



Master - QSAR anul I



Quantitative Structure Property Relationships (QSPR): From *Pearson-* to *Spectral-*, to *Quantum-* Correlations

Mihai V. Putz

Laboratory of Computational and Structural Physical Chemistry,
Chemistry Department, West University of Timișoara, Str. Pestalozzi
No.16, Timisoara, RO-300115, Romania, Tel: +40-256-592633; Fax. +40-
256-592620, Ems: mvputz@cbg.uvt.ro, mv_putz@yahoo.com;
www.cbg.uvt.ro/mvputz

1. Variance, Dispersion and Correlation

x y	$x_1 \dots$	$x_k \dots$	x_n
y_1	$p_{11} \dots$	$p_{1k} \dots$	p_{1n}
y_2	$p_{21} \dots$	$p_{2k} \dots$	p_{2n}
...
y_k	$p_{k1} \dots$	$p_{kk} \dots$	p_{kn}
...
y_m	$p_{m1} \dots$	$p_{mk} \dots$	p_{mn}

$$\sum_{i=1}^n \sum_{j=1}^m p_{ij} = 1$$

n/N mărimi de măsurat (x)
m/M stații în care se efectuează
măsurători (y)

$$P[(x, y) | \text{domeniu interes}] = \int \int_{x, y} f(x, y) dx dy = 1$$

■ Media măsurabilei:

$$\langle \hat{A} \rangle = \int_D \psi^* \hat{A} \psi d\tau = \int \hat{A} \psi^* \psi d\tau = \int \hat{A} |\psi|^2 d\tau = \int \int_{x,y} \hat{A} f(x,y) dx dy$$

$|\psi|^2$ Se numeste DENSITATE DE PROBABILITATE

■ Media pentru x:

$$\langle x \rangle = \int \int_D x f(x,y) dx dy$$

$$\langle (x - \langle x \rangle)^2 \rangle = \int \int (x - \langle x \rangle)^2 f(x,y) dx dy$$

$$\overline{(x - \bar{x})} = \sum_{i,j} (x_i - \bar{x}) p_{ij} = \sum_{i,j} [x_i - \sum_{i,j} x_i, x_j] p_{ij}$$

- Mărimi de măsurat $\{x_i\}$ și stații de măsurat $\{y_j\} \rightarrow$ nu au nimic în comun! Atunci:

$$p_{ij} = \begin{cases} p_i, \text{ pentru setul } \{x_i\}_n \\ p_j, \text{ pentru setul } \{y_j\}_m \end{cases} = \begin{cases} \frac{1}{n}, \text{ pentru setul } \{x_i\}_n \\ \frac{1}{m}, \text{ pentru setul } \{y_j\}_m \end{cases}$$

- Pentru seturi independente $\{x_i\}$ și $\{y_j\}$ și probabilitati uniforme:

$$\overline{(x - \bar{x})} = \sum_i \frac{1}{n} (x_i - \frac{1}{n} \sum_i x_i) = \frac{\sum_i (x_i - \frac{1}{n} \sum_i x_i)}{n}$$

$$P_x = \frac{1}{b_x - a_x}$$



- Abaterea lui x_1 fata de medie: $x_1 - \langle x \rangle$

- Abaterea lui x_2 fata de medie: $x_2 - \langle x \rangle$

.....

- Abaterea lui x_n fata de medie: $x_n - \langle x \rangle$

- Se folosesc pătratele pentru ușurința calculului și îndepărtarea erorilor:

$$(x_1 - \langle x \rangle)^2, (x_2 - \langle x \rangle)^2, \dots, (x_n - \langle x \rangle)^2$$

- se numește **DISPERSII**. **Dispersia lui x_i :**

$$D_x = \langle (x - \langle x \rangle)^2 \rangle = \int (x - \langle x \rangle)^2 f(x, y) dx dy$$

■ Abaterea standard (σ):

$$\sigma_x = \sqrt{D_x} = \begin{cases} \sqrt{\int (x - \langle x \rangle)^2 f(x, y) dx dy} & \text{varianta continua} \\ \sum_{i,j} (x_i - \sum_i x_i p_{ij})^2 p_{ij} & \text{varianta discreta} \\ \frac{1}{n} \sum_i (x_i - \frac{1}{n} \sum_i x_i)^2 & \text{pentru probabilitate uniforma} \end{cases}$$

$$D_x = \langle (x - \langle x \rangle)^2 \rangle$$

$$= \langle x^2 \rangle - \langle x \rangle^2$$

$$= \langle (x - \langle x \rangle)(x - \langle x \rangle) \rangle$$

$$= \langle x^2 - 2\langle x \rangle x + \langle x \rangle^2 \rangle$$

$$= \langle x^2 \rangle - 2\langle x \rangle \langle x \rangle + \langle x \rangle^2$$

$$= \langle x^2 \rangle - 2\langle x \rangle^2 + \langle x \rangle^2$$

$$\Rightarrow D_x = \iint x^2 f(x, y) dx dy - [\iint x f(x, y) dx dy]^2$$

- În unele cărți D_x se mai numește VARIANȚĂ



- Se introduce o nouă mărime
COVARIANȚA

$$\begin{aligned}C_{x,y} &= \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle \\&= \langle xy - x\langle y \rangle - \langle x \rangle y + \langle x \rangle \langle y \rangle \rangle \\&= \langle xy \rangle - \langle x \rangle \langle y \rangle - \langle x \rangle \langle y \rangle + \langle x \rangle \langle y \rangle \\&= \langle xy \rangle - 2\langle x \rangle \langle y \rangle + \langle x \rangle \langle y \rangle \\&= \langle xy \rangle - \langle x \rangle \langle y \rangle \\&= \sum_{i,j} x_i y_j p_{ij} - \left(\sum_i x_i p_i \right) \left(\sum_j y_j p_j \right) \\&= \sum_{i,j} x_i y_j p_{ij} - \frac{1}{n^2} \left(\sum_i x_i \right) \left(\sum_j y_j \right)\end{aligned}$$

$$C_{xy} = 0 \Rightarrow \langle xy \rangle = \langle x \rangle \langle y \rangle \Rightarrow f(x, y) = f(x)f(y)$$

Correlation Coefficient

$$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} = \frac{\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle}{\sqrt{\langle (x - \langle x \rangle)^2 \rangle} \sqrt{\langle (y - \langle y \rangle)^2 \rangle}}$$

■ Consideram:

$$\begin{cases} p_{ij} = p_i - p_j = \frac{1}{n} \\ n = m \end{cases}$$

$$r_{xy} = \frac{\frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}, |r_{xy}| \leq 1$$

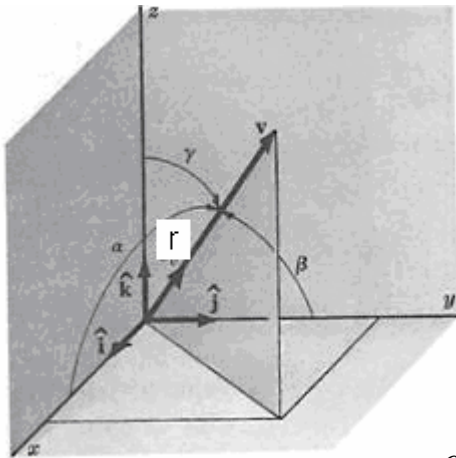
$$|x\rangle = |x_1, x_2, \dots, x_n\rangle, |y\rangle = |y_1, y_2, \dots, y_n\rangle$$

■ *Produs scalar:*

$$\langle x | y \rangle \stackrel{\text{def}}{=} \sum_i x_i y_i = x_1 y_1 + x_2 y_2 + \dots + x_n y_n$$

$$\Rightarrow \langle x | x \rangle = \sum_i x_i^2$$

- Modulul lui x = NORMA lui x : $\|x\| = \sqrt{\langle x | x \rangle} = \sqrt{\sum_i x_i^2}$



$$|\vec{r}| = \sqrt{x_1^2 + x_2^2 + x_3^2} = \sqrt{\langle x_1, x_2, x_3 | x_1, x_2, x_3 \rangle} = \sqrt{\langle r | r \rangle} = \|r\|$$

Distanța dintre vectori:

$$d(|x\rangle, |y\rangle) = \| |x\rangle - |y\rangle \| = \| x - y \| = \sqrt{\langle x - y | x - y \rangle} = \sqrt{\sum_i (x_i - y_i)^2}$$

■ Cauchy-Schwartz inequality:

$$\begin{aligned}
 \langle x - y | x - y \rangle &\geq 0 \\
 &= \langle x - ty | x - ty \rangle \geq 0 \\
 0 &\leq (\langle x | x \rangle - t \langle x | y \rangle)(\langle x | x \rangle - t \langle y | y \rangle) \\
 0 &\leq \langle x | x \rangle - t \langle x | y \rangle - t \langle y | x \rangle + t^2 \langle y | y \rangle \\
 0 &\leq t^2 \langle y | y \rangle - 2t \langle x | y \rangle + \langle x | x \rangle
 \end{aligned}$$

