大規模構造クラスタリング統計量の予言 ―機械学習的アプローチ

西 道 啓 博

〈東京大学　国際高等研究所　カブリ数物連携宇宙研究機構　〒277-8583 千葉県柏市柏の葉 5-1-5〉

e-mail: takahiro.nishimichi@ipmu.jp

観測データを説明する数理モデルや内包されるパラメータのベイズ推定には，適切な理論テンプレートが必要である．宇宙大規模構造の理論予言には，N体シミュレーションが威力を発揮するが，多次元パラメータ空間を埋め尽くすように多数のシミュレーションを用意し，そのままテンプレートとするのは現実的ではない．本稿では，すばる望遠鏡Hyper Suprime-Camによる銀河・銀河レンズ効果の解析を念頭に構築した，「ダークエミュレータ」を題材に，ベイズの枠組みでパラメータ推定や予測を行う方法論について紹介する．特に，その基礎となる多次元パラメータ空間の効率的なサンプリングや，ガウス過程を用いた回帰について詳しく解説する．

1. はじめに

宇宙論は今，データが物を言う時代にある．およそ100年前，アインシュタインが一般相対性理論を発表してから現在に至るまで，宇宙膨張，ブラックホール，さらには重力波の存在が理論的に予言され，それらの言わば「宿題」は後になって観測的に実証されていった．これとは全く逆に，ダークマターやダークエネルギーについては，これらを第一原理的に予言，記述する基礎的な理論は存在せず，観測先行の状況が続いている．数々の大型観測ミッションを引っ提げ，世界は今データの力にすがって宇宙のダーク成分の正体を絞り込もうとしている．宇宙論におけるデータ科学化の波はここ日本にも押し寄せており，Subaru Measurement of Images and Redshifts（SuMIRe）プロジェクト[1]がまさに今進行中である．観測の大型化に伴い，得られたデータをくまなく適切に解釈する理論的・統計的枠組みの整備がますます重要となっている．

本稿では「データ科学」シリーズの特集記事として，宇宙論の分野でベイズ推定がどのように使われているか，SuMIReの前半部分を担うすばる望遠鏡Hyper Suprime-Cam（HSC）による測光サーベイから測定した，銀河・銀河重力レンズ効果の解析に向けた理論研究を題材として紹介する．本稿は以下のように構成されている．まず，第2章ではベイズ推定の基本から実際の運用までを簡単に紹介する．併せて，宇宙論的な揺らぎを用いた統計解析のあらましと困難について述べる．次に，第3章では「エミュレータ」の考え方を導入し，これに必要な統計的手法について順次議論する．第4章はわれわれが構築したダークエミュレータについて紹介する．最後に第5章は今後の宇宙論統計解析に向けた展望について述べる．ここで触れる内容は宇宙論に限らず，一般的に観測データからモデルパラメータを推定する問題に通じている．分野を問わず，読者の皆様の研究の一助となれば幸いである．

2. ベイズ推定

手始めに，ベイズ推定の基礎についてまとめ，

宇宙論的揺らぎの統計解析における課題について述べることで，後半の議論の導入としたい．

2.1 ベイズの定理

われわれが計算したいのは「与えられたデータの元で，どのモデルがどの程度確からしいか」という確率的指標である．これは端的に言えばベイズの定理に集約される：

$$P(M|D)=\frac{P(D|M)P(M)}{P(D)}. \qquad (1)$$

ここで，Dは観測データを，Mは何らかの数理モデルを意味し，モデルがパラメータを含む場合には，その自由度も含めてMと書くことにする．左辺を事後確率と呼び，右辺は尤度$P(D|M)$，事前確率$P(M)$と証拠$P(D)$により書かれている．今，モデルMに関する確率にしか興味がないので，$P(D)$は単なる規格化因子としての意味しか持たない．ベイズ的アプローチにおいては，事前確率$P(M)$にわれわれがモデルMについてもつ主観が入ってしまうことがしばしば批判の対象となるが，本稿の議論の本筋からは外れるので深入りはしない．

尤度$P(D|M)$は，物理モデルに基づいて観測量を予言するステップに対応し，通常はここに最も多くの計算時間がかかる．大雑把に言えば，ベイズ推定とは，検証したい数々のモデルとそこに含まれるパラメータの組に対して$P(D|M)$をひたすら計算し，これに事前確率$P(M)$をかけて得られる式(1)の左辺を最大化するMを見つける作業である．ここで問題となるのは，この最適化問題を解く際に，モデルが多数のパラメータを有すると，パラメータの数に対して計算コストが指数関数的に増大してしまうという，一種の次元の呪いである．多次元パラメータ空間から現実的な計算時間で答えを見つけ出すために，通常，以下で述べるような効率的なサンプリングアルゴリズムが利用される．

2.2 マルコフ連鎖モンテカルロ法

さて，サンプリング法の代表格と言えば，マルコフ連鎖モンテカルロ法（Markov-Chain Monte Carlo; MCMC）であろう．これは，何らかのモデルM_0からスタートしてマルコフ的[*1]に順次モデル$M_i(i=1,\cdots,N)$を発生させ，モデルの列$\{M_i\}$が最終的に知りたい事後確率$P(M|D)$に従うよう条件を課した一連のサンプリング法である．得られたモデルの列$\{M_i\}$は，サンプル数Nを大きく取れば，事後確率を十分良く代表する標本とみなせ，そこからパラメータの最尤推定値や信頼区間を見積もることができる．

