Prediction of Extreme Events

Thanks to
Sarah Hallerberg,
Eduardo G. Altmann, Jochen Bröcker, Detlef Holstein
Volker Jentsch, Mario Ragwitz, Nikolay K. Vitanov

Max Planck Institute for the Physics of Complex Systems, Dresden

Outline

1. Prediction and scoring
2. Performance of predictors: Model processes
3. Extreme events in experimental data
   - Free- Jet Experiment
   - Wind speed prediction
   - Predicting Failures of Weather Forecasts
4. Summary

Classification Extreme Events

- **No unique definition of the term "extreme events"!**

"Dresden-Classification" of Extreme Events

We are interested in events,

- which are rare
- which occur irregularly due to a complex stochastic or deterministic dynamic
- which are recurrent (here: do not end the lifetime of the system)
- which are inherent to the system under study (endogenous), not due to strong external perturbation
- to which we can assign a variable ("magnitude")

Key issues

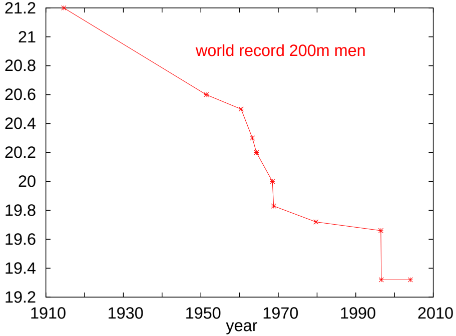
- What is the magnitude distribution of events?
  What are the most extreme events in a given system?
- Are there temporal (or spatial) correlations between EE?
- What does a sequence of "records" tell about drifts or trends?
- Can we predict the next EE?
- What are the costs caused by wrong predictions?
- Can one control/manipulate the system to avoid a predicted event?

Answers require understanding of the **dynamics** of EE!

Example: Trends from records

Records:

overcoming all prior values (e.g., sports, daily maximum temperatures, floods)

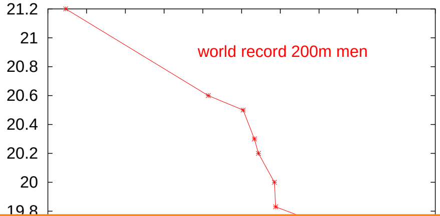


Does a sequence of ever increasing records reflect a trend?

Example: Trends from records

Records:

overcoming all prior values (e.g., sports, daily maximum temperatures, floods)



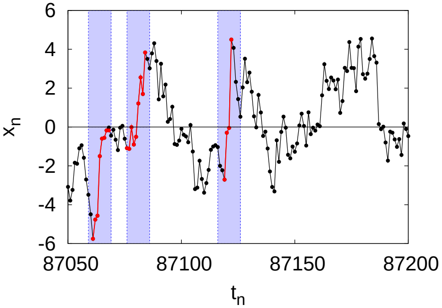
Why potentially no trend?

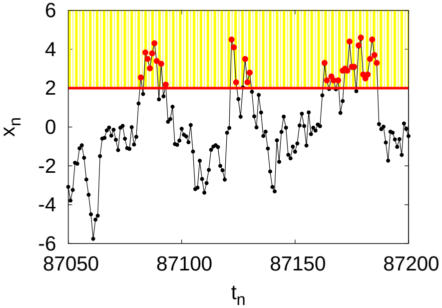
Finite sample effect: Growing sample size implies an increase of the observed largest (decrease of the smallest) number!

Does a sequence of ever increasing records reflect a trend?

