

2011 年度修士論文

Border Bases による
代数曲線回帰について

早稲田大学基幹理工学術院 修士課程 2 年

5110A030-8

手島 悠人

指導教員名 楫 元

Abstract

\mathbb{R}^2 上にいくつか観測値 (x_i, y_i) が与えられているとする。統計学では、これらの観測値を何らかの代数曲線で近似するという問題が良く扱われる。既存の方法として最小二乗法があり、多くの相関分析の手法がこれに帰着される。これらはまず代数曲線の次数を適当に決めた後、最もデータと良く当てはまるよう係数を決めていくものである。ここで、代数曲線の次数は“人間の目”を用いて決めなくてはならない。つまり、どのような代数曲線に近似させるかを決めなくてはならず、その数学的な方針は無い、という欠点を持つ。

この論文では border bases という道具を用いる事により、与えられた観測値を次数を仮定せずに代数曲線で回帰する方法を提案しそれを考察する。

1 Introduction

統計学とは、ある対象について法則性を見出し、確定する事が目的である。統計学の重要なテーマの一つに、実験的に得られたデータに最も良く当てはまるような曲線を求める曲線回帰がある。これは学問的にも実務的にも大変必要性の高いものである。

曲線回帰には多くの手法がある。最も良く用いられるのは最小二乗法であり、多くの手法はこれに帰着される。具体的には後述するが、これは次のようなステップで計算される。

step1. データを収集する。

step2. 回帰すべきモデル（本論文では代数曲線に限る）を定める。つまり代数曲線の次数を決める。

step3. モデルの係数を定める。

ここで step2 についてあるが、曲線の次数を求める際の数学的な指針は無い。計算する人の経験に基づいて行う必要がある。例えば2次曲線とも3次曲線とも考えられるような観測値の扱いは難しい。

本論文は、border bases という代数幾何学や計算代数の分野で近年注目され始めた道具を用いて、曲線回帰を行う方法を提案する。この手法を border bases 法と呼ぶ事にする。border bases 法は step2 で行う次数の決定を数学的な指標に基づいて行う事ができる。それと同時に、十分に観測値が多く与えられていれば、任意の次数の曲線で回帰を行う事ができる。この論文では border bases 法の理論的背景と計算方法を述べ、いくつかの例を用いてこの手法の長所短所、課題などを考察していく。その結果、border bases 法は曲線回帰の方法として、問題点や課題がまだまだ多い反面、独自の強みを持つものであることがわかる。

2 Border Bases と諸定義

この章では border bases の定義と基本的な性質について述べる。以下、 $P = \mathbb{R}[x_1, \dots, x_n]$ とする。また \mathbb{T}^n を power product 全体、即ち、 $\mathbb{T}^n = \{x_1^{a_1} \cdots x_n^{a_n} \mid a_i \in \mathbb{Z}_{\geq 0}\}$ とする。

Definition 2.1 $\mathbb{X} = \{p_1, \dots, p_s\}$ を \mathbb{R}^n の空でない部分集合とし、 $G = \{g_1, \dots, g_k\}$ を P の空でない部分集合とする。

(a) $\mathcal{I}(\mathbb{X}) = \{f \in P \mid f(p_i) = 0 \forall p_i \in \mathbb{X}\}$ を **vanishing ideal of \mathbb{X}** という。

(b) \mathbb{R} -linear map として $\text{eval}_{\mathbb{X}} : P \rightarrow \mathbb{R}^s$ を $\text{eval}_{\mathbb{X}}(f) = (f(p_1), \dots, f(p_s))$ で定義する。この \mathbb{R} -linear map を \mathbb{X} に関する **evaluation map** という。 $\text{eval}_{\mathbb{X}}(f)$ を $f(\mathbb{X})$ と書くこともある。

(c) G の \mathbb{X} に関する **evaluation matrix** とは

$$M_G(\mathbb{X}) = \begin{bmatrix} g_1(p_1) & \cdots & g_k(p_1) \\ \vdots & \ddots & \vdots \\ g_1(p_s) & \cdots & g_k(p_s) \end{bmatrix} \in \text{Mat}_{s,k}(\mathbb{R})$$

である。

Definition 2.2 \mathcal{O} を \mathbb{T}^n の空でない部分集合とする。

(a) \mathcal{O} の **closure** とは、ある \mathcal{O} の元を割り切るような power product 全体の事である。 $\overline{\mathcal{O}}$ と書く。