最も単純なMCMCの実装である，メトロポリス法では，現在の状態M_iから次の状態M_{i+1}の候補を提案するステップと，提案を採用，または棄却するステップとに分かれる．最初のステップでは提案確率を対称（$M\to M'$と$M'\to M$が等しい）に取る．後半のステップでは採択率$\min[1, P(M_{i+1}|D)/P(M_i|D)]$にて採否を決める．棄却された場合には，元のモデル$M_i$に戻り$M_{i+1}=M_i$とする．実際，この方法で生成した連鎖は事後確率$P(M|D)$からの標本となっている．この方法は，データに対するモデルの当てはまりの良し悪しに応じた確率で遷移が行われ，比較的直感に即した手法となっている．

この例でもわかるように，多くのMCMC法においては入力次元数によらず，比較的データとの当てはまりの良いモデルを重点的に調査することで，次元の呪いの問題をうまく回避している．具体的にどのようにMCMCを実装するのが最も効率が良いかは事後確率分布の形状によるので一概には言えないが，メトロポリス法以外にもさまざまな方法が運用されている[*2]．

[*1] 次に遷移するモデルM_{i+1}が現在のモデルM_iのみに依存し，過去の履歴によらないと言う意味．

[*2] 多峰性の分布に強い入れ子サンプリング[2]や，入力次元同士が複雑に相関している場合に有効なアフィン変換不変サンプリング[3]などがよく使われている．

2.3 尤度計算によらない近似的手法

MCMCによるサンプリングでは，繰り返し尤度を計算することが必要となる．最近では，この計算が解析的にはできず，コストの高い数値計算を要する場合を念頭に，近似的にベイズ推定を行う方法論が議論されるようになってきた（近似ベイズ計算：approximate Bayesian computation; ABC）．この方法では，まず，事前確率$P(M)$に従ってMを選択し，順モデル$\hat{D}(M)$を発生させる．ここで，順モデルとは観測データと全く同等な条件で生成された擬似的なシミュレーションデータのことを呼ぶ．そして，擬似データ$\hat{D}$と観測データDの間に何らかの距離指標$\rho(\hat{D}, D)$を導入する．ρは$\hat{D}$とDが完全に一致するときのみゼロになる連続的な正値の関数であれば何でも良く，簡単には通常のユークリッド距離で良い．問題によっては，データベクトル空間に適切な計量を導入するなどして，次元ごとに重みを異にする距離を定義することで，サンプリングを効率化できる．次に，閾値ϵを用意し，$\rho(\hat{D}, D)<\varepsilon$を満たせば順モデル$\hat{D}$（あるいは，それを発生させる元となったモデル$M$）を採用，そうでなければ棄却する．この一連の操作を何度も繰り返し，観測データとεの範囲で近いモデルMの集合を得る．面白いことに，$\varepsilon\to 0$の極限では，この集合は事後確率$P(M|D)$に従うことが知られている．

モデルを特徴付けるパラメータ空間が連続的であれば，閾値εを0に取ると発生させたすべてのサンプルが棄却されてしまう．そこで，実用上は，事後確率分布の評価に十分な大きさのサンプルを残すことができ，なおかつできるだけ小さな正の値をεとして選ぶことで，近似的な事後確率$P_\varepsilon(M|D)$を得る．一般に，データのもつ次元が高い問題では，提案された順モデル$\hat{D}$が採用される確率が低くなるため，適切な要約統計量$S(D)$を導入してデータの次元を削減することで，より効率的に近似的な事後確率分布をサンプリングできる．いずれにせよ，近似的な分布$P_\varepsilon(M|D)$と真の分布$P(M|D)$とがどの程度ずれるか，別途調査しておくことが必要となる．

ABCでは，尤度$P(D|M)$を計算する代わりに，擬似データ$\hat{D}(M)$をバイパスするところが大きな特徴となっている．当然，それを実現するためには$\hat{D}$を（できれば大量に）生成可能である必要がある．尤度を経由する従来の推定では，検証したい1つ1つのモデルMに対して多数の順モデル$\hat{D}(M)$を発生させることが，尤度$P(D|M)$を評価するステップに当たる．ABCでは1つのモデルについて順モデルを多数回発生することはせず，より多くのモデルを1度ずつざっと調べることで，より素早く，直接的に事後確率$P(M|D)$をサンプリングしているのである．

2.4 宇宙論的揺らぎの統計解析

標準的な宇宙の構造形成シナリオでは，インフレーション期に微小な揺らぎが生成され，これが成長して現在の宇宙のあらゆる構造を形作ったと考えられている．宇宙の原始揺らぎは非常にいい近似でガウス確率場に従うとされる．空間座標$\boldsymbol{x}$における揺らぎの値をyなどと書き[*3]，期待値を$\langle\cdots\rangle$と書くと，ガウス場とはすなわち，2点相関関数

$$C(\boldsymbol{x}, \boldsymbol{x}') = \langle yy' \rangle, \tag{2}$$

だけを持ち，高次のキュムラント[*4]がゼロであるようなものを呼ぶ．このとき，任意の自然数N個の空間座標$\boldsymbol{X}_N = \{\boldsymbol{x}^{(i)} | i=1, \cdots, N\}$における揺らぎ$\boldsymbol{Y}_N = \{y^{(i)} | i=1, \cdots, N\}$の従う同時確率分布は正規分布

[*3] 揺らぎと呼んでいるからには，$\langle y\rangle$は恒等的にゼロ．

[*4] n次のキュムラントとは，n次のモーメント$\langle y^{(1)}\cdots y^{(n)}\rangle$から低次のモーメントの積で表される冗長な成分を差し引いたもの．

$$P(\boldsymbol{Y}_N) \propto \exp\left[-\frac{1}{2}\boldsymbol{Y}_N^T \boldsymbol{C}_N^{-1}\boldsymbol{Y}_N\right], \qquad (3)$$

