Time Series Analysis in Astronomy

(Aikasarja-analyysi tähtitieteessä)

Code: PAP 312
Credits: 5

Lauri Jetsu
Department of Physics
University of Helsinki

Introduction

- **Lecturer:** Lauri Jetsu (lauri.jetsu@helsinki.fi)
- **Assistant:** Ari Leppälä (ari.leppala@helsinki.fi)
- **Magenta** colour www-links: symbols ☠ highlight
- **Lecturer's homepage** ☠

- Homepage **"Time Series Analysis in Astronomy"** ☠

Paper I **"Discrete Chi-square Method for Detecting Many Signals" ([2] Jetsu 2020, OJAp)** ☠

- **ONLY** 1 **Paper I**: Print, read and take to lectures
- **Introduces** Discrete Chi-square Method (DCM)
- **Applies** DCM to **simulated data**
- **Compares** DCM to other period analysis methods

Introduction ...

- Homepage **"Variable Stars"** ☠

Paper II **"Say hello to Algol's new companion candidates" ([3]Jetsu 2021)** ☠

- **2 PAPERS Paper I** and **Paper II**: Print, read and take to lectures
- **Introduces** DCM
- **Applies** DCM and other methods to **simulated** and **real** variable star data (e.g. **Paper II**)

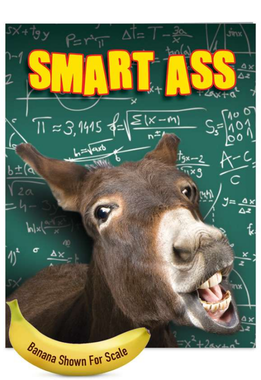
- **Different exercises** in courses **"Time Series Analysis in Astronomy"** and **"Variable Stars"**

$\longrightarrow$ **Study order** of **"Time Series Analysis in Astronomy"** and **"Variable Stars"** courses **flexible**

Introduction

Figure: @www.nobleworkscards.com

- **Question:** Do last term students have an advantage, because they already know DCM?
- **Answer 1:** Hopefully, they remember something about DCM, because all exercises are different.
- **Answer 2:** Next year: You will have the same advantage $\rightarrow$ Order of courses irrelevant
- **Answer 3:** Your future in Science? Good to learn DCM thoroughly: Artificial and real data analysis & DCM performance versus other methods, like DFT



Introduction ...

Status of papers

- **Paper I**: **accepted & published**
- **Paper II**: **accepted & published**

In all lectures

- Both **“Time Series Analysis in Astronomy”** and **“Variable Stars”** courses:
- Symbols of variables
- Equation, Figure, Table and Section numbers
- References
- Abbreviations ...

same as in Paper I and **Paper II**:

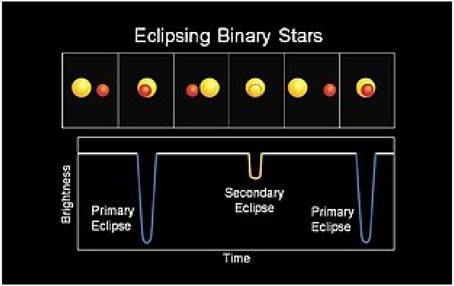
$\rightarrow$ **We save a lot of time**

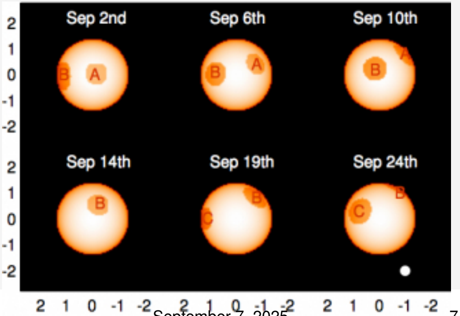
Introduction

- **Exercises** in **python**
- **We try** to use **same** symbols in all **python** program exercises, like
  T= t_i = time, **Y**= y_i = observation
- Important variables are written in **VIOLET capital or small letters** → Use same notations → Assistant can find them in your **python** programs
- DCM is an abstract method. It can be used to analyse arbitrary periodic, not only astronomical, phenomena
- **Observable** variability time scale
  → Can be observed in human time scale
  → DCM analysis possible

