Regression, Mutual Information and Point Processes

David R. Brillinger
Statistics Department
University of California, Berkeley

www.stat.berkeley.edu/~brill

$2\pi \neq 1$

$2\pi \neq 1$

$2\pi \neq 1$
"I hate getting all these Canadian coins, but I guess that's the price of living in Toronto."
Layout.

1. Introduction
2. Regression
3. Mutual information
4. Examples - neuron, soccer, river flow
5. Extensions
6. Summary
7. References
8. Some proofs
1. **Introduction.**

Science studies relationships

Regression analysis studies relations between variables \( Y \) and \( X \)

Question: what is the strength of a particular relationship?

One answer: the coefficient of determination

Another answer: the coefficient of mutual information
2. Classical regression.

Coefficient of determination

$$\rho_{XY}^2 = \text{corr}\{X,Y\}^2 \quad X, Y \text{ real-valued}$$

Symmetric and invariant for:

1) independence
2) explained variation
3) linear dependence
4) uncertainty of estimates

One wishes for more!

Discrete variates.

\[ \text{Prob} \{ X_j = j, Y_k = k \} = p_{jk} \]

\[ I_{XY} = \sum_{j,k} p_{jk} \log \frac{p_{jk}}{p_j p_k}, \quad p_{jk} \neq 0 \]

Continuous variates.

Given \( x_j \in \delta_j, y_k \in \Delta_k \) and \( |\delta|, |\Delta| \) small,

\[ \text{Prob} \{ X \in \delta_j, Y \in \Delta_k \} \approx p(x_j, y_k) |\delta| |\Delta| \]

leading to

\[ \iint p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)} \, dx \, dy, \quad p \neq 0 \]
Prediction - lower bound

\[ E \{ Y - g(X) \}^2 \geq \frac{1}{2\pi e} \exp\{2(I_{YY} - I_{XY})\} \]

Incorrect model - upper bound

\[ \iint p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)} \, dx \, dy \geq \iint p(x,y) \log \frac{q(x,y)}{q_X(x)q_Y(y)} \, dx \, dy \]

\[ \geq E_{X,Y} \left[ \log \frac{q_Y|X(X|X)}{q_Y(Y)} \right] \]

E.g. take \( q \) to be bivariate normal
Properties of $I_{XY}$.

1) Non-negative, $I_{XY} \geq 0$

2) Invariant, $I_{XY} = I_{UV}$ if 1-1 transformations

3) Measures strength of dependence
   
   i) $I_{XY} = 0 \iff X \text{ indep } Y$
   
   ii) $I_{XY} = \infty$ if $Y = g(X)$
   
   iii) $I_{XZ} \leq I_{XY}$ if $X \text{ indep } Z | Y$ \quad $X - Y - Z$

4) Bivariate normal, $I_{XY} = .5 * \log(1 - \rho_{XY}^2)$
Disadvantage - *IM* is "just" a number

**Uses.**

*Questions* - change? trend? serial correlation?  
dimension? model fit? ...

*Estimation* -

lag  
image registration  
selection of variables  
model selection  
association  

...
**Parametric estimation.** Data \((x_i, y_i, i=1,...,n)\)

Model \(p(x, y \mid \theta)\), with \(X\) and \(Y\) independent when \(\theta = 0\)

\[
p(x, y \mid 0) = p_X(x)p_Y(y)
\]

Test of independence of \(X\) and \(Y\)

\[
\hat{I}_{XY} = -\log(likelihood \ ratio)/n
\]

\((n \ \text{sample size})\)

Approximate null distribution, when \((X_i, Y_i)\) independent

\[
\chi^2_v / 2n, \ \nu = \text{dim}(\theta)
\]

\[
\int \int p(x, y) \log p(x, y) dx dy - \int \int p(x)p(y) \log p(x)p(y) \ dx dy
\]
Non-parametric estimation.