に従う．ここで，$\boldsymbol{C}_N$は式(2)の$\boldsymbol{x}$および$\boldsymbol{x}'$に$\boldsymbol{X}_N$の各座標を代入して得られる共分散行列である．さらに，宇宙の一様等方性を仮定すると，式(2)は2点間の距離$r=|\boldsymbol{x}-\boldsymbol{x}'|$だけに依存する．このような確率場は，2点相関関数$C(r)$，またはそのフーリエ変換であるパワースペクトル$P(k)$（$k$は波数）という1変数関数で完全に特徴づけられる．

宇宙マイクロ波背景放射（cosmic microwave background; CMB）の観測は，揺らぎがまだ微小な時期を見ており，各フーリエモードは独立に成長しガウス性を保っている．この成長率の計算には線形近似[*5]で十分であり，高速に実行できる．観測データからパワースペクトルの不偏推定量を作ると，与えられた宇宙論モデルの元でこれが真の値の周りにどのように分布するか解析的に評価でき，尤度が得られる．宇宙論パラメータ（標準的なモデルでは6つ）に加えてさまざまな不定性を説明する数十個の追加のパラメータを設けても，MCMCにより長さ百万程度の連鎖を現実的な時間で発生させ，宇宙論パラメータをベイズ推定することができる．

一方で，宇宙大規模構造から見た宇宙の揺らぎの統計性は非線形効果を受けて極めて非自明なものとなる．それでも，大きなスケールに目を向けると揺らぎはガウス統計に近いと期待できるので，引き続き2点相関関数が議論の主役となる．よって，CMBと同様な手続きを踏むが，通常，2点相関関数の不偏推定値が従う分布を正規分布で近似し[*6]，期待値と共分散行列とで特徴づける．そして，それぞれについて，非線形効果を適切に取り入れたモデルを構築することになる．このうち，共分散行列は解析的な評価が難しいため，数値シミュレーションにより多数の順モデルを作って推定する．通常，これを精度良く決定するには数千ものシミュレーションを要する[4]．この数字は1つのモデルに対するものであるが，共分散行列自体，宇宙論パラメータに依存すると考えるのが自然で，この依存性を適切に理解するためには，さらにずっと多くのシミュレーションが必要となる．これはまさにABCの節で述べた，尤度計算が困難な状況と一致する．実際のところ，共分散行列の推定には尤もらしい1つのモデルで代表して宇宙論パラメータ依存性は無視し，しかも，高速に発生することができる近似的な順モデルを使うことが多い．

さて，共分散行列の話題は他に譲るとして，本稿では統計量の期待値を非線形効果を取り入れて高速に予言する点に集中したい．これには，1モデルにつき数千ものシミュレーションは不要である．かつて大規模構造シミュレーションの金字塔だったMillennium Simulation[5]と同等の大規模

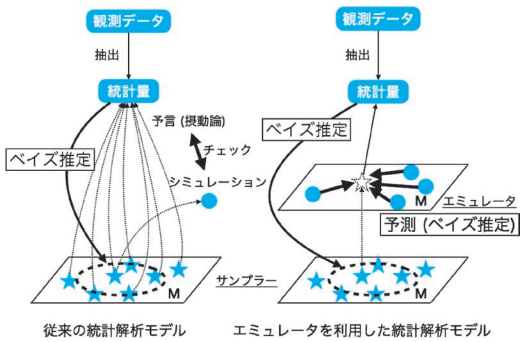

図1　従来の統計解析（左）とエミュレータによる解析（右）との比較．ラベルMで示した面は多次元宇宙論パラメータ空間を表し，その上の破線の楕円は推定されたパラメータの信頼区間を表す．星印はパラメータ推定のためのサンプル，丸印はシミュレーションを実行したサンプル．エミュレータを用いた解析は，理論予測とパラメータ推定の2カ所でベイズ推定を利用している．

[*5] 揺らぎの2乗以上の高次の項を無視した近似．

[*6] 統計量を推定する際，通常，多数の確率変数に対する平均を取るため，元となるデータ（揺らぎの場そのもの）が著しく非正規分布であっても中心極限定理から正規分布に近くなる．

な計算が，近代的なスーパーコンピュータではたった数日で可能となった．ギガパーセクにも及ぶ大きな体積をカバーしつつ，銀河スケールの構造（暗黒物質ハロー，以下単にハロー）まで解像するN体計算を多数実行することができる．それでも，このスピードではMCMCやABCのようなスキームに直接乗せることはまだ叶わない．そこで，実際に観測データから宇宙論パラメータの推定を行う際には，高次の摂動論などを用いて理論予言を与えることが多い．その際，摂動展開がどこでどのように破れ，それにより生じた誤差が最終的なパラメータ推定にどう伝播するか理解するために，言うなればシミュレーションデータは補助的に利用されてきた（図1左参照）．以下で解説する「エミュレータ」は，このような状況を打破し，シミュレーションデータをより積極的に使って，観測データとの直接比較に基づくベイズ推定を目指すものとして位置づけられる．

3. エミュレータ

宇宙論パラメータを変化させながら多数回シミュレーションを実行した後，新たな宇宙論パラメータについて，追加のシミュレーションをすることなく，手持ちのデータから予測を行うのがエミュレータである．この手法は，Katrin Heitmann, Salman Habibらにより宇宙論業界に持ち込まれ[6]，物質の密度揺らぎのパワースペクトルの予言に応用された（cosmic emulators[7]）．この方法では，ベイズ推定の力を積極的に利用することで，理論予測からパラメータ推定までを一気に結ぶ．図1の右に示す通り，MCMCなどでサンプリングされたパラメータ空間の各点各点で新たなシミュレーションを実行することなく，既に完了しているシミュレーションデータに基づくベイズ推定を行い，これを理論予測として利用する．

