Darwinian Model Building

Do Kester

SRON
Netherlands Institute for Space Research

Introduction

Darwins evolution theory applied to model building

- Darwin
  - Variation of the genotype
  - Selection of the phenotype
- Model
  - relation between some input(s) and some output.
  - function $y = f(x_1, x_2, \ldots : \vartheta)$

Cervical Cancer

- In the Netherlands every woman between 30 and 60 is invited to participate in a free test every 5 years to detect cervical cancer.
- These pap smear tests yield 2 numbers
  - O-value: 9 unrelated inflammatory events.
    - bacterial, viral, fungal
    - only one is present
    - a value of 6 represents a healthy status
  - P-value: indicator for stages in cervical cancer development (1-9)
    - increasing in severity
    - at a value above 5 the woman is sent to a gyneacologist
- Sometimes also a Human Papilloma Virus (HPV) test is done.
  - HPV is associated with cervical cancer.

Data I

- The Leiden Cytology and Pathology Laboratory (LCPL) has a database with pap smear tests for 300000 women.
- We selected those where at least 2 HPV tests could be found: 1750 in total.
- On average there are 5 tests per woman.
- The case history for a woman form a small time series with data:
  - P-values
  - O-values
  - HPV
  - age
- Half of the data were used in modeling. The other half was for testing.

SRON
Netherlands Institute for Space Research

Data II

- Can we predict the next P-value from the previous data in the time series
- As inputs we have:
  - age at time of test — real — measured in decades
  - time to the next test — real
  - P-value — real
  - O-values — 9 booleans
  - HPV test — integer with 3 values
- Output
  - next P-value — real

SRON
Netherlands Institute for Space Research

Model

The model is defined by the genotype. In this case one chromosome containing one gene. Very, very simple.

The gene is a string (or equivalently a tree) of bases.

A base is a data item, a model parameter, an operator, etc. Each base is assigned a ascii character.

The chromosome is a string of ascii characters which is interpreted as a program in Reverse Polish Notation (RPN).

aX*b+R translates in $\sqrt{(a*X+b)}$

From these genotypes individuals (phenotypes) must be grown which interact with the environment: either die or survive and reproduce.

RPN

- RPN (or postfix notation) uses stack based calculations.
  - HP calculators
- Each item in the string changes the stack count by a fixed amount. (+1, 0, -1 or -2)
- At the end the stack count needs to be 1. Only one item is left on the stack which is the result of the calculation.

Genotype Bases

Bases for the LCPL dataset

X, Y, Z	+1	real	P-value, age, time to next test
K	+1	integer	HPV
B .. J	+1	boolean	O-values
a .. h	+1	real	model parameters
+, -, *, /	-1	operators	
<, >, ~, &, |	-1	operators	
R, S, L, A	0	functions	sqrt, sign, log, atan
?	-2	if – branch	abc? => if c then b else a
p, q, m	+1	real	read from memory
P, Q, M	0	null	write to memory

Memory location

- Memory can be written to and read from.
- When writing before reading a value is stored to be used later in the algorithm.
  - aX*LPb+ZB?p+
- When reading is done before writing, the value stored in the previous cycle of the time series is used.
- Some value is needed for the first cycle in each time series. It is added to the list of free model parameters.
- This way information can be passed from one test to the next.

Phenotype

- The phenotype is how the genotype manifests itself in the environment. The environment is the data. With different data one would get different individuals.
- The model is fit to the data where the fitting is done over
  - model parameters
  - memory locations
- Nested sampling is used to find the best set of parameters
- Evidence represents the fitness in the environment.

A bit of Bayes

For parameters ϑ, data D and model M:

$\Pr(\vartheta|M) * \Pr(D|\vartheta M) = \Pr(D|M) * \Pr(\vartheta|DM)$

prior * likelihood = evidence * posterior

The evidence is obtained directly from Nested Sampling.

As nothing is known about the models or its parameters we take a classic Gaussian error distribution as likelihood.

