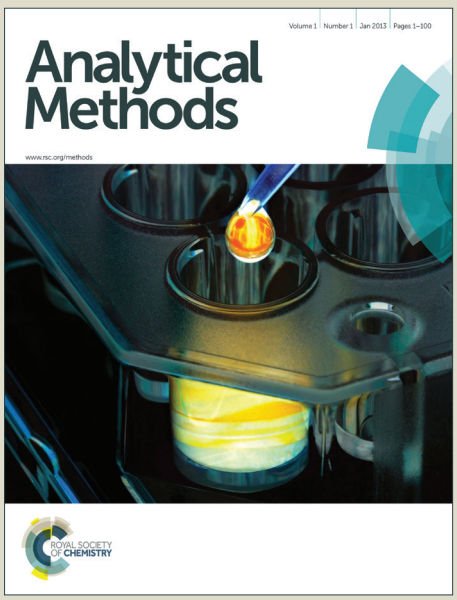
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Assessing the discrimination potential of linear and non-linear supervised chemometrics methods on a filamentous fungi FTIR spectral database

V. Gaydou [1,2], A. Lecellier [1,2], D. Toubas [1,2,3], J. Mounier [4], L. Castrec [4], G. Barbier [4], W. Ablain [5], M. Manfait [1,2], G.D. Sockalingum [1,2]*

[1] Université de Reims Champagne-Ardenne, MéDIAN-Biophotonique et Technologies pour la Santé, UFR de Pharmacie, 51 rue Cognacq-Jay, 51096 REIMS cedex, France

[2] CNRS UMR7369, Matrice Extracellulaire et Dynamique Cellulaire, MEDyC, Reims, France

[3] Laboratoire de Parasitologie Mycologie, CHU de Reims, Hôpital Maison Blanche, 45 rue Cognacq-Jay, 51092 Reims cedex, France

[4] Laboratoire Universitaire de Biodiversité et Ecologie Microbienne (EA3882), SFR148 SclnBioS, Université Européenne de Bretagne, Université de Brest, ESIAB, Technopôle de Brest Iroise, 29280 Plouzané, France

[5] AES CHEMUNEX/BIOMERIEUX, Rue Maryse Bastié, CS17219 Ker Lann, 35172 Bruz cedex, France

*Corresponding author:

Ganesh D. Sockalingum

Université de Reims Champagne-Ardenne

MéDIAN, Biophotonique et Technologies pour la Santé

Unité MEDyC, CNRS FRE3481

UFR Pharmacie, SFR CAP-Santé FED4231

51 rue Cognacq-Jay, Reims, France.

Tel: +33 3 26 91 35 53

Fax: +33 3 26 91 35 50

ganesh.sockalingum@univ-reims.fr

Abstract

This study proposes a comparative investigation of different linear and non-linear chemometrics methods applied to the same database of infrared spectra for filamentous fungi discrimination and identification. The database was comprised of 277 strains, (14 genus, 36 species), identified and validated by DNA sequencing, and analyzed by high-throughput Fourier Transform Infrared (FTIR) spectroscopy in the 4000-400 cm^{-1} wavenumber range. A cascade of 20 supervised models based on taxonomic ranks was constructed to predict classes until the species taxonomic rank. The cascade modeling was used to test 11 algorithms (5 linear and 6 non-linear) of supervised classification methods. To assess these algorithms, indicators of classification rates and McNemar's tests were defined and applied in same way to each of them. For non-linear algorithms, the KNN (K Nearest Neighbors) method proved to be the best classifier (78%). Linear algorithms, PLS-DA (Partial Least Square - Discriminant Analysis) and SVM (Support Vector Machine) showed better performances than non-linear methods with the best classification potential (~93%). SVM and PLS-DA were comparable and a possible complementarity between these two algorithms was highlighted.

Keywords: Supervised classification, cascade models, infrared spectroscopy, fungi identification

1 Introduction

Spectrometric techniques play an important role in both research and industrial applications. The development of these techniques has continuously progressed in order to exploit at best their capacities. Among these, infrared spectroscopy has emerged as a promising approach for rapid analysis. In mid-infrared spectroscopy, the molecular fundamental vibrational modes are measured and involve wavelengths between 2.5 and 25 micrometers corresponding to the wavenumber range of 4000-400 cm^{-1}. It relies on the absorption of mid-infrared light by vibrational transitions in covalent bonds. Fourier transform infrared (FTIR) spectroscopy has high molecular sensitivity and reveals numerous types and modes of vibrations. It is fast, label-free, cost-effective, easy to use, and applicable to various fields. However, it is perturbed by aqueous states and by the atmospheric water vapor and carbon dioxide. In near-infrared (NIR) spectroscopy, the sample receives wavelengths in the range of 800-2500 nm, whereby molecular overtone and combination vibrations are measured. The spectra are more complex and it can be difficult to assign specific features to specific chemical components. The molar absorptivity in the NIR region is typically quite small but NIR has the advantage that it can typically penetrate much farther into a sample than mid infrared radiation. NIR spectroscopy is, therefore, not a particularly sensitive technique, but it can be very useful in probing bulk material with little or no sample preparation. Numerous analytical attempts have been described in the literature with their advantages and disadvantages [1]. FTIR and NIR approaches have proved to be very effective in the characterization, differentiation, and identification of various fungi [2,3,4,5]. Another analytical method that has emerged as a new discovery tool for bacterial characterization (Lay, 2001) is matrix-assisted laser desorption/ionization time-of-flight mass spectroscopy (MALDI-TOF MS). It is an analytical tool sensitive to molecular composition and distinct mass signals can be observed in a mass-

to-charge (m/z) ratio. It allows molecular profiling such as protein profiling. Its potential to discriminate filamentous fungi of clinical origin at the species level has been demonstrated giving comparable results as with molecular identification methods (Cassagne et al., 2011; De Carolis et al., 2012). However, MALDI-TOF spectrometric databases for filamentous fungi, particularly from the food industry, are still under development (Santos et al., 2010b). A recent work, reports on the useful integration of different analytical imaging techniques, including FTIR and MALDI-TOF, in a multimodality platform for a deeper characterization of the potential medicinal fungus *Hericium* coralloides [6].

The progress in analytical spectroscopy and speed of data acquisition has also led to the construction of large and complex datasets. In order to exploit these large datasets sophisticated statistical methods were developed [7,8,9]. The field of chemometrics has thus emerged as a powerful approach for data mining, interpretation, and understanding; specifically for extracting relevant molecular information in different fields of spectroscopy. Recent advances in computing and chemometric allow choosing a wide variety of statistical algorithms to analyze the same spectral data bank.

The aim of this study was to compare the discriminating potential of 11 algorithms on the same dataset of 5960 FTIR spectra of filamentous fungi collected from 277 fungal strains belonging to 14 genera and 36 species. Among these, 194 strains (4159 spectra) were used for the model optimization and calibration steps and 83 strains (1801 spectra) were used for the external validation step. The assessed methods were all supervised discrimination methods requiring a calibration step and grouped in two categories. The first category concerns the linear methods with the Factorial Discriminant Analysis (FDA) method which is the most famous method and was introduced by Fisher in 1936 [10]. Then comes the Linear Discriminant Analysis (LDA) method, with a discrimination rule quite equivalent to that of

FDA [11]. The Partial Least Square-Discriminant Analysis (PLS-DA) method is more recent, ensues from the PLS algorithm, and was reported by Wold and Martens in 1983 [12]. The Soft Independent Modeling of Class Analogies (SIMCA) method was described by Wold and Sjöström in 1977 [13], which is a less used method. The second category concerns the non-linear methods comprised the Support Vector Machine (SVM) method was proposed by Vapnik et al. in 1963 [14]. The K-Nearest Neighbors (K-NN) methods, introduced by J. H. Friedman in 1975 [15], was developed to answer discrimination challenge with various kind of data. The Probalistic Neural Network (PNN) method is relatively recent and was presented by Specht in 1990 [16]. It is based on the analogy with the functioning of the brain of superior organisms. Networks are formed by small units, called neurons connected them. Quadratic Discriminant Analysis (QDA) method, proposed by Wold in 1976, is based on a quadratic function to apply the discrimination law [9].

The chemometrics question raised in this study was to select the most appropriate statistical method able to discriminate and identify an unknown strain of filamentous fungi from its FTIR spectrum and using a pre-established and non-exhaustive spectral library recently generated in our group [2,3].

In order to assess these supervised discrimination methods in terms of statistical significance, indicators of classification rates and McNemar's tests were defined and applied in same way to each of the studied algorithm.

2 Materials and Methods

2.1 Sample preparation and FTIR analysis

Two hundred and seventy-seven fungi strains belonging to 14 genera and 36 species and yielding 5960 spectra were used in this study. They were from the Université de Bretagne Occidentale and Centraalbureau voor Schimmelcultures culture collections and were identified by sequencing of specific DNA region like the rDNA internal transcribed spacer (ITS) region.

Cryopreserved strains were first sub-cultured on Sabouraud agar slants (Becton Dickinson, Le Pont de Claix, France) and incubated for 4 to 7 days at 25 °C depending on the strain. The cultures were transferred to an M tube, each sample was dissociated using a cycle of 100 seconds. Two milliliters of dissociated mycelia suspension were then transferred into an Eppendorf tube. The culture medium was then eliminated by centrifugation, the mycelia were washed and suspended in 1 ml of 0.9 % physiological saline and the supernatant was eliminated by another centrifugation. The mycelia pellets were recovered in 300 µl of physiological saline. Finally the samples were deposited on an IR-transparent 384-well silicon plate and dried into thin films. For reproducibility concerns, 3 independent cultures of each strain prepared on 3 different days (biological replicates), were performed and for each culture several instrumental replicates were recorded. The plate was then analyzed with a high-throughput module (HTS-XT) coupled with a Tensor 27 FTIR spectrometer (Bruker Optics, Etlingen, Germany).

