Discrimination of *Salix caprea*, *Salix gracilistyla*, and Their Interspecific Hybrid Using Vegetative Characteristics and Partial Least Squares Discriminant Analysis

Han-Na Seo, Hyo-In Lim, Yong-Yul Kim, and Seung-Beom Chae

Forest Bioinformation Division, National Institute of Forest Science, Suwon, Korea 16631

Wonwoo Cho

Forest Tree Improvement Division, National Institute of Forest Science, Suwon, Korea 16631

Additional index words. accuracy, classification, DUS test, hybridization, variable influence on projection, VIP, variety

Abstract. Identifying the morphological characteristics that distinguish plant varieties is an important issue for plant breeders and researchers. The objective of the present study was to create a partial least squares discrimination analysis (PLS-DA) model with morphological characteristics for species discrimination and to select the characteristics most important for species discrimination. Data for 27 vegetative characteristics were obtained from *Salix caprea* and *Salix gracilistyla*, and their interspecific hybrid (*S. caprea* × *S. gracilistyla*), and used for PLS-DA. According to this analysis, seven of the 27 characteristics were identified as those that most influenced species discrimination, and the PLS-DA model with these seven characteristics had a classification accuracy of 86% to 100%. The classification performance of this model was not significantly different from that of the model with all 27 characteristics (full model). Therefore, these results indicated that the three species can be relatively well distinguished by the seven characteristics extracted by PLS-DA. In addition, the selected characteristics can be used to select cross-breeding parents in subsequent breeding programs and to test the distinction, uniformity, and stability (DUS test) of the hybrid variety. From this perspective, PLS-DA is thought to be a useful methodology for classifying new plant varieties and providing information for breeding.

According to the International Union for the Protection of New Varieties of Plants (UPOV), protection of new varieties can only be granted if the DUS test proves that their expression characteristics differ from that of any other variety (UPOV, 2002). Therefore, plant breeders and researchers are focused on finding morphological characteristics that can distinguish a new variety from other varieties and can explain the overall features of the variety well. This is mainly because these characteristics can be used to test the DUS of different breeds as well as to select cross-breeding parents in subsequent breeding programs and to preserve genetic resources (Korir et al., 2012). From a statistical point of view, the process of extracting characteristics that can distinguish a given variety from others belongs to a main topic dealt with in a

discrimination analysis rather than in a cluster analysis (Kuhn and Johnson, 2013).

Linear discriminant analysis (LDA) is the most commonly used method to find a linear combination of characteristics that can be used to discriminate two or more classes of varieties (Galdón et al., 2012). The resulting linear combination can be used as a classifier for the classification of varieties. In addition, because LDA is performed as a multiple linear regression model using characteristics as explanatory variables, it has the advantage of being able to compare the relative influence of each characteristic on the classification of the varieties (Bruce and Bruce, 2017). However, LDA has a drawback in that the accuracy of the model is decreased by the multicollinearity and dimensionality occurring when multiple correlated variables outnumber the number of observations used. As an alternative method, principal component analysis and linear discriminant analysis (PCA-LDA) has often been used; this analysis applies the LDA on principal components (latent variables) from the PCA rather than on the original variables (De Luca et al., 2012).

On the other hand, in the field of chemometrics and metabolomics research, PLS-DA has been widely used for discrimination,

classification, and authenticity identification of a target object (Fonville et al., 2010; Hur et al., 2015; Kwon et al., 2014; Yan et al., 2014). Recently, a classification research of cultures using PLS-DA is also being conducted in the plant field (Kong et al., 2013; Shrestha et al., 2016). PLS-DA is effective in selecting remarkable characters for solving classification problems (Ruiz-Perez et al., 2020). In particular, PLS-DA has an advantage in that it is free of multicollinearity and dimensionality problems (Barker and Rayens, 2003).

S. caprea and *S. gracilistyla* are deciduous broadleaf willow species native to Korea (Lee, 2003). *S. caprea* is a small tree growing in wetlands or lower parts of mountains, and it is known to be suitable for landscape restoration (Vaculík et al., 2012; Wu and Raven, 1999). *S. gracilistyla* is a shrub that grows in wetlands (or by the water) and mountain valleys, and it is known to invade the restored areas quickly after the restoration of wetlands (Cho et al., 2008; Choi and Kim, 2015) and to have flowering precocious characteristics (Wu and Raven, 1999). Recently, the National Institute of Forest Research has cross-bred *S. caprea* and *S. gracilistyla* to develop high biomass productivity varieties. A study using PCA to analyze 21 flower characteristics (12 for female flowers and nine for male) showed that *S. caprea*, *S. gracilistyla*, and their interspecific hybrid were distinguishable from each other (Seo et al., 2021).

The characteristics of vegetative organs are also very important for testing discrimination, uniformity, and stability (the DUS test) according to the International Union for the Protection of New Varieties of Plants (UPOV) Convention. For example, in the guidelines for conducting DUS tests for willow (*Salix* L.) developed by the UPOV, 20 of 23 characteristics are those of vegetative organs, such as leaves and branches (UPOV, 2006). The guidelines for goat willow (*S. caprea* L.) developed by the Korea Forest Service also presented 14 characteristics of vegetative organs (Korea NFSV, 2019). Nevertheless, to date, no studies have been conducted to discriminate and classify *S. caprea*, *S. gracilistyla*, and their interspecific hybrid (*S. caprea* × *S. gracilistyla*) using vegetative characteristics.