- ecuatie de gr II in $t \rightarrow$ este pozitiva daca are semnul lui a , adica for $\langle y | y \rangle$ (care este +), deci $\Delta \leq 0$

$$a = \langle y | y \rangle, b = -2 \langle x | y \rangle, c = \langle x | x \rangle$$

$$\Delta = b^2 - 4ac = 4(\langle x | y \rangle)^2 - 4 \langle y | y \rangle \langle x | x \rangle \leq 0$$

$$\Rightarrow (\langle x | y \rangle)^2 \leq \langle y | y \rangle \langle x | x \rangle \quad \Rightarrow |\langle x | y \rangle| \leq \sqrt{\langle y | y \rangle} \sqrt{\langle x | x \rangle}$$

$$|x \rangle = |x - \bar{x} \rangle, \quad |y \rangle = |y - \bar{y} \rangle$$

$$|\langle x - \bar{x} | y - \bar{y} \rangle| \leq \sqrt{\langle y - \bar{y} | y - \bar{y} \rangle} \sqrt{\langle x - \bar{x} | x - \bar{x} \rangle}$$

$$\Rightarrow \left| \sum_i \langle x_i - \bar{x} | y_i - \bar{y} \rangle \right| \leq \sqrt{\sum_i (y_i - \bar{y})^2} \sqrt{\sum_i (x_i - \bar{x})^2}$$

$$\Rightarrow \frac{\left| \sum_i \langle x_i - \bar{x} | y_i - \bar{y} \rangle \right|}{\sqrt{\sum_i (y_i - \bar{y})^2} \sqrt{\sum_i (x_i - \bar{x})^2}} \leq 1 \quad \Rightarrow \mathbf{r_{xy}} \leq \mathbf{1} \quad \Rightarrow \mathbf{-1 \leq r_{xy} \leq 1}$$

$$d(|x\rangle; |y\rangle) = \sqrt{\langle x-y | x-y \rangle} = \sqrt{\sum_i (x_i - y_i)^2}$$

$$t \in \mathfrak{R} \Rightarrow 0 \leq \langle x - ty | x - ty \rangle = \langle x | x \rangle - 2t \langle x | y \rangle + t^2 \langle y | y \rangle = \|x\|^2 - 2t \langle x | y \rangle + t^2 \|y\|^2$$

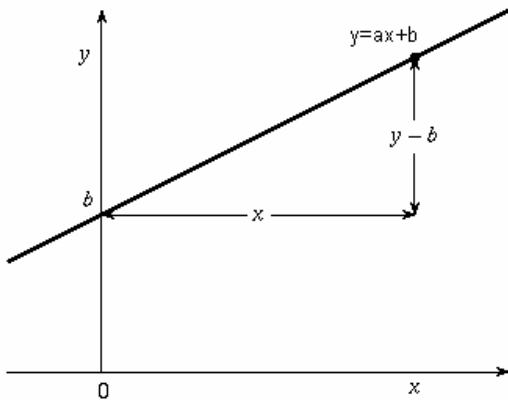
$$\text{pentru } \Delta \leq 0 \Rightarrow \Delta = 4 \langle x | y \rangle^2 - 4 \|y\|^2 \|x\|^2$$

$$\Rightarrow \langle x | y \rangle^2 \leq \|y\|^2 \|x\|^2 \quad \Rightarrow (\langle x | y \rangle)^2 \leq \langle y | y \rangle \langle x | x \rangle$$

$$\left| \sum_i (x_i - \bar{x})(y_i - \bar{y}) \right| \leq \sqrt{\sum_i (y_i - \bar{y})^2} \sqrt{\sum_i (x_i - \bar{x})^2}$$

2. Pearson Correlation: traditional QSP(A)R

x	x_1	x_2	...	x_n
y	y_1	y_2	...	y_n



$$y_i^{obs} = ax_i + b + e_i$$

$$y_{\text{calculat}} = ax_i + b$$

$$e_i^2 = [y_i - (ax_i + b)]^2 \Rightarrow \sum_i e_i^2 \rightarrow \text{minim}$$

$$f(a,b) = \sum_i e_i^2 = \sum_i (y_i - ax_i - b)^2 \rightarrow \text{MIN},$$

$$\begin{cases} \frac{\partial f(a,b)}{\partial a} = 0 \\ \frac{\partial f(a,b)}{\partial b} = 0 \end{cases} \Leftrightarrow \begin{cases} 2 \sum_i (y_i - ax_i - b)(-x_i) = 0 \\ 2 \sum_i (y_i - ax_i - b)(-1) = 0 \end{cases}$$

$$\begin{cases} \sum_i y_i x_i = a \sum_i x_i^2 + b \sum_i x_i \cdot n \\ \sum_i y_i = a \sum_i x_i + bn \cdot \left(-\sum_i x_i\right) \end{cases}$$

$$a = \frac{n \sum_i y_i x_i - (\sum_i y_i)(\sum_i x_i)}{n \sum_i x_i^2 - (\sum_i x_i)^2}, \quad b = \frac{\sum_i y_i \sum_i x_i^2 - \sum_i x_i \sum_i x_i y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

- Dacă înmulțim ecuația 2 din sistem cu $1/n$ obținem:

$$\frac{1}{n} \sum_i y_i = a \frac{1}{n} \sum_i x_i + b \Rightarrow \bar{y} = a\bar{x} + b$$

Correlation through Great Numbers' Law

$$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} = \frac{\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle}{\sqrt{\langle (x - \langle x \rangle)^2 \rangle} \sqrt{\langle (y - \langle y \rangle)^2 \rangle}}$$
$$= \frac{\sum_{ij} x_i y_j p_{ij} - (\sum_{ij} x_i p_{ij})(\sum_{ij} y_j p_{ij})}{\sqrt{\sum_{ij} x_i^2 p_{ij} - (\sum_{ij} x_i p_{ij})^2} \sqrt{\sum_{ij} y_j^2 p_{ij} - (\sum_{ij} y_j p_{ij})^2}}$$
$$i = j \Rightarrow p_i = p_j = \frac{1}{n}$$

$$r_{xy} = \frac{\frac{1}{n} \sum_i x_i y_i - \frac{1}{n^2} (\sum_i x_i) (\sum_i y_i)}{\sqrt{\frac{1}{n} \sum_i x_i^2 - \frac{1}{n^2} (\sum_i x_i)^2} \sqrt{\frac{1}{n} \sum_i y_i^2 - \frac{1}{n^2} (\sum_i y_i)^2}}$$

$$r_{xy} \frac{\sigma_y}{\sigma_x} = \frac{C_{xy}}{\sigma_x \sigma_y} \frac{\sigma_y}{\sigma_x} = a = \frac{C_{xy}}{\sigma_x^2} \Rightarrow$$

$$y = r_{xy} \frac{\sigma_y}{\sigma_x} x + b \quad (y - \bar{y}) = r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x}) + b$$

$$\Rightarrow \bar{y} = r_{xy} \frac{\sigma_y}{\sigma_x} \bar{x} + b \quad \Rightarrow y = \overset{\text{calculat}}{\bar{y}} + r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$\underbrace{|r_{xy}|}_{\in(0,1)} \sqrt{n-1} \geq 3$$

Correlation by Distribution Functions

- momentul de ordine K al parametrului x :

$$\mu_K = \langle (x - \langle x \rangle)^K \rangle = \int (x - \langle x \rangle)^K f(x) dx$$

- condiția de normalizare a funcției de distribuție:

$$\mu_0 = \langle (x - \langle x \rangle)^0 \rangle = \langle 1 \rangle = \int f(x) dx \Rightarrow \mu_0 = \int f(x) dx$$

$$\mu_1 = \langle (x - \langle x \rangle)^1 \rangle = \langle x \rangle - \langle x \rangle = 0$$

$$\mu_1 = \int (x - \langle x \rangle) f(x) dx = \int x f(x) dx - \int \langle x \rangle f(x) dx$$

$$\mu_2 = \langle (x - \langle x \rangle)^2 \rangle = \sigma_x^2 = \int (x - \bar{x})^2 f(x) dx$$

- Rezumand avem:
$$\int f = 1$$
$$\int (x - \langle x \rangle) f = 0$$
$$\int (x - \langle x \rangle)^2 f = \sigma^2$$




- **f = ? Trebuie gasita functia f(x) !!!**

- **Gaussian Nature:**

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \cdot e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \quad x \in R$$

- funcția de distribuție normală a probabilităților de existență a variabilei x pe domeniul $(+\infty, -\infty)$.

$$\mu_0 = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \int_{-\infty}^{+\infty} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} d(x - \bar{x}) = 1$$


$$\mu_1 = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \int_{-\infty}^{+\infty} (x - \bar{x}) e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} d(x - \bar{x}) = 0$$

$$\mu_2 = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \int_{-\infty}^{+\infty} (x - \bar{x})^2 e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} d(x - \bar{x}) = \sigma_x^2$$

$$I_0 = \int_{-\infty}^{+\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$$

$$I_1 = \int_{-\infty}^{+\infty} x e^{-ax^2} dx = 0$$

$$I_2 = \int_{-\infty}^{+\infty} x^2 e^{-ax^2} dx = \frac{1}{2a} \sqrt{\frac{\pi}{a}}$$

- For that, let firstly rewrite

$$I_0^2 = I_0 \cdot I_0 = \left(\int_{-\infty}^{+\infty} e^{-ax^2} dx \right) \left(\int_{-\infty}^{+\infty} e^{-ay^2} dy \right) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-a(x^2+y^2)} dx dy$$