The FTIR acquisition parameters were 64 accumulations per well with a spectral resolution of 4 cm^{-1}, in the spectral range of 4000-400 cm^{-1}. The spectral recording and preprocessing procedures were carried out by the OPUS 6.5. The preprocessing included baseline correction, second derivative, and vector normalization. The wavenumber ranges selected for the data bank were 800-800 cm^{-1} and 2800-3200 cm^{-1}. Details of the experimental protocol were reported recently [2,3].

2.2 Cascade modeling and building of the calibration and validation sets

For FTIR spectroscopic data, the establishment of a single model of discrimination, parameterized by more than around thirty clusters is quite challenging, particularly for linear algorithms. Such one-model procedure is difficult to implement since the zones of variance and co-variance overlap and become inconsistent with the number of clusters. For this reason, a modeling called “in cascade” has been developed [17] to circumvent the problem in this study (figure 2). The particularity of the cascade modeling is that it is parameterized from a reference arborescence and for the study presented here, it is the taxonomic classification of fungi that is used in this respect. At every taxonomic rank, samples were distributed in subphylum, class, order, family, genus, subgenus, section, serial, and species. In so doing, several “subgroups” were established at every rank and for each model the number of clusters was around 3 and so on, until the last rank called “species” rank is reached. The taxonomic tree is thus used to structure the data matrix in a subgroup and cluster cascade. We call "taxonomic nodes" the subgroups highlighted by the taxonomic tree. For every taxonomic node, a discrimination model was built. So, this technique allowed constructing the discrimination model in cascade including not less than 20 models with a maximum of 7 models required to reach the species taxonomic rank as regards to *Camemberti* strains.

The main advantage of the cascade modeling is that it allows obtaining a strong method of discrimination although the final number of clusters is high. On the other hand, this method is completely parameterized and thus totally dependent on the cascade reference to which it is associated. Yet, the fungal taxonomy is in constant evolution and consequently training variation on taxonomic nodes can influence the outcome in a significant way.

The data matrix was split into 2 sets; about 4159 spectra (194 strains) of samples were attributed to the calibration set and the rest (1801 spectra, 83 strains) to the validation set. In fact, one third of strains of each of 36 species represented in spectral data bank was randomly selected to constitute the validation spectral data set and to ensure that the relative variance of the validation set is inferior to that of the calibration set [18].

2.3 Partial cross validation for parameter optimization

Fundamentally, cross validation was developed for chemometrics experiments with a low sample population [19]. Because of this low population it is impossible to split the data matrix into calibration and validation sets while keeping a representative sample set. Thus, the cross validation allows to estimate the accuracy and robustness with one sample set only. For the present study, the calibration set was used with cross validation to optimize chemometrics parameters of all the studied algorithms presented in table 1 [20].

A large number of spectra are available in the calibration set. However, although the number of samples is quite high, the proportion between the number of strains (194) and that of the number of species (36) is close to five. But many species present in the calibration set is represented by only 2 different strains. Therefore, the use of cross validation is justified. The cascade structure of all models is complex and the calibration set is constituted of biological and technical replicates. These two features must be taken into account in the implementation of the cross validation.

The partial cross validation by culture was chosen and scripted such that in every cross validation iteration, all replicate spectra associated to the same culture were removed, then used as test in the validation phase. By applying this partial cross validation, it was able to test all cultures of the calibration set. The partial cross validation by culture allows estimating

(partially) the intra-species, intra-strains, and intra-cultures co-variances. Further, concerning the species represented by only 2 strains, this cross validation algorithm was more stable less over-fitted than partial cross validation by strains.

2.4 Percentage of Good Prediction (PGP) and McNemar test

The accuracy of each used algorithm was computed during the three following steps: the cross validation step to explore the algorithm's parameters, the computing model step to build discrimination models with optimized parameters, and the validation step to evaluate the final expected accuracy.

The statistical indices here called the Percentage of Good Prediction (PGP) were calculated at the end of these three steps. This index was calculated by dividing the number of well identified spectra by the total number of spectra to predict. They allowed estimating the accuracy of the discrimination models at each step. It is possible to calculate for these three steps the PGP for each model independently. However, during the validation step and for better presentation, the PGP was calculated only by taxonomic rank.

McNemar's test is a statistical procedure that allows estimating if the prediction powers of two methods are significantly different. This test is based on a χ^2 with one degree of freedom because the sample number of each model is always higher than twenty (α: type I error). The χ^2 critical value with a 95% level of significance, written $\chi^2_{(1,0.95)}$ is equal to 3.8414.

The McNemar's test was chosen because a unique training set was used for each model and each algorithm. In this condition, the McNemar's test allows low probability of Type I error and presents a powerful ability to differentiate between two algorithms. [21].

2.5 Linear and non-linear chemometrics methods

The methodological rules of these two categories are entirely different and the data are not visualized in the same way. For linear methods, the variance of the explanatory variables is considered as linear and a proportionality relationship between them and the variables to explain is assumed. Non-linear methods take into account two types of variances, the global variance of the explanatory variables and variance of variables to explain, and then try to correlate these by means of a non-linear function such as the polynomial Kernel function for SVM algorithm. Also, for these two categories of algorithms, chemometrics models were not built around the same statistical rules. For supervised discrimination studies, the variety of chemometrics methods available is quite diverse. The linear methods are generally the most used with spectroscopic data. Indeed, the linearity relationship put forward by the Beer-Lambert expression, linking concentration and absorption, implies that the linear approach appears better [22]. However, the evolution of non-linear methods has allowed the elaboration of effective approaches such as SVM or Neural Network, which have been successfully applied to numerous experimental cases, including complex biological spectral data [23]. In order to optimize data mining and improve the understanding of biological phenomena from spectral results, it becomes essential to evaluate both linear and non-linear methods. Many of these linear and non-linear algorithms were declined in various specific algorithms, e.g., for the PLS algorithm, it was declined in robust or double PLS, quadratic PLS, splines function PLS or GIFI-PLS and many algorithms were combined such as the neural networks PLS or the least squares SVM [24]. For this study, only the “classical” (not “declined”) algorithms were used in order to assess the fundamental computing methodology of each of the following described algorithms.

2.5.1 Linear algorithms

LDA is a linear method of supervised discrimination that can improve the spreading of the sample distribution. The aims of this method are to maximize the ratio of the inter- to intra-class distances and to find a linear transformation allowing to achieve the maximum class discrimination. However, for the classical LDA the scatter matrices must be non-singular, which is well-known as the under sampling problem. To get round this problem many solutions exist. One of them is to precede LDA by a Principal Component Analysis (PCA) in order to extract the discriminant information. For this study PCA-LDA was tested although this algorithm may lead to a loss of discriminant information during the PCA step [25].

FDA aims at finding the subspace of the original variable space that best separates clusters by maximizing the inter-class variance with regard to the total variance [26]. This descriptive analysis builds a discriminant model to determine which cluster a new sample belongs to. This is simply done by projecting this sample onto the eigenvectors space and by selecting the nearest cluster. Several distances can be used for this decision, the Euclidean distance was preferred.

Wold and Sjöström were the first to describe the SIMCA chemometrics method [13]. It is a supervised classification method which considers every “cluster of samples” or “groups” separately. This method is very useful for classifying high-dimensional observations because it incorporates PCA for dimension reduction. So for every cluster, decomposition into principal components (PC) is carried out providing a matrix of scores and loadings for each. The most practical interest of this analysis is that each cluster can be reduced to a set of PCs [27] and during the calibration step, the optimal PCs are determined by means of their explained variance. After PCA steps, the discrimination models are built using Euclidean distance between clusters and PCA subspaces, taking into account the information and properties of clusters.

PLS-DA is a supervised classification method based on the multivariate PLS regression algorithm [28]. This algorithm allows to mathematically maximize the variance-covariance between the explanatory variable matrix and the property variable matrix. PLS-DA applies the multivariate PLS algorithm to establish discrimination rules by means of a binary matrix. The validation samples were attributed by means of the predicted binary code. The highest predicted binary code variable gives the predicted cluster [29].

2.5.2 Non-linear chemometrics methods

QDA is non-linear algorithm because it is based on a quadratic function but it is not very much different from LDA except that it is assumed that the covariance matrix can be different for each cluster, where it is estimated separately as a Gaussian distribution. The Gaussian parameters for each cluster are computed from training points with maximum likelihood estimation [30]. For this study, this method was applied on the PCA scores of the data matrix. KNN techniques were developed to answer challenges about density estimation and pattern classification [31]. Processing of this algorithm consists of basically ordering the training samples in a d-dimensional unit hypercube by means of a metrics distance measure. Then, for each tested sample, the training matrix is examined in the order of their projected distance from the tested sample on the sorted coordinate. The prediction of the unknown sample is determined by the most representative cluster of the k nearest neighbors [32]. To optimize the training model, the k integer and the metrics of distance can be adjusted.

Neural networks were successfully used to solve complicated pattern recognition and classification problems in different domains. The probabilistic neural networks (PNN) method presents a few advantages over the conventional neural network [33]. It provides a robust classification with noisy data. PNN combines different concepts: neural computing, Bayes

classification rule, and non parametric estimation of the probability density function. In this study, the PNN method was employed on the eigenvalues of the data matrix, after PCA preprocessing, and the Mahalanobis method was used for distance computing.

SVM is a supervised method originally proposed by Vapnik et al. in 1963. Fifty years later, many publications reporting on SVM and its extensions as a multiclass classification method can be found in literature [34]. The SVM algorithm classifies data by finding the best hyperplane that separates all data points of one class from the others classes. The best hyperplane for an SVM corresponds to the one with the largest margin between the two classes. In this study, the nu-SVM algorithm was employed and this algorithm could be used with many Kernel functions (table 1). When the SVM was used with a linear Kernel function, this algorithm was considered as a linear algorithm, and for this study, the results of SVM with a linear Kernel function, also called linear SVM, were associated with the results obtained by other linear methods.