In the present study, a PLS-DA model was created to discriminate and classify the two willow species and their interspecific hybrid using 27 characteristics of vegetative organs. In addition, a set of characteristics that most influenced the discrimination and classification of *S. caprea*, *S. gracilistyla*, and their interspecific hybrid was extracted so that it can be used to select cross-breeding parents in subsequent breeding programs and to test the DUS of the hybrid variety.

Materials and Methods

Sample collection and measurement of vegetative characteristics. A total of 100 trees of *S. caprea* × *S. gracilistyla* (SH) were used in this study. They were sampled from a population of single full-sib progenies obtained

Received for publication 24 May 2021. Accepted for publication 29 June 2021.

Published online 3 September 2021.

H.-I.L. is the corresponding author. E-mail: iistorm@korea.kr.

This is an open access article distributed under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0/>).

Table 1. Twenty-seven vegetative characteristics (19 quantitative and eight qualitative) of *Salix caprea* (SC), *Salix gracilistyla* (SG), and their interspecific hybrid (SH) along with measurement units or states of expression.

Organ	Characteristics	Abbreviation	Measurement units or states of expression
Leaf	Leaf length	LL	cm
	Leaf width	LW	cm
	Leaf length/width ratio	LR	ratio
	Leaf width upper 1/3	LWU	cm
	Leaf width lower 1/3	LWL	cm
	Leaf base angle	LB	degree (°)
	Leaf head angle	LH	degree (°)
	Petiole length	LPL	mm
	Petiole width	LPW	mm
	Leaf thickness	LT	mm
	Lateral vein number	LVN	number
	Leaf lower hair length	HL	mm
	Number of leaf lower hairs per unit area	HN	number
	Leaf margin type	LM	irregular or serrate
Stipule	Lateral vein type	LV	joining together or not joining together
	Leaf lower hair type	LBH	curly or straight
	Stipule length	SL	mm
	Stipule width	SW	mm
	Stipule length/width ratio	SR	ratio
	Stipule serration number	SN	number
Branchlet	Stipule margin type	SM	irregular or serrate
	Branchlet hair type	BH	glabrous, glabrous or tomentose, and tomentose
Winter bud	Branchlet color	BC	yellow or red
	Winter bud length	BL	mm
	Winter bud width	BW	mm
	Winter bud hair type	WBH	glabrous or tomentose
	Winter bud color	WBC	red or yellow

in 2015 by a cross between one female tree of *S. caprea* (SC) and one male tree of *S. gracilistyla* (SG). The progenies were 5 years old and grew at an experimental site of the National Institute of Forest Science in Suwon City, Korea. Thirty-five trees of each species (SC and SG) were sampled from two natural populations at Gangneung City (for SC) and Chuncheon City (for SG) in Gangwon Province, Korea. When possible, mature trees were selected to minimize observation of immature characteristics.

Twenty-seven characteristics of four vegetative organs (leaves, stipules, branchlets, and winter buds) in 170 trees (100 for SH and 35 for each SC and SG) were measured (Table 1) as described in Wu and Raven (1999), UPOV (2006), and Korea NFSV (2019). Nineteen of the 27 characteristics were quantitative, and eight were qualitative. Details of the names, abbreviations, and measurement units (expression states for qualitative characteristics) of the 27 characteristics are given in Table 1, and the relevant characteristics are shown in Fig. 1 (for 19 quantitative characteristics) and Fig. 2 (for eight qualitative characteristics). All measurements were completed between July and August 2020.

Statistical analysis. The *agricolae* package in R (De Mendiburu and Simon, 2015) was used to calculate basic descriptive statistics for the 19 characteristics and to conduct analysis of variance (ANOVA) and Duncan's multiple range test.

Before conducting the PLS-DA, a set of data for the 27 characteristics of the 170 trees of SC, SG, and SH (three classes) was divided

into two subsets: training (70%) and testing set (30%). This data partition was implemented based on a method of species-level stratified random sampling without replacement using the *caret* package in R (Kuhn, 2008). The training set comprised 120 observations (25 observations for each SC and SG, and 70 for SH), and the testing set comprised 50 observations (10 observations for each SC and SG, and 30 for SH).

All PLS-DA processes were performed using the *mdatools* package in R (Kucheryavskiy, 2020). The following model equation was used for the PLS-DA as described in Brereton et al. (2018): $Y = XB + E$, where Y is a matrix of the response (the three classes), X is a matrix of centered and scaled predictor variables (27 characteristics), B is a matrix of regression coefficients of the predictor variables, and E is a matrix of error terms (residuals).

An algorithm, which was a statistically inspired modification of the PLS method (SIMPLS) in the *mdatools* package in R was used to decompose the X and Y matrices and to compute scores, loadings, and residuals according to the following equations, as described in Kucheryavskiy (2021) and Peerbhay et al. (2013): $X = TP + E_x$ and $Y = UQ + E_y$, where T and U are the factor score matrices, P and Q are the loading matrices, and E_x and E_y are the residuals.

Cross-validation was conducted on the training set using the leave-one-out cross-validation (LOOCV) method (Kucheryavskiy, 2021; Mabood et al., 2017). An optimal number of components (latent variables) was selected by comparing the root mean square

error (RMSE), coefficient of determination (R^2), and classification accuracy of each model generated by the LOOCV method.