Introduction

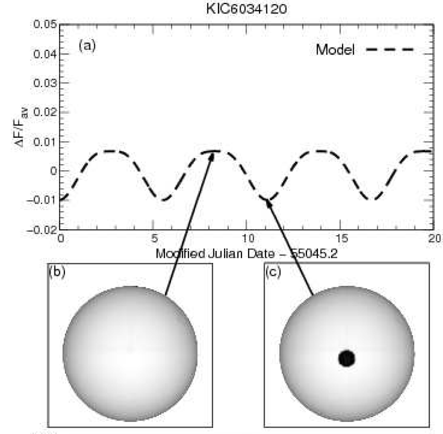
- For example, stars are variable, not constant, because they evolve
- **Observable periodic changes in variable stars:**
    - Eclipses
    - Starspots
    - Activity cycles
- DCM is general $\rightarrow$ Can be applied to **many periodic** phenomena





Background

- **First studies**
  [4](e.g. Jetsu et al. 1990)
- **Power spectrum analysis**
  [9] (Scargle 1982)
- Aug, 2021: 4741 citations
- Sinusoidal light curve
  $g(t) = A \sin 2\pi f(t - t_{min})$
  $A =$ Amplitude = spot size
  $P = 1/f =$ Rotation period
  $t_{min} =$ Minimum epoch
- **One constant** period for **one starspot**
  Figure from [10](Shibayama et al. 2013: their Fig. 3)



Background

- **Next studies** [6](e.g. Jetsu et al. 1999)
- **T**hree **S**tage **P**eriod **A**nalysis ([5]Jetsu & Pelt 1999: **TSPA**)
- Data divided into segments (seasons)
- Second order $g(t)$ light curve (double wave)
- P, A, $t_{min,1}$ and $t_{min,2}$ for **two** starspots
- **One constant** period for **two starspots**

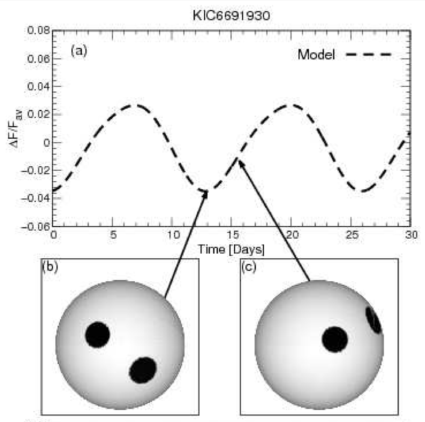
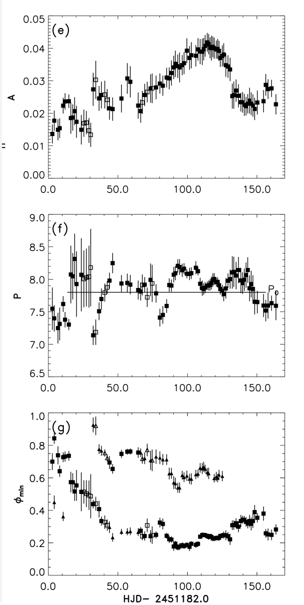


Figure from [10](Shibayama et al. 2013: their Fig. 4)

Background

- **Next studies** [8](e.g. Lehtinen et al. 2016)
- **C**ontinuous **P**eriod **S**earch ([7] Lehtinen et al. 2011: **CPS**)
- Sliding model window
- Best model identified
  - Constant
  - Sine wave
  - Double wave
- **One constant** period for **two starspots**

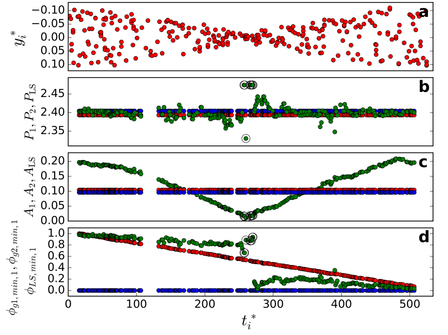
Figure [7](Lehtinen et al. 2011: their Fig. 7)



Background

- **Next studies** [1]
  (e.g. Jetsu 2019a)
- **Preliminary**
  Discrete
  Chi-square
  Method version
  ([2] Jetsu 2020 **DCM**)
- **Two constant**
  period light curves
  superimposed on a polynomial trend
- **Incompatibility** of one- and two-dimensional period finding methods, e.g. there are no **“flip-flops”**

