

Flux 校正と誤差

八木雅文

概要

観測をして得たデータから天文学に繋げるには、他の観測結果や理論と比較可能な量を求めなければならない。ここで必要となるのが「校正」である。本資料ではこの「校正」についてと、解析や校正の際に重要な誤差の概念を概説する

1 いんどろだくしょん

本文書では主に撮像観測にまつわる校正と誤差の話が書かれている。撮像観測とは絵を撮る観測である。もう少し詳しく定義すれば、「ある決められた時刻の間に、ある波長範囲（＝エネルギー範囲）の光子が、天球上のある位置の範囲から、単位時間にどれだけの数来ているか」を計測し「光の強度の天球上での分布」として見る観測である。撮像と分光との違いは「波長範囲の切り方」と「天球上での分布」とどちらのパラメータを優先的に考えるかの違いである。

撮像観測の際、必要な各物理量は以下のように求められている。

- 時刻
可視だとシャッターで制御。赤外だと読み出しタイミングで制御。
- 波長範囲（＝エネルギー範囲）
光学フィルターを使う。大気や光学系も影響。
- 天球上の位置
望遠鏡の視野でまず区切る。その上で CCD や赤外アレイの画素（ピクセル）が天のある一部だけを見ているようにする。
- 単位時間にどれだけの数の光子
光子を光電効果で電荷に変え、一定時間貯めた後、電圧に変換して計測する。

観測を行なって、データを得て、そこから最終的に求めたいのは物理量だが、「観測データから物理量へ変換する」のが「校正」の話で、「物理量がどれだけ正しく推定できたか？」が「誤差評価」の話となる。

2 校正

ここは講義では資料では広辞林の定義を出していたのだが、それを全世界に公開してしまう資料でそのまま出しておくのは、どうも著作権が微妙な気がする（正当な引用の範囲かどうか微妙な気がする）ので、ここでは省略する。各自辞書で確認してもらいたい。このような事があったりするので、講義資料を全てウェブ公開する事について最近私は消極的になっているのだが、それはそれとして、光赤外のデータ解析で言う「校正」は辞書の定義より広い範囲の操作も含まれている。例えば、データ整約・解析の講義で説明された、ダーク補正、フラット補正も校正と呼ばれたりする。本資料では flux(光量) の校正について以下で検討するが、この他、校正には分光観測での波長校正、また、撮像での位置校正などもある。

さて、校正の目的は、観測された量が他の観測や理論と比較できるようにする事であり、このような比較が不要な量に関しては、校正は必ずしも必要とは限らない。例えば超新星を見つけるのは、時期の異なる 2 つの画像がありさえすれば、そして、画像の中での相対校正ができていれば、その画像の中の天体の等級がわかっていなくても可能。つまり、（絶対）校正は必ずしも必要ではない。一方、その超新星の光度を求めたくなるときには（絶対）flux 校正が必要となる。

2.1 flux

Flux(正しくは flux 密度) とは、(ある天体・領域から) 観測者の側の単位面積を、単位時間にどれだけの光が通過するかという流量である。Flux を測る単位としては、エネルギーを用いた $[erg/sec/cm^2]$ という単位が光赤外分野では用いられ、分光の場合は、これを単位波長当たりで測った $erg/sec/cm^2/\text{\AA}$ や単位周波数当たりで測った $erg/sec/cm^2/Hz$ が用いられる場合が多い。

一方、光赤外分野で天体の明るさの指標として用いられる「等級」は、歴史的には人の目で見た明るさを元に決められたものだが、現在は Flux の常用対数を -2.5 倍したものに定数を加えた値が用いられる。加える定数としては、以下の 2 種がよく用いられる。

- Vega

こと座の Vega のみかけ等級が 0 等となるように定数を加える (はずだったのだがその後の補正で Vega は +0.03 等)

- AB

単位周波数当たりの有効フラックス密度を cgs 単位系で F_ν [erg/cm²/s/Hz] として表現した後、

$$m(AB) = -2.5 \log F_\nu - 48.6$$

で求める。

これら 2 つの等級は、V-band で一致するが、赤外で数等異なるので注意が必要である。

2.2 フィルター

撮像観測の場合、ある波長域の光だけを通すフィルターを通して観測を行うのが普通である。しかし、フィルターを通して測定した Flux は、そのままでは観測所や観測装置ごとに異なってしまう、互いに比較できない。そこで、例えばすばる + Suprime-Cam のデータを「仮に SDSS のシステムで観測したらどうなるはずか」という値に変換する。この変換先の標準測光システムとしては、可視では Johnson, Kron-Cousins, SDSS, HST などがよく使われる。赤外ではより多種多様なシステムがあり、例えば 2MASS, MKO, CIT, UKIRT などがある。

