UNITED APPLICABLE STATISTICS: MID-DISTRIBUTION, MID-QUANTILE, MID P CONFIDENCE INTERVALS PROPORTION p

by

Emanuel Parzen

Department of Statistics, Texas A&M University, College Station, Texas, USA

Abstract

We believe that approaches to analysis of complex data can be developed from our United Applicable Statistics ("analogies between analogies") concept of a unified learning framework for almost all of the Science of Statistics, which we distinguish from the Statistics of Science. Important tools are the novel probability and statistical theory of mid-distributions, mid-quantiles, new way to calculate (for data with ties) sample quantiles and median (mid), asymptotic normality of mid-distributions of Binomial, Poisson, hypergeometric distributions. We advocate statistical inference by mid-PVALUE function of a parameter whose inverse (under a stochastic order condition) is defined to be confidence quantile (of a confidence distribution). We show mid-P frequentist confidence intervals for discrete data have endpoint function equal to confidence quantile, which is algorithmically analogous to Bayesian posterior quantile. One computes frequentist (without assuming prior) but interprets Bayesian. We conclude with $0-1$ *data inference, and quasi-exact (Beta distribution based) confidence quantiles of parameters* p *and log-odds* (p)*. We claim quasi-identity of frequentist mid-P confidence intervals* *and Bayesian posterior credible intervals with uninformative Jeffrey's prior. For parameters of standard probability models, calculating confidence quantiles yields Bayesian posterior quantiles for non-informative conjugate priors and provides frequentist motivation for conjugate priors.*

Keywords: Mid-distribution, mid-quantile, inversion formula from characteristic function, mid-PVALUE, confidence intervals for parameter p, normal approximation to discrete distributions, convergence in mid-distribution.

1 In Honor of Professor Ben Kedem

I am honored to speak at the celebration of the outstanding career of Ben Kedem, and to thank him for his leadership and pioneering contributions to Statistical Analysis of Time Series and Spatial data.

One way to show that Ben is a Leading Expert on Time Series Analysis is to compare what he teaches in his course on Time Series Analysis. I find the syllabus of Ben's 2009 course Stat 730: Time Series Analysis a role model of a comprehensive course. It covers:

(1) Spectral Analysis

(2) Filtering

(3) ARMA modeling

(4) Model selection criteria AIC

(5) Box Jenkins modeling

(6) State space modeling, Kalman filtering

(7) Regression methods for time series

(8) Spatial prediction

(9) Higher order crossings

Only a few important topics might be missing:

(10) Reproducing kernel Hilbert space inference

(11) Long memory models

(12) Non-linear models ARCH GARCH

(13) Copula nonparametric models

Ben's leadership in research and pedagogy is demonstrated by his emphasizing in his teaching analogies between methods for time series and spatial data.

I have been an admirer and friend of Ben for many years because of our interests in time series. This conference in Ben's honor provides additional evidence to the administrators of the University of Maryland of the eminence of Ben Kedem. He provides its statistics program with unique strength in Statistical Time Series Analysis that I believe ranks Maryland in the Top Five American programs in Time Series Analysis.

Since Parzen (1977), (1979) my research interests have been quantiles and their role in the unification and synthesis of diverse statistical methods: frequentist, Bayesian, nonparametric, functional, concrete (continuous and discrete) data analysis. I believe that formulas which apply to both continuous and discrete variables can be especially applicable to high-dimensional data.

Many statistical methods are analogous because they are about comparing distributions and testing the equality of two distributions:

$$H_0 : P = F(y) = G(y), \text{ for all } y,$$

which I express in terms of the quantile (inverse distribution) function $G^{-1}(P)$

$$D(P; G, F) = F(G^{-1}(P)) = P \text{ for } 0 < P < 1.$$

I call $D(P; G, F)$ a comparison distribution. In the continuous case it has comparison density (or relative density)

$$d(P; G, F) = f(G^{-1}(P))/g(G^{-1}(P)).$$

To define the comparison density in the discrete case replace the probability density functions f and g by probability mass functions p_F and p_G.

Kedem's research on combination of information from many samples can be viewed as estimating comparison (or relative or grade) densities. These (neglected) problems are very important for applications and deserve study by statisticians who worry that there are no open problems for statistical research.

I recommend that a cure to the feeling of the statistical community that it gets "no respect" is to define a *frontier* of statistical research. This is difficult because the discipline of statistics is composed of a few mainstream "fads" and many mini-communities which pay no attention to many advances in the Science of Statistics that are ready to be applied to other fields of science (which I call the Statistics of Science).

Statisticians are in a battle for leadership of the “Statistics of Science” (applications). In my view young statisticians (who may be concerned less with the Science of Statistics and more with the practice of the Statistics of Science) should realize that they can remain relevant (and competitive with econometricians and computer scientists) only by emphasizing their expertise in the “Science of Statistics” (applicable theory, emphasizing “analogies between analogies” which provide solutions in one field by technology transfer of solutions of problems in other fields that have been successfully solved statistically).

EXAMPLE of inappropriate practice of “analogies between analogies” reasoning: An alleged cause of the current Wall Street caused economic crisis is practice of analogies between measuring risks of (1) deaths of related people and (2) defaults of related securities, described in newspaper articles as from “couples to copulas”.

Academic statisticians need to practice United Statistics to fulfill their applied research and teaching missions, to innovate applicable methods for analysis of complex data, and to fulfill their enormous responsibilities of teaching statistical thinking to millions of students seeking (or forced) to learn statistical methods and modeling for success in their careers and daily lives.

To solve statisticians’ public relations problem of “more respect” they should advertise to the world that they integrate and connect the three circles of statistical practice, research, and education.

2 United Applicable Statistics, Learning Framework for Almost All of the Science of Statistics

As I evaluate my research career I believe that its approach has been to be comprehensive (a fox) rather than concentrate on a narrow problem (a hedgehog). Some recent commentators allege that in predicting and planning the future, and also in innovating new scientific ideas, “big picture” foxes do better than “specialized publication centered” hedgehogs. The concepts “hedgehog” and “fox” are very popular (and were advocated by the British philosopher Isaiah Berlin whom I met when I was a Harvard undergraduate).

An informative game (that I think would be beneficial for the health of the discipline of Statistics to play) is to choose a list of statisticians and debate who is a fox and who is a hedgehog. In my view Ben Kedem is a fox because he has helped innovate new fields of research. Should we be concerned to inspire young statisticians to be foxes by improving career rewards for foxes?

I describe my recent research (driven by intellectual curiosity but I believe ultimately very practical for teaching and applied research) as seeking a unified approach to the discipline and profession of statistics which I call "United Applicable Statistics". It aims to develop theory to practice (1) unified rather than ad hoc problem solving strategies, and (2) reasoning called "analogies between analogies". Details of this approach to statistical thinking are described on my website in my "Last Lecture 1949-2009: Quantiles are Optimal", and in recent papers Parzen (2004), (2008), (2009).

A main tool of my research, which is the focus of this paper, is the theory of mid-P inference, mid-distribution functions, and mid-quantile functions of discrete random variables (which I believe provide a key to unifying discrete and continuous data modeling and to handling ties in data).

I believe that for the important frontier problem of high dimensional data analysis and classification we can develop nonparametric methods that apply our "correlation" and "mid-distribution" unification of traditional nonparametric methods. My personal motto is, "I have answers to statistical questions; what is your problem?"

Our practical goal is to enable simultaneous practice in statistical inference of the approaches of Fisher, Neyman, Tukey, and modern (hierarchical) Bayesians. An important conclusion is that frequentist inference (confidence intervals and hypotheses tests) can be expressed in terms of a "knowledge distribution" for a parameter (while emphasizing that the parameter is an unknown constant and not a random variable with a prior distribution). We describe the frequentist knowledge distribution of a parameter by a confidence quantile (a confidence interval endpoint function) which has same mathematical (algorithmic) properties as a posterior quantile used to describe the Bayesian posterior knowledge distribution of a parameter which is assumed to have a prior (knowledge) distribution.

We claim that unification can be accomplished in practice for the usual introductory probability models because with a conjugate prior distribution:

(1) The Bayesian posterior distribution is identical with the frequentist confidence distribution for "augmented data", and

(2) The confidence distribution for the actually observed data is identical with the Bayesian posterior distribution for an "uninformative flat" prior.

(3) We propose to practice statistical inference by thinking Bayesian and computing frequentist (and when prior knowledge is available comparing its answers with answers not assuming prior knowledge).

A problem often encountered in the practice of statistics is not that we don't have an answer to your question, but that we have too many answers and don't know which ones to choose as our "final" answer. A problem with an extensive literature, and many competing answers, is inference for parameters of discrete data, such as the true population proportion p when one observes K successes in n trials. What may be novel is our claim that for a proportion p of $0-1$ data the mid-P frequentist confidence interval is approximately identical with the Bayesian Jeffrey's prior credible interval.

An important inference method that is unknown (and perhaps difficult to accept) to many statisticians is the "mid-P" approach (usually credited to Lancaster, 1961). This paper presents theory to justify this frequentist approach and argues that it can be recommended as the "final" (benchmark) answer because it is identical with the Bayesian answer for a Jeffrey's Beta(.5,.5) prior for p. While for large samples other popular answers are approximately numerically equivalent, introductory courses will be happier if we teach only one way, the "right" way, the way that is accurate for small samples and zero successes. It is easy to compute from software for the quantile function of the Beta distribution.

3 MID-Distribution, MID-Quantiles

A random variable Y can be described by

(1) Distribution function $F(y) = F(y;Y) = Pr[Y \leq y]$

(2) Quantile function (inverse distribution function)

$$Q(P) = Q(P;Y) = \inf\{y : F(y) \geq P\}, 0 < P < 1$$

(3) stochastic model

$$Y = h(\theta, V)$$

where θ is a parameter and V is a random variable with known distribution $F(y;V)$. When h is increasing function of V, continuous from the left,

$$Q(P;Y|\theta) = h(\theta, Q(P;V))$$

Many applications assume important stochastic model, called location-scale parameter model,

$$Y = \mu + \sigma V, Q(P;Y|\mu,\sigma) = \mu + \sigma Q(P;V).$$

In terms of U =Uniform(0,1) in distribution, one can always represent $Y = Q(U;Y)$ in distribution. Less well known is an important representation, useful for computing conditional quantiles $Q(P;Y|X)$:

$$Y = Q(F(Y)) \text{ with probability 1.}$$

When Y is continuous we define $U = F(Y;Y)$ to be the probability integral transform of Y, satisfying U =Uniform(0,1) in distribution.

When Y is discrete we recommend mid-probability integral transform $U = F\text{mid}(Y;Y)$ where $F\text{mid}(y;Y)$ is mid-distribution function defined in terms of probability mass function $p(y;Y) = Pr[Y = y]$ by

$$F \text{ mid}(y;Y) = F(y;Y) - .5p(y;Y) = \text{mid}Pr[Y \leq y]$$

The mean and variance of $F\text{mid}(Y;Y)$ have formulas with elegant proofs given in Parzen (2004).

DEFINE EXTENDED MEDIAN OF DISCRETE RANDOM VARIABLE: Verify

$$Pr[Y > y] - P[Y < y] = 1 - 2Pr[Y < y] - p(y;Y) = 1 - 2F\text{mid}(y;Y).$$

Median $Q2$ can be defined intuitively as number satisfying $Pr[Y > Q2] = Pr[Y < Q2]$; an equation to compute median is $1 - 2F\text{mid}(Q2;Y) = 0$ which may not have a solution. We therefore define "extended median" $Qm2$ by equation

$$1 - 2Fmidc(Qm2;Y) = 0,$$

where $F\text{midc}(y;Y)$ is continuous version of $F\text{mid}(y;Y)$ defined below. Its inverse, denoted $Q\text{mid}(P;Y)$, is called mid-quantile.

When Y is discrete, $F\text{mid}(y;Y)$ is piecewise constant equal, at points of discontinuity, to the average of the left hand and right hand limits of $F(y;Y)$. I believe that a theoretical justification for statistical practice of mid-probability inference is that inversion formulas for distribution functions from characteristic function, usually stated for $F(y;Y)$ at continuity points y, are true for $F\text{mid}(y;Y)$ for all values of y. Section 4 discusses a proof essentially already in my classic introductory probability textbook Parzen (1960).

The concepts of population median $Q2$ and quartiles $Q1$ and $Q3$ are defined for Y continuous

$$Q2 = Q(.5;Y), Q1 = Q(.25;Y), Q3 = Q(.75;Y)$$

For Y discrete we recommend extended definitions (in terms of the mid-quantile $Q\text{mid}(P;Y)$ to be defined below)

$$Qm2 = Q\text{mid}(.5;Y), Qm1 = Q\text{mid}(.25;Y), Qm3 = Q\text{mid}(.75;Y)$$

The sample distribution function $\tilde{F}(y;Y)$ is discrete; we recommend that its sample quantiles should be defined and computed using the extended definition in terms of sample mid-quantile function $\tilde{Q}\text{mid}\ (P;Y)$.

DEFINITION OF MID-QUANTILE: Define $Q\text{mid}(P;Y)$ to be the continuous function which is the inverse of the continuous version $F\text{midc}(y;Y)$ of the mid-distribution $F\text{mid}(y;Y)$ of the discrete random variable Y. Denote by $y_1 < \ldots, y_r$ the jump points (probable values) of Y. Define

$$P_j = F\text{mid}(y_j;Y), p_j = p(y_j;Y).$$

At P_j define

$$Q\text{mid}(P_j) = y_j$$

For $P < P_1$, define $Q\text{mid}(P;Y) = y_1$. For $P > P_r$, define $Q\text{mid}(P;Y) = y_r$. For $P_j < P < P_{j+1}$, define $Q\text{mid}(P;Y)$ by linear interpolation,

$$Q\text{mid}(P;Y) = y_j + (y_{j+1} - y_j)(P - P_j)/(P_{j+1} - P_j).$$

A very useful identity:

$$P_{j+1} - P_j = .5(p_j + p_{j+1}).$$

Define $F\text{mid}c(y; Y)$ to be inverse of $Q\text{mid}(P; Y)$; it obeys for $y_j < y < y_{j+1}$

$$F\text{mid}c(y; Y) = P_j + (P_{j+1} - P_j)(y - y_j)/(y_{j+1} - y_j),$$

For $y < y_1$, $F\text{mid}c(y; Y) = 0$. For $y > y_r$, $F\text{mid}c(y; Y) = 1$.

Note formula for probability density function for y between y_j and y_{j+1} and P between P_j and P_{j+1}

$$f\text{mid}c(y; Y) = (P_{j+1} - P_j)/(y_{j+1} - y_j) = f\text{mid}c(Q\text{mid}(P; Y); Y).$$

Exploratory data analysis seeks to describe a sample $Y_1, \ldots, Y_n$ of Y. We have following functions:

sample distribution function $\tilde{F}(y; Y)$,

sample quantile $\tilde{Q}(P; Y)$,

sample mid-distribution $\tilde{F}$ mid$(y; Y)$,

sample continuous version of mid-distribution $\tilde{F}\text{mid}c(y; Y)$,

sample mid-quantile $\tilde{Q}\text{mid}(P; Y)$,

sample median(mid) $\tilde{Q}m2$, sample quartiles (mid) $\tilde{Q}m1$ and $\tilde{Q}m3$.

EXAMPLE: A Bernoulli $0 - 1$ random variable Y, with $Pr[Y = 1] = p$, satisfies: $F\text{mid}(0) = .5(1 - p)$, $F\text{mid}(1) = 1 - (p/2)$, $Q\text{mid}(.5) = (.5 - .5(1 - p))/.5((1 - p) + p) = p$.

4 Order Statistics, Example Sample Median (mid)

The values in a sample arranged in nondecreasing order is denoted $Y(j; n)$ and called the order statistics. The sample quantile function $\tilde{Q}(P; Y)$ can be expressed

$$\tilde{Q}(P; Y) = Y(j; n), (j - 1)/n < P \leq j/n$$

Statisticians are not in consensus about how to define sample medians and quartiles. They can be viewed as values at $P = .25, .5, .75$ of a continuous version $\tilde{Q}c(P; Y)$ with possible definitions (we assume all values in sample are distinct, no ties)

$$\begin{aligned}\tilde{Q}5(P;Y) &= Y(nP+.5;n), R \text{ type 5 and Parzen;}\\ \tilde{Q}7(P;Y) &= Y((n-1)P+1;n), R \text{ type } 7, R \text{ default}, S \text{ and Excel;}\\ \tilde{Q}6(P;Y) &= Y((n+1)P;n), R \text{ type 6, Minitab, and SPSS}\end{aligned}$$

defining fractional order statistic $Y(k+r;n) = Y(k;n) + r(Y(k+1;n) - Y(k;n))$

When all values are distinct mid$P$ values are $P_j = (j - .5)/n$, $\tilde{F}\text{mid}(Y(j;n);P) = P_j$, $\tilde{Q}\text{mid}(P_j;Y) = Y(j;n)$, $\tilde{Q}\text{mid}(P;Y) = \tilde{Q}5(P;Y)$.

EXAMPLE: Sample median (of a sample with ties) where extended definition provides an answer different (in value and interpretation) from standard answer which does not take account of ties in the data.

An automobile dealer with 9 new cars available for sale advertises that his cars are "fuel efficient" with average miles per gallon 18 mpg. The actual mpg ratings of his cars are listed: 16, 21, 20, 25, 20, 13, 20, 15, 15. The order statistics $Y(j;n)$ are 13, 15, 15, 16, 20, 20, 20, 21, 25. The usual median $Q2$ equals 20, the middle value. It is very different from our median $\tilde{Q}m2 = 18$, closer to the sample average 18.3. To calculate our median, determine: (1) the distinct values in the sample, 13, 15, 16, 20, 21, 25; (2) mid-probabilities P_1,P_2,P_3,P_4,P_5,P_6 equal 1/18, 4/18, 7/18, 11/18, 15/18, 17/18. Because .5 is the average of P_3 and P_4, we compute $\tilde{Q}m2 = \tilde{Q}\text{mid}(.5;Y)$ as the average of 16 and 20, equal to 18!

5 Mid Distribution Asymptotic Normal Theorems, Characteristic Function Inversion Formulas

Frequentist inference of the population parameter p of 0-1 variable Y with sample proportion $\tilde{p} = K/n$ starts with exact or approximate sampling distribution given p of the discrete random variable $\tilde{p}$. We express this distribution approximately in terms of a continuous random variable Z by using a transformation called a pivot. Our definition of pivot is expressed:

$$T\text{in } (p|\tilde{p}) = (p - \tilde{p})/\sqrt{(p(1-p)/n)}$$

We write $T\text{in}$ to denote that it is an increasing function of p; this can be proved by showing that its derivative with respect to p is positive. The other condition on a pivot is that its distribution does not depend on the parameter p.

Let Z be a generic symbol for a Normal(0,1) random variable. From the Asymptotic Normal Theorem (Central Limit Theorem) we conclude that when p is the true value of the parameter

$$T\text{in } (p|\tilde{p}) = Z$$

in distribution approximately. A usual (but often inaccurate) interpretation is for every y

$$F(y; T\text{in } (p|\tilde{p})) = F(y; Z)$$

approximately.

To obtain an improved approximation of the discrete random variable $T\text{in } (p|\tilde{p})$ by the continuous random variable Z we have a choice of two approaches:

(1) a continuity correction,

(2) an approximation of mid-distribution functions:

$$Fmid(y; T\text{in } (p|\tilde{p})) = F(y; Z)$$

approximately.

My probability textbook Parzen (1960) has a direct geometric proof of the approximate normal distribution of $\tilde{p}$ which demonstrates why the mid-distribution normal approximation is very accurate, and why the continuity correction works to approximate the distribution function.

More research is needed to show why mid-probability approximation is accurate in general; we conjecture it is true because (as we next show) the inversion formula of distribution functions from characteristic functions actually holds for mid-distributions. We state three general theorems about the relation between characteristic functions and mid-distribution functions which we believe will help explain the increased accuracy of the mid-distribution normal approximation.

Parzen (1960) has detailed proofs (of inversion formulas and convergence in distribution of a sequence Z_n to a limit Z) that can be immediately interpreted to provide proofs of the following theorems about mid-distributions.

Convergence in mid-distribution Theorem. If Z_n converges in distribution to Z then at every continuity point y of $F(y;Z)$

$$Fmid(y;Z_n) \text{ converges to } F(y;Z)$$

Proof: in Eq. (5.6), p. 435 replace $F(b;Z_n)$ by $Fmid(b;Z_n)$

Extended Inversion Formulas. In statement of Theorem 3A on p. 401 replace $F(b;X)-F(a;X)$ by $Fmid(b;X)-Fmid(a;X)$ and drop condition that a and b are continuity points. In statement of eq. (3.12) on p. 402 replace $F(x;X)$ by $Fmid(x;X)$ and drop condition that x is a continuity point.

Proof of extended inversion formula Extended Theorem 3A follows from eq. (3.8) by replacing $F(b;X)-F(a;X)$ by $Fmid(b;X)-Fmid(a;X)$. The proof of extended (3.16) follows from the equation following eq. (5.17) on p. 412 and the fact

$$Pr[X>x]-Pr[X<x]=1-2Fmid(x;X).$$

More research is required on Berry-Esseen theorems for mid-distributions, and on Bahadur representations for sample mid-quantiles. The asymptotic normal distribution of the sample mid-quantiles is studied for both continuous and discrete data in the paper Ma, Genton, Parzen (2009).

DIRECT PROOFS WITH ERROR BOUND OF ASYMPTOTIC NORMALITY OF MID-DISTRIBUTION OF BINOMIAL, HYPERGEOMETRIC, POISSON: Let K be integer-valued random variable, and Z Normal(0,1).

STARTING LEMMA: "probability mass function lemma". By interpreting usual calculations show that for a suitable constant c and large values of $\sigma[K]$, the probability mass function $p(k)$ of K satisfies, for $y=(k-E[K])/\sigma[K]$,

$$\sigma[K]p(E[K]+y\sigma[K])/f(y;Z)=1+|y|^3c/\sigma[K].$$

Therefore conclude that for almost all y (not a probable value)

$$fmidc(y;(K-E[K])/\sigma[K])/f(y;Z)=1+|y|^3c/\sigma[K].$$

THEOREM: A Berry-Esseen type supremum bound for all y and k

$$|Fmidc(y;(K-E[K])/\sigma[K])-F(y;Z)| \leq c/\sigma[K]$$

$$|Fmid(k;K)-F(k;E[K]+\sigma[X]Z)| \leq c/\sigma[K].$$

We believe that the foregoing facts (proved separately for each model) can be applied in practice as an example of unifying analogous problems by "analogies between analogies". A proof for hypergeometric probability of the "probability mass function lemma" is contained in Lahiri and Chatterjee (2007). For binomial and Poisson proof is outlined in Parzen (1960), p. 242, eq. (2.2). Guidelines for how small can be value of $\sigma[K]$ for accurate normal approximation should be numerically determined by comparing true values of mid-distribution with approximate normal values.

6 Confidence Quantile, Inverse of MID-PVALUE

Television ads often tell us that only one doctor out of ten does not prefer an advertised product. We teach our students that to interpret this information we should ask "What is the 95% confidence interval for true proportion p of all doctors who do not prefer the advertised product?" This is a typical (analogous) problem of applied statistics.

We observe a sample of size n (here $n = 10$) of a 0-1 variable Y with true population probability $p = Pr[Y = 1]$, and K values 1 in the sample (here $K = 1$). Sample probability $\tilde{p} = K/n$ is an estimator of p which is regarded as an unknown constant. The numerical value of K is denoted Kobs, which yields a numerical value $\tilde{p}$obs=Kobs$/n$ for the random variable $\tilde{p}$. We desire an interval estimator of the parameter p. We obtain this from a formula for the quantile of a probability distribution for our knowledge of p given the observed data.

Modern Bayesian inference using conjugate priors assumes a Beta prior distribution Beta(a,b) for p, and reports a Beta posterior distribution Beta(a^*,b^*) for p, with hyper-parameter update formulas $a^* = a+K$ and $b^* = b+n-K$. We call a and b hyper-parameters whose "update formulas" are central tools of Bayesian inference with conjugate priors.

Uninformative prior or Jeffrey's prior assumes $a = b = .5$. To describe posterior distribution of p we recommend posterior quantile

$$Q(P;p|\tilde{p},\text{ Jeffrey's prior}) = Q(P;\text{ Beta}(K+.5,n-K+.5)).$$

because it provides the most convenient way to describe a 95% credible interval for the parameter p:

$$Q(.025;\ \text{Beta}(K+.5, n-K+.5)) < p < Q(975;\ \text{Beta}(K+.5, n-K+.5)).$$

By introducing the concept of "endpoint function" of frequentist confidence interval, and denoting it $Q(P; p|\tilde{p}\text{obs})$, $0 < P < 1$, we can express the 95% level confidence interval in a similar form

$$Q(.025; p|\tilde{p}\text{obs}) < p < Q(.975; p|\tilde{p}\text{obs})).$$

To use the concept of endpoint function we have to answer three questions:
how to define it;
how to compute it;
how to interpret it.

We define confidence interval endpoint function below in terms of mid-probability as the inverse of MID-PVALUE which is an increasing function of p by a stochastic order condition; we compute it by the quasi-exact (approximately accurate) formula

$$Q(P; p|\tilde{p}\text{obs}) = Q(P; \text{Beta}(K+.5, n-K+.5));$$

we interpret it in the same way that we interpret a Bayesian credible interval.

CONFIDENCE QUANTILE: We call $Q(P; p|\tilde{p}\text{obs})$ a confidence quantile; it has the same mathematical properties as a posterior quantile, and is the quantile function of the confidence distribution of the random variable $p|\tilde{p}\text{obs}$, representing our uncertain knowledge (given the data) of the unknown constant p.

MID PVALUE AND STOCHASTIC ODER CONDITION: The frequentist definition of confidence quantile starts with the concept of Mid-PVALUE function of the parameter p, given $\tilde{p}$obs:

$$\begin{aligned} Mid - \text{PVALUE}(p; \tilde{p}\text{obs}) &= MidPr[K \geq K\text{obs}|p] = \\ &= 1 - Fmid(K\text{obs}; K|p) \end{aligned}$$

We require FUNDAMENTAL STOCHASTIC ORDER ASSUMPTION: Mid-PVALUE is assumed to be an increasing function of p, for fixed value of Kobs.

The confidence quantile $Q(P; p|\tilde{p}\text{obs})$ is defined to be the INVERSE Mid-PVALUE, the inverse function of the Mid-PVALUE function of p. For fixed P, $Q(P; p|\tilde{p}\text{obs})$ is the value of p such that $P =$Mid-PVALUE$(p; \tilde{p}\text{obs})$, and satisfies estimating equation

$$Fmid(K\text{obs}; K|Q(P; p|\tilde{p}\text{obs})) = 1 - P.$$

One can numerically compute and plot the solution of this equation by plotting, for $0 < p < 1$, $(Fmid(K\text{obs};\ K|p), p)$.

THEOREM: An analytic formula for the confidence quantile of the parameter p is obtained by "proving" the quasi-exact approximate formula

$$MidPr[\text{Binomial}(n, p) \geq k] = Pr[\text{Beta}(k + .5, n - k + .5) \leq p]$$

Proof: Justify by proving the inequalities (noted by Leonard (1999, p. 136))

$$\begin{aligned} Pr[\text{Binomial}(n, p) \geq k] &\leq Pr[\text{Beta}(k + .5, n - k + .5) \leq p] \\ &\leq Pr[\text{Binomial}(n, p) \geq (k + 1)]. \end{aligned}$$

Note the well known identity between the Binomial and Beta distributions:

$$Pr[\text{Binomial}(n, p) \geq k] = Pr[\text{Beta}(k, n - k + 1) \leq p]$$

7 Confidence Quantiles of Parameters p, logodds(p)

We conclude our discussion of confidence quantiles with formulas for confidence quantiles of p and logodds (p) that we believe deserve to be widely practiced in applied statistics (the Statistics of Science).

Theorem A: Frequentist confidence quantile of parameter p is quasi-identical with Bayesian Jeffrey's posterior quantile

$$\begin{aligned} P &= Pr[\text{Beta}(K\text{obs} + .5, n - K\text{obs} + .5) \leq p = Q(P; p|\tilde{p}\text{obs})] \\ Q(P; p|\tilde{p}\text{obs}) &= Q(P; \text{Beta}(K\text{obs} + .5, n - K\text{obs} + .5)). \end{aligned}$$

Define $a^* = K\text{obs}+.5, b^* = n - K\text{obs} +.5, n^* = a^* + b^*$, $p^* = a^*/n^*$, $n^{**} = n^*p^*(1 - p^*)$. Note $1/n^{**} = (1/a^*) + (1/b^*)$.

Theorem B: Confidence quantile of parameter logodds(p)

$$Q(P; \text{logodds}(p)|\tilde{p}\text{obs}) = \text{logodds}(p^*) + Q(P; \log F(2a^*, 2b^*))$$

Theorem C: Novel exact normal approximation to confidence quantile for logodds(p)

$$Q(P; \log F(2a^*, 2b^*)) = (-1/3a^*) + (1/3b^*) + \sqrt{(1/n^{**})}Q(P; Z)$$

The problem of Normal approximations for a log F random variable has an extensive literature outlined in the book by Kendall and Stuart.

Theorem D: Wilson (1927) or Score confidence interval for p has endpoint function $Q(P; p|\tilde{p}\text{obs}, n)$ which can be computed from estimating equation in terms of approximately normal increasing pivot $T\text{in}(p; \tilde{p})$:

$$T\text{in}(Q(P; p|\tilde{p}\text{obs}, n); \tilde{p}\text{obs}) = Q(P; Z)$$

We regard Wilson confidence quantile as a large sample approximation to Mid$P$/Beta confidence quantile.

EXAMPLES p CONFIDENCE INTERVALS: $n = 10$, $K = 1$; $n = 5$, $K = 4$; $n = 100$, $K = 10$. When we observe that $K = 1$ doctors out of $n = 10$ do not favor a product, we seek a 95% confidence interval for p, the population proportion of doctors favoring the product. We recommend the mid-P confidence interval (equivalently Bayesian credible interval with Jeffrey's prior), whose endpoints are computed from confidence quantile $Q(P;$Beta(1.5,9.5)). The following table compares endpoints of intervals computed by various popular formulas.

Recommended Mid$P$/Beta	WilsonScore	ExactClopper	AgrestiCoull	Wald
Lower .025 endpoint .011	.018	.0025	-.004	-.086
Upper .975 endpoint .38	.404	.445	.426	.285

Quality control engineers seek 95% confidence intervals for the true probability p of completing a task when in a small sample of $n = 5$ one observes an 80% completion rate ($K = 4$). The Mid$P$ confidence quantile is $Q(P;$Beta(1.5,4.5)).

Recommended Mid$P$/Beta	WilsonScore	ExactClopper	AgrestiCoull	Wald
Lower .025 endpoint .321	.376	.284	.365	.449
Upper .975 endpoint .971	.964	.995	.983	1.00

We conclude with the endpoints of the 95% confidence interval for p when $n = 100$, $K = 10$. The mid$P$ confidence quantile of the parameter p is $Q(P;\text{Beta}(10.5,90.5))$.

Recommended Mid$P$/Beta	WilsonScore	ExactClopper	AgrestiCoull	Wald
Lower .025 endpoint .053	.055	.049	.054	.041
Upper .975 endpoint .170	.174	.176	.176	.159

References

Agresti, Alan and Gottard, Anna. (2007). Reducing conservatism of exact small sample methods of inference for discrete data. *Computational Statistics and Data Analysis*, 51, 6447–6458.

Lahiri, S. N. and Chatterjjee, A (2007). A Berry-Esseen theorem for hypergeometric probabilities under minimal conditions. *Proceedings of the American Mathematical Society*, 133, 1335–1345.

Lancaster, H. O. (1961). Significance tests for discrete distributions. *Journal of the American Statistical Association*, 58, 223–234.

Leonard, T. and Hsu, John S. J. (1999). *Bayesian Methods.* Cambridge University Press.

Koenker, Roger. (2005). *Quantile Regression.* Cambridge University Press.

Ma, Y., Genton, M. G., Parzen, E. (2009). Asymptotic properties of sample quantiles of discrete distributions. *Annals of the Institute of Statistical Mathematics.*

Parzen, E. (1960). *Modern Probability Theory and its Applications.* Wiley: New York.

Parzen, E. (1977). *Nonparametric statistical data science: A unified approach based on density estimation and testing for white noise.* Technical report. Statistical Science Division. SUNY at Buffalo.

Parzen, E. (1979). Nonparametric statistical data modeling. *Journal American Statistical*

Association, 74, 105–131.

Parzen, E. (2004). Quantile probability and statistical data modeling. *Statistical Science*, 19, 652–662.

Parzen, E. (2008). United statistics, confidence quantiles, Bayesian statistics. *Journal Statistical Planning and Inference*, 138, 2777–2785.

