特集：データサイエンスと幾何的方法　「積分幾何と統計推測」

栗木 哲 (統計数理研究所)*

1　はじめに

統計学やデータサイエンスにおいて，幾何的手法はいろいろな場面で現れる．例えば微分幾何に基づく統計理論として成功を収めた情報幾何学では，統計モデルや統計推測手順が持つ幾何構造を顕わにすることにより，より進んだ統計手法の研究の枠組みを提供する．

本稿の表題にある積分幾何も，微分幾何の一分野である．ここではその確率バージョンである「確率場の幾何」の統計推測への応用について，著者の研究に関わる範囲で概観したい．扱う対象である確率場 $X(t)$ とは，ベクトル t を添字に持つような確率変数のことであり，多くの確率変数についての推論を同時に行う多重検定・多重比較の応用を念頭においている．

また本稿は，本特集「データサイエンスと幾何的方法」の他の 2 稿 ([16],[15]) とも密接に関係する．このことについては本稿の最後で触れることにする．

2　積分幾何「超」入門

ここでは積分幾何の入門部分を 2 次元図形に即して説明する．教科書，解説書に [8], [14], [12] などがある．

積分幾何では，図形を合同変換しても変わらない特徴量 (不変量) を扱う．ここで合同変換とは平行移動と回転，鏡像の合成である．そのような不変量に図形 S の面積 $\mathrm{Area}(S)$ $(=\varphi_2(S))$ や S の周囲長 $\mathrm{Len}(\partial S)$ $(=2\varphi_1(S))$，ならびに S のオイラー数 (Euler characteristic) $\chi(S)$ $(=\varphi_0(S))$

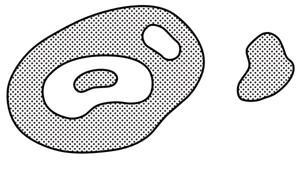

図 1: オイラー数の計算例
(連結成分 3, ホール数 2，オイラー数 $\chi=3-2=1$)

がある．図形 S のオイラー数とは，S の連結成分の個数 $\beta_0(S)$ から S のホール (穴) の個数 $\beta_1(S)$ を引いた量

$$\chi(S)=\beta_0(S)-\beta_1(S)$$

である (図 1)．特に S が空集合のときはオイラー数は 0 となる．

これら φ_i は集合を引数とする関数であるので汎関数である．特に 3 つ組 $(\varphi_2,\varphi_1,\varphi_0)$ をミンコフスキー汎関数 (Minkowski functional) とよぶ．φ_i はそれぞれ加法性とよばれる性質

$$\varphi_i(A)+\varphi_i(B)=\varphi_i(A\cup B)+\varphi_i(A\cap B)$$

を持つ．図形 S を k 倍に拡大した図形を kS と書くとき，$\varphi_i(kS)=k^i\varphi_i(S)$ となり，S の面積，周囲長，オイラー数はそれぞれの 2 次元的汎関数，1 次元的汎関数，0 次元的汎関数である．

逆に，加法性を持ち合同変換不変な汎関数 $\varphi(S)$ は $\varphi_i(S)$ の線形結合

$$\varphi(S)=c_0\varphi_0(S)+c_1\varphi_1(S)+c_2\varphi_2(S)$$

(c_i は S とは無関係) の形でかけることが知られている (Hadwiger の定理)．この定理より，i 次

* 〒 190-8562 東京都立川市緑町 10-3
kuriki@ism.ac.jp

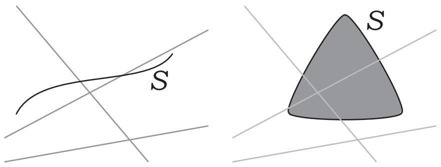

図 2: クロフトンの公式
(グレーのランダム直線 E と曲線 S の交点数の平均をとる (左)，E と図形 S の共通部分の長さの平均をとる (右))