Using the selected optimal number of latent variables for the 27 characteristics, the first PLS-DA model (full model) was created and then fit to the training set. The overall performance of the first model was evaluated by reviewing statistics, such as the values of RMSE, R^2 , and accuracy. In particular, the scores of variables important for projection (VIP) of each characteristic were computed and then used to select the most influential characteristics that can simplify the PLS-DA model and improve performance (Chong and Jun, 2005; Peerbhay et al., 2013; Pérez-Enciso and Tenenhaus, 2003). Regression coefficients and their corresponding P values were used along with the VIP scores to select the predictor variables. The criterion for variable selection used in this study was that the VIP score is greater than 1.0, and the P value of the regression coefficients is less than 0.05, for at least two of the three classes.

The second PLS-DA model (reduced model) was created using the selected optimal number of components and a set of most influential characteristics, and it was then fitted to the training set. The overall performance of the second model was evaluated as described for the first model.

The second model was fitted to the testing set, and the predicted values for each observation included in the testing set were computed and used to create a confusion matrix. The confusion matrix was structured with four cases of classification: true positive (TP), false negative (FN), false positive (FP), and true negative (TN). TP is the number of cases in which the given class is correctly classified as in-class, TN is the number of cases when the other class is correctly classified as out-class, FN is the number of cases when the given class is incorrectly classified as out-class, and FP is the number of cases when the other class is incorrectly classified as in-class (Ballabio and Consonni, 2013; Srout et al., 2020). The values of specificity, sensitivity, and accuracy were computed and used to evaluate the classification performance of the second model.

Results

Comparison of vegetative characteristics. Means, standard deviations, one-way ANOVAs, and Duncan's multiple range tests of the 19 quantitative characteristics of the three species are shown in Table 2. There were significant mean differences among the three species in 17 of the 19 characteristics. As shown in Fig. 2, SC and SG differed in leaf size and shape; SC had large sized and ovate-oblong shaped leaves, whereas SG had relatively small and narrow elliptic-oblong shaped leaves, and SH had intermediate-formed leaves. These differences were reflected in the five characteristics related to leaf size (LL, LW, LWU, LWL, and LB); the mean values of these characteristics were higher in SC than in SG and SH, and the

Fig. 1. Quantitative morphological characteristics of leaves, stipules, and winter buds of the three studied species. (A) *Salix caprea* leaf, (B) interspecific hybrid leaf, (C) *Salix gracilistyla* leaf, (D) *Salix caprea* stipule, (E) interspecific hybrid stipule, (F) *Salix gracilistyla* stipule, (G) *Salix caprea* winter bud, (H) interspecific hybrid winter bud, and (I) *Salix gracilistyla* winter bud. Abbreviations of flower characteristics are listed in Table 1.

differences were significant according to Duncan's multiple range test (Table 2). In the other five characteristics (LPL, LT, LVN, HL, and SW), SC also had significantly higher mean values than those in SG and SH. In only four characteristics (LR, LPW, HN, and SR), SG had higher mean values than those in SC, whereas SH showed intermediate characteristics between SC and SG. On the other hand, in another three characteristics (SN, BL, and BW), SH had higher mean values than those in SC and SG according to Duncan's multiple range test.

In five qualitative characteristics (LV, LBH, BC, WBH, and WBC), all the SCs showed only the SC type, indicating 100% uniformity. In another three qualitative characteristics (SM, LM, and BH), SC showed 77%, 89%, and 97% uniformity, respectively (Fig. 3). SG also showed only the SG type in four qualitative characteristics (LV, LBH, SM, and WBH). Among

three qualitative characteristics (LM, BH, and BC), SG had 97%, 83%, and 60% uniformity, respectively. In the WBC of SG, the frequency of the SG type was less than 14%. SH had either SC or SG types in seven qualitative characteristics, except for WBC. However, the proportions of the SC and SG types in the SH population varied by characteristics: in three qualitative characteristics (LV, LBH, and SM), the proportions of the SHs with the SC and SG type were similar; on the other hand, in another four qualitative characteristics (LM, BH, BC, and WBH), the proportion of the SHs with the SG type was higher than the proportion of the SHs with the SC type. Overall, there seemed to be many SHs more similar to SG than to SC in seven qualitative characteristics except for WBC.

Partial least squares discrimination analysis. The values of the RMSE and accuracy for each PLS-DA model generated from the cross-validation (LOOCV) performed with

the maximum number of latent variables (components), which was seven, are given in Table 3 and Fig. 4. For all species (three classes), the decreasing rate of the RMSE values for each model gradually slowed down in more than four latent variables (0.2286 for SC, 0.3478 for SG, and 0.4067 for SH), and the discriminant accuracy of each model in more than four latent variables showed no significant difference (1.0 for SC, 0.992 for SG, and 1.0 for SH). In terms of model interpretation, stability, and classification performance, four seemed to be the optimal number of latent variables (Ballabio and Consonni, 2013). Thus, four latent variables were used for the subsequent PLS-DA in the present study.