2.6 Computing

All the chemometrics analyses were performed with Matlab R2013a (32-bit) (Mathwork, USA) and was used to classify the samples using their explanatory variables. The algorithms used for LDA, QDA and KNN were available in the pure Matlab. The algorithms used for SIMCA were developed by Cleiton A. Nunes, Brazil (available on Mathwork/matlabcentral). The algorithms used for SVM called lib-SVM were developed by Chih-Chung Chang and Chih-Jen Lin, China [34]. The algorithms used for FDA, PLS-DA and PNN were developed by Dominique Bertrand and Christophe Cordella, INRA, France [35].

The computing was realized on a personal computer with 2Go RAM, an Intel Core2 Duo 2.66GHz as processor and Microsoft Vista (32 bit). The required time for models computing

step and validation step was negligible (only few decades of seconds) compared to that of the optimization step. The number of parameters to optimize was the factor which had the most influence on the total required time. For example, the PLS-DA algorithm required to optimize only one parameter and took twelve hours to explore this parameter from 1 to 35, corresponding to 1 845 900 computing models (20 models x 35 range parameter x 879 total cross validated strains x 3 cultures per strain). On the other hand, for polynomial SVM, four days were needed to optimize four parameters (113 127 300 computing models).

3 Results and discussion

For all linear and non-linear methods described previously, the results obtained took into account the three following steps: optimization of the chemometric parameters, model computing, and validation in cascade. The details of the models are presented in table 2 and each model was validated taking into account the taxonomic tree as described in figure 2. Twenty models were built to complete the cascade and one cascade was built for each tested algorithm (a total of 220 optimized models). During the optimization and computing steps, the models were built independently by means of the calibration set (or part of the calibration set), but during the validation step, the models were tested by the validation set and were interlocked into each other.

3.1 Optimization of chemometric parameters and model computing

The various methods of discrimination compared in this work required an optimization of parameters. These parameters were different from each other and were directly associated to

the chemometrics method used (table 1). They naturally have a strong influence on the final results and it was thus essential to optimize these parameters in the most rigorous way.

The parameters were optimized by means of the calibration set only and for each algorithm, the influence of the variability of the various parameters or combination of parameters (e. g., SVM with polynomial Kernel function) was explored culture wise by partial cross validation. In the scope of this study, it is not possible to describe all the chemometric parameters for all algorithms and all models. Thus, we have therefore chosen to illustrate with the PLS-DA algorithm.

Concerning this algorithm, the parameter to optimize was the used number of latent variables (LV). It corresponds to the number of computed regression vectors. In this study, the LV parameter varied from 1 to 35. The limit of 35 was chosen principally for computing reasons. These parameters were optimized by partial cross validation and each LV was tested with each culture. The average of Percentage of Good Prediction (PGP) as a function of the LV was computed and plotted highlighting a maximum of PGP (figure 3). The LV corresponding to this maximum was taken as the best parameter. In fact, the LV optimization was important in order to minimize model under- and over-fitting [9]. All of the LV optimal number were defined for each model of the cascade and each model was built with its appropriate LV as presented in table 3 (see the LV column). It was possible to observe that the LV parameters highlighted the complexity of the model because the more the model took into account a high number of clusters, the higher was the LV parameter.

After parameter optimization, the models were computed taking into account these optimized parameters. During the model computing step, the discrimination abilities of each model and each algorithm was evaluated by means of the PGP of calibration. The calibration results concerning PLS-DA and linear SVM algorithms are presented in table 3.

3.2 Validation step

The prediction capacity of all the classification models was evaluated by means of the sample validation set. This step allowed observing, in real conditions, the behavior of the various models tested in this investigation. Figures 4.1 and 4.2 show the broken curves corresponding to the PGP of validation spectra of each tested algorithm versus the taxonomic rank. Concerning the linear methods, the best results were obtained with the PLS-DA method (Figure 4.1). This method allowed reaching a PGP of 98.9% for the genus taxonomic rank and 93.2% for the species taxonomic rank. The LDA and FDA methods respectively gave a PGP around 3% and 6% less than the PLS-DA method, with 96.4% and 95% for the genus taxonomic rank and 89.6% and 85.8% for the species taxonomic rank. The broken curve of SIMCA is not shown so as to preserve the best scale for PGP. This method showed the worst results with PGP of 66.5% for the genus taxonomic rank and less than 50% for the species taxonomic rank. The linear SVM gave very good results with a PGP of 99.8% for the genus taxonomic rank and 91.3% for the species taxonomic rank. This algorithm was able to reach the second best PGP for the last taxonomic rank and it appears also adapted to the identification of fungi described in this study.

The PLS-DA and linear SVM algorithm were equivalent and showed a superior ability for correct identification than the other linear methods. This is because LDA and FDA methods use a mathematical algorithm based mainly on variance of the spectral data matrix, whereas the PLS-DA and SVM algorithms are based on combination of the spectral and the reference matrices. Concerning LDA and FDA algorithms, the validation results showed a relative similarity for both methods. This particularity could be explained by the methodology employed, which is based on minimization of the Euclidian distance between validation samples and the cluster barycenters. On the opposite, the SIMCA method uses the internal

variance of each cluster separately. The pertinent variance searched for each model becomes finer when the modeling is near to the taxonomic species rank. When the SIMCA method estimates the internal variance of each cluster, the variance of interest is probably occulted by other internal cluster variances.

In figure 4.2, the 3 non-linear SVM (RBF, sigmoid, and polynomial) showed PGP values near to 100% only down to the family taxonomic rank. For the following ranks, these PGP decreased strongly, at the genus taxonomic rank (92%, 82% and 43%) and at species taxonomic rank (42%, 51% and 25%). Concerning the other non-linear algorithms, the best result was obtained with the KNN algorithm and gave a PGP of 90.4% and 78.2% respectively for genus and species taxonomic ranks. The PGP of this algorithm was close to 100% down to the family taxonomic rank and was about 15% less than the PLS-DA algorithm from the genus to the species rank. The second best non-linear algorithm was the QDA algorithm. This algorithm gave PGP values close to the KNN algorithm, nearly 5% less, with a PGP of 71.5% for the species taxonomic rank. Finally, the PNN algorithm gave the worst results, comparable to SVM with the polynomial Kernel function, with PGP values around 50% for the class taxonomic rank, which then decreased substantially down to the species taxonomic rank.

Non linear SVM did not perform very well probably because the variance of interest was not adapted to these parameters. The optimization step could also induce an over-fitting in the model. In addition, the interlocked cascade models could most certainly exacerbate this effect. Concerning the other non-linear algorithms, a plausible explanation is that the variance of interest could not be efficiently extracted by these non-linear algorithms probably due to the linearity rules associating the spectral data matrix with the taxonomy.

In order to evaluate if the prediction power of two methods was significantly different, the McNemar's test was applied to the investigated methods pairwise. The results are displayed in table 4 in the form of a two-dimension correlation matrix. The tests were computed for the species taxonomic rank. This test showed that all these algorithms were significantly different except for LDA versus linear SVM and PNN versus SVM (polynomial Kernel function). For the linear and non-linear methods all the curves presented in figure 4.1 and 4.2 showed a decreasing tendency. This decrease could be correlated with the variance sought at each taxonomic rank. We noticed that the change from the genus to the subgenus and from the section to the taxonomic species rank induced complications. These may be associated firstly, to the morphological proximity at the subgenus rank and to the closeness of the biochemical structures at the species rank. These observations seemed to converge with difficulties encountered during morphological identification, in particular for *Aspergillus* 2, *Camenberti*, *Chrysogena* and *Roquefortorum* species models [36, 37, 38].

3.3 Combined cascade

Table 3 underlines some problematical models for both PLS-DA and linear SVM algorithms. The concerned models were principally *Aspergillus* 2, *Camemberti*, *Chrysogena*, and particularly *Roquefortorum*, presenting respectively validation PGP values of 85.7, 83.3, 86.5 and 61% for PLS-DA and 0, 37.5, 75 and 59% for SVM. These 4 models gave, during the calibration step, a PGP of 100% (or close to 100%) but their validation was low and illustrated a lack of accuracy. These inconveniencies could be explained for *Aspergillus* 2, *Chrysogena* and *Camemberti* models, by the low number of strains and by the high number of needed taxonomic nodes. For the *Roquefortorum* model, the taxonomy of strains concerned by this model is still in evolution [36] and the difficulties to discriminate by sequencing and

the genetic proximity of these concerned strains were real and correlated with the outcome of our chemometrics models.

On other hand, a complementarity between PLS-DA and linear SVM algorithms was suggested through the data presented in figure 4.1. At the subgenus rank, these two algorithms gave similar accuracy with respectively PGP of 98.9 and 99.0%. At the genus rank, the PGP of these two algorithms were 98.9 and 99.9% respectively, and at the species rank, 93.2 and 90.1% respectively. The remarkable intersection point between these two broken curves is pointed by a circle in figure 4.1. Due to this singularity between these two algorithms, a "combined cascade" model was built. The linear SVM algorithm was used to elaborate the first 8 models from the subphylum to the genus taxonomic ranks while the PLS-DA algorithm was used to build 12 models from the sub-genus to species taxonomic ranks. The validation results of this combined cascade are presented in figure 5.