- That in polar coordinates becomes

$$I_0^2 = \int_0^{2\pi} \int_0^{\infty} e^{-ar^2} r dr d\varphi = \left(\int_0^{\pi} r e^{-ar^2} dr \right) \left(\int_0^{2\pi} d\varphi \right) = 2\pi \int_0^{\infty} \left(-\frac{1}{2a} \right) d(e^{-ar^2}) = -\frac{\pi}{a} e^{-ar^2} \Big|_0^{\infty} = \frac{\pi}{a}$$

- Leaving with

$$I_0 = \int_{-\infty}^{+\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}} \Rightarrow \mu_0 = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-a\frac{(x-\bar{x})^2}{2\sigma^2}} d(x-\bar{x}) = \frac{1}{\sqrt{2\pi\sigma}} \sqrt{\frac{\pi}{\frac{1}{2\sigma^2}}} = 1$$

$$I_1 = \int_{-\infty}^{+\infty} x e^{-ax^2} dx = -\frac{1}{2a} \int_{-\infty}^{+\infty} d(e^{-ax^2}) = -\frac{1}{2a} e^{-ax^2} \Big|_{-\infty}^{+\infty} = 0 \Rightarrow \mu_1 = 0$$

$$I_2 = \int_{-\infty}^{+\infty} u^2 e^{-au^2} du = \int_{-\infty}^{+\infty} u u e^{-au^2} du = -\frac{1}{2a} \int_{-\infty}^{+\infty} u d(u e^{-au^2}) du$$

$$I_2 = -\frac{1}{2a} \left[\int_{-\infty}^{+\infty} \frac{d}{du} (u \cdot e^{-au^2}) du - \int_{-\infty}^{+\infty} e^{-au^2} \frac{du}{du} du \right] = -\frac{1}{2a} \sqrt{\frac{\pi}{a}}$$

$$\Rightarrow \mu_2 = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \int_{-\infty}^{+\infty} (x - \bar{x})^2 e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} d(x - \bar{x}) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \cdot \frac{1}{2 \cdot \frac{1}{2\sigma^2}} \cdot \sqrt{2\pi\sigma_x^2} = \sigma_x^2$$

Funcția de distribuție bidimensională

$$f(x, y) = \frac{1}{2\pi\sigma_x \sigma_y \cdot \sqrt{1-r_{xy}^2}} \exp\left\{\frac{-1}{2(1-r_{xy}^2)} \left[\frac{(x-\bar{x})^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_y^2} - 2r_{xy} \frac{(x-\bar{x})(y-\bar{y})}{\sigma_x \sigma_y} \right]\right\}$$

- Observație : Când cele 2 variabile sunt total independente

$$r_{xy} = 0 \Rightarrow f(x, y) = \left[\frac{1}{\sqrt{2\pi} \cdot \sigma_x} \cdot e^{-\frac{(x-\bar{x})^2}{2\sigma_x^2}} \right] \left[\frac{1}{\sqrt{2\pi} \cdot \sigma_y} \cdot e^{-\frac{(y-\bar{y})^2}{2\sigma_y^2}} \right] = f(x)f(y)$$

Funcția de distribuție (probabilitate) condiționată : $g(y/x)$

$$g(y/x) = \frac{f(x, y)}{f(x)}$$

$$\langle y \rangle = y_{\text{calculat}} = \int yg(y/x)dy$$

$$g(y/x) = \frac{f(x, y)}{f(x)} =$$

$$= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r_{xy}^2}} \frac{\sqrt{2\pi}\sigma_x}{1} \exp\left\{ \frac{-1}{2(1-r_{xy}^2)} \left[\frac{(x-\bar{x})^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_y^2} - 2r_{xy} \frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} \right] + \frac{(x-\bar{x})^2}{2\sigma_x^2} \right\}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-r_{xy}^2}} \exp\left\{\frac{-1}{2(1-r_{xy}^2)}\left[\frac{(x-\bar{x})^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_y^2} - 2r_{xy}\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} - \frac{2(1-r_{xy}^2)(x-\bar{x})^2}{2\sigma_x^2}\right]\right\}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-r_{xy}^2}} \exp\left\{\frac{-1}{2(1-r_{xy}^2)}\left[r_{xy}^2\frac{(x-\bar{x})^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_y^2} - 2r_{xy}\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}\right]\right\} =$$

$$= \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-r_{xy}^2}} \exp\left\{\frac{-1}{2(1-r_{xy}^2)}\left[\frac{(y-\bar{y})}{\sigma_y} - r_{xy}\frac{(x-\bar{x})}{\sigma_x}\right]^2\right\} =$$

$$= \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-r_{xy}^2}} \exp\left\{\frac{-1}{2(1-r_{xy}^2)\sigma_y^2}\left[y - \bar{y} - r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x})\right]^2\right\}$$

$$g(y/x) = \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-r_{xy}^2}} \exp\left\{\frac{-1}{2(1-r_{xy}^2)\sigma_y^2}\left[y - \bar{y} - r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x})\right]^2\right\}$$

■ Pentru valoare lui y avem relatia

$$\langle y \rangle = \int_{-\infty}^{+\infty} yg(y/x)dy = \int_{-\infty}^{+\infty} [y - \bar{y} - r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x})] g(y/x) d(y - \bar{y} - r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x})) +$$

$$+ \int_{-\infty}^{+\infty} [+ \bar{y} - r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x})] g(y/x) d(y - \bar{y} - r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x})) =$$

$$= 0 + [y - \bar{y} - r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x})] \underbrace{\int_{-\infty}^{+\infty} g(y/x) d(y - \bar{y} - r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x}))}_{=1} \Rightarrow$$

■ **ecuația de corelare a lui y cu x**

$$y^{calc} = \bar{y} + r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$



- $SR_y =$ suma erorilor

$$SR_y = \text{Min} \left[\sum_i (y_i^{obs} - y_i^{calc})^2 \right]$$

$$= \text{Min} \langle (y_i^{obs} - y_i^{calc})^2 \rangle = \langle y_i^{obs} - \bar{y} - r_{xy} \cdot \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \rangle^2 =$$

$$= \int [y_i^{obs} - \bar{y} - r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x})]^2 g(y/x) d(y - \bar{y} - r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x}))$$

$$\Rightarrow SR_y = \text{Min} \left(\sum_i (y_i^{obs} - y_i^{calc})^2 \right) = (1 - r_{xy}^2) \sigma_y^2$$

$$\frac{SR}{\sigma_y^2} = 1 - r_{xy}^2 \quad \Rightarrow \quad r_{xy} = \sqrt{1 - \frac{SR}{\sigma_y^2}}$$

- **factorul de corelare standard**

$$r_{xy} = \sqrt{1 - \frac{\sum_i (y_i^{obs} - y_i^{calc})^2}{\sum_i (y_i^{obs} - \bar{y})^2}}$$

- **Covarianța**

$$C_{xy} = \langle \langle xy \rangle - \langle x \rangle \langle y \rangle \rangle$$

$$C_{xy} = \sum_{ij} x_i y_j P_{ij} - \frac{1}{n^2} \left(\sum_i x_i \right) \left(\sum_j y_j \right)$$

- **Coeficientul de corelare:**

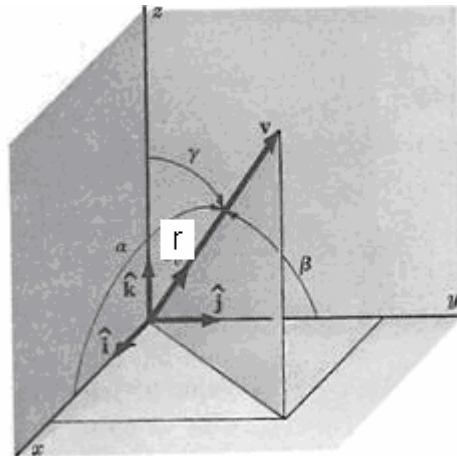
$$\frac{C_{xy}}{\sigma_x \sigma_y} = \frac{\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle}{\sqrt{\langle (x - \langle x \rangle)^2 \rangle \langle (y - \langle y \rangle)^2 \rangle}} = r_{x,y}$$

$$r_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \leq 1$$

The vectorial interpretation:

$$\vec{x} \cdot \vec{y} = |\vec{x}| \cdot |\vec{y}| \cos(\vec{x}, \vec{y})$$

$$\Rightarrow \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| |\vec{y}|} \Rightarrow \cos(\langle x | y \rangle) = \frac{\langle x | y \rangle}{\| |x\rangle \| \| |y\rangle \|} = \frac{\langle x | y \rangle}{\sqrt{\langle x | x \rangle} \sqrt{\langle y | y \rangle}} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \sqrt{\sum_i y_i^2}}$$



Multilinear Correlation

	x_0	$x_k \dots$	$x_k \dots$	x_M
y_1	1	$x_{11} \dots$	$x_{1k} \dots$	x_{1M}
y_2	1	$x_{21} \dots$	$x_{2k} \dots$	x_{2M}
...
y_k	1	$x_{k1} \dots$	$x_{kk} \dots$	x_{kM}
...
y_N	1	$x_{N1} \dots$	$x_{Nk} \dots$	x_{NM}

$$y = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_M X_M$$

$$\begin{cases} y_1^{obs} = b_0 + b_1 X_{11} + b_2 X_{12} + \dots + b_M X_{1M} + e_1 \\ y_2^{obs} = b_0 + b_1 X_{21} + b_2 X_{22} + \dots + b_M X_{2M} + e_2 \\ \dots \\ y_N^{obs} = b_0 + b_1 X_{N1} + b_2 X_{N2} + \dots + b_M X_{NM} + e_N \end{cases}$$

