Problem Formulations for `dsdp`[*]

Satoshi Kakihara[†] Takashi Tsuchiya[‡]

October 26th, 2022

Abstract

This vignette discusses problem formulations used in the package `dsdp`. The main task of `dsdp` is to estimate a density function using a maximum likelihood method whose models are a family of exponential distributions with polynomial correction terms. In this vignette, we discuss the procedure to transform the maximum likelihood problems to a variant of semidefinite programming (SDP) problems. Detailed discussions of SDP and implementations of solvers will be found in another vignette.

1 Maximum Likelihood Methods

Let $g(x)$ be an unknown univariate density function over the support $S \subset \mathbb{R}$, where $\mathbb{R}$ denotes a set of real numbers. For an n data set $\{x_1, \ldots, x_n\}$, realizations of the random variable whose density is $g(x)$, we try to estimate the density g using the model:

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

where $p(x; \alpha)$ is a univariate nonnegative polynomial over S with coefficients α, and $K(x; \beta)$ is a density function over the support S. As a base function, the density $K(x; \beta)$ is an instance of an exponential family of distributions, namely,

1. Gaussian distribution with mean μ and variance σ^2
$$N(x; \mu, \sigma^2) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in S = (-\infty, \infty).$$
2. An exponential distribution with rate parameter $\lambda > 0$
$$\mathrm{Exp}(x; \lambda) := \lambda e^{-\lambda x}, \quad x \in S = [0, \infty).$$

In the sequel, we call Gaussian distribution with a polynomial **Gaussian-based model** and an exponential distribution with a polynomial **Exponential-based model**, respectively.

One of the approaches for this problem is a maximum likelihood method (MLE). In MLEs, to estimate the model with a data set $\{x_1, \ldots, x_n\}$, it maximizes the product of the probabilities (1) as follows:

$$\max_{\alpha,\beta} \prod_{i=1}^{n} f(x_i; \alpha, \beta), \tag{2}$$

[*] This research was supported in part with Grant-in-Aid for Scientific Research(B) JP18H03206, JP21H03398 and Grant-in-Aid for Exploratory Research JP20K21792 from the Japan Society for the Promotion of Sciences.

[†] National Graduate Institute for Policy Studies, skakihara@gmail.com

[‡] National Graduate Institute for Policy Studies, tsuchiya@grips.ac.jp

or, in the form of a log likelihood:

$$\max_{\alpha,\beta} \sum_{i=1}^{n} \log f(x_i;\alpha,\beta). \tag{3}$$

The principle of MLE is that good parameters, in this case, α and β, maximize the probability of realization of observed samples $\{x_1,\ldots,x_n\}$. The log likelihood (3) is also concerned with a maximum entropy principle, and they are discussed in, for example, [Aka73; KK08].

As for a constraint condition, we need to guarantee that f is a probability density:

$$\int_S p(x;\alpha)\cdot K(x;\beta)\mathrm{d}x = 1. \tag{4}$$

In computation, it is customary to take a minus of a log likelihood, which flips the sign of the objective and change a maximization problem to a minimization problem. In addition, we introduce a regularization term rAIC(α,β) to avoid overfitting. We will discuss AIC later in detail.

The skeleton of the estimation problem becomes:

$$\begin{array}{ll} \displaystyle\min_{\alpha,\beta} & \displaystyle -\sum_{i=1}^{n}\left[\log p(x_i;\alpha)+\log K(x_i;\beta)\right]+\mathrm{rAIC}(\alpha,\beta) \\ \text{s.t.} & \begin{array}{l} p(x;\alpha)\geq 0, \quad x\in S, \\ \int_S p(x;\alpha)\cdot K(x;\beta)\mathrm{d}x = 1. \end{array} \end{array} \tag{5}$$

To compare different models, we need to consider some regularization regarding the complexities of models; otherwise complex models get better in specialization than simpler ones, but complex models fail in generalization. In our case, we adopt Akaike Information Criterion (AIC) [Aka74]. With AIC, the regularization term rAIC(α,β) are put on the objective of (5), the adjusted which penalizes the number of free parameters, i.e., the number of parameters minus the number of constraints.

In the following cases, considering the fact that the number of parameters of coefficient is $k+1$ for polynomials of degree k, the number of parameters of Gaussian distribution is 2, and that of an exponential distribution is 1, and the number of the constraints in (5), with the presence of (4), is 1, we have

1. Gaussian-based model
$$\mathrm{rAIC}(\alpha,\mu,\sigma^2) = k+2.$$
2. Exponential-based model
$$\mathrm{rAIC}(\alpha,\lambda) = k+1.$$

In practice, we restrict the set of degrees Θ_{deg} and the set of parameters of base functions Θ_β to be finite cardinality, and compute the coefficients α from these values using SDP.

