Introduction to VIPs

Variable importances for the projections in PLS

matthieu.lesnoff@cirad.fr

ChemHouse 06/03/2023

Some useful references (not exhaustive)

- **Chong, I.-G., Jun, C.-H., 2005**. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems 78, 103–112. https://doi.org/10.1016/j.chemolab.2004.12.011

- **Cocchi, M., Biancolillo, A., Marini, F., 2018.** Chapter Ten - Chemometric Methods for Classification and Feature Selection, in: Jaumot, J., Bedia, C., Tauler, R. (Eds.), Comprehensive Analytical Chemistry, Data Analysis for Omic Sciences: Methods and Applications. Elsevier, pp. 265–299. https://doi.org/10.1016/bs.coac.2018.08.006

- **Mehmood, T., Sæbø, S., Liland, K.H., 2020**. Comparison of variable selection methods in partial least squares regression. Journal of Chemometrics 34, e3226. https://doi.org/10.1002/cem.3226

- **Tenenhaus, M., 1998**. La régression PLS: théorie et pratique. Editions Technip, Paris.

Chong, I.-G., Jun, C.-H., Chemolab 2005

The VIP score of a predictor, first published in [6], is a summary of the importance for the projections to find *h* latent variables.

[6] Wold S., Johansson E., Cocchi M., 3D QSAR in Drug Design; Theory, Methods, and Applications, ESCOM, Leiden, Holland,**1993**, pp. 523– 550.

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Chemometrics and Intelligent Laboratory Systems 78 (2005) 103–112

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Performance of some variable selection methods when multicollinearity is present

Il-Gyo Chong, Chi-Hyuck Jun*

Department of Industrial Engineering, Pohang University of Science and Technology, San 31 Hyoja-dong, Pohang 790-784, Republic of Korea

Received 22 April 2004; accepted 22 December 2004
Available online 7 March 2005

$$\mathrm{VIP}_j = \sqrt{p \sum_{k=1}^{h} \left(SS(b_k \boldsymbol{t}_k) (\boldsymbol{w}_{jk} / \|\boldsymbol{w}_k\|)^2 \right) \Big/ \sum_{k=1}^{h} SS(b_k \boldsymbol{t}_k)},$$

where $$SS(b_k \boldsymbol{t}_k) = b_k^2 \boldsymbol{t}_k^t \boldsymbol{t}_k \qquad (2)$$

Ratio of two weighted sums

$$VIP(\boldsymbol{x}_j)^2 = \frac{\sum_{a=1}^{h} \frac{{w_{a,j}}^2}{\|\boldsymbol{w}_a\|^2} \times SS(\boldsymbol{y}, \boldsymbol{t}_a)}{\sum_{a=1}^{h} \frac{1}{p} \times SS(\boldsymbol{y}, \boldsymbol{t}_a)}$$

Theory

Notations

$\mathbf{X}$ n,p $\mathbf{y}$ n,1

Model $\mathbf{y}$ = PLS1($\mathbf{X}$) with h LVs

Score matrix $\mathbf{T}$ n,h

$\mathbf{x}_j$ n,1 column j of $\mathbf{X}$

$\mathbf{t}$ n,1 column of $\mathbf{T}$

General idea

VIP($\mathbf{x}_j$) = Compromise

how much is important

1) column $\mathbf{x}_j$ to build score $\mathbf{t}_a$

2) score $\mathbf{t}_a$ for explaining $\mathbf{y}$

Aggregation over scores (LVs) $\mathbf{t}_1, \ldots \mathbf{t}_h$ of the fitted PLS model

1) Importance of x_j to build a given score t

Use of the PLS weights **w**

- $\mathbf{t} = w_1 \, \mathbf{x}_{1,\text{ deflated}} + \ldots + w_p \, \mathbf{x}_{p,\text{ deflated}}$

$$w_j \propto \operatorname{cov}(\mathbf{x}_{j,\text{deflated}}, \mathbf{t})$$

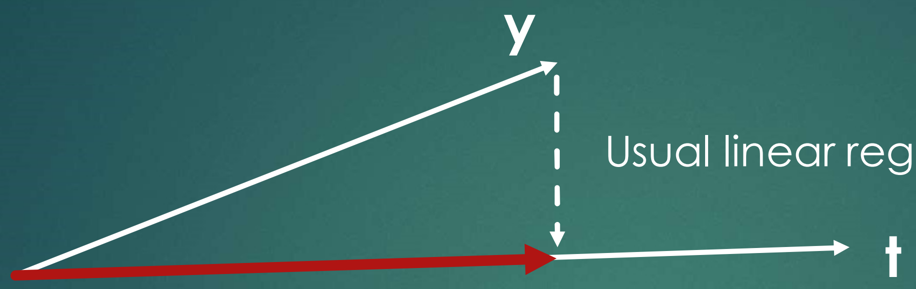
- $\underbrace{\text{Weight}(x_j)}_{\alpha_j} = \dfrac{w_j{}^2}{w_1{}^2+\dots+w_p{}^2} = \dfrac{w_j{}^2}{\|\boldsymbol{w}\|^2} = \big(w_j/\underbrace{\|\boldsymbol{w}\|}_{}\big)^2$

