Regularized Canonical Correlation Analysis for studying relationships between sensorial and physicochemical measurements of rabbit meat

S. Déjean[(1)], S. Combes[(2)], I. González[(1,3)], C. Larzul[(4)], L. Cauquil[(2)] and A. Baccini[(1)]

(1) Institut de Mathématiques de Toulouse UMR 5219
Université de Toulouse et CNRS

(2) Tissus Animaux, Nutrition, Digestion, Ecosystème et Métabolisme UMR 1289
INRA et Université de Toulouse

(3) Departamento de Matemática - Facultad de Ciencias y Tecnología
Universidad de Carabobo - Venezuela

(4) Station d'Amélioration Génétique des Animaux - INRA

Data acquired on rabbit meat

96 experimental units (rabbits) on which were acquired 2 sets of variables :

Sensory analysis

10 variables evaluated in back and leg by trained tasters

- tenderness
- sticky
- fibrous
- juiciness
- flavour

Physicochemical measurements

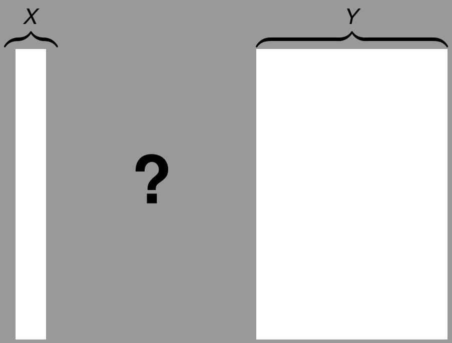
63 variables concerning

- pH
- weight
- color
- water content
- cooking loss
- shear test
- ...

☞ *Physicochemical measurements are detailed in Combes et al. (2008)*.

Question

How to highlight relationships between sensory analysis (X) and physicochemical measurements (Y) ?



Principle of CCA

Canonical Correlation Analysis (CCA) is the appropriate tool to deal with such a problem. The general objective of CCA is to explore linear relationships between two sets of quantitative variables observed on the same experimental units. This is achieved by finding the largest correlation between a linear combination of the variables in the first set and a linear combination of the variables in the second set.

CCA : mathematical aspects

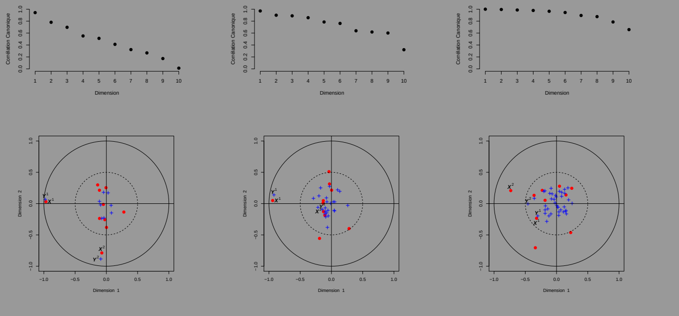
1. Orthogonal projections in $\mathbb{R}^n$ onto the respective column-space of X and Y :

$$P_X = X(X'X)^{-1}X'$$
$$P_Y = Y(Y'Y)^{-1}Y'$$

2. Spectral analysis of $P_X P_Y$ (or $P_Y P_X$)
   - canonical correlations ρ_s are the positive square roots of the eigenvalues λ_s of $P_X P_Y$ (which are the same as those of $P_Y P_X$) : $\rho_s = \sqrt{\lambda_s}$;
   - vectors $U^1, \ldots, U^p$ are the standardized eigenvectors corresponding to the decreasing eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ of $P_X P_Y$;
   - vectors $V^1, \ldots, V^p$ are the standardized eigenvectors corresponding to the same decreasing eigenvalues of $P_Y P_X$.

Limitation of CCA : simulation study

Effect of the ratio $n/(p+q)$ ($n = 50$) on the scree graph of canonical correlations and variables plots. True underlyning cross-correlations are determined by $\rho(X_1, Y_1) = 0.9$ and $\rho(X_2, Y_2) = 0.7$.



$p = 10, q = 10$ | $p = 10, q = 25$ | $p = 10, q = 39$

☞ Even if CCA can be performed in the three cases, the true underlying correlation can only be highlighted in the first one.

Principle of RCCA

- The principle of Regularized CCA was first proposed by Vinod (1976) then developped by Leurgans *et al.* (1993)
- Ill-conditioned covariance matrices $X'X$ and $Y'Y$ require regularization :

$$P_X = X(X'X + \lambda_1 I_p)^{-1} X'$$
$$P_Y = Y(Y'Y + \lambda_2 I_q)^{-1} Y'$$

$\lambda_1, \lambda_2 \geq 0$, enable inversibility

- As in standard CCA, spectral analysis of $P_X P_Y$ (or $P_Y P_X$) is performed.