\( \hat{p}(x, y) \) estimate of \( p(x, y) \), e.g. histogram or kernel-based

\[
I_{XY} = \sum_{j,k} \hat{p}(u_j, v_k) \log \frac{\hat{p}(u_j, v_k)}{\hat{p}_X(u_j)\hat{p}_Y(v_k)}
\]

Approx null distribution

\[
\chi^2_v / 2n, \quad v = (J-1)*(K-1)
\]

Non-null

\[
\hat{I}_{XY} \to I_{XY} \text{ in prob, etc.}
\]

(Entropy - Bilmes, Fernandes, Hall & Morton, Joe, Kozachenko & Leonenko, Parzen, Robinson)
The point process case. Isolated points

Univariate. points \( \{\tau_k\} \), counts \( N(t) = \#\{\tau_k \leq t\} \),

intervals \( \{Y_k = \tau_{k+1} - \tau_k\} \), history \( H^t_N = B(\tau_k \leq t) \)

conditional intensity

\[ \text{Prob}\{dN(t) = 1 \mid H^t_N\} \approx \mu_N(t)dt \]

likelihood

\[ \Pi_k \mu_N(\tau_k)\exp\left\{-\int_0^T \mu_N(t)dt\right\} \]

entropy

\[ E\left\{ \int_0^T \log \mu_N(t)\;dN(t) - \int_0^T \mu_N(t)dt \right\} \]
Bivariate. \( M(t) = \# \{ \sigma_j \leq t \} \), \( N(t) = \# \{ \tau_k \leq t \} \)

history \( H_{MN}^t = B(\sigma_j, \tau_k \leq t) \)

\[
\text{Prob} \{ dM(t) = 1 \mid H_{MN}^t \} \approx \gamma_M(t) dt, \quad \text{Prob} \{ dN(t) = 1 \mid H_{MN}^t \} \approx \gamma_N(t) dt
\]

entropy

\[
E_{M,N} \left\{ \int_0^T \log \gamma_M(t) dM(t) + \int_0^T \log \gamma_N(t) dN(t) - \int_0^T \gamma_M(t) dt - \int_0^T \gamma_N(t) dt \right\}
\]

mutual information

\[
E_{M,N} \left\{ \int_0^T \frac{\gamma_M(t) \gamma_N(t)}{\mu_M(t) \mu_N(t)} dt \right\}
\]
Examples.

a) *Interval analysis - Aplysia californica*

Neuron, *L*10, discharging spontaneously

Spike train \( \{\tau_k\} \)

\( \{Y_k = \tau_{k+1} - \tau_k\} \) intervals between firings

Many models imply that intervals i.i.d. (renewal)

Estimate coefficient of determination and MI, as functions of lag, \( k \)
Intervals between firings of an Aplysia neuron

Coefficient of determination

The dashes give the critical level of 99%

Mutual information

The data
We estimate the 99% critical level via random permutations of the intervals.

There is evidence against the assumption of a renewal process

Unusual test of normality
The world's first game of soccer...

AHAJOKES.COM
b. Two discrete variables - soccer

Question - In which country is the relationship strongest between the number of goals teams score and and their playing at home?

Data for the Premier Leagues, 2001-2002

http://sunsite.tut.fi/rec/riku/soccer2.html

\[ Y = 0,1,2,3,4+ \text{ goals} \]
\[ X=1,0 - \text{ playing at home or away} \]
### Brazilian Serie A - Games played so far