■ Suma erorilor sa fie minima:


$$\left\{ \begin{array}{l} \frac{\partial}{\partial b_0} \left[\sum_{i=1}^N e_i^2 \right] = 0 \\ \frac{\partial}{\partial b_1} \left[\sum_{i=1}^N e_i^2 \right] = 0 \\ \dots \\ \frac{\partial}{\partial b_M} \left[\sum_{i=1}^N e_i^2 \right] = 0 \end{array} \right. \longrightarrow \left\{ \begin{array}{l} -2 \sum_{i=1}^N [y_i - (b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_M X_{iM})] \cdot 1 = 0 \\ -2 \sum_{i=1}^N [y_i - (b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_M X_{iM})] \cdot x_{i1} = 0 \\ -2 \sum_{i=1}^N [y_i - (b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_M X_{iM})] \cdot x_{i2} = 0 \\ \dots \\ -2 \sum_{i=1}^N [y_i - (b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_M X_{iM})] \cdot x_{iM} = 0 \end{array} \right.$$

■ Suma erorilor sa fie minima:

$$\left\{ \begin{array}{l} \sum_{i=1}^N y_i = b_0 N + b_1 \sum_{i=1}^N x_{i1} + b_2 \sum_{i=1}^N x_{i2} + \dots + b_M \sum_{i=1}^N x_{iM} \\ \sum_{i=1}^N y_i x_{i1} = b_0 \sum_{i=1}^N x_{i1} + b_1 \sum_{i=1}^N x_{i1}^2 + \dots + b_M \sum_{i=1}^N x_{iM} x_{i1} \\ \sum_{i=1}^N y_i x_{i2} = b_0 \sum_{i=1}^N x_{i2} + \dots + b_M \sum_{i=1}^N x_{iM} x_{i2} \\ \dots\dots\dots \\ \sum_{i=1}^N y_i x_{iM} = b_0 \sum_{i=1}^N x_{iM} + b_1 \sum_{i=1}^N x_{i1} x_{iM} + \dots + b_M \sum_{i=1}^N x_{iM}^2 \end{array} \right.$$

$$[\hat{x}] = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1M} \\ 1 & x_{21} & x_{22} & \dots & x_{2M} \\ \dots & & & & \\ 1 & x_{N1} & x_{N2} & \dots & x_{NM} \end{pmatrix} \quad [\hat{y}] = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}, \quad [\hat{b}] = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_M \end{pmatrix} \quad [\hat{e}] = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_N \end{pmatrix}$$

$$[\hat{y}] = [\hat{x}][\hat{b}] + [\hat{e}]$$


$$[\hat{x}]^T [\hat{y}] = [\hat{x}]^T [\hat{x}] [\hat{b}]$$

$$\frac{\partial}{\partial [\hat{b}]} ([\hat{e}]^T [\hat{e}]) = 0$$

- Matricea **Moare-Penrose**:

$$([x]^T [x])^{-1} [x]^T [y] = [b]$$

- **Exemplu**: sa verificam pentru regresia simpla liniara:
 $y=b+ax$

$$\begin{cases} y_1^{obs} = b + aX_1 + e_1 \\ y_2^{obs} = b + aX_2 + e_2 \\ \dots \\ y_N^{obs} = b + aX_N + e_N \end{cases}$$

$$[\hat{y}] = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} \quad [\hat{b}] = \begin{pmatrix} b = b_0 \\ a = b_1 \end{pmatrix} \quad [\hat{x}] = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \\ 1 & x_N \end{pmatrix}$$

■ We successively have

$$[\hat{x}]^T [\hat{x}] = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \\ 1 & x_N \end{pmatrix} = \begin{pmatrix} N & \sum_i^N x_i \\ \sum_i^N x_i & \sum_i^N x_i^2 \end{pmatrix}$$

$$|\hat{x}^T \hat{x}| = N \sum_i^N x_i^2 - \left(\sum_i^N x_i \right)^2$$

$$\tilde{A}_{11} = (-1)^{1+1} \sum_i^N x_i^2 = \sum_i^N x_i^2$$

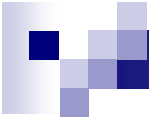
$$\tilde{A}_{12} = (-1)^{1+2} \sum_i^N x_i = -\sum_i^N x_i$$

$$\tilde{A}_{21} = (-1)^{2+1} \sum_i^N x_i = -\sum_i^N x_i$$

$$\tilde{A}_{22} = (-1)^{2+2} N = N$$

$$\hat{A}^* = \begin{pmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{pmatrix} = \begin{pmatrix} \sum_i^N x_i^2 & -\sum_i^N x_i \\ -\sum_i^N x_i & N \end{pmatrix}$$


$$\hat{A}^{-1} = (\hat{x}^T \hat{x})^{-1} = \frac{\hat{A}^*}{|\hat{A}|} = \frac{(\hat{x}^T \hat{x})^*}{|\hat{x}^T \hat{x}|}$$



$$(\hat{x}^T \hat{x})^{-1} = \begin{pmatrix} \frac{\sum_i^N x_i^2}{N \sum_i^N x_i^2 - (\sum_i^N x_i)^2} & \frac{-\sum_i^N x_i}{N \sum_i^N x_i^2 - (\sum_i^N x_i)^2} \\ \frac{-\sum_i^N x_i}{N \sum_i^N x_i^2 - (\sum_i^N x_i)^2} & \frac{N}{N \sum_i^N x_i^2 - (\sum_i^N x_i)^2} \end{pmatrix}$$

$$\hat{b} = (\hat{x}^T \hat{x})^{-1} \hat{x}^T \hat{y}$$

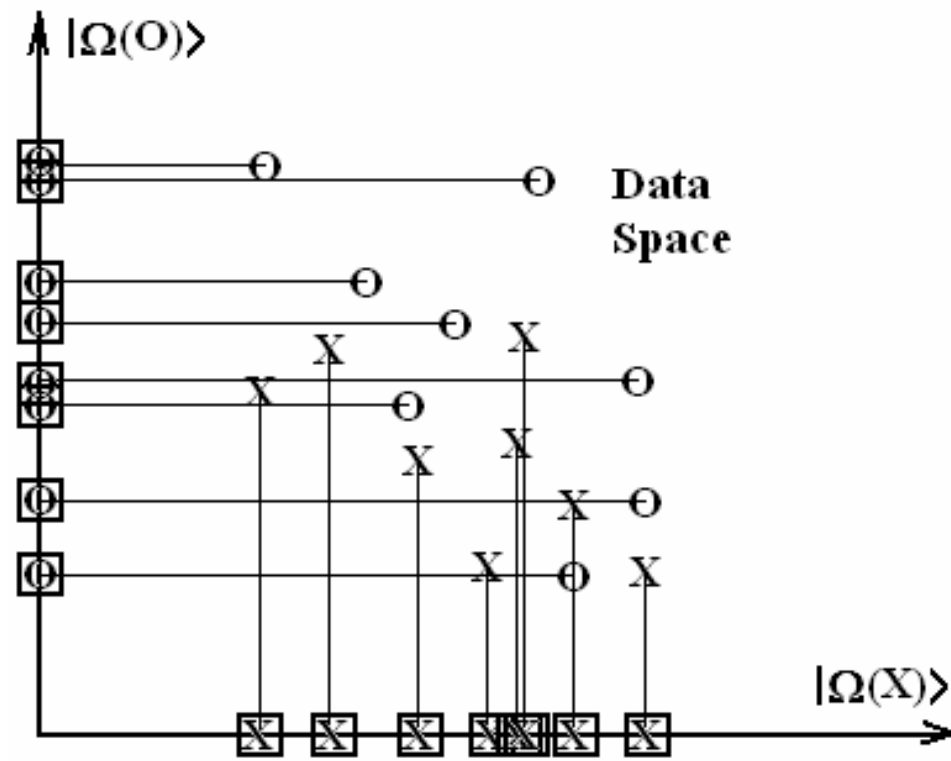
$$\hat{x}^T \hat{y} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} \sum_i^N y_i \\ \sum_i^N x_i y_i \end{pmatrix}$$



$$\hat{b} = \begin{pmatrix} \frac{\sum_i^N x_i^2 \sum_i^N y_i - \sum_i^N x_i \sum_i^N x_i y_i}{N \sum_i^N x_i^2 - (\sum_i^N x_i)^2} \\ -\frac{\sum_i^N x_i \sum_i^N y_i + N \sum_i^N x_i y_i}{N \sum_i^N x_i^2 - (\sum_i^N x_i)^2} \end{pmatrix} = \begin{pmatrix} b_0 = b \\ b_1 = a \end{pmatrix}$$

- Da! Se verifica pentru regresia liniara simpla.