理想的には、フィルターも含めた透過率（どの波長の光をどれだけの効率で検出するか）と、検出器がどれだけの flux を受けたかが分かれば標準システムへの変換は可能である。しかし、感度特性はフィルターだけでなく大気状態、望遠鏡、観測装置、などに影響される。検出器も本当にどれだけの flux を受けたかは簡単には測定できない。そこで、予め別の方法で標準システムでの flux が求められた「測光標準天体」を観測して、それとの比較で標準システムでの flux を求めるのが典型的な flux 較正となる。

2.3 データ整約

一般に CCD 撮像のデータは

$$data = \frac{((object + sky) \times t_{obs} \times flat + dark \times t_{dark})}{gain} + bias$$

だとした場合、整約は

$$object = \frac{(data - bias) \times gain - dark \times t_{dark}}{flat \times t_{obs}} - sky$$

となるが、右辺のうち、明らかなのは data と、時刻関連の t_{dark}, t_{obs} だけである。他の情報はどうやって求めるかといえば、

- bias, sky: 撮ったデータ自身から求める
- (bias), dark, flat: 較正用の別のデータを取っておく
- gain, (flat): 標準星で較正

であろう。

この較正用データとしては、可視近赤外撮像の場合は例えば以下になる。

- BIAS: 0 秒露出。CCD をクリアして直後に読み出す。
- DARK: シャッターを開かず暗電流だけを貯める。
- FLAT: ドーム内のスクリーンを撮る。あるいは夕暮れや夜明けの明るい空を撮る。

- STANDARD_STAR : 等級が分かっている天体を撮る

アーカイブデータを解析・較正するときも、目的天体だけでなく、これらのデータも必要になる。さて、

$$object = \frac{(data - bias) \times gain - dark \times t_{dark}}{flat \times t_{obs}} - sky$$

ここで、較正用データなどを使いつつ、検出器出力を光子の数に戻すには、実用上は以下のようなになる。

1. ゲタ (bias) を引く
2. 暗電流 (dark) の効果を引く
3. ピクセル間の相対感度を補正する (flat)
4. スカイを引く
5. カウントを Flux (光子の数) に直す

となる。

なお、整約の際に注意すべき点は

「本当に欲しい情報は何か？」を見失うな！

という点である。データ整約では全ての情報の誤差を増やさず進めるのは難しい。欲しい情報は位置なのか、色なのか、形状なのか、絶対 Flux なのか、その情報の誤差を増やさない手法をとることが重要である。

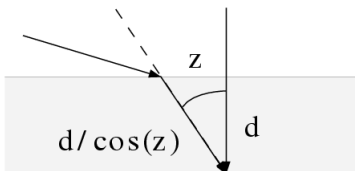
2.4 標準測光システムへの変換

ここはなぜ標準星が必要なのかという話である。前節で触れた bias,dark,flat だけでは標準測光システムに変換可能ではない。この一つの理由は flat の絶対値が明らかではないからである。更に、ズレを生じさせる（主な）要因として、

- 大気による吸収・散乱（大気減光）
- フィルターの透過特性の違い

などがあり、これらをそれぞれ考慮し、必要な場合は必要に応じて補正してやる必要がある。これらの補正をどのように行なうかについて、観測前に考えた上で、後で必要となるデータを、観測中に取得できるうちにしておくことが大切である。

2.4.1 大気の影響



大気は光を吸収・散乱で減光し、自分も光る。大気が光る分は、空間的な構造を見て天体のない領域の観測値を用いて補正する。（うまく補正できない場合もある）吸収・散乱による減光は特に短い波長域で影響が大きく、また大気の厚みによって変わる。Flux は厚みの指数で変化するので、

$$F_{obs} = F_0 \exp(-kx)$$

となり、観測値を等級に変換した値（機器等級）は、大気の厚み x に対して線形に変化する。

$$m_{obs} = m_0 + k'x + \Delta m$$

この厚み（エアマス）は、天頂からの角度（天頂距離）を z として近似的に $1/\cos(z) = \sec(z)$ で表される。異なった高度で標準天体を観測すれば係数 k' と Δm が求められる。この係数を使って、任意の天頂距離での大気の吸収を求めることができる。