Our estimation (5) is now more concisely presented as follows:

$$\begin{array}{ll} \displaystyle\min_{\substack{k\in\Theta_{\mathrm{deg}}\\ \beta\in\Theta_\beta}} \min_{\alpha\in\mathbb{R}^{k+1}} & \displaystyle -\sum_{i=1}^{n}\left[\log p(x_i;\alpha)+\log K(x_i;\beta)\right]+\mathrm{rAIC}(\alpha,\beta) \\ \qquad\qquad\text{s.t.} & \begin{array}{l} p(x;\alpha)\geq 0, \quad x\in S, \\ \int_S p(x;\alpha)\cdot K(x;\beta)\mathrm{d}x = 1, \end{array} \end{array} \tag{6}$$

where Θ_{deg} and Θ_β are finite sets.

2 Semidefinite Programming Problems

This section is the main topic of this vignette where we discuss the estimation of coefficients of polynomials with fixed degrees of polynomials and parameters of base functions.

Before that, we introduce several notations. For a positive integer d, we define a (d+1)-dimensional column vector $\mathbf{x}_d := (1, x, \dots, x^d)^T$ for $x \in \mathbb{R}$, where the superscript $\square^T$ denotes the transpose of a vector or a matrix, and also define a (d+1) by (d+1) matrix $X_d := \mathbf{x}_d \mathbf{x}_d^T$.

Q, Q_1, Q_2 and Q_3 denote symmetric matrices with appropriate sizes, and $Q \succeq 0$ denotes the semidefiniteness of the symmetric matrix Q, i.e., all eigenvalues of Q are nonnegative. $\mathrm{trace}(Q)$ denotes the trace of the matrix Q, i.e., the sum of the diagonal elements of Q.

Using these notations, nonnegativity of k-th order polynomials is expressed in quadratic forms as follows.

Proposition 2.1. *(e.g. [Nes00; FHT06]) Let $p(x;\alpha)$ be a univariate k-th degree polynomial with coefficient α. Then, there exists a unique quadratic form corresponding to α in each of the following cases.*

1. *$p(x;\alpha) \geq 0$ over $S = (-\infty, \infty)$*

- *k is even and $d = \frac{k}{2}$*

$$p(x;\alpha) = \boldsymbol{x}_d^T Q \boldsymbol{x}_d = \mathrm{trace}(X_d Q),$$

  for some unique symmetric matrix $Q \succeq 0$.

2. *$p(x;\alpha) \geq 0$ over $S = [0, \infty)$*

- *k is odd and $d = \frac{k-1}{2}$*

$$p(x;\alpha) = \boldsymbol{x}_d^T Q_1 \boldsymbol{x}_d + x \cdot \boldsymbol{x}_d^T Q_2 \boldsymbol{x}_d = \mathrm{trace}(X_d Q_1) + \mathrm{trace}(x X_d Q_2)$$

  for some unique symmetric matrices $Q_1, Q_2 \succeq 0$.

- *k is even and $d = \frac{k}{2}$*

$$p(x;\alpha) = \boldsymbol{x}_d^T Q_1 \boldsymbol{x}_d + x \cdot \boldsymbol{x}_{d-1}^T Q_3 \boldsymbol{x}_{d-1} = \mathrm{trace}(X_d Q_1) + \mathrm{trace}(x X_{d-1} Q_3),$$

  for some unique symmetric matrices $Q_1, Q_3 \succeq 0$.

Using this proposition, the positivity constraint (4) with k-th order polynomials ($k \geq 1$) is now written down to as follows.

1. Gaussian-based model

- k is even and $d = \frac{k}{2}$

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} p(x;\alpha) N(x;\mu,\sigma^2)\mathrm{d}x = \int_{-\infty}^{\infty} \mathbf{x}_d^T Q \mathbf{x}_d \cdot N(x;\mu,\sigma^2)\mathrm{d}x \\ &= \mathrm{trace}\left(\int_{-\infty}^{\infty} Q \mathbf{x}_d \mathbf{x}_d^T N(x;\mu,\sigma^2)\mathrm{d}x\right) = \mathrm{trace}\left(Q \int_{-\infty}^{\infty} X_d N(x;\mu,\sigma^2)\mathrm{d}x\right) \\ &= \mathrm{trace}(QM), \end{aligned}$$

  where $M := \int_{-\infty}^{\infty} X_d N(x;\mu,\sigma^2)\mathrm{d}x$.

