*RIMS Kôkyûroku Bessatsu*
**B41** (2013), 065–072

# 離散可積分系と事前情報付き線形回帰モデル
# Discrete integrable systems and linear regression with prior information

By

關戸 啓人 (Hiroto Sekido)*

**Abstract**

In design of experiments, D-optimal designs are multisets of experimental conditions which give us the highest accuracy estimators based on a particular optimality criterion. One of the approaches for D-optimal designs is to use canonical moments. On the other hand, a relationship between discrete integrable systems and canonical moments is investigated. In this paper, by using the relationship, an algorithm for calculating D-optimal designs for some linear regression models is proposed.

## §1. はじめに

本稿では，実験計画法における D-optimal design について取り上げる．ここで，design とは，回帰モデルの未知係数を推定する際の，説明変数の組のことであり，D-optimal design とは，ある基準を用いて，最も精度良く未知係数を推定されるような design である．一部のモデルに対しては，D-optimal design を解析的に，または，数値的に求める手法はいくつか提案されており，その中の一つにカノニカルモーメントを用いる方法がある．また，カノニカルモーメントは，離散 Lotka-Volterra 方程式の行列式解と見なせ，その事実を用いて，固定点を通る場合の多項式回帰モデルに対する D-optimal design を求める方法が提案されている [6].

本稿では，固定点を通る場合の多項式回帰モデルに対する D-optimal design の一般化として，事前情報付き多項式回帰モデル対する D-optimal design の計算法を提案する．また，多項式回帰モデルにおいて係数関数を推定する場合の Maximin optimal design，多

Received November 1, 2012.
2000 Mathematics Subject Classification(s): 62K05
*Key Words*: *Key Words*: D-optimal design，非自励離散時間戸田方程式，カノニカルモーメント
*Depart. of Appl. Math. and Phys., Grad. School of Informatics, Kyoto Univ., Kyoto 606-8502, Japan.
e-mail: sekido@amp.i.kyoto-u.ac.jp

項式回帰モデルに対するロバストな D-optimal design に対して，事前情報付きに拡張した場合の計算法を提案する．多項式回帰モデルにおいて係数関数を推定する場合の特別な場合の Maximin optimal design は [2]，多項式回帰モデルに対するロバストな D-optimal design は [3] によって，カノニカルモーメントを用いることにより数値的に計算する方法が提案されている．

## §2. 準備

最初に，線形回帰モデルとカノニカルモーメントの定義，および，性質を簡単に紹介する．線形回帰モデルは

$$Y = \theta^{\mathrm{T}} f(x) + \varepsilon,\ \ E[\varepsilon] = 0,\ \ V[\varepsilon] = \sigma^2$$

で表され，ここで，$f(x) = (f_0(x)\ f_1(x)\ \cdots\ f_{m-1}(x))^{\mathrm{T}}$ は既知関数からなるベクトル，$\theta = (\theta_0\ \theta_1\ \cdots\ \theta_{m-1})^{\mathrm{T}}$ は未知係数からなるベクトル，$\varepsilon$ は誤差項である．ここで，誤差項 $\varepsilon$ は，実験ごとに独立であると仮定する．有限区間 $I$ 上の確率測度全体の集合を $\mathcal{P}_I$ と書き，$\mu \in \mathcal{P}_I$ の $k$ 次モーメントを $c_k = c_k(\mu) = \int_I x^k d\mu(x)$ と書く．実験回数は $n$ 回とし，$n$ 個の説明変数 $x_1, x_2, \ldots, x_n$ はそれぞれ区間 $[0,1]$ に属さなければならないとする．ここで，design $x_1, x_2, \ldots, x_n$ と確率測度 $\mu \in \mathcal{P}_{[0,1]}$ とを，

$$\mu(\{x\}) = \frac{\#\{k \mid x_k = x\}}{n} \tag{2.1}$$

で対応付ける．D-optimal design はフィッシャーの情報行列 $M_f(x) = \int_0^1 f(x) f(x)^{\mathrm{T}} d\mu(x)$ の行列式を最大化する [0,1] 上の確率測度と定義される．特に，多項式回帰モデル，つまり $f_k(x) = x^k$ の場合の D-optimal design は，Hankel 行列式 $|c_{i+j}|_{i,j=0}^{m-1}$ を最大化する [0,1] 上の確率測度である．つまり，多項式回帰モデルに対する D-optimal design は最適化問題

$$\text{maximize } |c_{i+j}(\mu)|_{i,j=0}^{m-1} \quad \text{s.t. } \mu \in \mathcal{P}_{[0,1]} \tag{2.2}$$