Spectral Correlation: S-SP(A)R



■ The spectral (vectorial) version of SAR descriptors

<i>Activity</i>	<i>Structural predictor variables</i>					
$ Y\rangle$	$ X_0\rangle$	$ X_1\rangle$...	$ X_k\rangle$...	$ X_M\rangle$
y_1	1	x_{11}	...	x_{1k}	...	x_{1M}
y_2	1	x_{21}	...	x_{2k}	...	x_{2M}
....
y_N	1	x_{N1}	...	x_{Nk}	...	x_{NM}

$$|Y\rangle = b_0|X_0\rangle + b_1|X_1\rangle + \dots + b_k|X_k\rangle + \dots + b_M|X_M\rangle + |e\rangle$$



$$|y_i^{obs}\rangle = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

$$|x_0\rangle = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$|x_1\rangle = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{N1} \end{pmatrix}$$

$$|x_M\rangle = \begin{pmatrix} x_{1M} \\ x_{2M} \\ \vdots \\ x_{NM} \end{pmatrix}$$

- However, before applying it effectively one has to introduce the generalized scalar product throughout the basic rule:

$$\langle \Psi_l | \Psi_k \rangle = \sum_{i=1}^N \psi_{il} \psi_{ik} = \langle \Psi_k | \Psi_l \rangle$$

$$|\Psi_l\rangle = |\psi_{1l} \quad \psi_{2l} \quad \dots \quad \psi_{Nl}\rangle \quad |\Psi_k\rangle = |\psi_{1k} \quad \psi_{2k} \quad \dots \quad \psi_{Nk}\rangle$$

- Briefly, remember that the orthogonal condition requires that the scalar product of type to be zero, the orthogonal basis

$$\{|\Omega_0\rangle, |\Omega_1\rangle, \dots, |\Omega_k\rangle, \dots, |\Omega_M\rangle\}$$

can be constructed from the set $\{|X_0\rangle, |X_1\rangle, \dots, |X_k\rangle, \dots, |X_M\rangle\}$

according with the iterative recipe:

- 
- Choose $|\Omega_0\rangle = |X_0\rangle$

- Then, by picking $|X_1\rangle$ as the next vector to be transformed, one can write that

$$|\Omega_1\rangle = |X_1\rangle - r_0^1 |\Omega_0\rangle \quad r_0^1 = \frac{\langle X_1 | \Omega_0 \rangle}{\langle \Omega_0 | \Omega_0 \rangle}$$

- Next

$$|\Omega_k\rangle = |X_k\rangle - \sum_{i=0}^{k-1} r_i^k |\Omega_i\rangle \quad r_i^k = \frac{\langle X_k | \Omega_i \rangle}{\langle \Omega_i | \Omega_i \rangle}$$

- repeated and extended until the last orthogonal predictor vector $|\Omega_M\rangle$ is obtained.

- Within the constructed orthogonal space, the vector activity

$$|Y\rangle = \omega_0 |\Omega_0\rangle + \omega_1 |\Omega_1\rangle + \dots + \omega_k |\Omega_k\rangle + \dots + \omega_M |\Omega_M\rangle$$

- At this point, since there is no residual vector remaining one can consider that the SAR problem is in principle solved once the new coefficients in $\omega_0, \omega_1, \dots, \omega_k, \dots, \omega_M$ are determined. These new coefficients can be immediately deduced based on the orthogonal peculiarities of the spectral decomposition grounded on the fact that:

$$\langle \Omega_k | \Omega_l \rangle = 0, \quad k \neq l$$

$$\omega_k = \frac{\langle \Omega_k | Y \rangle}{\langle \Omega_k | \Omega_k \rangle} \quad k = \overline{0, M}$$

SAR algorithm

- It consists in going back from the orthogonal to the initial basis of data through the system of coordinate transformations:

$$\left\{ \begin{array}{l} |Y\rangle = \omega_0 |\Omega_0\rangle + \omega_1 |\Omega_1\rangle + \dots + \omega_k |\Omega_k\rangle + \dots + \omega_M |\Omega_M\rangle \\ |X_0\rangle = 1 \cdot |\Omega_0\rangle + 0 \cdot |\Omega_1\rangle + \dots + 0 \cdot |\Omega_k\rangle + \dots + 0 \cdot |\Omega_M\rangle \\ |X_1\rangle = r_0^1 |\Omega_0\rangle + 1 \cdot |\Omega_1\rangle + \dots + 0 \cdot |\Omega_k\rangle + \dots + 0 \cdot |\Omega_M\rangle \\ \dots \\ |X_k\rangle = r_0^k |\Omega_0\rangle + r_1^k |\Omega_1\rangle + \dots + 1 \cdot |\Omega_k\rangle + \dots + 0 \cdot |\Omega_M\rangle \\ \dots \\ |X_M\rangle = r_0^M |\Omega_0\rangle + r_1^M |\Omega_1\rangle + \dots + r_k^M |\Omega_k\rangle + \dots + 1 \cdot |\Omega_M\rangle \end{array} \right.$$

- Finally, the system is algebraically true if and only if the associated augmented determinant disappears,

$$\begin{vmatrix}
 |Y\rangle & \omega_0 & \omega_1 & \cdots & \omega_k & \cdots & \omega_M \\
 |X_0\rangle & 1 & 0 & \cdots & 0 & \cdots & 0 \\
 |X_1\rangle & r_0^1 & 1 & \cdots & 0 & \cdots & 0 \\
 \vdots & \vdots & \vdots & \vdots & & \vdots & \\
 |X_k\rangle & r_0^k & r_1^k & \cdots & 1 & \cdots & 0 \\
 \vdots & \vdots & \vdots & \vdots & & \vdots & \\
 |X_M\rangle & r_0^M & r_1^M & \cdots & r_k^M & \cdots & 1
 \end{vmatrix} = 0$$

Check with Linear Correlation

- Verificăm pentru cea monoliniară

$$|y\rangle = b_0|x_0\rangle + b_1|x_1\rangle \text{ este de forma } y = b + ax$$

$$0 = \begin{vmatrix} |y^P\rangle & \omega_0 & \omega_1 \\ |x_0\rangle & \mathbf{1} & \mathbf{0} \\ |x_1\rangle & r_0^1 & \mathbf{1} \end{vmatrix} = |y^P\rangle \begin{vmatrix} \mathbf{1} & \mathbf{0} \\ r_0^1 & \mathbf{1} \end{vmatrix} - |x_0\rangle \begin{vmatrix} \omega_0 & \omega_1 \\ r_0^1 & \mathbf{1} \end{vmatrix} + |x_1\rangle \begin{vmatrix} \omega_0 & \omega_1 \\ \mathbf{1} & \mathbf{0} \end{vmatrix}$$

$$|y^P\rangle = \underbrace{(\omega_0 - r_0^1 \omega_1)}_b |x_0\rangle + \underbrace{\omega_1}_a |x_1\rangle$$

$$a = \omega_1, b = \omega_0 - r_0^1 \omega_1$$

■ Se realizează tabelul:

$ y\rangle$	$ x_0\rangle$	$ x_1\rangle$
y_1	1	x_{11}
y_2	1	x_{21}
...
y_N	1	x_{N1}

$$|\Omega_0\rangle = |1 \ 1 \ 1 \ \dots 1\rangle, |\Omega_1\rangle = |x_1\rangle - r_0^1 |\Omega_0\rangle$$


$$r_0^1 = \frac{\langle x_1 | \Omega_0 \rangle}{\langle \Omega_0 | \Omega_0 \rangle} = \frac{\sum_{i=1}^N x_i}{N}$$

$$|\Omega_1\rangle = |x_1, x_2, \dots, x_N\rangle - \frac{\sum_{i=1}^N x_i}{N} |1 \ 1 \ 1 \ \dots 1\rangle = \left| x_1 - \frac{\sum_{i=1}^N x_i}{N}, \dots, x_N - \frac{\sum_{i=1}^N x_i}{N} \right\rangle$$

$$\begin{aligned}
\omega_1 &= \frac{\langle \Omega_1 | y \rangle}{\langle \Omega_1 | \Omega_1 \rangle} = \frac{\left\langle x_1 - \frac{1}{N} \sum x_i, \dots, x_N - \frac{1}{N} \sum x_i \middle| y_1, \dots, y_N \right\rangle}{\sum_i \left(x_i - \frac{1}{N} \sum x_i\right)^2} \\
&= \frac{\sum_i y_i \left(x_i - \frac{1}{N} \sum x_i\right)}{\sum_i \left(x_i - \frac{1}{N} \sum x_i\right)^2} = \frac{\sum_i y_i x_i - \frac{1}{N} \left(\sum_i y_i\right) \left(\sum_i x_i\right)}{\sum_i \left(x_i^2 + \frac{1}{N^2} \sum x_i^2 - \frac{2}{N} x_i \sum_i x_i\right)}
\end{aligned}$$

$$= \frac{N \sum_i y_i x_i - \left(\sum_i y_i\right) \left(\sum_i x_i\right)}{N \sum_i x_i^2 - \left(\sum_i x_i\right)^2} = a$$

$$\omega_0 = \frac{\langle \Omega_0 | y \rangle}{\langle \Omega_0 | \Omega_0 \rangle} = \frac{\sum_i y_i}{N} = \frac{1}{N} \sum_i y_i$$



$$b = \omega_0 - r_0^1 \omega_1 = \frac{1}{N} \sum_i y_i - \frac{1}{N} \sum_i x_i \frac{N \sum_i y_i x_i - (\sum_i y_i)(\sum_i x_i)}{N \sum_i x_i^2 - (\sum_i x_i)^2}$$

$$= \frac{(\sum_i y_i)(\sum_i x_i)^2 - (\sum_i x_i)(\sum_i y_i x_i)}{N \sum_i x_i^2 - (\sum_i x_i)^2}$$

- DA!!! Se verifică ecuația monolinară!