(b) $\mathcal{O} = \overline{\mathcal{O}}$ であるとき、 \mathcal{O} を **order ideal** という。

(c) $I \subseteq P$ を zero-dimensional ideal とし、 $s = \dim(P/I)$ とする。このとき、 \mathcal{O} が order ideal でその剰余類が P/I の (\mathbb{R} の vector space としての) basis であるとき、 \mathcal{O} を **quotient basis for I** という。

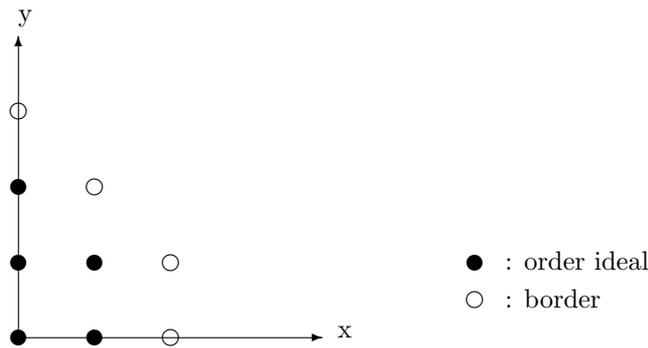
(d) \mathcal{O} を order ideal とする。 \mathcal{O} の **border** とは、

$$\partial\mathcal{O} = (x_1\mathcal{O} \cup \dots \cup x_n\mathcal{O}) \setminus \mathcal{O}$$

である。

次の例をみると, order ideal と border が直感的に理解できると思う.

Example 2.3 $\mathcal{O} = \{1, x, y, xy, y^2\}$ とすると, これは order ideal であり, border は $\partial\mathcal{O} = \{x^2, x^2y, xy^2, y^3\}$ となる. これは下の図のようになる. 横軸は x の指数, 縦軸は y の指数である.



Definition 2.3 $\mathcal{O} = \{t_1, \dots, t_\mu\}$ を order ideal とし, その border を $\partial\mathcal{O} = \{b_1, \dots, b_\nu\}$ とする. $g_i = b_j - \sum_{i=1}^{\mu} \alpha_{ij} t_i$ とし, $\mathcal{B} = \{g_1, \dots, g_\nu\}$ とおく. (ただし $\alpha_{ij} \in \mathbb{R}$). $I \subseteq P$ を \mathcal{B} を含む ideal とする. このとき, \mathcal{O} が, P/I の basis であるとき, \mathcal{B} を \mathcal{O} に基づく I の border basis という. 簡単に, I の \mathcal{O} -border basis という.

Proposition 2.4 (Existence and Uniqueness of Border Bases)

$I \subseteq P$ を zero-dimensional ideal とし, $\mathcal{O} = \{t_1, \dots, t_\mu\}$ を quotient basis for I とする. このとき, I の \mathcal{O} -border basis は一意的に存在する.

proof. [2] の Proposition 6.4.17 を見よ.

Proposition 2.5 (The Border Basis Algorithm)

zero dimensional ideal $I \subset P$ が与えられているとき, I の border basis を求めるアルゴリズムが存在する.

proof. [2] の Theorem 6.4.36 を見よ

最後に, 次の section の準備として admissible perturbation 等を定義しておく. まず, \mathbb{R}^n 上の norm を定義しておこう. $v = {}^t(v_1, \dots, v_n) \in \mathbb{R}^n$ とし, E を成分が正の $n \times n$ 対角行列とする. このとき, 以下のように norm を定義する.

$$\|v\| = \sqrt{\sum_{j=1}^n v_j^2}$$

$$\|v\|_E = \|Ev\|$$

Definition 2.6 p を \mathbb{R}^n 上の点, $\varepsilon \in \mathbb{R}^+$ とし, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ とする. このとき, 組 (p, ε) を **empirical point** と言い, p^ε と書く. また, p を **specified value**, ε を **tolerance** と言う.