Parzen, E. (2009). Quantiles, conditional quantiles, confidence quantiles for p, logodds(p). *Communications in Statistics: Theory and Methods.*

Wilson, E. B. (1927). Probable inference, the Law of Succession, and statistical inference. *Journal of the American Statistical Association*, 22, 209–212.

Linear and Loglinear Poisson Autoregression

K. Fokianos
University of Cyprus

Joint work with A. Rahbek and D. Tjøstheim

Part of this work is supported by Cyprus Research Foundation

- Ben Kedem influenced my research on this area by suggesting to work for my Ph.D. on categorical time series.

- Ben Kedem influenced my research on this area by suggesting to work for my Ph.D. on categorical time series.
- His initial contribution to the subject matter is Slud & Kedem (1994), Statistica Sinica.

- Ben Kedem influenced my research on this area by suggesting to work for my Ph.D. on categorical time series.
- His initial contribution to the subject matter is Slud & Kedem (1994), Statistica Sinica.
- Worked with him around 10 years–quite an experience.

- Ben Kedem influenced my research on this area by suggesting to work for my Ph.D. on categorical time series.
- His initial contribution to the subject matter is Slud & Kedem (1994), Statistica Sinica.
- Worked with him around 10 years–quite an experience.
- Published several results for regression models for time series, including the book Kedem and Fokianos (2002), Wiley.

- Ben Kedem influenced my research on this area by suggesting to work for my Ph.D. on categorical time series.
- His initial contribution to the subject matter is Slud & Kedem (1994), Statistica Sinica.
- Worked with him around 10 years–quite an experience.
- Published several results for regression models for time series, including the book Kedem and Fokianos (2002), Wiley.
- He also influenced my research on semi-parametrics (with J. Qin).

- Ben Kedem influenced my research on this area by suggesting to work for my Ph.D. on categorical time series.
- His initial contribution to the subject matter is Slud & Kedem (1994), Statistica Sinica.
- Worked with him around 10 years–quite an experience.
- Published several results for regression models for time series, including the book Kedem and Fokianos (2002), Wiley.
- He also influenced my research on semi-parametrics (with J. Qin).

Table of contents

Introduction

Motivation

Ergodicity Results for the Linear Model

Log–Linear Models for Time Series of Counts

Likelihood Inference

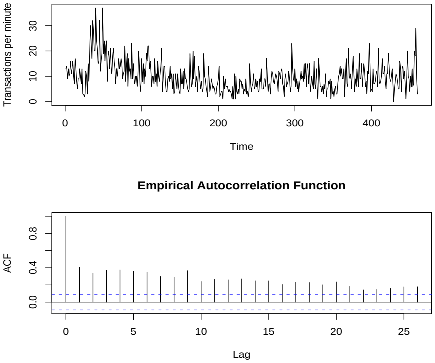
Inference for the Log-Linear Model

Simulations

Transactions Data

Transactions Data

Number of transactions per minute for the stock Ericsson B during July 2nd, 2002. The bottom plot shows their autocorrelation function.



Previous Work 1

Suppose that $\{Y_t, t = 1, 2, \ldots, n\}$ is a count time series and let $\mathcal{F}_t^{Y,\lambda}$ stands for the σ–field generated by $\{Y_0, \ldots, Y_t, \lambda_0\}$. Rydberg and Shephard (2000) and Streett (2000) have studied the following linear model

$$\begin{aligned} Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda} &\sim \text{Poisson}(\lambda_t), \\ \lambda_t &= d + a\lambda_{t-1} + bY_{t-1}, \end{aligned} \tag{1}$$

for $t \geq 1$ and the parameters d, a, b are assumed to be positive. In addition assume that λ_0 and Y_0 are fixed.

Previous Wok 2

1. For the Poisson distribution, the conditional mean is equal to the conditional variance, that is

$$\mathrm{E}[Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \mathrm{Var}[Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \lambda_t.$$

Previous Wok 2

1. For the Poisson distribution, the conditional mean is equal to the conditional variance, that is

$$\mathrm{E}[Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \mathrm{Var}[Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \lambda_t .$$

2. It is tempting to call (1) an INGARCH(1,1)–that is integer GARCH model.

Previous Wok 2

1. For the Poisson distribution, the conditional mean is equal to the conditional variance, that is

$$\mathrm{E}[Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \mathrm{Var}[Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \lambda_t.$$

2. It is tempting to call (1) an INGARCH(1,1)–that is integer GARCH model.
3. Proposed modeling is based on the evolution of the mean of the Poisson instead of its variance.

Previous Work 3

1. Second order properties of model (1) have been studied by Rydberg and Shephard (2000).

Previous Work 3

1. Second order properties of model (1) have been studied by Rydberg and Shephard (2000).
2. Streett (2000) shows existence and uniqueness of a stationary distribution.

Previous Work 3

1. Second order properties of model (1) have been studied by Rydberg and Shephard (2000).
2. Streett (2000) shows existence and uniqueness of a stationary distribution.
3. Ferland et al (2006) consider the general INGARCH(p,q)

$$\begin{aligned} Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda} &\sim \text{Poisson}(\lambda_t), \\ \lambda_t &= d + \sum_{i=1}^{p} a_i \lambda_{t-i} + \sum_{j=1}^{q} b_j Y_{t-j}, \end{aligned}$$

   and show second order stationarity provided that

$$0 < \sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_j < 1.$$

Previous Work 4

1. It can be shown that

$$\mathrm{E}\left[Y_t\right] = \mathrm{E}[\lambda_t] \equiv \mu = d/(1 - a - b)$$

Previous Work 4

1. It can be shown that

$$\mathrm{E}\left[Y_t\right] = \mathrm{E}[\lambda_t] \equiv \mu = d/(1 - a - b)$$

2. The autocovariance function of Y_t is

$$\mathrm{Cov}\left[Y_t, Y_{t+h}\right] = \begin{cases} \dfrac{(1-(a+b)^2+b^2)\mu}{1-(a+b)^2}, & h = 0, \\[2ex] \dfrac{b(1-a(a+b))(a+b)^{h-1}\mu}{1-(a+b)^2}, & h \geq 1. \end{cases}$$

Previous Work 5

1. All moments of model (1) are finite if and only if $0 \leq a + b < 1$.

Previous Work 5

1. All moments of model (1) are finite if and only if $0 \leq a + b < 1$.
2. Notice that

$$\mathrm{Var}[Y_t] = \mu \left(1 + \frac{b^2}{1 - (a+b)^2} \right).$$

Therefore $\mathrm{Var}[Y_t] \geq \mathrm{E}[Y_t]$ with equality when $b = 0$. This is a case of overdispersion.

Previous Work 5

1. Model (1) is related to the theory of generalized linear models for time series, see Ch. 1 and 4 of Kedem and Fokianos (2002). They fall under the framework of observation driven models, Cox (1981).

Previous Work 5

1. Model (1) is related to the theory of generalized linear models for time series, see Ch. 1 and 4 of Kedem and Fokianos (2002). They fall under the framework of observation driven models, Cox (1981).
2. Observation driven models for time series of counts have been studied by several authors including Zeger and Qaqish (1988), Li (1994), Brumback et al (2000), Fahrmeir and Tutz (1994), Benjamin et al (2003), Davis et al (2003) and Jung et al (2006).

Previous Work 5

1. Model (1) is related to the theory of generalized linear models for time series, see Ch. 1 and 4 of Kedem and Fokianos (2002). They fall under the framework of observation driven models, Cox (1981).
2. Observation driven models for time series of counts have been studied by several authors including Zeger and Qaqish (1988), Li (1994), Brumback et al (2000), Fahrmeir and Tutz (1994), Benjamin et al (2003), Davis et al (2003) and Jung et al (2006).
3. A log–linear model for the mean of the observed process is usually assumed and its structure is composed by past values of the response, moving average terms and other explanatory variables.

Previous Work 5

1. Model (1) is related to the theory of generalized linear models for time series, see Ch. 1 and 4 of Kedem and Fokianos (2002). They fall under the framework of observation driven models, Cox (1981).
2. Observation driven models for time series of counts have been studied by several authors including Zeger and Qaqish (1988), Li (1994), Brumback et al (2000), Fahrmeir and Tutz (1994), Benjamin et al (2003), Davis et al (2003) and Jung et al (2006).
3. A log–linear model for the mean of the observed process is usually assumed and its structure is composed by past values of the response, moving average terms and other explanatory variables.
4. Davis et al (2003) considers a simple but important case of a log-linear model and shows ergodicity of the observed process.

Our Contribution

(1) Show geometric ergodicity for INGARCH(1,1) models.

Our Contribution

(1) Show geometric ergodicity for INGARCH(1,1) models.

(2) Study likelihood inference for INGARCH(1,1) models.

Our Contribution

(1) Show geometric ergodicity for INGARCH(1,1) models.

(2) Study likelihood inference for INGARCH(1,1) models.

(3) Study geometric ergodicity for a log–linear model to be discussed next.

Our Contribution

(1) Show geometric ergodicity for INGARCH(1,1) models.

(2) Study likelihood inference for INGARCH(1,1) models.

(3) Study geometric ergodicity for a log–linear model to be discussed next.

(4) Study likelihood inference for a log–linear models.

Our Contribution

(1) Show geometric ergodicity for INGARCH(1,1) models.

(2) Study likelihood inference for INGARCH(1,1) models.

(3) Study geometric ergodicity for a log–linear model to be discussed next.

(4) Study likelihood inference for a log–linear models.

(5) Apply all models to the transactions data.

Rephrasing the model

To study model (1) for each time point t, introduce a Poisson process $N_t(\cdot)$ of unit intensity. Then, we can assume that Y_t is equal to the number of events $N_t(\lambda_t)$ of $N_t(\cdot)$ in the time interval $[0, \lambda_t]$. Let therefore $\{N_t(\cdot), t = 1, 2, \ldots\}$ be a sequence of independent Poisson process of unit intensity and rephrase (1) as

$$Y_t = N_t(\lambda_t), \quad \lambda_t = d + a\lambda_{t-1} + bY_{t-1}, \tag{2}$$

for $t \geq 1$ and with Y_0, λ_0 fixed.

Perturbed Linear Models 1

We resort to the perturbed chain (Y_t^m, λ_t^m) defined by

$$Y_t^m = N_t\left(\lambda_t^m\right), \; \lambda_t^m = d + a\lambda_{t-1}^m + bY_{t-1}^m + \varepsilon_{t,m}, \tag{3}$$

with λ_0^m, Y_0^m fixed, and

$$\varepsilon_{t,m} = c_m 1\left(Y_{t-1}^m = 1\right) U_t, \quad c_m > 0\,, \quad c_m \to 0, \quad \text{as} \quad m \to \infty,$$

where $1(\cdot)$ is the indicator function, and where $\{U_t\}$ is a sequence of iid uniform random variables on $(0, 1)$ and such that the $\{U_t\}$ is independent of $\{N_t(\cdot)\}$.

Perturbed Linear Models 2

1. The perturbation in (3) is a purely auxiliary device to obtain ϕ–irreducibility.

Perturbed Linear Models 2

1. The perturbation in (3) is a purely auxiliary device to obtain ϕ–irreducibility.
2. The perturbation can be introduced in many other ways. For, instance, it is enough to set $\{U_t\}$ to be an i.i.d sequence of positive random variables with bounded support possessing density on the positive real axis with respect to the Lebesgue measure and finite fourth moment.

Perturbed Linear Models 2

1. The perturbation in (3) is a purely auxiliary device to obtain ϕ–irreducibility.
2. The perturbation can be introduced in many other ways. For, instance, it is enough to set $\{U_t\}$ to be an i.i.d sequence of positive random variables with bounded support possessing density on the positive real axis with respect to the Lebesgue measure and finite fourth moment.
3. The form of the likelihood functions for $\{Y_t\}$ and $\{Y_t^m\}$ as far as dependence on $\{\lambda_t\}$ is concerned will be the same for both models (2) and (3).

Linear Model 3

Proposition

Consider model (3) and suppose that $0 < a + b < 1$. *Then the process* $\{\lambda_t^m, t \geq 0\}$ *is a geometrically ergodic Markov chain with finite moments of order* k*, for an arbitrary* k*.*

We show that

- $\{\lambda_t^m, t \geq 0\}$ is aperiodic and ϕ–irreducible.

Linear Model 3

Proposition

Consider model (3) and suppose that $0 < a + b < 1$*. Then the process* $\{\lambda_t^m, t \geq 0\}$ *is a geometrically ergodic Markov chain with finite moments of order* k*, for an arbitrary* k*.*

We show that

- $\{\lambda_t^m, t \geq 0\}$ is aperiodic and ϕ–irreducible.
- There exists a small set C and a test function $V(\cdot)$ which satisfies

$$E[V(\lambda_t^m)|\lambda_{t-1}^m = \lambda] \leq (1 - k_1)V(\lambda) + k_2 1(\lambda \in C)$$

  for some constants k_1, k_2 such that $0 < k_1 < 1$, $0 < k_2 < \infty$.

Linear Model 4

Proposition

Consider model (3) and suppose that the conditions of Proposition 3.1 hold. Then the process $\{(Y_t^m, \lambda_t^m, U_t), t \geq 0\}$ is a $V_{(Y,U,\lambda)}$–geometrically ergodic Markov chain with $V_{Y,U,\lambda}(Y, U, \lambda) = 1 + Y^k + \lambda^k + U^k$.

Use the method of Meitz and Saikokonen (2008) to show that geometric ergodicity of the $\{\lambda_t^m\}$ process implies geometric ergodicity of the chain $\{(Y_t^m, U_t, \lambda_t^m)\}$.

Linear Model 5

However, the following holds:

Lemma

Suppose that (Y_t, λ_t) *and* (Y_t^m, λ_t^m) *are defined by (2) and (3) respectively. If* $0 \leq a + b < 1$*, then the following statements hold:*

1. $|E(\lambda_t^m - \lambda_t)| = |E(Y_t^m - Y_t)| \leq \delta_{1,m}$,

Linear Model 5

However, the following holds:

Lemma

Suppose that (Y_t, λ_t) *and* (Y_t^m, λ_t^m) *are defined by (2) and (3) respectively. If* $0 \leq a + b < 1$*, then the following statements hold:*

1. $|E(\lambda_t^m - \lambda_t)| = |E(Y_t^m - Y_t)| \leq \delta_{1,m}$,
2. $E(\lambda_t^m - \lambda_t)^2 \leq \delta_{2,m}$,

Linear Model 5

However, the following holds:

Lemma

Suppose that (Y_t, λ_t) *and* (Y_t^m, λ_t^m) *are defined by (2) and (3) respectively. If* $0 \leq a + b < 1$*, then the following statements hold:*

1. $|E(\lambda_t^m - \lambda_t)| = |E(Y_t^m - Y_t)| \leq \delta_{1,m}$,
2. $E(\lambda_t^m - \lambda_t)^2 \leq \delta_{2,m}$,
3. $E(Y_t^m - Y_t)^2 \leq \delta_{3,m}$,

where $\delta_{i,m} \to 0$ *as* $m \to \infty$ *for* $i = 1, 2, 3$.

Linear Model 5

However, the following holds:

Lemma

Suppose that (Y_t, λ_t) *and* (Y_t^m, λ_t^m) *are defined by (2) and (3) respectively. If* $0 \le a + b < 1$*, then the following statements hold:*

1. $|E(\lambda_t^m - \lambda_t)| = |E(Y_t^m - Y_t)| \le \delta_{1,m}$,
2. $E(\lambda_t^m - \lambda_t)^2 \le \delta_{2,m}$,
3. $E(Y_t^m - Y_t)^2 \le \delta_{3,m}$,

where $\delta_{i,m} \to 0$ *as* $m \to \infty$ *for* $i = 1, 2, 3$*. Furthermore, almost surely, with m large enough*

$$\left|\lambda_t^m - \lambda_t\right| \le \delta \text{ and } \left|Y_t^m - Y_t\right| \le \delta, \quad \text{for any } \delta > 0.$$

Log–Linear Model 1

Suppose again that Y_t is a time a series of counts and set

$$\nu_t = \log \lambda_t.$$

We study the following family of log-linear models

$$Y_t \mid \mathcal{F}_{t-1}^{Y,\nu} \sim \text{Poisson}(\nu_t), \quad \nu_t = d + a\nu_{t-1} + b \log(Y_{t-1} + 1), \tag{4}$$

for $t \geq 1$.

Log–Linear Model 2

- We choose to work with a log-linear model which includes an one-to-one transformation of the data.

Log–Linear Model 2

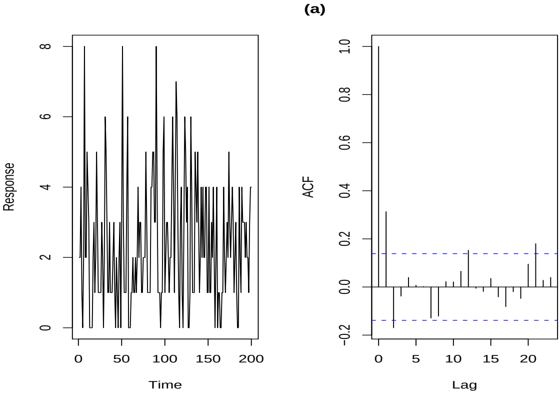
- We choose to work with a log-linear model which includes an one-to-one transformation of the data.
- Each datum is increased by one unit. Hence, we avoid zero data values.

Log–Linear Model 2

- We choose to work with a log-linear model which includes an one-to-one transformation of the data.
- Each datum is increased by one unit. Hence, we avoid zero data values. In this sense both λ_t and Y_{t-1} are transformed into the same scale.
- In addition, we note that covariates can be accommodated by model (2) by including them in the second equation.

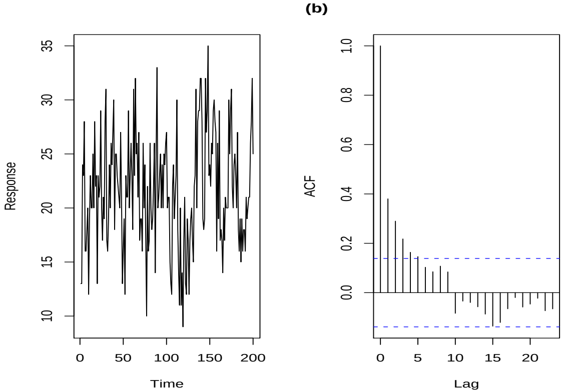
Log–Linear Model 3

Two hundred realizations and their sample autocorrelation function from model (4 for different parameter values. $d = 0.5$,$a = -0.5$ and $b = 2/3$.



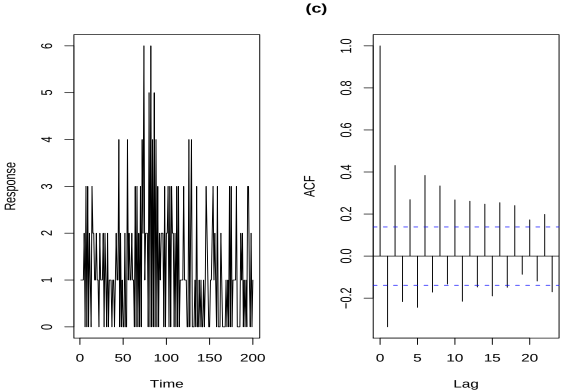
Log–Linear Model 4

Two hundred realizations and their sample autocorrelation function from model (4 for different parameter values. $d = 0.5$, $a = 0.5$ and $b = 1/3$.



Log–Linear Model 5

Two hundred realizations and their sample autocorrelation function from model (4 for different parameter values. $d = 0.5$, $a = -3/4$, and $b = -3/8$.



Log–Linear Model 6

The perturbed chain (Y_t^m, ν_t^m) is defined by

$$Y_t^m = N_t(\lambda_t^m) = N_t(\exp(\nu_t^m)) \quad \nu_t^m = d + a\nu_{t-1}^m + b\log(Y_{t-1}^m + 1) + \varepsilon_{t,m}, \tag{5}$$

with ν_0^m, Y_0^m fixed, and

$$\varepsilon_{t,m} = c_m 1\left(Y_{t-1}^m = 1\right) U_t, \quad c_m > 0\,, \quad c_m \to 0, \quad \text{as} \quad m \to \infty,$$

Log–Linear Model 7

Proposition

Assume model (3) and suppose that $\mid a \mid < 1$. In addition, assume that when $b > 0$ then $|a+b| < 1$, and when $b < 0$ then $|a||a+b| < 1$. Then, the following conclusions hold:

1. The process $\{\nu_t^m, t \geq 0\}$ is a geometrically ergodic Markov chain with finite moments of order k, for an arbitrary k.
2. The process $\{(Y_t^m, U_t, \nu_t^m), t \geq 0\}$ is a $V_{(Y,U,\nu)}$–geometrically ergodic Markov chain with $V_{Y,U,\lambda}(Y, U, \nu) = 1 + \log^{2k}(1+Y) + \nu^{2k} + U^{2k}$, k being a positive integer.

Log–Linear Model 8

It can be also proved that the difference between (4) and (5) is negligible, as $m \to \infty$ such that $c_m \to 0$. This fact is proved under the conditions that

$$|a + b| < 1,$$

if a and b have the same sign, and

$$a^2 + b^2 < 1$$

if they have different signs. These conditions are quite restrictive when compared to the conditions for geometric ergodicity. It is likely that they can be weakened to at least $|a + b| < 1$ for all possible cases of signs and possibly to the generality of the ergodicity conditions. In many applications it seems that $a > 0$ and $b > 0$ in which case, of course, the above condition is the same as the ergodic one, that is $|a + b| < 1$

Inference for the Linear Model 1

The log–likelihood function is given up to a constant, by

$$l_n(\theta) = \sum_{t=1}^{n} l_t(\theta) = \sum_{t=1}^{n} \left(y_t \log \lambda_t(\theta) - \lambda_t(\theta) \right), \tag{6}$$

and the score function is defined by

$$S_n(\theta) = \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\theta)} - 1 \right) \frac{\partial \lambda_t(\theta)}{\partial \theta}, \tag{7}$$

where $\partial \lambda_t(\theta)/\partial \theta$ is a three-dimensional vector with components given by

$$\begin{aligned} \frac{\partial \lambda_t}{\partial d} &= 1 + a\frac{\partial \lambda_{t-1}}{\partial d}, \quad \frac{\partial \lambda_t}{\partial a} = \lambda_{t-1} + a\frac{\partial \lambda_{t-1}}{\partial a}, \\ \frac{\partial \lambda_t}{\partial b} &= Y_{t-1} + a\frac{\partial \lambda_{t-1}}{\partial b}. \end{aligned} \tag{8}$$

Inference for the Linear Model 2

The Hessian matrix is

$$
\begin{aligned}
H_n(\theta) &= \sum_{t=1}^{n} \frac{Y_t}{\lambda_t^2(\theta)} \left(\frac{\partial \lambda_t(\theta)}{\partial \theta}\right) \left(\frac{\partial \lambda_t(\theta)}{\partial \theta}\right)' \\
&- \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\theta)} - 1\right) \frac{\partial^2 \lambda_t(\theta)}{\partial \theta \partial \theta'}. \qquad (9)
\end{aligned}
$$

Inference for the Linear Model 3

- We do not know precisely what conditions guarantee ergodicity of (2).

Inference for the Linear Model 3

- We do not know precisely what conditions guarantee ergodicity of (2).
- However, the assumptions of Proposition 3.2 guarantee geometric ergodicity of the perturbed model (Y_t^m, λ_t^m).

Inference for the Linear Model 3

- We do not know precisely what conditions guarantee ergodicity of (2).
- However, the assumptions of Proposition 3.2 guarantee geometric ergodicity of the perturbed model (Y_t^m, λ_t^m).
- In addition, Lemma 1 shows that λ_t^m approaches λ_t, for large *m*.

It is rather natural to use the ergodic properties of the perturbed process (Y_t^m, λ_t^m) to study the asymptotic properties of the maximum likelihood estimators analogous to (7) and then use Lemma 1.

Inference for the Linear Model 3

Define the counterparts of expressions (6)-(9) for model (3).
The log likelihood function is given up to a constant by

$$l_n^m(\theta) = \sum_{t=1}^{n} \big(y_t \log \lambda_t^m(\theta) - \lambda_t^m(\theta)\big) + \sum_{t=1}^{n} \log f_u(u_t), \tag{10}$$

whereas the score function is equal to

$$S_n^m(\theta) = \sum_{t=1}^{n} \left(\frac{Y_t^m}{\lambda_t^m(\theta)} - 1 \right) \frac{\partial \lambda_t^m(\theta)}{\partial \theta}, \tag{11}$$

and is seen to have exactly the same form as (7), but with $\lambda_t(\theta)$ replaced by $\lambda_t^m(\theta)$.

Inference for the Linear Model 4

Similarly,

$$\begin{aligned} H_n^m(\theta) &= \sum_{t=1}^{n} \frac{Y_t^m}{(\lambda_t^m(\theta))^2} \left(\frac{\partial \lambda_t^m(\theta)}{\partial \theta} \right) \left(\frac{\partial \lambda_t^m(\theta)}{\partial \theta} \right)' \\ &- \sum_{t=1}^{n} \left(\frac{Y_t^m}{\lambda_t^m(\theta)} - 1 \right) \frac{\partial^2 \lambda_t^m(\theta)}{\partial \theta \partial \theta'}. \end{aligned} \tag{12}$$

Inference for the Linear Model 5

To study the asymptotic properties of the maximum likelihood estimator $\hat{\theta}$, for the linear model (2) we derive and use the asymptotic properties of the maximum likelihood estimator $\hat{\theta}^m$ for the perturbed linear model (3).

Proposition

(Prop. 6.3.9 of Brockwell and Davis (1991)) Let $\mathbf{X}_n$, $n = 1, 2, \ldots$ *and* $\mathbf{Y}_{nm}$, $m = 1, 2, \ldots$, $n = 1, 2, \ldots$ *be random* k*-vectors such that*

1. $\mathbf{Y}_{nm} \xrightarrow{D} \mathbf{Y}_m$, *as* $n \to \infty$, *for each* $m = 1, 2, \ldots,$
2. $\mathbf{Y}_m \xrightarrow{D} \mathbf{Y}$, *as* $m \to \infty$, *and*
3. $\lim_{m\to\infty} \limsup_{n\to\infty} P[|\mathbf{X}_n - \mathbf{Y}_{nm}| > \epsilon] = 0$, *for every* $\epsilon > 0$.

Then

$$\mathbf{X}_n \xrightarrow{D} \mathbf{Y} \ \ as \ \ n \to \infty.$$

Inference for the Linear Model 6

Theorem

Consider model (2) and suppose that at the true value θ_0, $0 < a_0 + b_0 < 1$. Then, there exists a fixed open neighborhood $O = O(\theta_0)$ of θ_0 such that with probability tending to one, as $n \to \infty$, the log likelihood function (6) has a unique maximum point $\hat{\theta}$.

Inference for the Linear Model 6

Theorem

Consider model (2) and suppose that at the true value θ_0, $0 < a_0 + b_0 < 1$. Then, there exists a fixed open neighborhood $O = O(\theta_0)$ of θ_0 such that with probability tending to one, as $n \to \infty$, the log likelihood function (6) has a unique maximum point $\hat{\theta}$. Furthermore, $\hat{\theta}$ is consistent and asymptotically normal,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{D} \mathcal{N}(0, G^{-1})$$

Inference for the Linear Model 6

Theorem

Consider model (2) and suppose that at the true value θ_0, $0 < a_0 + b_0 < 1$. Then, there exists a fixed open neighborhood $O = O(\theta_0)$ of θ_0 such that with probability tending to one, as $n \to \infty$, the log likelihood function (6) has a unique maximum point $\hat{\theta}$. Furthermore, $\hat{\theta}$ is consistent and asymptotically normal,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{D} \mathcal{N}(0, G^{-1})$$

A consistent estimator of G is given by $G_n(\hat{\theta})$, where

$$G_n(\theta) = \sum_{t=1}^{n} Var\left[\frac{\partial l_t(\theta)}{\partial \theta} \mid \mathcal{F}_{t-1}\right] \quad = \quad \sum_{t=1}^{n} \frac{1}{\lambda_t(\theta)} \left(\frac{\partial \lambda_t(\theta)}{\partial \theta}\right) \left(\frac{\partial \lambda_t(\theta)}{\partial \theta}\right)'$$

Inference for the Linear Model 7

Lemma

Define the matrices

$$\begin{aligned} G^m(\theta) &= E\left(\frac{1}{\lambda_t^m}\left(\frac{\partial\lambda_t^m}{\partial\theta}\right)\left(\frac{\partial\lambda_t^m}{\partial\theta}\right)'\right) \\ G(\theta) &= E\left(\frac{1}{\lambda_t}\left(\frac{\partial\lambda_t}{\partial\theta}\right)\left(\frac{\partial\lambda_t}{\partial\theta}\right)'\right). \end{aligned}$$

Under the assumptions of Theorem 2, the above matrices evaluated at the true value $\theta = \theta_0$ satisfy

$$G^m \rightarrow G,$$

as $m \rightarrow \infty$. In addition, G^m and G are positive definite.

Inference for the Linear Model 8

Lemma

Under the assumptions of Theorem 2, the score functions defined by (7) and (11) and evaluated at the true value $\theta = \theta_0$ satisfy the following:

1. $\frac{1}{\sqrt{n}} S_n^m \xrightarrow{D} S^m := \mathcal{N}\left(0, G^m\right)$, *as* $n \to \infty$ *for each* $m = 1, 2 \ldots$

Inference for the Linear Model 8

Lemma

Under the assumptions of Theorem 2, the score functions defined by (7) and (11) and evaluated at the true value $\theta = \theta_0$ satisfy the following:

1. $\dfrac{1}{\sqrt{n}} S_n^m \xrightarrow{D} S^m := \mathcal{N}\left(0, G^m\right)$, *as* $n \to \infty$ *for each* $m = 1, 2 \ldots$
2. $S^m \xrightarrow{D} \mathcal{N}(0, G)$ *as* $m \to \infty$

Inference for the Linear Model 8

Lemma

Under the assumptions of Theorem 2, the score functions defined by (7) and (11) and evaluated at the true value $\theta = \theta_0$ satisfy the following:

1. $\dfrac{1}{\sqrt{n}} S_n^m \xrightarrow{D} S^m := \mathcal{N}\left(0, G^m\right)$*, as* $n \to \infty$ *for each* $m = 1, 2 \ldots$
2. $S^m \xrightarrow{D} \mathcal{N}\left(0, G\right)$ *as* $m \to \infty$
3. $\lim_{m\to\infty} \limsup_{n\to\infty} P\left(|S_n^m - S_n| > \varepsilon\sqrt{n}\right) = 0$*, for every* $\varepsilon > 0$*.*

Inference for the Linear Model 9

Lemma

Under the assumptions of Theorem 2, the Hessian matrices defined by (9) and (12) and evaluated at the the true value $\theta = \theta_0$ satisfy the following:

1. $\frac{1}{n} H_n^m \xrightarrow{P} G^m$ *as* $n \to \infty$ *for each* $m = 1, 2 \ldots,$

Inference for the Linear Model 9

Lemma

Under the assumptions of Theorem 2, the Hessian matrices defined by (9) and (12) and evaluated at the the true value $\theta = \theta_0$ satisfy the following:

1. $\frac{1}{n} H_n^m \xrightarrow{P} G^m$ *as* $n \to \infty$ *for each* $m = 1, 2 \ldots$,
2. $G^m \ \to \ G$, *as* $m \to \infty$,

Inference for the Linear Model 9

Lemma

Under the assumptions of Theorem 2, the Hessian matrices defined by (9) and (12) and evaluated at the the true value $\theta = \theta_0$ satisfy the following:

1. $\frac{1}{n} H_n^m \xrightarrow{P} G^m$ *as* $n \to \infty$ *for each* $m = 1, 2 \ldots,$
2. $G^m \;\to\; G$, *as* $m \to \infty$,
3. $\lim_{m\to\infty} \limsup_{n\to\infty} P(\|H_n^m - H_n\| > \varepsilon n) = 0$, *for every* $\varepsilon > 0$.