The first PLS-DA model with 27 predictor variables (characteristics) using four latent variables explained 85.6% of the total variance in the Y response variable (the three classes) (Table 4). The values of the coefficient of

Characteristics		SC	SH	SG	
LM	margin irregularly notched				serrate at margin
LV	lateral veins joining together before reaching margin				lateral veins not joining together
LBH	curly hair				straight hair
SM	margin irregularly notched				serrate at margin
BH	glabrous or downy				tomentose
BC	yellow				red
WBH	glabrous				tomentose
WBC	red				red or yellowish brown

Fig. 2. Qualitative morphological characteristics of leaves, stipules, branchlets, and winter buds of *Salix caprea* (SC), *Salix gracilistyla* (SG), and their inter-specific hybrid (SH). Abbreviations of flower characteristics are listed in Table 1.

determination (R^2) and RMSE of this model varied by class, where the SC had higher R^2 and lower RMSE values than those of SG and SH. This model also showed 100% classification accuracy for both SC and SH, but a relatively lower accuracy (99.2%) for SG.

The VIP scores of each of the 27 characteristics for the three classes, which were obtained from the first PLS-DA, are shown in Fig. 5. These values varied according to class and characteristics. Given that the VIP value of 1.0 was a cutoff criterion for variable

selection, as suggested in many related studies (Chong and Jun, 2005; Rajalahti et al., 2009; Wold et al., 2001), a total of 14 characteristics in SC, nine in SG, and 10 in SH could be selected based on these criteria. Only six characteristics (LR, BL, LV, LBH, WBH, and BC) had VIP values higher than 1.0 in all three classes. Although it seemed reasonable to use only six characteristics to create a new reduced PLS-DA model according to the widely used method of VIP-based variable selection, it is possible that such an extremely reduced number of characteristics would decrease the discrimination performance of the new model (Rajalahti et al., 2009; Villa et al., 2019). Thus, in the present study, only the characteristics with VIP values higher than 1.0 and P values of the regression coefficient less than 0.05 in at least two classes were selected and used to create the second model (i.e., the reduced model). Based on the variable selection using both VIP values and P values, seven characteristics (LR, SN, BL, LV, LBH, WBH, and BC) were finally selected; the first three were quantitative, and the remaining four were qualitative.

Compared with the first PLS-DA model, the second PLS-DA model with seven characteristics (LL, LR, HL, BL, LV, LBH, and BC) using four latent variables showed lower values in all statistics, including the total

Table 2. Summary of quantitative vegetative characteristics of *Salix caprea* (SC), *Salix gracilistyla* (SG), and their interspecific hybrid (SH).

Organ	Characteristics ^z	SC	SH	SG
Leaf	LL*** ^y	11.05 ± 1.70 ^x a ^w	9.34 ± 1.38 c	10.04 ± 1.37 b
	LW***	4.96 ± 0.67 a	3.20 ± 0.56 b	2.80 ± 0.50 c
	LR***	2.26 ± 0.38 c	2.95 ± 0.36 b	3.63 ± 0.39 a
	LWU***	3.85 ± 0.60 a	2.80 ± 0.49 b	2.60 ± 0.46 b
	LWL***	4.03 ± 0.67 a	2.93 ± 0.56 b	2.54 ± 0.49 c
	LB***	114.57 ± 28.28 a	86.74 ± 20.81 b	65.35 ± 16.72 c
	LH NS	55.58 ± 31.85	50.37 ± 16.60	52.25 ± 12.82
	LPL***	18.86 ± 4.02 a	13.25 ± 2.83 b	11.43 ± 2.83 c
	LPW**	1.51 ± 0.38 b	1.50 ± 0.26 b	1.72 ± 0.28 a
	LT***	0.11 ± 0.03 a	0.09 ± 0.02 b	0.11 ± 0.02 a
	LVN**	12.80 ± 1.98 a	11.83 ± 1.66 b	12.46 ± 1.56 ab
	HL***	0.57 ± 0.12 a	0.45 ± 0.13 b	0.32 ± 0.09 c
	HN***	16.86 ± 8.95 b	45.51 ± 22.57 a	41.23 ± 16.23 a
	SL NS	6.26 ± 1.68	6.56 ± 1.88	7.25 ± 1.70
Stipule	SW***	4.20 ± 1.13 a	3.32 ± 1.06 b	2.97 ± 0.85 b
	SN***	10.06 ± 3.23 b	13.63 ± 3.80 a	10.89 ± 4.86 b
	SR***	1.51 ± 0.22 c	2.02 ± 0.31 b	2.53 ± 0.58 a
	BL***	7.85 ± 1.99 c	16.29 ± 3.58 a	12.82 ± 2.70 b
Winter bud	BW***	3.99 ± 0.82 b	4.42 ± 0.81 a	3.63 ± 0.53 c

^zAbbreviations of flower characteristics are the same as those in Table 1.

^yAnalysis of variance test (nonsignificant, NS; significance levels: * P < 0.05, ** P < 0.01, *** P < 0.001).

^xMean ± SD.

^wDuncan's multiple range test (significant at P < 0.05).