Definition 2.7 p^ε を empirical point とし, tolerance を $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ とする. E を対角成分が正の \mathbb{R} 上 $n \times n$ 行列とし, $E = \text{diag}(1/\varepsilon_1, \dots, 1/\varepsilon_n)$ とする. p^ε の **ellipsoid of perturbations** を次のように定義する.

$$N(p^\varepsilon) = \{\tilde{p} \in \mathbb{R}^n \mid \|\tilde{p} - p\|_E \leq 1\}$$

Definition 2.8 $\mathbb{X} = (p_1, \dots, p_s)$ を \mathbb{R}^n の部分集合とする. ε を tolerance とし, empirical point の集合 $\mathbb{X}^\varepsilon = \{p_1^\varepsilon, \dots, p_s^\varepsilon\}$ を考える. このとき,

$$(\tilde{p}_1, \dots, \tilde{p}_s) \in \prod_{i=1}^s N(p_i^\varepsilon)$$

を満たす \mathbb{R}^n の部分集合 $\tilde{\mathbb{X}} = \{\tilde{p}_1, \dots, \tilde{p}_s\}$ を \mathbb{X}^ε の **admissible perturbation** と言う.

つまり, **admissible perturbation** とは \mathbb{X} を微小変化させたもの全体である.

Definition 2.9 $p_1, p_2 \in \mathbb{R}^n$ の empirical point $p_1^\varepsilon, p_2^\varepsilon$ が

$$N(p_1^\varepsilon) \cap N(p_2^\varepsilon) = \emptyset$$

を満たす時, $p_1^\varepsilon, p_2^\varepsilon$ は **distinct** であるという.

3 Stable Border Bases

この section では, section2 で定義した border bases 等の道具を用いて, 本論文の中心的役割を担う stable border bases を定義し, その性質について述べる. 以下では, $\mathbb{X} = \{p_1, \dots, p_s\} \subset \mathbb{R}^n$ とし, その empirical point の集合を $\mathbb{X}^\varepsilon = \{p_1^\varepsilon, \dots, p_s^\varepsilon\} \subset \mathbb{R}^n$ とし, さらに \mathbb{X}^ε の admissible perturbation を $\tilde{\mathbb{X}}$ とする. また, \mathbb{X} の任意の二つの元は ε において distinct であるとする.

Definition 3.1 \mathcal{O} を order ideal とする. このとき, \mathcal{O} が \mathbb{X}^ε に関して stable とは, \mathbb{X}^ε の 任意の admissible perturbation $\tilde{\mathbb{X}}$ において, evaluation matrix $M_{\mathcal{O}}(\tilde{\mathbb{X}})$ が full rank である事を言う.

この定義の条件で, 何故 \mathcal{O} が “ stable ” と呼ばれるかについては後述する.

Proposition 3.2 $\mathcal{O} = (t_1, \dots, t_s)$ を $\mathcal{I}(\mathbb{X})$ の quotient basis とし, \mathbb{X}^ε に関して stable とする. \mathcal{O} の border を $\partial\mathcal{O} = \{b_1, \dots, b_\nu\}$ とする. このとき, 任意の \mathbb{X}^ε の admissible perturbation $\tilde{\mathbb{X}}$ について, $\mathcal{I}(\tilde{\mathbb{X}})$ は \mathcal{O} -border basis を持つ. さらに, $\mathcal{I}(\tilde{\mathbb{X}})$ の border basis は, $\tilde{\mathcal{B}} = \{g_1, \dots, g_\nu\}$ である. ただし, $g_i = b_j - \sum_{i=1}^s \alpha_{ij} t_i$ であり, $b_j(\tilde{\mathbb{X}}) = \sum_{i=1}^s \alpha_{ij} t_i(\tilde{\mathbb{X}})$ が成り立つ (α_{ij} はある \mathbb{R} の元).

proof. [3] の Proposition 3.2 を見よ.

この Proposition から \mathcal{O} が stable である時, 次の事が分かる. \mathbb{X} と, それを微小変化させた, $\tilde{\mathbb{X}}$ においても, vanishing ideal の border basis は, border 以外の項の係数部分が変化するだけなのである. さらにこの係数部分の変化も, tolerance ε に依存するものであり, これが十分小さければ係数の変化も小さくなるのである (この議論については [3] の Proposition 3.2 の後の部分を参照). border basis が安定的と言う意味で, \mathcal{O} は stable なのである.

Definition 3.3 \mathcal{O} を $\mathcal{I}(\mathbb{X})$ の quotient basis とする. この時, \mathcal{O} が \mathbb{X}^ε に関して stable のとき, $\mathcal{I}(\mathbb{X})$ の \mathcal{O} -border basis は \mathbb{X}^ε に関して stable であるという. \mathcal{O} が明確である場合は, $\mathcal{I}(\mathbb{X})$ の \mathbb{X}^ε に関する stable order basis と言う事もある.