エミュレータを構築する上で重要となるのは，少数のサンプル点で多次元パラメータ空間を効率良く調査することと，得られたシミュレーションデータから適切に予測を行うことである．以下，これらについて順次解説する．

3.1 ラテン超方格サンプリング

n次元実数入力空間からN点をサンプルした集合，$\boldsymbol{X}_N=\{\boldsymbol{x}^{(i)}|i=1,\cdots,N,\boldsymbol{x}^{(i)}\in \mathrm{R}^n\}$，を考える．ラテン超方格法（Latin Hypercube Design; LHD[8]）は多次元入力空間を効率的にカバーする実験計画法としてよく知られている．まず，各入力次元について，探索する最小値，最大値を設定し，そうして決められた超直方体領域をN^n個の等間隔の格子に区切る．N^n格子すべてについてシミュレーションすることは，nが大きくなると途端に難しくなるが，そうする代わりに格子の各行各列から必ず1度ずつ，計N回選ぶのがラテン超方格である（図2）．この方法は，次元数nとは無関係に計算時間の都合などから決めた，任意の実験数Nについて実現可能で，運用上大変都合が良い．定義から，どの入力次元についても小さな値から大きな値まで満遍なくサンプルしており，出力$y(\boldsymbol{x})$の各入力次元に対する依存性を効率良く捉えることができる．

実はそのようなデザインは複数存在する．よって，追加の条件を課してより「効率的な」ものを見つける必要がある．図2の左のパネルを見ると，対角線上に並べた単純なデザインもラテン方格の条件を満たしているが，これは空間充填度という観点からうまくサンプリングできているとは言い難い．よく用いられる基準はマクシミン距離設計と呼ばれるもので，最も近いサンプル間の距離を最大にしなさいという条件である．実装の都合上，空間充填度の指標として

$$\phi(\boldsymbol{X}_N)=\left(\frac{2}{N(N-1)}\sum_{i\neq j}\frac{1}{d^r(\boldsymbol{x}^{(i)},\boldsymbol{x}^{(j)})}\right)^{1/r},\quad(4)$$

を定め，これを最小化するデザイン$\boldsymbol{X}_N$を見つける．これは，$r\to\infty$の極限でマクシミン距離設計に帰着する．ここで，2点間の距離dとしては，

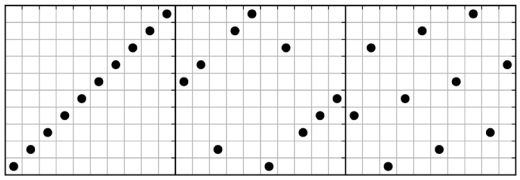

図2　10サンプル点からなる2次元LHDの例．左から，対角線上のデザイン，ランダムに生成したLHD，式(4)を最小化して得られたLHD（r=15を採用）．右のデザインは良い空間充填設計となっている．

通常ユークリッド距離を用いるが，より一般的な距離を用いても良い．図2の中央は無作為に生成したLHDデザインで，サンプリングの密度にムラがある．式(4)に基づいて最適化を行ったものが図2右のデザインである．

機械学習を念頭に置くと，訓練データ，検定データなど，異なる用途に対応できるようデザインしておくと都合が良い．そこで，複数の層をもつマクシミン距離設計LHD[9]を紹介しておく．すなわち，全Nサンプル点をm個の層に分けておき，各層の内部を見ても，全Nサンプル全体を見ても最適になっているようなサンプリング法である．式(4)と同様の充填度の指標を，サンプル全体（$\phi_{\rm all}$）および各層（ϕ_t）について定義し，これらを重み付きで合計した

$$\Phi(\boldsymbol{X}_N)=\frac{1}{2}\left(\phi_{\rm all}+\frac{1}{m}\sum_{t=1}^{m}\phi_t\right), \qquad (5)$$

を最小化することで，そのようなデザインが実現できる（図3）．

3.2　ガウス過程による回帰

ラテン超方格法などにより効率的デザイン$\boldsymbol{X}_N$を発生させ，デザイン上の各点$\boldsymbol{x}^{(i)}$について実験を行い，出力$\boldsymbol{Y}_N=\{y_i|i=1,\cdots,N, y^{(i)}\in\mathbb{R}\}$を得たとしよう．ここでは入力は多次元だが出力は1次元を考えている．このデータから未知の関数$y=f(\boldsymbol{x})$を推定するにはどうしたら良いだろうか？やはりベイズ推定の枠組みから考えてみよう．本来，関数$f(\boldsymbol{x})$は入力$\boldsymbol{x}$に対して決定論的に決まるべきものであるが，確率密度汎関数$P[f(\boldsymbol{x})]$を導入することで，確率的な解釈を持ち込む．そして，ベイズの定理（1）に従って，手持ちのデータ$\boldsymbol{Y}_N$のもつ情報を，$f(\boldsymbol{x})$へと伝播させ，事後確率$P[f(\boldsymbol{x})|\boldsymbol{Y}_N]$を計算するのがここでのゴールとなる．

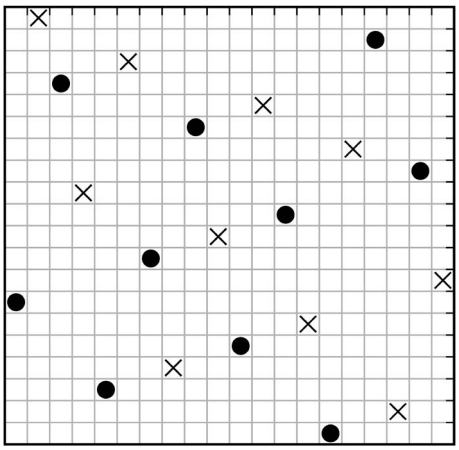

図3　20サンプル点を2層に分けたマクシミン距離設計LHD．式(5)を最小化して得た（r=15を採用）．同じ記号同士で見たときにも，すべてのサンプルで見たときにも偏りのないサンプリングとなっている．