の最適解として定義される．確率測度と design の対応関係 (2.1) より，$\mu(\{x\})$ は任意の $x$ で $1/n$ の倍数でなければならないが，その条件は考えず，緩和問題を解くことを考える．詳しいことは [5] などを参照されたい．

次にカノニカルモーメントを紹介する．確率測度 $\mu \in \mathcal{P}_{[0,1]}$ が与えられたとき，$c_k^+, c_k^-$ を

$$c_k^+ = \max_{\xi \in \mathcal{P}_{[0,1]}} \{c_k(\xi) \mid c_j(\xi) = c_j(\mu),\ 0 \le j < k\},$$
$$c_k^- = \min_{\xi \in \mathcal{P}_{[0,1]}} \{c_k(\xi) \mid c_j(\xi) = c_j(\mu),\ 0 \le j < k\}$$

で定め，確率測度 $\mu$ のカノニカルモーメント $p_k$ を

$$p_k = \frac{c_k - c_k^-}{c_k^+ - c_k^-},\ \ k = 1, 2, \ldots, N$$

で定義する．ここで，$N$ は $c^-_{j+1} = c^+_{j+1}$ となる最小の $j$ で，恒に $c^-_{j+1} < c^+_{j+1}$ であるなら $N = \infty$ とする．定義より，カノニカルモーメントは $0 \le p_k \le 1$ を満たす．

また，カノニカルモーメントは，モーメントのなす Hankel 行列式を用いて書き下すことができる．後々のために少し一般化した形で Hankel 行列式を定義しておく．モーメント列 $c_k$ が与えられたとき，$c^{(T)}_k$ を

$$c^{T\uplus\{\lambda\}}_k = c^{(T)}_{k+1} - \lambda c^{(T)}_k, \quad c^{(\phi)}_k = c_k \tag{2.3}$$

で定義する．ただし，上付き添え字は多重集合である．ここで，モーメントの変形 (2.3) は，直交多項式の Christoffel 変換，もしくは，非自励離散時間戸田方程式の時間発展に相当するものであることを注意しておく．また，$c^{(T)}_k$ のなす Hankel 行列式を $H^{(T)}_k = |c^{(T)}_{i+j}|^{k-1}_{i,j=0}$ と書く．ただし，$k \le 0$ のときは，$H^{(T)}_0 = 1$，$H^{(T)}_{-1} = H^{(T)}_{-2} = \cdots = 0$ とする．すると，カノニカルモーメントは Hankel 行列式表示

$$p_{2k} = -\frac{H^{(\phi)}_{k+1} H^{(\{0,-1\})}_{k-1}}{H^{(\{0\})}_k H^{(\{-1\})}_k}, \quad p_{2k+1} = -\frac{H^{(\{0\})}_{k+1} H^{(\{-1\})}_k}{H^{(\phi)}_{k+1} H^{(\{0,-1\})}_k}$$

をもち，更に $\zeta_0 = 0$，$\zeta_1 = p_1$，$\zeta_k = (1 - p_{k-1})p_k$，$k = 2, 3, \ldots, N$ で定義される量は，Hankel 行列式を用い

$$\zeta_{2k} = \frac{H^{(\phi)}_{k+1} H^{(\{0\})}_{k-1}}{H^{(\{0\})}_k H^{(\phi)}_k}, \quad \zeta_{2k+1} = \frac{H^{(\{0\})}_{k+1} H^{(\phi)}_k}{H^{(\phi)}_{k+1} H^{(\{0\})}_k}$$

と表すことができる．$\zeta_k$ がわかれば，カノニカルモーメント $p_k$ も簡単に計算できることに注意する．カノニカルモーメントのその他の性質などは [1] などを参照されたい．

