SAS/STAT® 14.2 User's Guide
The DISCRIM Procedure

This document is an individual chapter from *SAS/STAT® 14.2 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2016. *SAS/STAT® 14.2 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 14.2 User's Guide

Copyright © 2016, SAS Institute Inc., Cary, NC, USA

All Rights Reserved. Produced in the United States of America.

For a hard-copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2016

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to `http://support.sas.com/thirdpartylicenses`.

Chapter 36

The DISCRIM Procedure

Contents

Overview: DISCRIM Procedure

For a set of observations containing one or more quantitative variables and a classification variable defining groups of observations, the DISCRIM procedure develops a discriminant criterion to classify each observation into one of the groups. The derived discriminant criterion from this data set can be applied to a second data set during the same execution of PROC DISCRIM. The data set that PROC DISCRIM uses to derive the discriminant criterion is called the *training* or *calibration* data set.

When the distribution within each group is assumed to be multivariate normal, a parametric method can be used to develop a discriminant function. The discriminant function, also known as a classification criterion, is determined by a measure of generalized squared distance (Rao 1973). The classification criterion can be based on either the individual within-group covariance matrices (yielding a quadratic function) or the pooled covariance matrix (yielding a linear function); it also takes into account the prior probabilities of the groups. The calibration information can be stored in a special SAS data set and applied to other data sets.

When no assumptions can be made about the distribution within each group, or when the distribution is assumed not to be multivariate normal, nonparametric methods can be used to estimate the group-specific densities. These methods include the kernel and k-nearest-neighbor methods (Rosenblatt 1956; Parzen 1962). The DISCRIM procedure uses uniform, normal, Epanechnikov, biweight, or triweight kernels for density estimation.

Either Mahalanobis or Euclidean distance can be used to determine proximity. Mahalanobis distance can be based on either the full covariance matrix or the diagonal matrix of variances. With a k-nearest-neighbor method, the pooled covariance matrix is used to calculate the Mahalanobis distances. With a kernel method, either the individual within-group covariance matrices or the pooled covariance matrix can be used to calculate the Mahalanobis distances. With the estimated group-specific densities and their associated prior probabilities, the posterior probability estimates of group membership for each class can be evaluated.

Canonical discriminant analysis is a dimension-reduction technique related to principal component analysis and canonical correlation. Given a classification variable and several quantitative variables, PROC DISCRIM derives canonical variables (linear combinations of the quantitative variables) that summarize between-class variation in much the same way that principal components summarize total variation. (See Chapter 31, “The CANDISC Procedure,” for more information about canonical discriminant analysis.) A discriminant criterion is always derived in PROC DISCRIM. If you want canonical discriminant analysis without the use of a discriminant criterion, you should use the CANDISC procedure.

The DISCRIM procedure can produce an output data set containing various statistics such as means, standard deviations, and correlations. If a parametric method is used, the discriminant function is also stored in the data set to classify future observations. When canonical discriminant analysis is performed, the output data set includes canonical coefficients that can be rotated by the FACTOR procedure. PROC DISCRIM can also create a second type of output data set containing the classification results for each observation. When canonical discriminant analysis is performed, this output data set also includes canonical variable scores. A third type of output data set containing the group-specific density estimates at each observation can also be produced.

PROC DISCRIM evaluates the performance of a discriminant criterion by estimating error rates (probabilities of misclassification) in the classification of future observations. These error-rate estimates include error-count estimates and posterior probability error-rate estimates. When the input data set is an ordinary SAS data set, the error rate can also be estimated by cross validation.

Do not confuse discriminant analysis with cluster analysis. All varieties of discriminant analysis require prior knowledge of the classes, usually in the form of a sample from each class. In cluster analysis, the data do not include information about class membership; the purpose is to construct a classification.

See Chapter 10, "Introduction to Discriminant Procedures," for a discussion of discriminant analysis.

Getting Started: DISCRIM Procedure

The data in this example are measurements of 159 fish caught in Finland's Lake Laengelmaevesi; this data set is available from Puranen (1917). For each of the seven species (bream, roach, whitefish, parkki, perch, pike, and smelt) the weight, length, height, and width of each fish are tallied. Three different length measurements are recorded: from the nose of the fish to the beginning of its tail, from the nose to the notch of its tail, and from the nose to the end of its tail. The height and width are recorded as percentages of the third length variable. The fish data set is available from the Sashelp library. The goal now is to find a discriminant function based on these six variables that best classifies the fish into species.

First, assume that the data are normally distributed within each group with equal covariances across groups. The following statements use PROC DISCRIM to analyze the Sashelp.Fish data and create Figure 36.1 through Figure 36.5:

```
title 'Fish Measurement Data';

proc discrim data=sashelp.fish;
   class Species;
run;
```

The DISCRIM procedure begins by displaying summary information about the variables in the analysis (see Figure 36.1). This information includes the number of observations, the number of quantitative variables in the analysis (specified with the VAR statement), and the number of classes in the classification variable (specified with the CLASS statement). The frequency of each class, its weight, the proportion of the total sample, and the prior probability are also displayed. Equal priors are assigned by default.

Figure 36.1 Summary Information

Fish Measurement Data

The DISCRIM Procedure

Total Sample Size	158	**DF Total**	157
Variables	6	**DF Within Classes**	151
Classes	7	**DF Between Classes**	6

Number of Observations Read	159
Number of Observations Used	158

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion	Prior Probability
Bream	Bream	34	34.0000	0.215190	0.142857
Parkki	Parkki	11	11.0000	0.069620	0.142857
Perch	Perch	56	56.0000	0.354430	0.142857
Pike	Pike	17	17.0000	0.107595	0.142857
Roach	Roach	20	20.0000	0.126582	0.142857
Smelt	Smelt	14	14.0000	0.088608	0.142857
Whitefish	Whitefish	6	6.0000	0.037975	0.142857

The natural log of the determinant of the pooled covariance matrix is displayed in Figure 36.2.

Figure 36.2 Pooled Covariance Matrix Information

Pooled Covariance Matrix Information

Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
6	4.17613

The squared distances between the classes are shown in Figure 36.3.

Figure 36.3 Squared Distances

Fish Measurement Data

The DISCRIM Procedure

Generalized Squared Distance to Species

From Species	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish
Bream	0	83.32523	243.66688	310.52333	133.06721	252.75503	132.05820
Parkki	83.32523	0	57.09760	174.20918	27.00096	60.52076	26.54855
Perch	243.66688	57.09760	0	101.06791	29.21632	29.26806	20.43791
Pike	310.52333	174.20918	101.06791	0	92.40876	127.82177	99.90673
Roach	133.06721	27.00096	29.21632	92.40876	0	33.84280	6.31997
Smelt	252.75503	60.52076	29.26806	127.82177	33.84280	0	46.37326
Whitefish	132.05820	26.54855	20.43791	99.90673	6.31997	46.37326	0

The coefficients of the linear discriminant function are displayed (in Figure 36.4) with the default options METHOD=NORMAL and POOL=YES.

Figure 36.4 Linear Discriminant Function

Linear Discriminant Function for Species

Variable	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish
Constant	-185.91682	-64.92517	-48.68009	-148.06402	-62.65963	-19.70401	-67.44603
Weight	-0.10912	-0.09031	-0.09418	-0.13805	-0.09901	-0.05778	-0.09948
Length1	-23.02273	-13.64180	-19.45368	-20.92442	-14.63635	-4.09257	-22.57117
Length2	-26.70692	-5.38195	17.33061	6.19887	-7.47195	-3.63996	3.83450
Length3	50.55780	20.89531	5.25993	22.94989	25.00702	10.60171	21.12638
Height	13.91638	8.44567	-1.42833	-8.99687	-0.26083	-1.84569	0.64957
Width	-23.71895	-13.38592	1.32749	-9.13410	-3.74542	-3.43630	-2.52442

A summary of how the discriminant function classifies the data used to develop the function is displayed last. In Figure 36.5, you see that only three of the observations are misclassified. The error-count estimates give the proportion of misclassified observations in each group. Since you are classifying the same data that are used to derive the discriminant function, these error-count estimates are biased.

Figure 36.5 Resubstitution Misclassification Summary

Fish Measurement Data

The DISCRIM Procedure
Classification Summary for Calibration Data: SASHELP.FISH
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into Species

From Species	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish	Total
Bream	34 100.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	34 100.00
Parkki	0 0.00	11 100.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	11 100.00
Perch	0 0.00	0 0.00	53 94.64	0 0.00	0 0.00	3 5.36	0 0.00	56 100.00
Pike	0 0.00	0 0.00	0 0.00	17 100.00	0 0.00	0 0.00	0 0.00	17 100.00
Roach	0 0.00	0 0.00	0 0.00	0 0.00	20 100.00	0 0.00	0 0.00	20 100.00
Smelt	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	14 100.00	0 0.00	14 100.00
Whitefish	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	6 100.00	6 100.00
Total	34 21.52	11 6.96	53 33.54	17 10.76	20 12.66	17 10.76	6 3.80	158 100.00
Priors	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286	

Error Count Estimates for Species

	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish	Total
Rate	0.0000	0.0000	0.0536	0.0000	0.0000	0.0000	0.0000	0.0077
Priors	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	

One way to reduce the bias of the error-count estimates is to split your data into two sets. One set is used to derive the discriminant function, and the other set is used to run validation tests. Example 36.4 shows how to analyze a test data set. Another method of reducing bias is to classify each observation by using a discriminant function computed from all of the other observations; this method is invoked with the CROSSVALIDATE option.

Syntax: DISCRIM Procedure

The following statements are available in the DISCRIM procedure:

PROC DISCRIM < *options* > ;
  CLASS *variable* ;
  BY *variables* ;
  FREQ *variable* ;
  ID *variable* ;
  PRIORS *probabilities* ;
  TESTCLASS *variable* ;
  TESTFREQ *variable* ;
  TESTID *variable* ;
  VAR *variables* ;
  WEIGHT *variable* ;

Only the PROC DISCRIM and CLASS statements are required.

The following sections describe the PROC DISCRIM statement and then describe the other statements in alphabetical order.

PROC DISCRIM Statement

PROC DISCRIM < *options* > ;

The PROC DISCRIM statement invokes the DISCRIM procedure. Table 36.1 summarizes the options available in the PROC DISCRIM statement.

Table 36.1 Options Available in the PROC DISCRIM Statement

Option	Description
Input Data Sets	
DATA=	Specifies input SAS data set
TESTDATA=	Specifies input SAS data set to classify
Output Data Sets	
OUTSTAT=	Specifies output statistics data set
OUT=	Specifies output data set with classification results
OUTCROSS=	Specifies output data set with cross validation results
OUTD=	Specifies output data set with densities
SCORES=	Outputs discriminant scores to the OUT= data set

Table 36.1 *continued*

Option	Description
TESTOUT=	Specifies output data set with TEST= results
TESTOUTD=	Specifies output data set with TEST= densities
Method Details	
METHOD=	Specifies parametric or nonparametric method
POOL=	Specifies whether to pool the covariance matrices
SINGULAR=	Specifies the singularity criterion
SLPOOL=	Specifies significance level homogeneity test
THRESHOLD=	Specifies the minimum threshold for classification
Nonparametric Methods	
K=	Specifies k value for k nearest neighbors
KPROP=	Specifies proportion, p, for computing k
R=	Specifies radius for kernel density estimation
KERNEL=	Specifies a kernel density to estimate
METRIC=	Specifies metric in for squared distances
Canonical Discriminant Analysis	
CANONICAL	Performs canonical discriminant analysis
CANPREFIX=	Specifies a prefix for naming the canonical variables
NCAN=	Specifies the number of canonical variables
Resubstitution Classification	
LIST	Displays the classification results
LISTERR	Displays the misclassified observations
NOCLASSIFY	Suppresses the classification
TESTLIST	Displays the classification results of TEST=
TESTLISTERR	Displays the misclassified observations of TEST=
Cross Validation Classification	
CROSSLIST	Displays the cross validation results
CROSSLISTERR	Displays the misclassified cross validation results
CROSSVALIDATE	Specifies cross validation
Control Displayed Output	
ALL	Displays all output
ANOVA	Displays univariate statistics
BCORR	Displays between correlations
BCOV	Displays between covariances
BSSCP	Displays between SSCPs
DISTANCE	Displays squared Mahalanobis distances
FORMULA	Displays formulas in destinations other than LISTING
MANOVA	Displays multivariate ANOVA results
NOPRINT	Suppresses all displayed output
PCORR	Displays pooled correlations
PCOV	Displays pooled covariances
POSTERR	Displays posterior probability error-rate estimates

Table 36.1 *continued*

Option	Description
PSSCP	Displays pooled SSCPs
SHORT	Suppresses some displayed output
SIMPLE	Displays simple descriptive statistics
STDMEAN	Displays standardized class means
TCORR	Displays total correlations
TCOV	Displays total covariances
TSSCP	Displays total SSCPs
WCORR	Displays within correlations
WCOV	Displays within covariances
WSSCP	Displays within SSCPs

ALL

activates all options that control displayed output. When the derived classification criterion is used to classify observations, the ALL option also activates the POSTERR option.

ANOVA

displays univariate statistics for testing the hypothesis that the class means are equal in the population for each variable.

BCORR

displays between-class correlations.

BCOV

displays between-class covariances. The between-class covariance matrix equals the between-class SSCP matrix divided by $n(c-1)/c$, where n is the number of observations and c is the number of classes. You should interpret the between-class covariances in comparison with the total-sample and within-class covariances, not as formal estimates of population parameters.

BSSCP

displays the between-class SSCP matrix.

CANONICAL

CAN

performs canonical discriminant analysis.

CANPREFIX=*name*

specifies a prefix for naming the canonical variables. By default, the names are Can1, Can2, ..., Can$n$. If you specify CANPREFIX=ABC, the components are named ABC1, ABC2, ABC3, and so on. The number of characters in the prefix, plus the number of digits required to designate the canonical variables, should not exceed 32. The prefix is truncated if the combined length exceeds 32.

The CANONICAL option is activated when you specify either the NCAN= or the CANPREFIX= option. A discriminant criterion is always derived in PROC DISCRIM. If you want canonical discriminant analysis without the use of discriminant criteria, you should use PROC CANDISC.

CROSSLIST

displays the cross validation classification results for each observation.

CROSSLISTERR

displays the cross validation classification results for misclassified observations only.

CROSSVALIDATE

specifies the cross validation classification of the input DATA= data set. When a parametric method is used, PROC DISCRIM classifies each observation in the DATA= data set by using a discriminant function computed from the other observations in the DATA= data set, excluding the observation being classified. When a nonparametric method is used, the covariance matrices used to compute the distances are based on all observations in the data set and do not exclude the observation being classified. However, the observation being classified is excluded from the nonparametric density estimation (if you specify the R= option) or the k nearest neighbors (if you specify the K= or KPROP= option) of that observation. The CROSSVALIDATE option is set when you specify the CROSSLIST, CROSSLISTERR, or OUTCROSS= option. With these options, cross validation information is displayed or output in addition to the usual resubstitution classification results. Cross validation classification results are written to the OUTCROSS= data set, and resubstitution classification results are written to the OUT= data set.

DATA=*SAS-data-set*

specifies the data set to be analyzed. The data set can be an ordinary SAS data set or one of several specially structured data sets created by SAS/STAT procedures. These specially structured data sets include TYPE=CORR, TYPE=COV, TYPE=CSSCP, TYPE=SSCP, TYPE=LINEAR, TYPE=QUAD, and TYPE=MIXED. The input data set must be an ordinary SAS data set if you specify METHOD=NPAR. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

DISTANCE

MAHALANOBIS

displays the squared Mahalanobis distances between the group means, F statistics, and the corresponding probabilities of greater Mahalanobis squared distances between the group means. The squared distances are based on the specification of the POOL= and METRIC= options.

FORMULA

displays formulas in destinations other than LISTING. See the section “Formulas” on page 2300 for more information.

K=*k*

specifies a *k* value for the *k*-nearest-neighbor rule. An observation $\mathbf{x}$ is classified into a group based on the information from the *k* nearest neighbors of $\mathbf{x}$. Do not specify the K= option with the KPROP= or R= option.

KPROP=*p*

specifies a proportion, *p*, for computing the k value for the k-nearest-neighbor rule: $k = \max(1, \text{floor}(np))$, where n is the number of valid observations. When there is a FREQ statement, n is the sum of the FREQ variable for the observations used in the analysis (those without missing or invalid values). An observation $\mathbf{x}$ is classified into a group based on the information from the k nearest neighbors of $\mathbf{x}$. Do not specify the KPROP= option with the K= or R= option.

KERNEL=BIWEIGHT | BIW

KERNEL=EPANECHNIKOV | EPA

KERNEL=NORMAL | NOR

KERNEL=TRIWEIGHT | TRI

KERNEL=UNIFORM | UNI

specifies a kernel density to estimate the group-specific densities. You can specify the KERNEL= option only when the R= option is specified. The default is KERNEL=UNIFORM.

LIST

displays the resubstitution classification results for each observation. You can specify this option only when the input data set is an ordinary SAS data set.

LISTERR

displays the resubstitution classification results for misclassified observations only. You can specify this option only when the input data set is an ordinary SAS data set.

MANOVA

displays multivariate statistics for testing the hypothesis that the class means are equal in the population.

METHOD=NORMAL | NPAR

determines the method to use in deriving the classification criterion. When you specify METHOD=NORMAL, a parametric method based on a multivariate normal distribution within each class is used to derive a linear or quadratic discriminant function. The default is METHOD=NORMAL. When you specify METHOD=NPAR, a nonparametric method is used and you must also specify either the K= or R= option.

METRIC=DIAGONAL | FULL | IDENTITY

specifies the metric in which the computations of squared distances are performed. If you specify METRIC=FULL, then PROC DISCRIM uses either the pooled covariance matrix (POOL=YES) or individual within-group covariance matrices (POOL=NO) to compute the squared distances. If you specify METRIC=DIAGONAL, then PROC DISCRIM uses either the diagonal matrix of the pooled covariance matrix (POOL=YES) or diagonal matrices of individual within-group covariance matrices (POOL=NO) to compute the squared distances. If you specify METRIC=IDENTITY, then PROC DISCRIM uses Euclidean distance. The default is METRIC=FULL. When you specify METHOD=NORMAL, the option METRIC=FULL is used.

NCAN=*number*

specifies the number of canonical variables to compute. The value of *number* must be less than or equal to the number of variables. If you specify the option NCAN=0, the procedure displays the canonical correlations but not the canonical coefficients, structures, or means. Let v be the number of variables in the VAR statement, and let c be the number of classes. If you omit the NCAN= option, only $\min(v, c-1)$ canonical variables are generated. If you request an output data set (OUT=, OUTCROSS=, TESTOUT=), v canonical variables are generated. In this case, the last $v-(c-1)$ canonical variables have missing values.

The CANONICAL option is activated when you specify either the NCAN= or the CANPREFIX= option. A discriminant criterion is always derived in PROC DISCRIM. If you want canonical discriminant analysis without the use of discriminant criterion, you should use PROC CANDISC.

NOCLASSIFY

suppresses the resubstitution classification of the input DATA= data set. You can specify this option only when the input data set is an ordinary SAS data set.

NOPRINT

suppresses the normal display of results. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 20, "Using the Output Delivery System," for more information.

OUT=*SAS-data-set*

creates an output SAS data set containing all the data from the DATA= data set, plus the posterior probabilities and the class into which each observation is classified by resubstitution. When you specify the CANONICAL option, the data set also contains new variables with canonical variable scores. See the section "OUT= Data Set" on page 2292 for more information.

OUTCROSS=*SAS-data-set*

creates an output SAS data set containing all the data from the DATA= data set, plus the posterior probabilities and the class into which each observation is classified by cross validation. When you specify the CANONICAL option, the data set also contains new variables with canonical variable scores. See the section "OUT= Data Set" on page 2292 for more information.

OUTD=*SAS-data-set*

creates an output SAS data set containing all the data from the DATA= data set, plus the group-specific density estimates for each observation. See the section "OUT= Data Set" on page 2292 for more information.

OUTSTAT=*SAS-data-set*

creates an output SAS data set containing various statistics such as means, standard deviations, and correlations. When the input data set is an ordinary SAS data set or when TYPE=CORR, TYPE=COV, TYPE=CSSCP, or TYPE=SSCP, this option can be used to generate discriminant statistics. When you specify the CANONICAL option, canonical correlations, canonical structures, canonical coefficients, and means of canonical variables for each class are included in the data set. If you specify METHOD=NORMAL, the output data set also includes coefficients of the discriminant functions, and the output data set is TYPE=LINEAR (POOL=YES), TYPE=QUAD (POOL=NO), or TYPE=MIXED (POOL=TEST). If you specify METHOD=NPAR, this output data set is TYPE=CORR. This data set also holds calibration information that can be used to classify new observations. See the sections "Saving and Using Calibration Information" on page 2289 and "OUT= Data Set" on page 2292 for more information.

PCORR

displays pooled within-class correlations.

PCOV

displays pooled within-class covariances.

POOL=NO | TEST | YES

determines whether the pooled or within-group covariance matrix is the basis of the measure of the squared distance. If you specify POOL=YES, then PROC DISCRIM uses the pooled covariance matrix in calculating the (generalized) squared distances. Linear discriminant functions are computed. If you specify POOL=NO, the procedure uses the individual within-group covariance matrices in calculating the distances. Quadratic discriminant functions are computed. The default is POOL=YES.

The k-nearest-neighbor method assumes the default of POOL=YES, and the POOL=TEST option cannot be used with the METHOD=NPAR option.

When you specify METHOD=NORMAL, the option POOL=TEST requests Bartlett's modification of the likelihood ratio test (Morrison 1976; Anderson 1984) of the homogeneity of the within-group covariance matrices. The test is unbiased (Perlman 1980). However, it is not robust to nonnormality. If the test statistic is significant at the level specified by the SLPOOL= option, the within-group covariance matrices are used. Otherwise, the pooled covariance matrix is used. The discriminant function coefficients are displayed only when the pooled covariance matrix is used.

POSTERR

displays the posterior probability error-rate estimates of the classification criterion based on the classification results.

PSSCP

displays the pooled within-class corrected SSCP matrix.

R=*r*

specifies a radius r value for kernel density estimation. With uniform, Epanechnikov, biweight, or triweight kernels, an observation $\mathbf{x}$ is classified into a group based on the information from observations $\mathbf{y}$ in the training set within the radius r of $\mathbf{x}$—that is, the group t observations $\mathbf{y}$ with squared distance $d_t^2(\mathbf{x},\mathbf{y}) \le r^2$. When a normal kernel is used, the classification of an observation $\mathbf{x}$ is based on the information of the estimated group-specific densities from all observations in the training set. The matrix $r^2\mathbf{V}_t$ is used as the group t covariance matrix in the normal-kernel density, where $\mathbf{V}_t$ is the matrix used in calculating the squared distances. Do not specify the K= or KPROP= option with the R= option. For more information about selecting r, see the section "Nonparametric Methods" on page 2282.