  Figure [1](Jetsu 2019, his Fig. 11)



Background (Jetsu 2019a)

Imagine a face with a left eye ( ) and a right eye ( ). Both eyes can disappear and reappear. At any given moment, the number of eyes may be zero, one or two. The original stationary right eye can disappear and reappear only at fixed locations. The original non-stationary left eye rotates slowly around the head. We see this head spinning. Soon it is impossible to tell which eye is the original left or right eye. The only compatible pictures of this face are snapshots, but none of these snapshots can be used to recognize this constantly changing face. These snapshots can capture only one side of the head, or equivalently only half of the full visible surface of FK Com.

Abstract (Paper I)

Discrete Chi-Square Method for Detecting Many Signals

Unambiguous detection of signals superimposed on unknown trends is difficult for unevenly spaced data. Here, we formulate the Discrete Chi-square Method (DCM) that can determine the best model for many signals superimposed on arbitrary polynomial trends. DCM minimizes the Chi-square for the data in the multi-dimensional tested frequency space. The required number of tested frequency combinations remains manageable, because the method test statistic is symmetric in this tested frequency space. With our known tested constant frequency grid values, the non-linear DCM model becomes linear, and all results become unambiguous. We test DCM with simulated data containing different mixtures of signals and trends. DCM gives unambiguous results, if the signal frequencies are not too close to each other, and none of the signals is too weak. It relies on brute computational force, because all possible free parameter combinations for all reasonable linear models are tested. DCM works like winning a lottery by buying all lottery tickets. Anyone can reproduce all our results with the DCM computer code.

Files (Paper I)

- All **program**, **file** and **other related** items are printed in **violet** colour
- All necessary files are available in **Zenodo**

  dcm.pdf = **Paper I** manuscript

  dcm.py = DCM analysis **python** program

  dcm.dat = DCM control file

  TestData.dat = Simulated data file

  fisher.py = Fisher test **python** program
- **Copy four last files** from Zenodo to the same directory in your own computer
- **Do not use** Zenodo **Paper I** manuscript version (**dcm.pdf**), because it is the **submitted** version

Model (Paper I)

- **Observing times** = $t_i \rightarrow$ **Model zero** point $t = 0$ at t_1
- **Observations and errors** = $y_i = y(t_i) \pm \sigma_i$, $1 \leq i \leq n$
- **Mean** of y_i = m_y, **Standard deviation** of y_i = s_y
- **Model**

$$g(t) = g(t, K_1, K_2, K_3) = h(t) + p(t) \tag{1}$$

- **Periodic part** $h(t)$ is a sum of K_1 signals

$$h(t) = h(t, K_1, K_2) = \sum_{i=1}^{K_1} h_i(t) \tag{2}$$

- **i:th signal** is

$$h_i(t) = \sum_{j=1}^{K_2} B_{i,j} \cos(2\pi j f_i t) + C_{i,j} \sin(2\pi j f_i t) \tag{3}$$

- **Signal order** $= K_2$ (**dcm.py** can test only alternatives $1 \equiv$ sine wave and $2 \equiv$ double sine wave)

Model (Paper I)

- **Aperiodic part** is K_3 order polynomial

$$p(t) = p(t, K_3) = \sum_{k=0}^{K_3} p_k(t) \quad (4)$$

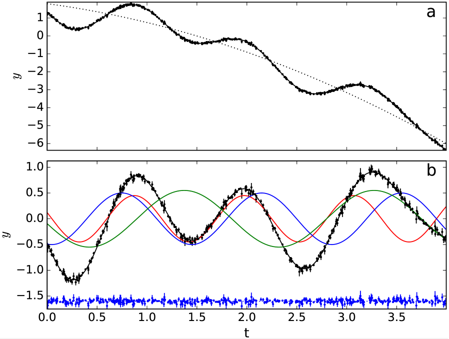
- **k:th** term is

$$p_k(t) = M_k \left[\frac{2t}{\Delta T} \right]^k \quad (5)$$

- It is difficult **to see** what this model means in reality
- **Figure on next page**
  - Three $h_i(t)$ signals ($K_1 = 3$)
  - All signals are sinusoids ($K_2 = 1$)
  - Signals superimposed on second order $p(t)$ polynomial ($K_3 = 2$)