Fig. 3. Results of qualitative vegetative characteristics frequency investigation of (A) *Salix caprea* (SC), (B) *Salix gracilistyla* (SG), and (C) their interspecific hybrid (SH). Qualitative vegetative characteristics were investigated based on the date shown in Fig. 2. Abbreviations of flower characteristics are listed in Table 1. Blue color indicates the SC type, orange indicates the SG type, and green indicates mixed A and B type.

variability explained by the model (77.7%), R^2 (90.0% for SC, 67.5% for SG, and 76.0% for SH), RMSE (0.2608 for SC, 0.4634 for SG, and 0.4840 for SH), and classification accuracy (100% for SC, 97% for SG, and 95% for SH) (Table 4). The decrease in all statistics in the second model seemed to be an inevitable consequence of using a reduced number of variables. However, the second model was selected and used for the subsequent classification of the three classes, mainly because this model showed a discrimination accuracy sufficient to be used for the classification, assuming that the error rate of classification is less than 5%.

The regression coefficients, by class, for the seven characteristics included in the second PLS-DA model are shown in Fig. 6. Because it represented the relative magnitude and direction of the effect of each characteristic in species discrimination, the regression coefficient plot indicated the following. First, the direction of effects of four characteristics (LR, LV, LBH, and SN) of SG on species classification was opposite to that of SC and SH. Thus, a species with characteristics such as longer leaf length compared with width (LR), lateral vein type joining together before reaching margin (LV), straight hair type of leaf lower part (LBH), and less number of

Table 3. Results of the leave-one-out cross-validation (LOOCV) for the partial least squares discrimination analysis (PLS-DA) model with all 27 predictor variables on the training dataset by species [*Salix caprea* (SC), *Salix gracilistyla* (SG), and their interspecific hybrid (SH)]. R^2 , the root mean square error (RMSE), and accuracy of each component are shown.

	Component	R^2	RMSE	Accuracy
SC	Comp 1	0.8038	0.3598	0.992
	Comp 2	0.8617	0.3020	1
	Comp 3	0.9147	0.2372	1
	Comp 4	0.9208	0.2286	1
	Comp 5	0.9359	0.2057	1
	Comp 6	0.9359	0.2056	1
	Comp 7	0.9393	0.2001	1
SG	Comp 1	0.2279	0.7137	0.792
	Comp 2	0.7397	0.4144	0.983
	Comp 3	0.7485	0.4074	0.992
	Comp 4	0.8167	0.3478	0.992
	Comp 5	0.8179	0.3466	0.992
	Comp 6	0.8282	0.3366	1
	Comp 7	0.8283	0.3366	1
SH	Comp 1	0.1192	0.9254	0.758
	Comp 2	0.7394	0.5033	0.967
	Comp 3	0.7520	0.4910	0.983
	Comp 4	0.8299	0.4067	1
	Comp 5	0.8352	0.4003	0.992
	Comp 6	0.8416	0.3924	1
	Comp 7	0.8434	0.3901	1

stipule serration (SN), was more likely to be classified as SG by the second PLS-DA model, but the reverse was likely for SH and SC. Second, the direction of effects of two characteristics (BC and BL) of SH was opposite to that of SC and SG. The species with red-colored branchlet and long-length winter bud was classified as SH, but the reverse was likely for SC and SG. Third, the direction of WBH type of SC was opposite to that of both SH and SG. The species with glabrous hair type of winter bud was more likely to be classified as SC, but the reverse was likely for SH.

The classification performance of the second model for the testing set is shown in Table 5. The second model showed a mean accuracy of 94% in the classification (86% for SH, 96% for SG, and 100% for SC), a mean sensitivity of 86% (80% for SG, 83.3% for SH, 100% for SC), and a mean specificity of 96.7% (90% for SH, 100% for SC and SG). The classification performance of the first model for the testing set is indicated in Table 5. Compared with the first model, the second model showed lower classification performance in terms of accuracy, sensitivity, and specificity. However, the classification performance was not very different between the two models, and in the consistent observation was that the misclassification of the two models was observed in both SG and SH. Therefore, considering these two facts, it seemed that the second PLS-DA model with seven characteristics could be used to discriminate and classify the three classes.

Discussion

The second PLS-DA model with seven characteristics (BL, SN, LR, LV, LBH, BC,

Fig. 4. The root mean square error (RMSE) value of the leave-one-out cross-validation (LOOCV) for the partial least squares discrimination analysis (PLS-DA) model with all 27 predictor variables. (A) *Salix caprea* (SC), (B) *Salix gracilistyla* (SG), and (C) their interspecific hybrid (SH).

and WBH) could discriminate SC, SG, and SH with an 86% to 100% accuracy (100% for SC, 96% for SG, and 86% for SH). This accuracy was lower than that obtained with the first PLS-DA model with 27 characteristics (100% for SC, 98% for SG, and 92% for SH). For more accurate classification of the three species, it was better to use all 27 characteristics included in the first PLD-DA model rather than seven characteristics in the second model. However, measuring all 27 characteristics is expensive; hence, the second PLS-DA model with seven characteristics appears to be more desirable and practical in terms of cost-effectiveness.

In addition, the second model showed lower discriminant accuracy for SG and SH than for SC (Table 5). It misclassified two SGs into SH and could not classify five SHs. The misclassification and nonclassification of the second model were caused due to similarity between SG and SH in the seven characteristics included in the model (Table 2, Fig. 3). This similarity could be due to the

unintentional use of SC similar to SG in the seven characteristics.