The "combined cascade" model showed the best performances compared to all the previous "regular cascades", with 94.2% of PGP at species taxonomic rank. The spectra that were wrongly predicted by PLS-DA were then correctly predicted by linear SVM at genus taxonomic rank, and were almost all correctly identified until the species taxonomic rank. The gain of one percent (illustrated by **ε** in figure 5), due to the combined cascade, was maintained from the genus down to the species taxonomic ranks. The "combined cascade" model revealed that linear SVM appeared to be the most pertinent algorithm to discriminate fungi strains until the genus taxonomic rank. This suggested that linear SVM could be better adapted than PLS-DA algorithm for voluminous sample sets. On the other hand, PLS-DA appeared as the most pertinent method to identify fungi at the species taxonomic rank, suggesting that PLS-DA could be adapted for reduced and clustered sample sets. Thus, the combination of both methods indicated an improvement of the identification capacity.

4 Conclusion

This is the first study that compares eleven linear and non linear supervised classification algorithms on such a large dataset of food-related fungi FTIR spectra. The results obtained highlight the suitability of the linear classification methods, in particular the PLS-DA and linear SVM algorithms, for discriminating and identifying from the family to the species taxonomic ranks. These findings are promising but also pointed out the dependence due to the taxonomic references and consequently the limits of the supervised cascade computing for the application to spectral data. These observations seem to corroborate with difficulties associated with the morphological and biochemical identification. The "combined cascade" modeling including the two well suited models, PLS-DA and linear SVM, gave an improvement of the identification accuracy from the subphylum to the species taxonomic ranks. This study also highlights the interest of the concept of cascade modeling based on taxonomy because of the size and nature of the data set. Indeed, extending a complex discrimination problem into several steps allowed to distribute the studied variance on several models and thus to target the adequate variance on every taxonomic node. Further, the supervised cascade model amplifies the discrimination capacity of each tested algorithm by means of the interlocked models. The McNemar's results pinpoint that the choice of the supervised cascade to develop chemometrics discrimination methods was appropriate to assess many discrimination algorithms. To perform knowledge about abilities of these supervised discrimination algorithms, linear and non-linear algorithms need to be assessed using other types of data, such as bio-morphological data or using other study cases. Also, the fungi spectral data bank could be used in a non-supervised way to define new clustering or new cascade of classification. In addition, by means of PLS-DA regression vector, rand

feature or ANOVA (ANalysis Of Variance) algorithms, it would very interesting to study spectroscopic markers for each model in order to link spectral, biological, and chemical properties of fungi.

6 Acknowledgements

The technological platform PICT-IBiSA “Imagerie Cellulaire et Tissulaire” and the “Pôle de compétitivité” Valorial, La Région Bretagne, La Région Champagne-Ardenne are gratefully acknowledged. Financial supports under the MOLDID project, project "Mycotech" of the European Union, the Région Bretagne, and the Conseil Général du Finistère are also gratefully acknowledged. The authors thank Marie-Anne Le Bras and Valérie Vasseur for their expertise and help in fungal identification, Amélie Weill and Olivia Le Bourhis for their excellent technical assistance, Cyril Gobinet for his reading of the chemometrics section, and Dominic Bertrand for his pertinent and constructive review of the manuscript.

Table 1: Optimized parameters used for the different chemometrics methods

Chemometrics methods	Used parameters	Chemometrics methods	Used parameters
Linear methods		*Non-linear methods*	
LDA	**Kdim** (positive integer included in 1 to 35): size of eigenvalues matrix	**QDA**	**Kdim** (positive integer included in 1 to 35): size of eigenvalues matrix
FDA	**maxscore** (integer included in 1 to 35): size of PCA-score matrix allowed to the model (PCA step) **Kdim** (positive integer included in 1 to 35): size of eigenvalues matrix	**KNN**	**NumNeighbors** (positive integer included in 1 to 30) : specifying the number of nearest neighbors in calibration data to find for classifying each point when predicting **Metric choice:** function use to specify the distance metric between neighbors (among 11 distances metric process)
SIMCA	**maxscores** (positive integer included in 1 to 35): size of PCA-score matrix allowed to each clusters (PCA step)	**PNN**	**FN** (positive integer included in 1 to 35): the number of computed iterations **σ2** (positive real included in 0 to + ∞): "smoothing parameter" of the probability function estimator
PLS-DA	**LV** (positive integer included in 1 to 35) is the numbers of computed regression vector defined by the Latent Variables	**SVM**	**Kernel function choice** (among 3 K-functions: **RBF**, **Sigmoïd** and **polynomial**) **ν** (positive real included in 0 to 1): "level of detail" or hyperplan resolution **γ** (positive real included in 0 to + ∞): selected value of γ in Kernel function (RBF, sigmoid and polynomial choice) **coef0** (positive real included in 0 to + ∞): selected value of coef0 in Kernel function (RBF and sigmoid choice) **d** (positive integer included in 1 to 5): selected degree in kernel function (polynomial choice)
SVM	**Linear Kernel function choice** **ν** (positive real included in 0 to 1): "level of detail" or hyperplan resolution		

Table 2: Details of the 20 discrimination models tested.

Taxonomic rank	Model name	Corresponding clusters names	Number of spectra (number of strains)	
			Optimization & calibration	External validation
Subphylum	*Micromycetes*	*Pezizomycotina* *Mucoromycotina*	4159 (194)	1801 (83)
Class	*Pezizomycotina*	*Eurotiomycetes* *Sordariomycetes* *Saccharomycetes* *Dothideomycetes*	3302 (154)	1430 (66)
Order	*Dothideomycetes*	*Dothideales* *Pleosporales*	144 (7)	39 (2)
	Sordariomycetes	*Xylariales* *Hypocreales*	1087 (51)	433 (19)
Family	*Mucorales*	*Mucoraceae* *Lichtheimiaceae*	857 (42)	371 (17)
	Hypocreales	*Cordycipitaceae* *Nectriaceae*	1046 (45)	412 (18)
Genus	*Mucoraceae*	*Rhizopus* *Mucor* *Actinomucor*	822 (40)	355 (16)
	Trichocomaceae	*Paecilomyces* *Penicillium 1* *Aspergillus 1*	1948 (91)	886 (42)
Subgenus	*Penicillium 1*	*Penicillium 2* *Aspergilloides*	1043 (49)	441 (22)
Section	*Aspergillus 1*	*Flavi* *Fumigati* *Nidulantes* *Nigri* *Aspergillus 2*	851 (39)	427 (19)
	Penicillium 2	*Brevicompacta* *Fasciculata* *Chrysogena* *Roquefortorum* *Penicillium 3*	808 (37)	337 (17)
Serial	*Fasciculata*	*Camenberti* *Verrucosa*	190 (9)	45 (2)
Species	*Nidulantes*	*E. nidulans* *A. versicolor*	147 (7)	57 (3)
	Aspergillus 2	*E. chevalieri* *E. amstelodami*	110 (5)	21 (1)
	Fusarium	*F. equiseti* *F. graminearum* *F. verticillioides* *F. oxysporum*	1001 (43)	392 (17)
	Mucor	*M. circinelloides* *M. racemosus* *M. spinosus*	647 (31)	320 (13)
	Camenberti	*P. biforme* *P. camenberti*	108 (5)	24 (1)
	Chrysogena	*P. nalgiovense* *P. chrysogenum*	141 (7)	84 (4)
	Roquefortorum	*P. roqueforti* *P. carneum* *P. paneum*	250 (12)	119 (6)
	Aspergilloides	*P. corylophilum* *P. glabrum* *P. oxalicum*	235 (11)	104 (5)

Table 3: Comparison of calibration and validation PGPs between PLS-DA and linear SVM models

Taxonomic rank	Model name	PLS-DA models			linear SVM models		
		LV	PGP Calibration	PGP Validation	γ	PGP Calibration	PGP Validation
Subphylum	*Micromycetes*	7	100	100	0,001	100	100
Class	*Pezizomycotina*	12	99,9	99,2	0,01	99,8	100
Order	*Dothideomycetes*	5	100	100	0,01	100	100
	Sordariomycetes	10	100	99,6	0,005	99,3	100
Family	*Mucorales*	16	100	100	0,005	98,4	100
	Hypocreales	5	100	99,5	0,001	100	100
Genus	*Mucoraceae*	20	100	100	0,005	99,1	100
	Trichocomaceae	25	99,9	98,8	0,05	95,7	99,9
Subgenus	*Penicillium 1*	13	100	99,6	0,2	98,7	95,7
Section	*Aspergillus 1*	20	100	97,7	0,05	99,2	99,5
	Penicillium 2	15	100	95,4	0,2	94,7	84,7
Serial	*Fasciculata*	5	100	100	0,2	100	100
Species	*Nidulantes*	6	100	100	0,01	95,8	100
	Aspergillus 2	5	100	85,7	0,8	100	0,0
	Fusarium	20	99,1	95,5	0,2	96,1	94,5
	Mucor	14	100	99,0	0,2	96,6	95,3
	Camenberti	10	100	83,3	0,1	93,3	37,5
	Chrysogena	5	100	86,5	0,1	95,0	75,0
	Roquefortorum	20	100	66,0	0,01	100	59,0
	Aspergilloides	5	100	100	0,01	100	100

Table 4: Correlation matrix of McNemar's test, presenting McNemar's value for each pair of chemometrics methods at the species level

Misclassified sample's number (species level)	Tested chemometrics algorithms	*linear methods*					*non-linear methods*					
		LDA	FDA	SIMCA	PLS-DA	SVM Linear	QDA	PNN	KNN	SVM RBF	SVM Sigmoid	SVM Polynomial
188	LDA	0	10	703	13	0,2	150	845	71	591	454	882
255	FDA	10	0	581	45	13	87	715	29	475	349	751
1164	SIMCA	703	581	0	840	723	251	9,2	382	6,8	37	14
123	PLS-DA	13	45	840	0	10	239	987	139	722	577	1026
178	SVM Linear	0,2	13	723	10	0	162	865	80	610	472	903
514	QDA	150	87	251	239	162	0	351	16	178	99	378
1316	PNN	845	715	9,2	987	865	351	0	499	32	83	0,6
392	KNN	71	29	382	139	80	16	499	0	293	191	531
1041	SVM RBF	591	475	6,8	722	610	178	32	293	0	12	41
888	SVM Sigmoid	454	349	37	577	472	99	83	191	12	0	97
1356	SVM Polynomial	882	751	14	1026	903	378	0,6	531	41	97	0

Figure 1: Raw and preprocessed FT-IR spectra of an *Alternaria alternata* strain with a tentative band assignment of major macromolecules.