Model (Paper I)

- **Time** = x-axis
- **Data** = y-axis



(a) Black dots = data = y_i

(a) Black curve $= g(t)$

(a) Dotted curve $= p(t)$

(b) **Removing** $p(t)$ trend

(b) Black dots = $y_i - p(t_i)$

(b) Black curve $= g(t) - p(t)$

(b) Red curve $= h_1(t)$ having period $1/f_1 = P_1 = 1.1$

(b) Blue curve $= h_2(t)$ having period $1/f_2 = P_2 = 1.4$

(b) Green curve $= h_3(t)$ having period $1/f_3 = P_3 = 1.9$

(b) Blue dots = Residuals $= \epsilon_i = y_i - g(t_i) =$ Data - model

Model (Paper I)

- **Problem:** If you only had **data**, black dots = y_i, **how** could you **unambiguously detect** $p(t)$ **trend** and three $h_i(t)$ **signals**?
- **DCM** succeeds in this!

Model (Paper I)

- DCM searches for combination of two patterns in data
- **Periodic** pattern $h(t)$ **repeating** itself
- **Aperiodic** pattern $p(t)$ **not repeating** itself
- **Sum of K_1 harmonic signals** $= h_i(t)$
  $f_i =$ signal frequency
  $K_2 =$ signal order
- **Polynomial K_3 order trend** $= p(t)$
- **Free parameters** of model

$$\begin{aligned}\bar{\beta} &= [\beta_1, \beta_2, ..., \beta_p] \\ &= [B_{1,1}, C_{1,1}, f_1, ..., B_{K_1,K_2}, C_{K_1,K_2}, f_{K_1}, M_0, ..., M_{K_3}]\end{aligned}$$

- **Number of free parameters**

$$p = K_1 \times (2K_2 + 1) + K_3 + 1 \qquad (6)$$

Linear and non-linear models

- **Main problem:** **Solution of best free parameter $\bar{\beta}$** values for analysed data $y_i \pm \sigma_i$?

  Definition: Model $g(t)$ has p free parameters $[\bar{\beta} = \beta_1, \beta_2, ..., \beta_p]$. This model is **linear**, **if all** $i = 1, ..., p$ model partial derivatives

$$\frac{\partial g(t)}{\partial \beta_i}$$

  do not contain any free parameter $\beta_1, ..., \beta_p$. The model is **non-linear**, **if any** of these partial derivatives contains any free parameter $\beta_1, ..., \beta_p$.

Linear and non-linear models

- Crucial difference between **linear** and **non-linear** models is
  - **Solution of free parameters** $\bar{\beta}$ is **ambiguous**, if the model is **non-linear**, because this solution depends on the chosen trial value $\bar{\beta}_{\mathrm{trial}}$. The final value $\bar{\beta}_{\mathrm{final}}$ is obtained from an iteration beginning from $\bar{\beta}_{\mathrm{trial}}$.
  - **Solution of free parameters** $\bar{\beta}$ is **unambiguous**, if the model is **linear**. No trial value $\bar{\beta}_{\mathrm{trial}}$ is required.
- **Conclusion:** If possible, analyse data with a **linear** model. Then all results are **unambiguous**. If a **non-linear** model is necessary, then some, or maybe even all, results are **ambiguous**.
- **Unambiguous = Unique**
- **Ambiguous = Not Unique**

Model (Paper I)

- DCM model $g(t)$ has p free parameters $\bar{\beta} = [B_{1,1}, C_{1,1}, f_1, ..., B_{K_1,K_2}, C_{K_1,K_2}, f_{K_1}, M_0, ..., M_{K_3}]$
- They belong to **two groups**

1st group = $\bar{\beta}_I = [f_1, ..., f_{K_1}]$

2nd group = $\bar{\beta}_{II} = [B_{1,1} C_{1,1}, ..., B_{K_1,K_2}, C_{K_1,K_2}, M_0, ..., M_{K_3}]$

- **1st group** $\bar{\beta}_I$ make model **non-linear**
- **If** $\bar{\beta}_I$ are fixed to constant known numerical values
  $\rightarrow$ Model becomes **linear**
  $\rightarrow$ Solution for remaining $\bar{\beta}_{II}$ free parameters becomes **unambiguous = unique**
- This is explained thoroughly Here: 08.09.2025 **ExerciseLinearNonlinear** ☠ **(A2024)** where linear and non-linear models are identified.