SCORES< = *prefix* >

computes and outputs discriminant scores to the OUT= and TESTOUT= data sets with the default options METHOD=NORMAL and POOL=YES (or with METHOD=NORMAL, POOL=TEST, and a nonsignificant chi-square test). Otherwise, or if no OUT= or TESTOUT= data set is specified, this option is ignored. The scores are computed by a matrix multiplication of an intercept term and the raw data or test data by the coefficients in the linear discriminant function. One score variable is created for each level of the CLASS variable. By default, the variables are named "Sc_" followed by the formatted class level. You can specify SCORES=prefix to use a prefix other than "Sc_". The specifications SCORES and SCORES=Sc_ are equivalent.

SHORT

suppresses the display of certain items in the default output. If you specify METHOD=NORMAL, then PROC DISCRIM suppresses the display of determinants, generalized squared distances between-class means, and discriminant function coefficients. When you specify the CANONICAL option, PROC DISCRIM suppresses the display of canonical structures, canonical coefficients, and class means on canonical variables; only tables of canonical correlations are displayed.

SIMPLE

displays simple descriptive statistics for the total sample and within each class.

SINGULAR=*p*

specifies the criterion for determining the singularity of a matrix, where $0 < p < 1$. The default is SINGULAR=1E–8.

Let **S** be the total-sample correlation matrix. If the R square for predicting a quantitative variable in the VAR statement from the variables preceding it exceeds $1 - p$, then **S** is considered singular. If **S** is singular, the probability levels for the multivariate test statistics and canonical correlations are adjusted for the number of variables with R square exceeding $1 - p$.

Let $\mathbf{S}_t$ be the group t covariance matrix, and let $\mathbf{S}_p$ be the pooled covariance matrix. In group t, if the R square for predicting a quantitative variable in the VAR statement from the variables preceding it exceeds $1 - p$, then $\mathbf{S}_t$ is considered singular. Similarly, if the partial R square for predicting a quantitative variable in the VAR statement from the variables preceding it, after controlling for the effect of the CLASS variable, exceeds $1 - p$, then $\mathbf{S}_p$ is considered singular.

If PROC DISCRIM needs to compute either the inverse or the determinant of a matrix that is considered singular, then it uses a quasi inverse or a quasi determinant. For details, see the section "Quasi-inverse" on page 2286.

SLPOOL=*p*

specifies the significance level for the test of homogeneity. You can specify the SLPOOL= option only when POOL=TEST is also specified. If you specify POOL= TEST but omit the SLPOOL= option, PROC DISCRIM uses 0.10 as the significance level for the test.

STDMEAN

displays total-sample and pooled within-class standardized class means.

TCORR

displays total-sample correlations.

TCOV

displays total-sample covariances.

TESTDATA=*SAS-data-set*

names an ordinary SAS data set with observations that are to be classified. The quantitative variable names in this data set must match those in the DATA= data set. When you specify the TESTDATA= option, you can also specify the TESTCLASS, TESTFREQ, and TESTID statements. When you specify the TESTDATA= option, you can use the TESTOUT= and TESTOUTD= options to generate classification results and group-specific density estimates for observations in the test data set. Note that if the CLASS variable is not present in the TESTDATA= data set, the output will not include misclassification statistics.

TESTLIST

lists classification results for all observations in the TESTDATA= data set.

TESTLISTERR

lists only misclassified observations in the TESTDATA= data set but only if a TESTCLASS statement is also used.

TESTOUT=*SAS-data-set*

creates an output SAS data set containing all the data from the TESTDATA= data set, plus the posterior probabilities and the class into which each observation is classified. When you specify the CANONICAL option, the data set also contains new variables with canonical variable scores. See the section "OUT= Data Set" on page 2292 for more information.

TESTOUTD=*SAS-data-set*

creates an output SAS data set containing all the data from the TESTDATA= data set, plus the group-specific density estimates for each observation. See the section "OUT= Data Set" on page 2292 for more information.

THRESHOLD=*p*

specifies the minimum acceptable posterior probability for classification, where $0 \leq p \leq 1$. If the largest posterior probability of group membership is less than the THRESHOLD value, the observation is labeled as `'Other'`. The default is THRESHOLD=0. In some cases, you might want to specify a THRESHOLD= value slightly smaller than the desired *p* so that observations with posterior probabilities within rounding error of *p* are classified. For example, you can specify **`threshold=%sysevalf(0.5 - 1e-8)`** instead of THRESHOLD=0.5 so that observations with posterior probabilities within 1E–8 of 0.5 and larger are classified.

TSSCP

displays the total-sample corrected SSCP matrix.

WCORR

displays within-class correlations for each class level.

WCOV

displays within-class covariances for each class level.

WSSCP

displays the within-class corrected SSCP matrix for each class level.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC DISCRIM to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement for the DISCRIM procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.

- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

If you specify the TESTDATA= option and the TESTDATA= data set does not contain any of the BY variables, then the entire TESTDATA= data set is classified according to the discriminant functions computed in each BY group in the DATA= data set.

If the TESTDATA= data set contains some but not all of the BY variables, or if some BY variables do not have the same type or length in the TESTDATA= data set as in the DATA= data set, then PROC DISCRIM displays an error message and stops.

If all BY variables appear in the TESTDATA= data set with the same type and length as in the DATA= data set, then each BY group in the TESTDATA= data set is classified by the discriminant function from the corresponding BY group in the DATA= data set. The BY groups in the TESTDATA= data set must be in the same order as in the DATA= data set. If you specify the NOTSORTED option in the BY statement, there must be exactly the same BY groups in the same order in both data sets. If you omit the NOTSORTED option, some BY groups can appear in one data set but not in the other. If some BY groups appear in the TESTDATA= data set but not in the DATA= data set, and you request an output test data set by using the TESTOUT= or TESTOUTD= option, these BY groups are not included in the output data set.

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

CLASS Statement

CLASS *variable* **;**

The values of the classification variable define the groups for analysis. Class levels are determined by the formatted values of the CLASS variable. The specified variable can be numeric or character. A CLASS statement is required.

FREQ Statement

FREQ *variable* **;**

If a variable in the data set represents the frequency of occurrence for the other values in the observation, include the variable's name in a FREQ statement. The procedure then treats the data set as if each observation appears n times, where n is the value of the FREQ variable for the observation. The total number of observations is considered to be equal to the sum of the FREQ variable when the procedure determines degrees of freedom for significance probabilities.

If the value of the FREQ variable is missing or is less than one, the observation is not used in the analysis. If the value is not an integer, it is truncated to an integer.

ID Statement

ID *variable* **;**

The ID statement is effective only when you specify the LIST or LISTERR option in the PROC DISCRIM statement. When the DISCRIM procedure displays the classification results, the ID variable (rather than the observation number) is displayed for each observation.

PRIORS Statement

PRIORS EQUAL ;

PRIORS PROPORTIONAL | PROP ;

PRIORS *probabilities* **;**

The PRIORS statement specifies the prior probabilities of group membership. To set the prior probabilities equal, use the following statement:

```
priors equal;
```

To set the prior probabilities proportional to the sample sizes, use the following statement:

```
priors proportional;
```

For other than equal or proportional priors, specify the prior probability for each level of the classification variable. Each class level can be written as either a SAS name or a quoted string, and it must be followed by an equal sign and a numeric constant between zero and one. A SAS name begins with a letter or an underscore and can contain digits as well. Lowercase character values and data values with leading blanks must be enclosed in quotes. For example, to define prior probabilities for each level of Grade, where Grade's values are `'A'`, `'B'`, `'C'`, and `'D'`, the PRIORS statement can be specified as follows:

```
priors A=0.1 B=0.3 C=0.5 D=0.1;
```

If Grade's values are `'a'`, `'b'`, `'c'`, and `'d'`, each class level must be written as a quoted string as follows:

```
priors 'a'=0.1  'b'=0.3  'c'=0.5  'd'=0.1;
```

If Grade is numeric, with formatted values of `'1'`, `'2'`, and `'3'`, the PRIORS statement can be written as follows:

```
priors '1'=0.3  '2'=0.6  '3'=0.1;
```

The specified class levels must exactly match the formatted values of the CLASS variable. For example, if a CLASS variable C has the format 4.2 and a value 5, the PRIORS statement must specify `'5.00'`, not `'5.0'` or `'5'`. If the prior probabilities do not sum to one, these probabilities are scaled proportionally to have the sum equal to one. The default is PRIORS EQUAL.

TESTCLASS Statement

TESTCLASS *variable* **;**

The TESTCLASS statement names the variable in the TESTDATA= data set that is used to determine whether an observation in the TESTDATA= data set is misclassified. The TESTCLASS variable should have the same type (character or numeric) and length as the variable given in the CLASS statement. PROC DISCRIM considers an observation misclassified when the formatted value of the TESTCLASS variable does not match the group into which the TESTDATA= observation is classified. When the TESTCLASS statement is missing and the TESTDATA= data set contains the variable given in the CLASS statement, the CLASS variable is used as the TESTCLASS variable. Note that if the CLASS variable is not present in the TESTDATA= data set, the output will not include misclassification statistics.

TESTFREQ Statement

TESTFREQ *variable* **;**

If a variable in the TESTDATA= data set represents the frequency of occurrence of the other values in the observation, include the variable's name in a TESTFREQ statement. The procedure then treats the data set as if each observation appears *n* times, where *n* is the value of the TESTFREQ variable for the observation.

If the value of the TESTFREQ variable is missing or is less than one, the observation is not used in the analysis. If the value is not an integer, it is truncated to an integer.

TESTID Statement

TESTID *variable* **;**

The TESTID statement is effective only when you specify the TESTLIST or TESTLISTERR option in the PROC DISCRIM statement. When the DISCRIM procedure displays the classification results for the TESTDATA= data set, the TESTID variable (rather than the observation number) is displayed for each observation. The variable given in the TESTID statement must be in the TESTDATA= data set.

VAR Statement

VAR *variables* **;**

The VAR statement specifies the quantitative variables to be included in the analysis. The default is all numeric variables not listed in other statements.

WEIGHT Statement

WEIGHT *variable* **;**

To use relative weights for each observation in the input data set, place the weights in a variable in the data set and specify the name in a WEIGHT statement. This is often done when the variance associated with each observation is different and the values of the weight variable are proportional to the reciprocals of the variances. If the value of the WEIGHT variable is missing or is less than zero, then a value of zero for the weight is used.

The WEIGHT and FREQ statements have a similar effect except that the WEIGHT statement does not alter the degrees of freedom.

Details: DISCRIM Procedure

Missing Values

Observations with missing values for variables in the analysis are excluded from the development of the classification criterion. When the values of the classification variable are missing, the observation is excluded from the development of the classification criterion, but if no other variables in the analysis have missing values for that observation, the observation is classified and displayed with the classification results.

Background

The following notation is used to describe the classification methods:

$\mathbf{x}$	a p-dimensional vector containing the quantitative variables of an observation
$\mathbf{S}_p$	the pooled covariance matrix
t	a subscript to distinguish the groups
n_t	the number of training set observations in group t
$\mathbf{m}_t$	the p-dimensional vector containing variable means in group t
$\mathbf{S}_t$	the covariance matrix within group t
$\lvert\mathbf{S}_t\rvert$	the determinant of $\mathbf{S}_t$
q_t	the prior probability of membership in group t
$p(t\lvert\mathbf{x})$	the posterior probability of an observation $\mathbf{x}$ belonging to group t
f_t	the probability density function for group t
$f_t(\mathbf{x})$	the group-specific density estimate at $\mathbf{x}$ from group t
$f(\mathbf{x})$	$\sum_t q_t f_t(\mathbf{x})$, the estimated unconditional density at $\mathbf{x}$
e_t	the classification error rate for group t

Bayes' Theorem

Assuming that the prior probabilities of group membership are known and that the group-specific densities at $\mathbf{x}$ can be estimated, PROC DISCRIM computes $p(t|\mathbf{x})$, the probability of $\mathbf{x}$ belonging to group t, by applying Bayes' theorem:

$$p(t|\mathbf{x}) = \frac{q_t f_t(\mathbf{x})}{f(\mathbf{x})}$$

PROC DISCRIM partitions a p-dimensional vector space into regions R_t, where the region R_t is the subspace containing all p-dimensional vectors $\mathbf{y}$ such that $p(t|\mathbf{y})$ is the largest among all groups. An observation is classified as coming from group t if it lies in region R_t.

Parametric Methods

Assuming that each group has a multivariate normal distribution, PROC DISCRIM develops a discriminant function or classification criterion by using a measure of generalized squared distance. The classification criterion is based on either the individual within-group covariance matrices or the pooled covariance matrix; it also takes into account the prior probabilities of the classes. Each observation is placed in the class from which it has the smallest generalized squared distance. PROC DISCRIM also computes the posterior probability of an observation belonging to each class.

The squared Mahalanobis distance from $\mathbf{x}$ to group t is

$$d_t^2(\mathbf{x}) = (\mathbf{x} - \mathbf{m}_t)' \mathbf{V}_t^{-1} (\mathbf{x} - \mathbf{m}_t)$$

where $\mathbf{V}_t = \mathbf{S}_t$ if the within-group covariance matrices are used, or $\mathbf{V}_t = \mathbf{S}_p$ if the pooled covariance matrix is used.

The group-specific density estimate at $\mathbf{x}$ from group t is then given by

$$f_t(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\mathbf{V}_t|^{-\frac{1}{2}} \exp\left(-0.5 d_t^2(\mathbf{x})\right)$$

Using Bayes' theorem, the posterior probability of $\mathbf{x}$ belonging to group t is

$$p(t|\mathbf{x}) = \frac{q_t f_t(\mathbf{x})}{\sum_u q_u f_u(\mathbf{x})}$$

where the summation is over all groups.

The generalized squared distance from $\mathbf{x}$ to group t is defined as

$$D_t^2(\mathbf{x}) = d_t^2(\mathbf{x}) + g_1(t) + g_2(t)$$

where

$$g_1(t) = \begin{cases} \ln |\mathbf{S}_t| & \text{if the within-group covariance matrices are used} \\ 0 & \text{if the pooled covariance matrix is used} \end{cases}$$

and

$$g_2(t) = \begin{cases} -2\ln(q_t) & \text{if the prior probabilities are not all equal} \\ 0 & \text{if the prior probabilities are all equal} \end{cases}$$

The posterior probability of $\mathbf{x}$ belonging to group t is then equal to

$$p(t|\mathbf{x}) = \frac{\exp\left(-0.5 D_t^2(\mathbf{x})\right)}{\sum_u \exp\left(-0.5 D_u^2(\mathbf{x})\right)}$$

The discriminant scores are $-0.5 D_u^2(\mathbf{x})$. An observation is classified into group u if setting $t = u$ produces the largest value of $p(t|\mathbf{x})$ or the smallest value of $D_t^2(\mathbf{x})$. If this largest posterior probability is less than the threshold specified, $\mathbf{x}$ is labeled as `'Other'`.

Nonparametric Methods

Nonparametric discriminant methods are based on nonparametric estimates of group-specific probability densities. Either a kernel method or the k-nearest-neighbor method can be used to generate a nonparametric density estimate in each group and to produce a classification criterion. The kernel method uses uniform, normal, Epanechnikov, biweight, or triweight kernels in the density estimation.

Either Mahalanobis distance or Euclidean distance can be used to determine proximity. When the k-nearest-neighbor method is used, the Mahalanobis distances are based on the pooled covariance matrix. When a kernel method is used, the Mahalanobis distances are based on either the individual within-group covariance matrices or the pooled covariance matrix. Either the full covariance matrix or the diagonal matrix of variances can be used to calculate the Mahalanobis distances.

The squared distance between two observation vectors, $\mathbf{x}$ and $\mathbf{y}$, in group t is given by

$$d_t^2(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})'\mathbf{V}_t^{-1}(\mathbf{x}-\mathbf{y})$$

where $\mathbf{V}_t$ has one of the following forms:

$$\mathbf{V}_t = \begin{cases} \mathbf{S}_p & \text{the pooled covariance matrix} \\ \text{diag}(\mathbf{S}_p) & \text{the diagonal matrix of the pooled covariance matrix} \\ \mathbf{S}_t & \text{the covariance matrix within group } t \\ \text{diag}(\mathbf{S}_t) & \text{the diagonal matrix of the covariance matrix within group } t \\ \mathbf{I} & \text{the identity matrix} \end{cases}$$

The classification of an observation vector $\mathbf{x}$ is based on the estimated group-specific densities from the training set. From these estimated densities, the posterior probabilities of group membership at $\mathbf{x}$ are evaluated. An observation $\mathbf{x}$ is classified into group u if setting $t = u$ produces the largest value of $p(t|\mathbf{x})$. If there is a tie for the largest probability or if this largest probability is less than the threshold specified, $\mathbf{x}$ is labeled as `'Other'`.

The kernel method uses a fixed radius, r, and a specified kernel, K_t, to estimate the group t density at each observation vector $\mathbf{x}$. Let $\mathbf{z}$ be a p-dimensional vector. Then the volume of a p-dimensional unit sphere bounded by $\mathbf{z}'\mathbf{z} = 1$ is

$$v_0 = \frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}+1\right)}$$

where Γ represents the gamma function (see *SAS Functions and CALL Routines: Reference*).

Thus, in group t, the volume of a p-dimensional ellipsoid bounded by $\{\mathbf{z} \mid \mathbf{z}'\mathbf{V}_t^{-1}\mathbf{z} = r^2\}$ is

$$v_r(t) = r^p|\mathbf{V}_t|^{\frac{1}{2}}v_0$$

The kernel method uses one of the following densities as the kernel density in group t:

Uniform Kernel

$$K_t(\mathbf{z}) = \begin{cases} \dfrac{1}{v_r(t)} & \text{if } \mathbf{z}'\mathbf{V}_t^{-1}\mathbf{z} \le r^2 \\ 0 & \text{elsewhere} \end{cases}$$

Normal Kernel (with mean zero, variance $r^2\mathbf{V}_t$)

$$K_t(\mathbf{z}) = \frac{1}{c_0(t)} \exp\left(-\frac{1}{2r^2}\mathbf{z}'\mathbf{V}_t^{-1}\mathbf{z}\right)$$

where $c_0(t) = (2\pi)^{\frac{p}{2}} r^p |\mathbf{V}_t|^{\frac{1}{2}}$.

Epanechnikov Kernel

$$K_t(\mathbf{z}) = \begin{cases} c_1(t)\left(1 - \frac{1}{r^2}\mathbf{z}'\mathbf{V}_t^{-1}\mathbf{z}\right) & \text{if } \mathbf{z}'\mathbf{V}_t^{-1}\mathbf{z} \le r^2 \\ 0 & \text{elsewhere} \end{cases}$$

where $c_1(t) = \frac{1}{v_r(t)}\left(1 + \frac{p}{2}\right)$.

Biweight Kernel

$$K_t(\mathbf{z}) = \begin{cases} c_2(t)\left(1 - \frac{1}{r^2}\mathbf{z}'\mathbf{V}_t^{-1}\mathbf{z}\right)^2 & \text{if } \mathbf{z}'\mathbf{V}_t^{-1}\mathbf{z} \le r^2 \\ 0 & \text{elsewhere} \end{cases}$$

where $c_2(t) = \left(1 + \frac{p}{4}\right) c_1(t)$.

Triweight Kernel

$$K_t(\mathbf{z}) = \begin{cases} c_3(t)\left(1 - \frac{1}{r^2}\mathbf{z}'\mathbf{V}_t^{-1}\mathbf{z}\right)^3 & \text{if } \mathbf{z}'\mathbf{V}_t^{-1}\mathbf{z} \le r^2 \\ 0 & \text{elsewhere} \end{cases}$$

where $c_3(t) = \left(1 + \frac{p}{6}\right) c_2(t)$.

The group t density at $\mathbf{x}$ is estimated by

$$f_t(\mathbf{x}) = \frac{1}{n_t} \sum_{\mathbf{y}} K_t(\mathbf{x} - \mathbf{y})$$

where the summation is over all observations $\mathbf{y}$ in group t, and K_t is the specified kernel function. The posterior probability of membership in group t is then given by

$$p(t|\mathbf{x}) = \frac{q_t f_t(\mathbf{x})}{f(\mathbf{x})}$$

where $f(\mathbf{x}) = \sum_u q_u f_u(\mathbf{x})$ is the estimated unconditional density. If $f(\mathbf{x})$ is zero, the observation $\mathbf{x}$ is labeled as `'Other'`.

The uniform-kernel method treats $K_t(\mathbf{z})$ as a multivariate uniform function with density uniformly distributed over $\mathbf{z}'\mathbf{V}_t^{-1}\mathbf{z} \le r^2$. Let k_t be the number of training set observations $\mathbf{y}$ from group t within the closed ellipsoid centered at $\mathbf{x}$ specified by $d_t^2(\mathbf{x},\mathbf{y}) \le r^2$. Then the group t density at $\mathbf{x}$ is estimated by

$$f_t(\mathbf{x}) = \frac{k_t}{n_t v_r(t)}$$

When the identity matrix or the pooled within-group covariance matrix is used in calculating the squared distance, $v_r(t)$ is a constant, independent of group membership. The posterior probability of $\mathbf{x}$ belonging to group t is then given by

$$p(t|\mathbf{x}) = \frac{\frac{q_t k_t}{n_t}}{\sum_u \frac{q_u k_u}{n_u}}$$

If the closed ellipsoid centered at $\mathbf{x}$ does not include any training set observations, $f(\mathbf{x})$ is zero and $\mathbf{x}$ is labeled as `'Other'`. When the prior probabilities are equal, $p(t|\mathbf{x})$ is proportional to k_t/n_t and $\mathbf{x}$ is classified into the group that has the highest proportion of observations in the closed ellipsoid. When the prior probabilities are proportional to the group sizes, $p(t|\mathbf{x}) = k_t/\sum_u k_u$, $\mathbf{x}$ is classified into the group that has the largest number of observations in the closed ellipsoid.

The nearest-neighbor method fixes the number, k, of training set points for each observation $\mathbf{x}$. The method finds the radius $r_k(\mathbf{x})$ that is the distance from $\mathbf{x}$ to the kth-nearest training set point in the metric $\mathbf{V}_t^{-1}$. Consider a closed ellipsoid centered at $\mathbf{x}$ bounded by $\{\mathbf{z} \mid (\mathbf{z}-\mathbf{x})'\mathbf{V}_t^{-1}(\mathbf{z}-\mathbf{x}) = r_k^2(\mathbf{x})\}$; the nearest-neighbor method is equivalent to the uniform-kernel method with a location-dependent radius $r_k(\mathbf{x})$. Note that, with ties, more than k training set points might be in the ellipsoid.