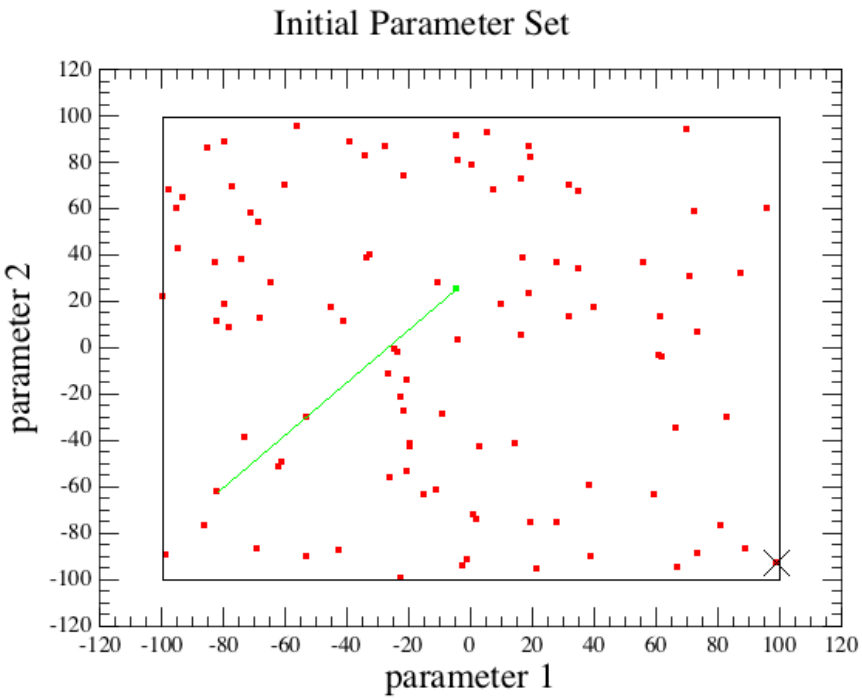
The priors on the parameters are uniform in [-100,100]. All our data are within the range [0,10]. Parameters should not be much different.

SRON
Netherlands Institute for Space Research

Nested Sampling I

Nested Sampling Algorithm

1. take N random points
2. calculate log likelihoods
3. select point with worst logL
4. store it with proper weight
5. replace by another
6. randomize the new point
7. goto 3.



Engines

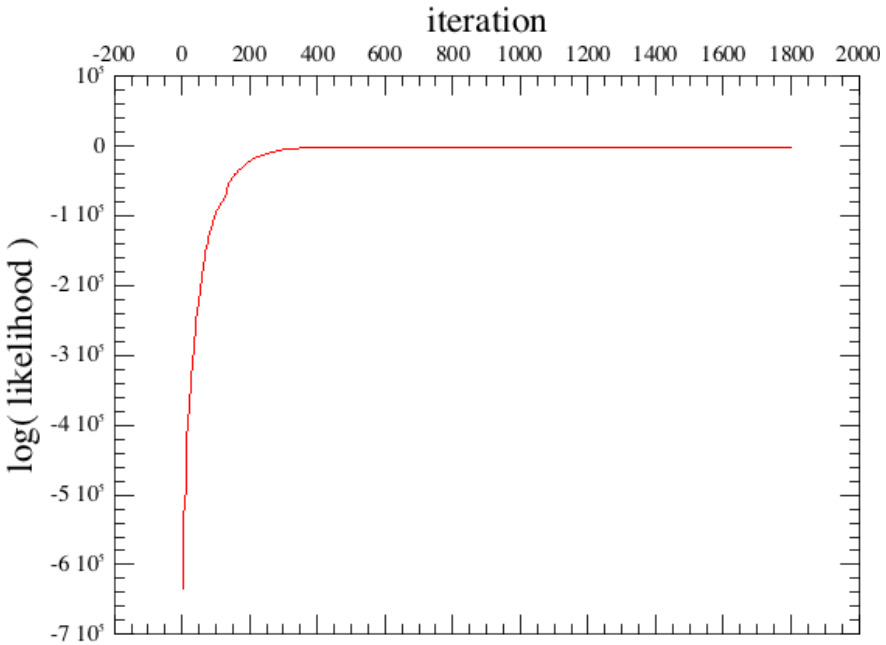
For randomizing a (multidimensional) point p we use 3 engines.