この係数は日によっても異なる場合があるので、対象天体を観測した日に標準星を取得しておく事が重要となる。

2.4.2 フィルターの違い

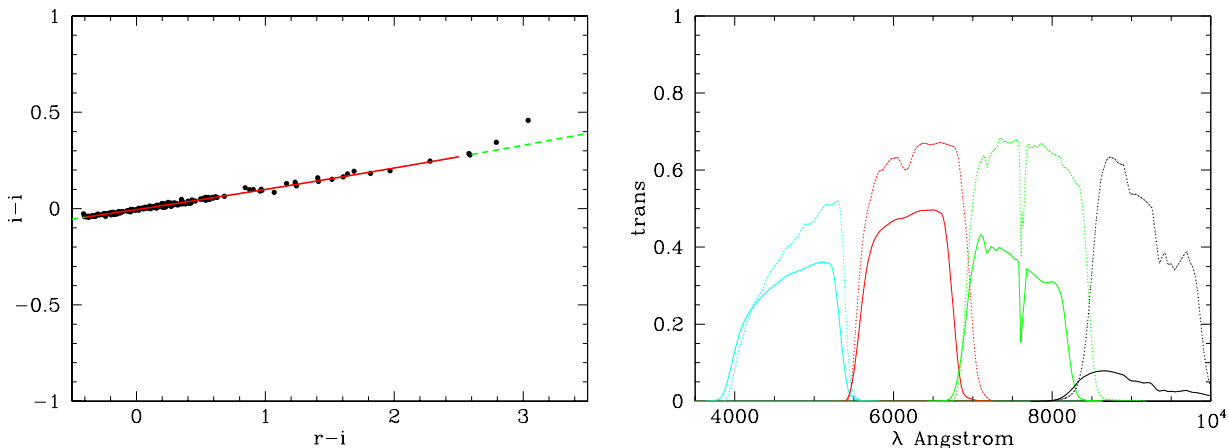
異なった波長特性のフィルターで観測したデータはそれだけでは標準システムに変換する事は不可能である。例えば標準フィルター R と、長波長側に寄った R' とで同じ天体を撮った場合、赤い天体は R' では R より flux が多くなり、青い天体では R より flux が少なくなる。

そこで、観測しているフィルターの波長域で、天体のスペクトルが大きく変化しないと仮定すれば、これを天体の色で補間できる、と仮定できる。例えば R' を求めるのに (V-R) の色を使って

$$R' = R + b(V - R)$$

して、係数 b を観測結果のフィットやモデル計算から求める事がよく行われる。このためには様々な色の標準天体の観測が必要となる。但し、この直線補間は近似であり、実際にはこの式のように線形になるとは限らない。

この関係をスペクトルライブラリの星のスペクトルを使い標準システムで測光した場合と Suprime-Cam で測光した場合を、大気吸収、主鏡、主焦点光学系、フィルター透過率、CCDの量子効率、を考慮して計算した例が以下の左図である。



Suprime-Cam i は SDSS i と同じ名前ではあるが、長波長側のカットオフが長い方に伸びている。このため、0.3 程度の系統差が出るので注意が必要である。

2.5 校正のまとめ

まず、自分が行いたいサイエンスには、どの程度の校正精度・正確度が必要か、予め見積もる。その精度・正確度に必要なデータを把握し、観測計画を立て実施することが重要である。データの後からの補完はできない事も多い。解析の際も、目標精度・正確度を念頭に置き、その精度を損なう手法を行っていないか常に気をつけること。その上で、手を抜いていい箇所は手抜きをしてよい。時間は有限だからである。

3 誤差

「測定値・理論的推定値また近似計算によって得られた値と、真の値との差（大辞林）」

ここは著作権法32条の引用の基準を満たしていると判断しそのまま残す。

さて、ここで「真の値」とは何かが問題である。測定値から、どのような値を求めたいと考えているのか？それを常に心に留めておく必要がある。ただ、実際の観測結果からは、真の値はわからず、観測値だけがあるので、厳密な事を言えば誤差は求められない。これが「不確かさ」の考え方が出てきた理由であるが、この「不確かさ」に関しては今回は深く触れない。

時間変化しないものは、真の値を言うのは簡単である。例えば真空中の光速度は厳密に 299792458 m/s なので、これと異なる測定値を得た場合は、その差が誤差となる。ところが、光の強さについてはそうはならない。光はミクロな確率過程を経て、天体から出てくるため「ある1秒に天体から出てくる光の量」は別の1秒のそれと同じではない。つまり、そもそも観測値は一定ではありえない。では、ある天体の Flux の値、あるいはある天体の明るさとは、何を求めているのかといえ、求めたい flux の値とは「長時間平均」である。この値は一般的には短時間に観測される光量の「確率分布」の「代表値」と見なすことができる。

明るさが変わらない天体を何度も観測して、その flux の観測値が $X_1, X_2, \dots, X_n$ だった場合、例えば、その平均を取れば、代表値の推定になる。測光の誤差とは、求めた値がどれだけ精度良くその分布の代表値を推定しているかという指標である。

