Multivariate methodologies

L. Villanova¹

¹Department of Statistical Science
University of Padua

10 December, 2009

Supervisors
Prof. Irene Poli (University Ca’ Foscari of Venice)

Cosupervisors
Prof. Kate Smith-Miles (Monash University)
Prof. Rob J Hyndman (Monash University)
Prof. Irene Poli
Department of Statistics,
University Ca’ Foscari of Venice

Prof. Kate Smith-Miles
Head of School of Mathematical Sciences,
Monash University

Prof. Rob Hyndman
Department of Econometrics and Business Statistics,
Monash University
Outline

The motivating problem

How it has been faced

Can the approach be improved?

Ongoing research
Purpose: improve coatings quality

1. functionalize a glass slide
   ▶ the glass slide is covered with a thin film with specific properties (6 chemical components are used)
   ▶ single-stranded DNA sequences can graft the surface

2. spot a glass slide
   ▶ a solution with single-stranded DNAs is deposited over the functionalized slide
   ▶ complementary DNA strands zipper up

3. measure the coating quality
   ▶ high quality coatings generate high quality spots (circular, high intensity, low background, homogeneous intensity)
Method: evolutionary design of experiments

- coatings quality as a surface to be optimized
- experiments as points in that surface

<table>
<thead>
<tr>
<th>exp</th>
<th>$X_1$</th>
<th>...</th>
<th>$X_p$</th>
<th>$Y_1$</th>
<th>...</th>
<th>$Y_q$</th>
<th>Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp_1</td>
<td>$x_{11}$</td>
<td>...</td>
<td>$x_{1p}$</td>
<td>$y_{11}$</td>
<td>...</td>
<td>$y_{1q}$</td>
<td>$f_1$</td>
</tr>
<tr>
<td>exp_2</td>
<td>$x_{21}$</td>
<td>...</td>
<td>$x_{2p}$</td>
<td>$y_{21}$</td>
<td>...</td>
<td>$y_{2q}$</td>
<td>$f_2$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>exp_n</td>
<td>$x_{n1}$</td>
<td>...</td>
<td>$x_{np}$</td>
<td>$y_{n1}$</td>
<td>...</td>
<td>$y_{nq}$</td>
<td>$f_n$</td>
</tr>
</tbody>
</table>
The ‘experiments evolution’: 3D simulation
The ‘experiments evolution’: 3D simulation
The ‘experiments evolution’: 3D simulation
The ‘experiments evolution’: 3D simulation
The ‘experiments evolution’: 3D simulation
The ‘experiments evolution’: 3D simulation
The ‘experiments evolution’: 3D simulation
The ‘experiments evolution’: 3D simulation
The ‘experiments evolution’: 3D simulation
The ‘experiments evolution’: 3D simulation
Results

- best coating: one variable at a time (std)
- best coating: evolutionary design of experiments

<table>
<thead>
<tr>
<th>Variable</th>
<th>Weight</th>
<th>Result wrt std</th>
</tr>
</thead>
<tbody>
<tr>
<td>stability (0; 1)</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>spot intensity</td>
<td>40%</td>
<td>380%</td>
</tr>
<tr>
<td>spot circularity</td>
<td>40%</td>
<td>unchanged</td>
</tr>
<tr>
<td>background</td>
<td>12%</td>
<td>39%</td>
</tr>
<tr>
<td>spot homogeneity</td>
<td>8%</td>
<td>89%</td>
</tr>
<tr>
<td>signal-to-noise ratio</td>
<td>-</td>
<td>246%</td>
</tr>
</tbody>
</table>
How to improve

- currently experiments are ‘evolved’ combining
  - nature-inspired rules of the evolutionary algorithm
  - predictions from a model fitted to the available observations (executed experiments)

- improvements are expected if
  - model with improved predictive accuracy is fitted
Open questions

- given an applicative problem, how do we select (a priori) the most appropriate model (or subset of models)?
- what makes a particular model work well (or not) for a particular problem?
- do the characteristics of a problem relate to the model performance?
- what if data are not normally distributed?
- what if data are not enough to obtain a good fit?