Using the k-nearest-neighbor rule, the k_n (or more with ties) smallest distances are saved. Of these k distances, let k_t represent the number of distances that are associated with group t. Then, as in the uniform-kernel method, the estimated group t density at $\mathbf{x}$ is

$$f_t(\mathbf{x}) = \frac{k_t}{n_t v_k(\mathbf{x})}$$

where $v_k(\mathbf{x})$ is the volume of the ellipsoid bounded by $\{\mathbf{z} \mid (\mathbf{z}-\mathbf{x})'\mathbf{V}_t^{-1}(\mathbf{z}-\mathbf{x}) = r_k^2(\mathbf{x})\}$. Since the pooled within-group covariance matrix is used to calculate the distances used in the nearest-neighbor method, the volume $v_k(\mathbf{x})$ is a constant independent of group membership. When $k = 1$ is used in the nearest-neighbor rule, $\mathbf{x}$ is classified into the group associated with the $\mathbf{y}$ point that yields the smallest squared distance $d_t^2(\mathbf{x},\mathbf{y})$. Prior probabilities affect nearest-neighbor results in the same way that they affect uniform-kernel results.

With a specified squared distance formula (METRIC=, POOL=), the values of r and k determine the degree of irregularity in the estimate of the density function, and they are called smoothing parameters. Small values of r or k produce jagged density estimates, and large values of r or k produce smoother density estimates. Various methods for choosing the smoothing parameters have been suggested, and there is as yet no simple solution to this problem.

For a fixed kernel shape, one way to choose the smoothing parameter r is to plot estimated densities with different values of r and to choose the estimate that is most in accordance with the prior information about the density. For many applications, this approach is satisfactory.

Another way of selecting the smoothing parameter r is to choose a value that optimizes a given criterion. Different groups might have different sets of optimal values. Assume that the unknown density has bounded and continuous second derivatives and that the kernel is a symmetric probability density function. One criterion is to minimize an approximate mean integrated square error of the estimated density (Rosenblatt 1956). The resulting optimal value of r depends on the density function and the kernel. A reasonable choice for the smoothing parameter r is to optimize the criterion with the assumption that group t has a normal distribution with covariance matrix $\mathbf{V}_t$. Then, in group t, the resulting optimal value for r is given by

$$\left(\frac{A(K_t)}{n_t}\right)^{1/(p+4)}$$

where the optimal constant $A(K_t)$ depends on the kernel K_t (Epanechnikov 1969). For some useful kernels, the constants $A(K_t)$ are given by the following:

$$A(K_t) = \frac{1}{p}2^{p+1}(p+2)\Gamma\left(\frac{p}{2}\right) \qquad \text{with a uniform kernel}$$

$$A(K_t) = \frac{4}{2p+1} \qquad \text{with a normal kernel}$$

$$A(K_t) = \frac{2^{p+2}p^2(p+2)(p+4)}{2p+1}\Gamma\left(\frac{p}{2}\right) \qquad \text{with an Epanechnikov kernel}$$

These selections of $A(K_t)$ are derived under the assumption that the data in each group are from a multivariate normal distribution with covariance matrix $\mathbf{V}_t$. However, when the Euclidean distances are used in calculating the squared distance ($\mathbf{V}_t = I$), the smoothing constant should be multiplied by s, where s is an estimate of standard deviations for all variables. A reasonable choice for s is

$$s = \left(\frac{1}{p}\sum s_{jj}\right)^{\frac{1}{2}}$$

where s_{jj} are group t marginal variances.

The DISCRIM procedure uses only a single smoothing parameter for all groups. However, the selection of the matrix in the distance formula (from the METRIC= or POOL= option), enables individual groups and variables to have different scalings. When $\mathbf{V}_t$, the matrix used in calculating the squared distances, is an identity matrix, the kernel estimate at each data point is scaled equally for all variables in all groups. When $\mathbf{V}_t$ is the diagonal matrix of a covariance matrix, each variable in group t is scaled separately by its variance in the kernel estimation, where the variance can be the pooled variance ($\mathbf{V}_t = \mathbf{S}_p$) or an individual within-group variance ($\mathbf{V}_t = \mathbf{S}_t$). When $\mathbf{V}_t$ is a full covariance matrix, the variables in group t are scaled simultaneously by $\mathbf{V}_t$ in the kernel estimation.

In nearest-neighbor methods, the choice of k is usually relatively uncritical (Hand 1982). A practical approach is to try several different values of the smoothing parameters within the context of the particular application and to choose the one that gives the best cross validated estimate of the error rate.

Classification Error-Rate Estimates

A classification criterion can be evaluated by its performance in the classification of future observations. PROC DISCRIM uses two types of error-rate estimates to evaluate the derived classification criterion based on parameters estimated by the training sample:

- error-count estimates
- posterior probability error-rate estimates

The error-count estimate is calculated by applying the classification criterion derived from the training sample to a test set and then counting the number of misclassified observations. The group-specific error-count estimate is the proportion of misclassified observations in the group. When the test set is independent of the training sample, the estimate is unbiased. However, the estimate can have a large variance, especially if the test set is small.

When the input data set is an ordinary SAS data set and no independent test sets are available, the same data set can be used both to define and to evaluate the classification criterion. The resulting error-count estimate has an optimistic bias and is called an *apparent error rate*. To reduce the bias, you can split the data into two sets—one set for deriving the discriminant function and the other set for estimating the error rate. Such a split-sample method has the unfortunate effect of reducing the effective sample size.

Another way to reduce bias is cross validation (Lachenbruch and Mickey 1968). Cross validation treats $n - 1$ out of n training observations as a training set. It determines the discriminant functions based on these $n - 1$ observations and then applies them to classify the one observation left out. This is done for each of the n training observations. The misclassification rate for each group is the proportion of sample observations in that group that are misclassified. This method achieves a nearly unbiased estimate but with a relatively large variance.

To reduce the variance in an error-count estimate, smoothed error-rate estimates are suggested (Glick 1978). Instead of summing terms that are either zero or one as in the error-count estimator, the smoothed estimator uses a continuum of values between zero and one in the terms that are summed. The resulting estimator has a smaller variance than the error-count estimate. The posterior probability error-rate estimates provided by the POSTERR option in the PROC DISCRIM statement (see the section "Posterior Probability Error-Rate Estimates" on page 2288) are smoothed error-rate estimates. The posterior probability estimates for each group are based on the posterior probabilities of the observations classified into that same group. The posterior probability estimates provide good estimates of the error rate when the posterior probabilities are accurate. When a parametric classification criterion (linear or quadratic discriminant function) is derived from a nonnormal population, the resulting posterior probability error-rate estimators might not be appropriate.

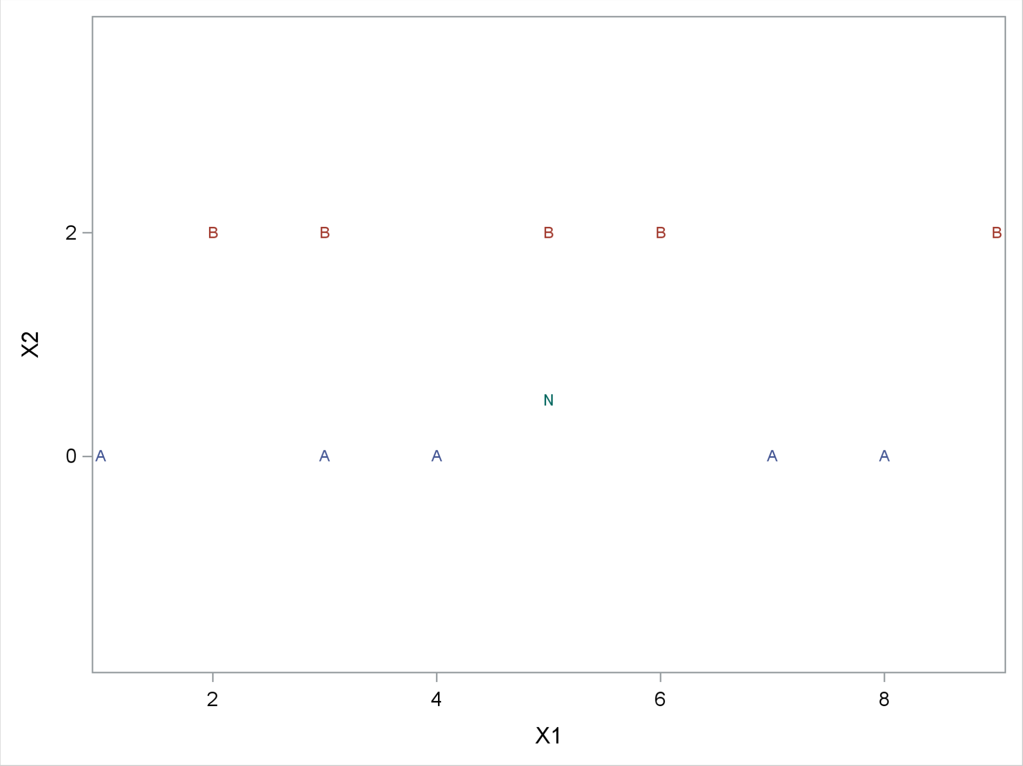
The overall error rate is estimated through a weighted average of the individual group-specific error-rate estimates, where the prior probabilities are used as the weights.

To reduce both the bias and the variance of the estimator, Hora and Wilcox (1982) compute the posterior probability estimates based on cross validation. The resulting estimates are intended to have both low variance from using the posterior probability estimate and low bias from cross validation. They use Monte Carlo studies on two-group multivariate normal distributions to compare the cross validation posterior probability estimates with three other estimators: the apparent error rate, cross validation estimator, and posterior probability estimator. They conclude that the cross validation posterior probability estimator has a lower mean squared error in their simulations.

Quasi-inverse

Consider the plot shown in Figure 36.6 with two variables, X1 and X2, and two classes, A and B. The within-class covariance matrix is diagonal, with a positive value for X1 but zero for X2. Using a Moore-Penrose pseudo-inverse would effectively ignore X2 in doing the classification, and the two classes would have a zero generalized distance and could not be discriminated at all. The quasi inverse used by PROC DISCRIM replaces the zero variance for X2 with a small positive number to remove the singularity. This permits X2 to be used in the discrimination and results correctly in a large generalized distance between the two classes and a zero error rate. It also permits new observations, such as the one indicated by N, to be classified in a reasonable way. PROC CANDISC also uses a quasi inverse when the total-sample covariance matrix is considered to be singular and Mahalanobis distances are requested. This problem with singular within-class covariance matrices is discussed in Ripley (1996, p. 38). The use of the quasi inverse is an innovation introduced by SAS.

Figure 36.6 Plot of Data with Singular Within-Class Covariance Matrix



Let **S** be a singular covariance matrix. The matrix **S** can be either a within-group covariance matrix, a pooled covariance matrix, or a total-sample covariance matrix. Let v be the number of variables in the VAR statement, and let the nullity n be the number of variables among them with (partial) R square exceeding $1 - p$. If the determinant of **S** (Testing of Homogeneity of Within Covariance Matrices) or the inverse of **S** (Squared Distances and Generalized Squared Distances) is required, a quasi determinant or quasi inverse is used instead. With raw data input, PROC DISCRIM scales each variable to unit total-sample variance before calculating this quasi inverse. The calculation is based on the spectral decomposition $\mathbf{S} = \mathbf{\Gamma\Lambda\Gamma}'$, where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues λ_j, $j = 1, \ldots, v$, where $\lambda_i \geq \lambda_j$ when $i < j$, and $\mathbf{\Gamma}$ is a matrix with the corresponding orthonormal eigenvectors of **S** as columns. When the nullity n is less than v, set $\lambda_j^0 = \lambda_j$ for $j = 1, \ldots, v - n$, and $\lambda_j^0 = p\bar{\lambda}$ for $j = v - n + 1, \ldots, v$, where

$$\bar{\lambda} = \frac{1}{v-n} \sum_{k=1}^{v-n} \lambda_k$$

When the nullity n is equal to v, set $\lambda_j^0 = p$, for $j = 1, \ldots, v$. A quasi determinant is then defined as the product of λ_j^0, $j = 1, \ldots, v$. Similarly, a quasi inverse is then defined as $\mathbf{S}^* = \mathbf{\Gamma}\mathbf{\Lambda}^*\mathbf{\Gamma}'$, where $\mathbf{\Lambda}^*$ is a diagonal matrix of values $1/\lambda_j^0$, $j = 1, \ldots, v$.

Posterior Probability Error-Rate Estimates

The posterior probability error-rate estimates (Fukunaga and Kessel 1973; Glick 1978; Hora and Wilcox 1982) for each group are based on the posterior probabilities of the observations classified into that same group.

A sample of observations with classification results can be used to estimate the posterior error rates. The following notation is used to describe the sample:

$\mathcal{S}$	the set of observations in the (training) sample
n	the number of observations in $\mathcal{S}$
n_t	the number of observations in $\mathcal{S}$ in group t
$\mathcal{R}_t$	the set of observations such that the posterior probability belonging to group t is the largest
$\mathcal{R}_{ut}$	the set of observations from group u such that the posterior probability belonging to group t is the largest

The classification error rate for group t is defined as

$$e_t = 1 - \int_{\mathcal{R}_t} f_t(\mathbf{x}) d\mathbf{x}$$

The posterior probability of $\mathbf{x}$ for group t can be written as

$$p(t|\mathbf{x}) = \frac{q_t f_t(\mathbf{x})}{f(\mathbf{x})}$$

where $f(\mathbf{x}) = \sum_u q_u f_u(\mathbf{x})$ is the unconditional density of $\mathbf{x}$.

Thus, if you replace $f_t(\mathbf{x})$ with $p(t|\mathbf{x}) f(\mathbf{x})/q_t$, the error rate is

$$e_t = 1 - \frac{1}{q_t} \int_{\mathcal{R}_t} p(t|\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

An estimator of e_t, unstratified over the groups from which the observations come, is then given by

$$\hat{e}_t \text{ (unstratified)} = 1 - \frac{1}{n q_t} \sum_{\mathcal{R}_t} p(t|\mathbf{x})$$

where $p(t|\mathbf{x})$ is estimated from the classification criterion, and the summation is over all sample observations of $\mathcal{S}$ classified into group t. The true group membership of each observation is not required in the estimation. The term $n q_t$ is the number of observations that are expected to be classified into group t, given the priors. If more observations than expected are classified into group t, then $\hat{e}_t$ can be negative.

Further, if you replace $f(\mathbf{x})$ with $\sum_u q_u f_u(\mathbf{x})$, the error rate can be written as

$$e_t = 1 - \frac{1}{q_t} \sum_u q_u \int_{\mathcal{R}_{ut}} p(t|\mathbf{x}) f_u(\mathbf{x}) d\mathbf{x}$$

and an estimator stratified over the group from which the observations come is given by

$$\hat{e}_t \text{ (stratified)} = 1 - \frac{1}{q_t} \sum_u q_u \frac{1}{n_u} \left(\sum_{\mathcal{R}_{ut}} p(t|\mathbf{x}) \right)$$

The inner summation is over all sample observations of $\mathcal{S}$ coming from group u and classified into group t, and n_u is the number of observations originally from group u. The stratified estimate uses only the observations with known group membership. When the prior probabilities of the group membership are proportional to the group sizes, the stratified estimate is the same as the unstratified estimator.

The estimated group-specific error rates can be less than zero, usually due to a large discrepancy between prior probabilities of group membership and group sizes. To have a reliable estimate for group-specific error rate estimates, you should use group sizes that are at least approximately proportional to the prior probabilities of group membership.

A total error rate is defined as a weighted average of the individual group error rates

$$e = \sum_t q_t e_t$$

and can be estimated from

$$\hat{e} \text{ (unstratified)} = \sum_t q_t \hat{e}_t \text{ (unstratified)}$$

or

$$\hat{e} \text{ (stratified)} = \sum_t q_t \hat{e}_t \text{ (stratified)}$$

The total unstratified error-rate estimate can also be written as

$$\hat{e} \text{ (unstratified)} = 1 - \frac{1}{n} \sum_t \sum_{\mathcal{R}_t} p(t|\mathbf{x})$$

which is one minus the average value of the maximum posterior probabilities for each observation in the sample. The prior probabilities of group membership do not appear explicitly in this overall estimate.

Saving and Using Calibration Information

When you specify METHOD=NORMAL to derive a linear or quadratic discriminant function, you can save the calibration information developed by the DISCRIM procedure in a SAS data set by using the OUTSTAT= option in the procedure. PROC DISCRIM then creates a specially structured SAS data set of TYPE=LINEAR, TYPE=QUAD, or TYPE=MIXED that contains the calibration information. For more information about these data sets, see Appendix A, "Special SAS Data Sets." Calibration information cannot be saved when METHOD=NPAR, but you can classify a TESTDATA= data set in the same step. For an example of this, see Example 36.1.

To use this calibration information to classify observations in another data set, specify both of the following:

- the name of the calibration data set after the DATA= option in the PROC DISCRIM statement
- the name of the data set to be classified after the TESTDATA= option in the PROC DISCRIM statement

Here is an example:

```
data original;
   input position x1 x2;
   datalines;
...[data lines]
;

proc discrim outstat=info;
   class position;
run;

data check;
   input position x1 x2;
   datalines;
...[second set of data lines]
;

proc discrim data=info testdata=check testlist;
   class position;
run;
```

The first DATA step creates the SAS data set Original, which the DISCRIM procedure uses to develop a classification criterion. Specifying OUTSTAT=INFO in the PROC DISCRIM statement causes the DISCRIM procedure to store the calibration information in a new data set called Info. The next DATA step creates the data set Check. The second PROC DISCRIM statement specifies DATA=INFO and TESTDATA=CHECK so that the classification criterion developed earlier is applied to the Check data set. Note that if the CLASS variable is not present in the TESTDATA= data set, the output will not include misclassification statistics.

Input Data Sets

DATA= Data Set

When you specify METHOD=NPAR, an ordinary SAS data set is required as the input DATA= data set. When you specify METHOD=NORMAL, the DATA= data set can be an ordinary SAS data set or one of several specially structured data sets created by SAS/STAT procedures. These specially structured data sets include the following:

- TYPE=CORR data sets created by PROC CORR by using a BY statement
- TYPE=COV data sets created by PROC PRINCOMP by using both the COV option and a BY statement
- TYPE=CSSCP data sets created by PROC CORR by using the CSSCP option and a BY statement, where the OUT= data set is assigned TYPE=CSSCP with the TYPE= data set option
- TYPE=SSCP data sets created by PROC REG by using both the OUTSSCP= option and a BY statement
- TYPE=LINEAR, TYPE=QUAD, and TYPE=MIXED data sets produced by previous runs of PROC DISCRIM that used both METHOD=NORMAL and OUTSTAT= options

When the input data set is TYPE=CORR, TYPE=COV, TYPE=CSSCP, or TYPE=SSCP, the BY variable in these data sets becomes the CLASS variable in the DISCRIM procedure.

When the input data set is TYPE=CORR, TYPE=COV, or TYPE=CSSCP, then PROC DISCRIM reads the number of observations for each class from the observations with _TYPE_=`'N'` and reads the variable means in each class from the observations with _TYPE_=`'MEAN'`. Then PROC DISCRIM reads the within-class correlations from the observations with _TYPE_=`'CORR'` and reads the standard deviations from the observations with _TYPE_=`'STD'` (data set TYPE=CORR), the within-class covariances from the observations with _TYPE_=`'COV'` (data set TYPE=COV), or the within-class corrected sums of squares and crossproducts from the observations with _TYPE_=`'CSSCP'` (data set TYPE=CSSCP).

When you specify POOL=YES and the data set does not include any observations with _TYPE_=`'CSSCP'` (data set TYPE=CSSCP), _TYPE_=`'COV'` (data set TYPE=COV), or _TYPE_=`'CORR'` (data set TYPE=CORR) for each class, PROC DISCRIM reads the pooled within-class information from the data set. In this case, PROC DISCRIM reads the pooled within-class covariances from the observations with _TYPE_=`'PCOV'` (data set TYPE=COV) or reads the pooled within-class correlations from the observations with _TYPE_=`'PCORR'` and the pooled within-class standard deviations from the observations with _TYPE_=`'PSTD'` (data set TYPE=CORR) or the pooled within-class corrected SSCP matrix from the observations with _TYPE_=`'PSSCP'` (data set TYPE=CSSCP).

When the input data set is TYPE=SSCP, the DISCRIM procedure reads the number of observations for each class from the observations with _TYPE_=`'N'`, the sum of weights of observations for each class from the variable INTERCEP in observations with _TYPE_=`'SSCP'` and _NAME_=`'INTERCEPT'`, the variable sums from the analysis variables in observations with _TYPE_=`'SSCP'` and _NAME_=`'INTERCEPT'`, and the uncorrected sums of squares and crossproducts from the analysis variables in observations with _TYPE_=`'SSCP'` and _NAME_=*variable-name*.

When the input data set is TYPE=LINEAR, TYPE=QUAD, or TYPE=MIXED, then PROC DISCRIM reads the prior probabilities for each class from the observations with variable _TYPE_=`'PRIOR'`.

When the input data set is TYPE=LINEAR, then PROC DISCRIM reads the coefficients of the linear discriminant functions from the observations with variable _TYPE_=`'LINEAR'`.

When the input data set is TYPE=QUAD, then PROC DISCRIM reads the coefficients of the quadratic discriminant functions from the observations with variable _TYPE_=`'QUAD'`.

When the input data set is TYPE=MIXED, then PROC DISCRIM reads the coefficients of the linear discriminant functions from the observations with variable _TYPE_=`'LINEAR'`. If there are no observations with _TYPE_=`'LINEAR'`, then PROC DISCRIM reads the coefficients of the quadratic discriminant functions from the observations with variable _TYPE_=`'QUAD'`.

TESTDATA= Data Set

The TESTDATA= data set is an ordinary SAS data set with observations that are to be classified. The quantitative variable names in this data set must match those in the DATA= data set. The TESTCLASS statement can be used to specify the variable containing group membership information of the TESTDATA= data set observations. When the TESTCLASS statement is missing and the TESTDATA= data set contains the variable given in the CLASS statement, this variable is used as the TESTCLASS variable. The TESTCLASS variable should have the same type (character or numeric) and length as the variable given in the CLASS statement. PROC DISCRIM considers an observation misclassified when the value of the TESTCLASS variable does not match the group into which the TESTDATA= observation is classified.

Output Data Sets

When an output data set includes variables containing the posterior probabilities of group membership (OUT=, OUTCROSS=, or TESTOUT= data sets) or group-specific density estimates (OUTD= or TESTOUTD= data sets), the names of these variables are constructed from the formatted values of the class levels converted to valid SAS variable names.