3.1 誤差の見積もり

誤差を系統誤差とランダム誤差に分けて考えることは良く行われている。

- 系統誤差：何回やっても同じ量ずれる誤差
- ランダム誤差：毎回再現性のない誤差

機器の目盛りが間違っていて毎回同じ量ずれるのは系統誤差であり、機器が不安定で結果に再現性がないのはランダム誤差となる。ある1秒間に受ける光量の平均値からのずれ（バラツキ）は、機器の不安定とは別種だが、ランダム誤差の一種ではある。

さて、誤差を評価する方法には以下の2通りが考えられる。

- その1) 観測されたデータ自体を使って、それにどれだけの「バラツキ」が含まれるかを測定する。最終的に求めた数値に誤差棒を付けるのは、この場合が多い（と思う）
- その2) モデルを立てて、それを処理することでどのように誤差が増えていくかを計算する。観測立案や解析手法の吟味などに役立つ。

その1の例としては、「データを複数に分け、それぞれで解析し、その差を見る」「十分独立に近くなる程度の広い範囲でサンプリングして、その値の分散を見る」などが考えられる。以下、その2のモデルを立てる方の話を行なう。

3.1.1 観測モデル

まず、パラメタを含んだ観測モデルを考える。このパラメタのうちいくつかは、最終的に求めたい物理量となる。

$$Y = f(X_1, X_2, \dots, X_n)$$

このモデルで Y が観測され、パラメタ X のうちいくつかは別途測定されとする。 Y の観測 $y_1, (x_1)_1, (x_2)_1, \dots, (x_k)_1$ から、例えば X_n を推定したとき、どれだけ良く X_n を推定できるかを評価すれば、これが誤差評価となる。

観測モデルとして先にも述べた以下を考える。

$$data = \frac{((object + sky) \times t_{obs} \times flat + dark \times t_{dark})}{gain} + bias$$

観測量はCCDの1ピクセルのカウント値である。パラメタは、天体の明るさや観測条件などとなる。object, sky, dark, bias は一定値ではなくバラついていると考えられるので、それぞれの分布を考えなければならない。モデルとしてはある確率分布に沿ったそれぞれの変数を観測していると考ええる。

さて、この確率分布はどうなっているかだが、恐らくは、素子（ピクセル）1つ当たりの光電荷の数はポアソン分布に従い、そこを読み出したカウントは正規分布（ガウス分布）に従うだろう、と「仮定」する。また、独立性に関しては、

あるピクセルでの光電荷の数は隣のピクセルの光電荷の数とは、独立だろうと普通は假定する（可視近赤外の場合）ただし、実際それぞれの装置で、ピクセル間の独立性がどこまで成り立っているかは、恐らく未確認、だと思われる。では、なぜポアソン分布・正規分布と思うのかの根拠だが、これは二項分布の話にまで遡る必要がある。ここらは受講者に聞いたところ既に大学等で習っている話のようなのでばっさり省略するが¹重要なことは、ポアソン分布

$$P(k)(x) = \frac{k^x e^{-x}}{x!}$$

の平均と分散は k であり、ポアソン分布に従う X 、 Y という 2 つの変数があったとき、和 $X+Y$ もポアソン分布に従うという性質があることである。ちなみに、差 $X-Y$ はポアソン分布に従わない。

可視や近赤外の場合、天体から出てくる光はポアソン分布に従い、空からの光もポアソン分布に従っていると考え、またダーク電流もポアソン分布に従っていると考え（ことにする）。この結果この独立な 3 つの量を足した光子数とそれが検出器の中で確率過程で化けた「電荷数」もポアソン分布に従うと考えて良い

一方、ポアソン分布の定数倍はポアソン分布にならない。電荷数はポアソン分布に従うが、FITS になって出てきた値は一般に定数倍されているため、ポアソン分布には従わない。

では、

$$data = \frac{((object + sky) \times t_{obs} \times flat + dark \times t_{dark})}{gain} + bias$$

この式の例えば $dark \times t$ などダメではないかと一瞬騙されそうになるが、この式は実は代表値の式であり、各確率変数の式ではない。例えば最初の 1 秒の露出のカウントを t 倍したらそれはポアソン分布には従わないが、上の式はポアソンに従う $dark$ を t 秒積分しているので、 $t \times P(dark)$ ではなく、 $P(dark \times t)$ に従う。ここは混乱しがちなので要注意である。また、その前の $\times flat$ も、あくまでこれは確率過程が一つ加わっているだけで、常に一定数倍ではない。結局この式の中では、