SIMULATION STUDY
Simulation study

- predict which method will perform best for unseen problems
- generalization of meta-learning for algorithm selection (StatLog and METAL european projects)
Multivariate regression

► \( Y = (Y_1, \ldots, Y_q)^T \), random \( q \)-vector-valued output variable with mean vector \( \mu_Y \) and covariance matrix \( \Sigma_{YY} \)

► \( X = (X_1, \ldots, X_p)^T \), nonstochastic \( p \)-vector-valued input variable (fixed-\( X \) case)

► \( n \) replications \( (X_k^T, Y_k^T)^T, k = 1, 2, \ldots, n \)

► purposes
  ► estimate \( q \) response variables using a common set of \( p \) input variables
  ► take into account the dependencies
    ► between \( X \) and \( Y \)
    ► within \( X \)
    ► within \( Y \)
Problems

- $N = 50, 200, 500$ observations
- $p = 5, 10, 70$ input variables (predictors)
- $q = 3, 8$ output variables (responses)
- predictors randomly generated (then considered fixed)
  - $X \sim \mathcal{N}(0, \Sigma_x)$,
  - $[\Sigma_x]_{ij} = \begin{cases} r|\!|i-j|\!, & r \sim U[-1, 1] \\ \sigma_x^{|i-j|}, & \sigma_x = (0.1, 0.8) \end{cases}$
- responses $Y = f(X) + E$
Problems - function structures

\[ f_k(X) = \sum_{i=1}^{p} a_i x_i + \sum_{i,j=1}^{p} a_{ij} x_i x_j \]

- \( i, j = 1, \ldots, p \) and \( k = 1, \ldots, q \)
- \( a_i \sim U[-r_1, r_2] \),

\[ a_{ij} = \begin{cases} 0, & a_{ij} \in A \\ U[-r_1, r_2] & a_{ij} \notin A \end{cases} \]

where \( A \in \{a_{ij}\} \) and \( |A| = \left\lceil \sqrt{p^2 \cdot 10/100} \right\rceil \),

\( r_{1,2} \sim U[0, 10] \).

- \( f_k(X) \) from WFG Toolkit
Problems - WFG Toolkit

- flexible tool to generate test problems for multiresponse optimization methods
- transformation function combined to add complexity (bias, multimodality, deceptivity, nonseparability...)

![Graphs and plots related to WFG Toolkit]
Problems - noise structure

- $\mathbf{E}$ randomly generated
  - $\mathbf{E} \sim \mathcal{N}(\mathbf{0}, \Sigma_e)$
  - $\Sigma_e = \begin{cases} \sigma^2 I_q, & \sigma = (0.1, 0.8) \\ \sigma^2 \text{diag}(\{j^2\}_1^q), & \sigma = (0.1, 0.8) \end{cases}$

- $\mathbf{E} \sim C(F_1, \ldots, F_q)$
  - $C$, copula
  - $F_j, j = 1, \ldots, q$ marginal distributions (normal, gamma, exponential...
Copulas

- multivariate distributions with support in \([0, 1]^n\) and with uniform marginals
- **Sklar’s theorem** given a \(n\)-dimensional distribution function \(F\) with continuous (cumulative) marginal distributions \(F_1, \ldots, F_n\), there exists a unique \(n\)-copula \(C\): \([0, 1]^n \rightarrow [0, 1]\) s.t.

\[
F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n))
\]

- given a multivariate distribution \(F\) with marginals \(F_1, \ldots, F_n\), the copula \(C\) is obtained

\[
C(u_1, \ldots, u_n) = F(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n))
\]

- standard copulas are obtained with Gaussian or Student’s \(F\)
Multivariate regression

- two approaches:
  - a separate model for each response (equation-by-equation/separate modelling)
  - a single model to estimate all the responses simultaneously (simultaneous modelling)
Separate modelling

- **shrinkage methods** (biased and regularized regression)
  - ridge regression (RR)
  - lasso regression
  - principal component regression (PCR)
  - partial least squares (PLS)

- **variable selection**
  - stepwise methods
  - all possible subsets
Simultaneous modelling

- multivariate ridge regression
- curds and whey (C&W)
- reduced-rank regression (RRR)
- filtered canonical y-variate regression (FCYVR)
- two-block partial least squares (PLS2)

- take advantage of the correlation between responses to improve predictive accuracy
- based on canonical correlation analysis (CCA)
Curds and whey

- linear combination of the OLS predictors $\tilde{y}$
  \[ \tilde{y} = B_{c&w} \hat{y} \]

- $B_{c&w}$, optimal shrinkage matrix
- $T$, $q \times q$ matrix whose rows are the response canonical co-ordinates
- $D = \text{diag}(d_1, \ldots, d_q)$, and $d_i = \frac{c_i^2}{c_i^2 + \frac{p}{N} (1-c_i^2)}$
- $c_i^2$, canonical correlations
**RRR and FCYVR**