Model

What causes nonlinearity?

- **Simple answer:** All trigonometric terms, like $B_{i,j} \cos(2\pi j f_i t)$. Its partial derivatives

$$\frac{\partial[B_{i,j} \cos(2\pi j f_i t)]}{\partial B_{i,j}} = \cos(2\pi j f_i t)$$

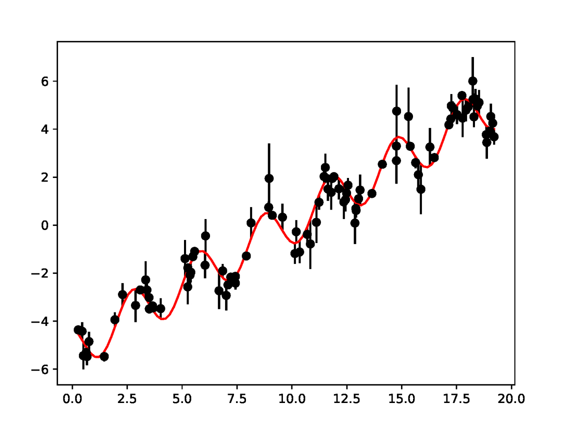
$$\frac{\partial[B_{i,j} \cos(2\pi j f_i t)]}{\partial f_i} = -B_{i,j}(\sin(2\pi j f_i t))(2\pi j t)$$

  contain free parameters f_i and $B_{i,j}$
- **If frequency f_i is fixed to a constant value**, frequency f_i is no longer a free parameter
  $\rightarrow$ The first partial derivative $\cos(2\pi j f_i t)$ no longer contains any free parameters, and there is no need for the second partial derivative
  $\rightarrow$ Model becomes **linear**
  $\rightarrow$ $B_{i,j}$ solution becomes **unambiguous = unique**

Model (Exercise)

- Simulating data using model **"Trend + Signal"**

ExerciseTrendSine ☠ (A2025)



Model (Paper I)

- **Paper I** statement
  "The first group of free parameters, the frequencies $\bar{\beta}_I = [f_1, ..., f_{K_1}]$, make this $g(t)$ model **non-linear**. *If these $\bar{\beta}_I$ are fixed to constant known numerical values, the model becomes* **linear**, *and the solution for the remaining second group of free parameters, $\bar{\beta}_{II} = [B_{1,1} C_{1,1}, ..., B_{K_1,K_2}, C_{K_1,K_2}, M_0, ..., M_{K_3}]$, is* **unambiguous**."
  should now be clear.
- In other words, **if** we test a frequency grid, where every tested frequency combination $\bar{\beta}_I = [f_1, ..., f_{K_1}]$ has fixed numerical constant values, then all these DCM models are **linear** and all results are **unambiguous**.

Model (Paper I)

- **Residuals**

$$\epsilon_i = y(t_i) - g(t_i) = y_i - g_i \qquad (7)$$

  are differences between **data** and **model**
- Residuals ϵ_i can be positive (**data** y_i above **model** g_i) or negative (**data** y_i below **model** g_i)
- **Good model**
  - **Mean** of ϵ_i residuals close to zero = Data at both sides of model = Model goes through data
  - **Standard deviation** of ϵ_i residuals equal to σ_i errors of data
  - **Absolute values** of individual ϵ_i residuals equal to errors of individual data $= |\epsilon_i| \approx \sigma_i =$ more accurate data closer to model

Model (Paper I)

- **Chi-square**

$$\chi^2 = \sum_{i=1}^{n} \frac{\epsilon_i^2}{\sigma_i^2} \quad (8)$$

  is sum of squared residuals divided by errors σ_i
- **Test statistic** χ^2 can be computed only if errors σ_i are **known**
- **Good** model has **small** χ^2
- **Bad** model has **large** χ^2
- **Reasonable** model has

$$\chi^2 \approx n,$$

  because $|\epsilon_i| \approx \sigma_i \Rightarrow \epsilon_i^2/\sigma_i^2 \approx 1$

Model (Paper I)

- **Sum of squared residuals**

$$R = \sum_{i=1}^{n} \epsilon_i^2. \quad (9)$$

- **Test statistic** R can be computed even when errors σ_i are **unknown**
- **Good** model has **small** R
- **Bad** model has **large** R
- **Least Squares Fit (LSF)** method gives **solution for free parameters** $\bar{\beta}$. This method solves $\bar{\beta}$ values that
  - **Minimize** χ^2 when errors σ_i are **known**
  - **Minimize** R when errors σ_i are **unknown**

Least Squares Fit = LSF

ExerciseSineFit ☠ **(A2024)** and **ExerciseTrendSineFit** ☠ **(A2025)** show how Least Squares Fit (LSF) is done in **python**.