元的汎関数は定数倍をのぞいて φ_i に限定されることが分かる．例えば E を平面上に同じ密度でランダムに配置される直線[1]とし，その直線と平面図形 S との交わり部分の長さを平均すれば S の面積となるという関係式

$$\int \mathrm{Len}(S \cap E)\,\mathrm{d}E = c \cdot \mathrm{Area}(S)$$

(c はある定数) は，上式左辺が S の合同変換不変な加法的汎関数であり，さらに 2 次元的特徴量であることから，Hadwiger の定理より直ちに従う (図 2 右)．また S が曲線の場合である場合，S とランダム直線 E との交点数 (オイラー数と同等) を平均すれば S の長さとなるという関係式

$$\int \#(S \cap E)\,\mathrm{d}E = c \cdot \mathrm{Len}(S)$$

も同様に示すことできる (図 2 左)．これらをクロフトン (Crofton) の公式という．

次に S_1, S_2 を長さが有限の曲線とする．g を平面 $\mathbb{R}^2$ の合同変換とし，S_2 を合同変換 g で移したものを gS_2 と書く．このとき

$$\int \#(S_1 \cap g\,S_2)\,\mathrm{d}g = c \cdot \mathrm{Len}(S_1)\mathrm{Len}(S_2)$$

がなりたつ．ここで積分 $\int \mathrm{d}g$ は，全ての合同変換について等ウエイトで平均をとる操作である．これは上式の左辺が，S_2 を固定したとき S_1 について合同変換不変な加法的汎関数であり，逆に S_1 を固定したとき S_2 についても同様であること，また S_1, S_2 を同時に k 倍すると，k^2 倍となる量であることから証明される．これをポアンカレ (Poincarè) の公式という．

[1] x 軸上に点 x_0 を同じ密度でランダムに選択し，x_0 を通る垂直な直線を原点を中心に角度 θ で回転させて得られる．ここで θ は $[0, 2\pi)$ に値をとる一様分布に従うとする．

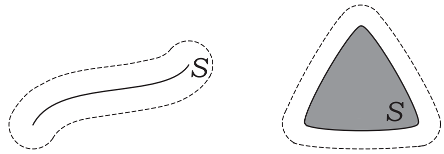

図 3: チューブ $\mathrm{Tube}(S, \rho)$
(点線で囲まれた領域がチューブ)

また別の例として，$B_x(\rho)$ を中心が x，半径が ρ の円盤とする．このとき

$$\varphi(S) = \int_{\mathbb{R}^2} \chi(S \cap B_x(\rho))\mathrm{d}x$$

という汎関数は合同変換不変で加法性を持つ．係数 c_i は S によらないので，積分が簡単に計算できるいくつかの具体的な S を代入し逆算すると

$$\int_{\mathbb{R}^2} \chi(S \cap B_x(\rho))\mathrm{d}x = \varphi_2(S) + 2\rho\varphi_1(S) + \pi\rho^2\varphi_0(S) \tag{1}$$

がわかる．これは積分幾何の基本定理 (Kinematic Fundamental Formula, KFF) の特別な場合である．さらに S が凸集合である場合には，この式の左辺は S から距離 ρ 以下の チューブ領域 $\mathrm{Tube}(S, \rho)$ の面積であり，また S のオイラー数は 1 なので，シュタイナー (Steiner) の公式

$$\mathrm{Area}(\mathrm{Tube}(S, \rho)) = \mathrm{Area}(S) + \rho\mathrm{Len}(\partial S) + \pi\rho^2$$

がなりたつ．

3　確率場の幾何：確率過程・確率場に対する積分幾何

3.1　ガウス型基本定理

このような古典的積分幾何の考え方をベースにして，確率過程や確率場に基づく積分幾何を

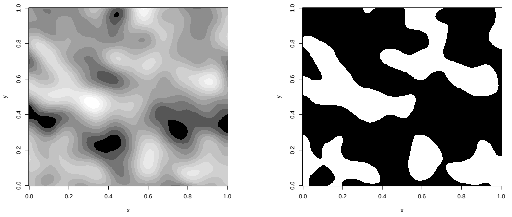

図 4: 確率場 (左) とエクスカーション集合 (右)

創始し，発展させたのは R. Adler, K. Worsley, ならびに J. Taylor である ([1]).