Inference for the Log–Linear Model 1

For the log–linear we obtain that the score function is derived as

$$S_n(\theta) \quad = \quad \frac{\partial l(\theta)}{\partial \theta} = \sum_{t=1}^{n} \frac{\partial l_t(\theta)}{\partial \theta} = \sum_{t=1}^{n} \left(Y_t - \exp(\nu_t(\theta))\right) \frac{\partial \nu_t(\theta)}{\partial \theta}, \tag{13}$$

Inference for the Log–Linear Model 1

For the log–linear we obtain that the score function is derived as

$$S_n(\theta) \quad = \quad \frac{\partial l(\theta)}{\partial\theta} = \sum_{t=1}^{n} \frac{\partial l_t(\theta)}{\partial\theta} = \sum_{t=1}^{n} \left(Y_t - \exp(\nu_t(\theta))\right) \frac{\partial\nu_t(\theta)}{\partial\theta}, \qquad (13)$$

where

$$\begin{aligned} \frac{\partial\nu_t(\theta)}{\partial d} &= 1 + a\frac{\partial\nu_{t-1}(\theta)}{\partial d} \\ \frac{\partial\nu_t(\theta)}{\partial a} &= \nu_{t-1}(\theta) + a\frac{\partial\nu_{t-1}(\theta)}{\partial a} \\ \frac{\partial\nu_t(\theta)}{\partial b} &= g(Y_{t-1}) + a\frac{\partial\nu_{t-1}(\theta)}{\partial b} \end{aligned} \qquad (14)$$

Inference for the Log–Linear Model 2

The Hessian matrix is given by

$$\begin{aligned} H_n(\theta) &= -\sum_{t=1}^{n} \frac{\partial^2 l_t(\theta)}{\partial\theta\partial\theta'} \\ &= \sum_{t=1}^{n} \exp(\nu_t(\theta)) \left(\frac{\partial \nu_t(\theta)}{\partial\theta}\right) \left(\frac{\partial \nu_t(\theta)}{\partial\theta}\right)' \qquad (15) \\ &- \sum_{t=1}^{n} (Y_t - \exp(\nu_t(\theta))) \frac{\partial^2 \nu_t(\theta)}{\partial\theta\partial\theta'}. \qquad (16) \end{aligned}$$

Prove consistency and asymptotic normality of MLE along the previous lines.

Simulation for Linear Model 1

Estimators and their mean square error (in parentheses) for model (2) when $(d_0, a_0, b_0) = (0.3, 0.4, 0.5)$ and for different sample sizes by both maximum likelihood and least squares methods. Results are based on 1000 simulations.

Sample Size	MLE			LSE		
n	$\hat{d}$	$\hat{a}$	$\hat{b}$	$\hat{d}$	$\hat{a}$	$\hat{b}$
200	0.3713	0.3756	0.4967	0.3909	0.3790	0.4863
	(0.1429)	(0.0940)	(0.0749)	(0.1589)	(0.1052)	(0.0841)
500	0.3271	0.3923	0.4971	0.3318	0.3922	0.4932
	(0.0803)	(0.0548)	(0.0443)	(0.0949)	(0.0657)	(0.0532)
1000	0.3148	0.3954	0.4985	0.3180	0.3951	0.4965
	(0.0505)	(0.0380)	(0.0314)	(0.0633)	(0.0452)	(0.0373)

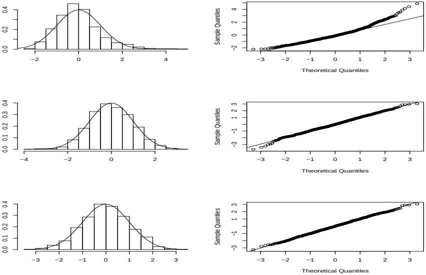
Simulation for Linear Model 2

Comparison of standard errors for model (2) with $(d_0, a_0, b_0) = (0.3, 0.4, 0.5)$.

Sample Size	Simulated standard errors			Standard Errors from $G(\theta_0)$		
n	d	a	b	d	a	b
200	0.1429	0.0940	0.0749	0.0937	0.0733	0.0593
500	0.0803	0.0548	0.0443	0.0574	0.0459	0.0372
1000	0.0505	0.0380	0.0314	0.0403	0.0323	0.0263

Simulation for Linear Model 3

Histograms and qq-plots of the sampling distribution of $\hat{\theta} = (\hat{d}, \hat{a}, \hat{b})$ for the linear model (2) when the true values are $(d_0, a_0, b_0) = (0.3, 0.4, 0.5)$. The results are based on 500 data points and 1000 simulations.



Simulations for Log-Linear Model 1

$d = 0.5$, $a = -0.50$, $b = 0.65$						
Parameters	Sample Size	MLE	Standard Error	Skewness	Kurtosis	p-value
d_0	200	0.501	0.187	0.269	3.226	0.443
a_0		-0.505	0.130	0.451	3.695	0.023
b_0		0.651	0.104	0.064	3.033	0.883
d_0	500	0.498	0.114	0.187	2.842	0.224
a_0		-0.497	0.081	0.208	3.502	0.578
b_0		0.649	0.063	0.087	2.942	0.556
d_0	1000	0.501	0.079	0.077	2.898	0.728
a_0		-0.500	0.055	0.155	3.254	0.936
b_0		0.649	0.045	0.022	2.819	0.477

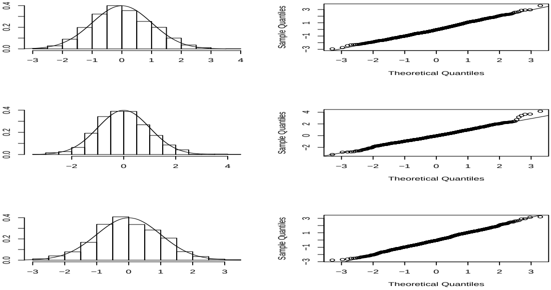
Simulations for Log-Linear Model 2

$d = 0.5$, $a = -0.50$, $b = -0.35$

Parameters	Sample Size	MLE	Standard Error	Skewness	Kurtosis	p-value
d_0	200	0.488	0.113	-0.458	3.902	0.078
a_0		-0.375	0.303	1.655	6.775	0.000
b_0		-0.370	0.123	-0.072	3.304	0.982
d_0	500	0.492	0.066	-0.019	2.927	0.957
a_0		-0.469	0.149	1.057	6.340	0.000
b_0		-0.353	0.075	-0.112	2.843	0.674
d_0	1000	0.499	0.046	-0.109	2.851	0.806
a_0		-0.485	0.102	0.494	3.961	0.295
b_0		-0.353	0.054	-0.082	2.871	0.697

Simulations for Log-Linear Model 3

From top to bottom: Histograms and qq-plots of the standardized sampling distribution of $\hat{\theta} = (\hat{d}, \hat{a}, \hat{b})$ for the log–linear model (4) when the true values are $(d_0, a_0, b_0) = (0.50, -0.50, 0.65)$. The results are based on 500 data points and 1000 simulations. From top to the bottom $\hat{d}$; $\hat{a}$; $\hat{b}$.



Transactions Data 1

The linear model (2) yields the following results:

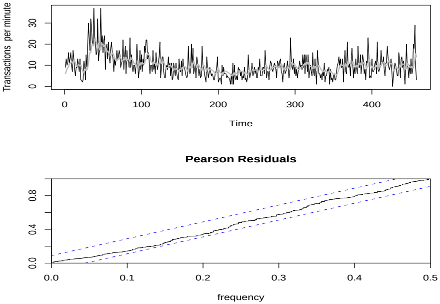
$$\begin{array}{cccccc} \hat{\lambda}_t = & 0.5808 & +0.7445 & \hat{\lambda}_{t-1} + & 0.1986 & Y_{t-1} \\ & (0.1628) & (0.0264) & & (0.0167) & \end{array}$$

Define the Pearson residuals

$$e_t = (Y_t - \lambda_t)/\sqrt{\lambda_t}$$

Transactions Data 2

Top: Observed and predicted (gray) number of transactions per minute using (2).
Bottom: Cumulative periodogram plot of the Pearson residuals.



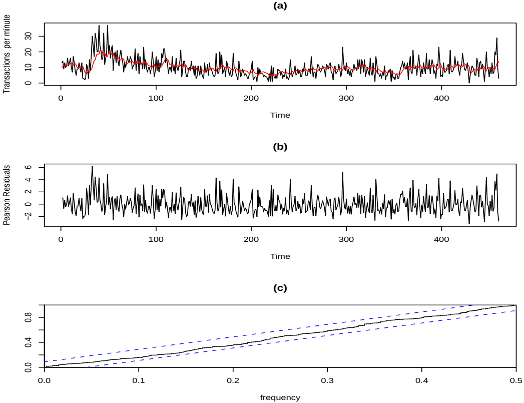
Transactions Data 3

The log–linear model is fitted as

$$\begin{array}{cccccc} \hat{\nu}_t = & 0.1051 & +0.7465 & \hat{\nu}_{t-1}+ & 0.2072 & Y_{t-1} \\ & (0.0345) & (0.0266) & & (0.0194) & \end{array}$$

Transactions Data 4

Top: Prediction for the log-linear model. Center: Time series plot of Pearson residuals
Bottom: Cumulative periodogram plot of the Pearson residuals.



Transactions Data 5

To compare the models, we calculate the mean square error of the Pearson residuals defined by

$$\sum_{t=1}^{N} e_t^2/(N-p),$$

where p is the number of estimated parameters.
It turns out that

1. For the linear model the mean square error of the Pearson residuals is equal to 2.3686

Transactions Data 5

To compare the models, we calculate the mean square error of the Pearson residuals defined by

$$\sum_{t=1}^{N} e_t^2/(N-p),$$

where p is the number of estimated parameters.
It turns out that

1. For the linear model the mean square error of the Pearson residuals is equal to 2.3686
2. For the log–linear model the mean square error of the Pearson residuals is equal to 2.3911

All of the models yield similar conclusions.

DEAR BEN

HAPPY BIRTHDAY !!!

T. J. Watson Research Center

Laplace Periodogram and Beyond

Ta-Hsin Li

Department of Mathematical Sciences
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598-0218, USA

thl@us.ibm.com

July 30-31, 2009, University of Maryland, College Park

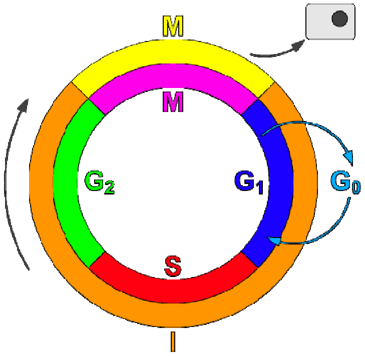
Dedicated to Dr. Kedem for His 65th Birthday

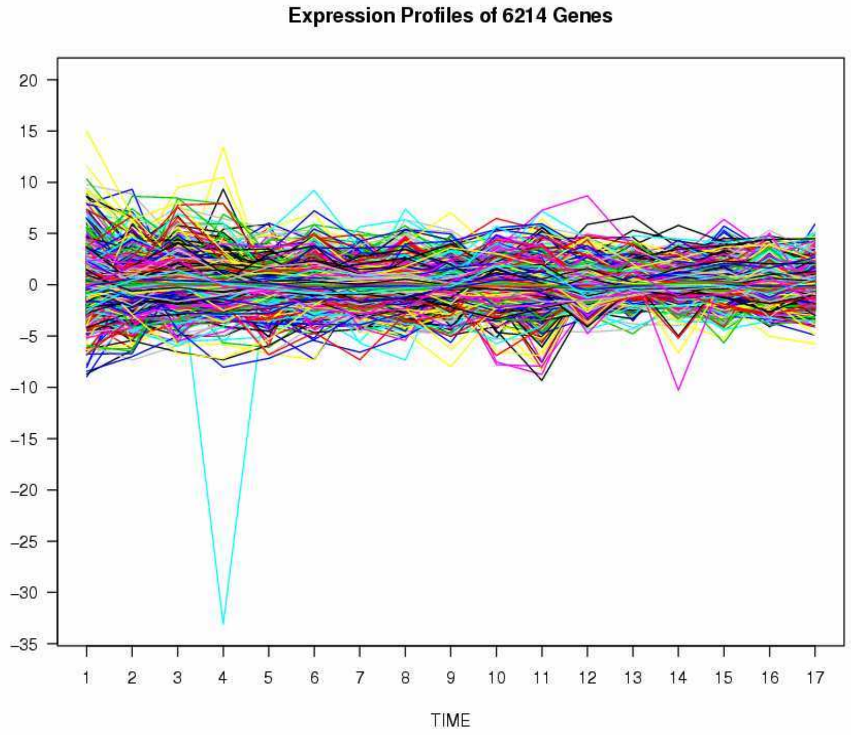
Also In Celebration of

Dr. Manny Parzen's 80th Birthday
and
Dr. Jerry North's 70th Birthday

Motivating Example 1

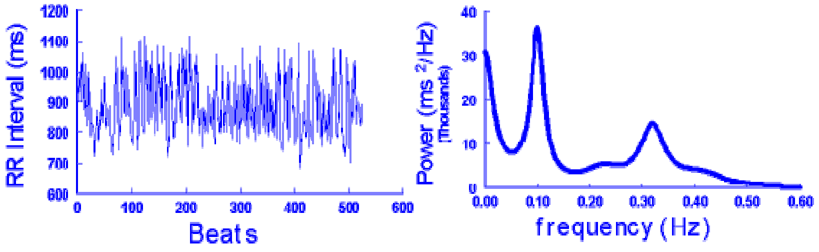
- **Discovery of Periodically Expressed Genes**
  - Periodically expressed genes contribute to the mechanisms that regulate the cell-division cycle which consists of a series of events that take place in an eukaryotic cell leading to its replication with 4 distinct phases: $\mathbf{G_1}$, **S**, $\mathbf{G_2}$, **M**.
  - Spectral domain techniques have been used for automatic transcription of gene expression profiles to identify periodically expressed genes in gnome-wide time-course studies of a wide range of organisms.
  - Outlier contamination degrades the identification results.

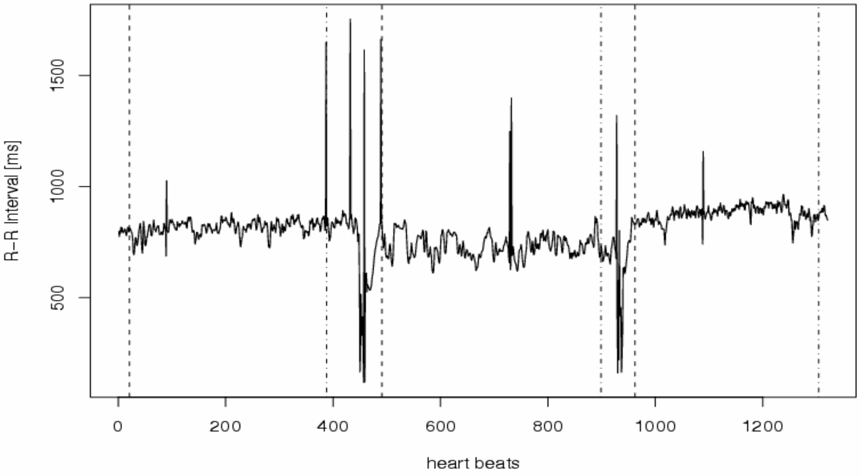




Motivating Example 2

- **Heart Rate Variability (HRV) Analysis**
  - The beat-to-beat alterations in heart rate, measured from ECG records, reflect the activity of the autonomic nervous system in regulating the cardiac rhythm.
  - Spectral analysis techniques have been applied to HRV data in studying the relationship between HRV and various physiological conditions.
  - Short-term studies rely on manual cleaning and editing of HRV records.
  - Automatic analysis of long-term HRV records requires robust methods to cope with inevitable contamination by ectopic events and artifacts.





Spangl & Dutter (2007)

Robust Spectral Analysis

- **Covariance Method: Fourier Analysis of Autocovariances**

$$\hat{r}(\tau) = \text{lag - } \tau \text{ sample autocovariance}$$

$$= \text{lag - } \tau \text{ sample autocorrelation} \div \text{sample variance}$$

$$\hat{f}(\omega) = \sum w(\tau)\, \hat{r}(\tau) \cos(\tau\omega)$$

$$f(\omega) = \sum r(\tau) \cos(\tau\omega) = \sigma^2 \sum \rho(\tau) \cos(\tau\omega)$$ (power spectrum)

- **Robustification Against Outliers**
  - Outliers due to measurement errors or heavy-tailed distributions of the underlying physical processes
  - Manually identify and clean anomalous data points
  - Replace sample autocovariances with robust alternatives insensitive to outliers

Direct Method of Spectral Analysis

Jean Baptiste Joseph Fourier
(1768-1830)

- **Fourier Analysis of Time Series Samples**

$$Y_t = A_0 + \sum_{k=1}^{(n-1)/2} [A_k \cos(\omega_k t) + B_k \sin(\omega_k t)] \quad (t = 1, \ldots, n)$$

$$\omega_k = 2\pi k / n \quad \text{(Fourier frequencies)}$$

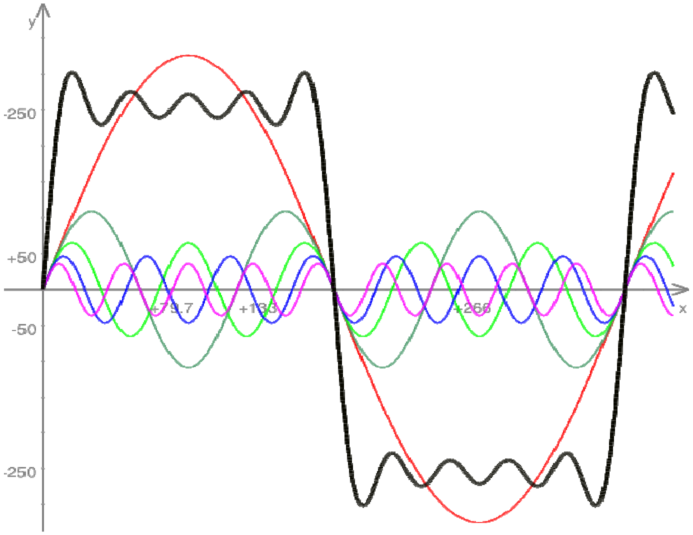
- **Calculation of Fourier Coefficients**

“Multiplying both sides by $\cos(\omega_k t)$ and then *summing* from *1* to *n* yields:” – Fourier, 1822

$$A_0 = n^{-1} \sum_{t=1}^{n} Y_t$$

$$A_k = 2n^{-1} \sum_{t=1}^{n} Y_t \cos(\omega_k t)$$

$$B_k = 2n^{-1} \sum_{t=1}^{n} Y_t \sin(\omega_k t)$$



Periodogram

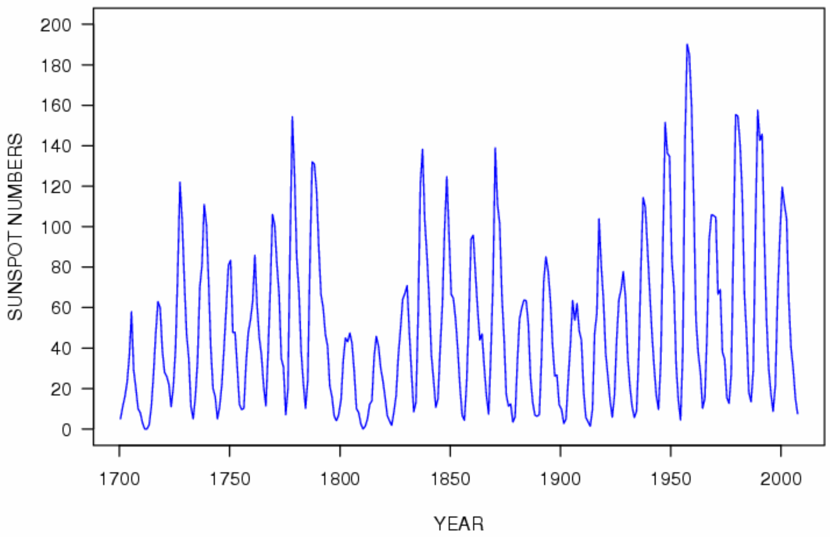
- **Schuster's Periodogram (1898)**

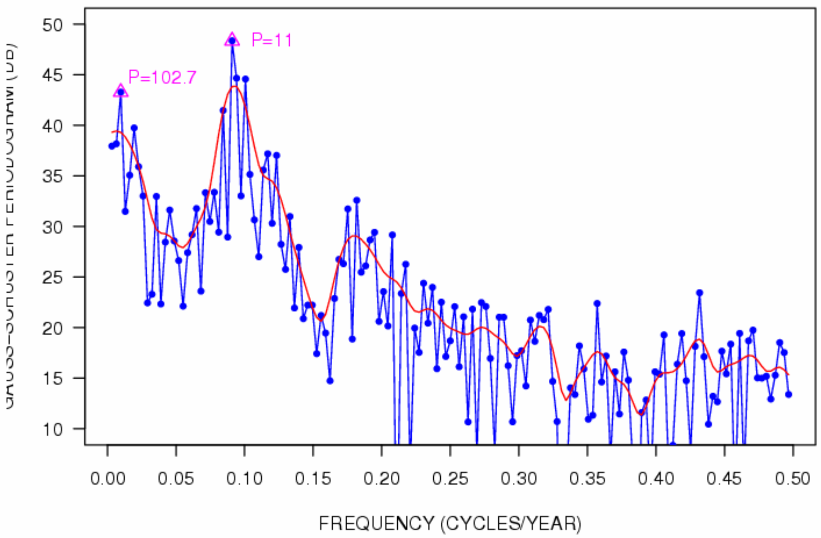
$$G_n(\omega_k) = \frac{1}{4} n\,(A_k^2 + B_k^2) = n^{-1} \left| \sum_{t=1}^{n} Y_t \exp(-it\omega_k) \right|^2$$

Sir Arthur Schuster
(1851-1934)

- **Periodogram of Sunspot Numbers**

"The periodogram of sunspots would show a 'band' in the neighborhood of a period of eleven years." – Schuster, 1898





Spectral Estimation

Emanuel Parzen

- **Periodogram Smoothing**

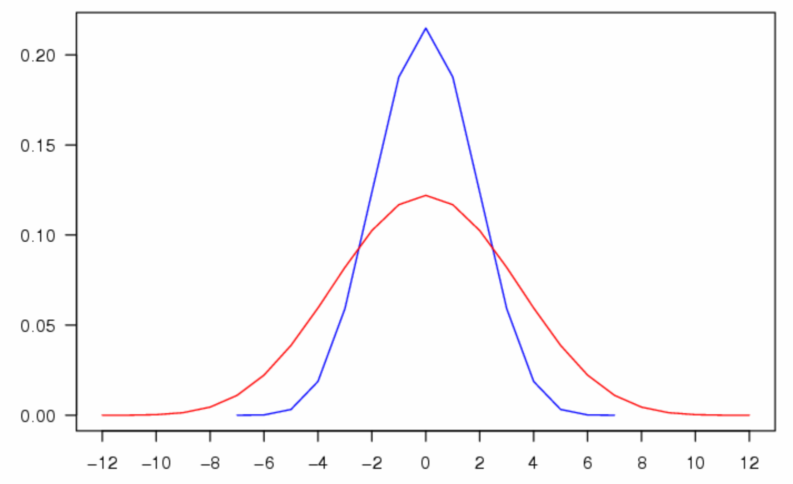
$$\tilde{G}_n(\omega_k) = \sum_j w_{mj} G_n(\omega_k - \omega_j), \quad w_{mj} \ge 0, \sum_j w_{mj} = 1$$

  - If $\{w_{mj}\} \to \{\delta_j\}$ but too fast as $m = m(n) \to \infty$, then

$$\tilde{G}_n(\omega_k) - f(\omega_k) \to 0 \quad \text{as } n \to \infty$$

- **Parzen’s Spectral Window**

$$w_{mj} \propto \left\{ \frac{\sin(\frac{1}{2}[n/m]\omega_j)}{\sin(\frac{1}{2}\omega_j)} \right\}^4, \quad |j| \le m$$



  - Smoothing parameter m can be selected by, e.g., minimizing the generalized cross-validation criterion (Ombao et al. 2001):

$$\mathrm{GCV}(m) = \sum_k \frac{\Phi_{KL}(G_n(\omega_k)/\tilde{G}_n(\omega_k))}{(1 - w_{m0})^2}$$

$$\Phi_{KL}(x) = x - \log x - 1 \ge 0 \quad \text{(Kullback - Leibler kernel)}$$

Statistical Distribution

- **Asymptotic Theory**
  - $\{Y_t\}$ is a zero-mean random process, stationary in second moments with a continuous spectrum $f(\omega)$. Under certain weak serial dependent conditions (e.g., m-dependent, mixing) and as $n \to \infty$, (Brockwell & Davis 1991)

$$G_n(\omega_k) \sim \tfrac{1}{2}\sigma^2 R(\omega_k)\chi_2^2$$

$$G_n(\omega_k) \perp G_n(\omega_{k'}) \quad (k \neq k')$$

$$R(\omega) = \sum \rho(\tau)\cos(\omega\tau) = \text{autocorrelation spectrum}$$

$$\rho(\tau) = E(Y_t Y_{t-\tau})/\sigma^2 = \text{lag-}\tau\text{ autocorrelation, } \sigma^2 = \mathrm{Var}(Y_t)$$

$$f(\omega) = \sigma^2 R(\omega) = E\{\tfrac{1}{2}\sigma^2 R(\omega)\chi_2^2\} = \text{power spectrum}$$

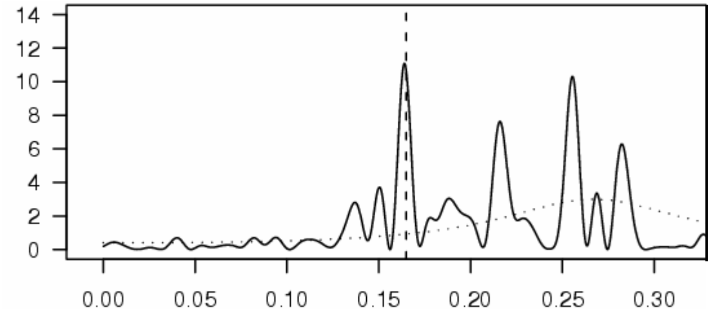
Application of Spectral Analysis

- **Detection of Hidden Periodicity**
  - $\{Y_t\}$ = sinusoidal signal with unknown frequency and amplitude + noise process with continuous spectrum $f(\omega)$. The sinusoid can be detected by Fisher’s test

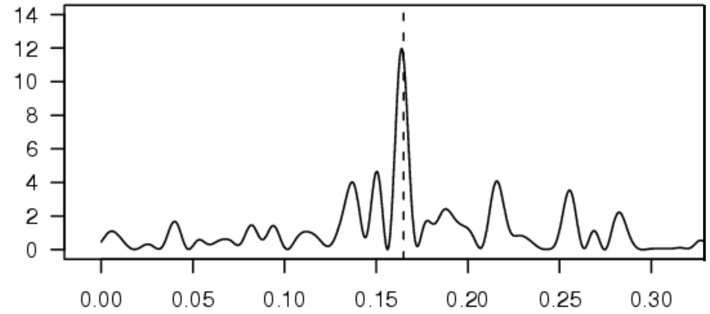
$$g = \frac{\max_k G_n(\omega_k)/\hat{f}(\omega_k)}{\sum_k G_n(\omega_k)/\hat{f}(\omega_k)} \ge \theta$$

Ronald A. Fisher
(1890-1962)

where $\hat{f}(\omega)$ is an estimator of the noise spectrum $f(\omega)$ from training data

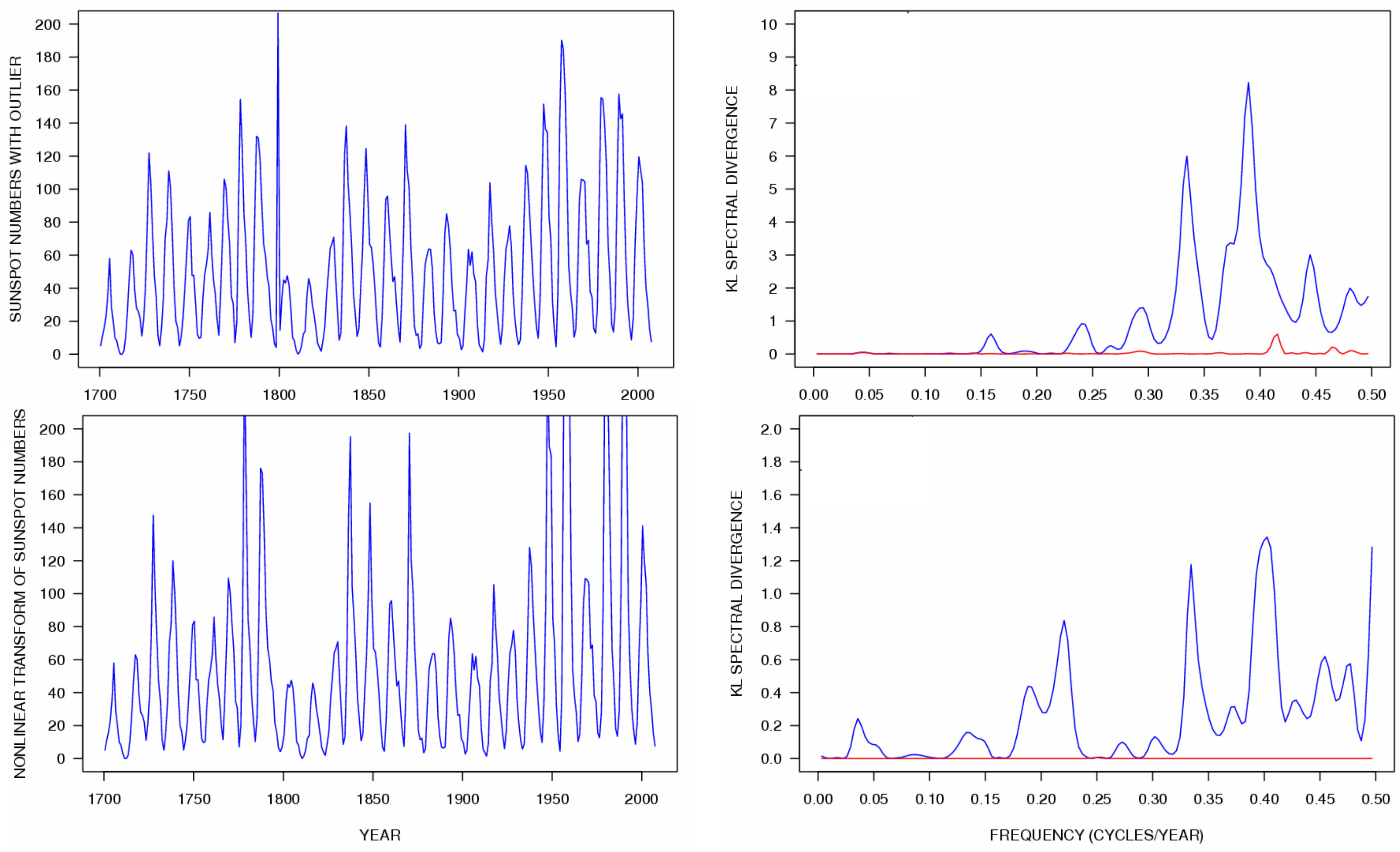


Without Normalization by Noise Spectrum



With Normalization by Noise Spectrum

Sensitivity to Outliers and Nonlinear Distortion



Measure of spectral divergence : $\Phi_{KL}(f_1(\omega)/f_0(\omega))$

Least Squares Reformulation

- **Method of Least Squares (LS)**

$$\tilde{\boldsymbol{\beta}}_n(\omega_k) = [A_k, B_k]^T = \underset{\boldsymbol{\beta} \in R^2}{\arg\min} \sum_{t=1}^{n} | Y_t - \mathbf{x}_t(\omega_k)^T \boldsymbol{\beta} |^2$$

$$\mathbf{x}_t(\omega_k) = [\cos(\omega_k t), \sin(\omega_k t)]^T$$

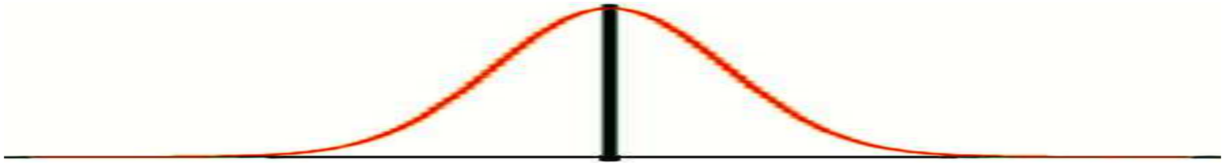
$$\boxed{G_n(\omega_k) = \frac{1}{4} n \| \tilde{\boldsymbol{\beta}}_n(\omega_k) \|^2}$$

(Gauss Periodogram)

- **Gauss Maximum Likelihood Estimation**

$$Y_t = \mathbf{x}_t(\omega_k)^T \boldsymbol{\beta}_0 + \varepsilon_t$$

$$\{\varepsilon_t\} \sim \text{IID } N(0, \sigma^2) \;\Rightarrow\; \tilde{\boldsymbol{\beta}}_n \text{ is the MLE of } \boldsymbol{\beta}_0$$



Carl Friedrich Gauss
(1777-1855)

Adrie Marie Legendre
(1752-1833)

Least Absolute Deviations

Pierre Simon Laplace
(1749-1827)