OUT= Data Set

The OUT= data set contains all the variables in the DATA= data set, plus new variables containing the posterior probabilities and the resubstitution classification results. The names of the new variables containing the posterior probabilities are constructed from the formatted values of the class levels converted to SAS names. A new variable, _INTO_, with the same attributes as the CLASS variable, specifies the class to which each observation is assigned. If an observation is labeled as `'Other'`, the variable _INTO_ has a missing value. When you specify the CANONICAL option, the data set also contains new variables with canonical variable scores. The NCAN= option determines the number of canonical variables. The names of the canonical variables are constructed as described in the CANPREFIX= option. The canonical variables have means equal to zero and pooled within-class variances equal to one.

An OUT= data set cannot be created if the DATA= data set is not an ordinary SAS data set.

OUTD= Data Set

The OUTD= data set contains all the variables in the DATA= data set, plus new variables containing the group-specific density estimates. The names of the new variables containing the density estimates are constructed from the formatted values of the class levels.

An OUTD= data set cannot be created if the DATA= data set is not an ordinary SAS data set.

OUTCROSS= Data Set

The OUTCROSS= data set contains all the variables in the DATA= data set, plus new variables containing the posterior probabilities and the classification results of cross validation. The names of the new variables containing the posterior probabilities are constructed from the formatted values of the class levels. A new variable, _INTO_, with the same attributes as the CLASS variable, specifies the class to which each observation is assigned. When an observation is labeled as `'Other'`, the variable _INTO_ has a missing value. When you specify the CANONICAL option, the data set also contains new variables with canonical variable scores. The NCAN= option determines the number of new variables. The names of the new variables are constructed as described in the CANPREFIX= option. The new variables have mean zero and pooled within-class variance equal to one.

An OUTCROSS= data set cannot be created if the DATA= data set is not an ordinary SAS data set.

TESTOUT= Data Set

The TESTOUT= data set contains all the variables in the TESTDATA= data set, plus new variables containing the posterior probabilities and the classification results. The names of the new variables containing the posterior probabilities are formed from the formatted values of the class levels. A new variable, _INTO_, with the same attributes as the CLASS variable, gives the class to which each observation is assigned. If an observation is labeled as `'Other'`, the variable _INTO_ has a missing value. When you specify the CANONICAL option, the data set also contains new variables with canonical variable scores. The NCAN= option determines the number of new variables. The names of the new variables are formed as described in the CANPREFIX= option.

TESTOUTD= Data Set

The TESTOUTD= data set contains all the variables in the TESTDATA= data set, plus new variables containing the group-specific density estimates. The names of the new variables containing the density estimates are formed from the formatted values of the class levels.

OUTSTAT= Data Set

The OUTSTAT= data set is similar to the TYPE=CORR data set produced by the CORR procedure. The data set contains various statistics such as means, standard deviations, and correlations. For an example of an OUTSTAT= data set, see Example 36.3. When you specify the CANONICAL option, canonical correlations, canonical structures, canonical coefficients, and means of canonical variables for each class are included in the data set.

If you specify METHOD=NORMAL, the output data set also includes coefficients of the discriminant functions, and the data set is TYPE=LINEAR (POOL=YES), TYPE=QUAD (POOL=NO), or TYPE=MIXED (POOL=TEST). If you specify METHOD=NPAR, this output data set is TYPE=CORR.

The OUTSTAT= data set contains the following variables:

- the BY variables, if any
- the CLASS variable
- _TYPE_, a character variable of length 8 that identifies the type of statistic
- _NAME_, a character variable of length 32 that identifies the row of the matrix, the name of the canonical variable, or the type of the discriminant function coefficients
- the quantitative variables—that is, those in the VAR statement, or, if there is no VAR statement, all numeric variables not listed in any other statement

The observations, as identified by the variable _TYPE_, have the following values:

TYPE	**Contents**
N	number of observations both for the total sample (CLASS variable missing) and within each class (CLASS variable present)
SUMWGT	sum of weights both for the total sample (CLASS variable missing) and within each class (CLASS variable present), if a WEIGHT statement is specified
MEAN	means both for the total sample (CLASS variable missing) and within each class (CLASS variable present)
PRIOR	prior probability for each class
STDMEAN	total-standardized class means
PSTDMEAN	pooled within-class standardized class means
STD	standard deviations both for the total sample (CLASS variable missing) and within each class (CLASS variable present)
PSTD	pooled within-class standard deviations
BSTD	between-class standard deviations
RSQUARED	univariate R squares
LNDETERM	the natural log of the determinant or the natural log of the quasi determinant of the within-class covariance matrix either pooled (CLASS variable missing) or not pooled (CLASS variable present)

The following kinds of observations are identified by the combination of the variables _TYPE_ and _NAME_. When the _TYPE_ variable has one of the following values, the _NAME_ variable identifies the row of the matrix:

TYPE	**Contents**
CSSCP	corrected SSCP matrix both for the total sample (CLASS variable missing) and within each class (CLASS variable present)
PSSCP	pooled within-class corrected SSCP matrix
BSSCP	between-class SSCP matrix
COV	covariance matrix both for the total sample (CLASS variable missing) and within each class (CLASS variable present)
PCOV	pooled within-class covariance matrix
BCOV	between-class covariance matrix
CORR	correlation matrix both for the total sample (CLASS variable missing) and within each class (CLASS variable present)
PCORR	pooled within-class correlation matrix
BCORR	between-class correlation matrix

When you request canonical discriminant analysis, the _NAME_ variable identifies a canonical variable, and _TYPE_ variable can have one of the following values:

TYPE	**Contents**
CANCORR	canonical correlations
STRUCTUR	canonical structure
BSTRUCT	between canonical structure
PSTRUCT	pooled within-class canonical structure
SCORE	standardized canonical coefficients
RAWSCORE	raw canonical coefficients
CANMEAN	means of the canonical variables for each class

When you specify METHOD=NORMAL, the _NAME_ variable identifies different types of coefficients in the discriminant function, and the _TYPE_ variable can have one of the following values:

TYPE	**Contents**
LINEAR	coefficients of the linear discriminant functions
QUAD	coefficients of the quadratic discriminant functions

The values of the _NAME_ variable are as follows:

NAME	**Contents**
variable names	quadratic coefficients of the quadratic discriminant functions (a symmetric matrix for each class)
LINEAR	linear coefficients of the discriminant functions
CONST	constant coefficients of the discriminant functions

Computational Resources

In the following discussion, let

n = number of observations in the training data set
v = number of variables
c = number of class levels
k = number of canonical variables
l = length of the CLASS variable

Memory Requirements

The amount of temporary storage required depends on the discriminant method used and the options specified. The least amount of temporary storage in bytes needed to process the data is approximately

$$c(32v + 3l + 128) + 8v^2 + 104v + 4l$$

A parametric method (METHOD=NORMAL) requires an additional temporary memory of $12v^2 + 100v$ bytes. When you specify the CROSSVALIDATE option, this temporary storage must be increased by $4v^2 + 44v$ bytes. When a nonparametric method (METHOD=NPAR) is used, an additional temporary storage of $10v^2 + 94v$ bytes is needed if you specify METRIC=FULL to evaluate the distances.

With the MANOVA option, the temporary storage must be increased by $8v^2 + 96v$ bytes. The CANONICAL option requires a temporary storage of $2v^2 + 94v + 8k(v + c)$ bytes. The POSTERR option requires a temporary storage of $8c^2 + 64c + 96$ bytes. Additional temporary storage is also required for classification summary and for each output data set.

Consider the following statements:

```
proc discrim manova;
   class gp;
   var x1 x2 x3;
run;
```

If the CLASS variable gp has a length of 8 and the input data set contains two class levels, the procedure requires a temporary storage of 1992 bytes. This includes 1104 bytes for processing data, 480 bytes for using a parametric method, and 408 bytes for specifying the MANOVA option.

Time Requirements

The following factors determine the time requirements of discriminant analysis:

- The time needed for reading the data and computing covariance matrices is proportional to nv^2. PROC DISCRIM must also look up each class level in the list. This is faster if the data are sorted by the CLASS variable. The time for looking up class levels is proportional to a value ranging from n to $n\ln(c)$.
- The time for inverting a covariance matrix is proportional to v^3.
- With a parametric method, the time required to classify each observation is proportional to cv for a linear discriminant function and cv^2 for a quadratic discriminant function. When you specify the CROSSVALIDATE option, the discriminant function is updated for each observation in the classification. A substantial amount of time is required.
- With a nonparametric method, the data are stored in a tree structure (Friedman, Bentley, and Finkel 1977). The time required to organize the observations into the tree structure is proportional to $nv\ln(n)$. The time for performing each tree search is proportional to $\ln(n)$. When you specify the normal KERNEL= option, all observations in the training sample contribute to the density estimation and more computer time is needed.
- The time required for the canonical discriminant analysis is proportional to v^3.

Each of the preceding factors has a different machine-dependent constant of proportionality.

Displayed Output

The displayed output from PROC DISCRIM includes the class level information table. For each level of the classification variable, the following information is provided: the output data set variable name, frequency sum, weight sum, proportion of the total sample, and prior probability.

The optional output from PROC DISCRIM includes the following:

- Within-class SSCP matrices for each group
- Pooled within-class SSCP matrix
- Between-class SSCP matrix
- Total-sample SSCP matrix
- Within-class covariance matrices, $\mathbf{S}_t$, for each group
- Pooled within-class covariance matrix, $\mathbf{S}_p$
- Between-class covariance matrix, equal to the between-class SSCP matrix divided by $n(c-1)/c$, where n is the number of observations and c is the number of classes
- Total-sample covariance matrix
- Within-class correlation coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the within-class population correlation coefficients are zero
- Pooled within-class correlation coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the partial population correlation coefficients are zero
- Between-class correlation coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the between-class population correlation coefficients are zero
- Total-sample correlation coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the total population correlation coefficients are zero
- Simple statistics, including N (the number of observations), sum, mean, variance, and standard deviation both for the total sample and within each class
- Total-sample standardized class means, obtained by subtracting the grand mean from each class mean and dividing by the total-sample standard deviation
- Pooled within-class standardized class means, obtained by subtracting the grand mean from each class mean and dividing by the pooled within-class standard deviation
- Pairwise squared distances between groups
- Univariate test statistics, including total-sample standard deviations, pooled within-class standard deviations, between-class standard deviations, R square, $R^2/(1-R^2)$, F, and $\mathrm{Pr} > F$ (univariate F values and probability levels for one-way analyses of variance)

- Multivariate statistics and F approximations, including Wilks' lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root with F approximations, numerator and denominator degrees of freedom (Num DF and Den DF), and probability values $(\text{Pr} > F)$. Each of these four multivariate statistics tests the hypothesis that the class means are equal in the population. See the section "Multivariate Tests" on page 88 in Chapter 4, "Introduction to Regression Procedures," for more information.

If you specify METHOD=NORMAL, the following three statistics are displayed:

- Covariance matrix information, including covariance matrix rank and natural log of determinant of the covariance matrix for each group (POOL=TEST, POOL=NO) and for the pooled within-group (POOL=TEST, POOL=YES)
- Optionally, test of homogeneity of within covariance matrices (the results of a chi-square test of homogeneity of the within-group covariance matrices) (Morrison 1976; Kendall, Stuart, and Ord 1983; Anderson 1984)
- Pairwise generalized squared distances between groups

If the CANONICAL option is specified, the displayed output contains these statistics:

- Canonical correlations
- Adjusted canonical correlations (Lawley 1959). These are asymptotically less biased than the raw correlations and can be negative. The adjusted canonical correlations might not be computable and are displayed as missing values if two canonical correlations are nearly equal or if some are close to zero. A missing value is also displayed if an adjusted canonical correlation is larger than a previous adjusted canonical correlation.
- Approximate standard error of the canonical correlations
- Squared canonical correlations
- Eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$. Each eigenvalue is equal to $\rho^2/(1-\rho^2)$, where ρ^2 is the corresponding squared canonical correlation and can be interpreted as the ratio of between-class variation to within-class variation for the corresponding canonical variable. The table includes eigenvalues, differences between successive eigenvalues, proportion of the sum of the eigenvalues, and cumulative proportion.
- Likelihood ratio for the hypothesis that the current canonical correlation and all smaller ones are zero in the population. The likelihood ratio for all canonical correlations equals Wilks' lambda.
- Approximate F statistic based on Rao's approximation to the distribution of the likelihood ratio (Rao 1973, p. 556; Kshirsagar 1972, p. 326)
- Numerator degrees of freedom (Num DF), denominator degrees of freedom (Den DF), and $\text{Pr} > F$, the probability level associated with the F statistic

The following statistic concerns the classification criterion:

- the linear discriminant function, but only if you specify METHOD=NORMAL and the pooled covariance matrix is used to calculate the (generalized) squared distances

When the input DATA= data set is an ordinary SAS data set, the displayed output includes the following:

- Optionally, the resubstitution results including the observation number (if an ID statement is included, the values of the ID variable are displayed instead of the observation number), the actual group for the observation, the group into which the developed criterion would classify it, and the posterior probability of membership in each group
- Resubstitution summary, a summary of the performance of the classification criterion based on resubstitution classification results
- Error count estimate of the resubstitution classification results
- Optionally, posterior probability error rate estimates of the resubstitution classification results

If you specify the CROSSVALIDATE option, the displayed output contains these statistics:

- Optionally, the cross validation results including the observation number (if an ID statement is included, the values of the ID variable are displayed instead of the observation number), the actual group for the observation, the group into which the developed criterion would classify it, and the posterior probability of membership in each group
- Cross validation summary, a summary of the performance of the classification criterion based on cross validation classification results
- Error count estimate of the cross validation classification results
- Optionally, posterior probability error rate estimates of the cross validation classification results

If you specify the TESTDATA= option, the displayed output contains these statistics:

- Optionally, the classification results including the observation number (if a TESTID statement is included, the values of the ID variable are displayed instead of the observation number), the actual group for the observation (if a TESTCLASS statement is included), the group into which the developed criterion would classify it, and the posterior probability of membership in each group
- Classification summary, a summary of the performance of the classification criterion
- Error count estimate of the test data classification results
- Optionally, posterior probability error rate estimates of the test data classification results

Formulas

By default, PROC DISCRIM displays formulas such as the following in the LISTING destination:

```
       2         _   _             -1   _   _
      D (i|j) = (X - X )' DIAG(COV )  (X - X )
                  i   j           j     i   j

                _                  _     2
               |       1        1   |  2P + 3P - 1
  RHO   = 1.0 - | SUM -----  -  ---  | -------------
               |_     N(i)      N  _|  6(P+1)(K-1)

 2         _       -1   _
D (X) = (X-X )' COV  (X-X ) + ln |COV | - 2 ln PRIOR
 j          j      j     j           j              j
```

These formulas are properly displayed only with uniform fonts. Formulas are not displayed by default in destinations such as HTML, RTF, and PDF, because destinations other than LISTING typically use proportional fonts. You can use the FORMULA option to display formulas in destinations other than LISTING.

In the LISTING destination, formulas are displayed as notes. In other destinations, formulas are displayed with batch capture (which provides uniform fonts). Therefore, the names that are used to select and exclude formulas from the LISTING destination are different from the names that are used with other destinations.

By default, batch capture results are displayed in a box. You can remove the box and change the color and font of the formulas as follows:

```
proc template;
   define style styles.batch; parent=styles.htmlblue;
      style fonts from fonts / "BatchFixedFont" =
         ("SAS Monospace, <monospace>, Courier, monospace", 2, bold);
      style batch from batch / frame=void foreground=colors('notefg');
   end;
run;

ods listing;
ods html style=batch body='dis.html';
ods rtf  style=batch body='dis.rtf';
ods pdf  style=batch body='dis.pdf';

proc discrim data=sashelp.iris formula;
   class Species;
   var Peta:;
run;

ods _all_ close;
```

The results of these steps are not displayed.

ODS Table Names

PROC DISCRIM assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in Table 36.2. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

Table 36.2 ODS Tables Produced by PROC DISCRIM

ODS Table Name	Description	PROC DISCRIM Option
ANOVA	Univariate statistics	ANOVA
AvePostCrossVal	Average posterior probabilities, cross validation	POSTERR & CROSSVALIDATE
AvePostResub	Average posterior probabilities, resubstitution	POSTERR
AvePostTestClass	Average posterior probabilities, test classification	POSTERR & TEST=
AveRSquare	Average R-square	ANOVA
BCorr	Between-class correlations	BCORR
BCov	Between-class covariances	BCOV
BSSCP	Between-class SSCP matrix	BSSCP
BStruc	Between canonical structure	CANONICAL
CanCorr	Canonical correlations	CANONICAL
CanonicalMeans	Class means on canonical variables	CANONICAL
ChiSq	Chi-square information	POOL=TEST
ClassifiedCrossVal	Number of observations and percent classified, cross validation	CROSSVALIDATE
ClassifiedResub	Number of observations and percent classified, resubstitution	default
ClassifiedTestClass	Number of observations and percent classified, test classification	TEST=
Counts	Number of observations, variables, classes, df	default
CovDF	DF for covariance matrices, not displayed	any *COV option
Dist	Squared distances	DISTANCE
DistFValues	*F* values based on squared distances	DISTANCE
DistGeneralized	Generalized squared distances	default
DistProb	Probabilities for *F* values from squared distances	DISTANCE
ErrorCrossVal	Error count estimates, cross validation	CROSSVALIDATE
ErrorResub	Error count estimates, resubstitution	default
ErrorTestClass	Error count estimates, test classification	TEST=

Table 36.2 *continued*

ODS Table Name	Description	PROC DISCRIM Option
Levels	Class level information	default
LinearDiscFunc	Linear discriminant function	POOL=YES
LogDet	Log determinant of the covariance matrix	default
MultStat	MANOVA	MANOVA
PCoef	Pooled standard canonical coefficients	CANONICAL
PCorr	Pooled within-class correlations	PCORR
PCov	Pooled within-class covariances	PCOV
PSSCP	Pooled within-class SSCP matrix	PSSCP
PStdMeans	Pooled standardized class means	STDMEAN
PStruc	Pooled within canonical structure	CANONICAL
PostCrossVal	Posterior probabilities, cross validation	CROSSLIST or CROSSLISTERR
PostErrCrossVal	Posterior error estimates, cross validation	POSTERR & CROSSVALIDATE
PostErrResub	Posterior error estimates, resubstitution	POSTERR
PostErrTestClass	Posterior error estimates, test classification	POSTERR & TEST=
PostResub	Posterior probabilities, resubstitution	LIST or LISTERR
PostTestClass	Posterior probabilities, test classification	TESTLIST or TESTLISTERR
RCoef	Raw canonical coefficients	CANONICAL
SimpleStatistics	Simple statistics	SIMPLE
TCoef	Total-sample standard canonical coefficients	CANONICAL
TCorr	Total-sample correlations	TCORR
TCov	Total-sample covariances	TCOV
TSSCP	Total-sample SSCP matrix	TSSCP
TStdMeans	Total standardized class means	STDMEAN
TStruc	Total canonical structure	CANONICAL
WCorr	Within-class correlations	WCORR
WCov	Within-class covariances	WCOV
WSSCP	Within-class SSCP matrices	WSSCP

Examples: DISCRIM Procedure

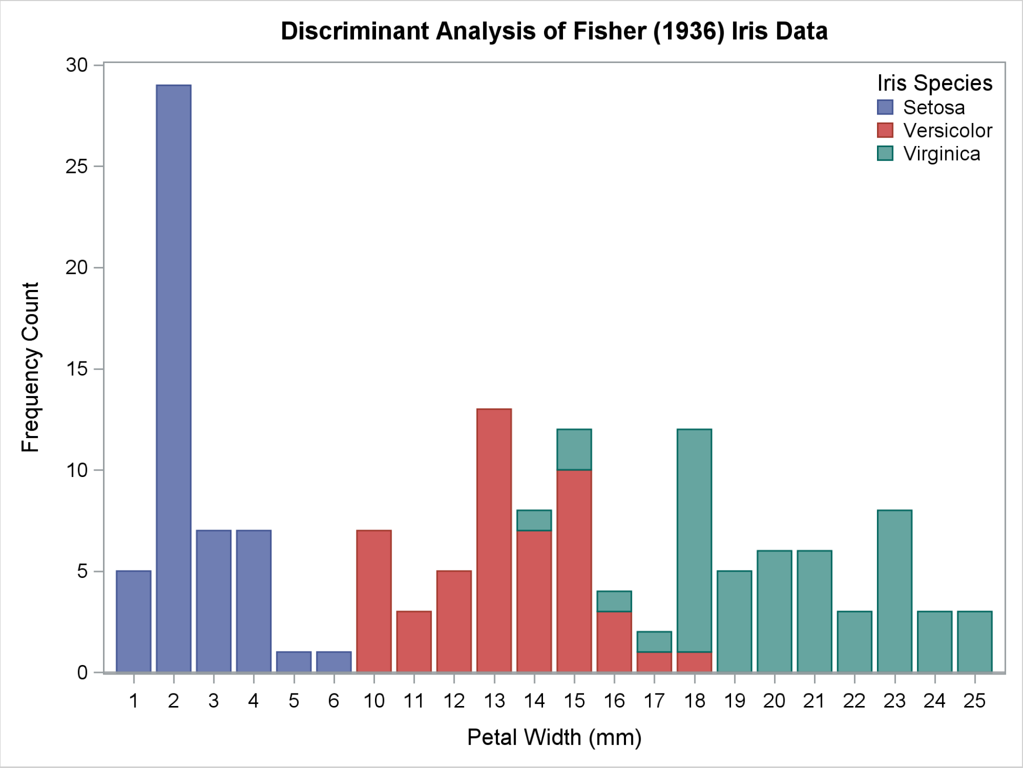
The iris data published by Fisher (1936) are widely used for examples in discriminant analysis and cluster analysis. The sepal length, sepal width, petal length, and petal width are measured in millimeters on 50 iris specimens from each of three species: *Iris setosa*, *I. versicolor*, and *I. virginica*. The iris data set is available from the Sashelp library.

Example 36.1: Univariate Density Estimates and Posterior Probabilities

In this example, several discriminant analyses are run with a single quantitative variable, petal width, so that density estimates and posterior probabilities can be plotted easily. The example produces Output 36.1.1 through Output 36.1.5. ODS Graphics is used to display the sample distribution of petal width in the three species. For general information about ODS Graphics, see Chapter 21, “Statistical Graphics Using ODS.” Note the overlap between the species *I. versicolor* and *I. virginica* that the bar chart shows. The following statements produce Output 36.1.1:

```
title 'Discriminant Analysis of Fisher (1936) Iris Data';

proc freq data=sashelp.iris noprint;
   tables petalwidth * species / out=freqout;
run;

proc sgplot data=freqout;
   vbar petalwidth / response=count group=species;
   keylegend / location=inside position=ne noborder across=1;
run;
```

Output 36.1.1 Sample Distribution of Petal Width in Three Species



In order to plot the density estimates and posterior probabilities, a data set called plotdata is created containing equally spaced values from –5 to 30, covering the range of petal width with a little to spare on each end. The plotdata data set is used with the TESTDATA= option in PROC DISCRIM. The following statements make the data set:

```
data plotdata;
   do PetalWidth=-5 to 30 by 0.5;
      output;
   end;
run;
```

The same plots are produced after each discriminant analysis, so macros are used to reduce the amount of typing required. The macros use two data sets. The data set plotd, containing density estimates, is created by the TESTOUTD= option in PROC DISCRIM. The data set plotp, containing posterior probabilities, is created by the TESTOUT= option. For each data set, the macros remove uninteresting values (near zero) and create an overlay plot showing all three species in a single plot.