Extreme Events in Time Series

General situation: event time series $\{Y_n\}$, $Y_n \in \{0, 1\}$,
Observation time series $\{x_n\}$.
Try to predict event with index $n+1$ from observations up to time index n.
Often: Events are defined on the observations $\{x_n\}$ themselves:

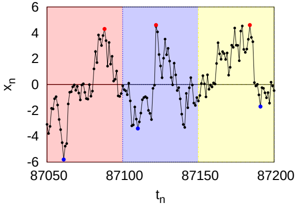




$$Y_{n+1}(\eta) = \begin{cases} 1: & x_{n+k} - x_n \geq \eta \\ 0: & \text{else} \end{cases}$$

$$Y_{n+1}(\eta) = \begin{cases} 1: & x_n \geq \eta \\ 0: & \text{else} \end{cases}$$

Extreme Value Statistics



- traditionally for i.i.d. random variables and block maxima
- generalization for threshold crossings, and correlated variables exist
- asymptotics of the cumulative distribution function $\mathbb{P}(M_n \leq z)$
- return levels z_p, which are exceed on averaged every $1/p$ time steps,
- no forecast!

References: E. J. Gumbel, S. Coles

Prediction

(Stochastic) dynamical system

State space plus evolution equations

Extreme event = large deviation from the system's normal behaviour

state vector far off its mean, but in a well defined subset of phase space

Ideal situation: detailed physical model, observed current state

Run the model to predict the future (on short times).

- computing orbits of astrophysical objects (satellites, meteorits)
- weather forecasts (really?)

Less ideal but more relevant situation: time series data.

Useful?

Prediction from time series data

general stochastic process

time series $\{x_i\}$, $i = 1, \ldots, N$:
Process is fully characterised by
all joint probabilities $p(x_{i_1}, x_{i_2}, \ldots, x_{i_l})$.
future is determined by
conditional probabilities $p(x_{i_1}|x_{i_2}, \ldots, x_{i_l}) := \frac{p(x_{i_1}, x_{i_2}, \ldots, x_{i_l})}{p(x_{i_2}, \ldots, x_{i_l})}$
Two more assumptions:
stationarity: only relative time indices are relevant
fast decay of dependence: good approximations by finite conditioning (Markov property).

Events

Event time series $\{Y_i\}$, $i = 1, \ldots, N$, $Y_i \in \{0, 1\}$.
$p(Y = 1|x_{i_2}, \ldots, x_{i_l})$ describes probability of an event to happen.

Prediction and cost functions

a) probabilistic forecasts

Probabilistic predictor

A map $(x_i, x_{i-1}, \ldots, x_{i-k+1}) \mapsto \hat{p}$,
probability of the event to happen, $\hat{p} \in [0, 1]$

Cost function (score)

Brier score: $S_B = \langle (\hat{p}_i - Y_i)^2 \rangle$.
benchmark: constant prediction $\hat{p} = r$, where r = event rate,
then $S_B = r(1 - r)$.

Two problems

Brier score depends explicitly of rate r,
Brier score has a bias towards trivial prediction for $r \to 0$.

Prediction and cost functions

b) Deterministic forecasts

deterministic predictor

A map $(x_i, x_{i-1}, \ldots, x_{i-k+1}) \mapsto \hat{Y}_i$,
Predicts a value of the event series, $\hat{Y}_i \in \{0, 1\}$.

Classical cost function

root mean squared (rms) error (for real-valued variables)
$\hat{s}_i$ prediction of s_i (the true observation)

$$\bar{e} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{s}_i - s_i)^2}$$

(for predicting chaos ([Farmer & Sidorowich 1987] and many others).

Prediction of extreme events as classification task

Three problems with rms-errors

- rare events contribute with a small weight
- involves a norm (symmetric)
- when $\hat{Y}$ is inferred from $\hat{x}$, a small error in x may change the value of $\hat{Y}$!