- **Method of Least Absolute Deviations (LAD)**

$$\hat{\boldsymbol{\beta}}_n(\omega_k) = \underset{\boldsymbol{\beta} \in R^2}{\arg\min} \sum_{t=1}^{n} | Y_t - \mathbf{x}_t(\omega_k)^T \boldsymbol{\beta} |$$

- **Laplace Periodogram**

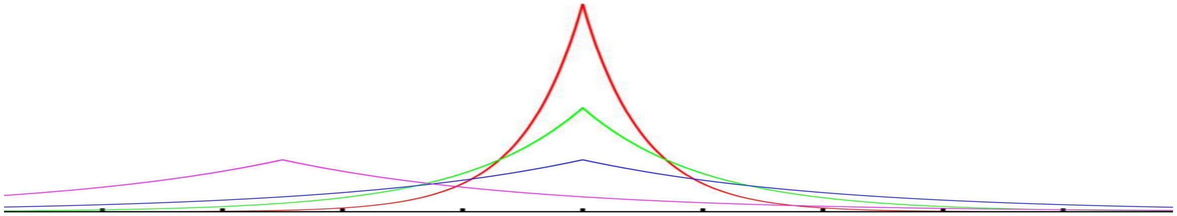
$$\boxed{L_n(\omega_k) = \frac{1}{4} n \, \| \hat{\boldsymbol{\beta}}_n(\omega_k) \|^2}$$

Roger Joseph Boscovich
(1711-1787)

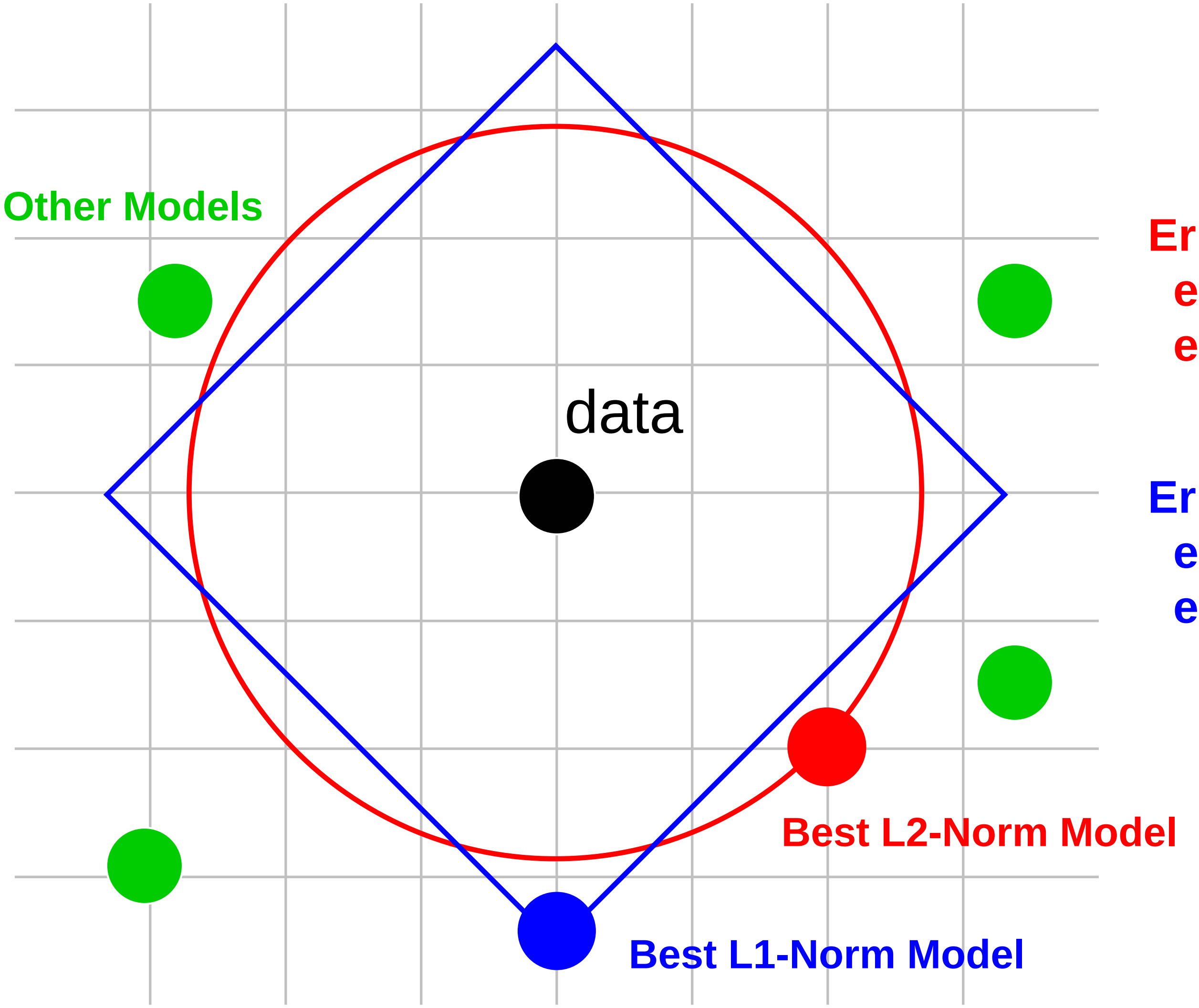
- **Laplace Maximum Likelihood Estimation**

$$Y_t = \mathbf{x}_t(\omega_k)^T \boldsymbol{\beta}_0 + \varepsilon_t$$

$$\{\varepsilon_t\} \sim \text{IID } L(0, \sigma^2) \;\Rightarrow\; \hat{\boldsymbol{\beta}}_n \text{ is the MLE of } \boldsymbol{\beta}_0$$



Measurement of Errors: L2 Norm vs L1 Norm



Other Models

data

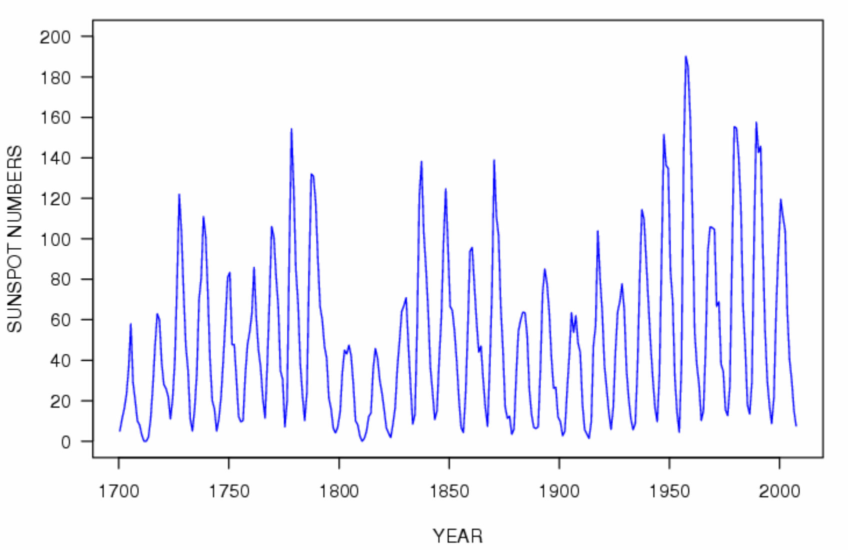
Best L2-Norm Model

Best L1-Norm Model

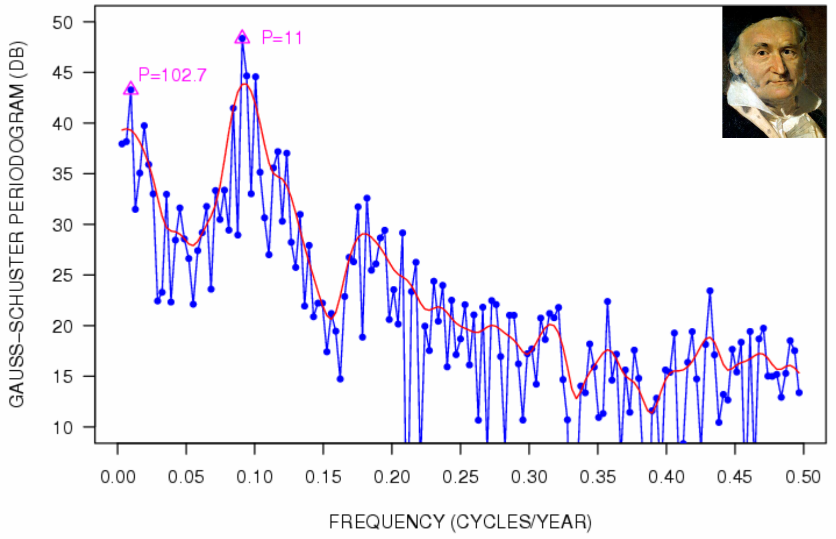
Error Composition:
$e_1 = -2$
$e_2 = 2$

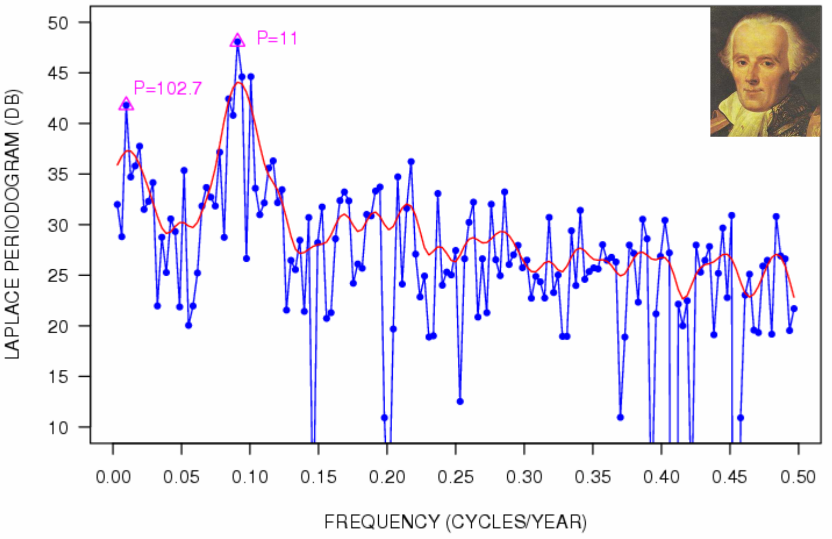
Error Composition:
$e_1 = 0$
$e_2 = 3.2$

Laplace Periodogram of Sunspot Numbers



Laplace periodogram need to be smoothed in the same way as the ordinary periodogram.





What Is Laplace Periodogram?

- **Question**
  - The ordinary (Gauss) periodogram estimates the power spectrum which is the Fourier transform of the autocovariance function of the underlying random process.
  - What does Laplace periodogram estimate?
  - Does it represent serial dependence in some way?

- **Challenge**
  - No closed-form expression
  - Cannot compute the mean and variance

Statistical Theory

- **Distribution of the Sample Median**
  - Let $\{Y_t\}$ be IID with $Y_t \sim F$, $F(0) = 1/2$, $F'(0) > 0$.

$$\hat{\beta}_n = \underset{\beta \in R}{\arg\min} \sum_{t=1}^{n} | Y_t - \beta |$$

$$\boxed{\sqrt{n}\,\hat{\beta}_n \xrightarrow{D} N(0, \eta^2), \quad \eta^2 = 1/\{2F'(0)\}^2 \qquad \text{(sparsity)}}$$

- **Distribution of LAD Regression (Bassett & Koenker 1978)**
  - Let $Y_t = \mathbf{x}_t^T \boldsymbol{\beta}_0 + \varepsilon_t$ with $\{\varepsilon_t\} \sim$ IID F, $F(0) = 1/2$, $F'(0) > 0$.

$$\hat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta} \in R^q}{\arg\min} \sum_{t=1}^{n} | Y_t - \mathbf{x}_t^T \boldsymbol{\beta} |$$

$$\boxed{\sqrt{n}\,\hat{\boldsymbol{\beta}}_n \xrightarrow{D} N(0, \eta^2 \mathbf{D}), \quad \eta^2 = 1/\{2F'(0)\}^2, \ \mathbf{D} = \lim_{n \to \infty} \left[n^{-1} \sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t^T \right]^{-1}}$$

Statistical Theory

- **Distribution of the Laplace Periodogram (Li 2008)**
  - Let $\{Y_t\}$ be a random process with $Y_t \sim F$, $F(0) = 1/2$, $F'(0) > 0$. Let

$$\hat{\boldsymbol{\beta}}_n(\omega) = \arg\min_{\boldsymbol{\beta} \in R^2} \sum_{t=1}^{n} | Y_t - \mathbf{x}_t^T(\omega)\boldsymbol{\beta} |, \quad \mathbf{x}_t(\omega) = [\cos(\omega t), \sin(\omega t)]^T$$

Under certain conditions of stationarity (e.g., strictly stationary) and weak serial dependence (e.g., m-dependent, mixing),

$$\sqrt{n}\,\hat{\boldsymbol{\beta}}_n(\omega) \xrightarrow{D} N(0, \eta^2 \mathbf{D}(\omega)), \quad \eta^2 = 1/\{2F'(0)\}^2, \quad \mathbf{D}(\omega) = 2\,S(\omega)\,\mathbf{I}$$

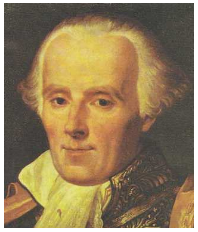
$$L_n(\omega) \sim \tfrac{1}{2}\eta^2\, Z(\omega)\chi_2^2$$

$$L_n(\omega_k) \perp L_n(\omega_{k'}) \quad (k \neq k')$$

  - **Important Note:** The result does not require the existence of moments and hence is applicable to heavy-tailed random processes of infinite variance.

Statistical Theory

- **Comparison with the Gauss Periodogram**



$G_n(\omega) \sim \frac{1}{2}\sigma^2 R(\omega)\chi_2^2$	$L_n(\omega) \sim \frac{1}{2}\eta^2 S(\omega)\chi_2^2$
$G_n(\omega_k) \perp G_n(\omega_{k'})\ \ (k \neq k')$	$L_n(\omega_k) \perp L_n(\omega_{k'})\ \ (k \neq k')$
$\boldsymbol{\sigma^2}$ (variance)	$\boldsymbol{\eta^2}$ (sparsity)
$\mathbf{R}(\omega)$ (autocorrelation spectrum)	**$\mathbf{S}(\omega)$???**
$\mathbf{f_G}(\omega) = \sigma^2 \mathbf{R}(\omega)$ (power or Gauss spectrum)	**$\mathbf{f_L}(\omega) = \eta^2 \mathbf{S}(\omega)$: Laplace spectrum**

Zero-Crossing Spectrum

- **Stationarity in Zero Crossings**
  - $\{Y_t\}$ is stationary in zero crossings iff for any t,

$$P(Y_t Y_{t-\tau} < 0) = \gamma(\tau) \qquad \text{(lag-}\tau\text{ zero-crossing rate)}$$

  - If $\{Y_t\}$ is stationary in zero crossings, then

$$\gamma(\tau) = (1 - r_Z(\tau))/2 \qquad r_Z(\tau) = \mathrm{Cov}(Z_t, Z_{t-\tau})$$

$$Z_t = 2\{I(Y_t < 0) - 1/2\} \qquad \text{(zero-crossing process)}$$

- **Zero-Crossing Spectrum (ZCS)**

$$\boxed{S(\omega) = \sum_{\tau=-\infty}^{\infty} (1 - 2\gamma(\tau)) \cos(\omega\tau) = \sum_{\tau=-\infty}^{\infty} r_Z(\tau) \cos(\omega\tau)}$$

For white noise, $\gamma(\tau) = (1 - \delta_\tau)/2 \Rightarrow S(\omega) = 1$

Benjamin Kedem

"The number of zero-crossings observed in a finitely long real-valued time series may be review as a measure of the oscillation exhibited by the time series."

"Zero-crossings have a certain advantage in the presence of outliers."

"Another example where zero-crossings have an advantage is when the process is strictly stationary but has no moments."

Kedem, 1994

Zero-Crossing Spectrum

- **Zero-crossing spectrum depicts the serial dependence of time series from a different perspective.**
  - For elliptically distributed processes that are stationary in second moments,

Arcsine formula: $\gamma(\tau) = 1/2 - (1/\pi)\arcsin(\rho(\tau))$

ZCS: $S(\omega) = \sum_{\tau=-\infty}^{\infty} (2/\pi)\arcsin(\rho(\tau))\cos(\omega\tau)$

Compare with $R(\omega) = \sum_{\tau=-\infty}^{\infty} \rho(\tau)\cos(\omega\tau)$

  - The one-to-one relationship has been exploited to estimate the power spectrum from clipped Gaussian processes (Hinich 1967, McNeil 1967, Brillinger 1968).

Zero-Crossing Spectrum

- **Invariance to Nonlinearity**
  - Let $\Phi(.)$ be a monotone function such that $\Phi(0)=0$ and $\Phi'(0)=c > 0$.

$$\boxed{S_{\Phi(Y)}(\omega) = S_Y(\omega)}$$

(unchanged)

$$\boxed{f_{L,\Phi(Y)}(\omega) = c^2 f_{L,Y}(\omega)}$$

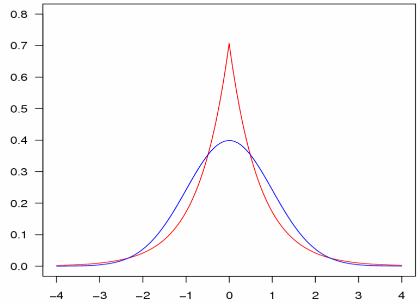
(simple scaling)

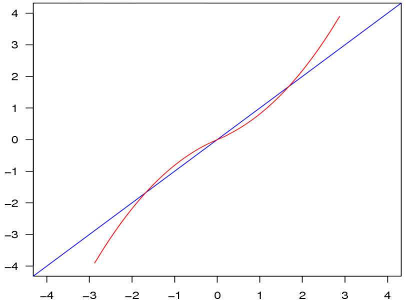
For autocorrelation spectrum and power spectrum,

$$R_{\Phi(Y)}(\omega) \not\propto R_Y(\omega), \quad f_{G,\Phi(Y)}(\omega) \not\propto f_{G,Y}(\omega)$$

(complicated functional)

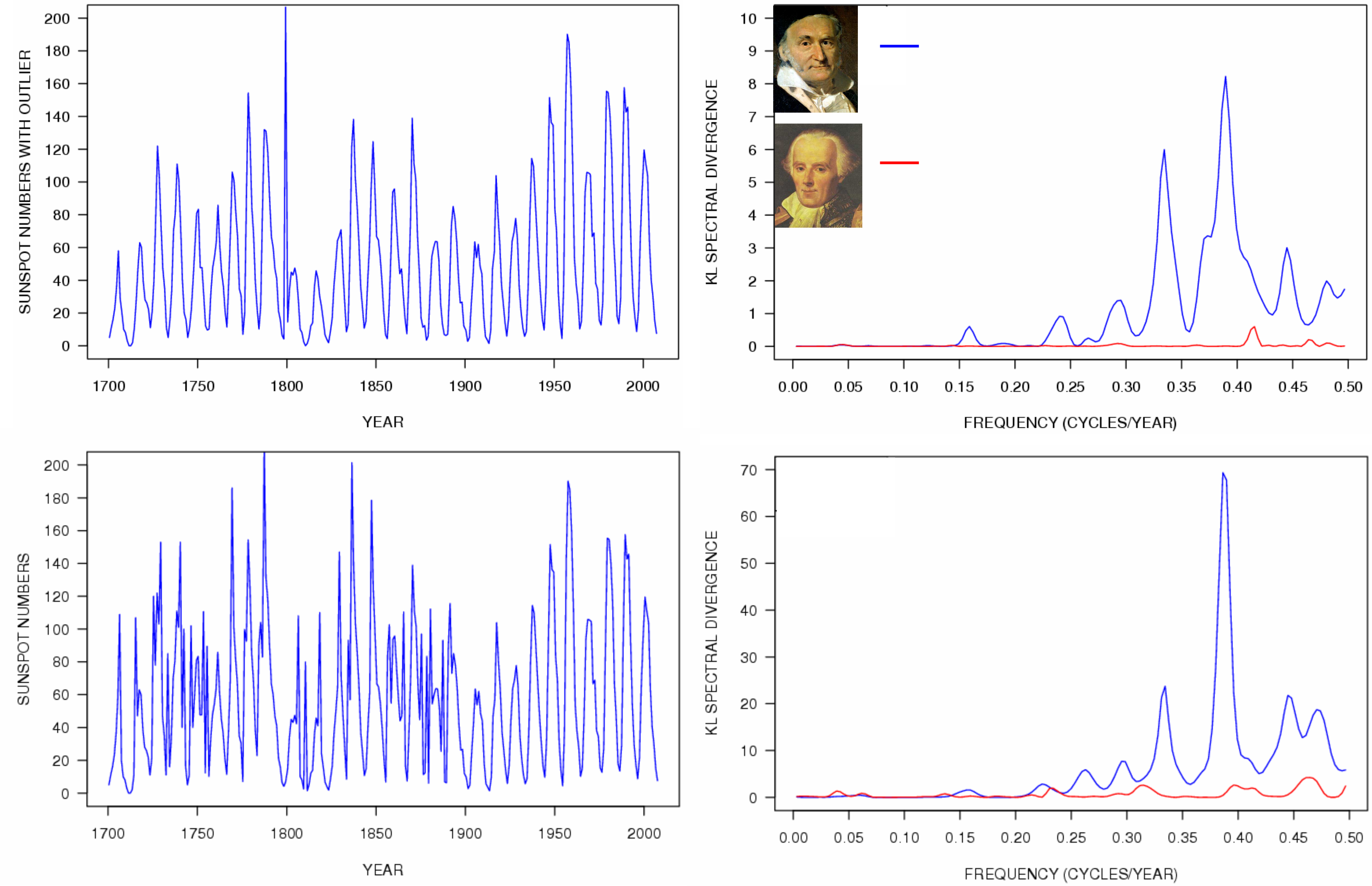
Gauss vs. Laplace



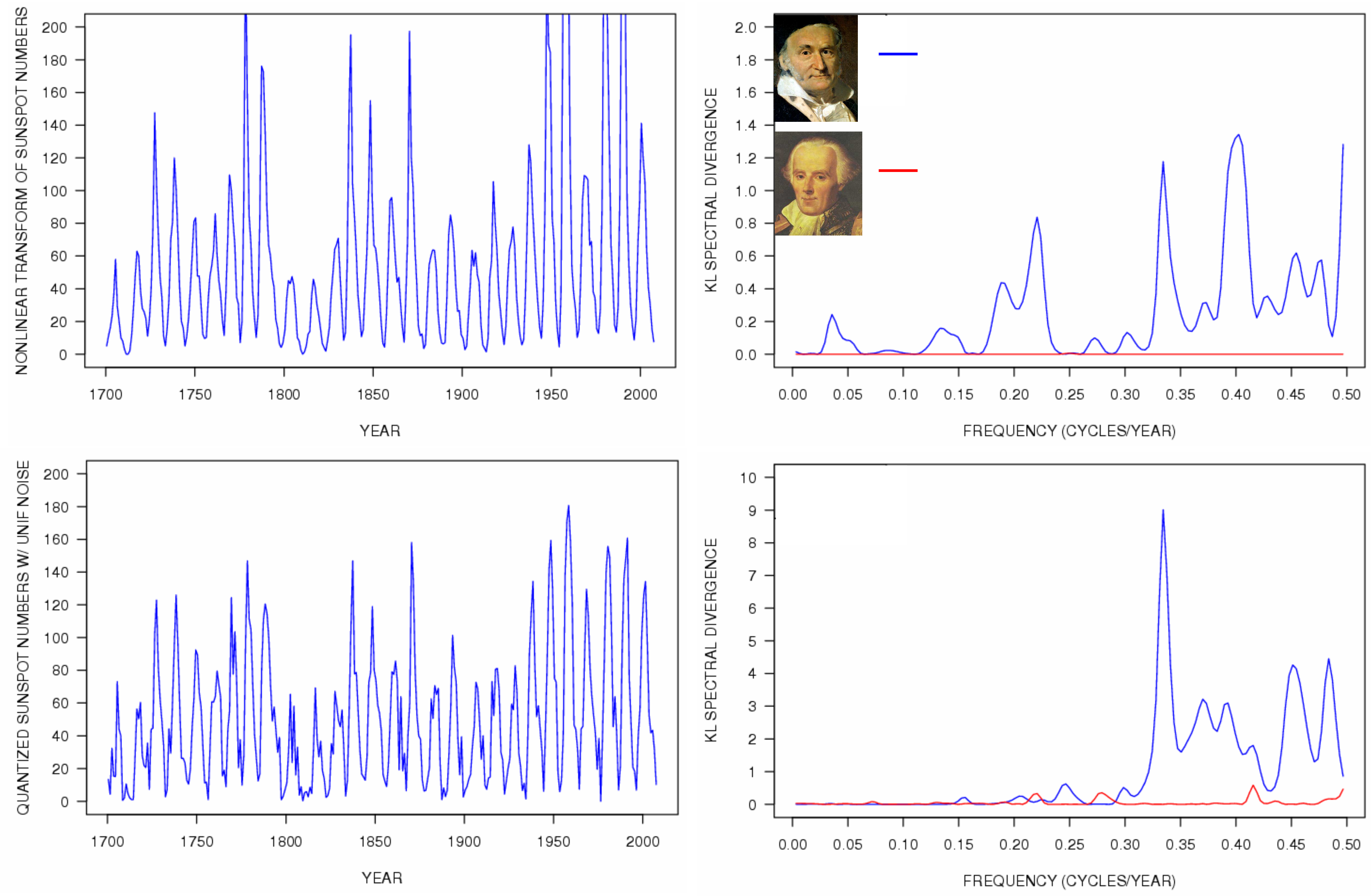


Gauss Periodogram	Laplace Periodogram
🙁 Sensitive to outliers	🙂 Robust to outliers
🙁 Sensitive to nonlinearity	🙂 Robust to nonlinearity
🙁 Less efficient for heavy tailed noise	🙂 More efficient for heavy tailed noise
🙂 More efficient for light-tailed noise	🙁 Less efficient for light-tailed noise
🙂 Unique solution	🙁 Possibly multiple solutions
🙂 Fast algorithm	🙁 Slower algorithm

Robustness to Outlier and Impulsive Noise



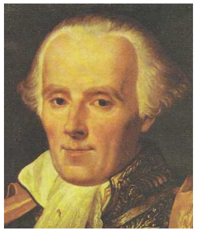
Robustness to Nonlinearity and Quantization Noise



Detection of Hidden Periodicity in Heavy-Tailed Noise

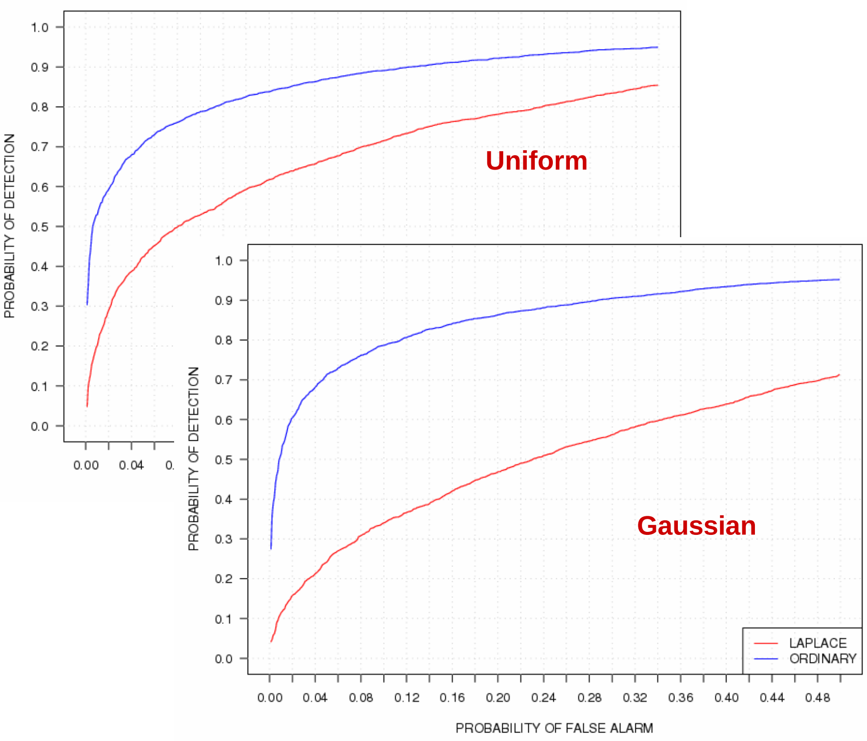
$$H_0 : Y_t = \varepsilon_t,\ \{\varepsilon_t\} \sim \text{IID white noise}$$

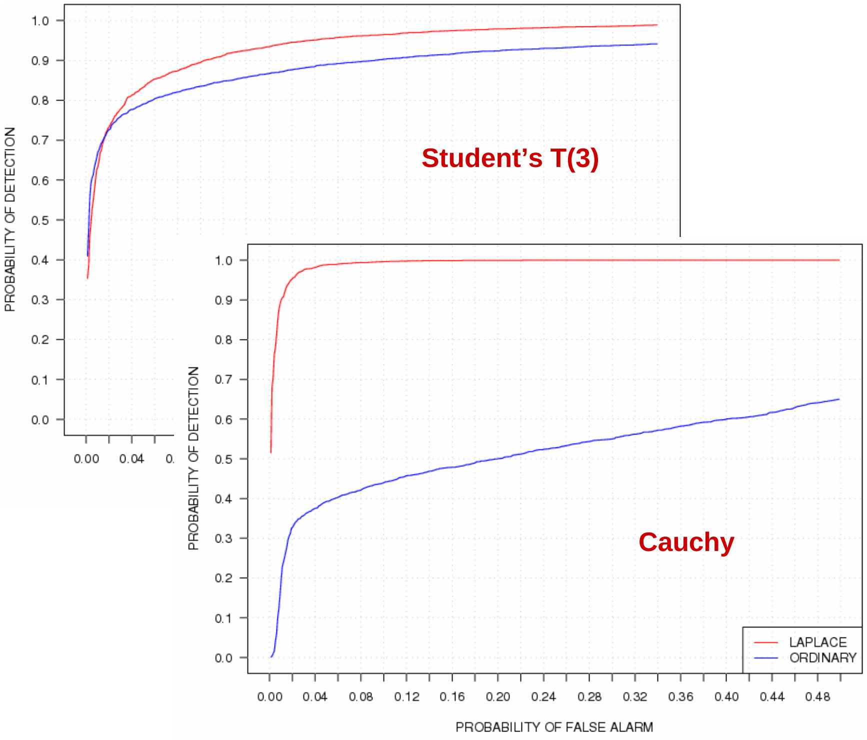
$$H_1 : Y_t = \mathbf{x}_t(\omega_0)^T \boldsymbol{\beta}_0 + \varepsilon_t,\ \omega_0, \boldsymbol{\beta}_0 \text{ unknown}$$



Gauss-Fisher Detector: H_1 iff $g_G > \theta_G$

Laplace-Fisher Detector: H_1 iff $g_L > \theta_L$





Generalization: Least Lp Norm Criterion

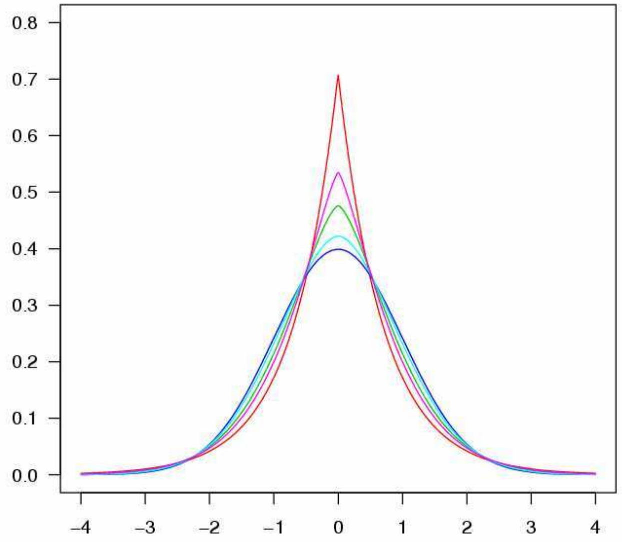
- **Method of Least Lp Norm (LLP)**

$$\hat{\boldsymbol\beta}_n^{(p)}(\omega_k) = \underset{\boldsymbol\beta\in R^2}{\arg\min} \sum_{t=1}^{n} | Y_t - \mathbf{x}_t(\omega_k)^T \boldsymbol\beta |^p$$

$$p = 2 \Rightarrow \text{LS}; \quad p = 1 \Rightarrow \text{LAD}$$

- **The Lp-Norm Periodogram**

$$\boxed{L_n^{(p)}(\omega_k) = \frac{1}{4} n \,||\, \hat{\boldsymbol\beta}_n^{(p)}(\omega_k) \,||^2}$$