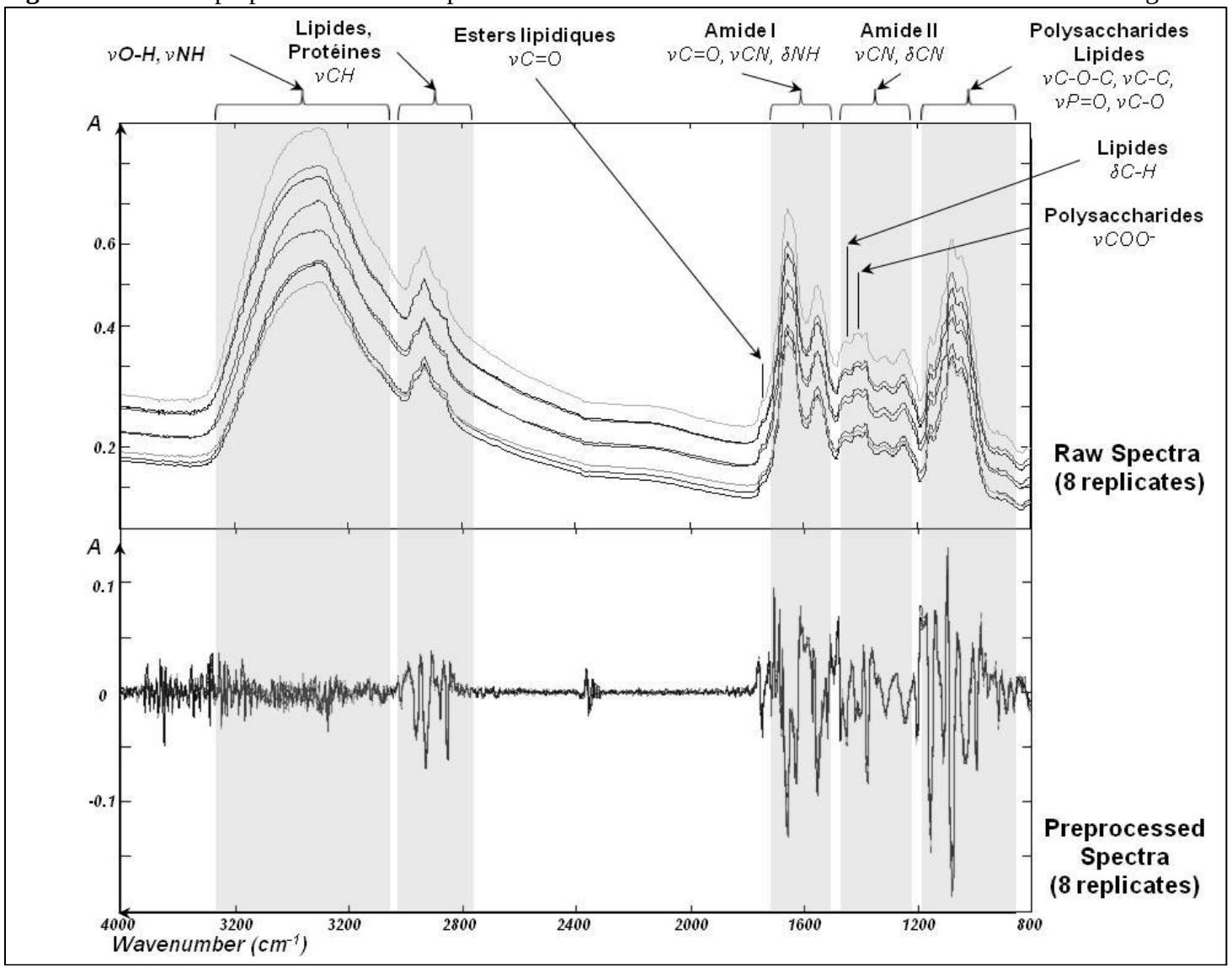


Figure 2: Organigram of the modeling cascade based on the current mold taxonomy

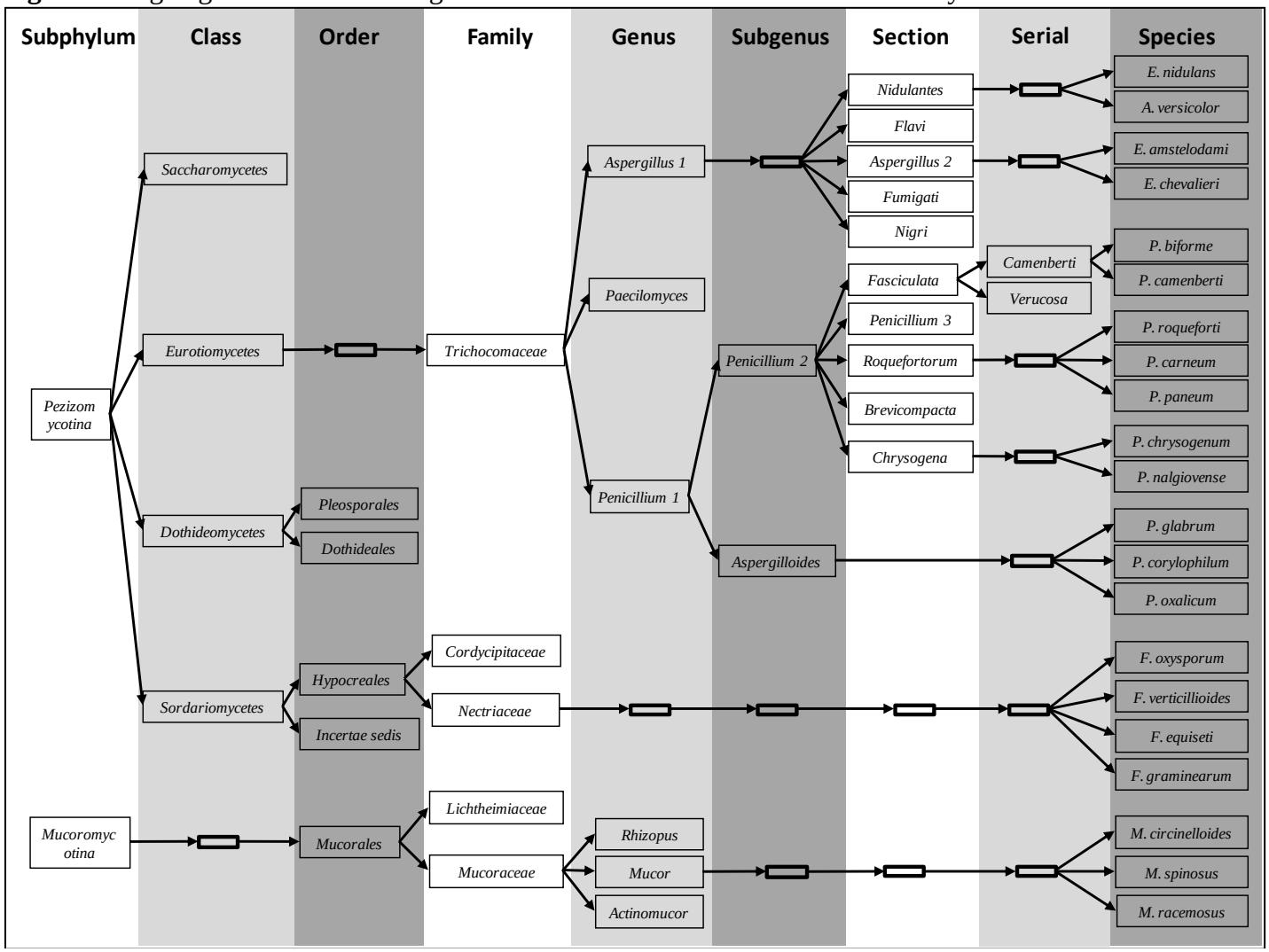


Figure 3: Number of Latent Variables (LV) optimization (PLS-DA methods) for the *Mucor* species model with 3 clusters (*M. racemosus*, *M. circinelloides*, *M. spinosus*)