2. Exponential-based model

- k is odd and $d = \frac{k-1}{2}$

$$
\begin{aligned}
1 = \int_0^\infty p(x;\alpha)\mathrm{Exp}(x;\lambda)\mathrm{d}x &= \int_0^\infty (\mathbf{x}_d^T Q_1 \mathbf{x}_d + x \cdot \mathbf{x}_d^T Q_2 \mathbf{x}_d) \cdot \mathrm{Exp}(x;\lambda)\mathrm{d}x \\
&= \mathrm{trace}\left(\int_0^\infty Q_1 \mathbf{x}_d \mathbf{x}_d^T \mathrm{Exp}(x;\lambda)\mathrm{d}x\right) + \mathrm{trace}\left(\int_0^\infty Q_2 x \mathbf{x}_d \mathbf{x}_d^T \mathrm{Exp}(x;\lambda)\mathrm{d}x\right) \\
&= \mathrm{trace}\left(Q_1 \int_0^\infty X_d \mathrm{Exp}(x;\lambda)\mathrm{d}x\right) + \mathrm{trace}\left(Q_2 \int_0^\infty x X_d \mathrm{Exp}(x;\lambda)\mathrm{d}x\right) \\
&= \mathrm{trace}(Q_1 M_1) + \mathrm{trace}(Q_2 M_2).
\end{aligned}
$$

where $M_1 := \int_0^\infty X_d \mathrm{Exp}(x;\lambda)\mathrm{d}x$, $M_2 := \int_0^\infty x X_d \mathrm{Exp}(x;\lambda)\mathrm{d}x$,

- k is even and $d = \frac{k}{2}$

$$
\begin{aligned}
1 &= \int_0^\infty p(x;\alpha)\mathrm{Exp}(x;\lambda)\mathrm{d}x \\
&= \mathrm{trace}\left(Q_1 \int_0^\infty X_d \mathrm{Exp}(x;\lambda)\mathrm{d}x\right) + \mathrm{trace}\left(Q_2 \int_0^\infty x X_{d-1} \mathrm{Exp}(x;\lambda)\mathrm{d}x\right) \\
&= \mathrm{trace}(Q_1 M_1) + \mathrm{trace}(Q_2 M_3),
\end{aligned}
$$

where $M_1 := \int_0^\infty X_d \mathrm{Exp}(x;\lambda)\mathrm{d}x$, $M_3 := \int_0^\infty x X_{d-1} \mathrm{Exp}(x;\lambda)\mathrm{d}x$.

Note that M is a moment matrix whose (i, j)-th element is an (i+j-2)-th moment of Gaussian distribution, and similarly, an (i, j)-th element of M_1 is an (i+j-2)-th moment of an exponential distribution, and an (i, j)-th element of M_2 or M_3 is an (i+j-1)-th moment of the same distribution.

We briefly summarize the computation of the moments. Using the moment generating function $M_X(t) = E[e^{tX}]$ of the random variable X, where E is an expectation operator, k-th degree of the moment m_k is obtained by differentiating $M_X(t)$ with respect to t k times at $t = 0$, i.e.,

$$
m_k := \left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0}.
$$

Thus, for each of the cases, its k-th moment is obtained as follows.

1. Gaussian distribution $N(x;\mu,\sigma^2)$
   The moment generating function is $M(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$, and its k-th moment is

$$
m_k = \left.\frac{d^k e^{\mu t + \frac{1}{2}\sigma^2 t^2}}{dt^k}\right|_{t=0}. \tag{7}
$$

   For example, we show k-th moments for $k = 0, 1, 2, 3, 4, 5$ in the table below

k	k-th moment(m_k)
0	1
1	μ
2	$\mu^2 + \sigma^2$
3	$\mu^3 + 3\mu\sigma^2$
4	$\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$

k	k-th moment(m_k)
5	$\mu^5 + 10\mu^3\sigma^2 + 15\mu\sigma^4$

2. An exponential distribution $\text{Exp}(x;\lambda) = \lambda \exp(-\lambda x)$
   The moment generating function is $M(t) = \frac{\lambda}{\lambda - t}$, for $t < \lambda$, and its k-th moment is
$$m_k = \frac{k!}{\lambda^k}. \tag{8}$$

Note that since the moment becomes easily too large as k becomes larger, a moment matrix tends to be highly ill conditioned for large degrees of polynomials, which makes the computation of coefficients difficult.

Before we show SDP formulations of density estimation, we introduce auxiliary variables $y_i = p(x_i;\alpha), i = 1,\dots,n$ and data matrices $X^{(1)},\dots,X^{(n)}$ generated by the elements of the data set $x_1,\dots,x_n$ as described in the beginning of the section.

Finally MLE (6) with a data set $x_1,\dots,x_n$ becomes as follows.