ガウス過程[10]（Gaussian Process; GP）では，$f(\boldsymbol{x})$の関数形を指定することをせず（ノンパラメトリック），$P[f(\boldsymbol{x})]$に対して事前情報としてガウス確率場に従うという条件を課す．すなわち，宇宙論的原始揺らぎの項で見たとおり，複数の入力地点における出力の同時確率分布が式(3)で書かれる確率場である*7．よって，その統計性は2点相関関数（GPの枠組みの中ではカーネルと呼ぶ）が決め，これが回帰曲線の特徴をコントロールする．宇宙の揺らぎの議論で一様等方性を課したのと同様，GPによる回帰でも2点間の距

*7　一般には，$y^{(i)}$の平均はゼロでなくても良い．

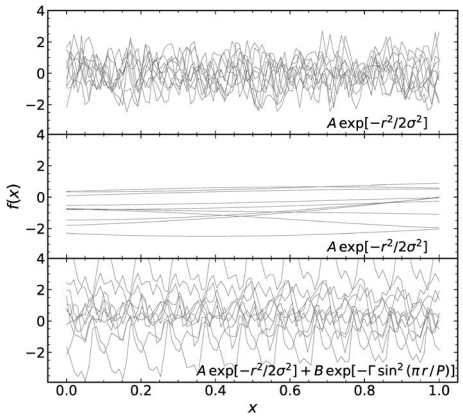

図4　さまざまなカーネル（各パネル内の数式）をもつガウス過程から発生させた10個の無作為なサンプル．上，中央はそれぞれガウシアンの幅をσ=0.01, 1と取った．下は中央パネルのカーネルにP=0.1の周期をもつ振動関数を加えた（パラメータΓ=4）．いずれも規格化A, Bは1とした．

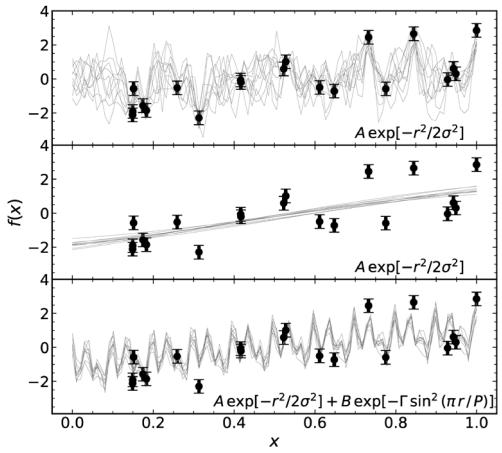

図5　図4の3つのカーネルについて，データ点（誤差棒）を得たときの事後分布から発生させた10のランダムサンプル．

離のみに依存するカーネルがしばしば用いられる．具体的なカーネルの形として，よく用いられるのはガウシアンなどの単調な減少関数である．その場合，ガウシアンの幅は関数の硬さを決める（図4上と中央を参照）．また，$f(\boldsymbol{x})$に何らかの周期性（日周変化，年周変化）が期待される場合には，適切な周期で変化する三角関数などを入れるのが良い（図4下）．得られた標本は，事前情報（カーブの硬さ，周期性）を反映して，異なる振る舞いを示すことがわかる．

次に，与えられた事前確率にデータ$\boldsymbol{Y}_N$の情報を加えることで，新しい入力値$\boldsymbol{x}^{(N+1)}$における出力$y^{(N+1)}$の事後確率を計算する．今，ベイズの定理（1）の分子に現れる尤度と事前確率との積は，データ$\boldsymbol{Y}_N$と予言値$y^{(N+1)}$の同時確率分布となっていることに注意しよう．GPでは，まさにこれが多変量正規分布だとしているので，事後確率$P(y^{(N+1)}|\boldsymbol{Y}_N) \propto P(\boldsymbol{Y}_N, y^{(N+1)})$は式(2)の$N$変数についての確率分布を，データ$\boldsymbol{Y}_N$に予言値$y^{(N+1)}$を合わせた$N$+1変数へと拡張したものになる．$N$+1点の共分散行列$\boldsymbol{C}_{N+1}$を

$$\boldsymbol{C}_{N+1} = \begin{pmatrix} C_N^{(11)} & \cdots & C_N^{(1N)} & k^{(1)} \\ \vdots & \ddots & \vdots & \vdots \\ C_N^{(N1)} & \cdots & C_N^{(NN)} & k^{(N)} \\ k^{(1)} & \cdots & k^{(N)} & \kappa \end{pmatrix} \tag{6}$$

と部分行列に分解して書くと，$\kappa = C(\boldsymbol{x}^{(N+1)}, \boldsymbol{x}^{(N+1)})$は予測値$y^{(N+1)}$に課された事前確率分布の分散，$N$要素のベクトル$\boldsymbol{k} = \{C(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(N+1)}) | i = 1, \cdots, N\}$は予測値と手持ちのデータ$\boldsymbol{Y}_N$の共分散であり，すでに知っている情報が新たな予測にどう伝播するのかを決定する．これらを用いると，データ$\boldsymbol{Y}_N$の元での$y^{(N+1)}$の期待値および分散は

$$\langle y^{(N+1)} | \boldsymbol{Y}_N \rangle = \boldsymbol{k}^T \boldsymbol{C}_N{}^{-1} \boldsymbol{Y}_N, \tag{7}$$

$$\langle (\Delta y^{(N+1)})^2 | \boldsymbol{Y}_N \rangle = \kappa - \boldsymbol{k}^T \boldsymbol{C}_N{}^{-1} \boldsymbol{k}, \tag{8}$$

