S-PLUS による和歌の解析

小林恒夫
福島県立医科大学医学部物理学講座

1. はじめに

2004 年 2 月 15 日日曜の夕刻、筆者は和歌等の日本語のひらがな表記に番号付けを行い、得られる数値列に対し時系列解析を行うことを思いついた。以来、百人一首等の短歌を中心に、和歌についての統計解析・時系列解析を行っている [1,2]。開始当時の筆者は S-PLUS を使い始めたばかりで、毎日、S-PLUS 関連の書籍に首っ引きでコマンドを探したり、わからなくなると、中園さん・亀川さんをはじめとする数理システム [3] のスタッフの方々に尋ねたりしていたものである。短歌についてはほぼ一日一首のペースで解析を進めていった。この習慣は今でも続いている。現在は、ひらがなの成立に由緒の深い、古今和歌集に取り組んでいるところである。

2. ひらがなの数値化

例えば、いろは歌を基準に数値化するときは、いろは歌の 「い、ろ、は、に、ほ、へ、と、・・・」 を数値列 「1、2、3、4、5、6、7、・・・」 に対応させる。ひらがなに濁点のつく場合は 0.5 を加え、半濁音の場合は 0.3 を加えた。例えば、「は、ぱ、ば」 を数値化すると 「3、3.3、3.5」 となる。和歌には登場しないが、現代文等を扱うため、促音の 「っ」 は 「つ」 に 0.3 を加え 19.3、また句点の 「、」 は 49、読点の 「。」 は 50 としてみた。

五十音図を基準とする場合は 「あ、い、う、え、お、か、き、・・・」 を 「1、2、3、4、5、6、7、・・・」 と対応させることになる。

数値化の基準として、いろは歌と五十音図の他に、乱数なども試みたが、結果に大きな差はないようである。

数値化されてしまえば、和歌に対して時系列解析に限らず、一般的な統計解析が使える。ふつうとは順序が逆になってしまったが、時系列解析の次に、ヒストグラム解析、正規性検定、相関解析と進んでいくこととなった。

最後の頁に、筆者の解析ノートをお見せする。この例は百人一首第三十五番、紀 貫之の一首について解析してある。何をやっているかは、以下に述べていくが、2 段組になっていて、最下段以外は、左半分がいろは歌基準、右半分が五十音図基準の結果を示している。以下では解析の方法と、百人一首についての結果を例として述べる。

3. 時系列解析

自己相関・偏自己相関

時系列解析 [4] で基本となるのが、自己相関関数 (autocorrelation function, ACF) と偏自己相関関数 (partial autocorrelation function, PACF) である。両者を比較検討することにより、定常時系列か何らかのトレンドがあるか、自己回帰モデルがよくあてはまるか移動平均モデルがよく当てはまるか、といった判定になくてはならないものである。S-PLUS の

コマンドは、

```
acf(x, type="correlation", plot=T)
```

で標本自己相関関数が求まり、type="partial" で標本偏自己相関関数が求まる。X はひらがなを数値化した数値列で、plot=T はプロットするかしないかのオプションである。

解析ノートの左半分、あるいは右半分の図の最上段の左側が ACF で右側が PACF になっている。短歌の場合はデータ数が少ないこともあって、この ACF や PACF に顕著な特長が現れることはほとんどない。

解析ノートの 2 段目の図の左側は、数値化した和歌の時系列をプロットしている。ここでの S-PLUS コマンドは、

```
plot(t, x, type="l", xlab="", ylab="字番")
```

で、t は単に「1、2、3・・・」で、type="l" はデータ点を直線で結ぶオプションで、xlab、ylab は x 軸 y 軸のラベルである。

1/f ゆらぎ

時系列解析をやろうということで、いちばん初めに思いついたのは、当時筆者に関心のあった、1/f ゆらぎ（エフぶんのいちゆらぎと読む、英語では one-over-f fluctuation） [5] が和歌にみられるかどうかということであった。使った S-PLUS のコマンドは、

```
spec.pgram(x, spans = 5, detrend = F, demean = F)
```

である。このコマンドでは、自己相関関数をフーリエ変換して得られるパワースペクトル、通称ピリオドグラムとよばれているものを求めることができる。ピリオドグラム自体ではスペクトル密度の一致推定量を得ることができないので、平滑化が行われる。上のコマンドの中で spans = 5 は修正 Daniell 平滑化法によって平滑化するときの平均の次数を 5 とした、すなわちピリオドグラムデータの両側 2 つを使った平均の意味である。端点では半分のウエイトが使われる。detrend = F と demean = F はトレンド除去も平均値除去も行ってはいないことを表す。