ここまでの議論をまとめると, $\mathcal{I}(\mathbb{X})$ の \mathbb{X}^ε に関する stable border basis を求めるには, 結局 quotient かつ \mathbb{X}^ε に関して stable な border を見つければ良い事になる. さらに次の命題で, 任意の \mathcal{O} -border basis は tolerance を十分小さく取れば \mathbb{X}^ε に関して stable にできる.

Definition 3.4 \mathcal{O} を $\mathcal{I}(\mathbb{X})$ の quotient basis とする. このとき, \mathcal{O} が \mathbb{X}^δ に

に関して, stable になるような tolerance δ が存在する.

proof. [3] の Proposition 3.4 を見よ.

Proposition 3.5 \mathbb{X} と tolerance ε を与えると, stable order ideal \mathcal{O} を与えるアルゴリズムが存在する. 又, $\#\mathbb{X} = \#\mathcal{O}$ であるとき, $\mathcal{I}(\mathbb{X})$ は \mathbb{X}^ε に関して stable な border basis を持つ.

proof. [3] の Proposition 4.3 と Theorem 4.4 をみよ.

ここで, stable order ideal が与えられているとき, stable border basis を求めることが出来る. ([3] 参照)つまり, 任意の \mathbb{X} に対して, 十分小さな tolerance ε を与えれば, \mathbb{X} に関して stable な border basis を求められる事が分かった.

4 主結果：代数曲線回帰の手法と考察

メインテーマであるこの section では, section3 の stable border bases を用いて代数曲線を近似する手法を述べ, といくつかの例からそれを考察する.

4.1 代数曲線回帰の方法

まず, どのような手順で代数曲線回帰を行うのかを述べる. 何が問題となっているのかははっきりさせておく.

Question 観測値 $\tilde{\mathbb{X}} = \{\tilde{p}_1, \dots, \tilde{p}_s\} \subset \mathbb{R}^n$ が与えられている. $\tilde{\mathbb{X}}$ はある代数曲線上の点の集合であると考えられるが, 多少のずれが生じているとする. このとき, ある代数曲線を求めよ.

section3 で, 任意の点集合は stable border basis をもつことを述べた. さらに, \mathcal{O} が stable である時, \mathbb{X} と, それを微小変化させた, $\tilde{\mathbb{X}}$ においても, vanishing ideal の border basis は, 係数が微小に異なるだけである事も分かっている. つまり, 次のような手順で近似的に回帰を行う事が出来る.

step.1 tolerance として十分小さな適当な値を取る.

step.2 観測値の stable border basis を求める. アルゴリズムがそれを返さない場合は tolerance としてさらに小さい値を選びなおす.

step.3 係数が小さい項を無視し, 最も良く観測値を反映していると考えられる多項式を選び出す.

実際に近似を行う際は工夫が必要であるが, それについては以下の具体例で説明する.

4.2 Examples

では, 具体例を用いて代数曲線回帰を実行してみよう. 比較のため, 最小二乗法で近似した場合の結果も記載しておく. ここでは簡単のため \mathbb{R}^2 上のデータを扱い $\mathbb{R}[x, y]$ 上の多項式で定義される代数曲線でのみ考える. また, 計算には ApCoCoA を用いた. 以下の全ての Example について, border bases 等の詳しいデータは論文の最後に資料として記した.

Example.1. 直線回帰

観測値 $\tilde{\mathbb{X}} = \{(0, 0), (1, 3), (2, 5.9), (3.1, 9)\}$ が与えられている. stable order ideal は

$$\{1, y, y^2, y^3\}$$

stable border basis は

$$B = \left\{ \begin{array}{l} x + 0.000031y^3 - 0.00022y^2 - 0.327y, \\ xy - 0.00017y^3 - 0.330y^2 + 0.0049y, \\ xy^2 - 0.360y^3 + 0.168y^2 - 0.267y, \\ y^4 - 17.9y^3 + 97.8y^2 - 159.3y, \\ xy^3 - 6.27y^3 + 34.93y^2 - 57.32y \end{array} \right\}$$

となる. なお, border basis の計算結果の数字に関しては, 煩雑さを防ぐため近似値を記している. ここで, stable border basis の一番目の式に注目してみる. 小さい係数を持つ項を無視することにより,

$$x + 0.000031y^3 - 0.00022y^2 - 0.327y \approx x - 0.327y$$

と近似できる. 実際, 観測値は $x - 0.327y$ で定義できる曲線に“ほぼ”乗っている事が分かる.