と計算できる．新たな入力$\boldsymbol{x}^{(N+1)}$は任意の地点に取れるので，これを動かすことで式(7)から回帰曲線が，さらに式(8)からその不定性が得られる．この際，データ$\boldsymbol{Y}_N$が誤差を含んでいるのであれば，その寄与を共分散行列$\boldsymbol{C}_N$に足せば良い．

図5は図4の各パネルの事前確率に対して，誤差棒で表されるデータを与えたときに得られる事

後確率から無作為に取ったサンプルである．いずれの場合も，データ点が拘束条件として働くことで，図4よりも狭い範囲に曲線が集まっている．上パネルでは相関長が短すぎてデータ点から離れるとすぐに予言力を失い，中央では逆に相関長が長すぎてデータ点を通るカーブが描けない．周期性をもつカーネルを用いた下パネルは，周期が合わないためにうまくデータ点を説明できない．

このように，GP回帰の性能はカーネルに大きく左右される．実は，カーネルの関数形さえ仮定すれば，これを特徴づけるパラメータ（超パラメータと呼んで，関数$f(\boldsymbol{x})$を特徴づけるパラメータとは区別する）に対してもう一度ベイズ推定を行い，最適化することができる．ここまでくるといささか冗長なので，再び詳しく説明することは避けるが，データ$\boldsymbol{Y}_N$の元での超パラメータの事後確率を計算すれば良い．超パラメータの不定性込みで$y^{(N+1)}$の事後分布を調べることも数値的には可能だが，多くの場合，単に超パラメータの最尤推定値を用いれば十分である．図6はそのようにして得られた最尤超パラメータの元でのGPの予言（実線）および不定性（影）を示す．データを生成する元となったカーブ（正弦関数と多項式を適当に組み合わせたもの）も合わせて破線で示している．いずれの場合も，関数の特徴を良く捉えている．周期性があるデータなので，カーネルに周期成分を入れておいた下の場合の方が性能が出る．

超パラメータ調整は，ニューラルネットワークなどでも登場する問題であり，これによりGPはデータの複雑さをデータ自体から学び取る一種の機械学習として機能する．実際，一つの隠れ層をもつネットワークは，重み関数に適切な事前確率を課して，ノードの数を大きくすると，GPに漸近することが知られている[11]（図7参照）．GPによる回帰は，ベイズ的なアプローチを取り，予測に不確かさが入る余地を残すことで，過学習を起こさずに済む利点がある．そうは言っても，カーネルの選び方や事前確率としてGPを与えることの妥当性は所詮経験則でしかないので，交差検定などを行って予言精度を検証することが望ましい．実際，図6に影で示されたGPの1σ信頼区間は必ずしも真の値（破線）を含んでいない．GPは，式(7)，(8)という極めて単純な行列演算に集約されるため，高速な予測を実現する．ただし，行列$\boldsymbol{C}_N$の反転（計算コストN^3）が必要となるため，データベクトルの大きさNがあまりに大きくなると取り扱いにくくなってしまうことを注意しておく．

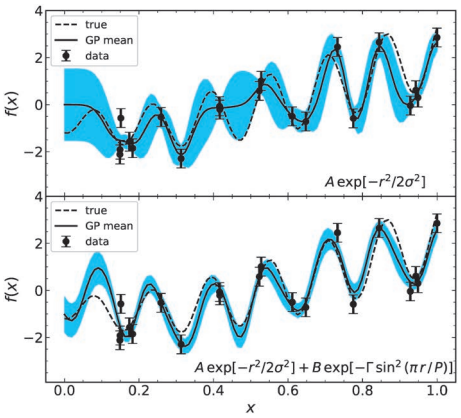

図6　2つのカーネルに対する，超パラメータ学習後のGPの予言と，真のモデルカーブとの比較．

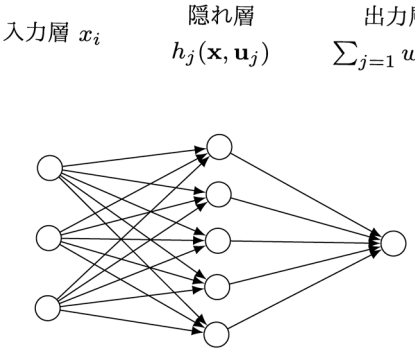

図7　単一の隠れ層をもつニューラルネットワークの一例．入力3次元，隠れ層5ノード，1出力の場合を示す．隠れ層の出力h_jが同一の分布に従い，重みw_jが同一の分布に独立に従うとき，隠れ層のノード数無限大の極限を取ると，中心極限定理からガウス過程に漸近する．

4. ダークエミュレータ

ここまでで，エミュレータ構築の肝となるサンプリングと機械学習について述べた．HSCの銀河・銀河レンズ効果の解析に向けて開発中の「ダークエミュレータ」もこれらの方法論を踏襲したものである．以下は，そのあらましと現在までに達成した性能について紹介する．

4.1 HSC銀河・銀河レンズの狙い

銀河クラスタリングを使った宇宙論では，物質全体（バリオン＋冷たい暗黒物質）が作る密度場と，それを離散的にトレースする銀河サンプルとの関係に大きな不定性があり，実際，色や明るさなどの性質が異なる銀河を選ぶと，クラスタリングの統計的振る舞いも変わってくる（銀河バイアス）．重力レンズ効果には，光る・光らないにかかわらずあらゆる重力源が寄与するので，この問題を避けることができる一方，揺らぎを視線に沿って射影した2次元的な情報しか得られない．