- **Generalized Gaussian Maximum Likelihood Estimation**

$$Y_t = \mathbf{x}_t(\omega_k)^T \boldsymbol\beta_0 + \varepsilon_t$$

$$\{\varepsilon_t\} \sim \text{IID } L_p(0, \sigma^2) \;\Rightarrow\; \hat{\boldsymbol\beta}_n^{(p)} \text{ is the MLE of } \boldsymbol\beta_0$$

Generalization: Spherical Processes

Gerald R. North

- **Laplace’s Spherical Harmonics Expansion**

$$Y(\theta,\varphi)=\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} C_{\ell m}\, Y_{\ell m}(\theta,\varphi)$$

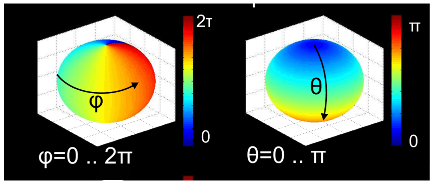
(real form)

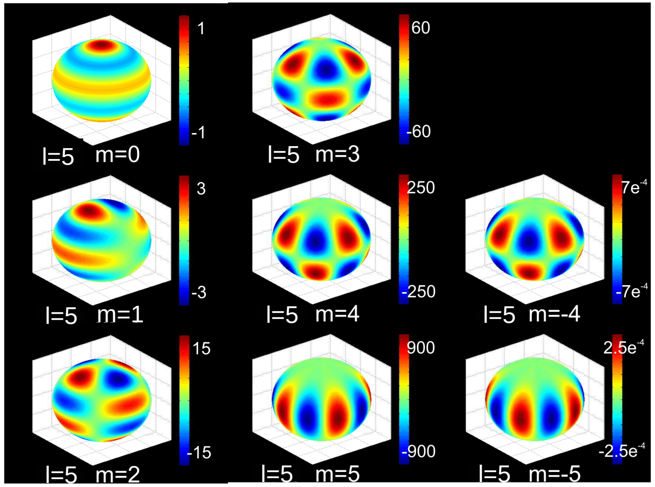
- **Least Lp-Norm Estimation**

$$\{\hat{C}^{(p)}_{\ell m}\} := \arg\min_{\{C_{\ell m}\}} \sum_{i=1}^{n} \left| Y(\theta_i,\varphi_i) - \sum_{\ell=0}^{L}\sum_{m=-\ell}^{\ell} C_{\ell m}\, Y_{\ell m}(\theta_i,\varphi_i)\right|^p$$

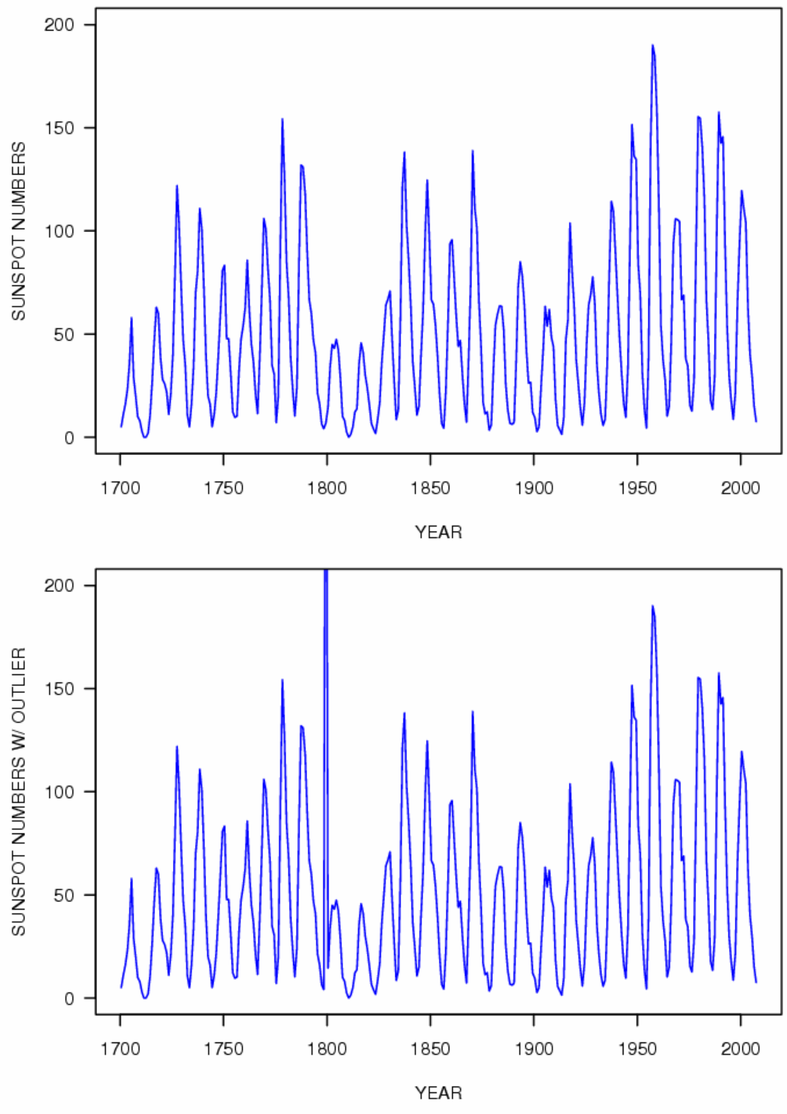
- **Lp-Norm Spherical Periodogram**

$$\boxed{L^{(p)}_{\ell m} := a\,|\,\hat{C}^{(p)}_{\ell m}\,|^2}$$

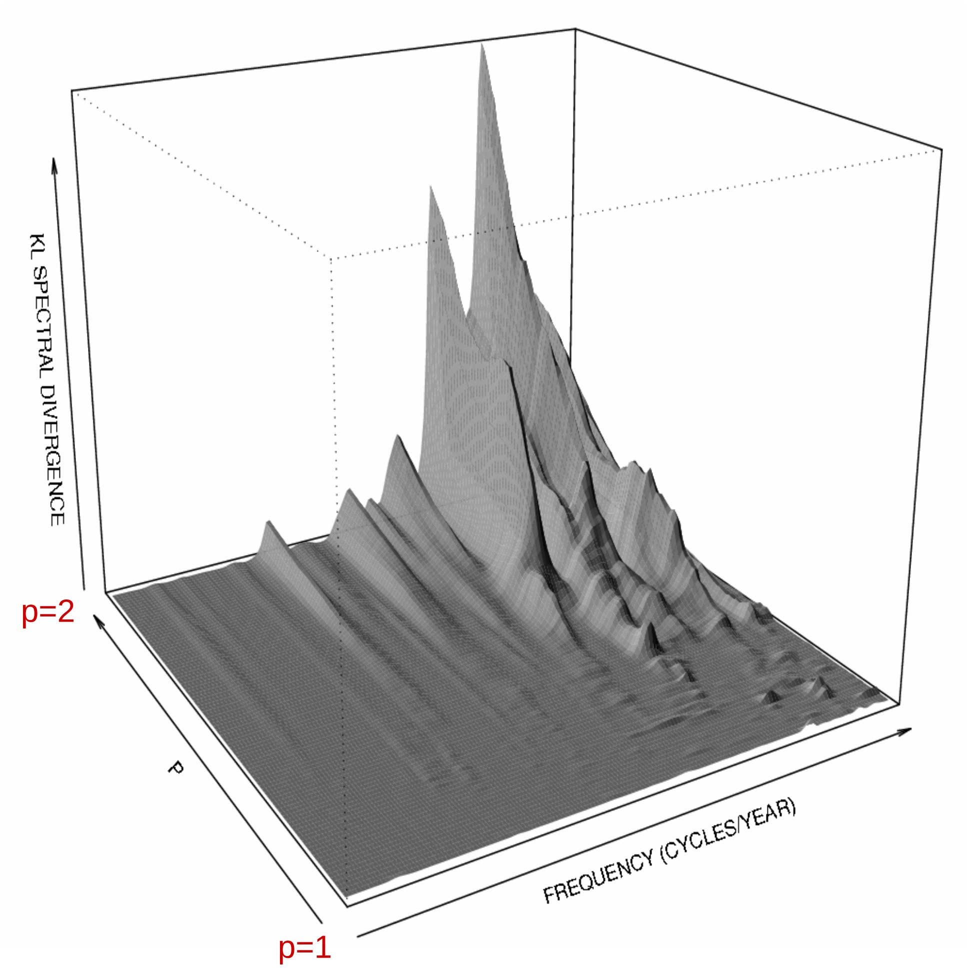




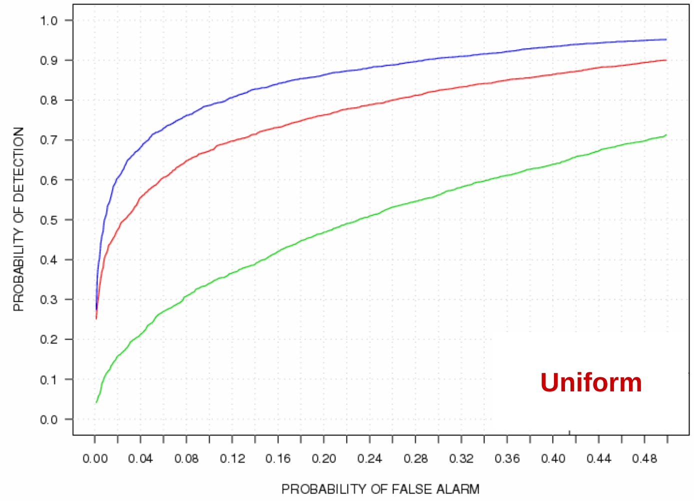
Sensitivity to Outliers

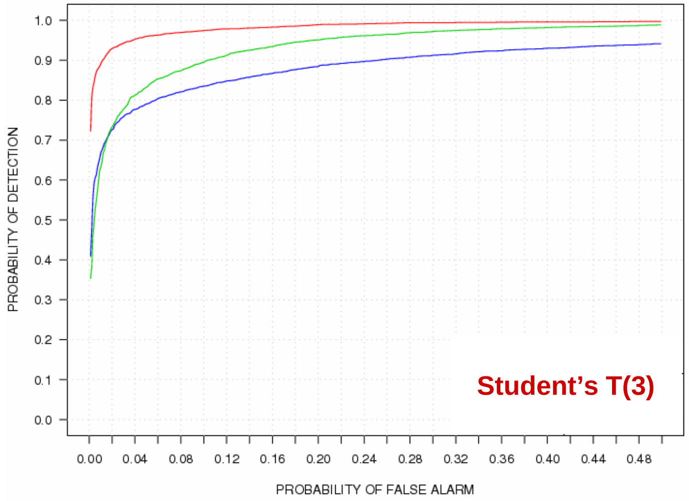


Spectral Distortion due to a Single Outlier

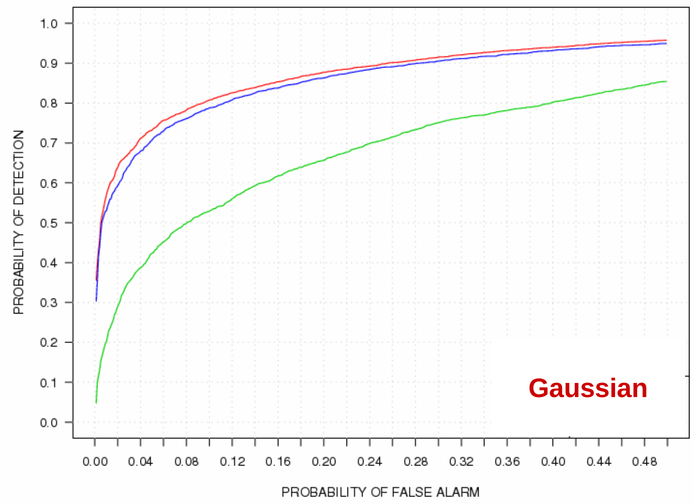


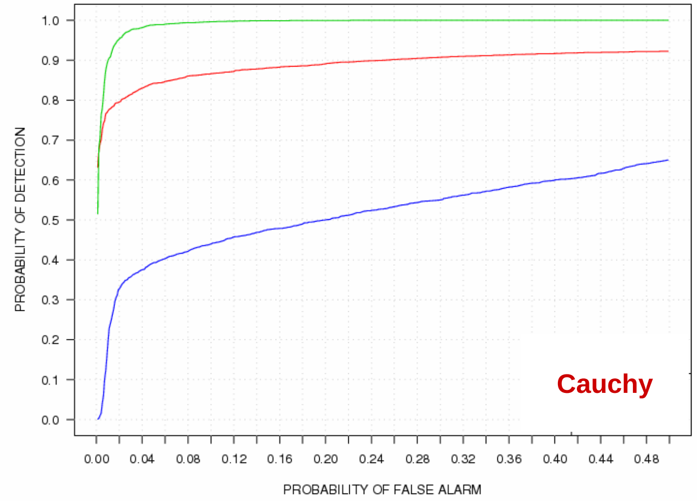
Tradeoff: Robustness vs. Efficiency for Periodicity Detection

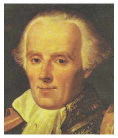




🟦





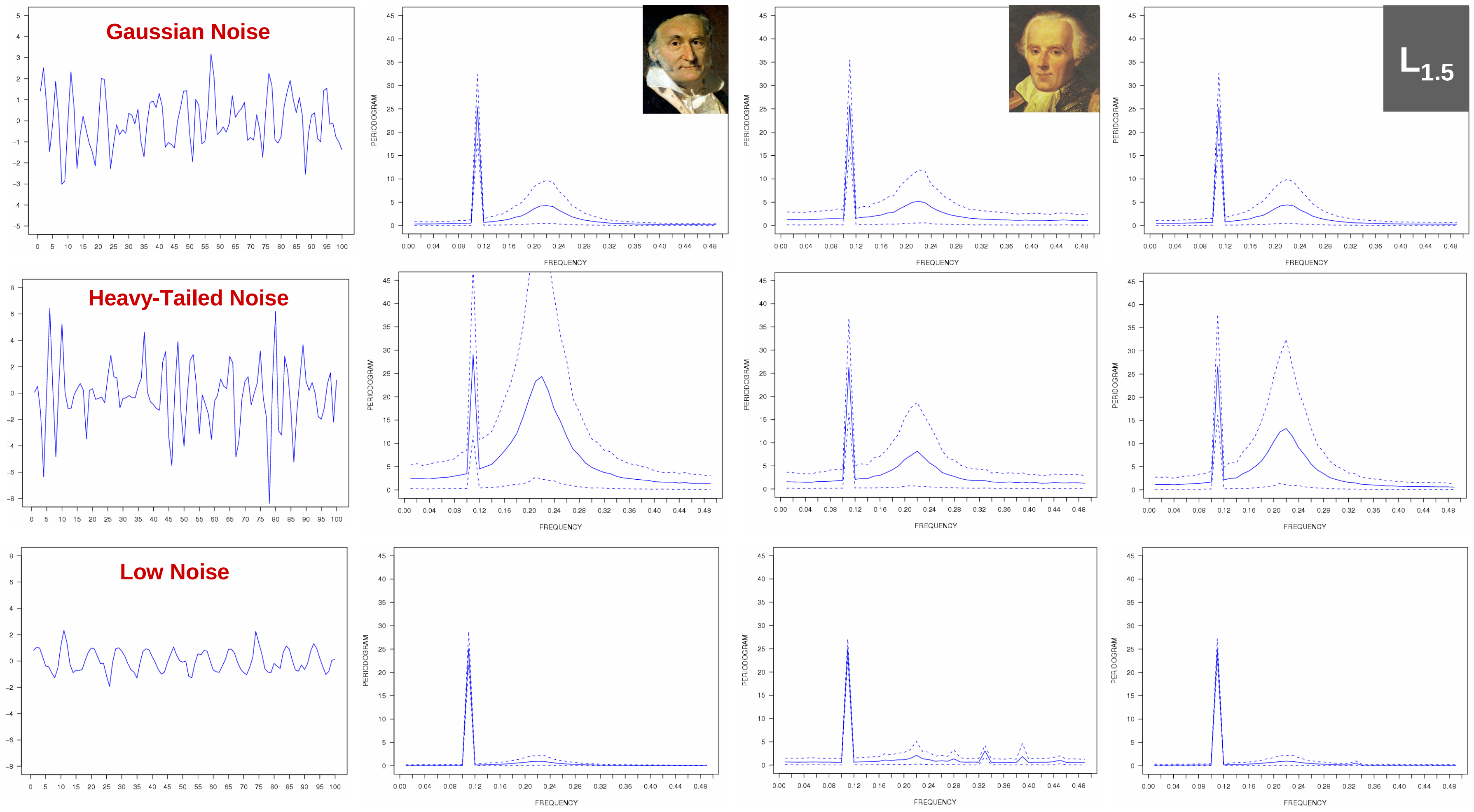


🟩

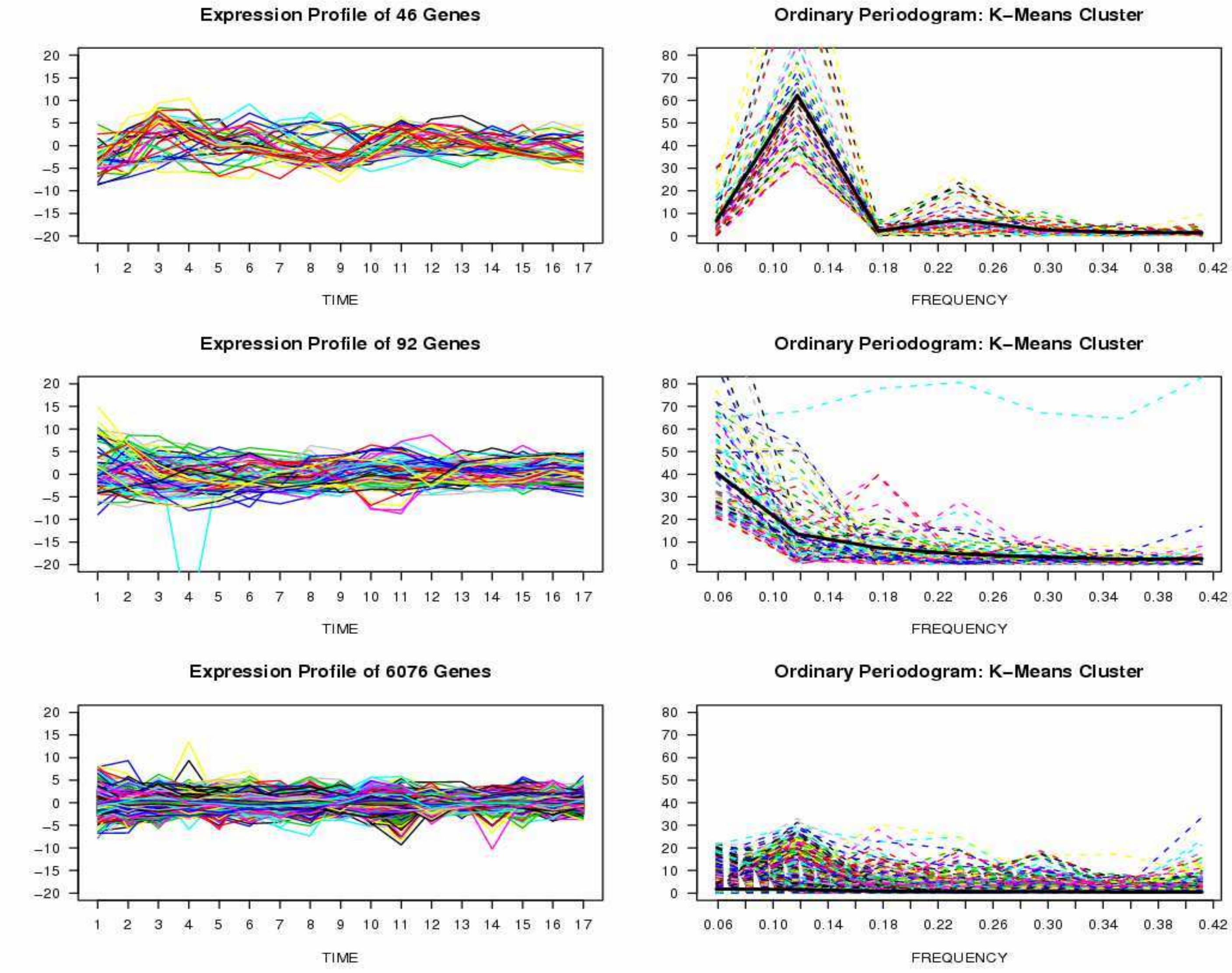
$L_{1.5}$

🟥

Tradeoff: Robustness vs. Leakage of Line Spectrum for Mixed Spectrum Analysis



Clustering of Gene Expression Profiles



Expression Profile of 46 Genes

Ordinary Periodogram: K–Means Cluster

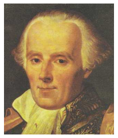
Expression Profile of 92 Genes

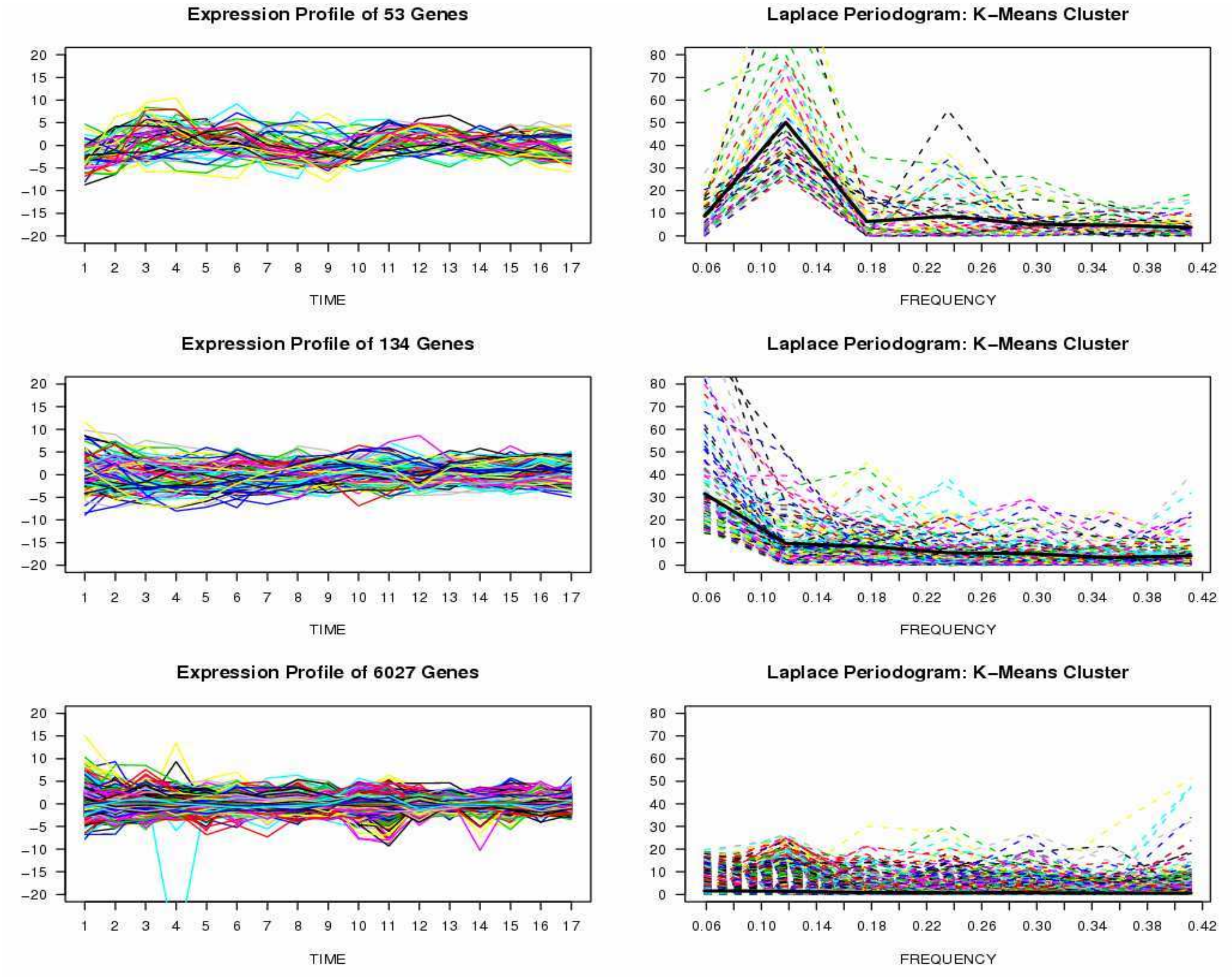
Ordinary Periodogram: K–Means Cluster

Expression Profile of 6076 Genes

Ordinary Periodogram: K–Means Cluster

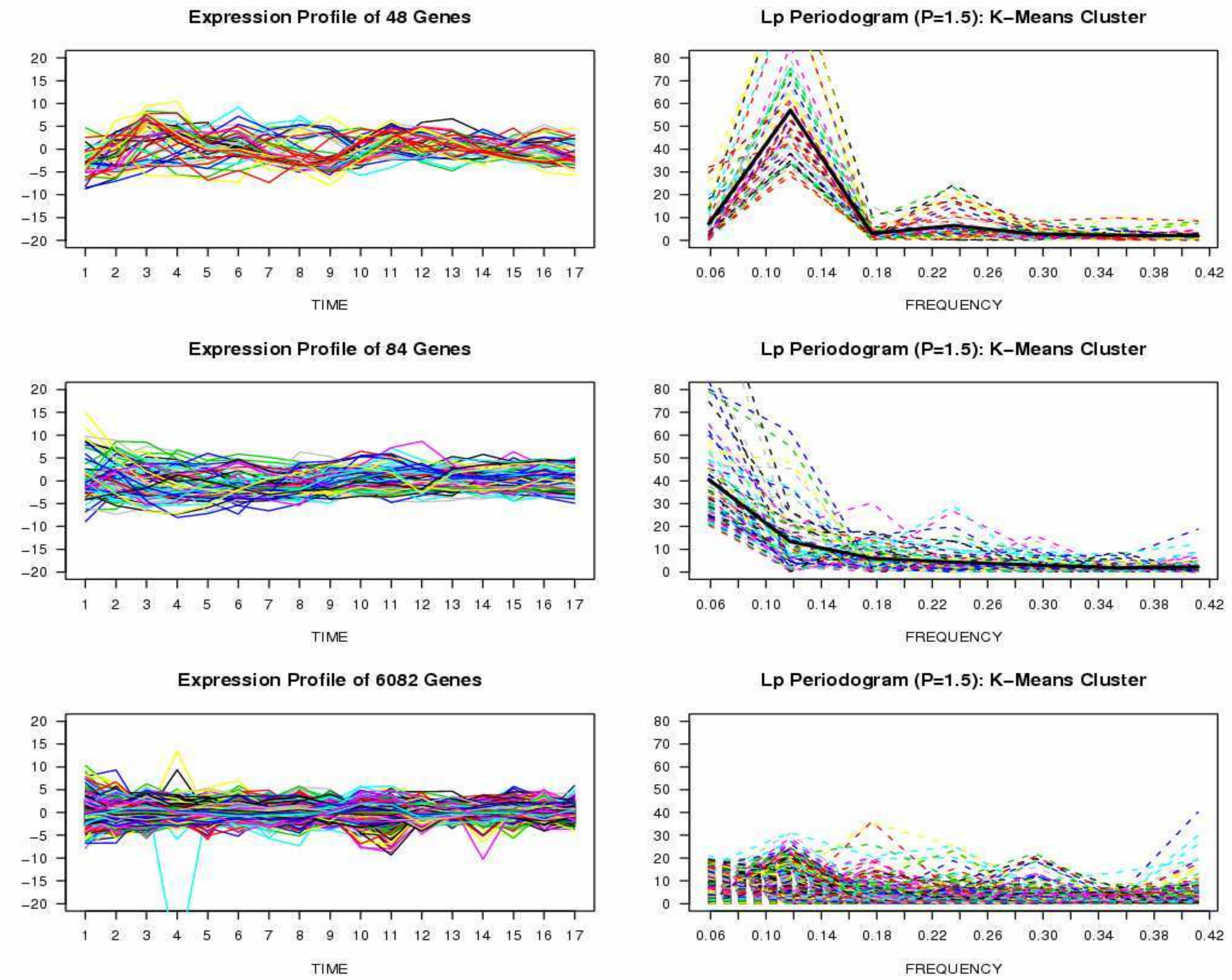
Clustering of Gene Expression Profiles



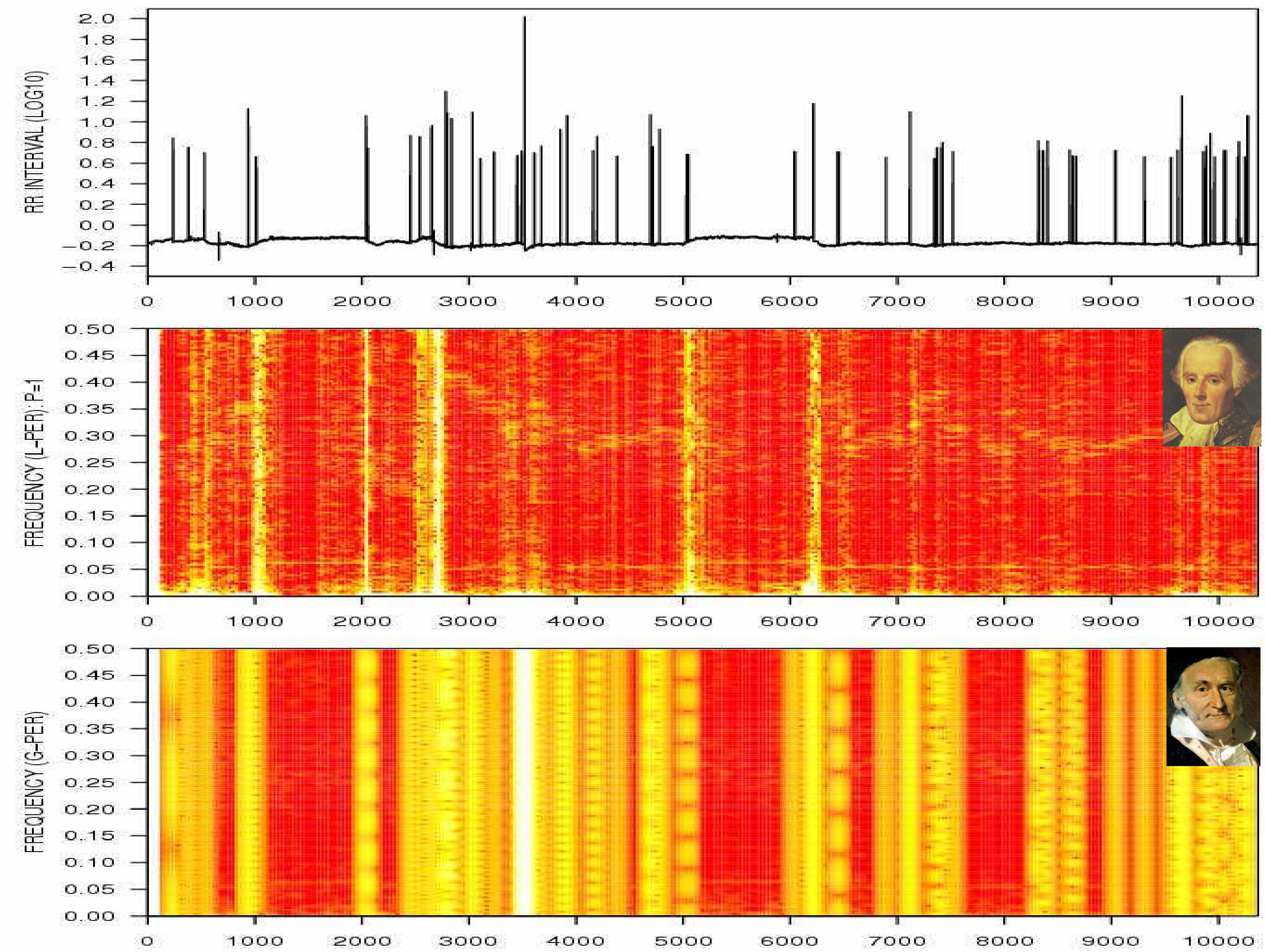

Expression Profile of 53 Genes
Laplace Periodogram: K–Means Cluster
Expression Profile of 134 Genes
Laplace Periodogram: K–Means Cluster
Expression Profile of 6027 Genes
Laplace Periodogram: K–Means Cluster
20
15
10
5
0
−5
−10
−15
−20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
TIME
80
70
60
50
40
30
20
10
0
0.06 0.10 0.14 0.18 0.22 0.26 0.30 0.34 0.38 0.42
FREQUENCY

Clustering of Gene Expression Profiles

$L_{1.5}$



Time-Frequency Analysis of Heart Rate Variability



Summary

- Obtained Laplace periodogram by replacing the L_2 norm (least squares) criterion with the L_1 norm (least absolute deviations) criterion in the linear harmonic regression
- Demonstrated robustness to outliers and nonlinear distortion
- Established the link to the zero-crossing spectrum
- Generalized to Lp norm periodogram
- Choice of p depends on the desirable performance tradeoffs in practice
  - Robustness, efficiency, leakage of line spectrum, computation, etc.
- Future work
  - Theory for long-range dependent processes
  - Parametric and nonparametric inference based on Laplace and Lp-Norm periodograms (Whittle’s pseudo likelihood)
  - FFT-like fast algorithms

Thank You!