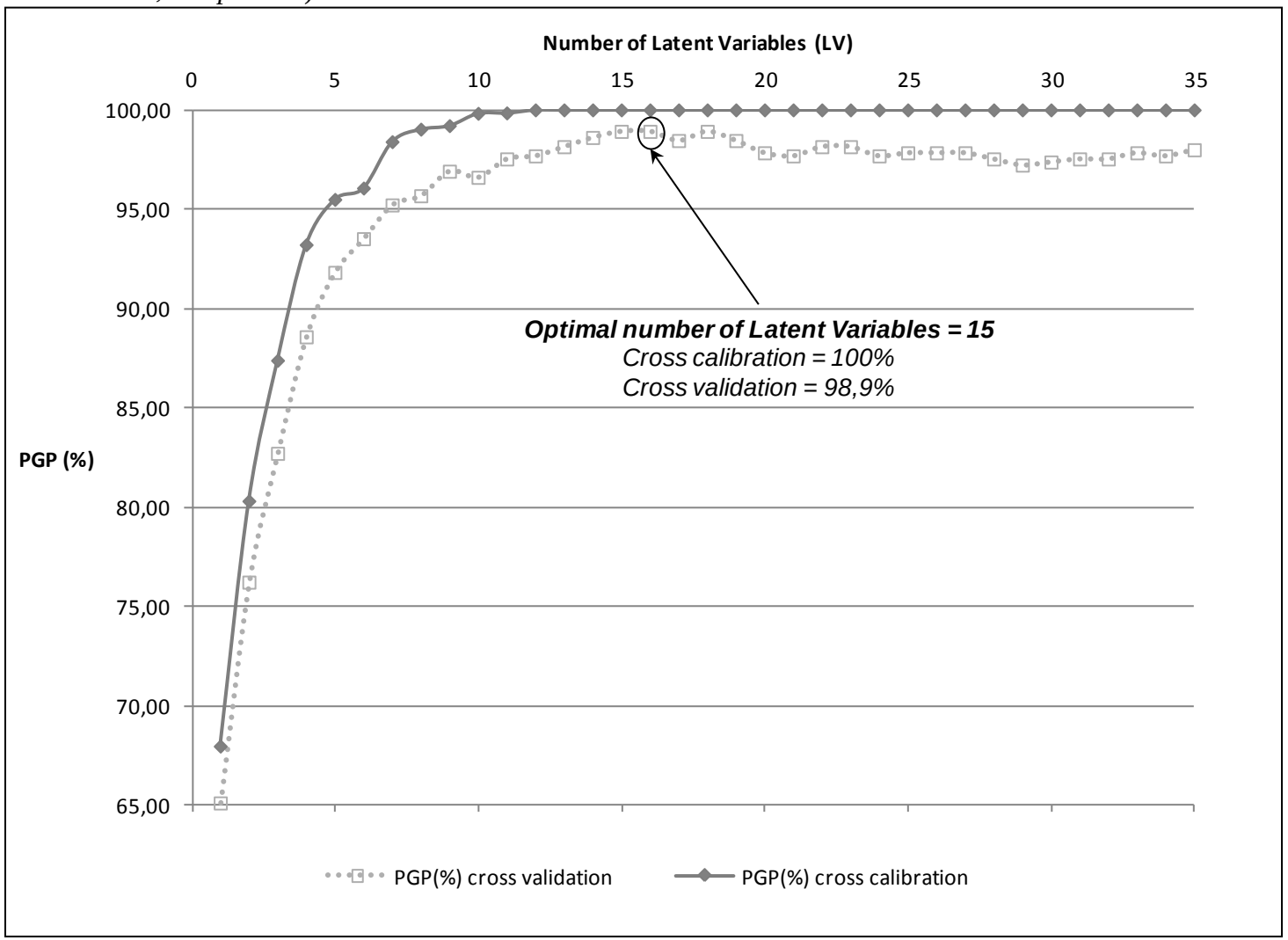
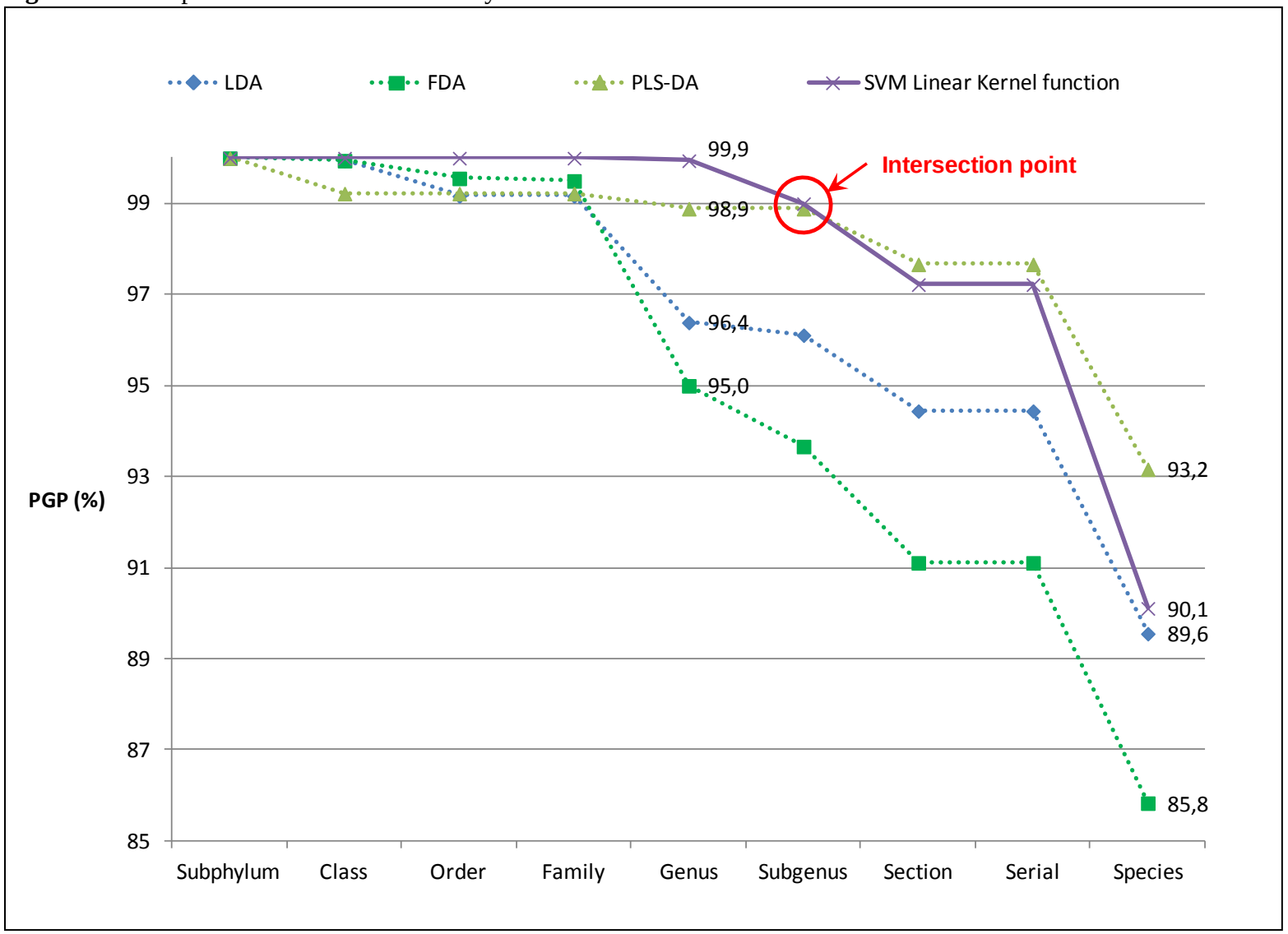


Figure 4.1: Comparison of validation PGPs by taxonomic rank between linear chemometrics methods