解析ノートの 2 段目の図の右側に、上のコマンドで求めたパワースペクトルがプロットされている。また、4 段目の左側には、パワースペクトルの両対数プロットが示されている。このプロットが 45° の右下がりの直線に近ければ 1/f ゆらぎということになる。傾きの値は 4 段目の図の下に Smoothed Periodogram と書かれたところに 1/f ゆらぎの判定という形で記されている。

1/f ゆらぎの探索結果は、和歌のうち短歌では、1/f ゆらぎを示すものが多く、とくに百人一首に多くみられることがわかった [1]。百人一首 100 首のうち 1/f ゆらぎと判定されたものは、いろは歌基準で 88 首、五十音図基準で 95 首であった [2]。

自己回帰モデル

次に、時系列の周期性の検出によく使われる自己回帰モデル（autoregressive model, AR）による解析を行った。使った S-PLUS のコマンドは、

```
ar(x, aic=T, method="burg")、
```

である。aic=T は、AR の次数として赤池情報量基準（Akaike's information criterion, AIC） [4] の値が最小となるものを自動的に選ぶオプションである。method="burg" は Burg のアルゴリズムを、method="yule-walker" は Yule-Walker のアルゴリズムを使う、というオプションである。AR を求めるには二通りの方法があり、データ数が少ないときは Burg 法のほうが有利であることが多い。解析ノートの 3 段目の図の左側は Burg 法、右側は Yule-Walker 法による結果を示している。この例でも、Burg 法ではパワースペクトルにピークが検出されているが、Yule-Walker 法ではピークは検出されないとの結果である。自己回帰モデルで計算されたパワースペクトルのピークの位置は 4 段目の図の下、3 行目の AR（Burg）のところに、周波数と周期の値として記されている。周波数と周期は互いに逆数の関係にある。

百人一首では、2.7 文字の周期が最も多く検出され、いろは歌基準で 33 首、五十音図基準で 29 首が 2.7 文字周期を含んでいた [2]。この 2.7 文字の周期にはどんな意味があるのであろうか。例えば“の”の字が 2.7 文字おきに出現することが考えられるが、2.7 文字周期の和歌の歌詞を眺めてみても、そのような傾向はみられない。和歌の歌詞で、五十音図のはじめの方から引用される文字と終わりの方から引用される文字の出現の有様が 2.7 文字くらいの間に周期的に変わる、といえるかもしれない。そのような傾向が、いろは歌基準でも五十音図基準でも同様にみられるというのも興味深いところである。

4. ヒストグラム解析と正規性検定

文字数のカウント

順序が前後してしまうが、解析ノートの左上に、和歌のひらがな読みがあってそこに、

全 31 文字　文字種：21 字、21 字

とある。これは、和歌に使われているひらがなの全文字数と、異なる文字数を求めている。S-PLUS のコマンドは、

```
cat(paste(as.character(dname$hiragana), collapse=""),
 "　全", x.ar.burg$n.used,"文字",
 " 文字種：", length(unique(x)), "字、",
 length(unique(trunc(x))), "字", "¥n", sep="")
```

である。文字列の操作は、どうも煩雑であるが、length(unique(x)) で異なる数値の数を（濁点を区別―例えば“そ”と“ぞ”は区別）、length(unique(trunc(x))で異なる整数の数を求めている（濁点を区別せず―例えば“そ”と“ぞ”は区別せず）。この例では、全 31 文字のうち異なるひらがなの数は 21 字で、濁点の付いているものを区別しなくとも 21 字である、という結果となっている。

百人一首の 100 首についてしらべた結果の平均値は、濁点を区別して 23.2 文字（最大 28 文字、最小 18 文字）、濁点を区別しないで 21.8 文字（最大 27 文字、最小 17 文字）となった [2]。濁点を区別しない場合、ひらがな全 47 文字のうち平均として 46%のものが使われていることになる。案外限られた文字が使われていることがわかる。

ヒストグラム解析と正規性検定

解析ノートの４段目右側の図は、ヒストグラムと正規性検定の結果が示されている。図では、ヒストグラムは確率分布に変換してあり、総面積は 1 である。図中の曲線は標本標準偏差と平均値から計算した正規分布曲線である。ここの S-PLUS コマンドは少し難しく、中園さんに教わったとおりに、

```
swtest <- shapiro.test(x)
pvalue <- format(round(swtest$p, digits=4))
a <- paste("五十音 "," p = ",pvalue, collapse="")
hist(x, probability=T, xlab="", col=0, main=a)
pp <- ppoints(100); me <- mean(x); sd <- stdev(x)
qqpoints <- qnorm(pp, me=me, sd=sd)
lines(qqpoints, dnorm(qqpoints, me=me, sd=sd))
```