## §3. 離散可積分系を用いた D-optimal design の計算法

通常の多項式回帰モデルに対する D-optimal design を求める方法として，最適化問題 (2.2) の目的関数 $H^{(\phi)}_{2m}$ をカノニカルモーメントで書きなおす方法がある [7]．一般に，目的関数をカノニカルモーメントで書きなおすことにより，制約条件が簡単になり，最適化問題を解きやすくなる．通常は，書きなおした後，数値計算で目的関数を最大化するカノニカルモーメントを求める必要があるが，通常の多項式回帰モデルの場合は解析的に計算できる．カノニカルモーメントから確率測度を計算する方法は，例えば [1] に書いてある．

本節では，最初に，事前情報付き多項式回帰モデルを定義し，その D-optimal design の計算法を提案する．その中で，事前情報を付加することと非自励離散時間戸田方程式の時間発展が対応することを用いる．次に，重み付き多項式回帰モデルにおいて係数関数を推定する場合の Maximin optimal design，事前情報付き多項式回帰モデルに対するロバストな D-optimal design の計算法を提案する．

## §3.1. 事前情報付き多項式回帰モデルに対する D-optimal design

最初に，$\zeta_k$ が非自励離散時間戸田方程式の行列式解 [4] と似ていることに着目し，非自励離散時間戸田方程式の時間発展を用いて拡張する．$\zeta_k^{(T,s)}$ を

$$\zeta_{2k}^{(T,s)} = \frac{H_{k+1}^{(T)} H_{k-1}^{(T\uplus\{s\})}}{H_k^{(T\uplus\{s\})} H_k^{(T)}},\quad \zeta_{2k+1}^{(T,s)} = \frac{H_{k+1}^{(T\uplus\{s\})} H_k^{(T)}}{H_{k+1}^{(T)} H_k^{(T\uplus\{s\})}} \tag{3.1}$$

で定義する．$\zeta_k^{(\phi,0)} = \zeta_k$ であることから，これは確かに $\zeta_k$ の拡張になっている．すると，非自励離散時間戸田方程式

$$\begin{aligned} &\zeta_{2k}^{(T\uplus\{\lambda_1\},\lambda_2)} + \zeta_{2k+1}^{(T\uplus\{\lambda_1\},\lambda_2)} + \lambda_2 = \zeta_{2k+1}^{(T,\lambda_1)} + \zeta_{2k+2}^{(T,\lambda_1)} + \lambda_1, \\ &\zeta_{2k+1}^{(T\uplus\{\lambda_1\},\lambda_2)} \zeta_{2k+2}^{(T\uplus\{\lambda_1\},\lambda_2)} = \zeta_{2k+2}^{(T,\lambda_1)} \zeta_{2k+3}^{(T,\lambda_1)} \end{aligned} \tag{3.2}$$

が成り立つことがわかる．また，(3.2) より，

$$\begin{aligned} &\zeta_{2k}^{(T,\lambda_1)} + \zeta_{2k+1}^{(T,\lambda_1)} + \lambda_1 = \zeta_{2k}^{(T,\lambda_2)} + \zeta_{2k+1}^{(T,\lambda_2)} + \lambda_2, \\ &\zeta_{2k+1}^{(T,\lambda_1)} \zeta_{2k+2}^{(T,\lambda_1)} = \zeta_{2k+1}^{(T,\lambda_2)} \zeta_{2k+2}^{(T,\lambda_2)} \end{aligned} \tag{3.3}$$

も成り立つことがわかる．なお，(3.2) の双線形形式は

$$H_{k-1}^{(T\uplus\lambda_1)} H_{k+1}^{(T\uplus\lambda_1)} - H_k^{(T)} H_{k+2}^{(T\uplus\lambda_1\uplus\lambda_2)} + H_k^{(T\uplus\lambda_1)} H_k^{(T\uplus\lambda_2)} = 0$$

となる．また，$\zeta_k^{(T,s)}$ の定義 (3.1) より，Hankel 行列式を $\zeta_k^{(T,s)}$ で書き表すと

$$H_k^{(T)} = \left(c_0^{(T)}\right)^k \prod_{j=1}^{k-1} \left(\zeta_{2j-1}^{(T,0)} \zeta_{2j}^{(T,0)}\right)^{k-j} \tag{3.4}$$