= 1 for usual PLS algorithms

- $\sum_{j=1}^{p} \alpha_j = 1$
- $\bar{\alpha} = \sum_{j=1}^{p} \alpha_j/p = 1/p$

If $\alpha_j > \bar{\alpha}$
$\Rightarrow$ $\mathbf{x}_j$ has a "high" influence on the building of **t**

2) Importance of score t to explain y



$\hat{\boldsymbol{y}} = \boldsymbol{t}\hat{\beta} =$ "**y** explained by **t**" $\quad \hat{\beta} = (\boldsymbol{t}'\boldsymbol{t})^{-1}\boldsymbol{t}'\boldsymbol{y}$

⇒ Proportion of variance(**y**) explained by **t**

$$R2(y,t) = \frac{Var(\hat{\boldsymbol{y}})}{Var(y)} = \frac{\|\hat{\boldsymbol{y}}\|^2}{\|\boldsymbol{y}\|^2}$$

(**y** and **t** are centered)

$$= \frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}}{\boldsymbol{y}'\boldsymbol{y}} = \frac{(\boldsymbol{t}\hat{\beta})'(\boldsymbol{t}\hat{\beta})}{\boldsymbol{y}'\boldsymbol{y}} = \frac{\hat{\beta}^2\boldsymbol{t}'\boldsymbol{t}}{\boldsymbol{y}'\boldsymbol{y}} = \frac{SS(\hat{\boldsymbol{y}})}{SS(\boldsymbol{y})} = \frac{SS(\boldsymbol{y},\boldsymbol{t})}{SS(\boldsymbol{y})}$$

1+2) over scores a = 1,...,h

$$VIP(\boldsymbol{x}_j)^2 = \frac{\sum_{a=1}^{h} \alpha_{a,j} \times R2(\boldsymbol{y}, \boldsymbol{t}_a)}{\sum_{a=1}^{h} \bar{\alpha} \times R2(\boldsymbol{y}, \boldsymbol{t}_a)}$$

Uniform weight for all the p variables

$$= \frac{\sum_{a=1}^{h} \left(w_{a,j}/\|\boldsymbol{w}_a\|\right)^2 \times SS(\boldsymbol{y}, \boldsymbol{t}_a)}{\sum_{a=1}^{h} 1/p \times SS(\boldsymbol{y}, \boldsymbol{t}_a)}$$

with $SS(\boldsymbol{y}, \boldsymbol{t}_a) = \hat{\beta}^2 \boldsymbol{t}_a{}' \boldsymbol{t}_a$

$\Rightarrow$ Same as in Chong & Jun 2005

- $\sum_{j=1}^{p} VIP(x_j)^2 = p$

- $\overline{VIP^2} = 1$ (p variables)

Chong, I.-G., Jun, C.-H., Chemolab 2005
Since the average of squared VIP scores equals 1, a "greater than one" rule is generally used as a criterion for variable selection.

- VIP : Same for PLSR, PCR, …

Can be computed for every LV-based regression methods

Case of PLSR2

multivariate response **Y** n,q

$$VIP(x_j)^2 = \frac{\sum_{a=1}^{h} \alpha_{a,j} \times R2(\boldsymbol{Y}, \boldsymbol{t}_a)}{\sum_{a=1}^{h} \bar{\alpha} \times R2(\boldsymbol{Y}, \boldsymbol{t}_a)} = \frac{\sum_{a=1}^{h} \alpha_{a,j} \times SS(\boldsymbol{Y}, \boldsymbol{t}_a)}{\sum_{a=1}^{h} \bar{\alpha} \times SS(\boldsymbol{Y}, \boldsymbol{t}_a)}$$

with

- $\widehat{\boldsymbol{\beta}} = (\boldsymbol{t}'\boldsymbol{t})^{-1}\boldsymbol{t}'\boldsymbol{Y} \qquad 1, \boldsymbol{q}$
- $\mathrm{SS}(\boldsymbol{Y}, \boldsymbol{t}) = \left\|\widehat{\boldsymbol{Y}}\right\|^2 = \mathbf{Tr}\left[(\boldsymbol{t}\widehat{\boldsymbol{\beta}})'(\boldsymbol{t}\widehat{\boldsymbol{\beta}})\right]$

Case of PLSDA

Categorical **y** $\Rightarrow$ Table of dummy variables **Y**

$\Rightarrow$ PLS2(Y)

$\Rightarrow$ use of the VIP-PLS2 formula

Redundancy VIP

Tenehaus 1998 (p.139) replaces the proportion of variance explained R2(**Y**,**t**) by the redundancy index Rd(**Y**,**t**)

- $Rd(\boldsymbol{Y},\boldsymbol{t}) = \frac{1}{q}\sum_{j=1}^{q} cor^2(\boldsymbol{y}_j,\boldsymbol{t})$

Used in package
R mixOmics

- Univariate case: Rd(**y**, **t**) = R2(**y**, **t**)