It is very difficult to obtain progenies that are distinct from their parents through just one breeding, as most characteristics of tree species are polygenic traits (Sewell and Neale, 2000; Weih et al., 2006). Furthermore, a specific genotype combination of the multiple genes related to the best performance of the given characteristics can be obtained only through repeated multiple-generation breeding between the highest-grade progenies. Thus, subsequent hybridization experiments are also needed to create SHs that are more distinct from SC and SG. Two characteristics (BC and BL) that significantly influenced the discrimination of SH from SC and SG can be used as criteria for selecting SH individuals as mating parents in the hybridization. Particularly, it would be desirable to hybridize the SH parents with BC and BL of higher grades, for the development of a more distinct SH variety.

If one of the SHs more distinct from SC and SG was applied to be registered for the

protection of new SH varieties, this SH would have to be tested for the DUS of its characteristics using the DUS test guidelines of the related available species according to act on the protection of new plant varieties (Korea Ministry of Agriculture Food and Rural Affairs, 2017). The DUS test guidelines for SH have not been prepared yet, so the guideline for SC, which was established by the Korea Forest Service in 2020, will inevitably have to be used as an alternative (Korea NFSV, 2019). However, the DUS test guidelines on SC do not include six of the seven characteristics that have significantly contributed to the discrimination among SC, SG, and SH (the six characteristics being LBH, LV, SN, BC, BL, and WBH). Consequently, the guidelines for SC need to be reestablished to include these six characteristics.

In conclusion, the results of the present study on the discrimination of SC, SG, and SH using 27 vegetative characteristics and PLS-DA methods clearly indicated the following two advantages of PLS-DA. First, PLS-DA can create a model with a linear combination of multiple intercorrelated characteristics relatively freely of multicollinearity and dimensionality, which are the main problems of LDA (Barker and Rayens, 2003). Second, PLS-DA had the advantages of facilitating the selection of characteristics that greatly influenced the discrimination of SC, SG, and SH, as well as comparing the relative importance and direction of influence of the selected characteristics using regression coefficients of these characteristics (Balabio and Consonni, 2013). Therefore, it is expected that PLS-DA methods will greatly contribute to related studies investigating identification, discrimination, classification,

Table 4. Results of the partial least squares discrimination analysis (PLS-DA) model with four-component latent variables of Xs predictors (27 variables) (first model) and the PLS-DA model with four-component latent variables of Xs predictors (seven variables) (second model) on the training dataset [120 observations: 25 observations for *Salix caprea* (SC), 25 for *Salix gracilistyla* (SG), and 70 for their interspecific hybrid (SH)] shown by three classes (SC, SG, and SH).

	Class	R ²	RMSE ^z	X cumulative	Y cumulative	Specificity	Sensitivity	Accuracy
First model	SC	0.9208	0.2286	55.12	85.58	1.00	1.00	1.00
	SG	0.8167	0.3478	55.12	85.58	1.00	0.96	0.99
	SH	0.8299	0.4067	55.12	85.58	1.00	1.00	1.00
Second model	SC	0.8969	0.2608	95.27	77.68	1.00	1.00	1.00
	SG	0.6745	0.4634	95.27	77.68	0.98	0.92	0.97
	SH	0.7590	0.4840	95.27	77.68	0.92	0.97	0.95

^zRMSE = root mean square error.

Fig. 5. The variable influence on projection (VIP) values by predictor obtained from the partial least squares discrimination analysis (PLS-DA) model with four-component latent variables of Xs predictors (27 variables) on the training dataset by species. (A) *Salix caprea* (SC), (B) *Salix gracilistyla* (SG), and (C) their interspecific hybrid (SH). Abbreviations of flower characteristics are the same as those listed in Table 1.

Table 5. Confusion matrix for the results of the partial least squares discrimination analysis (PLS-DA) model with four-component latent variables of Xs predictors (27 variables) (first model) and the PLS-DA model with four-component latent variables of Xs predictors (seven variables) (second model) on the test dataset [50 observations: 10 observations for *Salix caprea* (SC), 10 for *Salix gracilistyla* (SG), and 30 for their interspecific hybrid (SH)] shown by three classes (SC, SG, and SH). Bold numbers indicate misclassification, and italic numbers indicate nonclassification.

	Class	TP (true positives)	FP (false positives)	TN (true negatives)	FN (false negatives)	Specificity	Sensitivity	Accuracy
First model	SC	10	0	40	0	1.00	1.00	1.00
	SG	9	0	40	1	1.00	0.90	0.98
	SH	27	1	19	3	0.95	0.90	0.92
Second model	SC	10	0	40	0	1.00	1.00	1.00
	SG	8	0	40	2	1.00	0.80	0.96
	SH	25	2	18	5	0.90	0.83	0.86

Fig. 6. Regression coefficients plot obtained from the second partial least squares discrimination analysis (PLS-DA) model with seven predictor variables. Blue: *Salix caprea* (SC); Yellow: *Salix gracilistyla* (SG); Green: their interspecific hybrid (SH).

and breeding, if used along with cluster analysis and PCA.