となる．

次に事前情報付き多項式回帰モデルと，その D-optimal design を定義する．$m+S-1$ 次多項式回帰モデル $Y = \sum_{k=0}^{m+S-1} \theta_k x^k + \varepsilon$ において，$S$ 個の値

$$\frac{d^k}{dx^k} g(x)|_{x=\beta_j},\quad 1 \le j \le l,\ \ 0 \le k < b_j \tag{3.5}$$

が既知の場合を考える．ただし，$g(x) = E[Y|x] = \sum_{k=0}^{m+S-1} \theta_k x^k$，$S = \sum_{j=1}^{l} b_j$ であり，$b_1, b_2, \ldots, b_l$ は正整数，$\beta_1, \beta_2, \ldots, \beta_l$ は相異なる $l$ 個の実数である．事前情報付き多項式回帰モデルを線形回帰モデルとして定式化する方法は一意ではないが，線形回帰モデルに対する D-optimal design は基底関数 $f_k(x)$ のはる空間のみに依存することから，D-optimal design は一意に定まることがわかる．具体的には，D-optimal design は以下の定理で定義される．

**Theorem 3.1.** 事前情報付き多項式回帰モデルの *D-optimal design* は最適化問題

$$\text{maximize } H_m^{(T)}(\mu) \quad \text{s.t. } \mu \in \mathcal{P}_{[0,1]}$$

の最適解である．ただし，$T$ は $\beta_k$ をちょうど $2b_k$ 個だけ含むような多重集合である．

この定理は事前情報 (3.5) を多項式回帰モデルに代入し，変数変換することで得られる．

次に，D-optimal design を計算するために，目的関数 $H_m^{(T)}$ をカノニカルモーメントで書きなおすことを考える．目的関数は，(3.4) により，$c_0^{(T)}$ と $\zeta_k^{(T,0)}$ で書き表すことができる．$\zeta_k^{(T,0)}$ を $\zeta_k^{(\phi,0)}$，つまり $\zeta_k$ で書きなおすには，非自励離散時間戸田方程式 (3.2), (3.3) を繰り返し用いれば良い．また，$c_0^{(T)}$ に関して，漸化式

$$\zeta_1^{(T,s)} = \frac{c_0^{(T \uplus \{s\})}}{c_0^{(T)}} \tag{3.6}$$

を用いることで，同様に $\zeta_k$ で書きなおすことができる．

まとめると，事前情報付き多項式回帰モデルに対する D-optimal design は以下の方法で計算できる．

Step 1. (3.4) を用いて，目的関数を $c_0^{(T)}$ と $\zeta_k^{(T,0)}$ で書き表す．

Step 2. 非自励離散時間戸田方程式 (3.2)，(3.3) と，漸化式 (3.6) を用いて，目的関数を $\zeta_k$ で書きなおす．

Step 3. 目的関数をカノニカルモーメントで書きなおし，目的関数を最大化するカノニカルモーメントを求める．

## §3.2. 重み付き多項式回帰モデルにおいて係数関数を推定する場合の Maximin optimal design

事前情報付き多項式回帰モデルは，D-optimal design を考える上では，重み付き多項式回帰モデルとみなすこともできる [6]．事前情報付き多項式回帰モデルは定式化の際に未知係数を変数変換しており，係数関数を推定する場合には意味づけがわかりにくくなるため，本項では，重み付き多項式回帰モデルとして考える．

本項で考える問題は，重み付き多項式回帰モデル

$$Y = \sum_{k=0}^{m-1} \theta_k x^k + \varepsilon,$$

$$E[\varepsilon] = 0, \ \ V[\varepsilon] = \frac{\sigma^2}{w(x)}, \ \ w(x) = \prod_{k=1}^{l} (x - \beta_k)^{2b_k}$$

において，係数関数 $\sum_{k=0}^{m-1} g_k(\theta_k)$ を推定する場合の optimal design を考える．ただし，$g_k(\theta_k)$ は多項式とする．