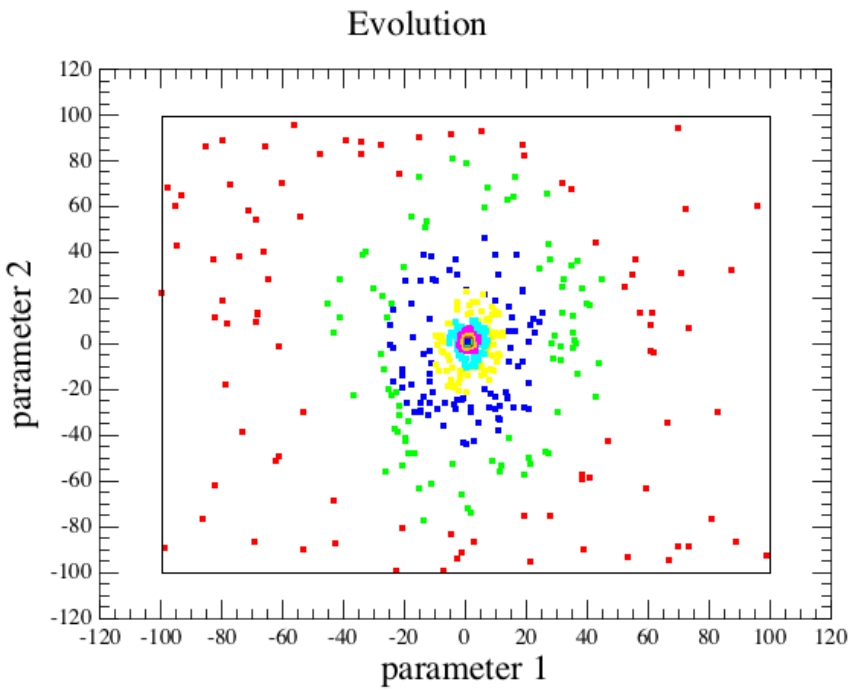
1. Step engine.
   - move each parameter of p by a random step.
1. Frog engine.
   - select a number (1-5) other points.
   - calculate the average of these points.
   - jump p_1 by a random amount to/from/over the average.
1. Cross engine.
   - Select another point.
   - take at random parameters from p and the other point.

All new points are subject to $\log L > \log L_{low}$.

SRON
Netherlands Institute for Space Research

Nested Sampling II





Evidence is the integral of the likelihood over the parameter space.

Translation

- Nested Sampling requires thousands of model evaluations.
- It needs something better than a RPN-interpreter.
- Translate the RPN string into C code.
- Compile and load into a dynamic library
- Link the library
- Find the handle to the function and run it.