Prediction of the occurence of events involves two types of errors

no event predicted, event takes place (missed hit)
event predicted, no event takes place (false alarm)

These two types of errors might cause very different costs.
(consider earthquake striking a city, costs for evacuation)

Prediction of events

Probabilistic prediction

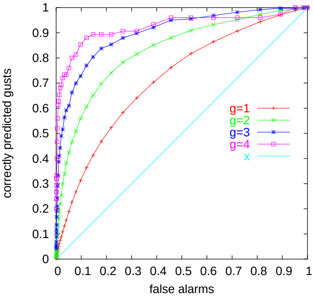
convert predicted $\hat{p}_n$ into a "warning" $\hat{Y}_n$ by threshold p_c:
if $\hat{p}_n \geq p_c$: $\hat{Y}_n = 1$, $\hat{p}_n < p_c$: $\hat{Y}_n = 0$

Deterministic prediction for $\hat{Y}_n$ through precursors

Precursor: Specific pattern of m succesive observations x_k which typically preceeds an event $Y_{n+1} = 1$, called $\mathbf{x}_{pre}$.
Alarm volume V_δ is the δ-neighbourhood of $\mathbf{x}_{pre}$.
(max-norm: a tube of diameter δ around the pattern.)
Event is likely to occur at time $n+1$ if $\mathbf{x}_n \in V_\delta$.
Randomness:
not every event is preceeded by the precursor.
not always is the precursor followed by an event.
Notice: precursors $\mathbf{x}_{pre}$ are elements of a delay embedding space.

Receiver Operating Characteristics

hit rate = (number correctly predicted events)/(all events in data set)
false alarm rate = (number of false alarms)/(all non-events in data set)
ROC-statistics:



hit rate versus false alarm rate as a function of the total number of alarms (sensitivity).
(useless) random predictor: hit rate = false alarm rate.
(compare: medical diagnostics: sensitivity versus specificity)

How to find "good" precursors?

Strategy I (the "intuitive" one)

find all events in the data base, study the preceeding time series segments.
define precursor as $\mathbf{x}_{pre} : \mathbb{P}(\mathbf{x}_{pre}|Y=1) = \max.$

Strategy II

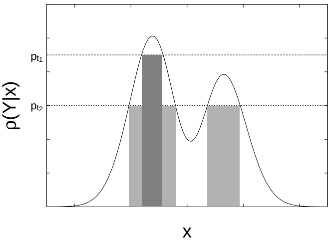
Study $\mathbb{P}(Y|\mathbf{x})$ for all possible values of $\mathbf{x}$,
define the precursor as $\mathbf{x}_{pre} : \mathbb{P}(Y=1|\mathbf{x}_{pre}) = \max.$

Remark: $p(a|b) = p(b|a)\frac{p(a)}{p(b)}$ Bayesian theorem.
Remark: Strategy I is used in machine learning ("learn pairs").
Is this the best we can do? Theoretical motivation?
Optimize the ROC statistics!

Result: Uniformly superior prediction scheme

Optimal probabilistic predictor

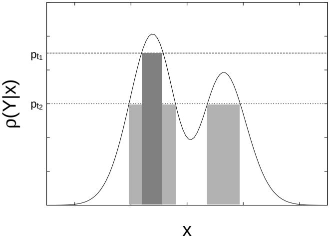
$\hat{p}_n = \mathbb{P}(Y_{n+1} = 1 | \mathbf{s}_{n,\tau})$
Possibly convert $\hat{p}_n$ into $\hat{Y}_n$ by threshold p_c.



Result: Uniformly superior prediction scheme

Optimal precursor for deterministic prediction

- Structure $\mathbf{x}$ which maximizes the conditional probability $\mathbb{P}(Y_{n+1} = 1|\mathbf{x}_{n,\tau})$ (*Bayesian estimate* for the optimal precursor)
- since τ is finite we neglect the past of the process, which is farther away than τ steps (pragmatic approach)
- superiority of $\mathbb{P}(Y_{n+1} = 1|\mathbf{x}_{n,\tau})$ to $\mathbb{P}(\mathbf{x}_{n,\tau}|Y_{n+1} = 1)$



Numerical algorithm

Fix the "embedding window" τ.
Estimate the conditional probability $\mathbb{P}(Y_{n+1} = 1|\mathbf{x}_{n,\tau})$ from data record

Two possibilities:

τ is small: Binning and counting
τ large: kernel estimator with kernel width δ:
$\mathbb{P}(Y_{n+1} = 1|\mathbf{x}_{n,\tau}) \approx \frac{1}{||\mathcal{U}_\epsilon(\mathbf{x}_n)||} \sum_{k:\mathbf{x}_k \in \mathcal{U}_\delta(\mathbf{x}_n)} Y_{k+1}$
(relative number of events in neighbourhood of $\mathbf{x}_n$)
Notice: Order of the Markov model/ memory depth τ enters through the definition of the neighbourhood $\mathcal{U}_\epsilon(\mathbf{x}_n)$
Compare: zeroth order predictor [Farmer & Sidorowivh, 1987], Local random analogue predictor [Paparella et al., 1997]

Extreme increments in a simple AR(1) model

Extreme event: increment $x_{k+1} - x_k > \eta$

process: $x_{n+1} = ax_n + \xi_n$, white noise ξ_k, $|a| < 1$.
Conditioning: $m = 1$, precursor is a single number
$x_k \in [x_{pre} - \delta, x_{pre} + \delta] \longrightarrow$ predict an event to follow at time $k + 1$.

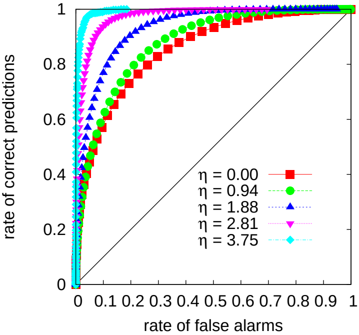
Extreme increments in a simple AR(1) model

Extreme event: increment $x_{k+1} - x_k > \eta$

process: $x_{n+1} = ax_n + \xi_n$, white noise ξ_k, $|a| < 1$.
Conditioning: $m = 1$, precursor is a single number
$x_k \in [x_{pre} - \delta, x_{pre} + \delta] \rightarrow$ predict an event to follow at time $k + 1$.

Analytical results [Hallerberg et al. (2007)]

Strategy II superior to strategy I
The more extreme the increment to be predicted (η), the better the predictability.



Extreme increments in simple models

More results:

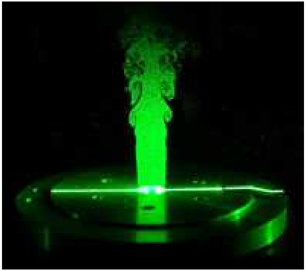
- Numerically equlivalent results for long-range correlated Gaussian data.
- more analytics (compute the slope of the ROC curve at the origin and its derivative with respect to η):
- symmetric exponential distribution: no systematic dependence of predictability on η.
- Power law tails: predictability drops with increasing η.

Threshold crossing in simple models

Feed AR(1) model with noises of different distributions, define events by threshold crossing:
Restrict prediction trials to situations where last observation is below threshold.
Larger Magnitude events are always better predictable.

A free-jet experiment

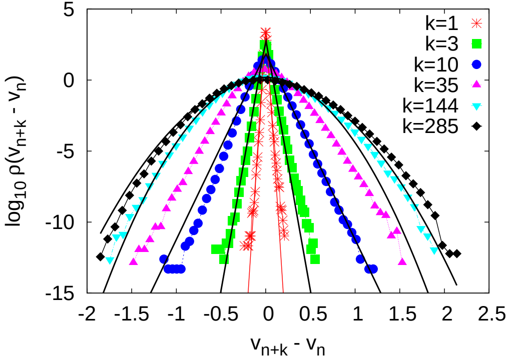
- free jet data (C. Renner, J. Peinke, R. Friedrich, *Experimental indications for Markov properties of small-scale turbulence*, J. Fluid Mech. (2001))



(fluid.jku.at/hp/images/stories/research/jet.jpg)