$$((object + sky) \times t_{obs} \times flat + dark \times t_{dark})$$

ここまではポアソン分布に従うと考えてよいが、これをゲインで割ったらもうポアソンではない。また、その後に足しているバイアスは正規分布に従うと、通常は假定している。

正規分布

$$N(m, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

の概説も同様に省くが、ポアソン分布と異なり、正規分布の定数倍は正規分布になること、

$$aX \sim N(am, (a\sigma)^2)$$

また、正規分布に従う独立な X 、 Y の和もまた正規分布に従うこと、

$$X + Y \sim N(m_X + m_Y, \sigma_X^2 + \sigma_Y^2)$$

は本文書後半で用いている。また、一般的な分布の場合、

$$E(aX + bY) = aE(X) + bE(Y)$$

$$V(aX) = a^2 V(X)$$

また、 X, Y の共分散を $Cov(X, Y)$ と書いて、

$$V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$$

がなりたつ。 X 、 Y が独立な場合は $Cov(X, Y) = 0$ であり、

$$V(X + Y) = V(X) + V(Y)$$

が常に成り立つこと、また共分散には線形性

$$Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$$

¹あの数式を全部 LaTeX で打ち直すのはしんどいです

と交換性

$$\text{Cov}(X, Y) = \text{Cov}(Y, X)$$

があること、自分との共分散は分散であること

$$\text{Cov}(X, X) = V(X)$$

も後半で用いている²。

さて、ここまでの準備で観測値

$$data = \frac{((object + sky) \times t_{obs} \times flat + dark \times t_{dark})}{gain} + bias$$

この式の右辺は正規分布に従うと考えて良いことがわかる。その平均は

$$\frac{((object + sky) \times t_{obs} \times flat + dark \times t_{dark})}{gain} + bias$$

このものズバリで³、分散は、

$$\frac{((object + sky) \times t_{obs} \times flat + dark \times t_{dark})}{gain^2} + V(bias)$$

となる。

このように観測値の確率分布が、予め何らかの方法でわかっていれば「誤差の典型値」を計算できる。良く使われるのが標準偏差（誤差の二乗平均）で、 σ などと書かれる。分布を $p(X)$ 、平均を $E(X)$ として、

$$\sigma = \sqrt{V(X)}$$

$$V(X) = \int_{-\infty}^{\infty} (X - E(X))^2 p(X) dX$$

となる。

観測値の母分布がわかっている場合は、誤差の分布も推定でき、典型値も求められる。また、この測光値の例のように観測値の母分布の関数形はわかっている、あるいは仮定できるが、パラメタが分からない場合は、母分布のパラメタを観測値から推定し、誤差を評価することができる。

3.1.2 具体例：限界等級

撮像データの測光の限界等級とは、天体の測光値 (Signal) が、測光誤差 (Noise) の何倍かという比、 S/N が、ある一定値、例えば、 S/N が 3 以上の天体が何等かという値である。

例えばある観測結果で 28 等の天体が $S/N=5$ の時、 $S/N=5$ での限界等級が 28 等、という。等級のランダム誤差 Δm と、 S/N との関係は、

$$\Delta m \sim \frac{1.086}{S/N}$$

である。

この S/N をモデルを仮定して計算してみる。まず、それぞれの項を分けて考えると、 $object \times t_{obs} \times flat$, $sky \times t_{obs} \times flat$, $dark \times t_{dark}$, はポアソン分布、 $bias$ は正規分布。 $bias$ の標準偏差を、読み出し雑音 (readout noise) と呼んで、 r と書くことにする。

S/N の S の部分は上の式の $object$ を含む項、 $object \times t_{obs} \times flat / gain$ である。

次に、 N は、 $object \times t_{obs} \times flat$, $sky \times t_{obs} \times flat$, $dark \times t_{dark}$, $bias$ このそれぞれが独立だと仮定して、分散は、先にも書いたように

$$\frac{((object + sky) \times t_{obs} \times flat + dark \times t_{dark})}{gain^2} + V(bias)$$

である。

²定義 $\text{Cov}(X, Y) \equiv E((X - E(X))(Y - E(Y)))$ から簡単に導けますけどね

³そもそも代表値の式ですしね

したがって、S/N は

$$S/N = \frac{object \times t_{obs} \times flat / gain}{\sqrt{((object + sky) \times t_{obs} \times flat + dark \times t_{dark}) / gain^2 + r^2}}$$

となる。この S/N だが、 $t = t_{obs} = t_{dark}$ と近似して考えると、露出時間が短い場合は読み出し雑音が効き、

$$S/N \sim \frac{object \times flat}{gain \times r} \times t$$

と、露出時間に比例して S/N が向上する。一方、露出時間が長くなると、ポアソン雑音が効き、

$$S/N \sim \frac{object \times flat}{\sqrt{(object + sky) \times flat + dark}} \times \sqrt{t}$$