3. Algebraic Correlation Factor

- Next, the vectorial norm gives the opportunity in advancing the so called algebraic correlation factor measuring the relative “intensity” or “amplitude” of the predicted to measured norm activities

$$r_{S-SAR}^{ALGEBRAIC} = \frac{\|Y^P\|}{\|Y^{Obs}\|}$$

- worth noting that the algebraic correlation factor gives systematic higher values respecting the assumed statistical one, computed on statistical definition of the data dispersion through the standard expressions

$$r_{QSAR}^{STATISTIC} = \sqrt{1 - \frac{SR}{SQ}}$$

$$SR = \sum_{i=1}^N [y_i^{Obs} - y_i^P]^2$$

$$SQ = \sum_{i=1}^N \left[y_i^{Obs} - N^{-1} \sum_{i=1}^N y_i^{Obs} \right]^2$$

- **Proposition (algebraic S-SAR vs. statistic Q-SAR correlations):** in any correlation analysis, considering the observed and predicted activity data as vectors $|Y^P\rangle$ and $|Y^{Obs}\rangle$ with the associate norms through the scalar product type, respectively, the algebraic correlation factor always exceeds the statistical QSAR correlation factor:

$$r_{S-SAR}^{ALGEBRAIC} \geq r_{QSAR}^{STATISTIC}$$

- **Proof:** by straight algebraic manipulation, condition firstly rewrites as:

$$\frac{\sum_{i=1}^N (y_i^2)^P}{\sum_{i=1}^N (y_i^2)^{Obs}} \geq \frac{\sum_{i=1}^N \left[y_i^P - N^{-1} \sum_{i=1}^N y_i^{Obs} \right] \left[2y_i^{Obs} - y_i^P - N^{-1} \sum_{i=1}^N y_i^{Obs} \right]}{\sum_{i=1}^N \left[y_i^{Obs} - N^{-1} \sum_{i=1}^N y_i^{Obs} \right]^2}$$

- Then, it can be conveniently arranged so that to separate the *predicted* and *measured* terms in left and right side of the inequality, respectively:

$$\frac{\sum_{i=1}^N (y_i^2)^P}{\sum_{i=1}^N \left[y_i^P - N^{-1} \sum_{i=1}^N y_i^{Obs} \right] \left[2y_i^{Obs} - y_i^P - N^{-1} \sum_{i=1}^N y_i^{Obs} \right]} \geq \frac{\sum_{i=1}^N (y_i^2)^{Obs}}{\sum_{i=1}^N \left[y_i^{Obs} - N^{-1} \sum_{i=1}^N y_i^{Obs} \right]^2}$$

- Now:
$$\frac{\sum_{i=1}^N (y_i^2)^{Obs} - \left[\sum_{i=1}^N (y_i^2)^{Obs} - \sum_{i=1}^N (y_i^2)^P \right]}{\sum_{i=1}^N \left[y_i^{Obs} - N^{-1} \sum_{i=1}^N y_i^{Obs} \right]^2 - \sum_{i=1}^N \left[y_i^{Obs} - y_i^P \right]^2} \geq \frac{\sum_{i=1}^N (y_i^2)^{Obs}}{\sum_{i=1}^N \left[y_i^{Obs} - N^{-1} \sum_{i=1}^N y_i^{Obs} \right]^2}$$

$$RS = \sum_{i=1}^N (y_i^2)^{Obs} - \sum_{i=1}^N (y_i^2)^P = \left\| \left\| Y^{Obs} \right\| \right\|^2 - \left\| \left\| Y^P \right\| \right\|^2$$

$$SR = \left\| \left\| Y^{Obs} \right\| - \left\| Y^P \right\| \right\|^2$$


$$SR \geq RS$$

$$SR - RS = \langle Y^{Obs} - Y^P | Y^{Obs} - Y^P \rangle - \langle Y^{Obs} | Y^{Obs} \rangle + \langle Y^P | Y^P \rangle$$

$$= 2(\langle Y^P | Y^P \rangle - \langle Y^{Obs} | Y^P \rangle)$$

$$= 2[\langle Y^P | Y^P \rangle - (\langle Y^P | + \langle e |) Y^P \rangle]$$

$$= -2\langle e | Y^P \rangle$$

$$\geq -2\|e\rangle\| \|Y^P\rangle\|$$

$$= 0$$

Spectral Paths

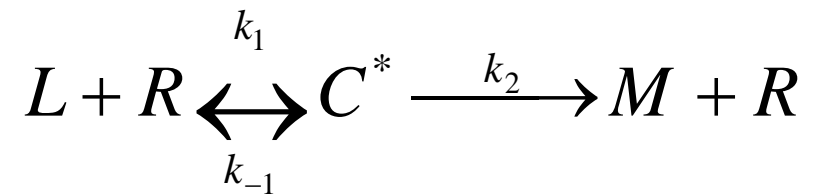
- the Spectral-SAR analysis employs the idea of “amplitude” or “intensity” or “length” of chemical-biological interaction and bonding. In this context, appears also the idea of introducing *the least path principle* that selects the optimal (shortest) paths among all computationally tested models towards the measured endpoint:

$$\delta[A, B] = 0; \quad A, B : \text{ENDPOINTS}$$

$$[A, B] = \sqrt{\left(\left\| \left\| Y^B \right\| \right\| - \left\| \left\| Y^A \right\| \right\| \right)^2 + \left(r_B^{\text{STATISTIC/ALGEBRAIC}} - r_A^{\text{STATISTIC/ALGEBRAIC}} \right)^2}$$

- $[A, B] = [A, C] + [C, B], \quad \delta[A, B] = 0$

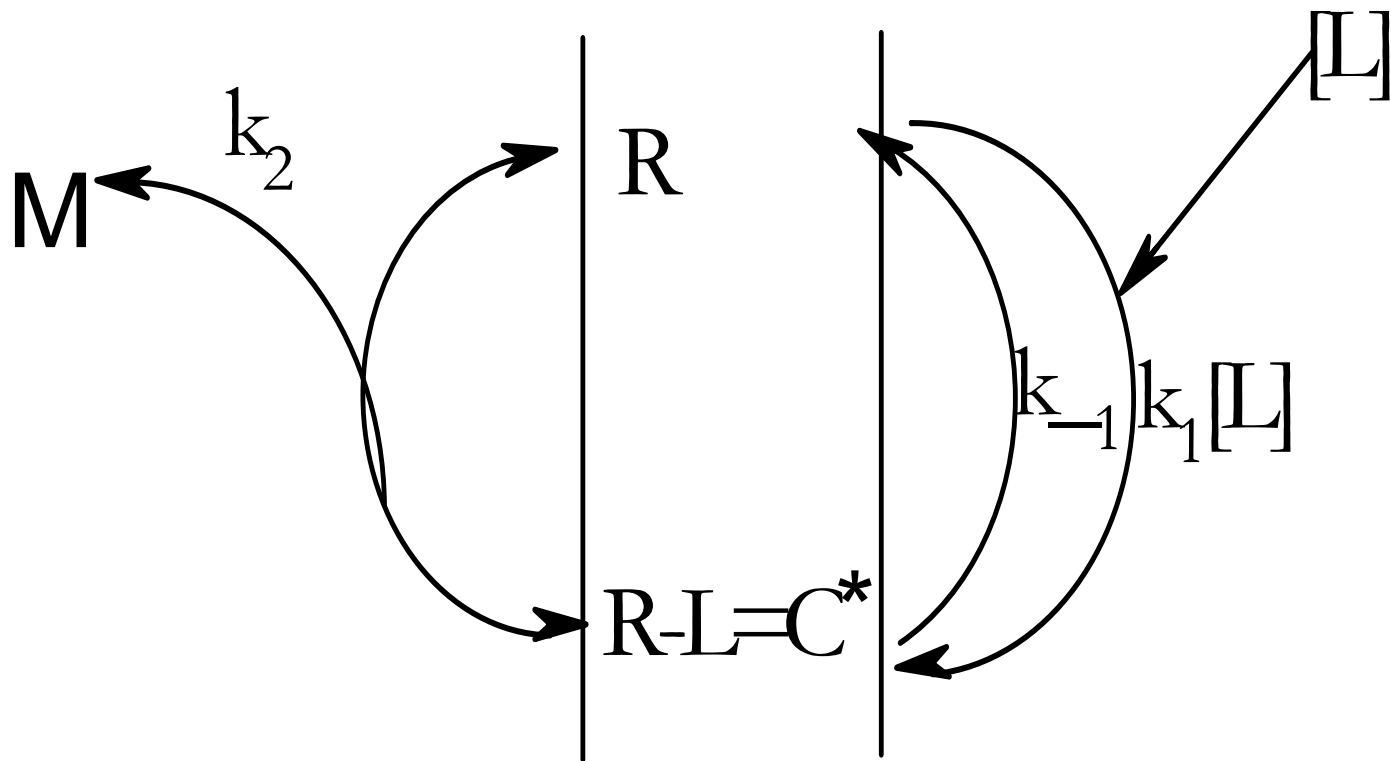
Temporal Spectral-SAR



$$K = \frac{[R][L]}{[C^*]} = \exp\left(-\frac{\Delta G}{RT}\right)$$

$$\frac{A}{A_{\max}} = \frac{[C^*]}{[R] + [C^*]}$$

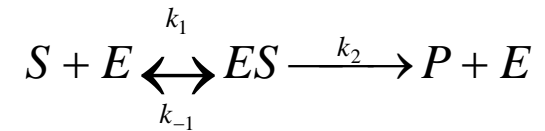
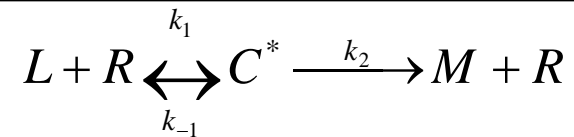
A=activitatea