The following statements create the macros:

```
%macro plotden;
   title3 'Plot of Estimated Densities';

   data plotd2;
      set plotd;
      if setosa     < .002 then setosa     = .;
      if versicolor < .002 then versicolor = .;
      if virginica  < .002 then virginica  = .;
      g = 'Setosa    '; Density = setosa;     output;
      g = 'Versicolor'; Density = versicolor; output;
      g = 'Virginica '; Density = virginica;  output;
      label PetalWidth='Petal Width in mm.';
   run;

   proc sgplot data=plotd2;
      series y=Density x=PetalWidth / group=g;
      discretelegend;
   run;
%mend;

%macro plotprob;
   title3 'Plot of Posterior Probabilities';

   data plotp2;
      set plotp;
      if setosa     < .01 then setosa     = .;
      if versicolor < .01 then versicolor = .;
      if virginica  < .01 then virginica  = .;
      g = 'Setosa    '; Probability = setosa;     output;
      g = 'Versicolor'; Probability = versicolor; output;
      g = 'Virginica '; Probability = virginica;  output;
      label PetalWidth='Petal Width in mm.';
   run;

   proc sgplot data=plotp2;
      series y=Probability x=PetalWidth / group=g;
      discretelegend;
   run;
%mend;
```

The first analysis uses normal-theory methods (METHOD=NORMAL) assuming equal variances (POOL=YES) in the three classes. The NOCLASSIFY option suppresses the resubstitution classification results of the input data set observations. The CROSSLISTERR option lists the observations that are misclassified under cross validation and displays cross validation error-rate estimates. The following statements produce Output 36.1.2:

```
title2 'Using Normal Density Estimates with Equal Variance';

proc discrim data=sashelp.iris method=normal pool=yes
   testdata=plotdata testout=plotp testoutd=plotd
   short noclassify crosslisterr;
   class Species;
   var PetalWidth;
run;

%plotden;
%plotprob;
```

Output 36.1.2 Normal Density Estimates with Equal Variance

Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Equal Variance

The DISCRIM Procedure

Total Sample Size	150	**DF Total**	149
Variables	1	**DF Within Classes**	147
Classes	3	**DF Between Classes**	2

Number of Observations Read	150
Number of Observations Used	150

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion	Prior Probability
Setosa	Setosa	50	50.0000	0.333333	0.333333
Versicolor	Versicolor	50	50.0000	0.333333	0.333333
Virginica	Virginica	50	50.0000	0.333333	0.333333

Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Equal Variance

The DISCRIM Procedure
Classification Results for Calibration Data: SASHELP.IRIS
Cross-validation Results using Linear Discriminant Function

Posterior Probability of Membership in Species

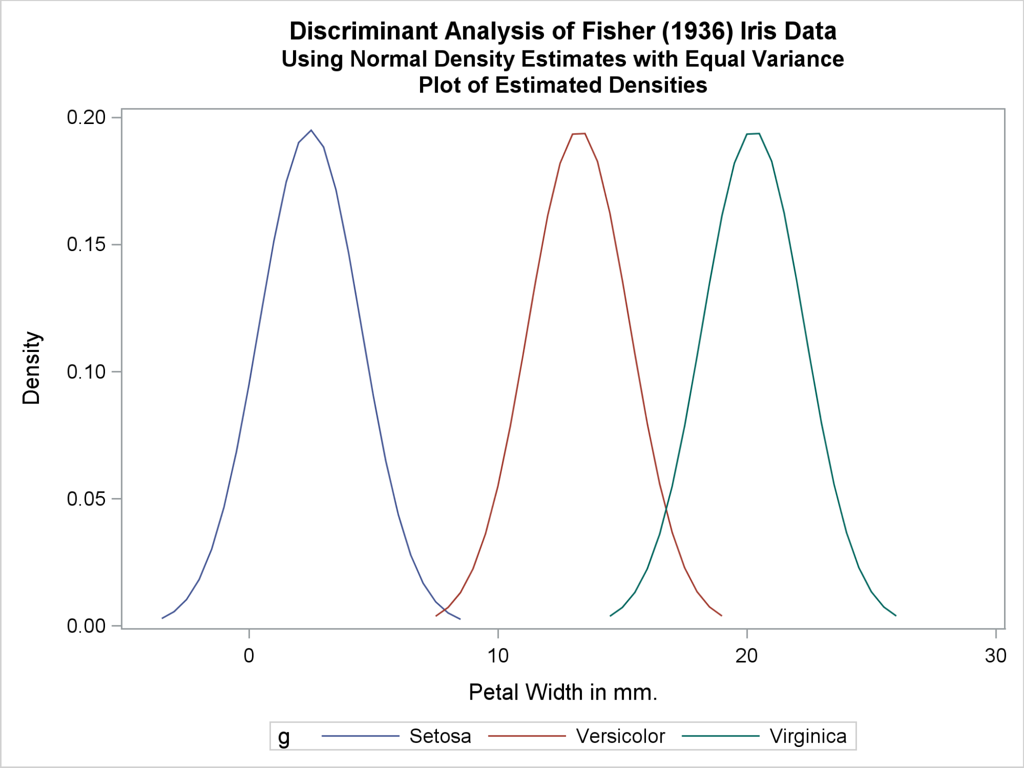
Obs	From Species	Classified into Species		Setosa	Versicolor	Virginica
53	**Versicolor**	**Virginica**	*	0.0000	0.0952	0.9048
100	**Versicolor**	**Virginica**	*	0.0000	0.3828	0.6172
103	**Virginica**	**Versicolor**	*	0.0000	0.9610	0.0390
124	**Virginica**	**Versicolor**	*	0.0000	0.9940	0.0060
130	**Virginica**	**Versicolor**	*	0.0000	0.8009	0.1991
136	**Virginica**	**Versicolor**	*	0.0000	0.9610	0.0390

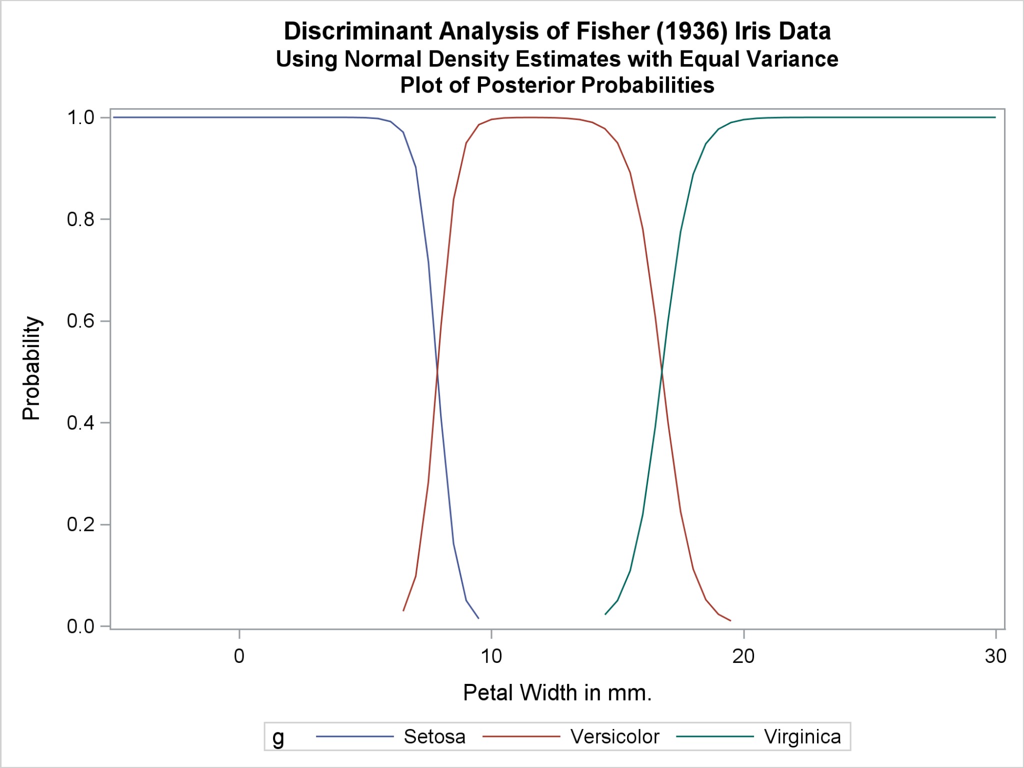
*** Misclassified observation**

Output 36.1.2 *continued*

```
            Discriminant  Analysis  of Fisher  (1936)  Iris Data
           Using  Normal Density  Estimates  with  Equal  Variance

                          The DISCRIM Procedure
             Classification  Summary  for Calibration  Data: SASHELP.IRIS
           Cross-validation  Summary  using Linear  Discriminant  Function

          Number of Observations and Percent Classified
                          into Species
          From
          Species       Setosa  Versicolor  Virginica    Total
          Setosa            50           0          0       50
                        100.00        0.00       0.00   100.00
          Versicolor         0          48          2       50
                          0.00       96.00       4.00   100.00
          Virginica          0           4         46       50
                          0.00        8.00      92.00   100.00
          Total             50          52         48      150
                         33.33       34.67      32.00   100.00
          Priors       0.33333     0.33333    0.33333

                 Error Count Estimates for Species
                   Setosa  Versicolor  Virginica   Total
          Rate     0.0000      0.0400     0.0800  0.0400
          Priors   0.3333      0.3333     0.3333

            Discriminant  Analysis  of Fisher  (1936)  Iris Data
           Using  Normal Density  Estimates  with  Equal  Variance

                          The DISCRIM Procedure
             Classification  Summary  for Test Data: WORK.PLOTDATA
           Classification  Summary  using Linear  Discriminant  Function

                  Observation Profile for Test Data
                  Number of Observations Read       71
                  Number of Observations Used       71

                 Number of Observations and Percent
                     Classified into Species
                     Setosa  Versicolor  Virginica    Total
          Total          26          18         27       71
                      36.62       25.35      38.03   100.00
          Priors    0.33333     0.33333    0.33333
```

Output 36.1.2 *continued*





The next analysis uses normal-theory methods assuming unequal variances (POOL=NO) in the three classes.

The following statements produce Output 36.1.3:

```
title2 'Using Normal Density Estimates with Unequal Variance';

proc discrim data=sashelp.iris method=normal pool=no
   testdata=plotdata testout=plotp testoutd=plotd
   short noclassify crosslisterr;
   class Species;
   var PetalWidth;
run;

%plotden;
%plotprob;
```

Output 36.1.3 Normal Density Estimates with Unequal Variance

Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Unequal Variance

The DISCRIM Procedure

Total Sample Size	150	**DF Total**	149
Variables	1	**DF Within Classes**	147
Classes	3	**DF Between Classes**	2

Number of Observations Read	150
Number of Observations Used	150

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion	Prior Probability
Setosa	Setosa	50	50.0000	0.333333	0.333333
Versicolor	Versicolor	50	50.0000	0.333333	0.333333
Virginica	Virginica	50	50.0000	0.333333	0.333333

Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Unequal Variance

The DISCRIM Procedure
Classification Results for Calibration Data: SASHELP.IRIS
Cross-validation Results using Quadratic Discriminant Function

Posterior Probability of Membership in Species

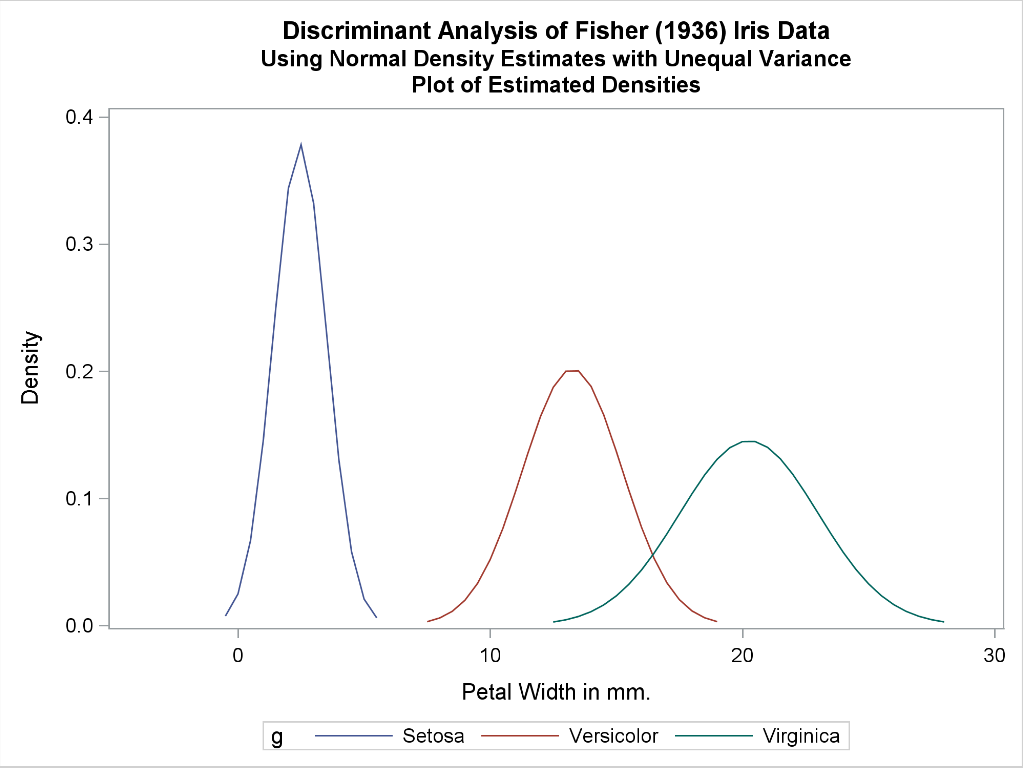
Obs	From Species	Classified into Species		Setosa	Versicolor	Virginica
10	**Setosa**	**Versicolor**	*	0.4923	0.5073	0.0004
53	**Versicolor**	**Virginica**	*	0.0000	0.0686	0.9314
100	**Versicolor**	**Virginica**	*	0.0000	0.2871	0.7129
103	**Virginica**	**Versicolor**	*	0.0000	0.8740	0.1260
124	**Virginica**	**Versicolor**	*	0.0000	0.9602	0.0398
130	**Virginica**	**Versicolor**	*	0.0000	0.6558	0.3442
136	**Virginica**	**Versicolor**	*	0.0000	0.8740	0.1260

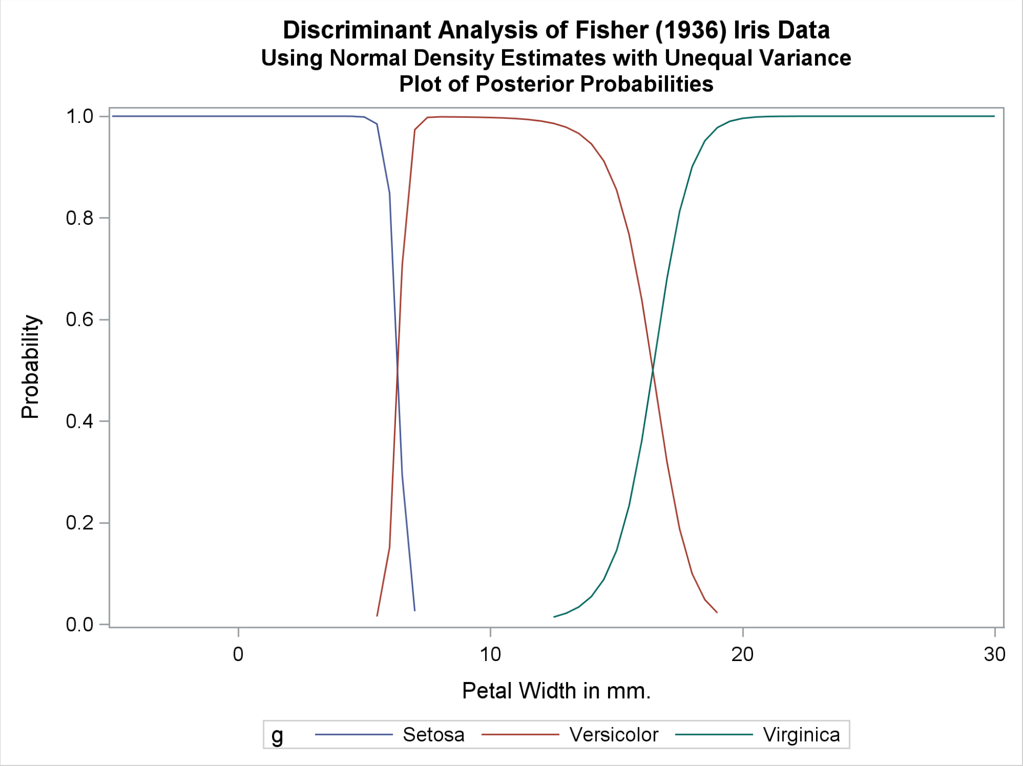
*** Misclassified observation**

Output 36.1.3 *continued*

```
           Discriminant  Analysis  of Fisher  (1936)  Iris Data
         Using  Normal Density  Estimates  with  Unequal  Variance

                          The  DISCRIM Procedure
            Classification  Summary  for Calibration  Data: SASHELP.IRIS
        Cross-validation  Summary  using Quadratic  Discriminant  Function

           Number of Observations and Percent Classified
                            into Species

           From
           Species      Setosa  Versicolor  Virginica   Total

           Setosa           49           1          0      50
                         98.00        2.00       0.00  100.00

           Versicolor        0          48          2      50
                          0.00       96.00       4.00  100.00

           Virginica         0           4         46      50
                          0.00        8.00      92.00  100.00

           Total            49          53         48     150
                         32.67       35.33      32.00  100.00

           Priors      0.33333     0.33333    0.33333


                  Error Count Estimates for Species

                    Setosa  Versicolor  Virginica   Total

             Rate   0.0200      0.0400     0.0800  0.0467

             Priors 0.3333      0.3333     0.3333


           Discriminant  Analysis  of Fisher  (1936)  Iris Data
         Using  Normal Density  Estimates  with  Unequal  Variance

                          The  DISCRIM Procedure
             Classification  Summary  for Test  Data: WORK.PLOTDATA
          Classification  Summary  using Quadratic  Discriminant  Function

                    Observation Profile for Test Data

                    Number of Observations Read      71
                    Number of Observations Used      71


                 Number of Observations and Percent
                      Classified into Species

                    Setosa  Versicolor  Virginica   Total

             Total      23          20         28      71
                     32.39       28.17      39.44  100.00

             Priors 0.33333    0.33333    0.33333
```

Output 36.1.3 *continued*





Two more analyses are run with nonparametric methods (METHOD=NPAR), specifically kernel density estimates with normal kernels (KERNEL=NORMAL). The first of these uses equal bandwidths (smoothing parameters) (POOL=YES) in each class. The use of equal bandwidths does not constrain the density estimates to be of equal variance. The value of the radius parameter that, assuming normality, minimizes an approximate mean integrated square error is 0.48 (see the section "Nonparametric Methods" on page 2282). Choosing $r = 0.4$ gives a more detailed look at the irregularities in the data. The following statements produce Output 36.1.4:

```
title2 'Using Kernel Density Estimates with Equal Bandwidth';

proc discrim data=sashelp.iris method=npar kernel=normal
   r=.4 pool=yes testdata=plotdata testout=plotp
   testoutd=plotd short noclassify crosslisterr;
   class Species;
   var PetalWidth;
run;

%plotden;
%plotprob;
```

Output 36.1.4 Kernel Density Estimates with Equal Bandwidth

```
            Discriminant Analysis of Fisher (1936) Iris Data
           Using Kernel Density Estimates with Equal Bandwidth

                         The DISCRIM Procedure

               Total Sample Size  150  DF Total              149
               Variables            1  DF Within Classes     147
               Classes              3  DF Between Classes      2

                   Number of Observations Read  150
                   Number of Observations Used  150

                          Class Level Information

             Variable                                               Prior
Species      Name         Frequency   Weight    Proportion    Probability
Setosa       Setosa              50  50.0000      0.333333       0.333333
Versicolor   Versicolor          50  50.0000      0.333333       0.333333
Virginica    Virginica           50  50.0000      0.333333       0.333333
```

Output 36.1.4 *continued*

Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Equal Bandwidth

The DISCRIM Procedure
Classification Results for Calibration Data: SASHELP.IRIS
Cross-validation Results using Normal Kernel Density

Posterior Probability of Membership in Species

Obs	From Species	Classified into Species		Setosa	Versicolor	Virginica
53	**Versicolor**	**Virginica**	*	0.0000	0.0438	0.9562
100	**Versicolor**	**Virginica**	*	0.0000	0.2586	0.7414
103	**Virginica**	**Versicolor**	*	0.0000	0.8827	0.1173
124	**Virginica**	**Versicolor**	*	0.0000	0.9472	0.0528
130	**Virginica**	**Versicolor**	*	0.0000	0.8061	0.1939
136	**Virginica**	**Versicolor**	*	0.0000	0.8827	0.1173

*** Misclassified observation**

Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Equal Bandwidth

The DISCRIM Procedure
Classification Summary for Calibration Data: SASHELP.IRIS
Cross-validation Summary using Normal Kernel Density

Number of Observations and Percent Classified into Species

From Species	Setosa	Versicolor	Virginica	Total
Setosa	50 100.00	0 0.00	0 0.00	50 100.00
Versicolor	0 0.00	48 96.00	2 4.00	50 100.00
Virginica	0 0.00	4 8.00	46 92.00	50 100.00
Total	50 33.33	52 34.67	48 32.00	150 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Species

	Setosa	Versicolor	Virginica	Total
Rate	0.0000	0.0400	0.0800	0.0400
Priors	0.3333	0.3333	0.3333	