C-code

```
/*
 * Automatically generated function by ga_make_function
 *   Copyright Do Kester, Zuidhorn  2010
 *
 * Original gene (genotype):
 *   pqa*L+K-RZ-ALKDIb*/-c*d+*AeX/I+*LLQfX/QZ*PR-S
 */

#include <math.h>
#include "user.h"

#define sqrt( x )       ( ( (x) >= 0 ) ? sqrt( x ) : 0 )
#define log( x )        ( ( (x) > 0 ) ? log( x ) : 0 )

double ga_function052(
    UserStr     *user )          // user struct
{
    int         i, k, K, D, I, *CVK;
    float       X, Z;
    double      a, b, c, d, e, f, p, q;
    double      aa, bb, cc, dd, zz;
    double      chisq = 0;

    a = user->param[0];          // parameters
    b = user->param[1];
    c = user->param[2];
    d = user->param[3];
    e = user->param[4];
    f = user->param[5];
    CVK = user->categ[0];
//  loop over all women
    for ( k = 0; k < user->data->length; k++ ){
        p = user->param[6];      // reset memory
        q = user->param[7];      // reset memory

//      loop over times series per woman
        for ( i = 0; i < user->data->nr[k]; i++ ) {
            X = user->data->f1[k][i];        // age
            Z = user->data->f3[k][i];        // P-value
            D = user->data->b3[k][i];        // O-3
            I = user->data->b8[k][i];        // O-8
            K = CVK[user->data->c1[k][i]];   // HPV

//          the algorithm starts here
            aa = q  * a  ;
            aa = log( aa );
            aa = p  + aa ;
            bb = aa - K  ;
            aa = sqrt( bb );
            bb = aa - Z  ;
            aa = atan( bb );
            aa = log( aa );
            bb = I  * b  ;
            bb = ( bb ) ? D  / bb : 0;
            bb = K  - bb ;
            cc = bb * c  ;
            dd = cc + d  ;
            bb = aa * dd ;
            aa = atan( bb );
            bb = ( X  ) ? e  / X  : 0;
            cc = bb + I  ;
            bb = aa * cc ;
            aa = log( bb );
            aa = log( aa );
            aa = q - aa ;
            bb = ( X  ) ? f  / X  : 0;
            bb = q - bb ;
            cc = bb * Z  ;
            bb = p - cc ;
            bb = sqrt( bb );
            bb = aa - bb ;
            aa = -( bb );

            user->data->mock[k][i] = zz = aa;  // result
            zz = zz - user->data->out[k][i];   // residual
            chisq += zz * zz;
        }
    }
    return chisq;  // sum of squared residuals
}
```


SRON
Netherlands Institute for Space Research

Another bit of Bayes

On the level of models Bayes rule:

Pr(M|M) * Pr(D|MM) = Pr(D|M) * Pr(M|DM)

M is the class of models accessable by the genotype definitions.

Pr(M|M): prior for the model

Pr(D|MM): likelihood == evidence of previous level

Pr(D|M): evidence for this class of models

Pr(M|DM): posterior for the model

The prior of a model is related to its length

$Pr(M|M) = \exp(- 0.1 * L_{genotype})$

Nested Sampling II

Play the nested sampling game again.

1. make ensemble of 100 models
2. calculate the evidence using nested sampling.
3. select the one with the lowest evidence.
4. copy one of the other onto it
5. make some variation of the model.
6. calculate the evidence.
7. if evidence is higher than it was, accept and go to 3
8. else reject and go to 5.

SRON
Netherlands Institute for Space Research

Variation I

Successful individuals reproduce by

- mutation: change an item into another of the same kind
  - lAY-aK-bXc*/d*-*e+ becomes lAY-aK-bXc*/d*-*e*
- addition: add some item(s) somewhere
  - lAY-aK-bXc*/d*-*e+ becomes lAY-aK-bXc*/d*-*J/e+
- deletion: delete some item(s)
  - lAY-aK-bXc*/d*-*e+ becomes lY-aK-bXc*/d*-*e+
- cross over: combine 2 chromosomes.
  - lAY-aK-bYR+*- and

    KZaXb*/c*-*J/d+ becomes lAY-aK-bXc*/d*-*J/e*

Variation II

- insertion: insert (part of) a chromosome into another.
  - lAY-aK-bYR+*- and
    KZaXb*/c*-*J/d+ becomes lAY-aK-bcXd*/e*-YR+*-
- memory: introduce a memory pair.
  - lAY-aK-bYR+*- becomes lAY-p+aK-bYR+*P-
- random: construct a random chromosome.
  - aX*b+R