- **reduced rank regression**, $r = \text{rank}(A)$

\[
\tilde{A}_r = B_r \hat{A}, \quad B_r = T^{-1} I_r T
\]

\[
I_r = \text{diag}\{1(i \leq r)\}^q
\]

- **filtered canonical y-variate regression**

\[
\tilde{A} = B_f \hat{A}, \quad B_f = T^{-1} F T
\]

\[
F = \text{diag}\{f_1, \ldots, f_q\}, \quad f_i = \frac{c_i^2 - \frac{p-q-1}{N}}{c_i^2 \left(1 - \frac{p-q-1}{N}\right)}
\]
Methods... in summary

- separate modelling (equation-by-equation)
- simultaneous modelling
  - RRR borrows strength among responses by truncating the CCA
  - C&W is a smooth version of RRR
  - FCYVR depends also on the number $q$ of responses
  - PLS2, iterative computational algorithm
- nonlinear fitting (i.e. neural networks, MARS, polynomial regression, penalized splines)
Performances

- predictive accuracy (MSEP) using cross-validation
Problem characterization

- features extraction
  - statistical measures of the data
  - indirect approach: relates features to the performance of a set of methods

- landmarking
  - measure the performance of simple and efficient learning algorithms
  - direct approach: relates performance of some learners - *landmarkers* - to the performance of some other algorithms
Features

- **dataset characteristics** (number of observations, input and output variables, collinearity)

- **basic statistical measures** (mean, standard deviation, quartiles, skewness, kurtosis) on
  - data
  - variance covariance matrix
  - euclidean distances
Features

- other features from PCA, CCA, decision trees
  - number of eigenvalues retained to explain the 95% of the total variability (if high → lot of high-variance directions → less shrinkage required)
  - number of pairs of canonical variates with nonzero canonical correlations (identify rank of coefficient matrix, canonical correlation coefficients → proportion of variance explained by regression of \( \alpha Y \) on \( Y \))
  - number of splits and surrogate splits, concordance primary-surrogate split (identify interactions, highly correlated variables, possible confounders)
Landmarking

- performance measures ($R^2$, AIC, MSEP) and their statistics of
  - linear regression
  - regression trees
Simulation: last step

- regression and classification methods
- obvious results, useful to validate the simulation method

<table>
<thead>
<tr>
<th>data characteristics</th>
<th>best methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>compressable data</td>
<td>RR, RRR</td>
</tr>
<tr>
<td>nonlinear relationships</td>
<td>MARS, NN</td>
</tr>
<tr>
<td>highly correlated responses</td>
<td>C&amp;W, FCYVR</td>
</tr>
</tbody>
</table>

- new, less intuitive, results
Example of expected results

▶ classification model

*C&W recommended if*  
\((\text{# responses} < 10) \text{ and } (\text{skewness} < 0.57) \text{ and } (\text{cor}^2(y_i, y_j) > 0.3, \ i \neq j)\)

*RR recommended if*  
\((\text{# input} > 20) \text{ and } (\text{# PCs} < 3)\)

▶ regression model

\[
\text{MSEP(C&W)} = 5.5 - 3 \cdot \text{Collinearity} + 0.72 \cdot R^2
\]

\[
\text{MSEP(RRR)} = 52 + 21 \cdot \text{Skewness} + 5.3 \cdot (\text{# nodes})
\]
Research status

- R code
  - developed for methods, performances, and WFG toolkit functions
  - under development for problem simulation and characterization
Research contributions

- tool to cope with multivariate problems
- R code
  - to model data with multiple responses
  - to predict the best model given an unseen problem
- cross-disciplinary perspective (borrowing from other literature)
- data generation process
Aknowledgements

▶ Prof. Kate Smith-Miles and Prof. Rob Hyndman
▶ Prof. Irene Poli
▶ Prof. Alessandra Salvan
▶ Ph.D course in statistics
▶ ECLT, CIVEN, Fondazione di Venezia
▶ Elena Zamborlin
Main references (ongoing research)

L. Breiman and J. H. Friedman
Predicting Multivariate Responses in Multiple Linear Regression

K.A. Smith-Miles
Cross-disciplinary perspectives on meta-learning for algorithm selection

A. J. Izenman
*Modern Multivariate Statistical Techniques*

T. Hastie, R. Tibshirani, and J. Friedman
*The Elements of Statistical Learning*