Literature Cited

- Ballabio, D. and V. Consonni. 2013. Classification tools in chemistry. Part I: Linear models PLS-DA. *Anal. Methods* 5:3790–3798, doi: 10.1039/C3AY40582F.
- Barker, M. and W. Rayens. 2003. Partial least squares for discrimination. *J. Chem.* 17:166–173, doi: 10.1002/cem.785.
- Brereton, R.G., J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J.M. Roger, B. Walczak, and R. Tauler. 2018. Chemometrics in analytical chemistry-part II: Modeling, validation, and applications. *Anal. Bioanal. Chem.* 410: 6691–6704, doi: 10.1007/s00216-018-1283-4.
- Bruce, P. and A. Bruce. 2017. Practical statistical for data scientists. O'Reilly Media, Inc., Sebastopol, CA.
- Cho, H.J., H. Woo, J. Lee, and K.H. Cho. 2008. Changes in riparian vegetation after restoration in a urban stream, Yangjae stream (in Korean with English abstract). *J. Wet. Res.* 10(3):111–124.
- Choi, H. and J.G. Kim. 2015. Study on characteristics of seed germination and seedling growth in *Salix gracilistyla* for invasive species management (in Korean with English abstract). *J. Korea. Soc. Environ. Restor. Technol.* 18(3): 79–95, doi: 10.13087/kosert.2015.18.3.79.
- Chong, I.G. and C.H. Jun. 2005. Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.* 78: 103–112, doi: 10.1016/j.chemolab.2004.12.011.
- De Luca, M., W. Terrouzi, F. Kzaiber, G. Ioele, A. Oussama, and G. Ragno. 2012. Classification of Moroccan olive cultivars by linear discriminant analysis applied to ATR-FTIR spectra of endocarps. *Int. J. Food Sci. Technol.* 47: 1286–1292, doi: 10.1111/j.1365-2621.2012.02972.x.
- De Mendiburu, F. and R. Simon. 2015. Agricolae - Ten years of an open source statistical tool for experiments in breeding, agriculture and biology. *PeerJ PrePrints* 3:e1404v1. <https://doi.org/10.7287/peerj.preprints.1404v1>.
- Fonville, J.M., S.E. Richards, R.H. Barton, C.L. Boulange, T.M.D. Ebbels, J.K. Nicholson, E. Holmes, and M.-E. Dumas. 2010. The evolution of partial least square models and related chemometric approaches in metabolomics and metabolite phenotyping. *J. Chemometr.* 24:636–649, doi: 10.1002/cem.1359.
- Galdón, B.R., L.H. Rodríguez, D.R. Mesa, H.L. León, N.L. Pérez, E.M.R. Rodríguez, and C.D. Romero. 2012. Differentiation of potato cultivars experimentally cultivated based on their chemical composition and by applying linear discriminant analysis. *Food Chem.* 133:1241–1248, doi: 10.1016/j.foodchem.2011.10.016.
- Hur, S.H., S.W. Kim, and B.W. Min. 2015. Discrimination of cultivars and cultivation origins from the sepals of dry persimmon using FT-IR spectroscopy combined with multivariate analysis (in Korean with English abstract). *Korean J. Food Sci. Technol.* 47:20–26, doi: 10.9721/KJFST.2015.47.1.20.
- Kong, W., C. Zhang, F. Liu, P. Nie, and Y. He. 2013. Rice seed cultivar identification using near-infrared hyperspectral imaging and multivariate data analysis. *Sensors (Basel)* 13:8916–8927, doi: 10.3390/s130708916.
- Korea Ministry of Agriculture Food and Rural Affairs. 2017. Act on the protection of new varieties of plants (Act No. 15075, 28 Nov. 2017). <https://www.law.go.kr/LSW/eng/engLsSc.do?menuId=2§ion=bodyText&query=15075&x=0&y=0#liBgcolor0>.
- Korea National Forest Seed and Variety Center (NFSV). 2019. Guidelines for measuring characteristics by crop for examination of new variety: *Salix caprea* L. Chungju, South Korea (in Korean).
- Korir, N.K., J. Han, L. Shangguan, C. Wang, E. Kayesh, Y. Zhang, and J. Fang. 2012. Plant variety and cultivar identification: Advances and prospects. *Crit. Rev. Biotechnol.* 15:111–125, doi: 10.3109/07388551.2012.675314.
- Kucheryavskiy, S. 2021. Getting started with mdatools for R. 29 Mar. 2021. <https://mdatools.com/docs/index.html>.
- Kucheryavskiy, S. 2020. mdatools - R package for chemometrics. *Chemom. Intell. Lab. Syst.* 198: 103937, doi: 10.1016/j.chemolab.2020.103937.
- Kuhn, M. 2008. Building predictive models in R using the caret package. *J. Stat. Softw.* 28(5):1–26, doi: 10.18637/jss.v028.i05.
- Kuhn, M. and K. Johnson. 2013. Applied predictive modeling. Springer, New York, NY, doi: 10.1007/978-1-4614-6849-3.
- Kwon, Y.K., M.S. Ahn, J.S. Park, J.R. Liu, D.S. In, B.W. Min, and S.W. Kim. 2014. Discrimination of cultivation ages and cultivars of ginseng leaves using Fourier transform infrared spectroscopy combined with multivariate analysis. *J. Ginseng Res.* 38(1):52–58, doi: 10.1016/j.jgr.2013.11.006.
- Lee, T.B. 2003. Coloured flora of Korea. Hayangmunsa, Seoul, Korea. Vol. 2. (in Korean).
- Mabood, F., F. Jabeen, J. Hussain, A. Al-Harrasi, A. Hamaed, S.A.A. Al Mashaykhi, Z.M.A. Al Rubaiey, S. Manzoor, A. Khan, Q.M.I. Haq, S.A. Gilani, and A. Khan. 2017. FT-NIRS coupled with chemometric methods as a rapid alternative tool for the detection & quantification of cow milk adulteration in camel milk samples. *Vib. Spectrosc.* 92:245–250, doi: 10.1016/j.vibspec.2017.07.004.
- Peerbhay, K.Y., O. Mutanga, and R. Ismail. 2013. Commercial tree species discrimination using airborne AISA Eagle hyperspectral imagery and partial least squares discriminant analysis (PLS-DA) in KwaZulu-Natal, South Africa. *ISPRS J. Photogramm. Remote Sens.* 79:19–28, doi: 10.1016/j.isprsjprs.2013.01.013.
- Pérez-Enciso, M. and M. Tenenhaus. 2003. Prediction of clinical outcome with microarray data: A partial least squares discriminant analysis (PLS-DA) approach. *Hum. Genet.* 112:581–592, doi: 10.1007/s00439-003-0921-9.
- Rajalahti, T., R. Arneberg, A. Kroksveen, M. Berie, K.M. Myhr, and M. Kvalheim. 2009. Discriminating variable test and selectivity ratio plot: Quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Anal. Chem.* 81:2581–2590, doi: 10.1021/ac802514y.
- Ruiz-Perez, D., H. Guan, P. Madhivanan, K. Mathee, and G. Narasimhan. 2020. So you think you can PLS-DA? *BMC Bioinformatics* 21:1–10, doi: 10.1186/s12859-019-3310-7.
- Seo, H.N., S.B. Chae, H.I. Lim, W. Cho, and W.Y. Lee. 2021. The flower morphological