- **Both** can be solved without presenting **the other exercise!**
  - **scipy** subroutine **optimize.leastsq** is **numerical** $\rightarrow$ No need to code model $g(t)$ partial derivatives
  - Only three subroutines are needed

    Model(T,BETA)
    Funct(BETA,T,Y,EY)
    LSF(T,Y,EY)
  - Many models can be applied in the same program by simply changing names of these three subroutines
  - Code **Model** $\rightarrow$ **Funct** always same $\rightarrow$ Only dimensions of **BETA** must be adjusted in **LSF**

Least Squares Fit = LSF

Memorize: **Least Squares Fit = LSF**

- Download **dcm.py**, **dcm.dat** and **TestData.dat** from **Zenodo** ☠
- Edit **only dcm.dat**. Do **NOT** edit **dcm.py**. Mistakenly edited? No worries, just download all files again.

ExampleDCMmodels ☠

- Explains **dcm.py linear** and **non-linear** model codes
- **Advice: Re-read** this example several times during this course → At this first time, you do not have to understand everything about this example → Print all seven pages of this example, **reread, reread, ...**

LSF of dcm.py in a nutshell

- Six **subroutines**: Two three **subroutine** models.
- Three free parameter groups:
  - **Frequencies**
  - **Signal amplitudes**
  - **Polynomial coeffients**

LinearLSF **Lfunct** **LinearModel**	**NonLinearLSF** **Nfunct** **NonLinearModel**
Frequencies: **Not free** fixed tested values	Frequencies: **Free** parameters
Signal amplitudes: **Free** parameters	Signal amplitudes: **Free** parameters
Polynomial coefficients: **Free** parameters	Polynomial coeffients: **Free** parameters

- Any K_1, K_2 and K_3 combination: All six subroutines work.

References I

[1] L. Jetsu.
Real light curves of FK Comae Berenices: Farewell flip-flop.
arXiv e-prints, page arXiv:1808.02221, Aug. 2019.

[2] L. Jetsu.
Discrete Chi-square Method for Detecting Many Signals (**Paper I**).
The Open Journal of Astrophysics, 3(1):4, Apr. 2020.

[3] L. Jetsu.
Say Hello to Algol's New Companion Candidates (**Paper II**).
ApJ, 920(2):137, Oct. 2021.

[4] L. Jetsu, J. Huovelin, I. Tuominen, O. Vilhu, B. . W. Bopp, and V. Piirola.
Rotational and long-term variations of HD 199 178 - A double period activity cycle.
A&A, 236:423–439, Sept. 1990.

[5] L. Jetsu and J. Pelt.
Three stage period analysis and complementary methods.
A&AS, 139:629–643, Nov. 1999.

References II

[6] L. Jetsu, J. Pelt, and I. Tuominen.
Time series analysis of V 1794 Cygni long-term photometry.
A&A, 351:212–224, Nov. 1999.

[7] J. Lehtinen, L. Jetsu, T. Hackman, P. Kajatkari, and G. W. Henry.
The continuous period search method and its application to the young solar analogue HD 116956.
A&A, 527:A136, Mar. 2011.

[8] J. Lehtinen, L. Jetsu, T. Hackman, P. Kajatkari, and G. W. Henry.
Activity trends in young solar-type stars.
A&A, 588:A38, Apr. 2016.

[9] J. D. Scargle.
Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data.
ApJ, 263:835–853, Dec. 1982.

References III

[10] T. Shibayama, H. Maehara, S. Notsu, Y. Notsu, T. Nagao, S. Honda, T. T. Ishii, D. Nogami, and K. Shibata.
Superflares on Solar-type Stars Observed with Kepler. I. Statistical Properties of Superflares.
ApJS, 209(1):5, Nov. 2013.