Free jet velocity increments

Use as time series data $x_n = v_{n+k} - v_n$, velocity increments



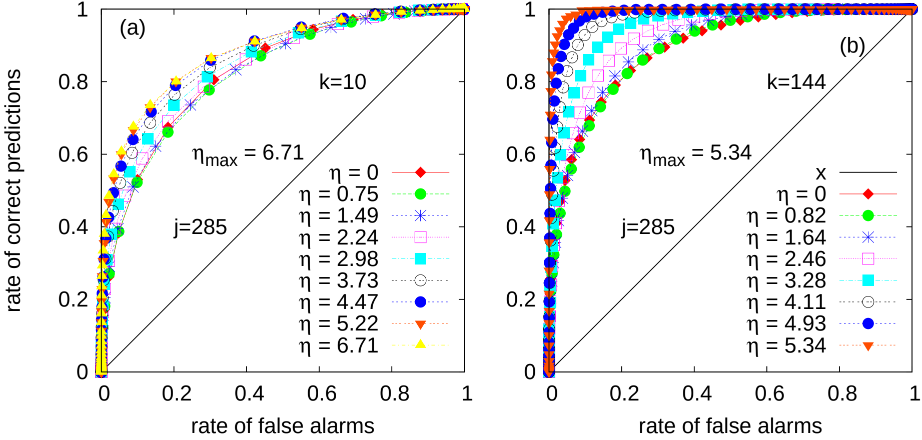
- predict increments of increments:, $a_n = v_{n+k} - v_n$;

$$Y_{n+j}(\eta) = \begin{cases} 1: & a_{n+j} - a_n \geq \eta \\ 0: & \text{else} \end{cases}$$

Prediction of free-jet velocity increments

- transition:

"exponential ROC" ⇒ "Gaussian ROC"



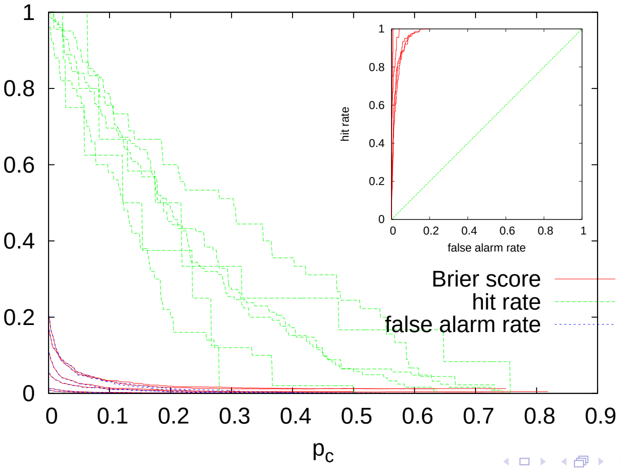
Wind speed prediction

- Lammefjord measurement site: recording wind velocity vectors
- Data: modulus of horizontal wind speed measured 20 m and 30 m above ground with 8Hz resolution, 1 day of data (691200 data items)
- events:
  a) threshold crossing from below at 2 time steps in the future
  b) large positive incrments (wind gusts),

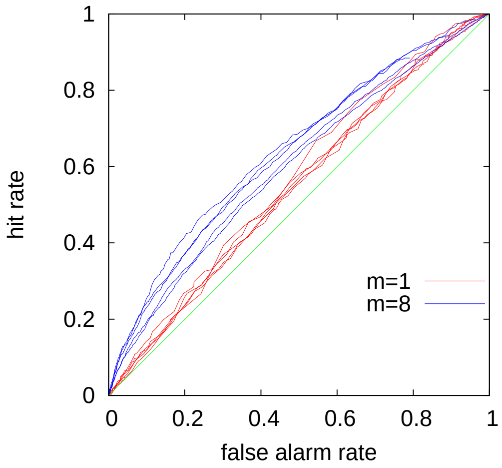
Wind speed prediction

Results

Threshold crossings:
deterministic prediction: threshold on predicted probability:
Brier score is dominated by false alarm rate



large increments (gusts): Conditioning improves forecasts, comparison of $m = 1$ to $m = 8$:

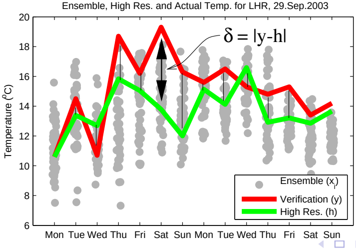


Predicting Failures of Weather Forecasts

- joint work of S. Hallerberg with **Jochen Broecker and Leonard A. Smith, LSE, London**
- Absolute error of high resolution forecast h with respect to verification y

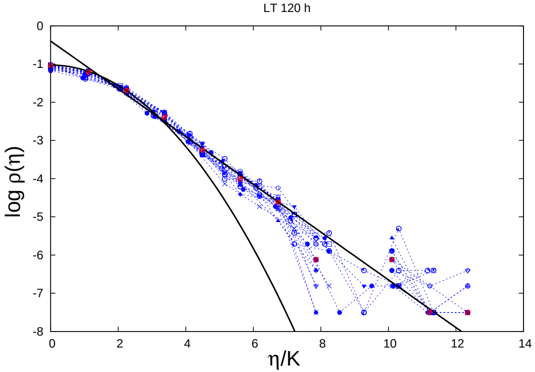
$$Y = \begin{cases} 0 & if \quad |y-h| < \eta \\ 1 & if \quad |y-h| \geq \eta \end{cases}$$

- Predictions are made using the number of ensemble members showing a large error $\rho = \#\{i, |y - x_i| \geq \eta\}$

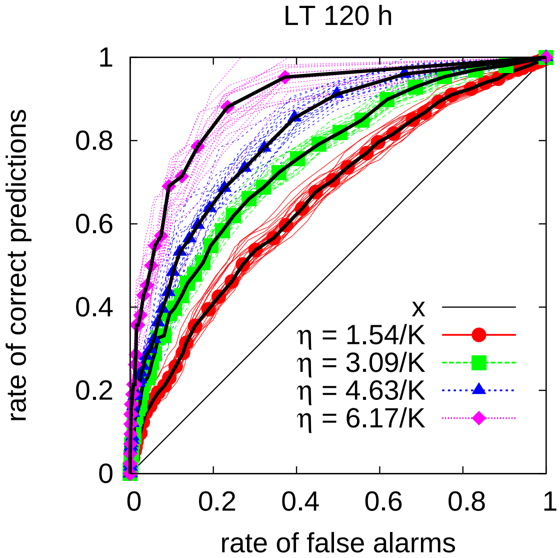


Predicting Failures of Weather Forecasts

- Distribution of $|y - h|$ and $|y - x_i|$ exhibit gaussian behavior for smaller η but have an exponential tail
- Weather data sets consist of only 1800 data



Predicting Failures of Weather Forecasts



Conclusions

- Complex dynamics generates rare and extreme events
- pseudo-embedding strategies for stochastic dynamics
- Different precursor strategies for predictions of extremes
- Dynamics enters through $P(X|\mathbf{x})$ and $P(\mathbf{x}|X)$
- The optimal strategy is different form standard machine learning rules
- Gaussian statistics: Larger events are better predictable than smaller events
- Statistically significant predictablity of wind speeds
- General flaw of this approach: cannot predict previously unobserved event magnitude

How to measure the quality of a prediction?

Overview of different measures

<table>
<tr><td>Predictability</td><td rowspan="2">study predictability as a property of the system</td><td rowspan="2">make use of the whole PDF $\mathbb{P}(Y_j = 1|\mathbf{s})$, for all $\mathbf{s}$
$\rightarrow$ do not consider the selection of the precursor</td></tr>
<tr><td>Kullback-Leibler distance</td></tr>
<tr><td>Brier Score</td><td rowspan="3">compare forecasts and observation of events</td><td rowspan="2">dependent on the relative frequency of an event, due to averaging $\frac{\sum_j f(\mathbb{P}(Y_j=1|\mathbf{s}))}{N}$</td></tr>
<tr><td>Ignorance</td></tr>
<tr><td>ROC-curve</td><td>independent on the relative frequency of an event</td></tr>
</table>