Output 36.1.4 *continued*

Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Equal Bandwidth

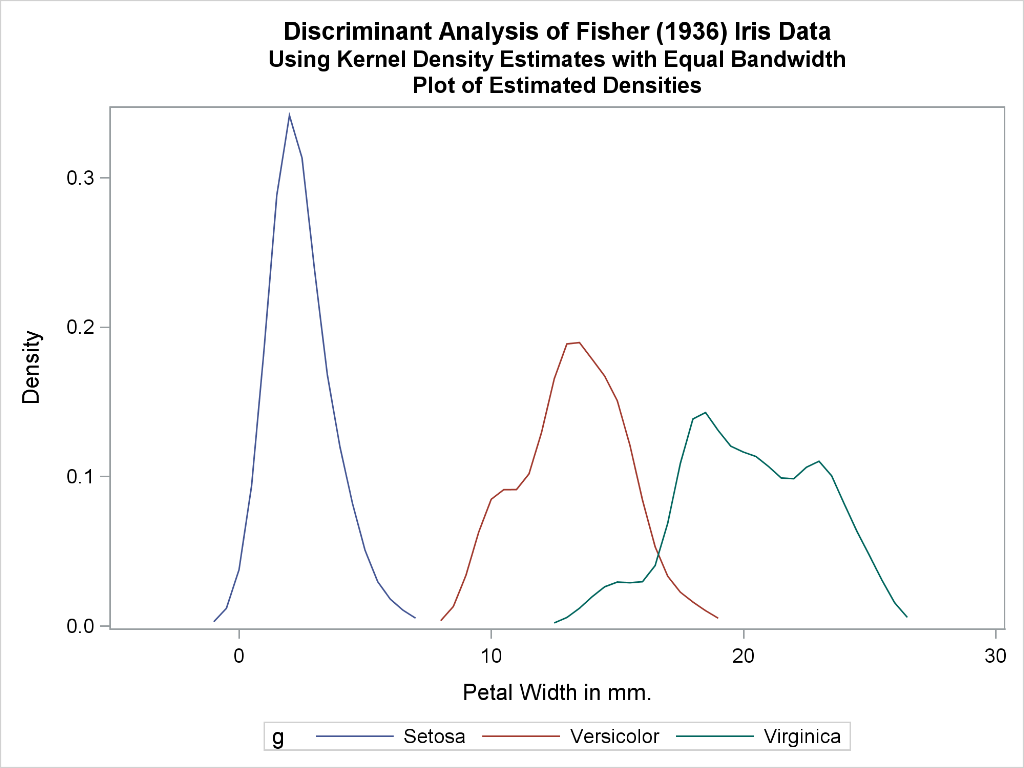
The DISCRIM Procedure
Classification Summary for Test Data: WORK.PLOTDATA
Classification Summary using Normal Kernel Density

Observation Profile for Test Data	
Number of Observations Read	71
Number of Observations Used	71

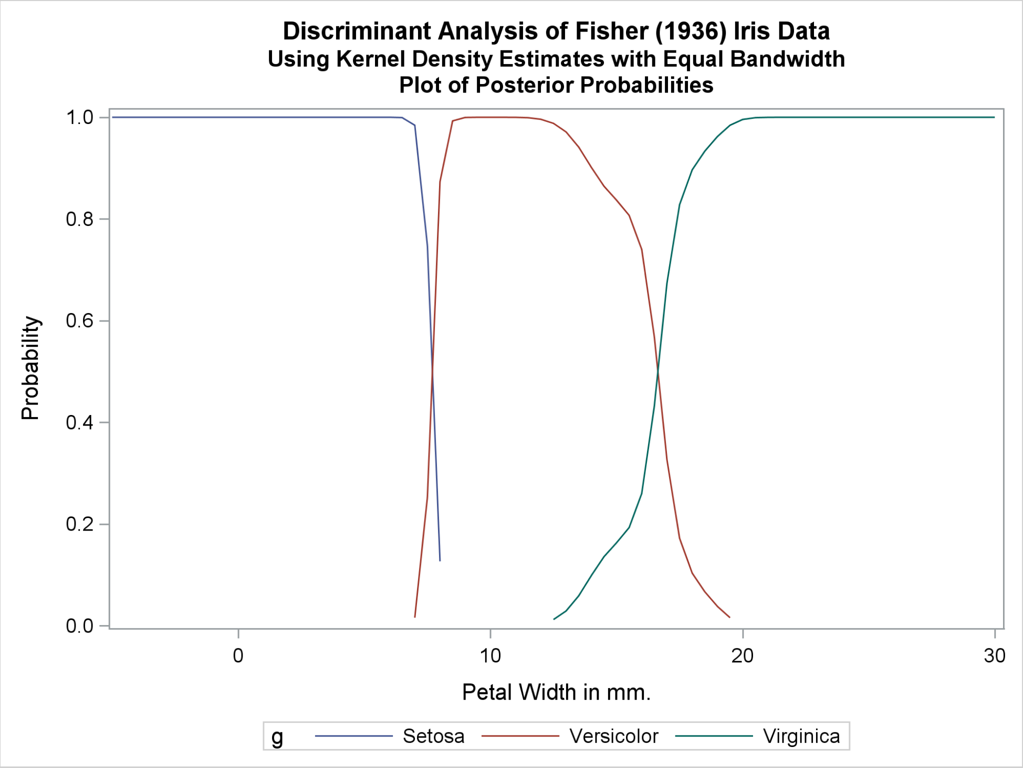
Number of Observations and Percent Classified into Species

	Setosa	Versicolor	Virginica	Total
Total	26 36.62	18 25.35	27 38.03	71 100.00
Priors	0.33333	0.33333	0.33333	

Output 36.1.4 *continued*



Output 36.1.4 *continued*



Another nonparametric analysis is run with unequal bandwidths (POOL=NO). The following statements produce Output 36.1.5:

```
title2 'Using Kernel Density Estimates with Unequal Bandwidth';

proc discrim data=sashelp.iris method=npar kernel=normal
   r=.4 pool=no testdata=plotdata testout=plotp
   testoutd=plotd short noclassify crosslisterr;
   class Species;
   var PetalWidth;
run;

%plotden;
%plotprob;
```

Output 36.1.5 Kernel Density Estimates with Unequal Bandwidth

Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Unequal Bandwidth

The DISCRIM Procedure

Total Sample Size	150	**DF Total**	149
Variables	1	**DF Within Classes**	147
Classes	3	**DF Between Classes**	2

Output 36.1.5 *continued*

Number of Observations Read	150
Number of Observations Used	150

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion	Prior Probability
Setosa	Setosa	50	50.0000	0.333333	0.333333
Versicolor	Versicolor	50	50.0000	0.333333	0.333333
Virginica	Virginica	50	50.0000	0.333333	0.333333

Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Unequal Bandwidth

The DISCRIM Procedure
Classification Results for Calibration Data: SASHELP.IRIS
Cross-validation Results using Normal Kernel Density

Posterior Probability of Membership in Species

Obs	From Species	Classified into Species		Setosa	Versicolor	Virginica
53	**Versicolor**	**Virginica**	*	0.0000	0.0475	0.9525
100	**Versicolor**	**Virginica**	*	0.0000	0.2310	0.7690
103	**Virginica**	**Versicolor**	*	0.0000	0.8805	0.1195
124	**Virginica**	**Versicolor**	*	0.0000	0.9394	0.0606
130	**Virginica**	**Versicolor**	*	0.0000	0.7193	0.2807
136	**Virginica**	**Versicolor**	*	0.0000	0.8805	0.1195

*** Misclassified observation**

Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Unequal Bandwidth

The DISCRIM Procedure
Classification Summary for Calibration Data: SASHELP.IRIS
Cross-validation Summary using Normal Kernel Density

Number of Observations and Percent Classified into Species

From Species	Setosa	Versicolor	Virginica	Total
Setosa	50 100.00	0 0.00	0 0.00	50 100.00
Versicolor	0 0.00	48 96.00	2 4.00	50 100.00
Virginica	0 0.00	4 8.00	46 92.00	50 100.00
Total	50 33.33	52 34.67	48 32.00	150 100.00
Priors	0.33333	0.33333	0.33333	

Output 36.1.5 *continued*

Error Count Estimates for Species				
	Setosa	**Versicolor**	**Virginica**	**Total**
Rate	0.0000	0.0400	0.0800	0.0400
Priors	0.3333	0.3333	0.3333	

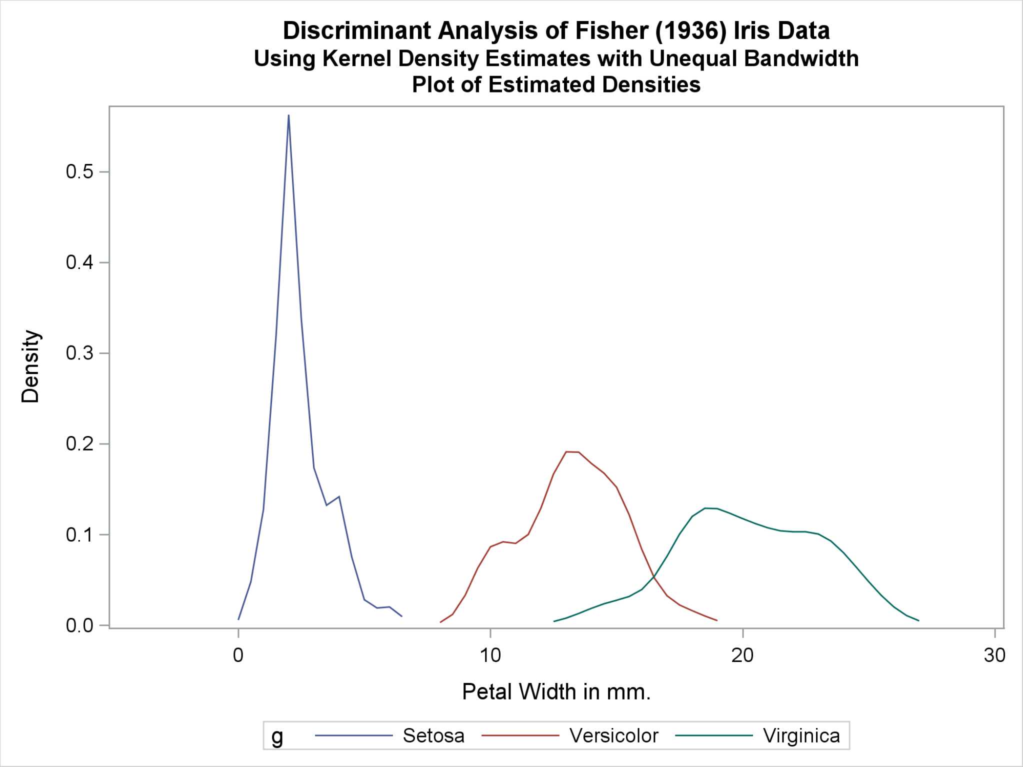
Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Unequal Bandwidth

The DISCRIM Procedure
Classification Summary for Test Data: WORK.PLOTDATA
Classification Summary using Normal Kernel Density

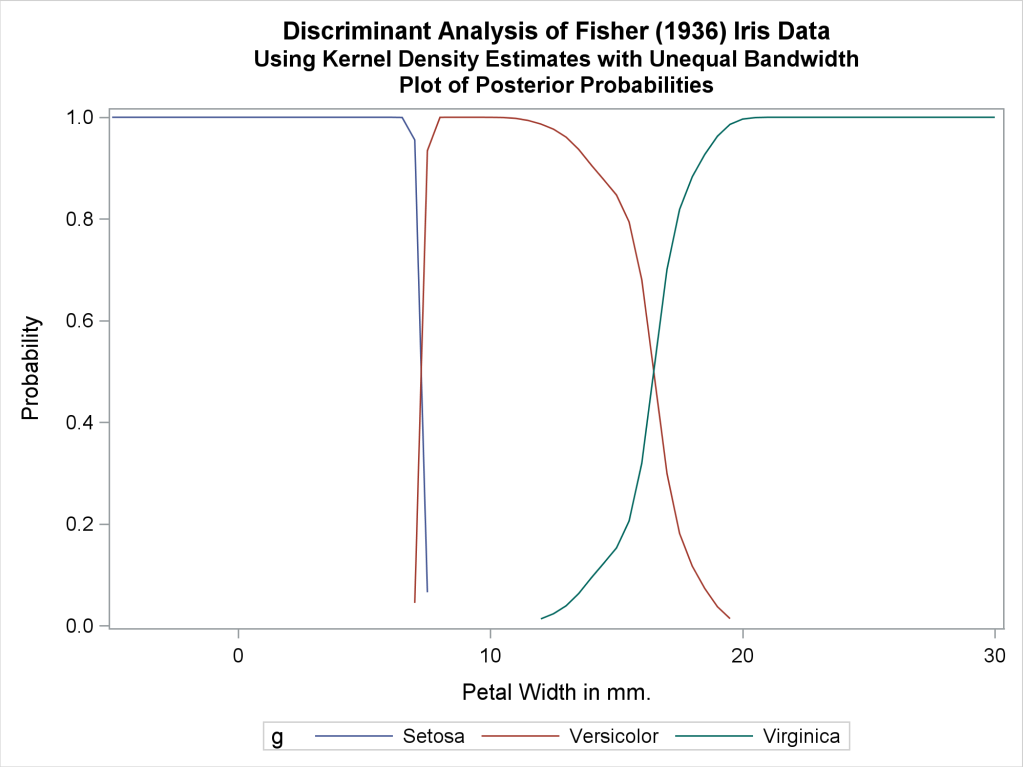
Observation Profile for Test Data	
Number of Observations Read	71
Number of Observations Used	71

Number of Observations and Percent Classified into Species				
	Setosa	**Versicolor**	**Virginica**	**Total**
Total	25 35.21	18 25.35	28 39.44	71 100.00
Priors	0.33333	0.33333	0.33333	

Output 36.1.5 *continued*



Output 36.1.5 *continued*



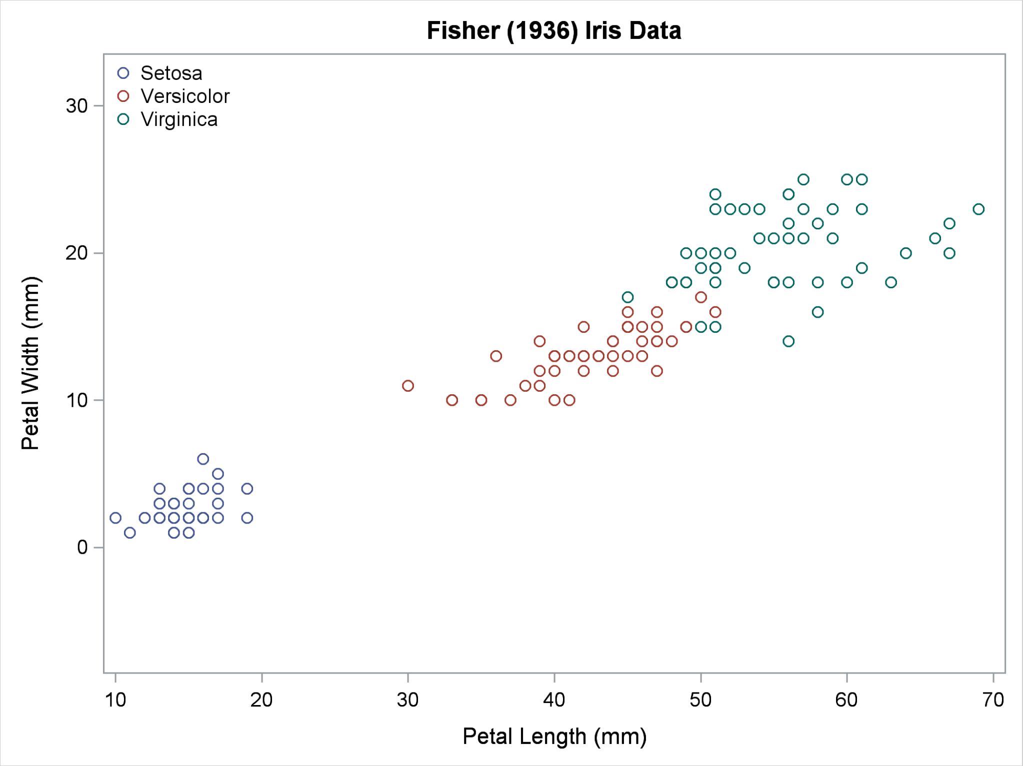
Example 36.2: Bivariate Density Estimates and Posterior Probabilities

In this example, four more discriminant analyses of iris data are run with two quantitative variables: petal width and petal length. The following statements produce Output 36.2.1 through Output 36.2.5:

```
title 'Discriminant Analysis of Fisher (1936) Iris Data';
proc template;
   define statgraph scatter;
      begingraph;
         entrytitle 'Fisher (1936) Iris Data';
         layout overlayequated / equatetype=fit;
            scatterplot x=petallength y=petalwidth /
                        group=species name='iris';
            layout gridded / autoalign=(topleft);
               discretelegend 'iris' / border=false opaque=false;
            endlayout;
         endlayout;
      endgraph;
   end;
run;

proc sgrender data=sashelp.iris template=scatter;
run;
```

The scatter plot in Output 36.2.1 shows the joint sample distribution.

Output 36.2.1 Joint Sample Distribution of Petal Width and Petal Length in Three Species



Another data set is created for plotting, containing a grid of points suitable for contour plots. The following statements create the data set:

```
data plotdata;
   do PetalLength = -2 to 72 by 0.5;
      do PetalWidth= - 5 to 32 by 0.5;
         output;
      end;
   end;
run;
```

Three macros are defined as follows to make contour plots of density estimates, posterior probabilities, and classification results:

```
%let close = thresholdmin=0 thresholdmax=0 offsetmin=0 offsetmax=0;
%let close = xaxisopts=(&close) yaxisopts=(&close);

proc template;
   define statgraph contour;
      begingraph;
         layout overlayequated / equatetype=equate &close;
            contourplotparm x=petallength y=petalwidth z=z /
                            contourtype=fill nhint=30;
            scatterplot x=pl y=pw / group=species name='iris'
                        includemissinggroup=false primary=true;
            layout gridded / autoalign=(topleft);
               discretelegend 'iris' / border=false opaque=false;
            endlayout;
         endlayout;
      endgraph;
   end;
run;

%macro contden;
   data contour(keep=PetalWidth PetalLength species z pl pw);
      merge plotd(in=d) sashelp.iris(keep=PetalWidth PetalLength species
                                    rename=(PetalWidth=pw PetalLength=pl));
      if d then z = max(setosa,versicolor,virginica);
   run;

   title3 'Plot of Estimated Densities';

   proc sgrender data=contour template=contour;
   run;
%mend;

%macro contprob;
   data posterior(keep=PetalWidth PetalLength species z pl pw into);
      merge plotp(in=d) sashelp.iris(keep=PetalWidth PetalLength species
                                    rename=(PetalWidth=pw PetalLength=pl));
      if d then z = max(setosa,versicolor,virginica);
      into = 1 * (_into_ =: 'Set') + 2 * (_into_ =: 'Ver') +
             3 * (_into_ =: 'Vir');
   run;

   title3 'Plot of Posterior Probabilities ';

   proc sgrender data=posterior template=contour;
   run;
%mend;
```

```
%macro contclass;
   title3 'Plot of Classification Results';

   proc sgrender data=posterior(drop=z rename=(into=z)) template=contour;
   run;
%mend;
```

A normal-theory analysis (METHOD=NORMAL) assuming equal covariance matrices (POOL=YES) illustrates the linearity of the classification boundaries. These statements produce Output 36.2.2:

```
title2 'Using Normal Density Estimates with Equal Variance';

proc discrim data=sashelp.iris method=normal pool=yes
   testdata=plotdata testout=plotp testoutd=plotd
   short noclassify crosslisterr;
   class Species;
   var Petal:;
run;

%contden
%contprob
%contclass
```

Output 36.2.2 Normal Density Estimates with Equal Variance

Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Equal Variance

The DISCRIM Procedure

Total Sample Size	150	**DF Total**	149
Variables	2	**DF Within Classes**	147
Classes	3	**DF Between Classes**	2

Number of Observations Read	150
Number of Observations Used	150

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion	Prior Probability
Setosa	Setosa	50	50.0000	0.333333	0.333333
Versicolor	Versicolor	50	50.0000	0.333333	0.333333
Virginica	Virginica	50	50.0000	0.333333	0.333333

Output 36.2.2 *continued*

```
           Discriminant Analysis of Fisher (1936) Iris Data
           Using Normal Density Estimates with Equal Variance

                        The DISCRIM Procedure
             Classification Results for Calibration Data: SASHELP.IRIS
           Cross-validation Results using Linear Discriminant Function

         Posterior Probability of Membership in Species
      From        Classified into
 Obs  Species     Species            Setosa  Versicolor  Virginica
  53  Versicolor  Virginica   *      0.0000      0.2130     0.7870
 100  Versicolor  Virginica   *      0.0000      0.3118     0.6882
 103  Virginica   Versicolor  *      0.0000      0.8453     0.1547
 113  Virginica   Versicolor  *      0.0000      0.8322     0.1678
 124  Virginica   Versicolor  *      0.0000      0.8057     0.1943
 136  Virginica   Versicolor  *      0.0000      0.8903     0.1097

                   * Misclassified observation

           Discriminant Analysis of Fisher (1936) Iris Data
           Using Normal Density Estimates with Equal Variance

                        The DISCRIM Procedure
            Classification Summary for Calibration Data: SASHELP.IRIS
           Cross-validation Summary using Linear Discriminant Function

          Number of Observations and Percent Classified
                          into Species
          From
          Species      Setosa  Versicolor  Virginica   Total
          Setosa           50           0          0      50
                       100.00        0.00       0.00  100.00
          Versicolor        0          48          2      50
                         0.00       96.00       4.00  100.00
          Virginica         0           4         46      50
                         0.00        8.00      92.00  100.00
          Total            50          52         48     150
                        33.33       34.67      32.00  100.00
          Priors      0.33333     0.33333    0.33333

              Error Count Estimates for Species
                   Setosa  Versicolor  Virginica   Total
          Rate     0.0000      0.0400     0.0800  0.0400
          Priors   0.3333      0.3333     0.3333
```