Reproduced in color on the Web and in black-and-white in print

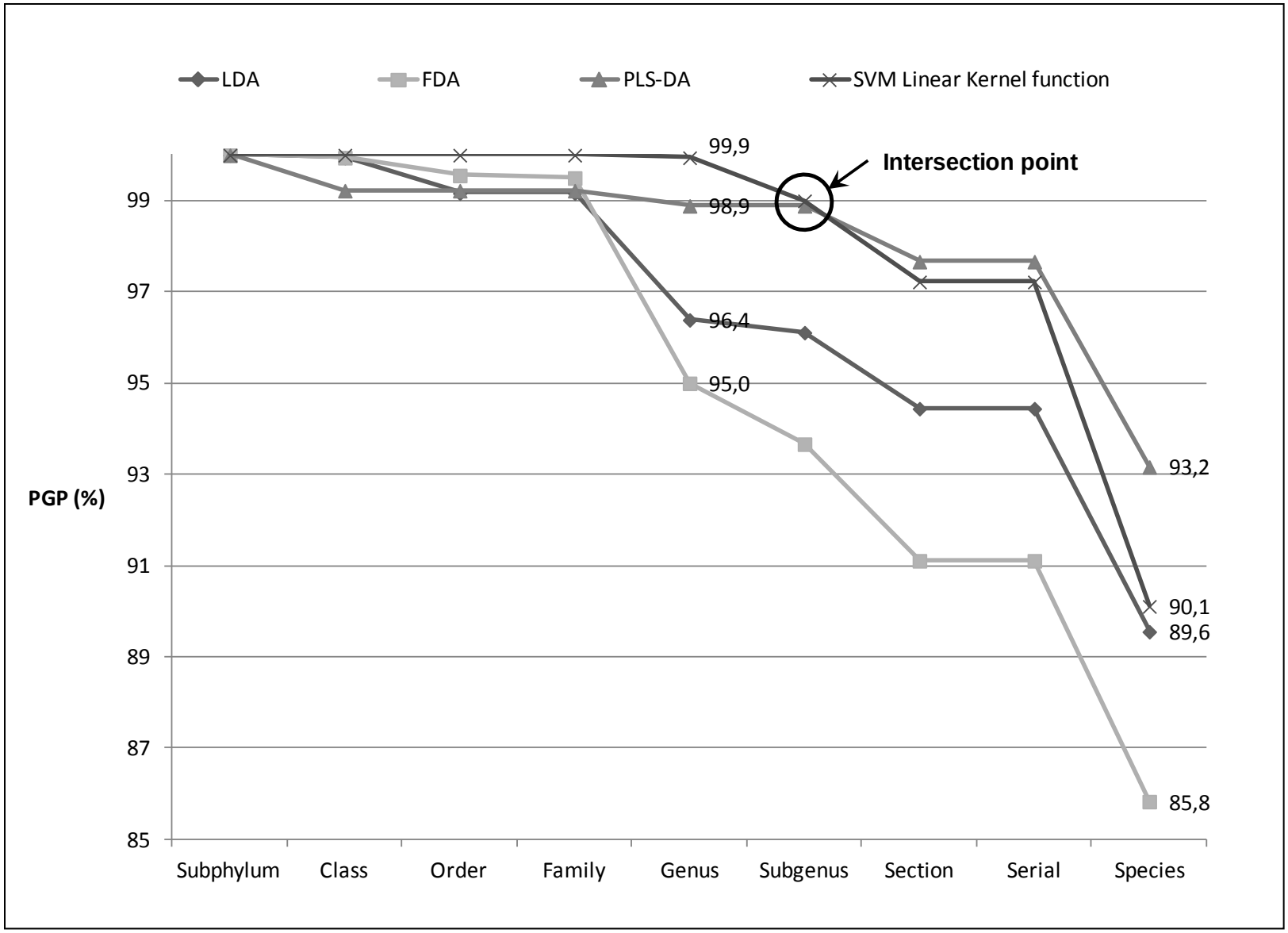
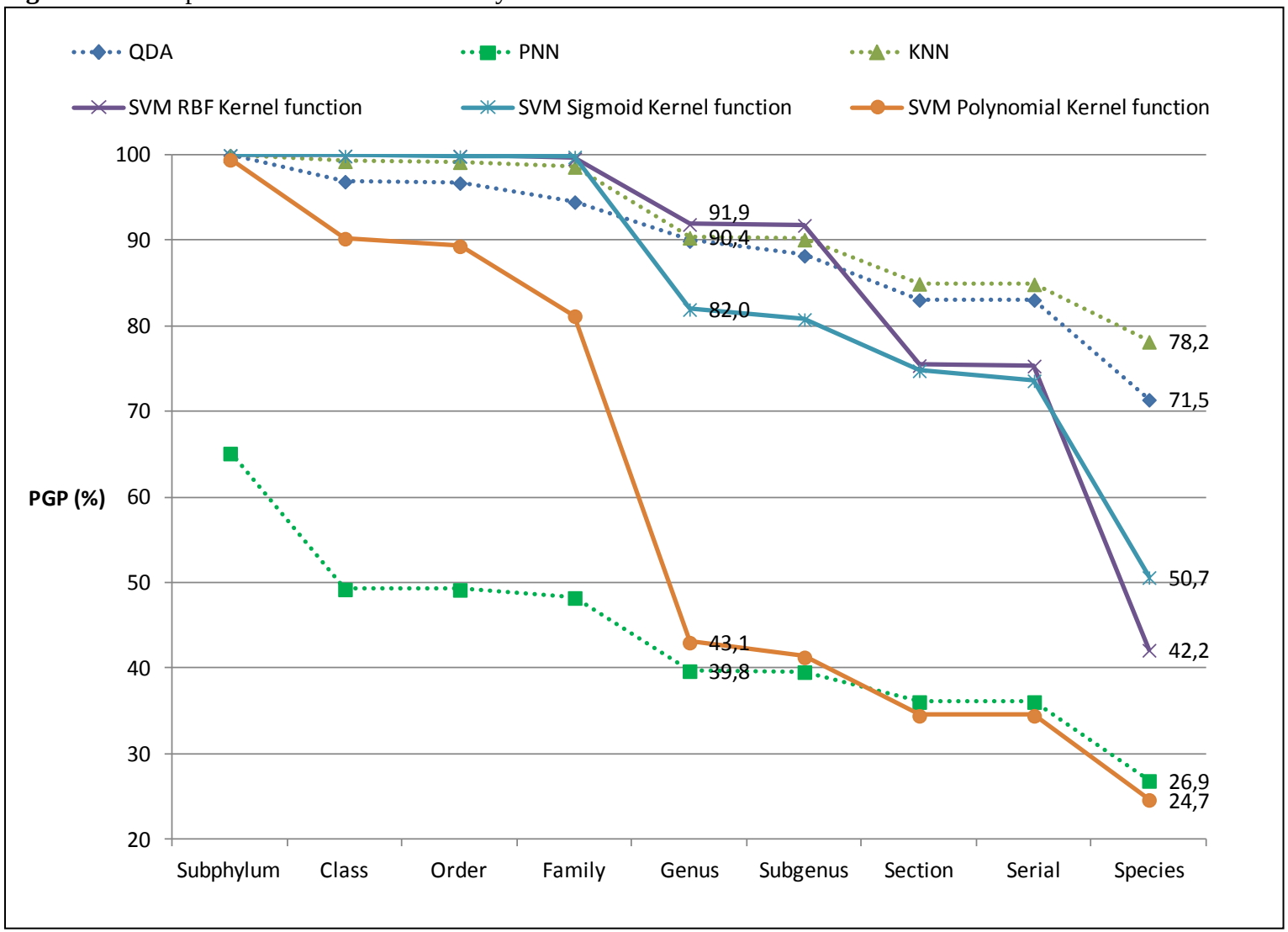


Figure 4.2: Comparison of validation PGPs by taxonomic rank between non-linear chemometrics methods



Reproduced in color on the Web and in black-and-white in print

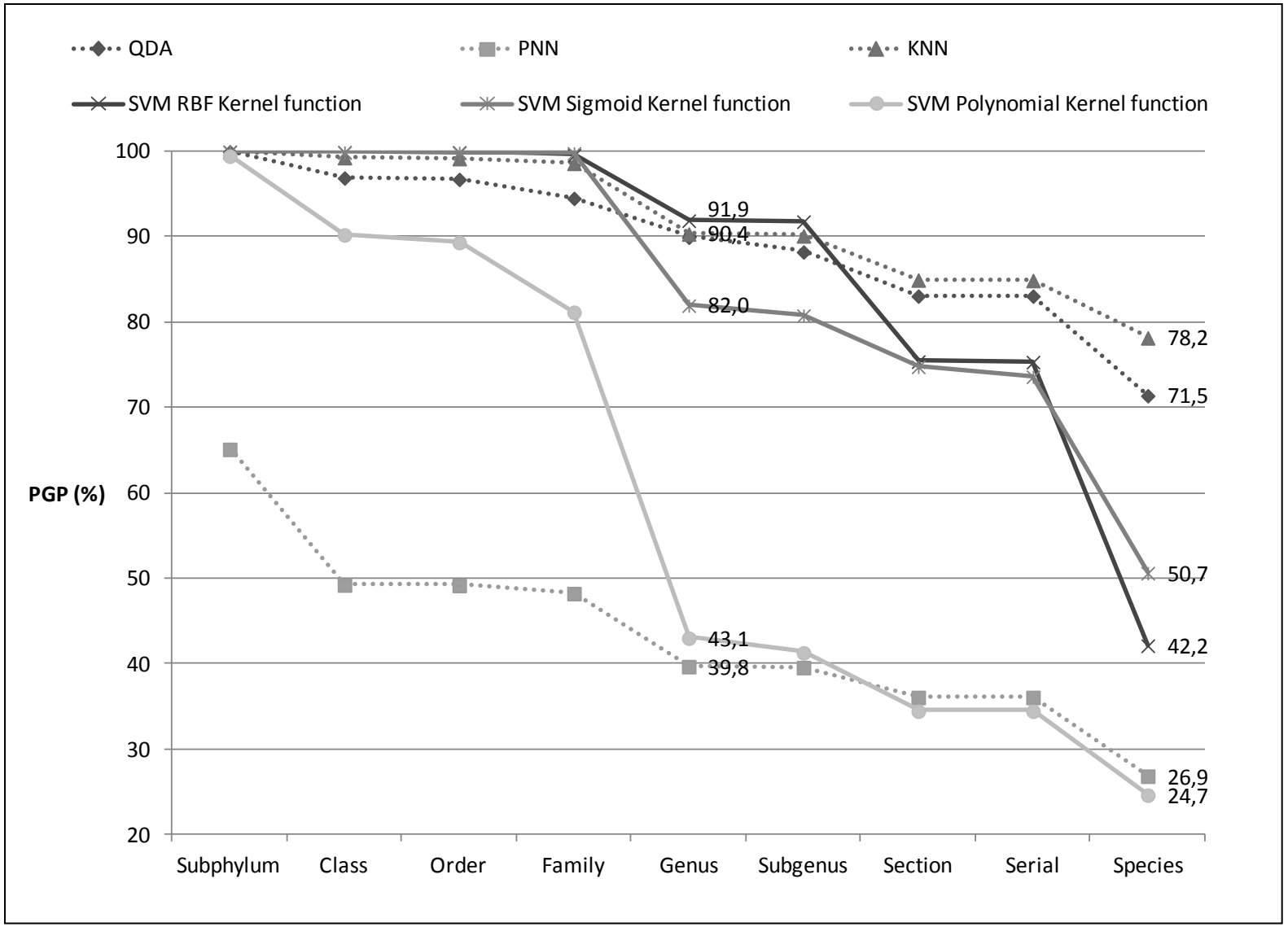
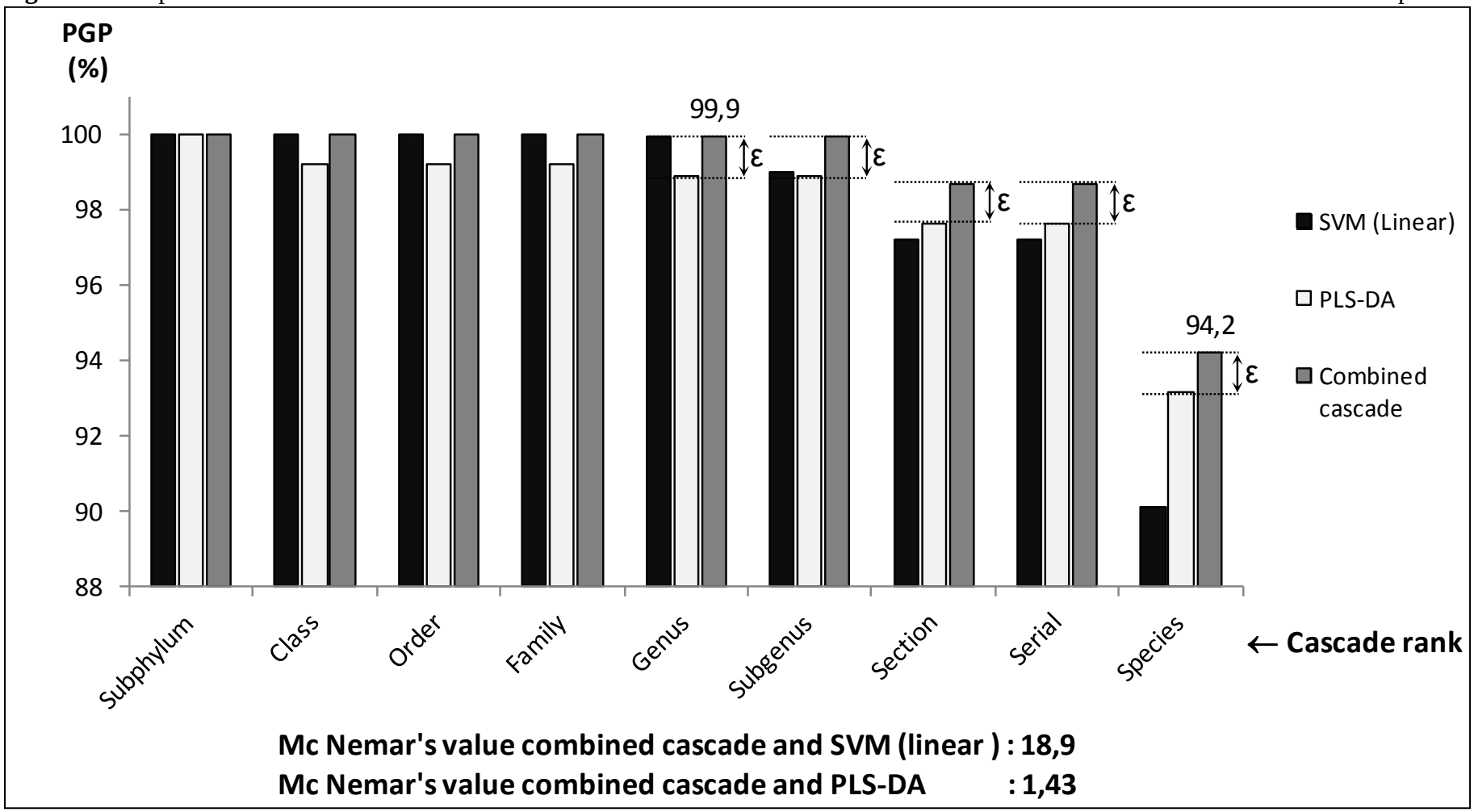