一方, 最小二乗法による回帰では,

$$y = 2.90 - 0.0494$$

という結果が得られる.

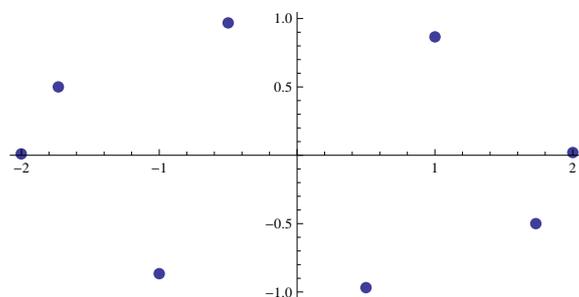
この例では, xy 平面上で傾きが 3 で, 原点を通る直線上の点を少しずらしたものを観測値として与えた. border bases 法でも最小二乗法でも良い近似が得られたと言えるだろう.

Example.2 二次曲線回帰 (楕円)

観測値

$$\tilde{X} = \{(1, 0.86603), (-1, -0.86603), (1.73205, -0.5), (-1.73205, 0.5), (2, 0.02), (-2.001, 0.01), (0.5, -0.96825), (-0.5, 0.96825)\}$$

が与えられている.



stable order ideal は

$$\{1, x, y, y^2, xy, x^2, y^3, xy^2\}$$

- 本論文では $\mathbb{R}[x, y]$ 上でのみ Example で検証したが, $\mathbb{R}[x_1, \dots, x_n]$ 上でも全く同様に, 理論上は曲線回帰は可能である. ただし, 処理量はかなり増加する.
- Example2 の結果のように, 得られるであろう代数曲線が既約で無い多項式の因数として現れるようなケースがある. 実験の結果, 観測値として与えられる点の量が多い場合はこのようなケースは少なくなると予想できたが, 数学的な説明を与えるには至っていない.
- 観測値として与えられる点が多くなると, border bases の計算量も増加し, 求める border bases の個数も増加する. これは, Proposition 3.5 の後半部分と, order ideal が大きくなれば その border も大きくなるからである. 結果, border bases のどの多項式を見れば良い回帰式が得られるのか調べる時間もかかる. 近似の信頼性を向上させるために観測値を増やす事は必要であるが, 処理に時間がかかるようになってしまう. 対策として, あらかじめ, 「出来るだけ次数の小さい代数曲線で回帰する」等の方針を決めておけば計算量を抑える事が出来る. また, step3 にもう少し明確な良い基準を与え, 回帰曲線の候補を返すアルゴリズムをつくる等の対処法も考えられる.
- 検定統計量が分からないため, 得られた代数曲線の確からしさを統計学的に測る事が出来ない. これは大きな課題と言える. (検定統計量については適当な統計の教科書等を参考にしてほしい)
- 観測値の偏りに影響を受けやすい. border bases 法は, 曲線の形を決めないため, 偏ったデータや他の数値と大きく異なるものが一つでもあると border bases を大きく変えてしまう. とびぬけた値であれば不適切なデータとして最初から観測値から除くことも考えられる.
- distinct で無くてはならないため, 同一の点 2 つ以上存在する時に, border bases を求める事が出来ない. この場合は微小に点を動かして, distinct になるように観測値に補正を加えれば良いと考えられる.

以上のように, 問題点や課題はまだ多いが, border bases 法は曲線回帰の方法として独自の強みを持つものである.

Acknowledgements

修士論文を執筆するにあたり, 毎週のセミナーだけでなく数多くのご助言をして頂いた楫元先生に心より感謝を申し上げます. また, 有益なアドバイスをして頂いた永井保成先生にも感謝を申し上げます.

楫研究室の同期, 先輩にはセミナーや論文執筆, それ以外の部分でも大変お世話になりました. この場を借りてお礼申し上げます.

References

- [1] Kreuzer, M., Robbiano, L. (2005). *Commutational Commutative Algebra 1*. Springer, Berlin.
- [2] Kreuzer, M., Robbiano, L. (2005). *Commutational Commutative Algebra 2*. Springer, Berlin.
- [3] J. Abbott, M. Fassino, and M. Torrente, (2007). *Stable Border Bases for Ideals of Points*. arXiv:0706.2316v2.
- [4] CoCoA Team. CoCoA: a system for doing computations in Commutative Algebra. Available at <http://cocoa.dima.unige.it/>
- [5] 東京大学教養学部統計学教室 (1991). 統計学入門. 東京大学出版会.
- [6] 田島 稔, 小牧 和雄 (1986). 最小二乗法の理論とその応用. 東洋書店.