Computation

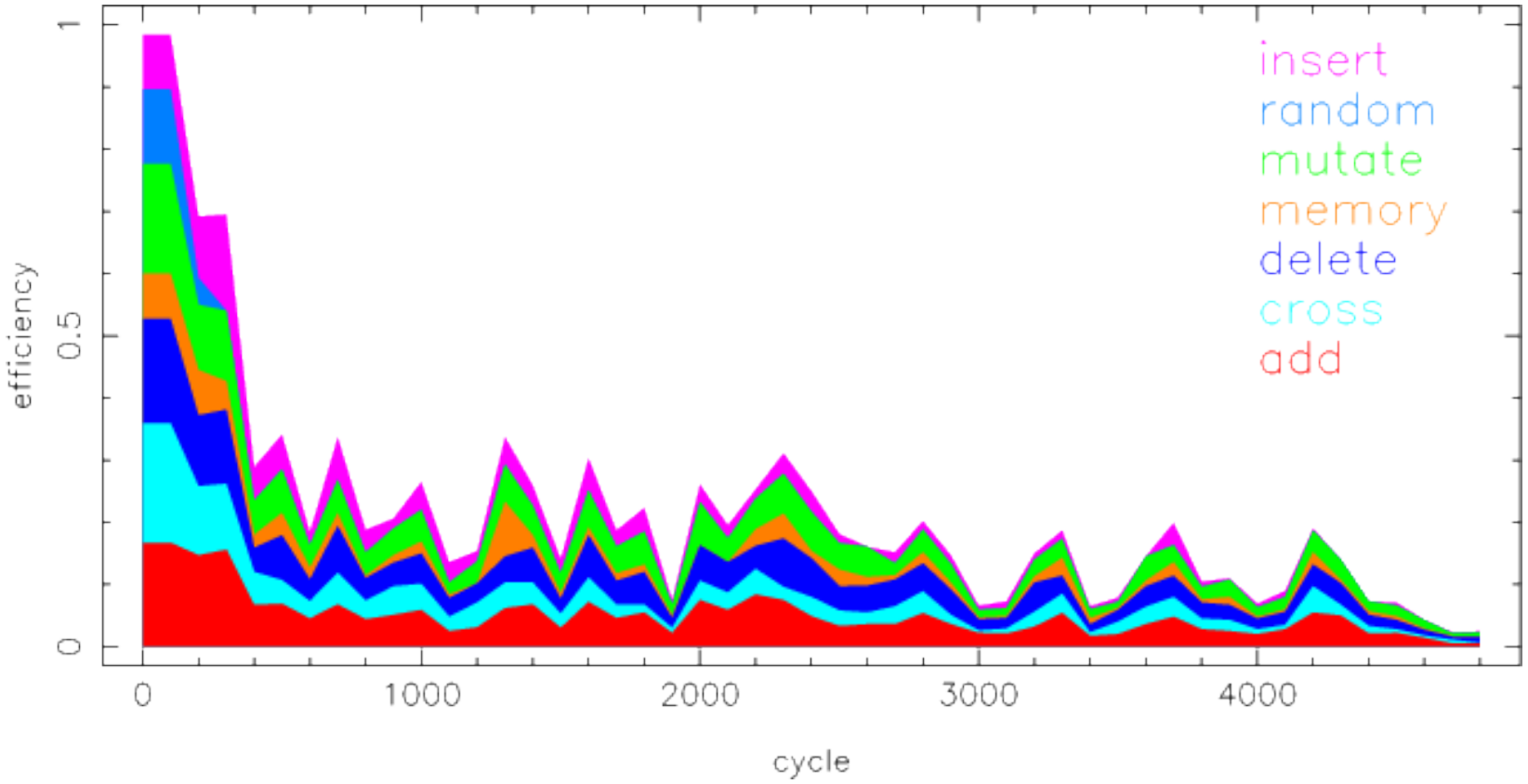
With all parts in place we can start the computation.

...

100 days of CPU later nested sampling actually converged.