確率過程 (stochastic process) とは添字 t を持った確率変数のことである．特に t がベクトルの場合，そのことを強調するために確率場 (random field) とよぶ．ガウス確率場とは有限個の異なる点の同時分布 ($(X(t_1),\dots,X(t_k))$ の分布) がガウス分布であることをいう．

ガウス確率場 $X(t)$ の一例を図 4 (左) に示す．ここで添字 $t=(t_1,t_2)$ は長方形の添字集合 S の点である．また図 4 (右) は，確率場の値がある閾値以上となるような添字の集合で，エクスカーション集合とよばれる．図 4 (右) では，黒地の連結成分数が $\beta_0=3$，白地のホールの数が $\beta_1=1$ であり，オイラー数は $\chi=\beta_0-\beta_1=2$ である．ここでホールとは，エクスカーション集合に周囲を囲まれた領域である．

閾値 a のエクスカーション集合 S_a は，数式では $S\cap X^{-1}([a,\infty))$ と表せる．ここで $X^{-1}([a,\infty))$ とは，$X(t)$ の値が半区間 $[a,\infty)$ に入るような添字 t の集合である．もし確率場 $X(t)$ が等方的，すなわち座標 t を平行移動したり回転，鏡像変換させたとしてもその統計的性質が変わらないならば，そのオイラー数の期待値は

$$\mathbb{E}[\chi(S\cap X^{-1}([a,\infty)))] = h_0(a)\varphi_2(S) + h_1(a)\varphi_1(S) + h_2(a)\varphi_0(S) \quad (2)$$

ただし $h_i(x) = (2\pi)^{-(i+1)/2}(-\mathrm{d}/\mathrm{d}x)^{i-1}e^{-x^2/2}$ となる．(2) 式と (1) 式は，対応

$$B_x(\rho) \ \leftrightarrow\ X^{-1}([a,\infty)),$$

$$\int_{\mathbb{R}^2}\cdot\,\mathrm{d}x \ \leftrightarrow\ \mathbb{E}[\,\cdot\,], \quad \omega_i\rho^i \ \leftrightarrow\ h_i(a)$$

(ただし $\omega_i=\pi^{i/2}/\Gamma(i/2+1)$) によって同じ形をしていることが分かる．(2) 式は KFF のガウス確率場バージョンであり，ガウス型基本定理 (Gaussian Kinematic Formula, GKF) とよばれる．$X(t)$ がガウスでない場合への拡張は現在でも研究されている ([2]).

3.2 確率場の最大値分布

閾値 a をだんだん大きくすると，エクスカーション集合 $S\cap X^{-1}([a,\infty))$ は単調減少し，最終的には空集合となる.(図 4 の右図では，黒地の部分が減少し，最終的には全部白地となる.) もし a が十分に大きく空集合となる直前とすると，エクスカーション集合は連結成分は 1 つであり，またどこにもホールは存在しないと考えられる．このときエクスカーション集合のオイラー数は 1 となる．またさらに大きい a を考えるとエクスカーション集合が空集合となり，そのオイラー数は 0 である．このような直感的な議論から，閾値 a が大きいとき

$$\chi(S_a)=0 \iff S_a=\emptyset$$
$$\chi(S_a)=1 \iff S_a\neq\emptyset$$

が近似的になりたつと考えられる．$X(t)$ の $t\in S$ での最大値が a 以上であることと，エクスカーション集合 S_a が空集合でないことが同値であるため，

$$\mathbb{P}\Bigl(\max_{t\in S} X(t)\ge a\Bigr) = \mathbb{P}(S_a\neq\emptyset) \approx \mathbb{E}[\chi(S_a)] \quad (a\text{ が大きいとき}) \qquad (3)$$