$$\frac{A}{A_{\max}} = \frac{[L]}{K + [L]}$$

Properties**L-R Kinetics****S-E Kinetics**

Reaction



Species

R

E

Substrate

L

S

Computation


$$\frac{A}{A_{\max}} = \frac{[L]}{[L] + [K]}$$

$$\frac{v}{V_{\max}} = \frac{[S]}{[S] + [K_M]}$$

Constants

$$K = \frac{[R][L]}{[C^*]}$$


$$K_M = \frac{K_1 + K_2}{K_1}$$


$$A = \frac{A_{\max} [L]}{[L] + EC_{50}}, \quad A(t) = -\frac{d}{dt} [L](t)$$

$$-\frac{d}{dt} [L](t) = \frac{A_{\max} [L](t)}{[L](t) + EC_{50}}$$

$$-\frac{d}{dt} [L](t) = \frac{A_{\max} [L](t)}{[L](t) + EC_{50}} \approx A_{\max} \left[1 - e^{-\frac{[L](t)}{EC_{50}}} \right]$$


$$[L](t) = EC_{50} \ln \left\{ 1 + e^{-\frac{A_{\max} t}{EC_{50}}} \left[e^{\frac{[L_0]}{EC_{50}}} - 1 \right] \right\}$$



$$A(t) = \frac{A_{\max} e^{-\frac{A_{\max} t}{EC_{50}} \left(e^{\frac{[L_0]}{EC_{50}} - 1} \right)}}{1 + e^{-\frac{A_{\max} t}{EC_{50}} \left(e^{\frac{[L_0]}{EC_{50}} - 1} \right)}}$$

$$[L_0] \rightarrow \left\| \left\| y^{PREDICTAL} \right\| \right\|$$

$$r_{stot} = \sqrt{\frac{SQ - SR}{SQ}}$$


$$A = \log\left(\frac{1}{EC_{50}}\right) = -\lg EC_{50}$$

$$EC_{50} = e^{-A} = e^{-\|y^{obs}\|}$$

$$A\left(t = e^{\frac{1}{1-\tau}-1}\right), \tau \in [0,1]$$

$$A(\tau) = 0 \underset{\text{soluti}}{\overset{\text{numeric}}{\implies}} \tau_{\infty}$$

$$\left. \begin{array}{l} A_{\max} \rightarrow \infty \\ t \rightarrow \infty \end{array} \right\} \implies e^{-\frac{A_{\max} t}{EC_{50}}} \rightarrow 0$$

4. Quantum Correlation: Qua-SP(A)R

- Paradoxically, the main problem for QSAR resides not in performing the correlation itself but setting the variable selection for it; the mathematical counterpart for such problem is known as the “factor indeterminacy” and affirms that the same degree of correlation may be reached with in principle an infinity of latent variable combinations.
- the main point is that given a set of N -molecules one can chose to correlate their observed activities $A_{i=1,\overline{N}}$ with M -selected structural indicators in as many combinations as

$$C = \sum_{k=1}^M C_M^k \quad C_M^k = \frac{M!}{k!(M-k)!}$$
$$K = \prod_{k=1}^M C_M^k$$



- *endpoint spectral norm*

$$\| | Y_l \rangle \| = \sqrt{\langle Y_l | Y_l \rangle} = \sqrt{\sum_{i=1}^N y_{il}^2} \quad l = \overline{1, C}$$

- *algebraic correlation factor*

$$R_{ALG,l} = \frac{\| | Y_l \rangle \|}{\| | A \rangle \|} = \sqrt{\frac{\sum_{i=1}^N y_{il}^2}{\sum_{i=1}^N A_i^2}} \quad l = \overline{1, C}$$

- *spectral path*, with the distance defined in the Euclidian sense as

$$[l, l'] = \sqrt{(\| | Y_l \rangle \| - \| | Y_{l'} \rangle \|)^2 + (R_l - R_{l'})^2} \quad \forall (l, l') = \overline{1, C}$$

- *least spectral path principle, formally shaped as*

$$\delta[l_1, \dots, l_k, \dots, l_M] = 0; \quad l_1, \dots, l_k, \dots, l_M : \text{ENDPOINTS}$$


- *inter-endpoint norm difference (IEND),*

$$\Delta Y_{l|l'} = \left\| \|Y_{l'}\rangle\right\| - \left\| \|Y_l\right\rangle\right\| \quad (l, l') \in \{\alpha_1, \dots, \alpha_M\}$$

- *inter-endpoint molecular activity difference (IEMAD),*

$$\Delta A_{i|j}^{l|l'} = A_j^{l'} - A_i^l = \ln \frac{1}{(EC_{50})_j^{l'}} - \ln \frac{1}{(EC_{50})_i^l} = \ln \frac{(EC_{50})_i^l}{(EC_{50})_j^{l'}}$$

$$\ln \frac{1}{q_{i|j}^{l|l'}} \equiv \Delta Y_{l|l'} - \Delta A_{i|j}^{l|l'} \quad (EC_{50})_i^l = (EC_{50})_j^{l'} q_{i|j}^{l|l'} \exp(i\Delta Y_{l|l'})$$

- 
- the amplitude of transformation driven by the so called *quantum-SAR factor* of an exponential form

$$q_{i|j}^{l|l'} = \exp\left(\Delta A_{i|j}^{l|l'} - \Delta Y_{l|l'}\right)$$

- *identity*

$$(EC_{50})_i^l = (EC_{50})_i^l$$

$$(EC_{50})_j^{l'} = (EC_{50})_i^l \frac{1}{q_{i|j}^{l|l'}} \exp(-i\Delta Y_{l|l'})$$

- “*real*” *quantum-SAR transformation*

$$\left| (EC_{50})_i^l \right| = q_{i|j}^{l|l'} \cdot \left| (EC_{50})_j^{l'} \right|$$



■ *multiple transformations*


$$q_{i|t}^{l|l''} = q_{i|j}^{l|l'} \cdot q_{j|t}^{l'|l''}$$

$$\left| (EC_{50})_i^l \right| = q_{i|t}^{l|l''} \cdot \left| (EC_{50})_t^{l''} \right| = q_{i|j}^{l|l'} \cdot \left| (EC_{50})_j^{l'} \right| = q_{i|j}^{l|l'} \cdot \left(q_{j|t}^{l'|l''} \cdot \left| (EC_{50})_t^{l''} \right| \right)$$

$$q_{i_1|i_M}^{l_1|l_M} = \prod_{w=2}^M q_{i_{w-1}|i_w}^{l_{w-1}|l_w}$$

■ *self-transformation*

$$q_{i|j=i}^{l|l'} = \exp\left(-\Delta Y_{l|l'}\right)$$

- 
- With the present Qua-SAR methodology one can appropriately identify the molecular pairs that drive certain bio-/eco- activities against given receptor by means of selected descriptors in a “wave”- or “quantum” mechanistic formal way.
 - The ultimate goal will be the computation of quantum-SAR factors along the least paths of actions that give the potential information of the conversion power of the fittest molecules in their specific bindings.