- characteristics of *Salix caprea* × *Salix gracilistyla*. J. For. Environ. Sci. 37:35–43, doi: 10.7747/JFES.2021.37.1.35.
- Sewell, M.M. and D.B. Neale. 2000. Mapping quantitative traits in forest trees, p. 407–423. In: S.M. Jain and S.C. Minocha (eds.). Molecular biology of woody plants. Forestry Sciences, Vol. 64. Springer, Dordrecht, doi: 10.1007/978-94-017-2311-4_17.
- Shrestha, S., L.C. Deleuran, and R. Gislum. 2016. Classification of different tomato seed cultivars by multispectral visible-near infrared spectroscopy and chemometrics. J. Spectral Imaging 5:1–9, doi: 10.1255/jsi.2016.a1.
- Srout, L., B.D. Byrd, and S.W. Huffman. 2020. Classification of mosquitoes with infrared spectroscopy and partial least squares-discriminant analysis. Appl. Spectrosc. 74:900–912, doi: 10.1177/0003702820915729.
- UPOV. 2002. General introduction to the examination of distinctness, uniformity and stability and the development of harmonized descriptions of new varieties of plants. TG/1/3. Union Internationale pour la Protection des Obtentions Végétales, Geneva, Switzerland. <https://www.upov.int/en/publications/tg-rom/tg001/tg_1_3.pdf>.
- UPOV. 2006. International Union for the protection of new varieties of plants. WILLOW. UPOV Code: SALIX. Guidelines for the conduct of tests for distinctness, uniformity and stability. TG/72/6. Union Internationale pour la Protection des Obtentions Végétales, Geneva, Switzerland. <<https://www.upov.int/edocs/tgdocs/en/tg072.pdf>>.
- Vaculík, M., C. Konlechner, I. Langer, W. Adlassnig, M. Puschenreiter, A. Lux, and M.T. Hauser. 2012. Root anatomy and element distribution vary between two *Salix caprea* isolates with different Cd accumulation capacities. Environ. Pollut. 163:117–126, doi: 10.1016/j.envpol.2011.12.031.
- Villa, J.E.L., N.R. Quiñones, F. Fantinatti-Garboggini, and R.J. Poppi. 2019. Fast discrimination of bacteria using a filter paper-based SERS platform and PLS-DA with uncertainty estimation. Anal. Bioanal. Chem. 411:705–713, doi: 10.1007/s00216-018-1485-9.
- Weih, M., A.C. Rönnberg-Wästljung, and C. Glynn. 2006. Genetic basis of phenotypic correlations among growth traits in hybrid willow (*Salix dasyclados* × *S. viminalis*) grown under two water regimes. New Phytol. 170:467–477, doi: 10.1111/j.1469-8137.2006.01685.x.
- Wold, S., M. Sjöström, and L. Eriksson. 2001. PLS-regression: A basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58(2):109–130, doi: 10.1016/S0169-7439(01)00155-1.
- Wu, Z.Y. and P.H. Raven. 1999. Flora of China. Vol. 4 (Cycadaceae through Fagaceae). Science Press, Beijing, and Missouri Botanical Garden Press, St. Louis, doi: 10.1111/j.1756-1051.1999.tb01142.x.
- Yan, S.M., J.P. Liu, L. Xu, X.S. Fu, H.F. Cui, Z.Y. Yun, X.P. Yu, and Z.H. Ye. 2014. Rapid discrimination of the geographical origins of an oolong tea (anxi-tieguanyin) by near-infrared spectroscopy and partial least squares discriminant analysis. J. Anal. Methods Chem. 1:704971, doi: 10.1155/2014/704971.