としてある。この例の正規性検定ではいろは歌基準で p = 0.0076、すなわち、正規分布とはいえず、五十音図基準では p = 0.26 > 0.05、すなわち正規分布といえる、という結果となっている。`shapiro.test(x)` が正規性検定のコマンドで、`hist(x,...)` はヒストグラムをプロットするコマンドである。

百人一首では、p > 0.05 だったものは、いろは歌基準で 50 首、五十音図基準で 74 首であった。また、p > 0.01 だったものは、いろは歌基準で 86 首、五十音図基準で 94 首であった [2]。五十音図基準の方が正規分布とみなせる和歌の数が多いという結果である。

5. 相関解析

解析ノートの最下段には同じ和歌をいろは歌基準で数値化したものと、五十音図基準で数値化したものとの相関解析を行っている。左の図は両者の散布図で、S-PLUS のコマンドは、

```
par(pty="s")
ls.out <- lsfit(x1, x2)
ls.results <- ls.print(ls.out)
plot(x1, x2, xlab="いろは", ylab="五十音", font=36)
abline(ls.results$coef.table)
```

である。`x1` はいろは歌基準、`x2` は五十音図基準のデータである。また、最下段の右には相関の有意性検定として、3 種類の相関係数による結果が記されている。S-PLUS のコマンドでは、

```
cor.test(x1, x2, alternative="two.sided", method="pearson")
```

で Pearson の積率相関が求まる。`method="kendall"` とすれば Kendall の順位相関が、`method="spearman"` とすれば Spearman の順位相関が求まる。ここでの例では、いずれも p > 0.05 で有意な相関は無い、という結果になっている。この相関が頻繁に有意になるようであれば、いろは歌基準と五十音図基準の二通りについて求める意味はあまりなくな

るのではないか、との発想でこの解析を行っている。百人一首では、93 首が $p > 0.01$ と、大部分が無相関であったので、その後もいろは歌基準と五十音図基準の二通りについて解析することとしている。

解析ノートからははずれるが、相関解析については、例えば百人一首の百首の中から二首をとりだし、二首のあいだに相関があるかどうか、ということもやっている。${}_{100}C_2 = 2450$ とおりの組み合わせに対する計算となり、筆者のパソコンでは for ループを使っているせいか 7 分くらいかかるが、百人一首の中で、いろは歌基準でも五十音図基準でも危険率 1%で有意な相関があるのは、

みかきもりゑじのたくひのよるはもえひるはきえつつものをこそおもへ
ながからむこころもしらずくろかみのみだれてけさはものをこそおもへ

の組と、

うらみわびほさぬそでだにあるものをこひにくちなむなこそをしけれ
おもひわびさてもいのちはあるものをうきにたへぬはなみだなりけり

の二組がある、などという興味深い結果を得ることができる。

5. おわりに

筆者は理系の研究が本業であるが、このように文系に関係した研究も始めることができ、幸運を感じている。

筆者のいつまでも初歩的な質問に辛抱強く答えていただいている数理ステムのスタッフの皆様に感謝いたします。また、今回、このような研究に目をとめてくださり、S-PLUS ユーザカンファレンスに発表の機会を与えてくださった数理システムの田澤氏に感謝いたします。

最後ながら、作品の選定や、ひらがなの入力等には全面的に妻に協力してもらっている。この場を借りて、感謝の意を表したい。

参考文献

[1] 小林恒夫、「和歌等における 1/f ゆらぎ」 比較文化研究, No.70, 13-22, 2005.
[2] 小林恒夫、「百人一首のスペトル解析およびヒストグラム解析」 比較文化研究, No.75, 1-8, 2007.
[3] ㈱数理システム http://www.msi.co.jp/splus/
[4] 北川源四郎、「時系列解析入門」 岩波書店、2005. ブロックウェル･デービス、逸見他訳、「入門時系列解析と予測」 シーエーピー出版、2004.
[5] 武者利光、「ゆらぎの世界」 講談社ブルーバックス、1980. 武者利光、「ゆらぎの発想」 日本放送出版協会、1998. 武者利光、「人が快不快を感じる理由」 河出書房新社、1999.