References

- Pogliani, L. *Numbers Zero, One, Two, and Three in Science and Humanities*. Mathematical Chemistry Monographs Vol. 2, University of Kragujevac-Faculty of Science, Kragujevac, 2006.
- Putz, M.V. Systematic Formulation for Electronegativity and Hardness and Their Atomic Scales within Density Functional Softness Theory. *Int. J. Quantum Chem.* **2006**, *106*, 361-386.
- Putz, M.V. Semiclassical Electronegativity and Chemical Hardness. *J. Theor. Comp. Chem.* **2007**, *6(1)*, 33-47.
- Delaney, J. S.; Mullaley, A.; Mullier, G. W.; Sexton, G. J.; Taylor, R.; Viner, R. C. Rapid construction of data tables for quantitative structure-activity relationship studies. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 174-178.
- Klopman, G.; Balthasar, D. M.; Rosenkranz, H. S. Application of the computer-automated structure evaluation (CASE) program to the study of structure-biodegradation relationships of miscellaneous chemicals. *Environ. Toxicol. Chem.* **1993**, *12*, 231-240.
- Basketter, D.; Dooms-Goossens, A.; Karlberg, A.-T.; Lepoittevin, J.-P. The chemistry of contact allergy: why is a molecule allergenic? *Contact Dermatitis* **1995**, *32*, 65-73.
- Feijtel, T. C. J. Evaluation of the use of QSARs for priority settings and risk assessment. *SAR and QSAR in Environmental Research* **1995**, *3*, 237-245.
- Hermens, J. L. M.; Verhaar, H. J. M. QSARs in environmental toxicology and chemistry. *ACS Symposium Series* **1995**, *606*, 130-140.
- Hermes, J. Prediction of environmental toxicity based on structure-activity relationships using mechanistic information. *Sci. Total Environ.* **1995**, *171*, 235-242.
- Hermens, J.; Balaz, S.; Damborsky, J.; Karcher, W.; Müller, M.; Peijnenburg, W.; Sabljic, A.; Sjöström, M. Assessment of QSARs for predicting fate and effects of chemicals in the environment: an international European project. *SAR and QSAR in Environmental Research* **1995**, *3*, 223-236.
- Ogihara, N. Drawing out drugs. *Mod. Drug Discovery* **2003**, *6 (9)*, 28-32.
- Hansch, C.; Hoekman, D.; Gao, H. Comparative QSAR: toward a deeper understanding of chemicobiological interactions. *Chem. Rev.* **1996**, *96*, 1045-1075.
- Kubinyi, H. Der Schlüssel zum Schloß I. Grundlagen der Arzneimittelwirkung. *Pharmazie in unserer Zeit* **1994**, *23 Jahrg. Nr.3*, 158-168.
- Liwo, A.; Tarnowska, M.; Grzonka, Z. Tempczyk, A. Modified Free-Wilson method for the analysis of biological activity data. *Computers Chem.* **1992**, *16*, 1-9.
- Schmidli, H. Multivariate prediction for QSAR. *Chemometrics and Intelligent Laboratory Systems* **1997**, *37*, 125-134.
- Lhuguenot, J.-C. Relation quantitative structure-activité (QSAR): une méthode mal reconnue car trop souvent mal utilisée. *Ann. Fals. Exp. Chim.* **1995**, *88*, 293-310.
- Crippen, G. M.; Bradley, M. P.; Richardson, W. W. Why are binding-site models more complicated than molecules? *Perspectives in Drug Discovery and Design* **1993**, *1*, 321-328.
- Kier, L.B.; Hall, L.H. *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, Letchworth, 1986.



References

- Balaban, A.T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological indices for structure-activity correlations. *Top. Curr. Chem.* **1983**, *114*, 21-55.
- Navia, M. A.; Peattie, D. A. Structure-based drug design: applications in immunopharmacology and immunosuppression. *Immunology Today* **1993**, *14*, 296-301.
- Perkins, T. D. J.; Dean, P. M. An exploration of a novel strategy for superposing several flexible molecules. *J. Comput.-Aided Mol. Design* **1993**, *7*, 155-172.
- Lemmen, C.; Lengauer, T. Time-efficient flexible superposition of medium-sized molecules. *J. Comput.-Aided Mol. Design* **1997**, *11*, 357-368.
- Balaban, A. T.; Chiriac, A.; Motoc, I.; Simon, Z. *Steric Fit in QSAR*; Springer, Berlin (Lecture Notes in Chemistry Series), 1980.
- Simon, Z; Chiriac, A.; Holban, S.; Ciubotariu, D.; Mihalas, G. I. *Minimum Steric Difference. The MTD Method for QSAR Studies*; Res. Studies Press (Wiley), Letchworth, 1984.
- Duda-Seiman C., Duda-Seiman D., Dragoş D., Medeleanu M., Careja V., Putz M.V., Lacrămă A.-M., Chiriac A., Nuţiu R., Ciubotariu D. Design of Anti-HIV Ligands by Means of Minimal Topological Difference (MTD) Method, *Int. J. Mol. Sci.* **2006**, *7*, 537-555.
- Cramer, R.D.III; Patterson, D.E.; Bunce, J.D. Comparative molecular field analysis (CoMFA). 1. Effect shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
- Cramer, R.D.III; DePriest, S.A.; Patterson, D.E.; Hecht, P. The developing practice of comparative molecular field analysis. In *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), Escom, Leiden, 1993, pp. 443-485.
- Sun, J.; Chen, H.F.; Xia, H.R.; Yao, J.H.; Fan, B.T. Comparative study of factor Xa inhibitors using molecular docking/SVM/HQSAR/3D-QSAR methods. *QSAR Comb. Sci.* **2006**, *25*, 25-45.
- Randic, M.; Jerman-Blazic, B.; Trinajstic, N. Development of 3-dimensional molecular descriptors. *Comput. Chem.* **1990**, *14*, 237-246.
- Randic, M.; Razinger, M. Molecular topographic indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 140-147.
- Manallack, D. T.; Livingstone, D. J. Artificial neural networks: application and chance effects for QSAR data analysis. *Med. Chem. Res.* **1992**, *2*, 181-190.
- Manallack, D. T.; Livingstone, D. J. Limitations of functional-link nets as applied to QSAR data analysis. *Quant. Struct-Act. Relat.* **1994**, *13*, 18-21.
- Marchant, C. A.; Combes, R. D. Artificial intelligence: the use of computer methods in the prediction of metabolism and toxicity, in *Bioactive Compound Design: Possibilities for Industrial Use*, M. G. Ford, R. Greenwood (eds.), G. T. Brooks and R. Franke BIOS Scientific Publishers Limited, 1996.
- Moriguchi, I.; Hirono, S.; Matsushita, Y.; Liu, Q.; Nakagome, I. Fuzzy adaptive least squares applied to structure-activity and structure-toxicity correlations. *Chem. Pharm. Bull.* **1992**, *40*, 930-934.
- Moriguchi, I.; Hirono, S. Fuzzy adaptive least squares and its use in quantitative structure-activity relationships, in *QSAR and Drug Design – New Developments and Applications*, T. Fujita (ed.), Elsevier Science B. V., 1995.
- Vapnik, V.N. *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.
- Vapnik, V.N. *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, Berlin, 1982.



References

- Schölkopf, B.; Burges, C.J.C.; Smola, A.J. (eds.) *Advances in Kernel Methods. Support Vector Learning*. MIT Press, Cambridge, MA, 1999.
- Schölkopf, B.; Smola, A.J. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
- Mangasarian, O.L.; Musicant, D.R. Successive overrelaxation for support vector machines. *IEEE Trans. Neural Networks* **1999**, *10*, 1032-1036.
- Mattera, D.; Palmieri, F.; Haykin, S. Simple and robust methods for support vector expansions. *IEEE Trans. Neural Networks* **1999**, *10*, 1038-1047.
- Luan, F.; Ma, W.P.; Zhang, X.Y.; Zhang, H.X.; Liu, M.C.; Hu, Z.D.; Fan, B.T. QSAR study of polychlorinated dibenzodioxins, dibenzofurans, and biphenyls using the Heuristic method and support vector machine. *QSAR Comb. Sci.* **25**, *25*, 46-55.
- Sutter, J. M.; Kalivas, J. H.; Lang, P. K. Which principal components to utilize for principal component regression. *J. Chemometrics* **1992**, *6*, 217-225.
- Nendza, M.; Wenzel, A. Statistical approach to chemicals classification. *Environ. Toxicol. Chem.* **1993**, *Supplement*, 1459-1470.
- Cash, G. G.; Breen, J. J. Principal component analysis and spatial correlation: environmental analytical software tools. *Chemosphere* **1992**, *24*, 1607-1623.
- Hemmateenejad, B.; Miri, R.; Jafarpour, M.; Tabarzad, M.; Foroumadi, A. Multiple linear regression and principal component analysis-based prediction of the anti-tuberculosis activity of some 2-aryl-1,3,4-thiadiazole derivatives. *QSAR Comb. Sci.* **2006**, *25*, 56-66.
- Randić, M. Resolution of ambiguities in structure-property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311-320.
- Randić, M. Orthogonal Molecular Descriptors. *New J. Chem.* **1991**, *15*, 517-525.
- Amić, D.; Davidović-Amić, D.; Trinajstić, N. Calculation of retention times of anthocyanins with orthogonalized topological indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 136-139.
- Lučić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D. The structure-property models can be improved using the orthogonalized descriptors. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532-538.
- Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, A.; Mihalić, Z. A Structure-Property Study of the Solubility of Aliphatic Alcohols in Water. *Croatica Chem. Acta* **1995**, *68*, 417-434.
- Lučić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D.; Jurić, A. A Novel QSPR Approach to Physicochemical Properties of the α -Amino Acids. *Croatica Chem. Acta* **1995**, *68*, 435-450.
- Šoškić, M.; Plavšić, D.; Trinajstić, N. Link between orthogonal and standard multiple linear regression models. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 829-832.
- Klein, D.J.; Randić, M.; Babić, D.; Lučić, B.; Nikolić, S.; Trinajstić, N. Hierarchical orthogonalization of descriptors. *Int. J. Quantum Chem.* **1997**, *63*, 215-222.
- Ivanciuc, O.; Taraviras, S.L.; Cabrol-Bass, D. Quasi-orthogonal basis sets of molecular graph descriptors as chemical diversity measure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 126-134.
- Putz M.V. A Spectral Approach of the Molecular Structure – Biological Activity Relationship Part I. The General Algorithm. *Annals of West University of Timișoara, Series of Chemistry* **2006**, *15*, 159-166.
- Putz M.V.; Lacrămă A.M. A Spectral Approach of the Molecular Structure – Biological Activity Relationship Part II. The Enzymatic Activity, *Annals of West University of Timișoara, Series of Chemistry* **2006**, *15*, 167-176.