と、露出時間の平方根に比例してしか S/N が上がらないことがわかる。

3.2 データ解析と誤差

1 回の観測の 1 ピクセルでの誤差はこのように求められるわけだが、実際には複数のピクセルの複数回の観測データから物理量を推定することが通常行なわれる。このような場合、各観測値を物理量に直してから典型値に戻すのか、観測値の典型値を求めてから物理量に直すのか、2 通りが考えられる。通常は後者（観測値の典型値）が使われる。複数回の露出を足し合わせる際の誤差は、本来は位置ずれや P S F の変動の影響など様々な要素を考えなければならないのだが、今回は、「あるピクセルの中に来る flux の分布は一定」という仮定で、mean、median、clipped mean を比較してみる。その次に、画像の並行移動の影響を吟味する。

3.2.1 平均と中央値と clipped mean

ここでは、プラス方向に偏った外れ値入りの正規分布で、パラメタ推定を行なうことを考える。これはまあ宇宙線が良く入る CCD の観測データのモデルだと思ってよい。具体的には X が N(0,1) に従う N 個と、大きな外れ値 x (例えば a=100) H 個を含むようなモデルを考える。確率密度関数は $h=H/(N+H)$ を使って、

$$f(x) = \frac{(1-h)}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} + h\delta(x-a)$$

と書ける。

標本平均は $E(X)=ah$ である。証明略。

標本の中央値は、正規分布から N 個サンプルして、その大きい方から (N-H+1)/2 番目 を取ることに相当する。離散効果を無視すると誤差関数を含む以下の方程式を解くことになる。

$$erf(x) = \frac{1}{2(1-h)}$$

誤差関数の定義は以下を用いている。

$$erf(x) = \int_{-\infty}^x e^{-\frac{z^2}{2}} dz$$

例えば h をいくつか例にとって解くと、以下の表のようになる。

H	x
0.001	0.00089
0.01	0.0090
0.05	0.047
0.1	0.099
0.2	0.23
0.3	0.40

例えば 10 枚に 1 枚高い側の外れ値が乗る場合、単純に 10 枚の中央値を取ると、およそ 0.1 σ ほど高いほうに偏る結果となることがわかる。

次に推定値の分散を計算してみる。議論を簡単にするため、一旦外れ値を除いて考えると、 X の分散は定義より 1。標本の標準偏差を計算すると、 N 個の標本平均 E は

$$E = \frac{1}{N} \sum_{i=1}^N X_i$$

であり、各 X_i は独立なので、

$$V(E) = \frac{1}{N^2} \sum_{i=1}^N V(X_i) = \frac{1}{N}$$

なので、標準偏差は $1/\sqrt{N}$ となる。

中央値の場合、分散を求めるのはまたも数値計算になってしまうのだが、まず、中央値の確率分布を求める。一般に $f(x)$ に従う変数 X に関し、標本数 N が奇数の場合の中央値 y の分布 $g(y)$ は 1 つが y 、 $(N-1)/2$ 個が y 以上、 $(N-1)/2$ 個が y 以下になる確率になるので、 $k=(N-1)/2$ と置いて、

$$g(y) = {}_N C_k {}_{N-k} C_k f(y) \left(\int_{-\infty}^y f(x) dx \right)^k \left(\int_y^{\infty} f(x) dx \right)^k$$

この分布について分散を求めると、 N が十分大きいところで、 $\pi/2N$ に漸近する。

従って、中央値の標準偏差は

$$\sqrt{\frac{\pi}{2N}} \sim \frac{1.25}{\sqrt{N}}$$

となって、平均よりも 25%ほど大きくなる。この比は N が小さくなると若干緩和し、 $N=3,5,7$ ではそれぞれ 16%, 20%, 21% 増しである。

このように平均は外れ値に弱く、中央値は標準偏差（バラツキ）が悪くなるのだが、この両方の欠点を補うような方法が clipped mean である。これは最初に代表値を推定し、そこから大きく外れた値は除去 (clip) して、残ったデータで平均を取る手法で、うまく clip できれば、外れ値の影響は 0、標準偏差も $1/\sqrt{N(1-h)}$ となる。

問題は「最初の代表値の推定」と、「大きく外れた」という、しきい値の設定方法であり、どの値を基準にどの範囲の平均を取って clipped mean とするか、その「標準的な方法」というのは、多分ないため、その辺りが「気持ち悪い」量ではある、と筆者は感じている。