このとき，推定量の分散の逆数は，design を $\mu$ とすると

$$\gamma(\mu,\theta)=\sum_{k=0}^{m-1} g_k'(\theta_k)^2\psi_k^{(1)}(\mu)$$

となり，これを最大化するのが目標である．ただし，$\psi_k^{(1)}(\mu)=H_{k+1}^{(T)}/H_k^{(T)}$ で，$T$ は $\beta_k$ をちょうど $2b_k$ 個だけ含む多重集合である．ここで，分散は design $\mu$ だけではなく，未知係数 $\theta_k$ にも依存することから，最適化問題

$$\text{maximize}\ \inf_{\theta\in\Theta}\gamma(\mu,\theta)\quad \text{s.t. } \mu\in\mathcal{P}_{[0,1]}$$

の最適解で定義される Maximin optimal design を求めることを考える．ただし，$\Theta$ は未知係数の取りうると思われる値からなる集合で，ここでは $\Theta=\{\theta\mid s_k\le\theta_k\le t_k\}$ とする．未知のパラメータ $\theta$ が最悪の場合に，最適な推定精度を得るという意味で Maximin という言葉が使われており，次の節で議論する D-optimal design とは別の意味でロバストな optimal design である．

ここで，[2, Theorem 3.1] を用いると，最適化問題

$$\begin{aligned}&\text{maximize}\ \int_\Theta \gamma(\mu,\theta)^p d\pi(\theta)\quad \text{s.t. } \mu\in\mathcal{P}_{[0,1]},\\ &d\pi(\theta)=d\theta\prod_k \frac{d}{d\theta_k}g_k'(\theta_k)^2\end{aligned}\tag{3.7}$$

の解は，$p\to-\infty$ で Maximin optimal design に弱収束することがわかり，十分小さな $p$ に対し，最適化問題 (3.7) を解けば良いことになる．ただし，$d\pi$ の定義の右辺の積は，$g_k'(\theta_k)^2$ が定数とならない全ての $k$ に対する積である．最適化問題 (3.7) の目的関数の積分は解析的に可能であり，目的関数に含まれる $\psi_k^{(1)}(\mu)=H_{k+1}^{(T)}/H_k^{(T)}$ は，3.1 項と同様の方法によりカノニカルモーメントで書きなおすことができる．

### §3.3. 事前情報付き多項式回帰モデルに対するロバストな D-optimal design

最初に，いくつか記号の定義をする．確率測度 $\mu(x)\in\mathcal{P}_{[0,1]}$ に対して，変数変換 $x=y^2$ により，原点について対称な確率測度 $\mu'(y)\in\mathcal{P}_{[-1,1]}$ が一意に定まる．$\mu'$ の $k$ 次モーメントを $c_k'=c_k(\mu')$ と書く．

本項では，対象が多項式とわずかに異なる場合を考える．具体的には，回帰モデル

$$\begin{aligned}&Y=\sum_{k=0}^{m+S-1}\theta_k x^k+x^{m+S}\psi(x)+\varepsilon,\\ &E[\varepsilon]=0,\ \ V[\varepsilon]=\sigma^2\end{aligned}$$

において，$S$ 個の値

$$\frac{d^k}{dx^k}g(x)|_{x=-\beta_j},\ \ \frac{d^k}{dx^k}g(x)|_{x=\beta_j},\ \ 1\le j\le l,\ \ 0\le k<b_j$$

が既知の場合を考える．ただし，$g(x)=\sum_{k=0}^{m+S-1}\theta_k x^k$，$S=\sum_{j=1}^{l}2b_j$ で，$\beta_1,\beta_2,\ldots,\beta_l$ は相異なる $l$ 個の正定数，$\psi(x)$ は未知関数である．また，本項では，説明変数の取りうる範囲を $[0,1]$ ではなく，$[-1,1]$ とする．事前情報は原点について対称に与えられていることに注意する．事前情報がない場合の D-optimal design の定式化については [3] なされており，同様の方法を用いて定式化してやると，事前情報付き多項式回帰モデルに対するロバストな D-optimal design は最適化問題

$$
\begin{aligned}
&\text{maximize } H_m^{(T')}\\
&\quad\text{s.t. } \mu'\in\mathcal{P}_{[-1,1]},\\
&\qquad {}^{\forall}x\in[-1,1],\ |\psi(x)|\le|x|^{\alpha}\ \text{となる任意の連続関数}\ \psi(x)\ \text{に対し}\\
&\qquad\quad (r(\psi)^{(T')})^{\mathrm{T}}(B_m^{(T')})^{-1}r(\psi)^{(T')}\le d.
\end{aligned}
\tag{3.8}
$$