$\Rightarrow$ we will use the Receiver Operator Characteristic Curve (ROC-Curve)

Predictability and Kullback- Leibler distance

Predictability

$$P(\mathbf{s}_n, Y_{n+\tau}) = 1 + \frac{H(Y_{n+\tau})}{H(\mathbf{s}_n)} - 2\frac{H(\mathbf{s}_n, Y_{n+\tau})}{H(\mathbf{s}_n)}$$

$$\text{with} \quad H(\mathbf{s}_n) = -\sum_{\mathbf{s}_n} p(\mathbf{s}_n)\log_2 p(\mathbf{s}_n);$$

Kullback- Leibler distance (relative Entropy)

$$D(\rho(\mathbf{s}|Y=1)||\rho(\mathbf{s}|Y=0)) = \sum_{\mathbf{s}} \rho(\mathbf{s}|Y=1)\log_2\left(\frac{\rho(\mathbf{s}|Y=1)}{\rho(\mathbf{s}|Y=0)}\right)$$

- both measures average over the whole possible range of precursory structures

Scores

Brier score

$$b(\mathbf{s}_n, Y_{n+1}) = \frac{1}{N}\sum_{n=0}^{N}(Y_{n+1} - \rho(Y_{n+1} = 1|\mathbf{s}_n))^2$$

- Relative brier score $b_{rel} = b_0 - b/b_0$;
  with b_0 calculated from the relative frequency of events

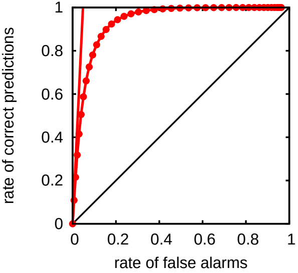
Ignorance (for a binary forecast)

$$I(\mathbf{s}, Y) = -\frac{1}{N}\sum_{n=0}^{N}\log\left[2\rho(Y_{n+1} = 1|\mathbf{s}_n)Y_{n+1} + 1 - Y_{n+1} - \rho(Y_{n+1} = 1|\mathbf{s}_n)\right]$$

- Relative ignorance $I_{rel} = I_0 - I/I_0$;
  with I_0 evaluated using the relative frequency of events

The Receiver Operating Characteristic Curve

- ROC-curve in signal detection theory (Egans 1975), medicine, machine learning, multi-dim. classification problems (Srinivasam 1999, Fieldsend et al. 2005)



$$r_c = \frac{\#\text{correctly predicted events}}{\#\text{events}}$$

$$r_f = \frac{\#\text{false alarms}}{\#\text{non-events}}$$

$\Rightarrow$ independent on the relative frequency of events

- slope m of the ROC-curve in the vicinity of the origin (*likelihood ratio*)

$$m(Y, \mathbf{s}_{(n,\tau)}) \approx \left.\frac{\Delta r_c}{\Delta r_f}\right|_{\delta=0} = \frac{\rho(\mathbf{s}_{(n,\tau)} | Y_{n+1} = 1)}{\rho(\mathbf{s}_{(n,\tau)} | Y_{n+1} = 0)}$$

Comparison

<table>
<tr><td>Predictability</td><td rowspan="2">study predictability as a property of the system</td><td rowspan="2">make use of the whole PDF $\mathbb{P}(Y_j|\mathbf{s})$, for all $\mathbf{s}$
$\rightarrow$ do not consider the selection of the precursor</td></tr>
<tr><td>Kullback- Leibler distance</td></tr>
<tr><td>Brier Score</td><td rowspan="3">compare forecasts and observation of events</td><td rowspan="2">dependent on the relative frequency of an event, due to averaging $\frac{\sum_j f(\mathbb{P}(Y_j|\mathbf{s}))}{N}$</td></tr>
<tr><td>Ignorance</td></tr>
<tr><td>ROC-curve</td><td>independent on the relative frequency of an event</td></tr>
</table>

$\Rightarrow$ we will use the Receiver Operator Characteristic Curve (ROC-Curve)