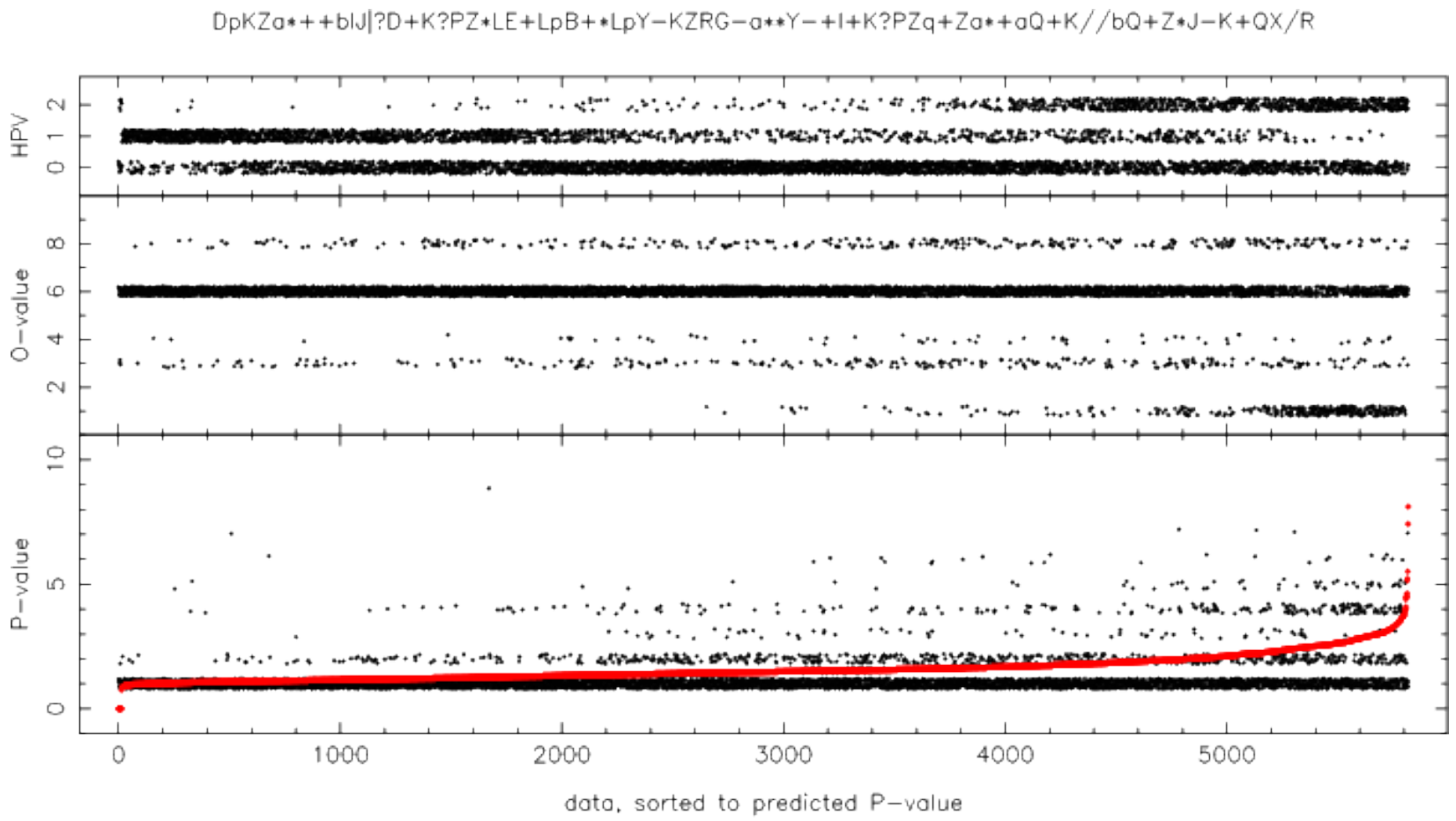
- 4800 iterations
- H = 23.9
- 40000 models were visited out of a possible 10^{60} models.