ただし $S_a=S\cap X^{-1}([a,\infty))$ であった．一般に確率過程や確率場の最大値の分布の導出は難しいが，a が大きい場合，最大値分布の上側裾確率は容易に計算可能な $\mathbb{E}[\chi(S_a)]$ で近似される．この近似法は期待オイラー標数法 (Expected Euler-characteristic method, EEC) とよばれる．

3.3　非等方確率場の場合

いままで確率場は等方的，すなわち添字の合同変換に対して確率的性質が変わらない，という仮定をおいていた．しかしこれらの仮定を外した場合でもエクスカーション集合のオイラー数の期待値 $\mathbb{E}[\chi(S_a)]$ は導出可能である．竹村と著者らは，確率場の平均が 0，分散 1 であるが相関構造は等方的でない場合の $\mathbb{E}[\chi(S_a)]$ を，添字集合 S を球面上の部分集合と考え，球面上チューブの体積評価を通して与えている．これはチューブ法 (tube method) とよばれる ([11, 13])．分散が 1 という仮定を外した場合については [3] を参照のこと．

4　統計的発見への応用

ここまでは数学 (積分幾何と確率論) の話であった．次に統計推測への利用について述べる．

S の各点 t で検定統計量 $X(t)$ が定義される場合，(3) 式の右辺は多重検定としての p 値，すなわち多重性調整 p 値に他ならない．(3) 式は，それが簡単に近似計算できることを意味している．この節では，そのようなデータをいくつか例示する．

4.1　イネの遺伝子ペア検出問題

遺伝学では標準的な考え方として，いくつかの遺伝子が特定の遺伝子型をとる場合に個体が生育できないような組合せが存在すると考えられている．この現象は生殖隔離障壁とよばれ，そのメカニズムが最終的に生物学的「種」を形成するとされる．

図 5 はイネにおいて，そのような生殖隔離障壁を検出するための実験データの要約図である．図の縦軸，横軸は第 1 染色体と第 6 染色体の位置を表す．致死 (不稔) 遺伝子が存在しないならば，生存する個体の遺伝子型の出現頻度は簡単に計算することができる．実際に生息する個体の遺伝子型をカウントし，その期待度数からの

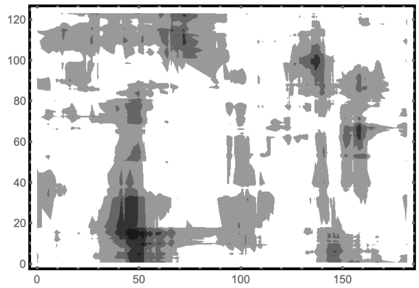

図 5: イネの生殖隔離障壁の探索
(統計量の値を濃淡で表示，横軸：第 1 染色体，縦軸：第 6 染色体, 単位：cM)

乖離を統計量 (具体的には自由度 4 のカイ 2 乗統計量) で尺度化し，色の濃淡で図示したものである．

この多くの統計量を確率場とみなした場合，確率場の相関構造は連鎖構造から確定される．期待オイラー標数法やチューブ法によって確率場の帰無仮説の下での最大値分布を求めることができ，データから観測された生殖隔離障壁の候補の多重性調整 p 値の見積もりを行うことができる ([4])．

検証実験によって，図中に濃く示された最大値は真の生殖隔離障壁であり，一方それ以外の極大点は偽陽性であることが判明している．

4.2　物理実験データからの信号検出

2013 年 3 月 14 日に，最後の素粒子であるヒッグス粒子の発見がアナウンスされた．図 6 はその根拠として CERN ホームページで公表されている実験データである．横軸は素粒子の質量であり，図の上側には素粒子の観測数とその平滑値，下段には共変量の効果 (系統誤差) を取り除いた残差がプロットされている．

図中には確かに一つのピークが存在するように見えるが，それがどのくらいの大きさであれば新発見と見なすことができるか，その合理的な判定基準が必要となる．このように信号の位置が事前に分かっていない問題では，2σ (標準