解析ノートの例

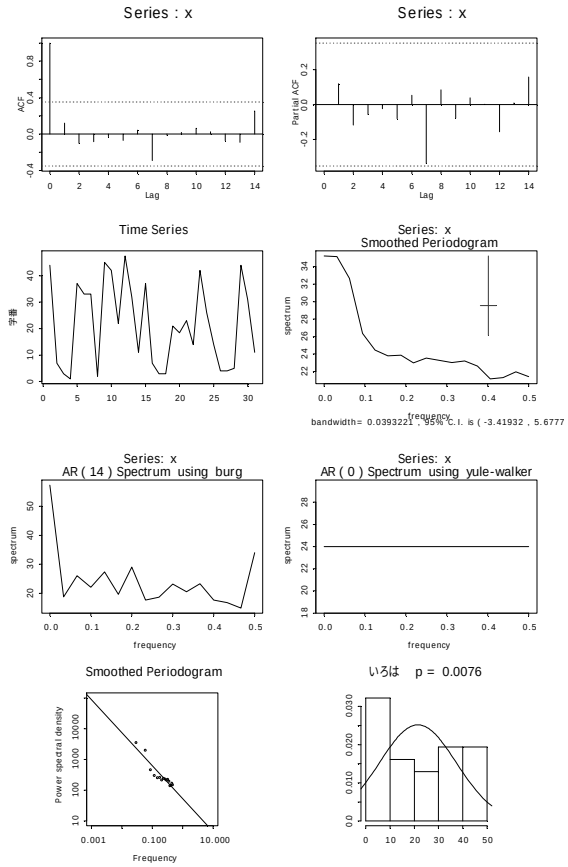
人はいさ心も知らずふるさとは花ぞ昔の香ににほひける

ひとはいさこころもしらずふるさとははなぞむかしのかににほひける (31字) 使用文字種: 21字、21字

百人一首　大岡 信　講談社文庫
紀貫之 (きのつらゆき)　三十五　古今集

2004-02-23

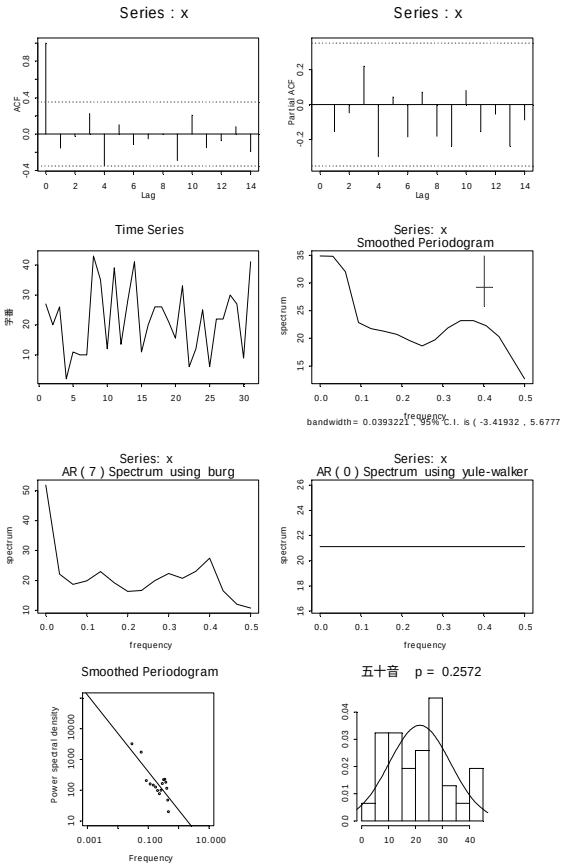
いろは歌を基準とする



Smoothed Periodogram
1/f ゆらぎの判定　-1.07 ± 0.11

AR (burg)
ピーク周波数　0.06666667　0.13333333　0.20000000　0.30000000　0.36666667
ピーク周期　15.000000　7.500000　5.000000　3.333333　2.727273

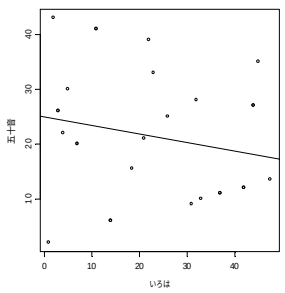
五十音図を基準とする



Smoothed Periodogram
1/f ゆらぎの判定　-1.24 ± 0.24

AR (burg)
ピーク周波数　0.1333333　0.3000000　0.4000000
ピーク周期　7.500000　3.333333　2.500000

いろは歌基準と五十音図基準との相関 (散布図)



いろは歌基準と五十音図基準との相関 (相関検定)

```
Pearson's product-moment cor.  p = 0.2403
Kendall's rank correlation tau p = 0.4738
Spearman's rank correlation    p = 0.407
```

(p > 0.05 で無相関)