<table>
<thead>
<tr>
<th>Date</th>
<th>Home Team</th>
<th>Away Team</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 10, 2002</td>
<td>Vasco</td>
<td>Figueirense</td>
<td>2 – 0</td>
</tr>
<tr>
<td></td>
<td>Sao_Paulo</td>
<td>Paysandu</td>
<td>4 – 2</td>
</tr>
<tr>
<td></td>
<td>Parana</td>
<td>Sao_Caetano</td>
<td>2 – 1</td>
</tr>
<tr>
<td></td>
<td>Santos</td>
<td>Botafogo-RJ</td>
<td>2 – 1</td>
</tr>
<tr>
<td></td>
<td>Goias</td>
<td>Portuguesa</td>
<td>3 – 1</td>
</tr>
<tr>
<td>Aug 11, 2002</td>
<td>Fluminense</td>
<td>Cruzeiro</td>
<td>5 – 1</td>
</tr>
<tr>
<td></td>
<td>Palmeiras</td>
<td>Gremio</td>
<td>1 – 1</td>
</tr>
<tr>
<td></td>
<td>At.Mineiro</td>
<td>Corinthians</td>
<td>1 – 2</td>
</tr>
<tr>
<td></td>
<td>Internacional</td>
<td>Flamengo</td>
<td>1 – 3</td>
</tr>
<tr>
<td></td>
<td>Guarani</td>
<td>Atletico-PR</td>
<td>2 – 1</td>
</tr>
<tr>
<td></td>
<td>Coritiba</td>
<td>Vitoria</td>
<td>1 – 0</td>
</tr>
<tr>
<td></td>
<td>Bahia</td>
<td>Gama</td>
<td>1 – 0</td>
</tr>
<tr>
<td></td>
<td>Juventude</td>
<td>Ponte_Preta</td>
<td>1 – 0</td>
</tr>
</tbody>
</table>

| Aug 14, 2002 | Sao_Caetano     | Fluminense        | 2 – 0  |
|              | Botafogo-RJ     | At.Mineiro        | 1 – 1  |
|              | Cruzeiro        | Palmeiras         | 1 – 1  |
|              | Gremio          | Vasco             | ? – ?  |
Discussion.

Marginal analyses

Games i.i.d. ?
Estimated MI - soccer

Argentina Brazil Uruguay Chile Mexico Canada England France Germany Italy Portugal Spain
The Great Mississippi River Flood of 1927, photographed in Illinois on March 25.

Dam 8 flow in 1960 & 1961

threshold at 10th percentile of yearly maxima

Cumulative count of upcrossings
c. Bivariate point process case. Mississippi River flow

Question - Strength of dependence between flows as function of distance between dams?

10 dams from St. Paul, Minnesota to Iowa

Daily flow from 1/1/1960 to 31/12/1997

\[ \{Y_i(t), t=0, \pm 1, \pm 2, \ldots, i=1, \ldots, 10\} \]

Have distances between dams

Upcrossing times of 10th percentile of series yearly maxima
Model.

Two 0-1 series $X_t$ and $Y_t$.

\[
\text{logit } \Pr(X_t=1|H_t) = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \beta_1 Y_{t-1} + \ldots + \beta_q Y_{t-q}
\]

\[
\text{logit } \Pr(Y_t=1|H_t) = \gamma_1 Y_{t-1} + \ldots + \gamma_p Y_{t-p} + \delta_1 X_{t-1} + \ldots + \delta_q X_{t-q}
\]

Include recent times and some from a year earlier.

Proceed via likelihood ratio test of no connection.
Discussion.

For asymptotics need full model to "fit".

Estimate spectral density matrix of residuals.
5. Extensions.

\( MI \) decomposes, \( I_{YX} = I_{YX_1} + I_{YX_2 | X_1} \)

Uncertainty (permutations, jackknife, bootstrap, ...)

Local estimates

Multivariate cases
Other estimates of entropy.

Discrete alphabet

$R_n$: time until the first $n$-string repeats

$$(\log R_n)/n \to \text{entropy} \quad a.s.$$

$MI$ is a concept extending correlation, substitutes for $r^2$ and $R^2$

"The hypothesis of independence is rejected."
becomes

"The estimated strength of dependence is $\hat{MI}$."

Functional forms useful, new parameters suggested

Google: corr 3,340,000 hits, MI 40,400
(5/31/2003)
7. References


8. Some proofs.

Jensen’s Inequality.

For $g$ convex

$$E\{g(Y)\} \geq g(E\{Y\})$$

with equality iff $Y$ is degenerate