3.2.2 具体例：重み付け平均と S/N

露出時間が違うデータを足す場合などには、それらをシグナルでスケールして単純平均を取るよりも、重みを付けて平均した方が S/N が良くなる。この節では具体的に重みをどう付ければ最適かを考えてみる。 N 回の露出でそれぞれのカウントが一定になるように規格化してから、重みを付けて足すようなモデルを考える。

$$X_i \sim N(m, \sigma_i^2)$$

に従う変数を考え、重み付け平均

$$X = \sum w_i X_i$$

を重みの総和が 1 になるよう、

$$\sum w_i = 1$$

という条件で考えてみる。ここで、 σ_i はそれぞれ異なることに注意したい。

この X の期待値 $E(X)$ は平均の線形性から

$$E(X) = \sum w_i E(X_i) = \sum w_i m = m$$

となって、母平均に一致する。というか、そうなるように重みの総和が 1 になるように選んだのであった。問題は分散だが、

$$V(X) = \sum w_i^2 V(X_i) = \sum w_i^2 \sigma_i^2$$

となるので、これが最小になるように、 w_i を選びたい。

これは条件付最小値問題なので、ラグランジュの未定乗数法を使って解ける。

$$L = \sum w_i^2 \sigma_i^2 - \lambda (\sum w_i - 1)$$

の極値を求めてみると、

$$\frac{\partial L}{\partial w_i} = 2w_i \sigma_i^2 - \lambda = 0$$

なので、

$$w_i = \frac{\lambda}{2\sigma_i^2}$$

また、

$$\sum w_i = 1$$

より、

$$\lambda \sum \frac{1}{2\sigma_i^2} = 1$$

となり、

$$\lambda = \frac{1}{\sum \frac{1}{2\sigma_i^2}}$$

となる。

したがって、

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}$$

の重みを付けて平均を取るのが分散最小であり、従って S/N 最適となる。

具体例としてスカイと天体の flux 比が一定で、Readout が無視できる場合を考えてみると、露出 i で観測されたあるピクセルでの電荷数を Y_i と置くと以下のポアソンに従う。

$$Y_i \sim p(object_i) + p(sky_i)$$

天体とスカイの明るさを object, sky と置き、それぞれの露出での相対 flux を a_i とし、 $m=object+sky$ とすると、

$$Y_i \sim p(a_i m) \sim N(a_i m, a_i m)$$

足し合わせる前に天体の明るさが同じになるようにそれぞれの露出を規格化する。規格化後 X と置くと、

$$X_i = \frac{Y_i}{a_i} \sim N(m, \frac{m}{a_i})$$

この場合、最適な重みは

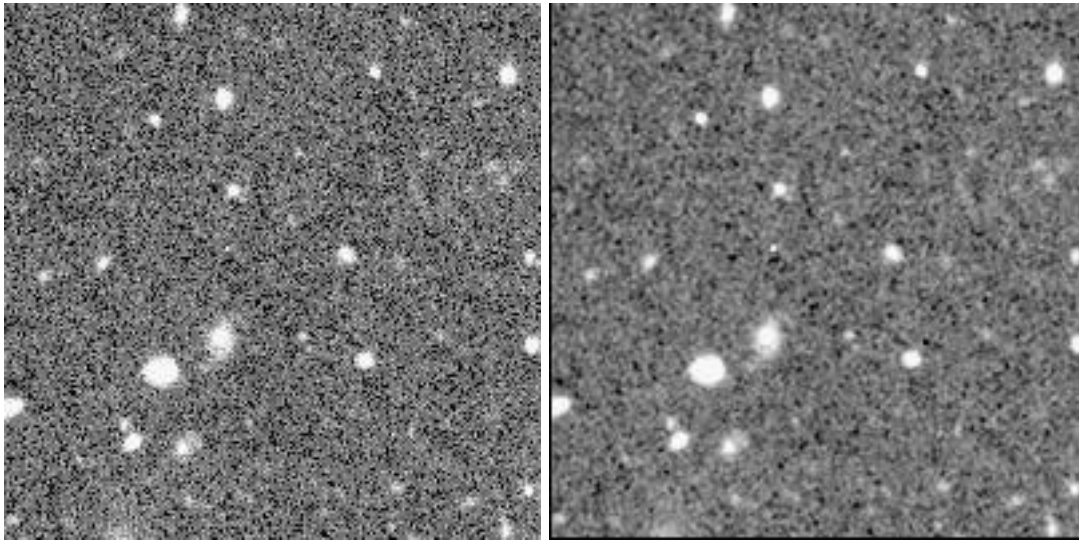
$$w_i = \frac{\frac{a_i}{m}}{\sum \frac{a_i}{m}} = \frac{a_i}{\sum a_i}$$