の最適解として定義される．ただし，$T'$ は $\pm\beta_k$ がちょうど $2b_k$ 個ずつ含む多重集合で，

$$
\begin{aligned}
&B_m^{(T')}=({c'}_{i+j}^{(T')})_{i,j=0}^{m-1},\\
&r(\psi)^{(T')}=\int_{-1}^{1}(1\ x\ \cdots\ x^{m-1})^{\mathrm{T}}x^m\psi(x)\left(\prod_{j=1}^{l}(x-\beta_j)(x+\beta_j)\right)d\mu(x),
\end{aligned}
$$

$\alpha$ は与えられた非負整数，$d$ は与えられた正定数である．この D-optimal design の正確な意味や，$\alpha,d$ の意味については，[3] を参照されたい．

最適化問題 (3.8) の最適解は，モデルの対称性と，最適化問題の凸性より，原点について対称な確率測度となることがわかる．その事実を用いて，最適化問題 (3.8) の目的関数と制約条件をカノニカルモーメントで書きなおしてやる．それは，事前情報がない場合は [3] なされているが，事前情報がある場合は，似たような計算と，3.1 項と同様の方法により達成される．目的関数を書きなおすのは簡単なので，制約条件を書きなおした結果のみ以下に述べておく．$S_{i,j}^{(T)}$ を漸化式

$$
\begin{cases}
S_{i,j}^{(T)}=0 & (j<i)\\
S_{i,j}^{(T)}=1 & (i=0,j>0)\\
S_{i,j}^{(T)}=S_{i,j-1}^{(T)}+\zeta_{j-i+1}^{(T)}S_{i-1,j}^{(T)} & (\text{otherwise})
\end{cases}
$$

で定義すると，

$$
\sup_{\psi^{(T')}}(r(\psi)^{(T')})^{\mathrm{T}}(B_m^{(T')})^{-1}r(\psi)^{(T')}=(c_0^{(T)})^{-1}\sum_{i=\lfloor\alpha/2\rfloor+1}^{\lfloor(m+\alpha)/2\rfloor}(S_{i,m+\alpha-i}^{(T)})^2\prod_{j=1}^{m+\alpha-2i}\zeta_j^{(T)}
$$

が成り立つ．ただし，$T$ は $\beta_k^2$ をちょうど $2b_k$ 個だけ含む多重集合であり，$c_0^{(T)}$，$\zeta_k^{(T)}$ は $x=y^2$ で変数変換された確率測度 $\mu(x)\in\mathcal{P}_{[0,1]}$ に対する量である．

## §4. まとめ

離散可積分系の時間発展を用いて一般化されたカノニカルモーメントと非自励離散時間戸田方程式を用いた，事前情報付き多項式回帰モデルの D-optimal design の計算法を提案した．また，カノニカルモーメントを用いて計算されていた様々な D-optimal design に対し，計算可能なモデルのクラスを広げることができ，その例として，重み付き多項式回帰モデルにおいて係数関数を推定する場合の Maximin optimal design，事前情報付き多項式回帰モデルに対するロバストな D-optimal design の計算法を提案した．

## References

[1] Dette H., Studden W., *The Theory of Canonical Moments with Applications in Statistics, Probability and Analysis*, Wiley, New York, 1997.
[2] Dette H., Haines L. M., Imhof L. A., Maximin and Bayesian optimal designs for regression models, *Statistica Sinica* **17** (2007), 463–480.
[3] Fang Z., Wiens D. P., Robust regression designs for approximate polynomial models, *J. Stat. Plann. Inference* **117**, (2003), 305–321.
[4] Hirota R., Conserved quantities of "random-time Toda equation", *J. Phys. Soc. Japan* **66**, (1997), 283–284.
[5] Pukelsheim F., *Optimal Design of Experiments*, Wiley, New York, 1993.
[6] Sekido H., An algorithm for calculating *D*-optimal designs for polynomial regression through a fixed point, *J. Stat. Plann. Inference* (2012), 935–943.
[7] Studden W. J., $D_s$-optimal designs for polynomial regression using continued fractions, *Ann. Stat.* **8** (1980), 1132–1141.