☞ How to set regularization parameters λ_1 and λ_2 ?

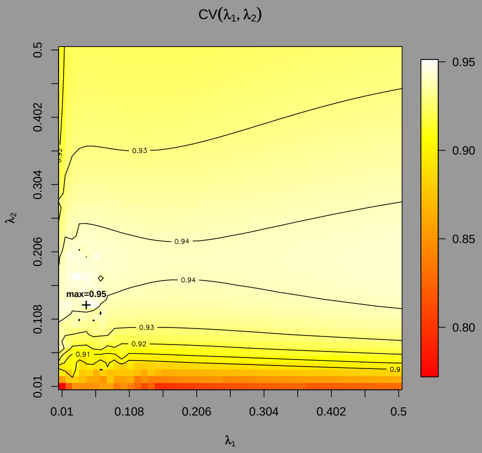
Tuning λ_1 and λ_2 by cross-validation

	X	Y
l_1	X_{l_1}	Y_{l_1}
l_2	X_{l_2}	Y_{l_2}
$\vdots$	$\vdots$	$\vdots$
l_k	X_{l_k}	Y_{l_k}
$\vdots$	$\vdots$	$\vdots$
l_M	X_{l_M}	Y_{l_M}

1. The dataset is randomly split into M roughly equal-size parts
2. For a fixed value of λ, compute $\rho_\lambda^{(-l_k)}$, the canonical correlation computed with the k^{th} part of the data removed. Let $a_\lambda^{(-l_k)}$ and $b_\lambda^{(-l_k)}$ be the maximizers of $\rho_\lambda^{(-l_k)}$, for $k = 1, \ldots, M$.
3. Define the CV-criterion $CV(\lambda_1, \lambda_2) = cor\left(\left\{X_{l_k} a_\lambda^{(-l_k)}\right\}_{k=1}^{M}, \left\{Y_{l_k} b_\lambda^{(-l_k)}\right\}_{k=1}^{M}\right)$
4. Optimal λ value is defined as : $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2) = arg\max_{\lambda_1,\lambda_2} \; CV(\lambda_1, \lambda_2)$

Cross-validation criterion map

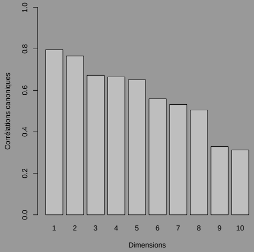
☞ *Processing is performed using the* `CCA` *package for R (González et al., 2008).*

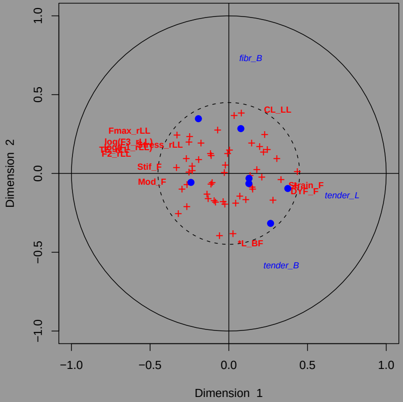


Cross-validation criterion map to determine optimal values for regularization parameters on a 50 $\times$ 50 grid between 0.01 and 0.5. The symbol "**+**" locate the point of the grid (0.05,0.13) on which the maximum (0.95) is reached.

Variables plot

Results of the RCCA : scree graph of canonical correlations (on the left) and variables plots on the plane defined by the first two dimensions (on the right).





☞ Negative correlations between tenderness (*tender_B* and *tender_L*) and measurements acquired from shear test (*Fmax_rLL*, *log(F3_rLL)*, *TE_rLL*, *log(F1_rLL)* and *F2_rLL*)

☞ Positive correlations between tenderness in the leg (*tender_L*) and variables (*DYF_F* and *Strain_F*) representing bone elasticity.

Bibliography

- González, I., Déjean, S., Martin, P., Baccini, A. (2008). CCA : an R package to extend canonical correlation analysis. *Journal of Statistical Software, to appear*.
- Combes, S., Larzul, C., Jehl, N., Cauquil, L., Gabinaud, B., Lebas, F. (2008). Ability of physicochemical measurements to discriminate rabbit meat from three different productive processes. *Journal of the Science of Food and Agriculture*, to appear.
- Leurgans, S.E., Moyeed, R.A., Silverman, B.W. (1993). Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society Series B*, 55, 725–740.
- Vinod, H.D. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 6, 129–137.