となる。よって、 X の重み付け平均は、

$$X = \sum \frac{a_i X_i}{\sum a_i} = \frac{\sum Y_i}{\sum a_i}$$

とするのが S/N 最適となる。

結局「スカイと天体の flux 比が一定で、Readout が無視できる」という条件では、観測値をそのまま足した後、適当な値で割るのが一番 S/N が良い、と言える。例えば相対 flux が、0.5, 1, 2 の3枚の画像があった場合は、その3枚を足して 3.5 で割るのが良い。ただしこの議論は外れ値の事をまったく考えてないのでそこには注意が必要である。



3.2.3 具体例：画像の並行移動

画像変形に伴う S/N の変化に関しては、回転等でも類似の問題があるのだが、ここでは並行移動を考える。画像・データを非整数ピクセルだけ並行移動させたい。例えば位置を合わせて足し合わせる場合など。方法として、線形補間のリサンプリングを考える。例えば x 軸方向に $-a$ ($0 < a < 1$) シフトさせるときは、

$$v'(x') = (1 - a)v(x) + av(x + 1)$$

である。実際にシフトさせてみたのが図である。

左がシフト前、右は $x+0.5, y+0.5$ シフトである。見た目でもざらつきが減っているのだが⁴、実際測るとピクセルあたりのノイズはシフト前 67ADU、シフト後 43ADU と、確かに減っている。すると並行移動で S/N が上げられる！と一瞬思うのは無理もない。

元画像での各点での標準偏差を σ とする。X 方向の移動だけ考えて

$$v'(x') = (1 - a)v(x) + av(x + 1)$$

とする。もし元画像で各点の誤差が独立であれば、並行移動後の各点での分散は

$$V((1 - a)v(x) + av(x + 1)) = (1 - a)^2\sigma^2 + a^2\sigma^2$$

となり、例えば $a=0.5$ の時、 $0.5\sigma^2$ となるので、確かに元画像の σ^2 よりも小さい。

しかし、並行移動操作の結果、隣り合った点の値は既に独立ではなくなっているため、広い範囲で見た場合の S/N は増えないのである。

具体的に $[0, N-1]$ の範囲の総和の誤差を計算してみる。元の誤差は独立な N 点の値の和なので分散は

$$V(v(0) + \dots + v(N - 1)) = V(v(0)) + \dots V(v(N - 1)) = N\sigma^2$$

次に、並行移動後の各点の値は

$$v'(0) = (1 - a)v(0) + av(1)$$

$$v'(1) = (1 - a)v(1) + av(2)$$

...

$$v'(N - 1) = (1 - a)v(N - 1) + av(N)$$

最初の 2 点の共分散を計算すると、

$$\text{Cov}(v'(0), v'(1)) = \text{Cov}((1 - a)v(0) + av(1), (1 - a)v(1) + av(2))$$

⁴見える？

ここで、共分散の線形性などを利用して展開する。

$$Cov(v'(0), v'(1)) = (1-a)^2 Cov(v(0), v(1)) + a^2 Cov(v(1), v(2)) + a(1-a)Cov(v(1), v(1)) + a(1-a)Cov(v(0), v(2))$$

ここで、 $v(x)$ は独立なので、 $Cov(v(1), v(1)) = V(v(1)) = \sigma^2$ だけが残って、

$$Cov(v'(0), v'(1)) = a(1-a)\sigma^2$$

この結果、

$$V(v'(0) + v'(1)) = V(v'(0)) + V(v'(1)) + 2Cov(v'(0), v'(1)) = 2(1-a(1-a))\sigma^2$$

となる。更に幅を広げて同様に計算できるが、

$$v'(0) + \dots + v'(N-1) = (1-a)v(0) + v(1) + \dots + v(N-1) + av(N)$$

と書いてやれば、各項は独立なので、この分散は

$$((1-a)^2 + (N-1) + a^2)\sigma^2$$

と簡単に計算できる。この結果、 N が大きくなると平行移動の効果が全くない事がわかる。並行移動で画像が綺麗になっても、ある範囲で積分した値、つまり天体で見た時の限界等級は上がらない。 S/N は上がってないのに、見かけ上、測定したピクセル当たりの RMS は小さくなるのであった。

一般的に、何らかの複数ピクセルを跨ぐような操作を行なった後の画像から測定した RMS から元の画像のピクセル当たりの誤差を推定するのは困難である。解析ソフトウェア等で、ピクセル当たりのノイズをパラメータとして与える事が求められたりするが、どのような数値が求められているのか？を正しく理解したうえで使わないと道を誤る。

このように、足し合わせや並行移動により、各ピクセルの誤差やある領域内の誤差は変わりうる。特に、バラツキ増加（ランダム誤差増加）ではなく、系統誤差が生じる場合は、データからは判断できない場合があるので、このようなモデルからの検証を行ない、どの操作がどう影響するかは把握し理解しておくことが望ましい。