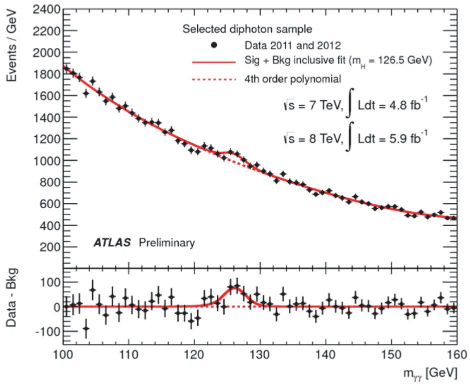

図 6: ヒッグス粒子の探索 (Copyright CERN)

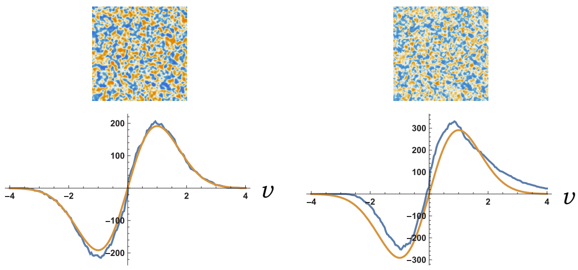

図 7: ガウス確率場 (左) と非ガウス確率場 (右)

誤差 σ の 2 倍) をもって有意と判断する手順では，しばしば見せかけの発見が生じる．この偽陽確率増大の現象は，統計の言葉では検定の多重性に他ならないが，実験物理の分野でも LEE (look-elsewhere effect; どこでも効果) の名前でよく知られている．そのため，実験物理の分野でも多重性調整の目的として (3) 式が使われている ([10])．

4.3 ミンコフスキー汎関数によるガウス性の検定

いままでは，確率場 $X(t)$ のエクスカーション集合の期待オイラー数を用いた最大値分布近似の信号検出への応用について紹介した．ここでは確率場のミンコフスキー汎関数やオイラー数そのものを統計量とする解析を紹介する．

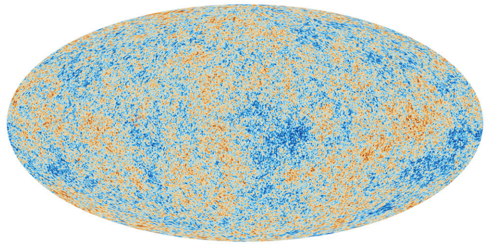

図 8: 宇宙マイクロ波背景放射 (CMB) (Copyright PLANCK))

図 7 上側パネルは，シミュレーションにより生成した 2 次元ガウス確率場 (左) と非ガウス確率場 (右) である．下側パネルは，閾値 v を -4 から 4 までの範囲で振らせたときのエクスカーション集合の標本オイラー数と，そのガウス性の仮定のものでの期待オイラー数をプロットしたものである．見かけ上同じような確率場であっても，オイラー曲線は確率場の非ガウス性を検出することが分かる．

このようなガウス性の検定は，宇宙論研究で用いられている．図 8 の宇宙マイクロ波放射 (CMB) は，宇宙の温度ゆらぎを観測しているもので，等方的ガウス確率場に非常に近いものであることが知られている．それがガウスであるか，そうでないならばどのような非ガウス性を持つか，は初期宇宙モデルで決定されるため，研究の重要なトピックである．その目的のために図 7 のオイラー曲線 (ジーナス統計量ともよばれる) やミンコフスキー汎関数が解析されている．

しかしながら，ガウス性の検定としてオイラー曲線を利用することの得失はよく分かっていない．ガウス確率場は共分散関数 (2 点相関関数) だけで決定される．松原 [6] は，非ガウスの下でのオイラー曲線の挙動の解析のために，ガウス確率場ならば 0 となる 3 次相関関数の存在を仮定したときの期待オイラー数を摂動展開の形で導出し，初期宇宙モデルとの対応を調べている．さらに松原と著者 [5, 7] は，次元を一般化し高次の相関関数の存在を仮定した摂動展開を導出