銀河クラスタリングと重力レンズ効果の中間的なものとして銀河・銀河レンズ効果[12]がある．これは，前景の赤方偏移既知の銀河サンプルの周りに，背景銀河の像をスタックして得られる弱重力レンズ効果を指す．個々の銀河が受ける重力レンズ効果は光の経路上の密度場を重み付きで積分したものだが，スタックすることで前景の銀河サンプルに付随する質量が引き起こした成分のみを引き出すことができる．これを利用すると，前景の銀河サンプルに付随するハローの典型的な質量や密度プロファイルを推定でき，銀河バイアスの不定性を大きく低減することができる．HSCが観測した遠方銀河をソース天体，スローンディジタルスカイサーベイ[13]が取得した分光銀河サンプルをレンズ天体として利用することで，バイアスの不定性に強い宇宙論解析を実現しようとしている．

4.2 ダークエミュレータ

このようなサイエンスを可能にするため，ダークエミュレータは，銀河から銀河団スケール（質量$\gtrsim 10^{12}\,h^{-1}\,M_\odot$）のハローに住む銀河サンプルに対して，銀河・銀河レンズ効果と銀河の2点相関関数を宇宙論パラメータの関数として予言する．われわれは，これを構築するために，2048^3体を用いた大規模なN体シミュレーションキャンペーンを推進しており，2018年3月現在で216試行のシミュレーションデータを蓄積している．

このシミュレーションには，多層のラテン超方格を用いてwCDM模型の6つの宇宙論パラメー

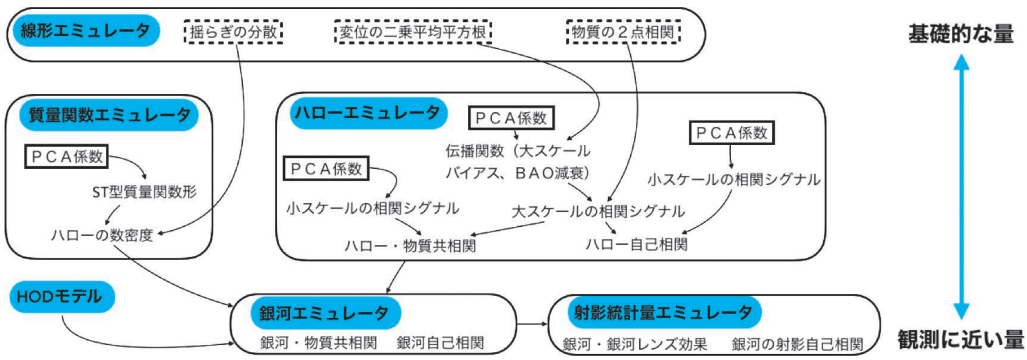

図8　ダークエミュレータの設計図．四角枠で囲まれた量はGPによりモデリング（破線：線形理論から計算した結果を元に学習，実線：シミュレーションデータを元に学習）．下部にある観測量を予言するエミュレータは，矢印で結ばれた，より基本的な要素を計算するブロックを呼び，その結果を解析的な表式により組み合わせることで予言を行う．計算の高速化のため，最上部の線形理論で計算可能な量までもGPで予測するようにデザインしている．

タθ_{6D}をサンプリングする．加えて，今，シグナルを測る距離R，赤方偏移z，銀河の環境を決める複数のパラメータ[*8]に対する依存性まで理解したいので，入力が20次元近くにまで大きくなってしまう．実は，これらの追加の入力次元については1度のシミュレーションから好きにサンプリングできるため，宇宙論パラメータ6次元部分はラテン超方格，残りの次元は正則格子という不均一なサンプリングに落ち着く．

レンズシグナル（$\Delta\Sigma$と書く）を多次元の入力空間の関数

$$\Delta\Sigma=\Delta\Sigma(\theta_{6D};R,z,\cdots)\,, \qquad (9)$$

と書くと，データ点の数が膨大になり，行列反転のコストからガウス過程への実装が難しくなる．ダークエミュレータでは，これらの追加の入力次元について，物理的考察から導かれた解析的な表式を手がかりにいくつかのビルディングブロックに分解する（図8）．そして，それらを主成分解析やモデルフィッティングを利用して少数の係数A_iに落とし込んでから，宇宙論パラメータθ_{6D}に対する依存性をGPでモデリングする：

$$\Delta\Sigma=\Delta\Sigma(A_1,A_2,\cdots)\,,\ A_i=A_i^{\mathrm{GP}}(\theta_{6D})\,. \quad (10)$$

図9は，ダークエミュレータのコアとなるハローの質量関数およびハロー・質量の共相関関数について予測の精度を検証したものである．それぞれ，上のパネルは訓練に用いた40モデルを，下のパネルは検定用の20モデルをエミュレータがどの程度よく予測しているかをプロットしていて，個々のモデルに対するシミュレーションデータと予測の比を実線で，それらの平均および標準偏差を誤差棒で示している．いずれの場合も，統計誤差の大きい高質量（$\gtrsim 2\times10^{14}\,\mathrm{h^{-1}}\,M_\odot$），長距離（$\gtrsim 20\,\mathrm{h^{-1}}\,\mathrm{Mpc}$）の領域を除けば比は1付近