1. Gaussian-based Model
   Let Θ_{deg} be a finite set of positive even integer, Θ_μ be a finite set of real numbers, and Θ_σ be a finite set of positive real numbers.
$$\min_{\substack{k\in\Theta_{\text{deg}}\\ \mu\in\Theta_\mu\\ \sigma\in\Theta_\sigma}} \text{SDP}_{\text{Gauss}}(k,\mu,\sigma;x_1,\dots,x_n)$$
   with $\text{SDP}_{\text{Gauss}}(k,\mu,\sigma;x_1,\dots,x_n)$:
$$\begin{array}{ll} \displaystyle\min_{\substack{y_1,\dots,y_n\\ Q}} & \displaystyle -\sum_{i=1}^{n}\Big[\log y_i + \log N(x_i;\mu,\sigma^2)\Big] + (k+2) \\ \text{s.t.} & \begin{array}{l} y_i = \text{trace}(X_d^{(i)}Q), \\ y_i \geq 0, \quad i = 1,\dots,n, \\ \text{trace}(MQ) = 1, \\ Q \succeq 0, \end{array} \end{array} \tag{9}$$
   where $d = \frac{k}{2}$, $Q \in \mathbb{R}^{(d+1)\times(d+1)}$ is a symmetric matrix, and M is the moment matrix whose (i,j) element is m_{i+j-2} in (7).
2. Exponential-based model
   Let Θ_{deg} be a finite set of positive integers, and Θ_λ be a finite set of positive real numbers.
$$\min_{\substack{k\in\Theta_{\text{deg}}\\ \lambda\in\Theta_\lambda}} \text{SDP}_{\text{Exp}}(k,\lambda;x_1,\dots,x_n)$$
   with $\text{SDP}_{\text{Exp}}(k,\lambda;x_1,\dots,x_n)$:

- k is odd

$$\begin{array}{rl} \min\limits_{\substack{y_1,\ldots,y_n\\Q_1,Q_2}} & -\sum_{i=1}^{n}[\log y_i + \log \mathrm{Exp}(x_i;\lambda)] + (k+1) \\ \text{s.t.} & \begin{array}{l} y_i = \mathrm{trace}(X_d^{(i)}Q_1) + \mathrm{trace}(x_i X_d^{(i)}Q_2), \\ y_i \geq 0, \quad i=1,\ldots,n, \\ \mathrm{trace}(M_1Q_1) + \mathrm{trace}(M_2Q_2) = 1, \\ Q_1 \succeq 0, \\ Q_2 \succeq 0, \end{array} \end{array} \tag{10}$$

where $d = \frac{k-1}{2}$, $Q_1, Q_2 \in \mathbb{R}^{(d+1)\times(d+1)}$ are symmetric matrices, and $M_1, M_2 \in \mathbb{R}^{(d+1)\times(d+1)}$ are the moment matrices whose (i,j) element are m_{i+j-2} and m_{i+j-1} in (8), respectively.

- k is even

$$\begin{array}{rl} \min\limits_{\substack{y_1,\ldots,y_n\\Q_1,Q_3}} & -\sum_{i=1}^{n}[\log y_i + \log \mathrm{Exp}(x_i;\lambda)] + (k+1) \\ \text{s.t.} & \begin{array}{l} y_i = \mathrm{trace}(X_d^{(i)}Q_1) + \mathrm{trace}(x_i X_{d-1}^{(i)}Q_3), \\ y_i \geq 0, \quad i=1,\ldots,n \\ \mathrm{trace}(M_1Q_1) + \mathrm{trace}(M_3Q_3) = 1, \\ Q_1 \succeq 0, \\ Q_3 \succeq 0, \end{array} \end{array} \tag{11}$$

where $d = \frac{k}{2}$, $Q_1 \in \mathbb{R}^{(d+1)\times(d+1)}$ and $Q_3 \in \mathbb{R}^{d\times d}$ are symmetric matrices, and $M_1 \in \mathbb{R}^{(d+1)\times(d+1)}$ and $M_3 \in \mathbb{R}^{d\times d}$ are the moment matrices whose (i,j) element are m_{i+j-2} and m_{i+j-1} in (8), respectively.

References

[Aka73] Hirotugu Akaike. “Information theory and an extension of the maximum likelihood principle”. In: *Second International Symposium in Information Theory.* Ed. by B.N. Petrov and F. Caski. Budapest: Akademiai Kiado, 1973, pp. 267–281.

[Aka74] Hirotugu Akaike. “A new look at the statistical model identification”. In: *IEEE transactions on automatic control* 19.6 (1974), pp. 716–723.

[FHT06] Tadayoshi Fushiki, Shingo Horiuchi, and Takashi Tsuchiya. “A maximum likelihood approach to density estimation with semidefinite programming”. In: *Neural computation* 18.11 (2006), pp. 2777–2812.

[KK08] Sadanori Konishi and Genshiro Kitagawa. “Information criteria and statistical modeling”. In: (2008).

[Nes00] Yurii Nesterov. “Squared functional systems and optimization problems”. In: *High performance optimization.* Springer, 2000, pp. 405–440.