Efficiency



SRON
Netherlands Institute for Space Research

Best Model



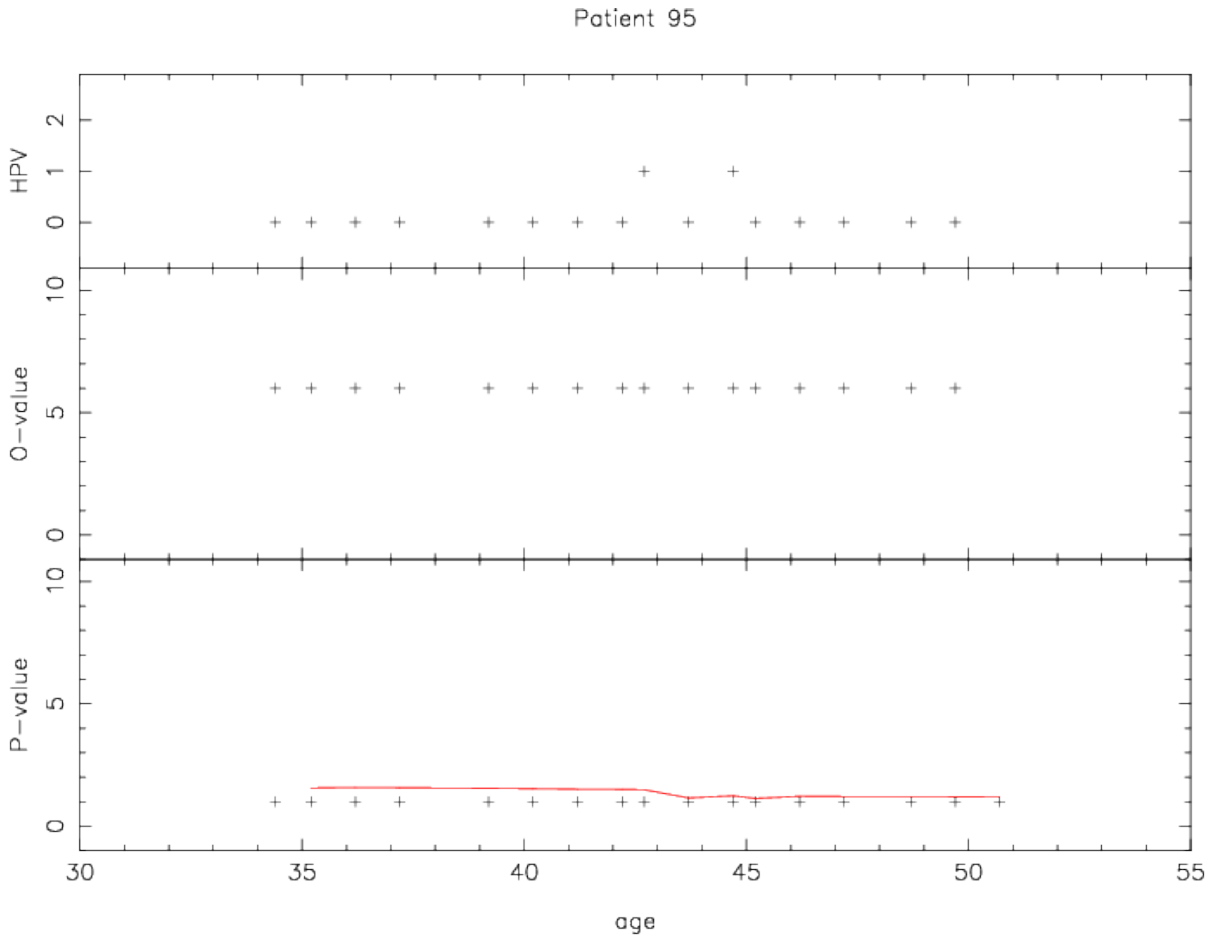
Bayes & MaxEnt

Chamonix 05-July-2010

SRON

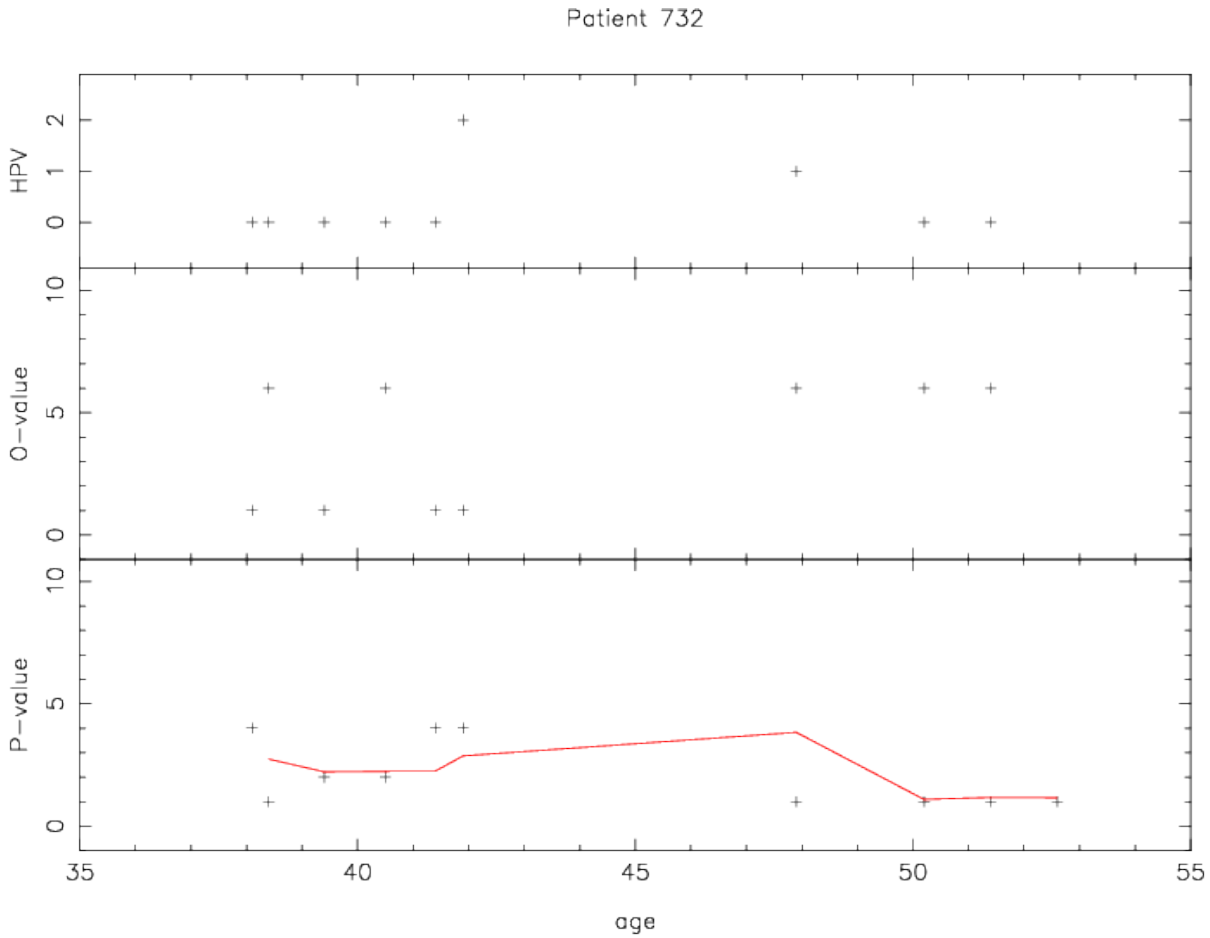
Netherlands Institute for Space Research

Time series I



SRON
Netherlands Institute for Space Research

Time Series II



SRON
Netherlands Institute for Space Research

Conclusions

- Data
  - All known connections were found.
  - The data has not enough information to find something new.
- Program
  - Models tend to get longer
  - It is harder to find new ones within the constraints
    - continuity in models ???
- Evolution
  - Introns (pieces of code which do nothing) appear
    - ...SSSS...
  - Ugly code, but it works.
    - Designed code looks better than evolved code