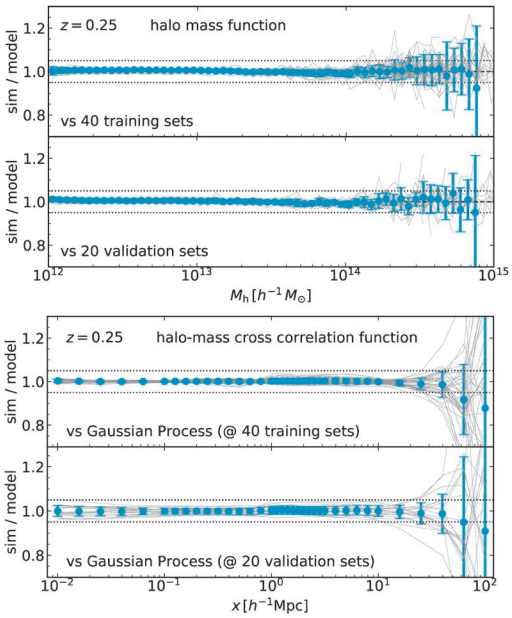

図9　ハローの質量関数（上）およびハロー・質量共相関関数（下）の精度のテスト．40モデルの訓練データと20モデルの検定データそれぞれについて，GPの予測値とシミュレーションとの比を示している．誤差棒はいずれもモデル間の平均および分散を示す．

にとどまり，誤差はおおむね2%程度，悪くても5%の範囲に収まっている．また，検定用データに対する予測の精度が訓練用データと比べてほとんど遜色ないことから，学習がうまくいっていることが確認できる．

5.　今後の展望

本稿では，少数の数値実験から多次元入力空間上でベイズ推定をすることで，新たな入力値に対する予測を与える方法論について紹介した．エミュレータはあらかじめ定められたパラメータの範囲内で高速かつ高精度な予言を与える．これをMCMCなどのサンプリング法に組み入れることで，観測データを説明するモデルパラメータのベ

[*8] ここでは，ハローの質量に応じて銀河の数を決める，5パラメータhalo occupation distribution（HOD）モデル[14]を採用する．さらに，中心銀河のハローの重心からのずれや，銀河サンプルの不完全性にかかわるパラメータ4つを導入する．

イズ推定ができる．興味あるモデルパラメータ空間をカバーできていれば，エミュレータはHSC以外にも使える汎用性の高いツールとなる．

ではひとたびデータを得て，そこからできるだけ正確な解析をしようと思った場合，エミュレータとは違った思想でシミュレーションセットを設計することになるかもしれない．エミュレータ構築のために広めに取った入力空間の中で，観測データと合う領域はほんの一部しかない．そこで例えばエミュレータを使って大まかにパラメータ範囲を狭めておき，モデルの当てはまりが良い領域のみに絞って細かくサンプリングを行い，シミュレーションデータを追加して予言精度の向上を図る．あるいは，そのような領域内で勾配法などを利用して，よりダイレクトにパラメータの最尤推定値に向かうようにシミュレーションを実行することなどが考えられる．HSCの初期データが得られた今，さまざまなアイディアを試すことで，統計解析の方法論自体を吟味することができるであろう．

謝　辞

本研究は日本学術振興会科学研究費補助金（15H05887, 15H05893, 17K14273），および科学技術機構CREST（JP-MJCR1414）のサポートの元で行われた．高田昌広さんからは原稿について貴重なコメントをいただいた．第4章で紹介したダークエミュレータは，高田昌広さんをはじめとするHSC弱重力レンズワーキンググループとの共同研究により開発を進めている．高橋龍一さん，大木平さん，白崎正人さん，大里健さんには，シミュレーションキャンペーンに参加・協力していただいた．機械学習の実装においては池田思朗さん，上田修功さんから貴重なご意見を頂戴した．最後に，本稿執筆の機会を与えてくださった編集委員の岡部信広さんに厚く御礼申し上げたい．

参考文献

1）http://sumire.ipmu.jp/
2）Skilling, J., 2004, AIP Conference Proceedings, 735, 395
3）Goodman, J., & Weare, J., 2010, Comm. App. Math. Comp. Sci., 5, 65
4）Takahashi, R., et al., 2009, ApJ, 700, 479
5）Springel, V., et al., 2005, Nature, 435, 629
6）Heitmann, K., et al., 2006, ApJ, 646, L1
7）http://www.hep.anl.gov/cosmology/CosmicEmu/index.html
8）McKay, M. D., et al., 1979, Technometrics, 21, 239
9）Ba, S., et al., 2015, Technometrics, 57, 479
10）平野照幸，2018, 天文月報111, 609
11）Neal, R. M., 1994, Technical Report CRG-TR-94-1, Dept. of Computer Science, University of Toronto
12）Tyson, J. A., et al., 1984, ApJ, 281, L59
13）http://www.sdss3.org/surveys/boss.php
14）Zheng, Z., et al., 2005, ApJ, 633, 791

Theoretical Predictions of Statistics of Cosmological Large Scale Structures: Machine-Learning Approach

Takahiro Nishimichi

Kavli Institute for the Physics and Mathematics of the Universe, The University of Tokyo Institutes for Advanced Study, The University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa, Chiba 277–8583, Japan

Abstract: We need appropriate theoretical templates for Bayesian inference to select a mathematical model or determine its parameters that explain the observed data. *N*-body simulation is a powerful tool to make predictions of cosmic large scale structures. However, using it to create theoretical templates is not realistic given that we have to cover typically a high-dimensional parameter space. Here, we introduce "dark emulator", which predicts in a Bayesian manner relevant statistics for the galaxy–galaxy lensing analysis from Hyper Suprime-Cam on Subaru Telescope. We explain basics of inference and prediction based on Bayes. Especially, we discuss in detail efficient methods to sample from a high-dimensional space as well as regression based on Gaussian Process.