資料

Examples で扱った stable border bases 等について ApCoCoA で計算した結果を資料として載せておく. Example.1 については出力そのまま載せておくが, 見やすさを考慮して Example.2 と 3 については係数を近似した値を載せておく. 以下, \mathcal{B} は stable border basis, \mathcal{O} は stable order ideal, $\tilde{\mathcal{X}}$ は観測値である.

Example.1

$$\tilde{\mathcal{X}} = \{(0, 0), (1, 3), (2, 5.9), (3.1, 9)\}$$

$$\mathcal{O} = \{1, y, y^2, y^3\}$$

$$\mathcal{B} = \{$$

77682742979400282241480559762643433765773088184

0532865211887100614688707/25000000000000000000

00

000000000 y^3

-27809112726912165082636981733906754174094533

43919134533941947075183628039/1250000000000000

00

00000000000 y^2

+ x

-163469402076800127364784893866170614346553714

212695001141674469854556956987/5000000000000000

00

00000000000 y ,

-1668520578420467185761957730812013348164627363

737486095661846496106785317/100000000000000000

00

000000000 y^3

Example.2

$$\begin{aligned}\tilde{\mathbb{X}} = \{ & (1, 0.86603), (-1, -0.86603), (1.73205, -0.5), \\ & (-1.73205, 0.5), (2, 0.02), (-2.001, 0.01), \\ & (0.5, -0.96825), (-0.5, 0.96825)\}\end{aligned}$$

$$\mathcal{O} = \{1, x, y, y^2, xy, x^2, y^3, xy^2\}$$

$$\begin{aligned}\mathcal{B} = \{ & x^2y + 0.0021871337xy^2 + 3.9957516075y^3 + 0.9466689208x^2 \\ & - 0.0029307704xy + 3.7882554218y^2 + 0.0006956227x \\ & - 3.9946312852y - 3.7896060902, \\ & x^3 + 4.0039960730xy^2 + 0.0075951544y^3 + 1.0052483208x^2 \\ & - 0.0001338544xy + 4.0210163921y^2 - 4.0022802441x \\ & - 0.0063467738y - 4.0211121695, \\ & y^4 + 0.3576069111xy^2 - 0.2307981167y^3 + 6.6399193424x^2 \\ & - 0.2563280911xy + 25.5142170749y^2 + 0.0074992208x \\ & + 0.3933741054y - 26.5827802561, \\ & xy^3 - 0.7258708041xy^2 + 0.4310045251y^3 \\ & - 13.8060136593x^2 - 0.0759315930xy - 54.8359181240y^2 \\ & - 0.0064826804x - 0.7588300477y + 55.2777294790 \\ & x^2y^2 - 1.4323220755xy^2 + 0.9268708409y^3 \\ & - 27.3974548338x^2 + 1.0278595468xy - 109.4093953618y^2 \\ & - 0.0306093698x - 1.5781783537y + 109.6847894609 \\ & \}\end{aligned}$$

Example.3

$$\tilde{\mathbb{X}} = \{(-8, 0), (-7, 0.484), (-3, -0.927), (2, 0.968), \\ (5, -0.781), (-6, 0.661)\}$$

$$\mathcal{O} = \{1, x, y, y^2, xy, y^3\}$$

$$\mathcal{B} = \{ \\ -0.0349497195y^3 + x^2 + 0.0042887300xy \\ +64.0130240699y^2 - 0.0005572217x + 0.0487914634y \\ -64.0044577737, \\ xy^2 - 6.1875522796y^3 + 0.0391632477xy \\ +3.7538885881y^2 - 0.3805374455x + 4.0809662963y \\ -3.0442995642, \\ x^2y + 64.0240374301y^3 - 0.0012612812xy \\ -0.0009955977y^2 + 0.0021524283x - 64.0108091778y \\ +0.0172194267, \\ y^4 - 0.4441632477y^3 - 0.0153391487xy \\ -0.9638888944y^2 + 0.0148901332x + 0.3190514473y \\ +0.1191210663, \\ xy^3 + 1.2479299148y^3 - 0.4769829902xy \\ -2.0301610985y^2 + 0.1070365604x - 1.2299759484y \\ +0.8562924838 \\ \}$$