Estimation of Death Rates in U.S. States With Small Subpopulations

Anastasia Voulgaraki[2], Rong Wei[1] and Benjamin Kedem[2]

[1]National Center for Health Statistics, Hyattsville, MD 20782
[2]University of Maryland, College Park, MD 20742

July 30, 2009

Introduction.

- NCHS publishes sex- and race-specific state decennial life tables for all states and DC.
- Problems arise in States with small subpopulations.
- In States with small subpopulations, observed mortality rates are often interrupted by gaps of zero death observations, especially in young ages.
- Mortality rates (death rates) are reported routinely in a log scale.
- Problem #1: How do we treat the zero values?

Introduction.

- NCHS publishes sex- and race-specific state decennial life tables for all states and DC.
- Problems arise in States with small subpopulations.
- In States with small subpopulations, observed mortality rates are often interrupted by gaps of zero death observations, especially in young ages.
- Mortality rates (death rates) are reported routinely in a log scale.
- Problem #1: How do we treat the zero values?

Introduction.

- NCHS publishes sex- and race-specific state decennial life tables for all states and DC.
- Problems arise in States with small subpopulations.
- In States with small subpopulations, observed mortality rates are often interrupted by gaps of zero death observations, especially in young ages.
- Mortality rates (death rates) are reported routinely in a log scale.
- Problem #1: How do we treat the zero values?

Introduction.

- NCHS publishes sex- and race-specific state decennial life tables for all states and DC.
- Problems arise in States with small subpopulations.
- In States with small subpopulations, observed mortality rates are often interrupted by gaps of zero death observations, especially in young ages.
- Mortality rates (death rates) are reported routinely in a log scale.
- Problem #1: How do we treat the zero values?

Introduction.

- NCHS publishes sex- and race-specific state decennial life tables for all states and DC.
- Problems arise in States with small subpopulations.
- In States with small subpopulations, observed mortality rates are often interrupted by gaps of zero death observations, especially in young ages.
- Mortality rates (death rates) are reported routinely in a log scale.
- Problem #1: How do we treat the zero values?

Introduction.

- In the process of generating the life tables, age-specific mortality rates are estimated and smoothed.
- One way to smooth mortality data is by using the Heligman-Pollard parametric model.
- Problem #2: Parametric models often use data on logarithmic scale.
- Problem #3: Small populations raise concerns regarding reliability of their mortality rate estimates and fidelity of mortality patterns after smoothing.
- In one fifth of the states, life tables were not published for some subpopulations previously due to small size of population.

Introduction.

- In the process of generating the life tables, age-specific mortality rates are estimated and smoothed.
- One way to smooth mortality data is by using the Heligman-Pollard parametric model.
- Problem #2: Parametric models often use data on logarithmic scale.
- Problem #3: Small populations raise concerns regarding reliability of their mortality rate estimates and fidelity of mortality patterns after smoothing.
- In one fifth of the states, life tables were not published for some subpopulations previously due to small size of population.

Introduction.

- In the process of generating the life tables, age-specific mortality rates are estimated and smoothed.
- One way to smooth mortality data is by using the Heligman-Pollard parametric model.
- Problem #2: Parametric models often use data on logarithmic scale.
- Problem #3: Small populations raise concerns regarding reliability of their mortality rate estimates and fidelity of mortality patterns after smoothing.
- In one fifth of the states, life tables were not published for some subpopulations previously due to small size of population.

Introduction.

- In the process of generating the life tables, age-specific mortality rates are estimated and smoothed.
- One way to smooth mortality data is by using the Heligman-Pollard parametric model.
- Problem #2: Parametric models often use data on logarithmic scale.
- Problem #3: Small populations raise concerns regarding reliability of their mortality rate estimates and fidelity of mortality patterns after smoothing.
- In one fifth of the states, life tables were not published for some subpopulations previously due to small size of population.

Introduction.

- In the process of generating the life tables, age-specific mortality rates are estimated and smoothed.
- One way to smooth mortality data is by using the Heligman-Pollard parametric model.
- Problem #2: Parametric models often use data on logarithmic scale.
- Problem #3: Small populations raise concerns regarding reliability of their mortality rate estimates and fidelity of mortality patterns after smoothing.
- In one fifth of the states, life tables were not published for some subpopulations previously due to small size of population.

Example: Black Females, CA, 2000.

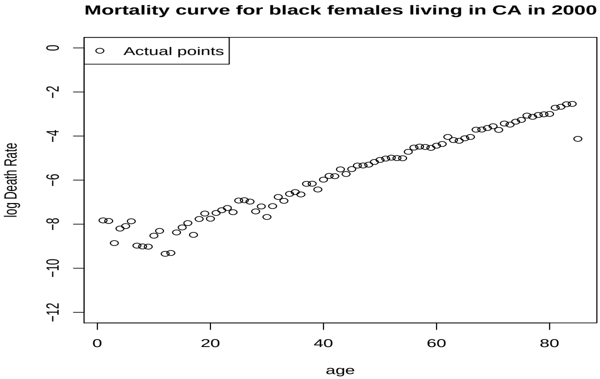


Figure 1: Observed non-zero log deathrates for black females living in California in 2000. Total Number of deaths: 6426, ages 1-84.

Example: Black Females, NV, 2000.

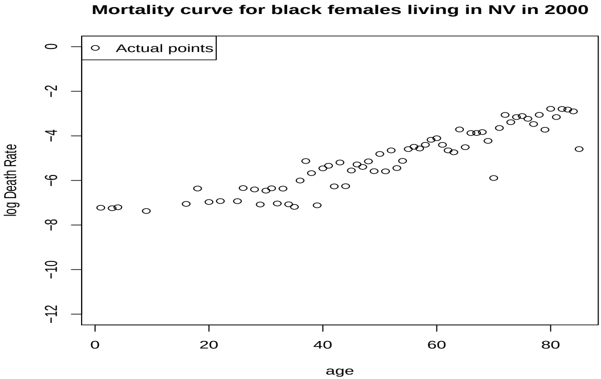


Figure 2: Observed non-zero log deathrates for black females living in Nevada in 2000. Total Number of deaths: 318, ages 1-84

Example: Black Females, NV, 2000.

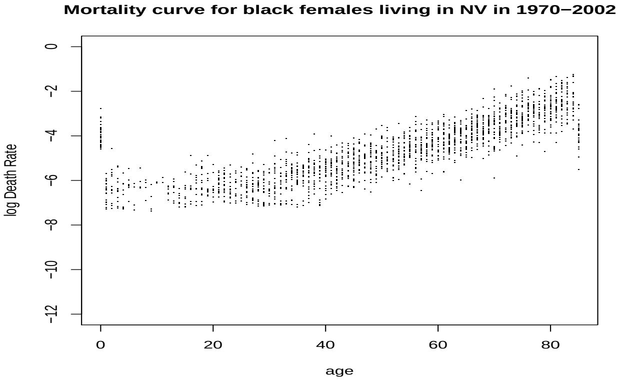


Figure 3: Observed non-zero log deathrates for black females living in Nevada in 1970-2002.

Introduction.

- Fact: As a biological feature of the human population, the age specific death rates should be continuous and nonzero without the interruption of a zero death rate.
- Problem: Reliable estimation of zero death rates in states with small population size.
- Solution: Fit appropriate probability models supported discretely at zero. Replace zero observations with expected values.

Outline of Part I

1. Theory and Methods
   - Main Idea: Mixed Distribution
   - Model 1: Mixed Lognormal Distribution
   - Model 2: Hurdle Model
   - Model 3: Zero-Inflated Model
   - Model 4: Poisson Regression
   - Construction of Confidence Intervals

Outline of Part II

2 Data Analysis

- Data and Selection of Models
- Example: Black Females, CA, 2000
- Example: Black Females, NV, 2000
- Model Comparison
- Smoothing the data: The Heligman-Pollard model
- Conclusions

Part I

Theory and Methods

Main Idea: Mixed Distribution.

Let Y be the variable of interest. A natural model for Y is a mixed distribution probability model:

$$Y = \begin{cases} 0, & \textit{with probability } 1-p \\ F(y,\theta_1), & \textit{with probability } p \end{cases}$$

Then,

$$P(Y \leq y) = G_m(y; p, \theta_1) = (1-p)H(y) + pF(y;\theta_1) \qquad (1)$$

where $H(y)$ is a step function:

$$H(y) = \begin{cases} 0, & y < 0 \\ 1, & y \geq 0 \end{cases}$$

Main Idea: Mixed Distribution.

The corresponding generalized pdf is:

$$g_m(y; p, \boldsymbol{\theta_1}) = (1-p)^{1-I[y>0]}[pf(y;\theta_1)]^{I[y>0]},\ y \geq 0 \qquad (2)$$

where $f(y;\theta_1)$ is a probability density function conditional on $Y > 0$ and corresponding to $F(y;\theta_1)$, and $I(A)$ is the indicator of the event A.
The goal is to estimate

$$\mathrm{E}(Y) = p\mathrm{E}(Y \mid Y > 0) \qquad (3)$$

Model 1: Mixed Lognormal Distribution.

Let Y denote death rate. The continuous part of the distribution of death rate is lognormal $LN(\mu, \sigma^2)$, with density,

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma y} \exp\{-(\log y - \mu)^2/(2\sigma^2)\}, \quad y > 0 \qquad (4)$$

The mean of Y is:

$$\mathrm{E}(Y) = p \exp\{\mu + \sigma^2/2\}, \qquad (5)$$

Using the maximum likelihood we can estimate (5).

Estimation of the Variance for Mixed Lognormal Distribution.

The Var.-Cov. matrix of estimates is obtained from $\boldsymbol{I}_f^{-1}$:

$$\boldsymbol{I}_f = -\mathrm{E}\left(\frac{\partial^2 \log g(y,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right) \tag{6}$$

Then

$$Var(\hat{p}) = p(1-p)/n,\ \ Var(\hat{\mu}) = \sigma^2/(np),\ \ Var(\hat{\sigma}) = \sigma^2/(2np)$$

For the mixed lognormal,

$$\mathrm{Var}(\hat{\mathrm{E}}(Y)) \approx \frac{1}{n}\exp(2\mu+\sigma^2)[p(1-p)+p\sigma^2+p\sigma^4/2] \tag{7}$$

Then approximate 95% confidence intervals can be calculated as

$$\hat{\mathrm{E}}(Y) \pm 1.96 \cdot \sqrt{\hat{\mathrm{Var}}(\hat{\mathrm{E}}(Y))}.$$

Model 2: The Hurdle Model.

Let Y denote the number of deaths. Hurdle models are two-component models:

$$f_{hurdle}(y;\mu) = \begin{cases} 1-p, & y=0 \\ p \cdot f_{count}(y,\mu)/(1-f_{count}(0,\mu)), & y>0 \end{cases} \tag{8}$$

If the count component is modeled as zero-truncated Poisson(μ), then

$$\mathrm{E}(Y) = \frac{p\mu}{1-\exp(-\mu)}$$

Estimate μ using a GLM model.
Estimate p using a binomial GLM model.

Estimation of the Variance for Hurdle Model.

It can be shown that:

$$\mathrm{Var}(Y) = \mathrm{P}(Y > 0)\mathrm{Var}(Y \mid Y > 0) + \mathrm{P}(Y > 0)(1 - \mathrm{P}(Y > 0))(\mathrm{E}(Y \mid Y > 0))^2. \quad (9)$$

If the count component is modeled as zero-truncated Poisson(μ):

$$\mathrm{Var}(Y) = p\left[\frac{2\mu^2}{1 - e^{-\mu}} - \frac{\mu^2}{(1 - e^{-\mu})^2}\right] + p(1 - p)\left(\frac{\mu}{1 - e^{-\mu}}\right)^2 \quad (10)$$

Then an approximate 95% confidence interval for the mean number of deaths for a given age and year is

$$\hat{Y} \pm 1.96\sqrt{\hat{\mathrm{Var}}(Y)}.$$

Model 3: The Zero-Inflated Model.

- Zero inflated models are two-component models where the count distribution is supported at zero as well.
- The probability of a zero is constructed from two sources:
  1. The point mass at zero
  2. The count distribution.

The distribution of the number of deaths Y is modeled as

$$f_{zeroinfl}(y;\mu) = \pi I_{\{0\}}(y) + (1-\pi) f_{count}(y;\mu), \quad y = 0, 1, 2, \ldots, \tag{11}$$

where $\pi \equiv 1 - p$ is the unobserved probability of belonging to the point mass component.

Model 3: The Zero-Inflated Model.

π can be modeled using binomial GLM.
The count component can be modeled as Poisson GLM. Then

$$\mathrm{E}(Y) = p\mu.$$

If the count component is Poisson distribution with mean μ then directly from (9) the approximate 95% confidence interval for the mean number of deaths is

$$\hat{Y} \pm 1.96\sqrt{\hat{p}\hat{\mu} + \hat{p}(1-\hat{p})\hat{\mu}^2}.$$

Model 4: Poisson Regression.

We assume that Y, the number of deaths, follows a Poisson(μ):

$$f(y;\mu) = \frac{\exp(-\mu)\mu^y}{y!}, \quad y = 0, 1, 2 \ldots$$

Then $E(Y) = \mu$.

- Estimate μ by using Poisson GLM with log link.
- Estimate $Var(\hat{\mu})$ using the Fisher Information matrix.
- If there is overdispersion in the data, then we assume that Y follows a negative binomial distribution.

Parametric Bootstrap.

Parametric Bootstrap was used to construct 95% Confidence Intervals on $\log(\hat{\mathrm{E}}(y))$.

1. Create a sample of size n from the density $f_{\hat{\theta}}$. Using the generated sample, calculate the maximum likelihood estimator $\tilde{\theta}$ of θ and estimate $\mathrm{E}_{\theta}(y)$ by $\tilde{\mathrm{E}}_{\tilde{\theta}}(y)$. Calculate $\log(\tilde{\mathrm{E}}_{\tilde{\theta}}(y))$.
2. Repeat Step 1, B times.
3. Calculate the sample variance, s^2, of the $\log(\tilde{\mathrm{E}}_{\tilde{\theta}}(y))$ estimators.
4. Using the sample variance, s^2, we can easily compute 95% confidence interval for $\log(\hat{\mathrm{E}}(y))$.

For this problem: $B = 1000$, $n = 33$

Part II

Data Analysis

Data and Selection of Models.

We used data from NCHS public-use mortality files

- Other Females, California, 1970-1998.
- Black Females, California, Iowa, Minnesota, Nevada, New Mexico, Nebraska, Oregon, and Rhode Island, 1970-2002.
- Black Males, Iowa, Minnesota, Nevada, New Mexico, Nebraska, Oregon, and Rhode Island, 1970-2002.
- Variables available: population size, number of deaths, and death rate for each year and age combination.

Selection and Comparison of Models: AIC ,BIC, RMSE, MAE for Poisson, hurdle and zero-inflated models.

Data and Selection of Models.

Root Mean Square Error (RMSE):

$$\sqrt{\frac{\sum_{ij}(deathrate - deat\hat{h}rate)^2}{n}} \tag{12}$$

Mean Absolute Error (MAE):

$$\frac{\sum_{ij} \mid deathrate - deat\hat{h}rate \mid}{n}, \tag{13}$$

where $i = 1 \ldots 84$ and $j = 1970 \ldots 2002$.

Example: Black Females, CA, 2000.

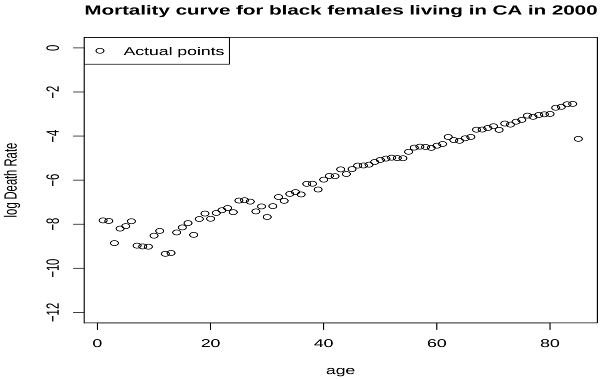


Figure 4: Observed non-zero log deathrates for black females living in California in 2000.

Example: Black Females, CA, 2000.

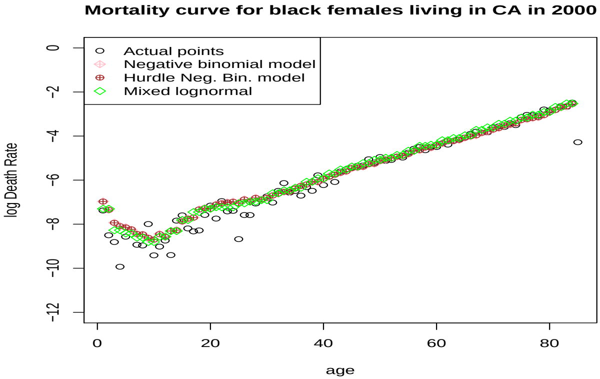


Figure 5: Model comparison for black females living in California in 2000.

Example: Black Females, NV, 2000.

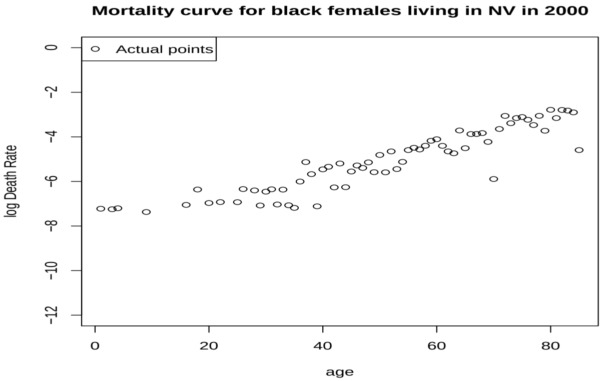


Figure 6: Observed non-zero log deathrates for black females living in Nevada in 2000.

Example: Black Females, NV, 2000.

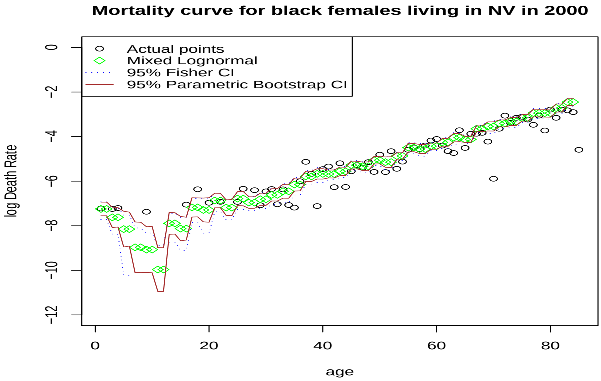


Figure 7: Mixed Lognormal model: Black females, Nevada, 2000.

Example: Black Females, NV, 2000.

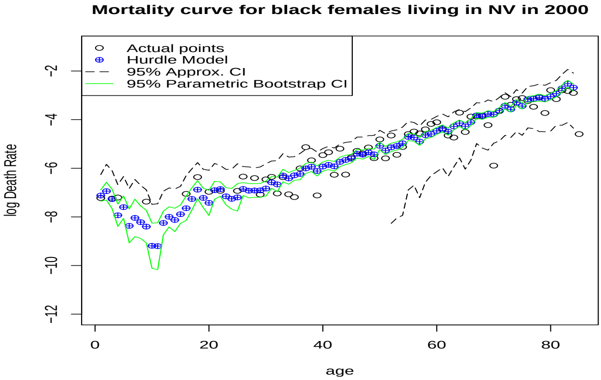


Figure 8: Hurdle model: Black females, Nevada, 2000.

Example: Black Females, NV, 2000.

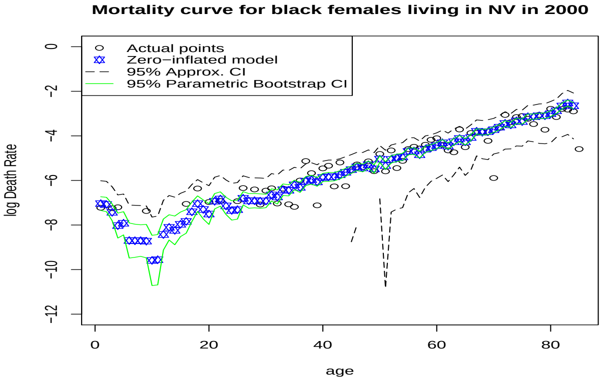


Figure 9: Zero-Inflated model: Black females, Nevada, 2000.

Example: Black Females, NV, 2000.

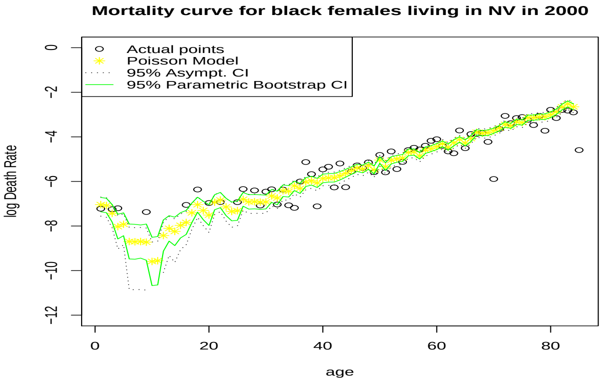


Figure 10: Poisson model: Black females, Nevada, 2000.

Example: Black Females, NV, 2000.

Age	Mixed Lognormal	Poisson	Hurdle	Zero-Inflated
2	-7.245430	-7.080092	-6.941776	-7.079647
5	-8.149839	-7.922335	-7.598458	-7.921949
6	-8.149839	-8.695307	-8.368559	-8.695012
7	-8.967177	-8.706949	-8.040790	-8.706646
8	-8.967177	-8.697627	-8.215434	-8.697402
10	-9.069759	-9.589797	-9.188796	-9.589588
11	-9.964206	-9.558039	-9.201704	-9.557741
12	-9.964206	-8.427220	-8.249543	-8.426834
13	-7.892806	-8.106477	-7.995877	-8.106061
14	-7.892806	-8.241683	-8.125688	-8.241226
15	-8.127934	-7.970045	-7.889547	-7.969611
17	-7.176140	-7.408133	-7.262139	-7.407598
19	-7.290718	-7.303220	-7.219202	-7.302631
21	-6.867852	-6.933962	-6.901483	-6.933353
23	-7.185431	-7.144202	-7.152747	-7.143573
24	-7.185431	-7.347249	-7.256396	-7.346656
27	-6.959963	-6.918943	-6.922451	-6.918379

Table 1: Estimated expected values of log(death rates) provided by the different models for black females living in NV, 2000.

Example: Black Females, NV, 2000.

RMSE for NV	Total	1-30 yrs	31-50 yrs	51-84 yrs
Mixed lognormal	*NA*	*NA*	3.582055	31.44193
Poisson	20.16510	1.143711	3.511821	31.56279
Hurdle Model	19.90017	1.139371	3.48776	31.14633
Zero-inflated Model	19.89754	1.139346	3.487707	31.14218

MAE for NV	Total	1-30 yrs	31-50 yrs	51-84 yrs
Mixed lognormal	*NA*	*NA*	2.559315	18.61506
Poisson	8.349927	0.8000284	2.535445	18.43189
Hurdle Model	8.233534	0.7848666	2.502147	18.17729
Zero-inflated Model	8.233423	0.7848975	2.502137	18.17700

Table 2: Nevada RMSE and MAE. Black female, age 1-84, period 1970-2002. Entries are multiples of 10^{-3}

Model Comparison

- The models give close estimates in all states and capture the pointed hook pattern for lower ages.
- In mixed lognormal, if the samples for some ages consist only of zeros, then the model cannot produce estimates for these particular ages.
- For a small to medium number of zeros the simpler models (mixed lognormal, poisson) perform sufficiently well.
- For a large number of zeros, hurdle and zero inflated models may be more appropriate.
- Poisson vs. hurdle and zero-inflated models.

Model Comparison

- The models give close estimates in all states and capture the pointed hook pattern for lower ages.
- In mixed lognormal, if the samples for some ages consist only of zeros, then the model cannot produce estimates for these particular ages.
- For a small to medium number of zeros the simpler models (mixed lognormal, poisson) perform sufficiently well.
- For a large number of zeros, hurdle and zero inflated models may be more appropriate.
- Poisson vs. hurdle and zero-inflated models.

Model Comparison

- The models give close estimates in all states and capture the pointed hook pattern for lower ages.
- In mixed lognormal, if the samples for some ages consist only of zeros, then the model cannot produce estimates for these particular ages.
- For a small to medium number of zeros the simpler models (mixed lognormal, poisson) perform sufficiently well.
- For a large number of zeros, hurdle and zero inflated models may be more appropriate.
- Poisson vs. hurdle and zero-inflated models.

Model Comparison

- The models give close estimates in all states and capture the pointed hook pattern for lower ages.
- In mixed lognormal, if the samples for some ages consist only of zeros, then the model cannot produce estimates for these particular ages.
- For a small to medium number of zeros the simpler models (mixed lognormal, poisson) perform sufficiently well.
- For a large number of zeros, hurdle and zero inflated models may be more appropriate.
- Poisson vs. hurdle and zero-inflated models.

Model Comparison

- The models give close estimates in all states and capture the pointed hook pattern for lower ages.
- In mixed lognormal, if the samples for some ages consist only of zeros, then the model cannot produce estimates for these particular ages.
- For a small to medium number of zeros the simpler models (mixed lognormal, poisson) perform sufficiently well.
- For a large number of zeros, hurdle and zero inflated models may be more appropriate.
- Poisson vs. hurdle and zero-inflated models.

Smoothing the data: The Heligman-Pollard model

If q_x is the mortality rate for a person aged x exactly, then

$$\frac{q_x}{1-q_x} = A^{(x+B)^C} + De^{-E(\log x - \log F)^2} + GH^x, \qquad (14)$$

The values of the parameters can be estimated by least squares using Gauss-Newton iteration:

$$S^2 \;=\; \sum_x (\log(q_x) - \log(\dot{q}_x))^2 \qquad (15)$$

where q_x at age x is given by (14) and $\dot{q}_x$ is a mixture of the observed and estimated mortality rates.

Smoothing the data: The Heligman-Pollard model

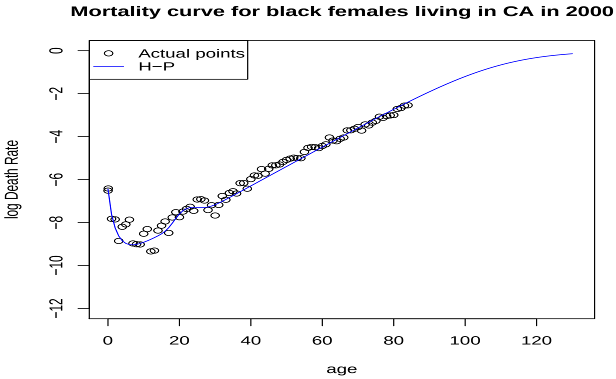


Figure 11: Fitted H-P curve: Black females, California, 2000.

Smoothing the data: The Heligman-Pollard model

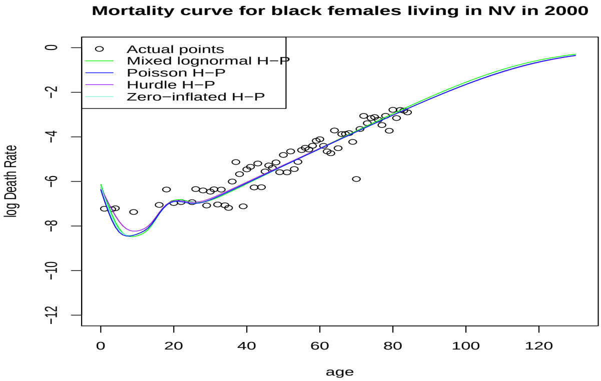


Figure 12: Comparison of H-P curves: Black females, Nevada, 2000.

Conclusions

Advantages of fitting the H-P model:

1. Smooth data
2. Extrapolate death rates in elder ages, i.e. 80+, where the reported ages being considered are not reliable.
3. Continuous interpolation of death rates between age intervals

Conclusions

Advantages of fitting the H-P model:

1. Smooth data
2. Extrapolate death rates in elder ages, i.e. 80+, where the reported ages being considered are not reliable.
3. Continuous interpolation of death rates between age intervals

Conclusions

Advantages of fitting the H-P model:

1. Smooth data
2. Extrapolate death rates in elder ages, i.e. 80+, where the reported ages being considered are not reliable.
3. Continuous interpolation of death rates between age intervals

Conclusions

We recommend a two-stage estimating/smoothing procedure:

1. Apply a suitable probability model on the data to get an estimate of the zero mortality rates.
2. Apply the Heligman-Pollard equation on a mixture of the estimated and actual data to obtain parameter estimates and smooth the mortality curve, covering the whole life span.

This procedure

1. permits more efficient, repeatable, and comparable results in generating US life tables.
2. allows for the publication of more life tables even for states with very small subpopulations.

Conclusions

We recommend a two-stage estimating/smoothing procedure:

1. Apply a suitable probability model on the data to get an estimate of the zero mortality rates.
2. Apply the Heligman-Pollard equation on a mixture of the estimated and actual data to obtain parameter estimates and smooth the mortality curve, covering the whole life span.

This procedure

1. permits more efficient, repeatable, and comparable results in generating US life tables.
2. allows for the publication of more life tables even for states with very small subpopulations.

Conclusions

We recommend a two-stage estimating/smoothing procedure:

1. Apply a suitable probability model on the data to get an estimate of the zero mortality rates.
2. Apply the Heligman-Pollard equation on a mixture of the estimated and actual data to obtain parameter estimates and smooth the mortality curve, covering the whole life span.

This procedure

1. permits more efficient, repeatable, and comparable results in generating US life tables.
2. allows for the publication of more life tables even for states with very small subpopulations.

Conclusions

We recommend a two-stage estimating/smoothing procedure:

1. Apply a suitable probability model on the data to get an estimate of the zero mortality rates.
2. Apply the Heligman-Pollard equation on a mixture of the estimated and actual data to obtain parameter estimates and smooth the mortality curve, covering the whole life span.

This procedure

1. permits more efficient, repeatable, and comparable results in generating US life tables.
2. allows for the publication of more life tables even for states with very small subpopulations.

References. I

1. Aitchison, J. (1955). On the distribution of a positive random variable having discrete probability mass at the origin. em J. American Statistical Association, 50, 901-908.

2. Aitchison, J. and Brown, J.A.C. (1963). The lognormal distribution. *Cambridge University Press, Cambridge, UK.*

3. Cameron, C.A. and Trivedi, P.K. (1998). Regression analysis of count data. *Cambridge University Press, UK.*

4. Curtin, L.R. (1983). Reliability considerations for state decennial life tables. *1983 Proceeding of the Social Statistics section of the American Statistical Associaion*, 161-166.

5. Heligman, L. and Pollard, J. H. (1980). The age pattern of mortality. *J. Inst. Actuaries*, 107, Part I, 659-671.

6. Kedem, B., Chiu, L.S., and North, G. (1990). Estimation of mean rain rate: Application to satellite observations. *J. Geophysical Res.*, 95, 1965-1972.

7. Kedem, B., and Fokianos, K. (2002). *Regression Models for Time Series Analysis*, Wiley, Hoboken.

8. Kedem, B., Pfeiffer, R., and Short, D.A. (1997). Variability of space-time mean rain rate. *J. Appl. Meteorology*, 36, 443-451.

9. Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1-14.

References. II

10 Mullahy, J. (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *J. Health Economics*, 17, 247-281.

11 Mullahy, J. (1986). Specification and testing of some modified count data models. *J. Econometrics*, 33, 341-365.

12 Panel on Nonstandard Mixtures of Distributions (1989). Statistical models and analysis in auditing. *Stat. Science*, 4, 2-33.

13 Rao, C.R. (1973). Linear Statistical Inference and its Applications. *Wiley, New York*.

14 Wei, R., Curtin, L. R. and Anderson R. (2003). Model US mortality data for building life tables and further studies. *2003 Joint Statistical Meetings of the American Statistical Association, Biometrics Section [CD-ROM], Alexandria, VA.*, American Statistical Association, 4458-4464.