Figure 5: Comparison between the combined linear SVM and PLS-DA cascade with the linear SVM and PLS-DA cascades taken independently.



7 References

[1] J. Laane, Frontiers of Molecular Spectroscopy, 1st ed. Elsevier Science, Texas, USA, **2008**.

[2] A. Lecellier, J. Mounier, V. Gaydou, L. Castrec, G. Barbier, W. Ablain, M. Manfait, D. Toubas, G.D. Sockalingum, Differentiation and identification of filamentous fungi by high-throughput FTIR spectroscopic analysis of mycelia, J Food Microbiol., 32 (**2014**) 168-169.

[3] A. Lecellier, V. Gaydou, J. Mounier, L. Castrec, G. Barbier, W. Ablain, M. Manfait, D. Toubas, G.D. Sockalingum, Implementation of an FTIR spectral library of 486 filamentous fungi strains for rapid identification of molds, J Food Microbiol., DOI information: 10.1016/j.fm.2014.01.002. In press (**2014**).

[4] V. Shapaval, J. Schmitt, T. Moretro, H.P. Suso, I. Skaar, A.W. Asli, D. Lillehaug, A. Kohler, Characterization of food spoilage fungi by FTIR spectroscopy, J. Appl. Microbiol. 114 (**2013**) 788-796.

[*5*] M. Decker , P.V. Nielsen, H. Martens, Near-infrared spectra of Penicillium camemberti strains separated by extended multiplicative signal correction improved prediction of physical and chemical variations, Appl Spectrosc. 59 (**2005**) 56-68.

[6] J.D. Pallua, W. Recheis, R. Poeder and al., Morphological and Tissue Characterization of the Medicinal Fungus Hericium corraloides by a Structural and Molecular Imaging Platform. Analyst 137 **(2012)** 1584-1595.

[*7*] B. R. Kowalski, Chemometrics : view and proposition, J. Chem Inf. Comput. Sci. 15 (**1975**) 201-203.

[8] A. Höskuldsson, Prediction methods in science and technology, Basic Theory vol.1; Thor Publishing: Copenhagen, Denmark, **1996**, pp. 245.

[9] D. Bertrand, E. Dufour, Chimiométrie appliquée à la spectroscopie infrarouge, La spectroscopie infrarouge et ses applications analytiques, 2nd ed. Lavoisier : Paris, **2006**, pp. 309-401.

[10] R. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 7 (**1936**) 179-188.

[11] D. H. Moore, Combining linear and quadratic discriminants, Comput. Biomed. Res. 6 (**1973**) 422-429.

[12] Wold, S.; Martens, H.; Wold, H. The multivariate calibration problem in chemistry solved by the PLS method, In: Ruhe, A., Kastrom, B. (Eds.) Springer, Heidelberg (**1983)**, 286-293.

[13] S. Wold, M. Sjostrom, SIMCA: A method for analyzing chemical data in terms of similarity and analogy, Chemometrics, 52 (**1977**) 243.

[14] V. Vapnik, A. Lerner, Pattern recognition using generalized portraitmethod, Automat. Rem. Contr. 24 (**1963**) 774-780.

[15] J.H. Friedman, F. Baskett, L.J. Shustek, An Algorithm for Finding Nearest Neighbors, Trans. Comput. 24 (**1975**) 1000-1006.

[16] D. Specht, Probabilistic neural networks, Neural networks 3 (**1990**) 110-118.

[17] D. Bertrand, E. Dufour, Identification et caractérisation des microorganismes, La spectroscopie infrarouge et ses applications analytiques, 2nd ed. Lavoisier : Paris, **2006**, pp. 561-581.

[18] L.J. Tashman, Out-of-samples tests of forecasting accuracy : an analysis and review, Int. J. Forecasting 16 (**2000**) 437-450.

[19] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Statistics Surveys 4 (**2010**) 40-79.

[20] M. Stone, Cross-Validatory Choice and Assessment of Statistical Predictions, J R Stat Soc Series B Stat Methodol. 36 (**1974**) 111-147.

[21] T.G. Dietterich, Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, Department of Computer Science, Oregon State University, 10 (**1998**) 1895-1923.

[22] J. Workman, Review of Chemometrics Applied to Spectroscopy: Quantitative and Qualitative Analysis, The Handbook of Organic Compounds, NIR, IR, Raman, and UV-Vis Spectra Featuring Polymers and Surfactants 1 (**2001**) 301-326.

[23] F. Chauchard, R. Cogdill, S. Roussel, J.M. Roger, V. Bellon-Maurel, Application of LS-SVM to non-linear phenomena in NIR spectroscopy : Development of a robust and portable sensor for acidity prediction in grapes, Chemometr. Intell. Lab. 71 (**2004**) 141-150.

[24] M. Mörtsell, M. Gulliksson, An overview of some non-linear techniques in Chemometrics, Rapportserie FSCN - ISSN 1650-5387 2001:6, Mid-Sweden University, **2001**.

[25] D.L. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, Pattern Anal and Machine Intelligence, 18 (**1996**) 831-836.

[26] Romeder, J. M. Méthodes et Programmes d'Analyse Discriminante, Dunod, Paris, France, **1973**.

[27] Sharaf, M. A.; Illman, D. L.; Kowalski B.R. Chemometrics, Wiley, New York, 1986.

[28] P.H. Garthwaite, An interpretation of partial least squares, J. Amer. Statist. Assoc. 89 (**1994**) 122-127.

[29] M. Tenenhaus, L'algorithme de régression PLS1, In Tenenhaus M. (ed.), Paris, France **1998** pp. 75-77.

[30] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag, New York, **2001**.

[31] D.O. Loftsgaarden, C.P. Quesenberry, A Nonparametric Estimate of a Multivariate Density Function, Ann. Math Statist. 36 (**1965**) 1049-1051.

[32] L. Labart, A. Morineau, N. Tabart, Technique de la description statistique, Méthodes et logiciels pour l'analyse des grands tableaux, Ed. Dunod, Paris, France, **1987**.

[33] P. Wasserman, Advanced methods in neural networks, Van Nostrand Reinhold, New York, USA, **1993**.

[34] C.C. Chang, C.J. Lin, T. ACM, LIBSVM : A Library for Support Vector Machines, Int. Sys. Techn, 2(27) (**2011**) 1-27, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[35] http://www.chimiometrie.fr/saisir_conceptors.html.

[36] M. Boysen, P. Skouboe, J. Frisvad, L. Rossen, Reclassification of the Penicillium roqueforti group into three species on the basis of molecular genetic and biochemical profiles, Microbiology, 142 (**1996**) 541-9.

[37] F. Giraud, T. Giraud, G. Aguileta, E. Fournier, R. Samson, C. Cruaud, S. Lacoste, J. Ropars, A. Tellier, J. Dupont, Microsatellite loci to recognize species for the cheese starter and contaminating strains associated with cheese manufacturing, Int. J. Food Microbiol. 137 (**2010**) 204-13.

[38] V. Hubka, M. Kolarik, A. Kubatova, S.W. Peterson, Taxonomic revision of Eurotium and transfer of species to Aspergillus, Mycologia 105 (**2013**) 912-37.