–346, 2013.

[10] E. L. Lehmann. *Elements of Large-Sample Theory*. Springer Texts in Statistics. Springer-Verlag, New York, 1999.

[11] H. Nakayama, K. Nishiyama, M. Noro, K. Ohara, T. Sei, N. Takayama, and A. Takemura. Holonomic gradient descent and its application to the Fisher-Bingham integral. *Adv. in Appl. Math.*, 47(3):639–658, 2011.

[12] T. Sei, H. Shibata, A. Takemura, K. Ohara, and N. Takayama. Properties and applications of Fisher distribution on the rotation group. *Journal of Multivariate Analysis*, 116:440–455, 2013.

[13] F. Xu, R. C. Mittelhammer, and L. A. Torell. Modeling nonnegativity via truncated logistic and normal distributions: An application to ranch land price analysis. *Journal of Agricultural and Resource Economics*, 19(1):102–114, 1994.

[14] D. Zeilberger. A holonomic systems approach to special function identities. *Journal of Computational and Applied Mathematics*, 32(3):321–368, 1990.