15 Wei, R., Curtin, L. R., Anderson R. and Arias E. (2006). Smoothing state life tables based on small numbers of death. *2006 Joint Statistical Meetings of the American Statistical Association, Biometrics Section [CD-ROM], Alexandria, VA.*, American Statistical Association, 2650-2656.

16 Zeileis, A., Kleiber, C., and Jackman S. (2007). Regression models for count data in R. *Research Report Series / Department of Statistics and Mathematics.*, 53, Wien, Wirtshaftsuniv., 2007.

Odds Ratio Bias in Case-Control Studies Using Robust Genetic Models

Neal Jeffries
Office of Biostatistics Research
National Heart, Lung, and Blood Institute

7/30/2009

Outline

- Discuss case-control models for SNP data
- Robust models when inheritance uncertain
- Show bias of robust models
  - Why does bias matter?
- Examine method of moments approach
- Examine conditional likelihood approach
- Conclusions/generalizations

Case-Control/SNP Data Structure

- Very common genetic design- unrelated individuals
- SNP – Genetic information at 1 point in $3x10^9$ genome. At that location each person receives 1 SNP from mother & 1 from father
- Each SNP has 2 flavors (alleles) denoted A and B => AA, or AB, or BB genotypes
- Q: Is one allele more often associated with a particular disease?
  - Can elucidate disease/treatment pathways

Case-Control Design

- Each person has 1 of 3 genotypes AA, AB, BB
- Obtain r cases (affected individuals) and s controls (healthy individuals)

	AA	AB	BB	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

- Let p_0, p_1, p_2 denote genotype probabilities for cases, i.e. p_i = Prob(i "B" alleles||Case)
- Estimate p_i by r_i/r
- Let q_0, q_1, q_2 denote probabilities for controls
- H_0: $p_0= q_0$, $p_1= q_1$, $p_2= q_2$

Case–Control Design

	AA	AB	BB	Total
Cases	p_0	p_1	p_2	1
Controls	q_0	q_1	q_2	1

$$p_i = \Pr(i \text{ "B" alleles } \|\text{Case}),$$

$$q_i = \Pr(i \text{ "B" alleles } \|\text{Control})$$

$$\sum p_i = \sum q_i = 1$$

- Could use Pearson χ^2 2 df
- Other tests have more power
- Depends on mode of inheritance:
  - Dominant
    - BB risk=AB risk > AA risk
  - Recessive
    - BB risk> AB risk = AA risk
  - Additive ~ allele based test
    - BB risk> AB risk > AA risk
    - Compare on basis of alleles, not genotypes

Tests for Inheritance Mode

	AA	AB	BB	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

Dominance 2x2 table:
Risk AB = Risk BB

	Low Risk	High Risk
	AA	AB&BB
Cases	r_0	r_1+r_2
Controls	s_0	s_1+s_2

Additive 2x2 table:
Compare “A” risk to “B” risk

	Low Risk	High Risk
	A	B
Cases	$r_1 + 2r_0$	$2r_2 + r_1$
Controls	$s_1 + 2s_0$	$2s_2 + s_1$

Recessive 2x2 table:
Risk AB = Risk AA

	Low Risk	High Risk
	AA&AB	BB
Cases	$r_0 + r_1$	r_2
Controls	$s_0 + s_1$	s_2

Each 2x2 table leads to χ^2 1 df test – more powerful than χ^2 2 df test

Generic Test for 2x2 Table

	Low Risk	High Risk	Total
Cases	r_L	r_H	r
Controls	s_L	s_H	$s=r$

For simplicity assume $r = s$.

Let $\hat{p}_L = r_L/r$, $\hat{p}_H = r_H/r$, $\hat{q}_L = s_L/r$, $\hat{q}_H = s_H/r$. Then

$$\frac{\sqrt{r}\,(\log \hat{OR} - \log OR)}{\sqrt{(1/\hat{p}_L + 1/\hat{p}_H + 1/\hat{q}_L + 1/\hat{q}_H)}} \sim N(0,1) \text{ and}$$

$$Z = \frac{\sqrt{r}\log \hat{OR}}{\hat{\sigma}} \sim N(\mu, 1)$$

$$\mu = \frac{\sqrt{r}\log(p_H q_L / p_L q_H)}{\sigma}$$

Note: this formula is not exactly right for additive case, but close

Robust Genetic Tests

- If inheritance mode known, could choose powerful Z test, Z_D (dominant mode), Z_A (Additive mode), Z_R (recessive mode)
- Inheritance mode typically not known – use robust procedure
  - Compute all three test statistics
  - select inheritance mode, I, by most extreme test statistic among Z_D, Z_A, and Z_R
- What are statistical properties of selected extreme statistic?

Robust Genetic Tests

Define the random variable I as

$$\mathrm{I} = \begin{cases} \mathrm{D}, \ \text{if} & = |Z_{\mathrm{D}}| > \max\{|Z_{\mathrm{A}}|, |Z_{\mathrm{R}}|\}; \\ \mathrm{A}, \ \text{if} & = |Z_{\mathrm{A}}| > \max\{|Z_{\mathrm{D}}|, |Z_{\mathrm{R}}|\}; \\ \mathrm{R}, \ \text{if} & = |Z_{\mathrm{R}}| > \max\{|Z_{\mathrm{D}}|, |Z_{\mathrm{A}}|\}. \end{cases}$$

- Z_I is most extreme test statistic
  - Infer inheritance mode from I
  - Infer effect size from $\log \hat{OR}_{\mathrm{I}}$ or $\frac{\log \hat{OR}_{\mathrm{I}}}{\hat{\sigma}_{\mathrm{I}}}$
- I varies in repeated sampling, is random

Properties of Z_I

- Though Z_D, Z_A, and Z_R are normally distributed, Z_I is not under H_0 or H_A
- Computing appropriate p-values for Z_I under H_0 is non-trivial (Zheng 2009)
- By selecting most extreme statistic, underlying observed effect size is likely biased

$$\text{standardized bias} = \mathrm{E}\left[Z_I - \mu_I\right] = \mathrm{E}\left[\frac{\sqrt{r}\ln \hat{\mathrm{OR}}_I}{\hat{\sigma}_I} - \frac{\sqrt{r}\ln \mathrm{OR}_I}{\sigma_I}\right]$$

$$\text{where } \hat{\sigma}_I = \sqrt{(1/\hat{p}_L + 1/\hat{p}_H + 1/\hat{q}_L + 1/\hat{q}_H)} \text{ for 2x2 table}$$

Two potential sources of bias

- Ranking bias – bias arising from selecting most extreme/highest ranked statistic – aka "Winner's curse"

- Significance bias – bias arising from conditioning on finding a significant p–value (a.k.a. publication bias)

- Both potentially present in genetic studies – little treatment of ranking bias

Simulated data showing ranking and significance bias

- $r = s = 500$
- Simulate case & control cohorts with different true inheritance and effect sizes
- Examine ranking bias alone

$$\text{ranking bias} = \mathrm{E}\left[Z_I - \mu_I\right] = \mathrm{E}\left[\frac{\sqrt{r}\ln \hat{OR}_I}{\hat{\sigma}_I} - \frac{\sqrt{r}\ln OR_I}{\sigma_I}\right]$$

- Examine combined ranking and significance bias

$$\text{comb. bias} = \mathrm{E}\left[\frac{\sqrt{r}\ln \hat{OR}_I}{\hat{\sigma}_I} - \frac{\sqrt{r}\ln OR_I}{\sigma_I} \parallel p(Z_I) < 0.05\right]$$

Demonstration of Standardized Log Odds Bias

10,000 Simulations, Sample size = 500, Minor Allele Frequency = 40%, Disease Prevalence = 10%, alpha = 0.05

	Empirical Power	OR	True Model	Prob. Dom. is Chosen	Dom. Bias When Chosen	Prob. Add. is Chosen	Add. Bias When Chosen	Prob. Rec. is Chosen	Rec. Bias When Chosen	Empirical Bias
Ranking Bias	NA	1.8	Dominant	78%	0.02	22%	0.62	< 1%	-1.52	0.153
Combined Ranking and Significance Bias	98%	1.8	Dominant	78%	0.07	22%	0.66	0%	NA	0.200
Ranking Bias	NA	1.8	Additive	15%	0.22	71%	0.24	14%	0.51	0.274
Combined Ranking and Significance Bias	87%	1.8	Additive	13%	0.58	74%	0.42	13%	0.86	0.496
Ranking Bias	NA	1.8	Recessive	< 1%	-1.28	12%	0.66	88%	0.07	0.137
Combined Ranking and Significance Bias	93%	1.8	Recessive	< 1%	1.21	12%	0.86	88%	0.20	0.280
Ranking Bias	NA	1.4	Dominant	68%	0.13	27%	0.52	4%	-0.42	0.211
Combined Ranking and Significance Bias	63%	1.4	Dominant	70%	0.68	29%	1.03	1%	1.46	0.790
Ranking Bias	NA	1.4	Additive	28%	0.25	46%	0.47	26%	0.34	0.373
Combined Ranking and Significance Bias	42%	1.4	Additive	23%	1.30	57%	1.15	20%	1.48	1.252
Ranking Bias	NA	1.4	Recessive	8%	-0.60	21%	0.63	71%	0.23	0.246
Combined Ranking and Significance Bias	44%	1.4	Recessive	1%	1.37	22%	1.39	77%	0.99	1.080

1) Bias exists 2) Ranking Bias ≠ Significance Bias

Does bias matter ?

- Not if solely interested in significance and p-value correctly computed
- When does magnitude matter?
  - Basis for power analysis for follow-up study
    - Overestimation of effect size => underpowered study
  - May not have opportunity for follow-up study to definitively establish effect size
  - Comparisons across studies/ Replication of previous study
  - Confidence intervals
    - Overestimate shifts entire interval
  - Do results have influence on whether further investigation warranted
    - e.g. pilot study results must exceed some threshold before proceeding

Method of moments approach

Let $\underline{Z} = (Z_D, Z_A, Z_R)$ – has trivariate normal distribution $N(\underline{\mu}, \mathbf{\Sigma})$ where $\underline{\mu} = (\mu_D, \mu_A, \mu_R)$, $\mathbf{\Sigma}$ is correlation matrix.

e.g. $z_D = \dfrac{\sqrt{r}\ln \hat{OR}_D}{\hat{\sigma}_D}$ and $\mu_D = \dfrac{\sqrt{r}\ln OR_D}{\sqrt{1/p_a + 1/p_b + 1/p_c + 1/p_d}}$ where the p terms correspond to the true cell probabilities in an associated 2 × 2 table.

Of interest is $E[Z_I - \mu_I]$ where I denotes the mode with the most extreme test statistic.

Method of moments

Decompose bias $\mathsf{E}\left[\mathsf{Z}_{\mathbf{I}} - \mu_{\mathbf{I}}; \underline{\mu}, \boldsymbol{\Sigma}\right]$ as a mixture

$$\mathsf{E}\left[\mathsf{Z}_{\mathbf{I}} - \mu_{\mathbf{I}}; \underline{\mu}, \boldsymbol{\Sigma}\right] =$$

$$
\begin{gathered}
\mathsf{E}\left[\mathsf{Z}_{\mathsf{D}} - \mu_{\mathsf{D}} \| \left|\mathsf{Z}_{\mathsf{D}}\right| > \left|\mathsf{Z}_{\mathsf{A}}\right|, \left|\mathsf{Z}_{\mathsf{D}}\right| > \left|\mathsf{Z}_{\mathsf{R}}\right|; \underline{\mu}, \boldsymbol{\Sigma}\right] \times \\
\mathsf{P}\left[\left|\mathsf{Z}_{\mathsf{D}}\right| > \left|\mathsf{Z}_{\mathsf{A}}\right|, \left|\mathsf{Z}_{\mathsf{D}}\right| > \left|\mathsf{Z}_{\mathsf{R}}\right|; \underline{\mu}, \boldsymbol{\Sigma}\right] \\
+ \\
\mathsf{E}\left[\mathsf{Z}_{\mathsf{A}} - \mu_{\mathsf{A}} \| \left|\mathsf{Z}_{\mathsf{A}}\right| > \left|\mathsf{Z}_{\mathsf{D}}\right|, \left|\mathsf{Z}_{\mathsf{A}}\right| > \left|\mathsf{Z}_{\mathsf{R}}\right|; \underline{\mu}, \boldsymbol{\Sigma}\right] \times \\
\mathsf{P}\left[\left|\mathsf{Z}_{\mathsf{A}}\right| > \left|\mathsf{Z}_{\mathsf{D}}\right|, \left|\mathsf{Z}_{\mathsf{A}}\right| > \left|\mathsf{Z}_{\mathsf{R}}\right|; \underline{\mu}, \boldsymbol{\Sigma}\right] \\
+ \\
\mathsf{E}\left[\mathsf{Z}_{\mathsf{R}} - \mu_{\mathsf{R}} \| \left|\mathsf{Z}_{\mathsf{R}}\right| > \left|\mathsf{Z}_{\mathsf{D}}\right|, \left|\mathsf{Z}_{\mathsf{R}}\right| > \left|\mathsf{Z}_{\mathsf{A}}\right|; \underline{\mu}, \boldsymbol{\Sigma}\right] \times \\
\mathsf{P}\left[\left|\mathsf{Z}_{\mathsf{R}}\right| > \left|\mathsf{Z}_{\mathsf{D}}\right|, \left|\mathsf{Z}_{\mathsf{R}}\right| > \left|\mathsf{Z}_{\mathsf{A}}\right|; \underline{\mu}, \boldsymbol{\Sigma}\right].
\end{gathered}
$$

Formula reflects the idea that bias arises from a mixture distribution of mode specific bias weighted by the probability that that a given inheritance mode generates the most extreme statistic.

Method of moments

The components of the formula, e.g.

$$\mathrm{E}\left[\mathrm{Z_D} - \mu_\mathrm{D} \| \left|\mathrm{Z_D}\right| > \left|\mathrm{Z_A}\right|, \left|\mathrm{Z_D}\right| > \left|\mathrm{Z_R}\right|; \underline{\mu}, \mathbf{\Sigma}\right] \text{ and}$$
$$\mathrm{P}\left[\left|\mathrm{Z_D}\right| > \left|\mathrm{Z_A}\right|, \left|\mathrm{Z_D}\right| > \left|\mathrm{Z_R}\right|; \underline{\mu}, \mathbf{\Sigma}\right]$$

can be computed analytically or via Monte Carlo methods if $\underline{\mu}$ and $\mathbf{\Sigma}$ known.

But $\underline{\mu}$ and $\mathbf{\Sigma}$ are not known.

Instead, estimate $\underline{\mu}$ by $\underline{\mathrm{z}}^\mathrm{o} = \left(\mathrm{z_d^o}, \mathrm{z_a^o}, \mathrm{z_r^o}\right)$.

$\mathbf{\Sigma}$ has closed form expression in terms of

$\mathrm{p_i} =$ Prob(i risk alleles $\|$ Case) and $\mathrm{q_i} =$ Prob(i risk alleles $\|$ Control).

which can also be estimated from data.

Method of moments

Compute the bias correction as

$$\mathsf{E}\left[\mathsf{Z}_{\mathbf{I}} - \mu_{\mathbf{I}}; \underline{\hat{\mu}}, \hat{\boldsymbol{\Sigma}}\right] =$$

$$\mathsf{E}\left[\mathsf{Z}_{\mathsf{D}} - \hat{\mu}_{\mathsf{D}} \| \, |\mathsf{Z}_{\mathsf{D}}| > |\mathsf{Z}_{\mathsf{A}}|\,, |\mathsf{Z}_{\mathsf{D}}| > |\mathsf{Z}_{\mathsf{R}}|\,; \underline{\hat{\mu}}, \hat{\boldsymbol{\Sigma}}\right] \times$$
$$\mathsf{P}\left[|\mathsf{Z}_{\mathsf{D}}| > |\mathsf{Z}_{\mathsf{A}}|\,, |\mathsf{Z}_{\mathsf{D}}| > |\mathsf{Z}_{\mathsf{R}}|\,; \underline{\hat{\mu}}, \hat{\boldsymbol{\Sigma}}\right] \; +$$

$$\mathsf{E}\left[\mathsf{Z}_{\mathsf{A}} - \hat{\mu}_{\mathsf{A}} \| \, |\mathsf{Z}_{\mathsf{A}}| > |\mathsf{Z}_{\mathsf{D}}|\,, |\mathsf{Z}_{\mathsf{A}}| > |\mathsf{Z}_{\mathsf{R}}|\,; \underline{\hat{\mu}}, \hat{\boldsymbol{\Sigma}}\right] \times$$
$$\mathsf{P}\left[|\mathsf{Z}_{\mathsf{A}}| > |\mathsf{Z}_{\mathsf{D}}|\,, |\mathsf{Z}_{\mathsf{A}}| > |\mathsf{Z}_{\mathsf{R}}|\,; \underline{\hat{\mu}}, \hat{\boldsymbol{\Sigma}}\right] +$$

$$\mathsf{E}\left[\mathsf{Z}_{\mathsf{R}} - \hat{\mu}_{\mathsf{R}} \| \, |\mathsf{Z}_{\mathsf{R}}| > |\mathsf{Z}_{\mathsf{D}}|\,, |\mathsf{Z}_{\mathsf{R}}| > |\mathsf{Z}_{\mathsf{A}}|\,; \underline{\hat{\mu}}, \hat{\boldsymbol{\Sigma}}\right] \times$$
$$\mathsf{P}\left[|\mathsf{Z}_{\mathsf{R}}| > |\mathsf{Z}_{\mathsf{D}}|\,, |\mathsf{Z}_{\mathsf{R}}| > |\mathsf{Z}_{\mathsf{A}}|\,; \underline{\hat{\mu}}, \hat{\boldsymbol{\Sigma}}\right] .$$

Incorporating significance bias

- To take into account significance bias, adjust the conditioning event to require p-value sufficiently small, e.g.

$$\mathsf{E}\left[Z_{\mathbf{I}} - \mu_{\mathbf{I}} \| \mathsf{p}(Z_{\mathbf{I}}) < 0.05; \underline{\hat{\mu}}, \hat{\boldsymbol{\Sigma}}\right].$$

To compute this modify the component terms, e.g.

$$\begin{aligned}&\mathsf{E}\left[Z_{\mathsf{D}} - \mu_{\mathsf{D}} \| \left|Z_{\mathsf{D}}\right| > \left|Z_{\mathsf{A}}\right|, \left|Z_{\mathsf{D}}\right| > \left|Z_{\mathsf{R}}\right|, \mathsf{p}(Z_{\mathsf{D}}) < 0.05; \underline{\hat{\mu}}, \hat{\boldsymbol{\Sigma}}\right] \text{ and} \\ &\quad \mathsf{P}\left[\left|Z_{\mathsf{D}}\right| > \left|Z_{\mathsf{A}}\right|, \left|Z_{\mathsf{D}}\right| > \left|Z_{\mathsf{R}}\right|, \mathsf{p}(Z_{\mathsf{D}}) < 0.05; \underline{\hat{\mu}}, \hat{\boldsymbol{\Sigma}}\right].\end{aligned}$$

- Use Monte Carlo techniques to evaluate these expectations and probabilities

Method of Moments and Nonparametric Bootstrap Correction

	Empirical Power	OR	True Model	Empirical Bias	Mean Meth. Of Moments Correction	Mean Nonparam. Bootstrap Correction
Ranking Bias	NA	1.8	Dominant	0.153		0.138
Combined Ranking and Significance Bias	98%	1.8	Dominant	0.200	0.245	0.230
Ranking Bias	NA	1.8	Additive	0.274		0.249
Combined Ranking and Significance Bias	87%	1.8	Additive	0.496	0.436	0.434
Ranking Bias	NA	1.8	Recessive	0.137		0.145
Combined Ranking and Significance Bias	93%	1.8	Recessive	0.280	0.310	0.300
Ranking Bias	NA	1.4	Dominant	0.211		0.182
Combined Ranking and Significance Bias	63%	1.4	Dominant	0.790	0.477	0.476
Ranking Bias	NA	1.4	Additive	0.373		0.246
Combined Ranking and Significance Bias	42%	1.4	Additive	1.252	0.606	0.607
Ranking Bias	NA	1.4	Recessive	0.246		0.184
Combined Ranking and Significance Bias	44%	1.4	Recessive	1.080	0.538	0.527

Conditional Likelihood Approach

- Present only to address ranking bias, not combined with significance bias
- Idea used in other contexts –
  - significance bias for single marker/SNP -- Prentice (2008), Ghosh (2008)
  - estimating parameters conditional on stopping clinical trial after interim analysis Liu, Troendle et al. (2004)

Conditional Likelihood

As before let $\underline{Z} = (Z_D, Z_A, Z_R)$ – has trivariate normal distribution $N(\underline{\mu}, \Sigma)$ where $\underline{\mu} = (\mu_D, \mu_A, \mu_R)$, Σ is correlation matrix.

$$\text{e.g. } z_D = \frac{\sqrt{r}\ln \hat{OR}_D}{\hat{\sigma}_D} \text{ and } \mu_D = \frac{\sqrt{r}\ln OR_D}{\sqrt{1/p_a + 1/p_b + 1/p_c + 1/p_d}}$$

Suppose observed data are $\underline{\underline{z}}^o = (z_D^o, z_A^o, z_R^o)$ with $|z_D^o| > |z_A^o| > |z_R^o|$

Conditional Likelihood

$\underline{Z} = (Z_D, Z_A, Z_R)$ – has trivariate normal distribution $N(\underline{\mu}, \boldsymbol{\Sigma})$ where $\underline{\mu} = (\mu_D, \mu_A, \mu_R)$, $\boldsymbol{\Sigma}$ is correlation matrix.

Denote the conditional likelihood of $\underline{\mu}$ and $\boldsymbol{\Sigma}$ given that $|Z_D| > |Z_A| > |Z_R|$ as

$$L_C(\underline{\mu}, \boldsymbol{\Sigma}; \underline{z}^o \| \, |Z_D| > |Z_A| > |Z_R|) =$$

$$\frac{f(\underline{z}^o; \underline{\mu}, \boldsymbol{\Sigma}) \cdot 1_{[|z_D^o|>|z_A^o|>|z_R^o|]}}{P(|Z_D| > |Z_A| > |Z_R| \, ; \underline{\mu}, \boldsymbol{\Sigma})} =$$

$$\frac{f(\underline{z}; \underline{\mu}, \boldsymbol{\Sigma}) \cdot 1_{[|z_D^o|>|z_A^o|>|z_R^o|]}}{\int_{-\infty}^{\infty} \int_{-|z_d|}^{|z_d|} \int_{-|z_a|}^{|z_a|} f(z_d, z_a, z_r; \underline{\mu}, \boldsymbol{\Sigma}) dz_r dz_a dz_d}$$

f( ) denotes multivariate normal density

Conditional Likelihood

Define $\hat{\underline{\mu}}_C, \hat{\Sigma}_C$ as the conditional maximum likelihood estimators of

$$L_C(\underline{\mu}, \Sigma; \underline{z} \| \, |Z_D| > |Z_A| > |Z_R|) =$$

$$\frac{f(\underline{z}^o; \underline{\mu}, \Sigma) \cdot 1_{[|z_D^o| > |z_A^o| > |z_R^o|]}}{\int_{-\infty}^{\infty} \int_{-|z_d|}^{|z_d|} \int_{-|z_a|}^{|z_a|} f(z_d, z_a, z_r; \underline{\mu}, \Sigma) dz_r dz_a dz_d}$$

These are natural estimators in some sense i.e.

$$(z_d^o, z_a^o, z_r^o) = E\left[\underline{Z} \| \, |Z_D| > |Z_A| > |Z_R| \, ; \mu = \hat{\underline{\mu}}_C, \Sigma = \hat{\Sigma}_C\right].$$

the conditional mean of Z_d, Z_a, Z_r (conditioned on $|Z_D| > |Z_A| > |Z_R|$) is equal to the obserserved z_d^o, z_a^o, z_r^o when the true parameters are $\hat{\underline{\mu}}_C, \hat{\Sigma}_C$.

Conditional Likelihood

How to find $\hat{\underline{\mu}}_c, \hat{\Sigma}_c$ that maximize

$$\frac{f(\underline{z}^o; \underline{\mu}, \Sigma) \cdot 1_{[|z_D^o| > |z_A^o| > |z_R^o|]}}{\int_{-\infty}^{\infty} \int_{-|z_d|}^{|z_d|} \int_{-|z_a|}^{|z_a|} f(z_d, z_a, z_r; \underline{\mu}, \Sigma) dz_r dz_a dz_d} ?$$

Both $\underline{\mu}, \Sigma$ are functions of p_1, q_1, p_2, q_2 where

$p_i = \text{Prob}(i \text{ risk alleles} \parallel \text{Case})$ and $q_i = \text{Prob}(i \text{ risk alleles} \parallel \text{Control})$.

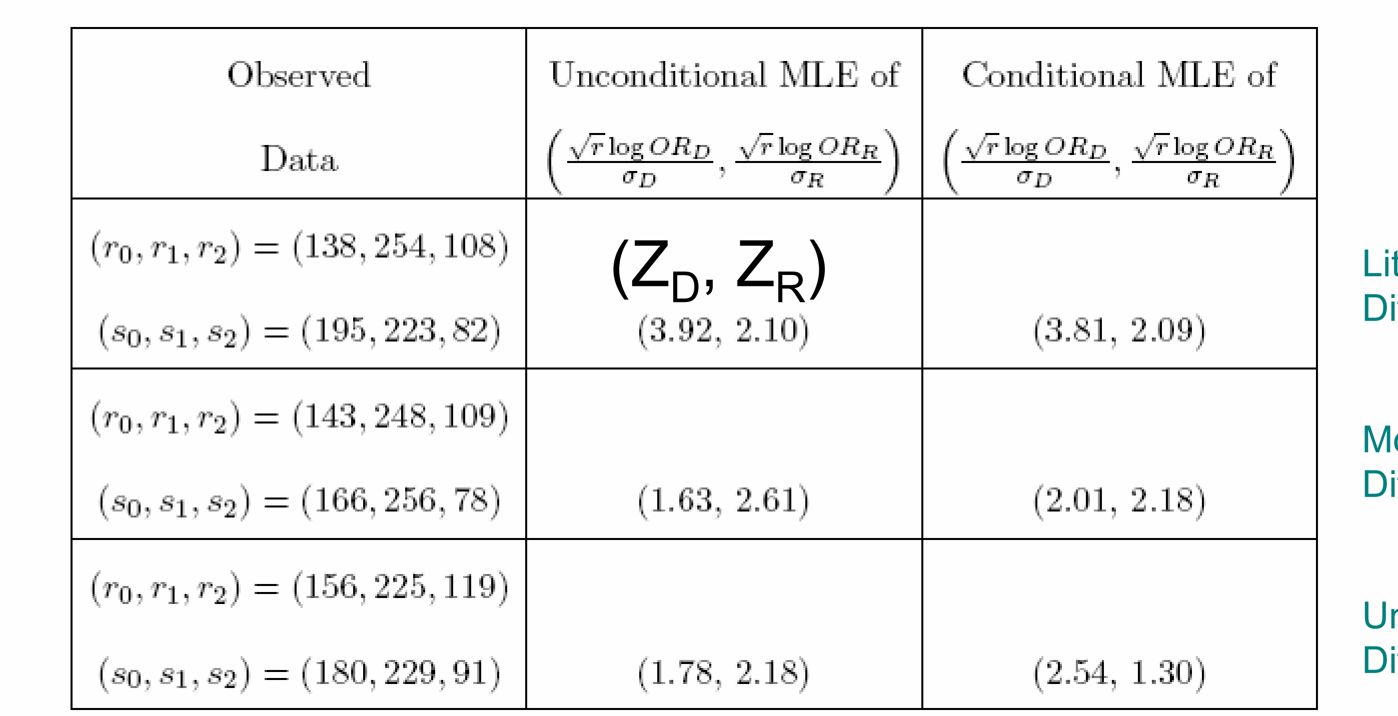
Implies 4 dimensional search for maximizing complicated function.

Consider a simplified situation for demonstration purposes.

Simplified CML Estimation

- Consider only choosing between dominant and recessive, i.e. $\underline{Z} = (Z_D, Z_R)$
- Take as known disease prevalence in population, minor allele frequency in population, and assume Hardy-Weinberg equilibrium
- Then only 2 dimensional search needed, integration is simpler
- Compare unconditional and conditional maximum likelihood estimates

MLE and CMLE comparison



Observed Data	Unconditional MLE of $\left(\frac{\sqrt{r}\log OR_D}{\sigma_D}, \frac{\sqrt{r}\log OR_R}{\sigma_R}\right)$	Conditional MLE of $\left(\frac{\sqrt{r}\log OR_D}{\sigma_D}, \frac{\sqrt{r}\log OR_R}{\sigma_R}\right)$
$(r_0, r_1, r_2) = (138, 254, 108)$ $(s_0, s_1, s_2) = (195, 223, 82)$	(Z_D, Z_R) (3.92, 2.10)	(3.81, 2.09)
$(r_0, r_1, r_2) = (143, 248, 109)$ $(s_0, s_1, s_2) = (166, 256, 78)$	(1.63, 2.61)	(2.01, 2.18)
$(r_0, r_1, r_2) = (156, 225, 119)$ $(s_0, s_1, s_2) = (180, 229, 91)$	(1.78, 2.18)	(2.54, 1.30)

Little Difference

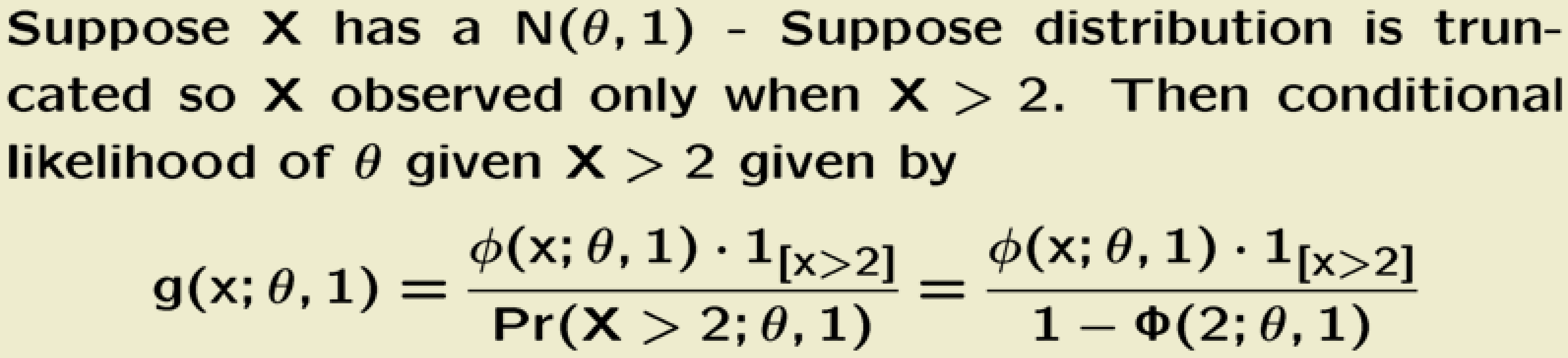
Moderate Difference

Unreasonable Difference

3 simulated datasets with parameters r = 500, MAF = 40%, prevalence=10%, true mode of inheritance = additive, True OR=1.8

Unusual estimates from CMLE

- Noted by Ghosh in context of significance bias in genetic markers (univariate)
- Overcorrection arises in very simple circumstances

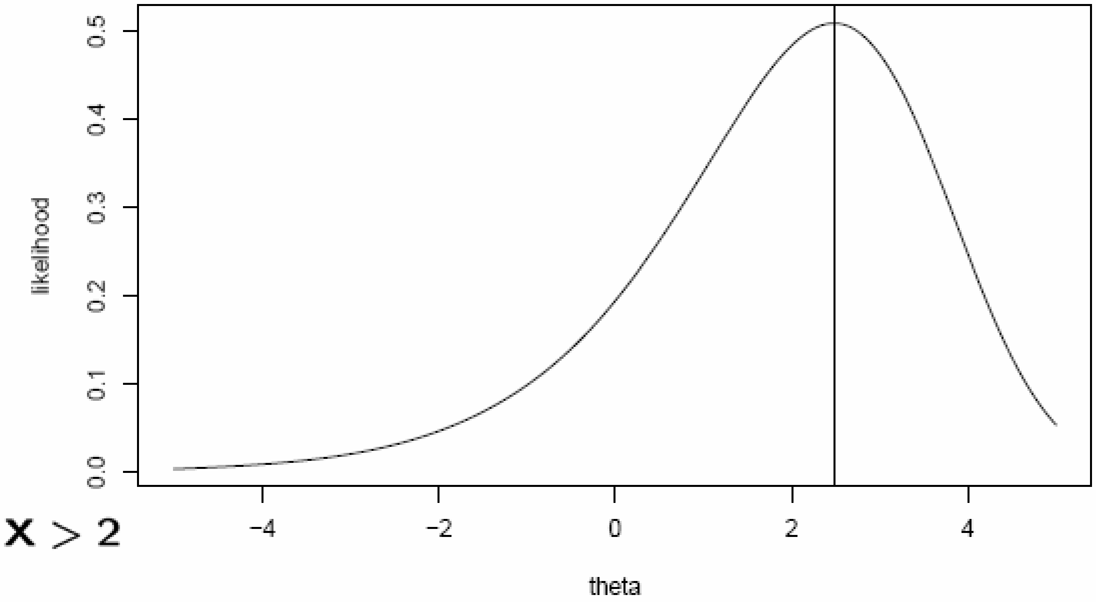


Suppose X has a $N(\theta,1)$ - Suppose distribution is truncated so X observed only when $X > 2$. Then conditional likelihood of θ given $X > 2$ given by

$$g(x;\theta,1) = \frac{\phi(x;\theta,1)\cdot 1_{[x>2]}}{Pr(X>2;\theta,1)} = \frac{\phi(x;\theta,1)\cdot 1_{[x>2]}}{1-\Phi(2;\theta,1)}$$

- See how g(x;θ,1) changes with observed x

Conditional likelihood for theta given X > 2, x = 3

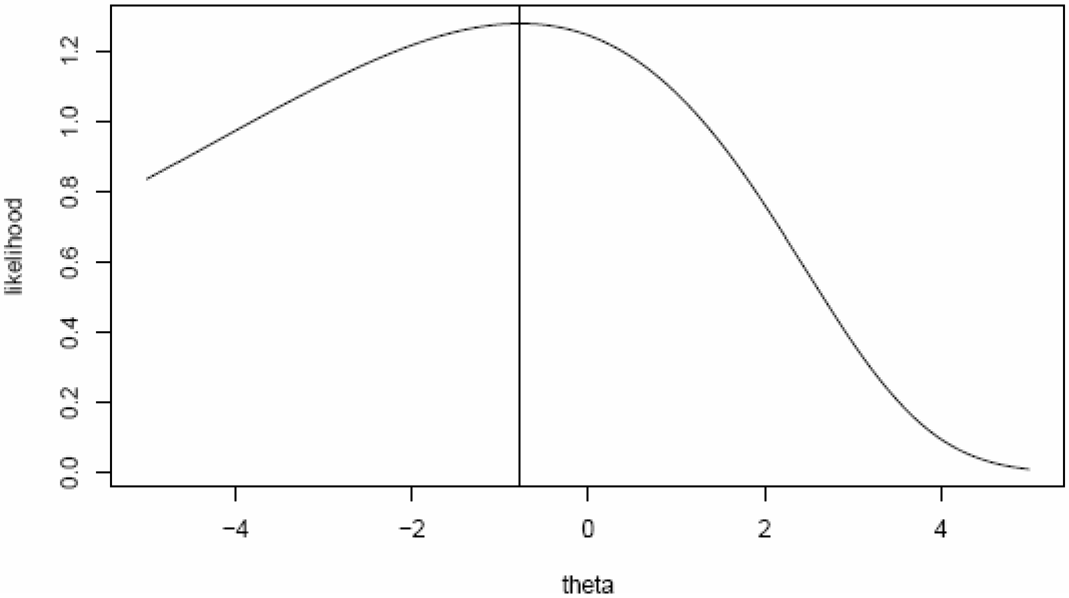


When x=3
CMLE for θ = 2.48
Reasonable Estimate

$\mathbf{X} \sim \mathrm{N}(\theta, 1)$
Condition on $\mathbf{X} > 2$

$$g(x; \theta, 1) = \frac{\phi(x;\theta,1)\cdot 1_{[x>2]}}{1-\Phi(2;\theta,1)}$$

Conditional likelihood for theta given X > 2, x = 2.3



When x=2.3
CMLE for θ = - 0.78
Unreasonable Estimate

Conclusions

- Simple case of ranking bias – easy to generalize to situations whenever look at extreme results among multiple tests
  - Genome-wide association – 1 million tests
- Bootstrap and Method of moments okay but not great
- Conditional likelihood unsatisfying
- Bayesian?