Output 36.2.2 *continued*

Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Equal Variance

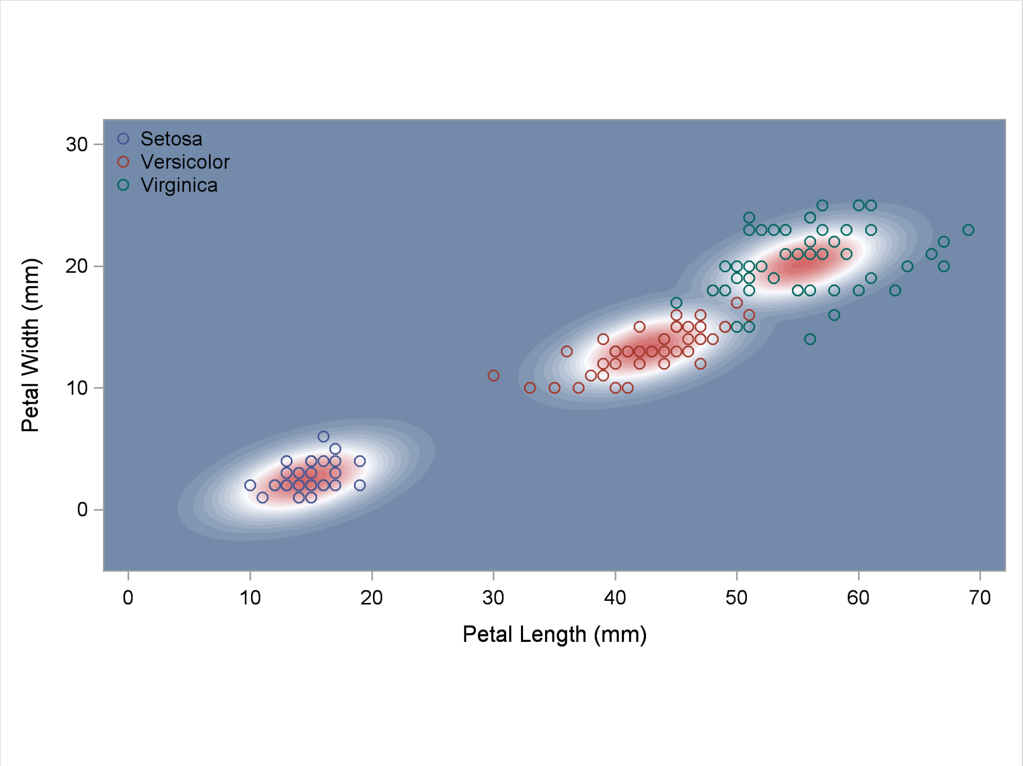
The DISCRIM Procedure
Classification Summary for Test Data: WORK.PLOTDATA
Classification Summary using Linear Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	11175
Number of Observations Used	11175

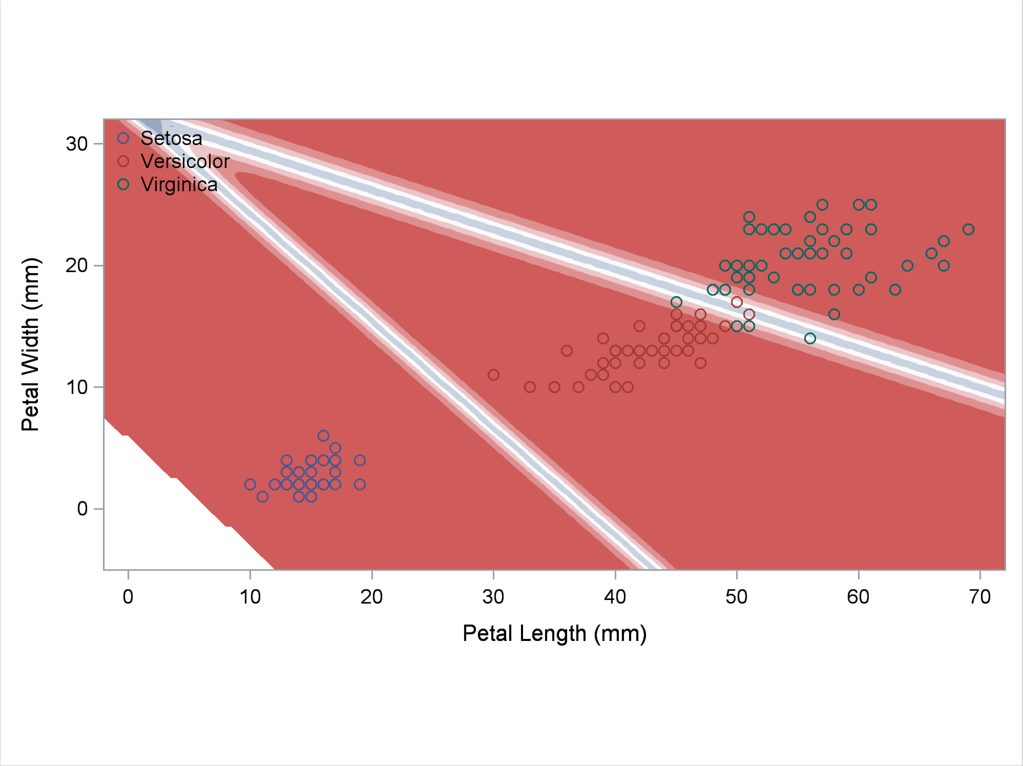
Number of Observations and Percent Classified into Species

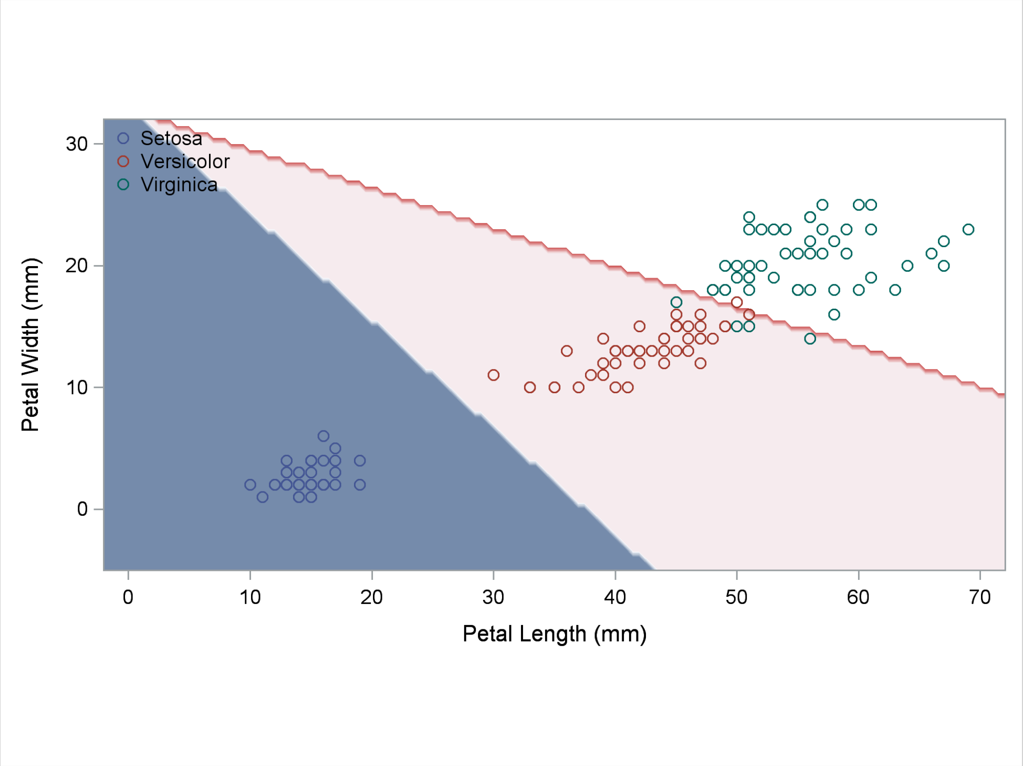
	Setosa	Versicolor	Virginica	Total
Total	3670 32.84	4243 37.97	3262 29.19	11175 100.00
Priors	0.33333	0.33333	0.33333	

Output 36.2.2 *continued*



Output 36.2.2 *continued*





A normal-theory analysis assuming unequal covariance matrices (POOL=NO) illustrates quadratic classification boundaries. These statements produce Output 36.2.3:

```
title2 'Using Normal Density Estimates with Unequal Variance';

proc discrim data=sashelp.iris method=normal pool=no
   testdata=plotdata testout=plotp testoutd=plotd
   short noclassify crosslisterr;
   class Species;
   var Petal:;
run;

%contden
%contprob
%contclass
```

Output 36.2.3 Normal Density Estimates with Unequal Variance

Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Unequal Variance

The DISCRIM Procedure

Total Sample Size	150	**DF Total**	149
Variables	2	**DF Within Classes**	147
Classes	3	**DF Between Classes**	2

Number of Observations Read	150
Number of Observations Used	150

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion	Prior Probability
Setosa	Setosa	50	50.0000	0.333333	0.333333
Versicolor	Versicolor	50	50.0000	0.333333	0.333333
Virginica	Virginica	50	50.0000	0.333333	0.333333

Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Unequal Variance

The DISCRIM Procedure
Classification Results for Calibration Data: SASHELP.IRIS
Cross-validation Results using Quadratic Discriminant Function

Posterior Probability of Membership in Species

Obs	From Species	Classified into Species		Setosa	Versicolor	Virginica
53	**Versicolor**	**Virginica**	*	0.0000	0.0903	0.9097
100	**Versicolor**	**Virginica**	*	0.0000	0.4675	0.5325
103	**Virginica**	**Versicolor**	*	0.0000	0.7288	0.2712
113	**Virginica**	**Versicolor**	*	0.0000	0.5196	0.4804
136	**Virginica**	**Versicolor**	*	0.0000	0.8335	0.1665

*** Misclassified observation**

Output 36.2.3 *continued*

Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Unequal Variance

The DISCRIM Procedure
Classification Summary for Calibration Data: SASHELP.IRIS
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Species

From Species	Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	50
	100.00	0.00	0.00	100.00
Versicolor	0	48	2	50
	0.00	96.00	4.00	100.00
Virginica	0	3	47	50
	0.00	6.00	94.00	100.00
Total	50	51	49	150
	33.33	34.00	32.67	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Species

	Setosa	Versicolor	Virginica	Total
Rate	0.0000	0.0400	0.0600	0.0333
Priors	0.3333	0.3333	0.3333	

Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Unequal Variance

The DISCRIM Procedure
Classification Summary for Test Data: WORK.PLOTDATA
Classification Summary using Quadratic Discriminant Function

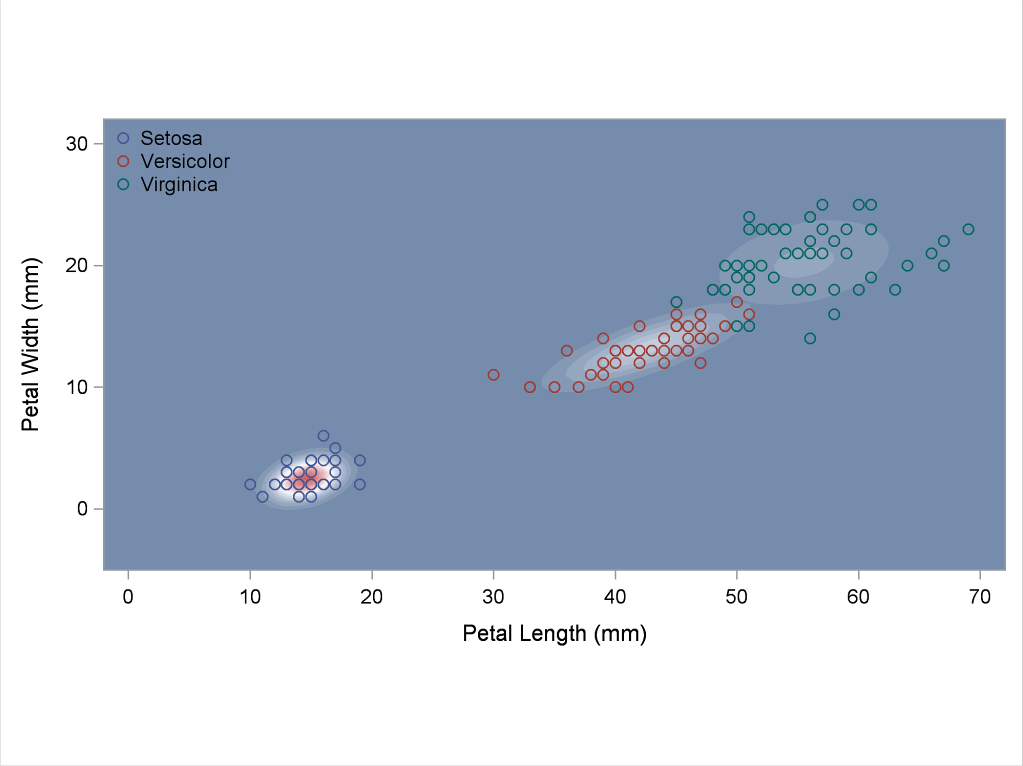
Observation Profile for Test Data

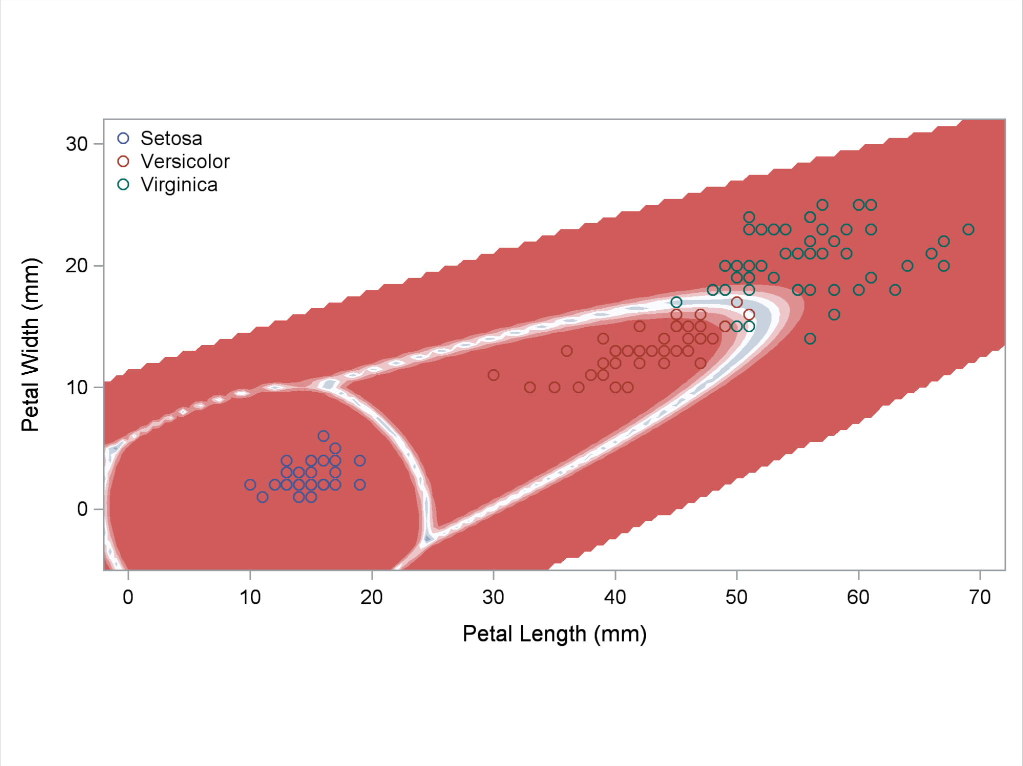
Number of Observations Read	11175
Number of Observations Used	11175

Number of Observations and Percent Classified into Species

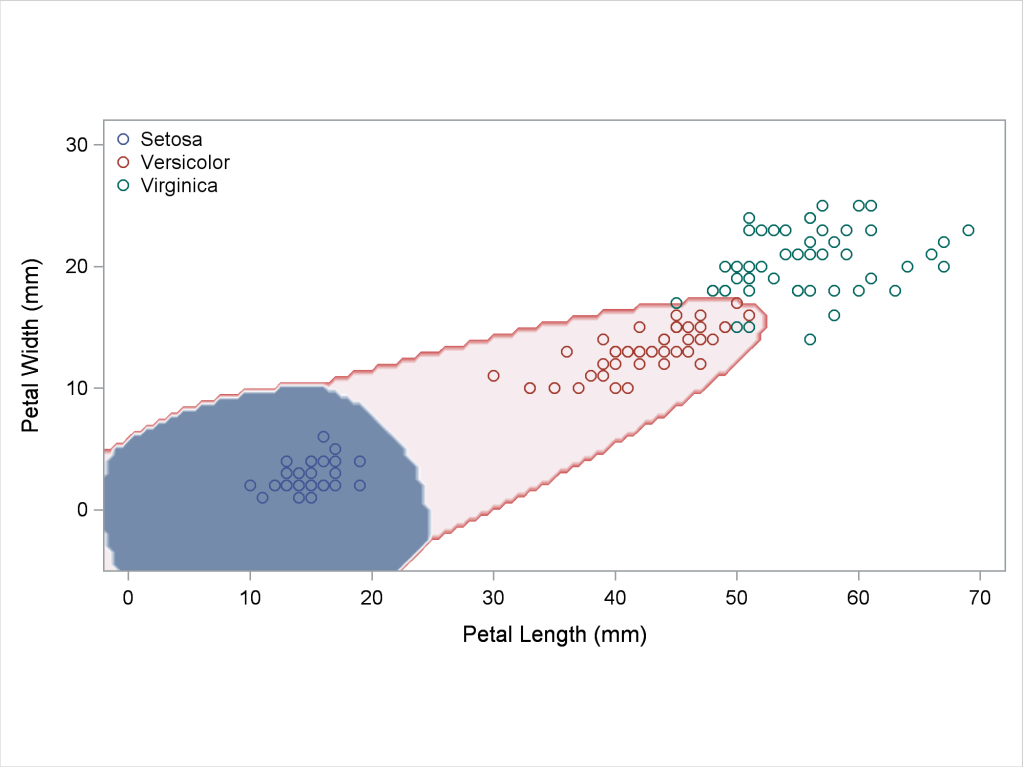
	Setosa	Versicolor	Virginica	Total
Total	1382	1345	8448	11175
	12.37	12.04	75.60	100.00
Priors	0.33333	0.33333	0.33333	

Output 36.2.3 *continued*





Output 36.2.3 *continued*



A nonparametric analysis (METHOD=NPAR) follows, using normal kernels (KERNEL=NORMAL) and equal bandwidths (POOL=YES) in each class. The value of the radius parameter *r* that, assuming normality, minimizes an approximate mean integrated square error is 0.50 (see the section "Nonparametric Methods" on page 2282). These statements produce Output 36.2.4:

```
title2 'Using Kernel Density Estimates with Equal Bandwidth';

proc discrim data=sashelp.iris method=npar kernel=normal
   r=.5 pool=yes testoutd=plotd testdata=plotdata testout=plotp
   short noclassify crosslisterr;
   class Species;
   var Petal:;
run;

%contden
%contprob
%contclass
```

Output 36.2.4 Kernel Density Estimates with Equal Bandwidth

Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Equal Bandwidth

The DISCRIM Procedure

Total Sample Size	150	**DF Total**	149
Variables	2	**DF Within Classes**	147
Classes	3	**DF Between Classes**	2

Number of Observations Read	150
Number of Observations Used	150

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion	Prior Probability
Setosa	Setosa	50	50.0000	0.333333	0.333333
Versicolor	Versicolor	50	50.0000	0.333333	0.333333
Virginica	Virginica	50	50.0000	0.333333	0.333333

Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Equal Bandwidth

The DISCRIM Procedure
Classification Results for Calibration Data: SASHELP.IRIS
Cross-validation Results using Normal Kernel Density

Posterior Probability of Membership in Species

Obs	From Species	Classified into Species		Setosa	Versicolor	Virginica
53	**Versicolor**	**Virginica**	*	0.0000	0.0800	0.9200
100	**Versicolor**	**Virginica**	*	0.0000	0.4123	0.5877
103	**Virginica**	**Versicolor**	*	0.0000	0.7474	0.2526
113	**Virginica**	**Versicolor**	*	0.0000	0.5863	0.4137
136	**Virginica**	**Versicolor**	*	0.0000	0.8358	0.1642

*** Misclassified observation**

Output 36.2.4 *continued*

Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Equal Bandwidth

The DISCRIM Procedure
Classification Summary for Calibration Data: SASHELP.IRIS
Cross-validation Summary using Normal Kernel Density

Number of Observations and Percent Classified into Species

From Species	Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	50
	100.00	0.00	0.00	100.00
Versicolor	0	48	2	50
	0.00	96.00	4.00	100.00
Virginica	0	3	47	50
	0.00	6.00	94.00	100.00
Total	50	51	49	150
	33.33	34.00	32.67	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Species

	Setosa	Versicolor	Virginica	Total
Rate	0.0000	0.0400	0.0600	0.0333
Priors	0.3333	0.3333	0.3333	

Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Equal Bandwidth

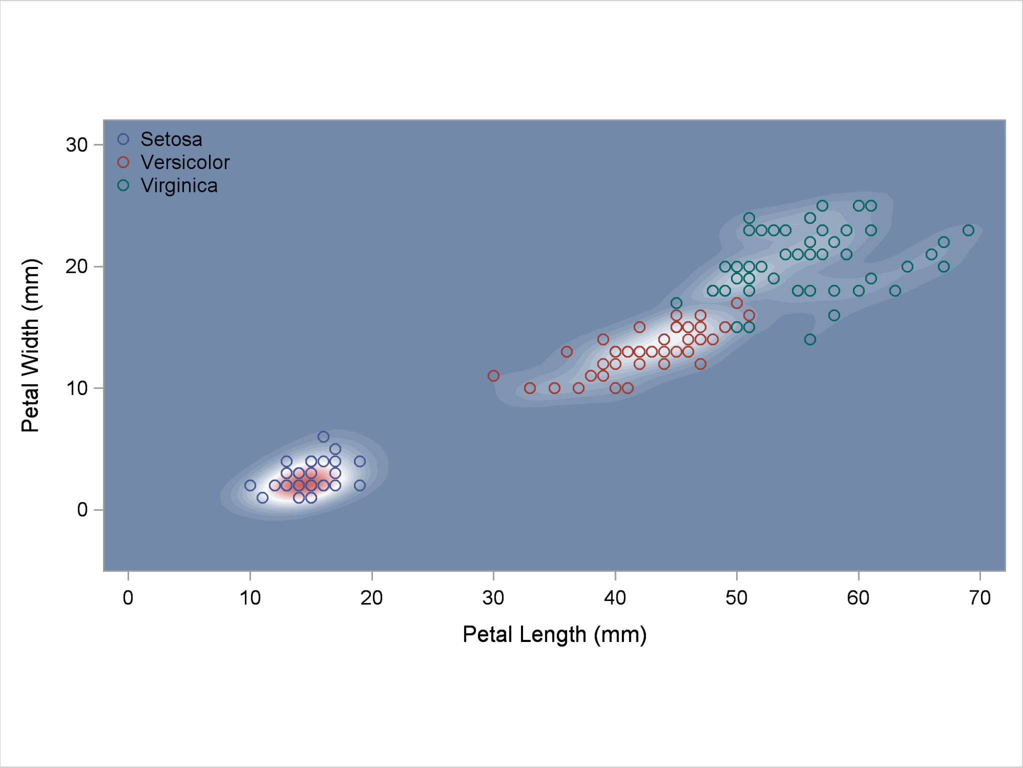
The DISCRIM Procedure
Classification Summary for Test Data: WORK.PLOTDATA
Classification Summary using Normal Kernel Density