している．

5 おわりに

本稿では積分幾何と確率場の幾何の考え方を概観し，信号検出や多重検定への適用について紹介した．本稿では触れることができなかったが，期待オイラー標数法やチューブ法は，同時信頼区間構成や特異モデルの尤度比検定にも適用される．これらについては，[13] を参照のこと．

また冒頭に述べたように，本稿は本特集の別論文 と関連している．本武 [16] が解説する位相的データ解析 (TDA) は，エクスカーション集合のオイラー数を構成するベッチ数 β_i を扱うが，TDA では閾値を変化させたときに β_i を増減させる生成元を解析対象とするため，オイラー曲線より数段細かな情報を扱う．そのため最近では，伝統的にミンコフスキー汎関数やオイラー曲線を用いてきた分野で，TDA を試みたという研究が多く報告されている．二宮 [15] の解説する選択的推論 (Selective inference) の分野の創始者は確率場の幾何の研究者 J. Taylor であり，選択的推論の最初の論文 [9] は，期待オイラー標数法で用いるカッツ・ライス公式 (モースの定理の積分型) を出発点とする論文である．そのため数理技法に多くの重なりがある．

有益なコメントを賜りました會田雅人氏に感謝いたします．

参考文献

[1] Robert J. Adler and Jonathan E. Taylor. *Random fields and geometry.* Springer, New York, 2007.

[2] Robert J. Adler and Jonathan E. Taylor. *Topological complexity of smooth random functions*, Vol. 2019 of *Lecture Notes in Mathematics.* Springer, Heidelberg, 2011.

[3] Satoshi Kuriki, Takemura Akimichi, and Jonathan E. Taylor. The volume-of-tube method for gaussian random fields with inhomogeneous variance. arXiv:2108.02118 [math.PR], 2021.

[4] Satoshi Kuriki, Yoshiaki Harushima, Hironori Fujisawa, and Nori Kurata. Approximate tail probabilities of the maximum of a chi-square field on multi-dimensional lattice points and their applications to detection of loci interactions. *Annals of the Institute of Statistical Mathematics*, Vol. 66, No. 4, pp. 725–757, 2014.

[5] Satoshi Kuriki and Takahiko Matsubara. Perturbation of the expected minkowski functional for weakly non-gaussian isotropic fields on a bounded domain. arXiv:2011.04953 [math.ST], 2020.

[6] Takahiko Matsubara. Statistics of smoothed cosmic fields in perturbation theory. I. Formulation and useful formulae in second-order perturbation theory. *The Astrophysical Journal*, Vol. 584, pp. 1–33, 2003.

[7] Takahiko Matsubara and Satoshi Kuriki. Weakly non-Gaussian formula for the Minkowski functionals in general dimensions. arXiv:2011.04954 [astro-ph.CO], 2020.

[8] Rolf Schneider and Wolfgang Weil. *Stochastic and integral geometry.* Springer, Berlin, 2008.

[9] Jonathan E. Taylor, Joshua R. Loftus, and Ryan J. Tibshirani. Inference in adaptive regression via the Kac-Rice formula. *The Annals of Statistics*, Vol. 44, No. 2, pp. 743–770, 2016.

[10] David A. van Dyk. The role of statistics in the discovery of a higgs boson. *Annual Review of Statistics and Its Application*, Vol. 1, pp. 41–59, 2014.

[11] 栗木哲, 竹村彰通. チューブの体積と正規確率場の最大値の分布. 数学, Vol. 60, No. 2, pp. 134–155, 2008.

[12] 腰塚武志. 応用のための積分幾何学—図形の測度：道路網・市街地・施設配置. 近代科学社, 2019.

[13] 栗木哲. チューブ法の理論・応用とその周辺. 統計数理, Vol. 67, No. 2, pp. 229–240, 2019.

[14] 田崎博之. 積分幾何学入門. 培風館, 2016.

[15] 二宮嘉行. 選択的推論：データサイエンスにおける quiet scandal の克服. エストレーラ, 2021 年 10 月.

[16] 本武陽一. 位相的データ解析法によるパターンダイナミクス分析のすすめ. エストレーラ, 2021 年 10 月.