Acknowledgements

- Ben Kedem and UMD professors
- John Kane – Summer student
- Gang Zheng – NHLBI colleague

A Semiparametric Model for Length-biased Data

Jing Qin
Biostatistics Research Branch, NIAID

July, 2009

Collaborator: Yu Shen (UT M.D. Anderson Cancer Center)

Outline

1. Background and Motivation
2. Model and Method
3. Large Sample Properties
4. Numerical Studies
5. Summary

Prevalent Sampling Bias

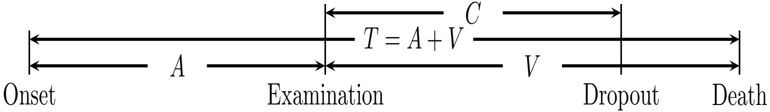
- Cancer screening studies: patients who are screening diagnosed (with a longer preclinical duration) often have more favorable disease prognosis
- Labor economy: subjects who have longer unemployment durations are more likely to be sampled into the studies
- Prevalent cohort studies: Individuals diagnosed with a disease (e.g dementia or HIV positive) are followed for the failure event (death)

Canadian Study of Health and Aging

- Dementia is a progressive degenerative medical condition
- The CSHA study was a multicenter epidemiologic study
- 14,000 $+$ subjects $\geq$ 65 years randomly chosen to receive an invitation for a health survey throughout Canada
- 10263 subjects agreed to participate $\Rightarrow$ screening for dementia in 1991

Prevalent Cohort

- 1132 people were identified as having the disease
- Dates of disease onset were retrieved from medical records
- Confirmed cases followed prospectively for death/censoring until 1996
- Patients with worse prognosis of dementia were more likely to die before the study recruitment



Length-biased Data

- Individuals have experienced the *initial event* but have not experienced the *failure event* at the time of recruitment
- Individuals diagnosed with the disease have to survive to the examination or sampling time (subject to left truncation)
- The “observed” time intervals from initiation to failure within the prevalent cohort tend to be longer than those arising from the underlying distribution of the general population

Biased and unbiased distributions

The observed failure time data $(T_1, \cdots, T_n)$ are a biased subset for population sample $\widetilde{T}$

- f: the unbiased pdf for $\widetilde{T}$
- g: the length-biased density for T

Under the stationarity assumption, given covariates, $Z = z$,

$$g(t|z) = \frac{tf(t|z)}{\int_0^\infty uf(u|z)du} := \frac{tf(t|z)}{\mu(z)}, \tag{1}$$

where $\mu(z) = \int_0^\infty uf(u|z)du$

Objective and Complications

Objective: estimate covariate effect on unbiased failure time $\widetilde{T}$ under the proportional hazards model

$$\lambda(t|Z) = \lambda_0(t)\exp(Z^T\beta)$$

Challenges:

- PH model structure assumed for $\widetilde{T}$ will not hold for the observed biased T
- Informative censoring due to left-truncation mechanism
- Naive approaches ignoring length-biased sampling ⇒ biased estimation

Literature on Length-Biased Data

- Nonparametric approaches

  Turnbull (1976), Lagakos, S. W., Barraj, L. M. and De Gruttola, V. (1988), Wang (1991)
  Vardi (1982, 1985), Gill, R. D., Vardi, Y. and Wellner, J. A. (1988), Asgharian, M., M'lan, C. M. and Wolfson, D. B. (2002), Asgharian, M. and Wolfson, D. B. (2005) among others

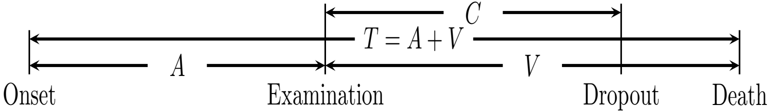
- Semiparametric Cox model to assess risk factors on length-biased data:

  Wang (1996) constructed a pseudo-likelihood method for length-biased data without right censoring
  Kalbfleisch and Lawless (1991), Keiding (1992), Wang, Brookmeyer and Jewell (1993) for left-truncated data

Notation

- $Y_i = \min(T_i, A_i + C_i)$, $T_i = A_i + V_i$
- $\delta_i = I(V_i \leq C_i)$
- Z_i is a vector of covariates

Assume that the censoring time measured from the recruitment time, C, and (A, V) are independent given covariate Z.

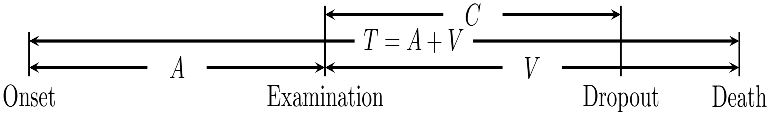


With Right-censoring

When the observed failure time T from length-biased sampling is subject to potential right censoring,

- the censoring variable can be dependent on the failure time, because

$$Cov(T, A+C) = Cov(A+V, A+C) = Var(A)+Cov(A, V) > 0$$



Model and Method

Under the stationarity assumption, the joint distribution of (A, V) and (A, T) given $Z = z$ follows (Zelen, 2004, Vardi, 1989)

$$f_{A,T}(a, t|z) = f_{A,V}(a, v|z) = f(a + v|z)I(a > 0, v > 0)/\mu(z)$$

Since C is assumed to be independent of (A, V)

$$\begin{aligned} &\Pr(A = a, T = t, C \geq t - a|z) \\ &= \quad f_{A,V}(a, t - a|z)\Pr(C \geq t - a) \\ &= \quad f(t|z)S_C(t - a)/\mu(z), \end{aligned}$$

where $S_C(t)$ is the survival function for C.

Estimating Equation I (EE-I)

Assume that C is independent of Z,

$$U_1(\beta) = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j=1}^{n} Z_j \exp(\beta^T Z_j) I(Y_j \geq Y_i) \delta_j \{Y_j S_C(Y_j - A_j)\}^{-1}}{\sum_{j=1}^{n} \exp(\beta^T Z_j) I(Y_j \geq Y_i) \delta_j \{Y_j S_C(Y_j - A_j)\}^{-1}} \right]$$

For unknown S_C, we can use its consistent Kaplan-Meier estimator, $\widehat{S}_C(t)$ for C

Estimating Equation II (EE-II)

With the observed data, we construct the unbiased estimating equation

$$U_2(\beta) = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j=1}^{n} I(Y_j \geq Y_i)\delta_j \{w_c(Y_j)\}^{-1} Z_j \exp(\beta^T Z_j)}{\sum_{j=1}^{n} I(Y_j \geq Y_i)\delta_j \{w_c(Y_j)\}^{-1} \exp(\beta^T Z_j)} \right]$$

By replacing $w_c(y) = \int_0^y S_C(t)dt$ with its consistent estimator $\hat{w}_c(t) = \int_0^t \widehat{S}(u)du$, we have an asymptotic unbiased estimating equation, EE-II

Asymptotic Properties

Under mild regularity conditions

- the estimating equations $U_1(\beta)$ and $U_2(\beta)$ have, asymptotically, a consistent and unique solution $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively
- $$\sqrt{n}(\hat{\beta}_1 - \beta_0) \Rightarrow N(0, \Sigma_1),$$
- $$\sqrt{n}(\hat{\beta}_2 - \beta_0) \Rightarrow N(0, \Sigma_2),$$

Simulation Set-up

The underlying population distribution of $\widetilde{T}$ follows a proportional hazards model

- with two independent covariates: $Z_1 \sim$ Bernoullli distribution and $Z_2 \sim$ a uniform variable on (-0.5,0.5)
- censoring percentages: 20, 35 and 50
- n=200 and repeated for 1000 times

Table 1. Proportional hazards model

		EE-I		EE-II	
β_0	Cen %	$\hat{\beta}_1$	95% CP	$\hat{\beta}_2$	95% CP
(1,1)	20%	0.991 0.972	.974 .959	1.020 1.016	.985 .970
	35%	0.936 0.920	.930 .912	1.024 1.026	.974 .958
	50%	0.838 0.813	.841 .866	1.019 1.003	.953 .944
(2,2)	20%	1.989 1.951	.968 .961	2.013 2.001	.972 .971
	35%	1.961 1.839	.963 .920	2.042 2.018	.971 .962
	50%	1.833 1.674	.894 .837	2.016 1.992	.962 .955

- Instability of EE-I due to $\hat{S}_C(t) \rightarrow 0$ in the denominator
- EE-II is more robust with $\int_0^t \hat{S}_C(u)du$ in the denominator

Canadian Study of Health and Aging

Excluding subjects with missing data on A and Z, a total of 818 patients remained

- date of onset, date of screening for dementia, date of death or censoring and death indicator variable
- three types dementia diagnosis: probable Alzheimer's disease (393), possible Alzheimer's disease (252) and vascular dementia (173)
- 638 out of 818 patients died at end of this study

Example Results

Table 2. Estimates (se) of regression coefficients for dementia data

	Length-bias adjusted Cox model		Naive Cox model
	EE-I	EE-II	
Vascular dementia	0.137 (.101)	0.074 (.101)	0.076 (.103)
Possible Alzheimer	-0.109 (.093)	-0.134 (.091)	-0.182 (.093)

probable Alzheimer's disease as the baseline

The diagnosis of three subtypes of dementia had little difference in long-term survival, which was consistent with the nonparametric survival estimators provided in Wolfson et al (2001).

Computation Algorithms

- Estimating covariate effects for *traditional* right-censored failure time data is easy for the end user with existing software
- Illustrate how to use coxph in S-PLUS/R for traditional right-censored data to analyze length-biased right-censored data under Cox model
- Slightly modified commands in S-PLUS/R for estimating β from EE-I and EE-II

Computation Algorithms

For length-biased data, use the function coxph, with the offset option

- define $\boldsymbol{Z}_V$ and $\boldsymbol{Z}_P$ as indicators of Vascular dementia and possible Alzheimer's disease
- EE-I: $\widehat{\boldsymbol{W}}_1$ is the Kaplan-Meier estimator of C on all $Y_i - A_i$ and $i = 1, \cdots, m$,

$$\widehat{W}_{1i} = \{Y_i \widehat{S}_C(Y_i - A_i)\}^{-1}.$$

- EE-II: $\widehat{\boldsymbol{W}}_2$ is the integral of the Kaplan-Meier estimator of C

$$\widehat{W}_{2i} = \{\hat{w}_c(Y_i)\}^{-1} = \{\int_0^{Y_i} \widehat{S}_C(u) du\}^{-1}$$

Computation Algorithms

Apply
> coxph(Surv(ftime,rep(1,m)) ~ Z_V + Z_P +
offset($\log(\widehat{W}_k)$)), data=fdata),

where $k = 1$ or 2, "ftime" is the observed failure times, m is the total number of the observed failure times, "fdata" is the subset of the whole data matrix among subjects with observed failure times only

offset term is used to include $(\log(\widehat{W}_k)))$ in the model as a *fixed* covariate with coefficient 1 in the model

Equivalence

Recall EE-I, $U_1(\beta)$ can be expressed as

$$\sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j=1}^{n} I(Y_j \geq Y_i)\delta_j Z_j \exp(\beta^T Z_j)\widehat{W}_{1j}}{\sum_{j=1}^{n} I(Y_j \geq Y_i)\delta_j \exp(\beta^T Z_j)\widehat{W}_{1j}} \right] = 0$$

$$\sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j=1}^{n} I(Y_j \geq Y_i)\delta_j Z_j \exp\{\beta^T Z_j + \log(\widehat{W}_{1j})\}}{\sum_{j=1}^{n} I(Y_j \geq Y_i)\delta_j \exp\{\beta^T \boldsymbol{Z}_j + \log(\widehat{W}_{1j})\}} \right] = 0$$

which is the same as the score equation used in the traditional Cox model with a linear predictor $\log(\boldsymbol{W}_1)$ restricting among the observed failure times.

Discussion

- Propose an inference method to evaluate covariate effects on unbiased data when the observed data are subject to length-biased sampling
- The semiparametric Cox model structure imposed on the population sample ⇒ straightforward interpretation of the regression coefficients
- The PH model assumption is not invariant for population data and length-biased data in general

Discussion

- $\hat{S}_C(t)$, as an inverse weight in EE-I $\Rightarrow$ instability of the estimating equation at the tail
- $\hat{w}_c(t) = \int_0^t \hat{S}_C(u)du$ in EE-II is the area under the survival curve $\Rightarrow$ robust
- Computational advantages: use modified standard software for traditional right-censored data in
    - S-PLUS/R: coxph with "offset" option
    - SAS: PROC PHREG with "OFFSET" option
- The consistent variance estimators of $\hat{\beta}_1$ and $\hat{\beta}_2$ can be obtained from the estimating equations or bootstrap method

Acknowledgment

We thank Professor Masoud Asgharian and investigators of the Canadian Study of Health and Aging (CHSA) for providing us the dementia data from CHSA.

additional equations

$$h(z|T=t)=\frac{g(t|\boldsymbol{z})h(\boldsymbol{z})}{\int g(t|\boldsymbol{z})h(\boldsymbol{z})d\boldsymbol{z}}=\frac{tf(t|\boldsymbol{z})h(Z)/\mu(Z)}{\int tf(t|Z)h(Z)/\mu(Z)dZ}.$$

Under the proportional hazards model, the conditional expectation of Z is

$$\begin{aligned}E[Z|T=t] &= \frac{\int ztf(t|z)h(z)/\mu(z)dz}{\int tf(t|z)h(z)/\mu(z)dz}\\ &= \frac{\int z\exp(\boldsymbol{\beta}^T z)S_f(t|z)h(z)/\mu(z)dz}{\int \exp(\boldsymbol{\beta}^T z)S_f(t|z)h(z)/\mu(z)dz}\\ &= \frac{E[Z\exp(\boldsymbol{\beta}^T Z)S_f(t|Z)/\mu(Z)]}{E[\exp(\boldsymbol{\beta}^T Z)S_f(t|Z)/\mu(Z)]}\end{aligned}$$

$$E[Z|y,\delta=1,a]=\frac{\int zf(y|z)S_C(y-a)h(z)/\mu(z)dz}{\int f(y|z)S_C(y-a)h(z)/\mu(z)dz}=\frac{E[Z\exp(\boldsymbol{\beta}^T Z)S_U(y|Z)/\mu(Z)]}{E[\exp(\boldsymbol{\beta}^T Z)S_U(y|Z)/\mu(Z)]}.$$

Estimating Equation I

The above joint distribution for (Y, A, C) leads to the following conditional expectation

$$\begin{aligned} & E\left[\delta I(Y \geq y)\{Y S_C(Y-A)\}^{-1}\Big| Z\right] \\ = & \int_y^\infty f(t|Z)\int_0^t S_C(t-a)t^{-1}S_C^{-1}(t-a)dadt/\mu(Z) \\ = & \int_y^\infty f(t|Z)/\mu(Z)dt = S_f(y|Z)/\mu(Z) \qquad (2) \end{aligned}$$

$$E[Z|Y=y,\delta=1,A=a] = \frac{E[Z\exp(\beta^T Z)S_f(y|Z)/\mu(Z)]}{E[\exp(\beta^T Z)S_f(y|Z)/\mu(Z)]}$$

Recall

$$\Pr(Y = y, A = a, C \geq y - a|z) = f(y|z)S_C(y-a)/\mu(z),$$

then the probability of observing the length-biased failure time at y

$$\Pr(Y = y, \delta = 1|z) = \frac{f(y|z)w_c(y)}{\mu(z)}, \tag{3}$$

Estimating Equation II

Based on (3), we have

$$
\begin{aligned}
& E[I(Y > y, \delta = 1)\{w_c(Y)\}^{-1}|Z] \\
= & \int_y^\infty f(t|Z)dt \int_0^t S_C(v)dv\{w_c(t)\}^{-1}/\mu(Z) \\
= & \int_y^\infty f(t|Z)dt/\mu(Z) = S_f(y|Z)/\mu(Z)
\end{aligned}
$$

where $w_c(y) = \int_0^y S_C(t)dt$.

Empirical Likelihood and the Nonparametric Behrens-Fisher Problem

James F. Troendle*

Biostatistics and Bioinformatics Branch
DESPR, NICHD, NIH, DHHS

*This research was supported by the Intramural Research Program of NICHD/NIH.

I. Nonparametric Behrens-Fisher Problem

Two sample problem with <u>no</u> distributional assumptions

$$X \sim F \quad Y \sim G$$

$$p = P\{\ X\ \ >\ \ Y\ \}\ +\ 1/2\ P\{\ X\ \ =\ \ Y\ \}$$

$$H_0 : \quad p = 1/2$$

I. Nonparametric Behrens-Fisher Problem (Cont)

- p is called the **relative treatment effect**

- Studied by many, especially Brunner & Munzel (2000)

$$p = \int F \; dG, \qquad \text{where F and G are the normalized df}$$

$$\hat{p} = \int \hat{F} \; d\hat{G}$$

- Note however that p is not transitive, i.e.

$F > G > H > F$ is possible

I. Nonparametric Behrens-Fisher Problem (Cont)

- Brunner & Munzel use $\hat{p}$ to estimate p, and derive tests of H_0 from the asymptotic normality of $\sqrt{N}(\hat{p} - 1/2)$

- This generalized Wilcoxon test (GW) works pretty well, but can one obtain an LRT for this problem?

II. A Semiparametric Solution

- Fokianos, Kedem, Qin, & Short (**FKQS**)(2001) solved

a similar problem using **Empirical Likelihood**, by assuming

$$f(x) = \exp(\alpha + \beta h(x))g(x)$$

- **FKQS** test $H_I : F = G$, not $H_0 : p = 1/2$

III. Empirical Likelihood

$$L = \prod_{i=1}^{n_1} [F(X_i) - F(X_i-)] \ \prod_{j=1}^{n_2} [G(Y_i) - G(Y_i-)]$$

- Maximization of L yields nonparametric MLE's of F and G

III. Empirical Likelihood (Cont)

- **FKQS** $\Rightarrow$ consider step cdf's with jumps at the observed values $x_1, \ldots, x_n$ where the size of the jumps are the unknown parameters. The problem becomes a constrained maximization problem with many parameters.

- For the two sample problem we get:

$$\mathcal{L} = \sum_{i=1}^{n} \log p_i^{m_i} + \sum_{j=1}^{n} \log q_j^{m'_j}$$

IV. Empirical Likelihood Ratio Test for NBFH

To get ELRT, we need to solve

$$\max \qquad \mathcal{L}$$

$$\text{subject to} \quad \sum_{i=1}^{n} p_i = 1$$

$$\sum_{j=1}^{n} q_j = 1$$

$$\sum_{j=1}^{n} q_j \sum_{i=j+1}^{n} p_i \; = \; \sum_{j=1}^{n} q_j \sum_{i=1}^{j-1} p_i$$

IV. Empirical Likelihood Ratio Test for NBFH (Cont)

Lagrange Multipliers $\Rightarrow$

$$m_i/p_i = \lambda_1 \left[\sum_{r=1}^{i-1} q_r - \sum_{r=i+1}^{n} q_r \right] + \lambda_2 \qquad i = 1, \ldots, n$$

$$m'_j/q_j = \lambda_1 \left[\sum_{s=r+1}^{n} p_s - \sum_{s=1}^{r-1} p_s \right] + \lambda_3 \qquad j = 1, \ldots, n$$

along with the constraint equations.

IV. Empirical Likelihood Ratio Test for NBFH (Cont)

- One can easily eliminate the q's. p's appear intractable.

- **Struggle** $\rightsquigarrow$

$$p_i = \frac{m_i}{n_1 + \lambda_1 \left[-1 + \sum_{j=1}^{i-1} \frac{m'_j}{n_2+\lambda_1\theta_j} + \sum_{j=1}^{i} \frac{m'_j}{n_2+\lambda_1\theta_j} \right]} \qquad i = 1, \ldots, n.$$

where

$$\theta_j = \sum_{i=j+1}^{n} p_i - \sum_{i=1}^{j-1} p_i$$

IV. Empirical Likelihood Ratio Test for NBFH (Cont)

- Problem of generating candidates for feasible solutions is reduced to getting numerical solution to roots of single variable (λ_1) equation.

- Numerical solutions to single variable problems work quite well; in this case feasible solutions are **almost** always found.

V. Approximating the Null Distribution

- Large Sample Approximations may or may not be ok in practice, but not ok for comparison of power

- $\tilde{F}$ and $\tilde{G}$ constrained EMLE's for F and G

- Samples from $\tilde{F}$ and $\tilde{G}$ should approximate data under H_0

- In practice this works like a random permutation test, but with random draws from $\tilde{F}$ and $\tilde{G}$ used to construct re-sampled datasets

VI. Simulations

(1) 1st Pop=N(0,1); 2nd Pop=N(μ_2, σ_2^2)

(2) 1st Pop=N(0,1); 2nd Pop=U(Mean=μ_2, Var=σ_2^2)

(3) 1st Pop=Poi(Mean=2); 2nd Pop=Poi(Mean=μ_2)

Table 1: Power of the NBFH tests for null configurations.*

			Null Distribution		
Dist.			Asympt.	Sim.	EMLE
Type	(μ_2, σ_2^2)	(n_1, n_2)	GW	GW	ELRT
(1)	(0,1)	(10,10)	**.0557**	.0472	.0494
	(0,2)	(10,10)	**.0543**	.0466	.0485
	(0,1)	(30,30)	**.0519**	.0499	.0501
	(0,2)	(30,30)	**.0515**	.0499	.0501
	(0,1)	(30,20)	**.0533**	.0496	.0497
	(0,2)	(30,20)	**.0527**	.0510	.0508
	(0,.5)	(30,20)	**.0519**	.0498	.0502
(2)	(0,1)	(10,10)	**.0540**	.0459	.0482
	(0,2)	(10,10)	**.0534**	.0468	.0487
	(0,1)	(30,30)	**.0521**	.0502	.0503
	(0,2)	(30,30)	**.0514**	.0507	.0508
	(0,1)	(30,20)	**.0528**	.0499	.0499
	(0,2)	(30,20)	**.0532**	.0510	.0511
	(0,.5)	(30,20)	**.0517**	.0499	.0501
(3)	(2,2)	(10,10)	**.0576**	.0469	.0484
	(2,2)	(30,30)	**.0524**	.0503	.0503
	(2,2)	(30,20)	**.0532**	.0495	.0496

*Estimated from 100000 replications, giving SE≈ .0007.

Table 2: Power of the NBFH tests for nonnull configurations.*

			Null Distribution	
Dist.			Sim.	EMLE
Type	(μ_2, σ_2^2)	(n_1, n_2)	GW	ELRT
(1)	(1.5,1)	(10,10)	.1642	**.1701**
	(1.5,2)	(10,10)	.0907	**.0934**
	(1.5,1)	(30,30)	.4494	.4508
	(1.5,2)	(30,30)	.1998	.2007
	(1.5,1)	(30,20)	.3685	.3700
	(1.5,2)	(30,20)	.1542	.1542
(2)	(1.5,1)	(10,10)	.1528	**.1595**
	(1.5,2)	(10,10)	.0734	**.0758**
	(1.5,1)	(30,30)	.4087	.4106
	(1.5,2)	(30,30)	.1425	.1435
	(1.5,1)	(30,20)	.3279	.3307
	(1.5,2)	(30,20)	.1138	.1141
(3)	(3,3)	(10,10)	.2384	**.2446**
	(3,3)	(30,30)	.6383	.6398
	(3,3)	(30,20)	.5312	.5339

*Estimated from 100000 replications. **Bold values significantly higher by Fisher's exact test at 5%.**

VII. Owen's Solution

- Owen (2001) is the classic reference for Empirical Likelihood.

- Section 11.4 gives a Multi-sample ELT with solution expressed via Taylor's series expansion.

- **Problem**: Truncating the series leaves a "solution" that is in general not feasible.

VIII. Density Ratio Model

- Assume the density ratio model:

$$f(x) = \exp(\alpha + \beta h(x))g(x)$$

- Given h, the EMLE's for F and G can be found in a straightforward manner using profiling, as in Qin and Zhang (1997) and **FKQS**.

- However, h is unknown. Estimate h from this Box-Cox family:

$$h(x,\lambda) = \begin{cases} \frac{(x-x_{min})^{\lambda}-1}{\lambda} & \text{when } \lambda \neq 0 \\ \log(x - x_{min}) & \text{when } \lambda = 0 \end{cases}$$

VIII. Density Ratio Model (Cont)

- Estimation of relative treatment effect, p, via

$$\hat{p} = \int \hat{F} \; d\hat{G}$$

- Fokianos and Troendle (2007) show that for known h,

$$\sqrt{n}(\hat{p} - p) \rightarrow \mathcal{N}(0, \sigma^2_{dr})$$

as $n_1, n_2 \rightarrow \infty$ such that $n_1/n_2 \rightarrow \rho$

IX. Tests of the NBFH

- Test $H_0 : p = 1/2$, by using

$$T_{dr} = \frac{\sqrt{n}(\hat{p} - 1/2)}{\hat{\sigma}_{dr}}$$

- One dimensional $h(x, \lambda)$ estimates
$\lambda \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$

- Two dimensional $h(x, \lambda_1, \lambda_2)$ estimates
$\lambda_1 \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and
$\lambda_2 \in \{1.0, 1.2, 1.4, 1.6, 1.8, 2.0\}$

X. Distribution under H_0

- Because of the uncertainty in the estimate of λ, we can't use the asymptotic distribution of $\hat{p}$

- Sampling from $\tilde{F}$ and $\tilde{G}$, **the constrained EMLE's for F and G under the nonparametric model**, is used to approximate the distribution of T_{dr} under H_0

Table 3: Power of the NBFH tests for non-null configurations.*

			Null Distribution Simulated EMLE			
(n_1, n_2)	F	G	GWT	ELRT	SEM1	SEM2
(20,20)	N(0,1)	N(0.95,1)	**.8044**	**.8050**	.7924	**.8068**
	N(0,1)	N(1.15,2)	**.7878**	**.7890**	.7500	**.7908**
	N(0,1)	U(0.95,1)	.7668	.7686	**.7894**	**.7904**
	N(0,1)	U(1.25,2)	.7740	.7754	**.8066**	**.7960**
	U(0,1)	U(1.25,2)	.7946	.7972	**.8184**	**.8086**
	G(1,1)	G(2,2)	**.8042**	**.8052**	.7806	**.8050**
	BIN(10,0.8)	BIN(50,0.2)	.6960	.6970	.6446	**.7164**
(20,30)	N(0,1)	N(0.95,1)	**.8690**	**.8698**	.8596	**.8756**
	N(0,1)	N(1.15,2)	**.8822**	**.8836**	.8570	**.8854**
	N(0,1)	U(0.95,1)	.8458	.8484	**.8640**	**.8664**
	N(0,1)	U(1.25,2)	.8820	.8838	**.8970**	**.8906**
	U(0,1)	U(1.25,2)	.8912	.8934	**.9034**	**.8992**
	G(1,1)	G(2,2)	**.8524**	**.8526**	.8338	**.8562**
	BIN(10,0.8)	BIN(50,0.2)	.8482	.8484	.8120	**.8602**
(30,20)	N(0,1)	N(0.95,1)	**.8680**	**.8694**	**.8604**	**.8714**
	N(0,1)	N(1.15,2)	**.8272**	**.8280**	.7906	**.8342**
	N(0,1)	U(0.95,1)	.8366	.8396	**.8646**	**.8628**
	N(0,1)	U(1.25,2)	.8110	.8160	**.8508**	**.8476**
	U(0,1)	U(1.25,2)	.8368	.8408	**.8672**	**.8610**
	G(1,1)	G(2,2)	**.8746**	**.8760**	.8562	**.8790**
	BIN(10,0.8)	BIN(50,0.2)	.7152	.7202	.6854	**.7504**

*Estimated from 5000 replications. **Bold values within 2 SE of highest power.**

XI. Conclusions

- **Empirical Likelihood** quite useful in two-sample problem

- EL yields **doubly robust** tests of NBFH

- **Struggle** $\rightsquigarrow$ **Success**

References

Brunner, E. and Munzel, U. (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biom. J. 42*, 17–25.

Fokianos, K., Kedem, B., Qin, J. and Short, D. (2001). A semiparametric approach to the one-way layout. *Technometrics 43*, 56–65.

Fokianos, K. and Troendle, J., F. (2007). Inference for the relative treatment effect with the density ratio model. *Statistical Modelling 7*, 155–173.

Owen, A., B. (2001). *Empirical Likleihood*. Boca Raton, Florida: Chapman and Hall/CRC.

Qin, J. and Zhang, B. (1997). A goodness of fit test for the logistic regression model based on case–control data. *Biometrika 84*, 609–618.

Troendle, J. F. (2002). A likelihood ratio test for the nonparametric Behrens-Fisher problem. *Biom. J. 44*, 813–824.

Troendle, J. F. and Yu, K. Y. (2006). Likelihood approaches to the non-parametric two-sample problem for right-censored data. *Stat. in Med. 25*, 2284–2298.