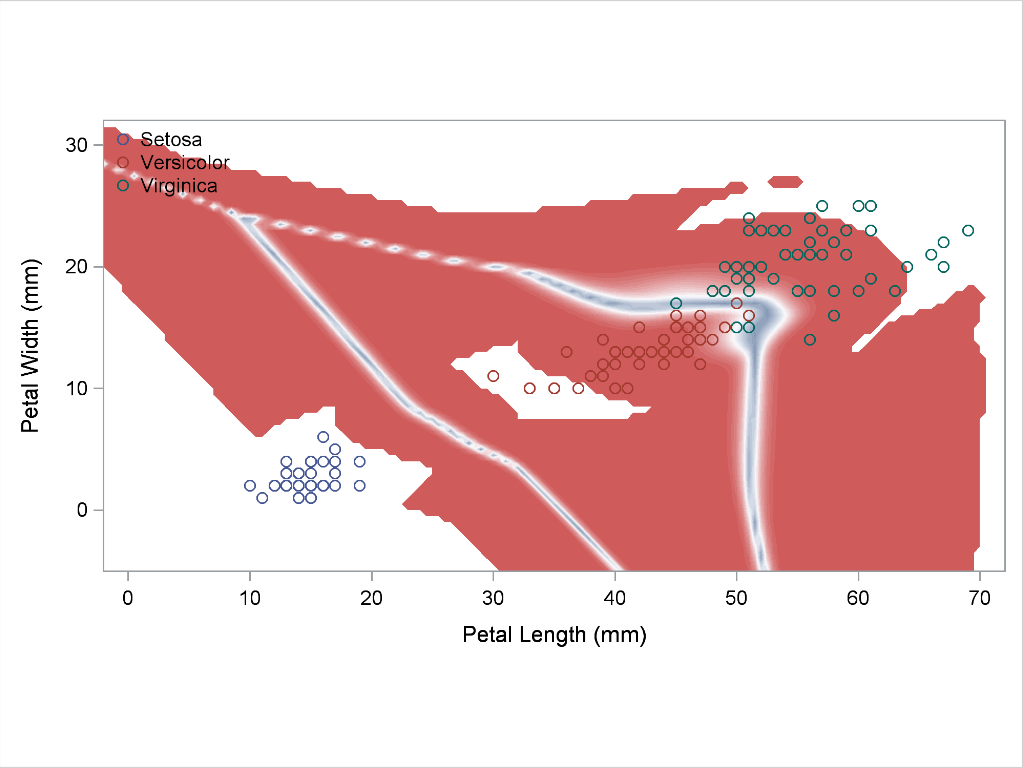
Observation Profile for Test Data	
Number of Observations Read	11175
Number of Observations Used	11175

Number of Observations and Percent Classified into Species

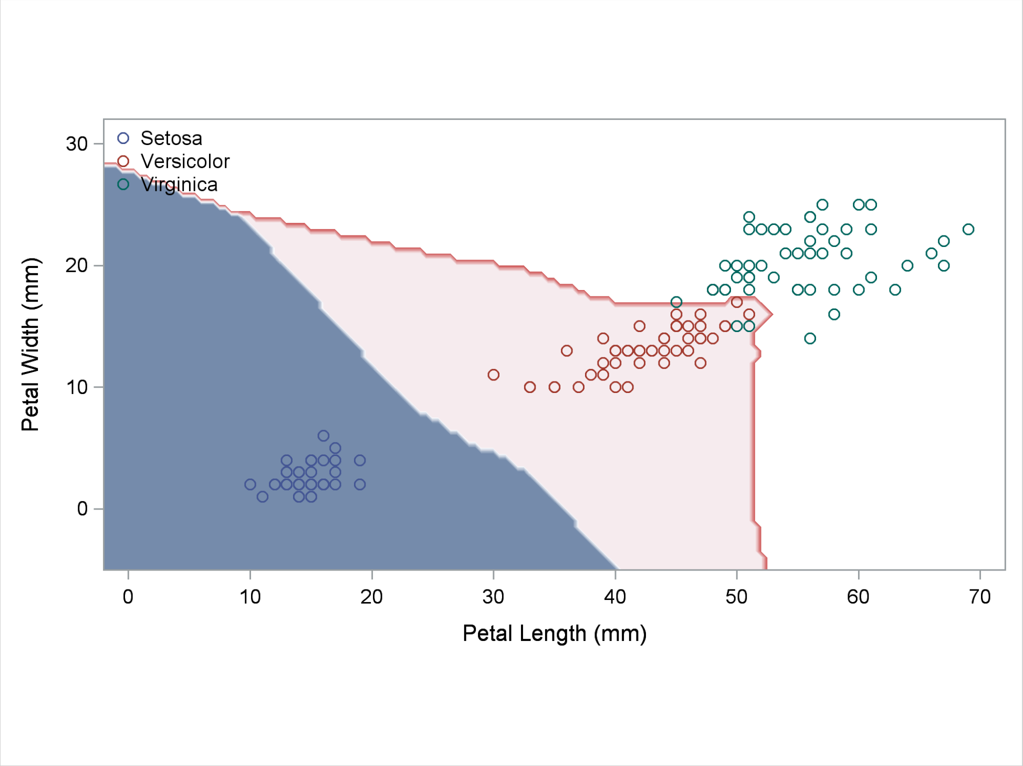
	Setosa	Versicolor	Virginica	Total
Total	3195	2492	5488	11175
	28.59	22.30	49.11	100.00
Priors	0.33333	0.33333	0.33333	

Output 36.2.4 *continued*





Output 36.2.4 *continued*



Another nonparametric analysis is run with unequal bandwidths (POOL=NO). These statements produce Output 36.2.5:

```
title2 'Using Kernel Density Estimates with Unequal Bandwidth';

proc discrim data=sashelp.iris method=npar kernel=normal
   r=.5 pool=no testoutd=plotd testdata=plotdata testout=plotp
   short noclassify crosslisterr;
   class Species;
   var Petal:;
run;

%contden
%contprob
%contclass
```

Output 36.2.5 Kernel Density Estimates with Unequal Bandwidth

Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Unequal Bandwidth

The DISCRIM Procedure

Total Sample Size	150	**DF Total**	149
Variables	2	**DF Within Classes**	147
Classes	3	**DF Between Classes**	2

Output 36.2.5 *continued*

Number of Observations Read	150
Number of Observations Used	150

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion	Prior Probability
Setosa	Setosa	50	50.0000	0.333333	0.333333
Versicolor	Versicolor	50	50.0000	0.333333	0.333333
Virginica	Virginica	50	50.0000	0.333333	0.333333

Discriminant Analysis of Fisher (1936) Iris Data Using Kernel Density Estimates with Unequal Bandwidth

The DISCRIM Procedure
Classification Results for Calibration Data: SASHELP.IRIS
Cross-validation Results using Normal Kernel Density

Posterior Probability of Membership in Species

Obs	From Species	Classified into Species		Setosa	Versicolor	Virginica
53	**Versicolor**	**Virginica**	*	0.0000	0.0516	0.9484
100	**Versicolor**	**Virginica**	*	0.0000	0.3773	0.6227
103	**Virginica**	**Versicolor**	*	0.0000	0.7826	0.2174
136	**Virginica**	**Versicolor**	*	0.0000	0.8802	0.1198

*** Misclassified observation**

Discriminant Analysis of Fisher (1936) Iris Data Using Kernel Density Estimates with Unequal Bandwidth

The DISCRIM Procedure
Classification Summary for Calibration Data: SASHELP.IRIS
Cross-validation Summary using Normal Kernel Density

Number of Observations and Percent Classified into Species

From Species	Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	50
	100.00	0.00	0.00	100.00
Versicolor	0	48	2	50
	0.00	96.00	4.00	100.00
Virginica	0	2	48	50
	0.00	4.00	96.00	100.00
Total	50	50	50	150
	33.33	33.33	33.33	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Species

	Setosa	Versicolor	Virginica	Total
Rate	0.0000	0.0400	0.0400	0.0267
Priors	0.3333	0.3333	0.3333	

Output 36.2.5 *continued*

Discriminant Analysis of Fisher (1936) Iris Data
Using Kernel Density Estimates with Unequal Bandwidth

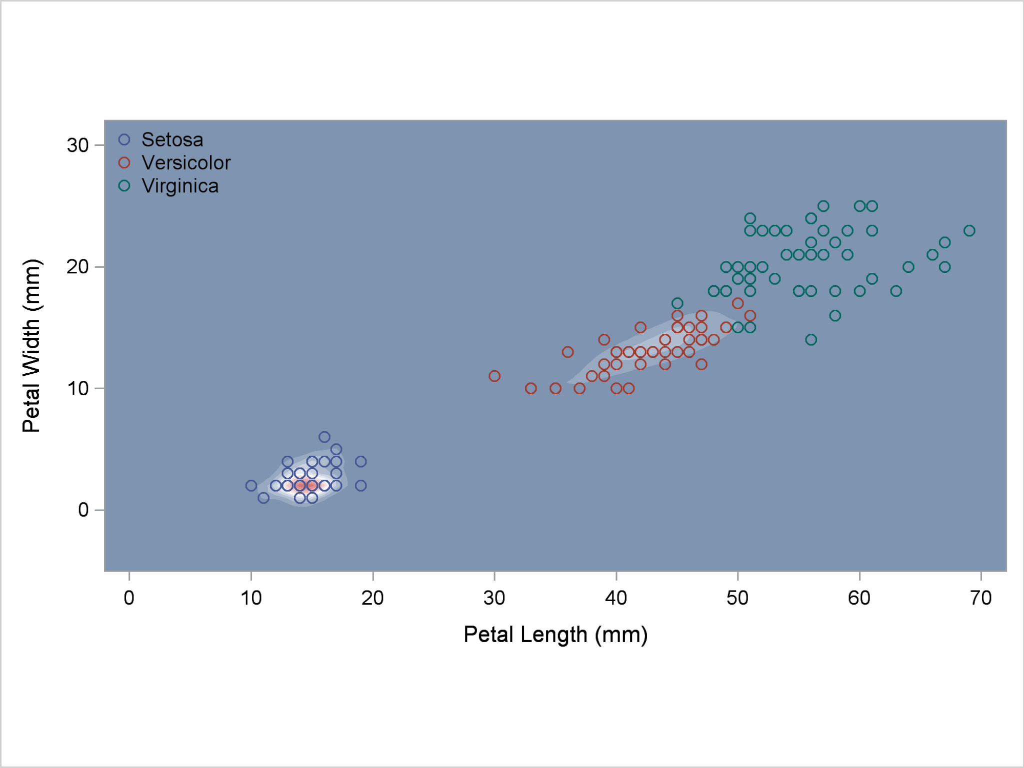
The DISCRIM Procedure
Classification Summary for Test Data: WORK.PLOTDATA
Classification Summary using Normal Kernel Density

Observation Profile for Test Data	
Number of Observations Read	11175
Number of Observations Used	11175

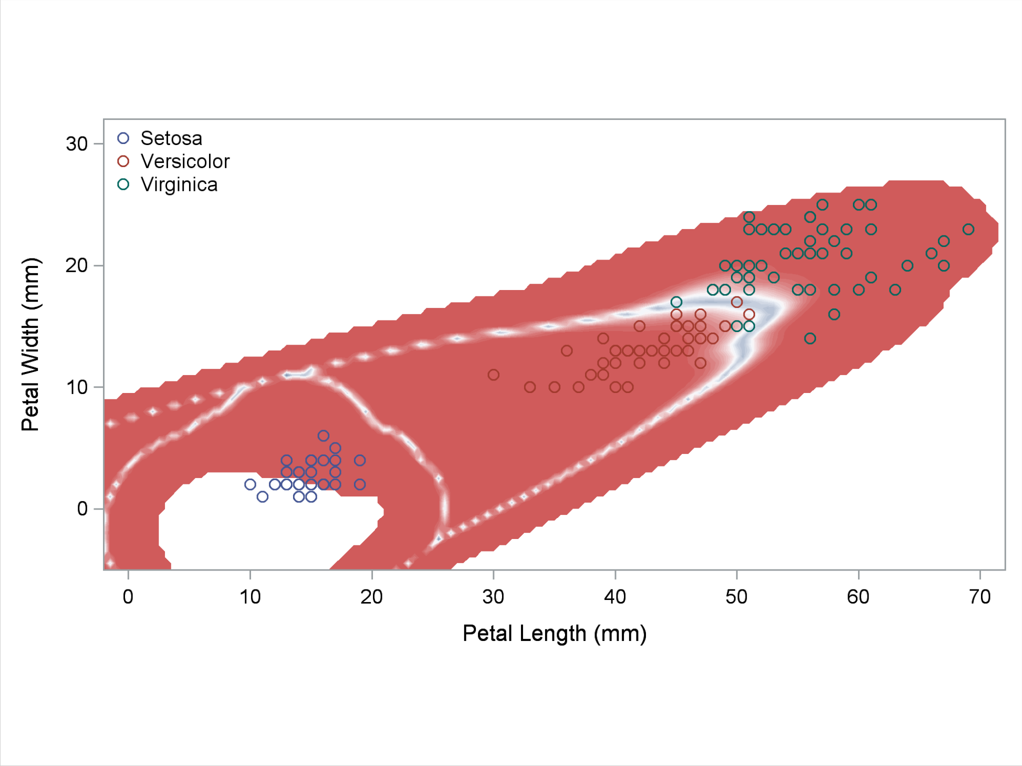
Number of Observations and Percent Classified into Species

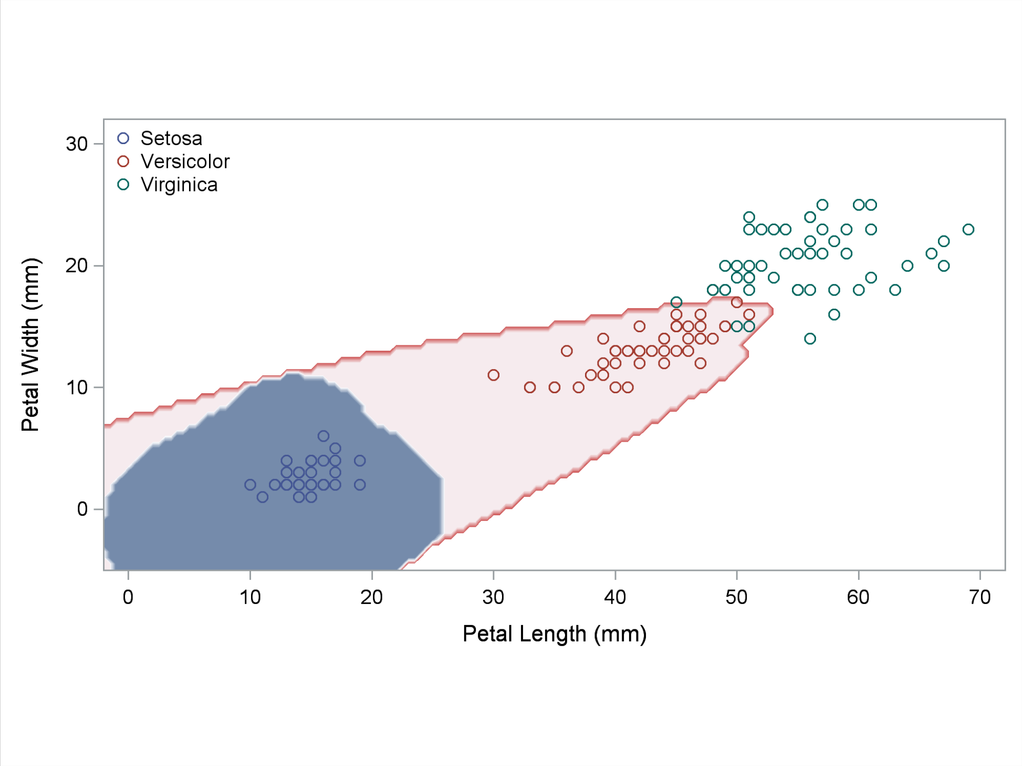
	Setosa	Versicolor	Virginica	Total
Total	1370 12.26	1505 13.47	8300 74.27	11175 100.00
Priors	0.33333	0.33333	0.33333	

Output 36.2.5 *continued*



Output 36.2.5 *continued*





Example 36.3: Normal-Theory Discriminant Analysis of Iris Data

In this example, PROC DISCRIM uses normal-theory methods to classify the iris data used in Example 36.1. The POOL=TEST option tests the homogeneity of the within-group covariance matrices (Output 36.3.3). Since the resulting test statistic is significant at the 0.10 level, the within-group covariance matrices are used to derive the quadratic discriminant criterion. The WCOV and PCOV options display the within-group covariance matrices and the pooled covariance matrix (Output 36.3.2). The DISTANCE option displays squared distances between classes (Output 36.3.4). The ANOVA and MANOVA options test the hypothesis that the class means are equal, by using univariate statistics and multivariate statistics; all statistics are significant at the 0.0001 level (Output 36.3.5). The LISTERR option lists the misclassified observations under resubstitution (Output 36.3.6). The CROSSLISTERR option lists the observations that are misclassified under cross validation and displays cross validation error-rate estimates (Output 36.3.7). The resubstitution error count estimate, 0.02, is not larger than the cross validation error count estimate, 0.0267, as would be expected because the resubstitution estimate is optimistically biased. The OUTSTAT= option generates a TYPE=MIXED (because POOL=TEST) output data set containing various statistics such as means, covariances, and coefficients of the discriminant function (Output 36.3.8).

The following statements produce Output 36.3.1 through Output 36.3.8:

```
title 'Discriminant Analysis of Fisher (1936) Iris Data';
title2 'Using Quadratic Discriminant Function';

proc discrim data=sashelp.iris outstat=irisstat
   wcov pcov method=normal pool=test
   distance anova manova listerr crosslisterr;
   class Species;
   var SepalLength SepalWidth PetalLength PetalWidth;
run;

proc print data=irisstat;
   title2 'Output Discriminant Statistics';
run;
```

Output 36.3.1 Quadratic Discriminant Analysis of Iris Data

Discriminant Analysis of Fisher (1936) Iris Data
Using Quadratic Discriminant Function

The DISCRIM Procedure

Total Sample Size	150	**DF Total**	149
Variables	4	**DF Within Classes**	147
Classes	3	**DF Between Classes**	2

Number of Observations Read	150
Number of Observations Used	150

Output 36.3.1 *continued*

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion	Prior Probability
Setosa	Setosa	50	50.0000	0.333333	0.333333
Versicolor	Versicolor	50	50.0000	0.333333	0.333333
Virginica	Virginica	50	50.0000	0.333333	0.333333

Output 36.3.2 Covariance Matrices

Discriminant Analysis of Fisher (1936) Iris Data Using Quadratic Discriminant Function

The DISCRIM Procedure
Within-Class Covariance Matrices

Species = Setosa, DF = 49

Variable	Label	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	Sepal Length (mm)	12.42489796	9.92163265	1.63551020	1.03306122
SepalWidth	Sepal Width (mm)	9.92163265	14.36897959	1.16979592	0.92979592
PetalLength	Petal Length (mm)	1.63551020	1.16979592	3.01591837	0.60693878
PetalWidth	Petal Width (mm)	1.03306122	0.92979592	0.60693878	1.11061224

Species = Versicolor, DF = 49

Variable	Label	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	Sepal Length (mm)	26.64326531	8.51836735	18.28979592	5.57795918
SepalWidth	Sepal Width (mm)	8.51836735	9.84693878	8.26530612	4.12040816
PetalLength	Petal Length (mm)	18.28979592	8.26530612	22.08163265	7.31020408
PetalWidth	Petal Width (mm)	5.57795918	4.12040816	7.31020408	3.91061224

Species = Virginica, DF = 49

Variable	Label	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	Sepal Length (mm)	40.43428571	9.37632653	30.32897959	4.90938776
SepalWidth	Sepal Width (mm)	9.37632653	10.40040816	7.13795918	4.76285714
PetalLength	Petal Length (mm)	30.32897959	7.13795918	30.45877551	4.88244898
PetalWidth	Petal Width (mm)	4.90938776	4.76285714	4.88244898	7.54326531

Discriminant Analysis of Fisher (1936) Iris Data Using Quadratic Discriminant Function

The DISCRIM Procedure

Pooled Within-Class Covariance Matrix, DF = 147

Variable	Label	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	Sepal Length (mm)	26.50081633	9.27210884	16.75142857	3.84013605
SepalWidth	Sepal Width (mm)	9.27210884	11.53877551	5.52435374	3.27102041
PetalLength	Petal Length (mm)	16.75142857	5.52435374	18.51877551	4.26653061
PetalWidth	Petal Width (mm)	3.84013605	3.27102041	4.26653061	4.18816327

Output 36.3.2 *continued*

Within Covariance Matrix Information

Species	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
Setosa	4	5.35332
Versicolor	4	7.54636
Virginica	4	9.49362
Pooled	4	8.46214

Output 36.3.3 Homogeneity Test

Discriminant Analysis of Fisher (1936) Iris Data
Using Quadratic Discriminant Function

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
140.943050	20	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Output 36.3.4 Squared Distances

Discriminant Analysis of Fisher (1936) Iris Data
Using Quadratic Discriminant Function

The DISCRIM Procedure

Squared Distance to Species

From Species	Setosa	Versicolor	Virginica
Setosa	0	103.19382	168.76759
Versicolor	323.06203	0	13.83875
Virginica	706.08494	17.86670	0

Generalized Squared Distance to Species

From Species	Setosa	Versicolor	Virginica
Setosa	5.35332	110.74017	178.26121
Versicolor	328.41535	7.54636	23.33238
Virginica	711.43826	25.41306	9.49362

Output 36.3.5 Tests of Equal Class Means

Discriminant Analysis of Fisher (1936) Iris Data Using Quadratic Discriminant Function

The DISCRIM Procedure

Univariate Test Statistics

F Statistics, Num DF=2, Den DF=147

Variable	Label	Total Standard Deviation	Pooled Standard Deviation	Between Standard Deviation	R-Square	R-Square / (1-RSq)	F Value	Pr > F
SepalLength	Sepal Length (mm)	8.2807	5.1479	7.9506	0.6187	1.6226	119.26	<.0001
SepalWidth	Sepal Width (mm)	4.3587	3.3969	3.3682	0.4008	0.6688	49.16	<.0001
PetalLength	Petal Length (mm)	17.6530	4.3033	20.9070	0.9414	16.0566	1180.16	<.0001
PetalWidth	Petal Width (mm)	7.6224	2.0465	8.9673	0.9289	13.0613	960.01	<.0001

Average R-Square

Unweighted	0.7224358
Weighted by Variance	0.8689444

Multivariate Statistics and F Approximations

S=2 M=0.5 N=71

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.02343863	199.15	8	288	<.0001
Pillai's Trace	1.19189883	53.47	8	290	<.0001
Hotelling-Lawley Trace	32.47732024	582.20	8	203.4	<.0001
Roy's Greatest Root	32.19192920	1166.96	4	145	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

Output 36.3.6 Misclassified Observations: Resubstitution

Discriminant Analysis of Fisher (1936) Iris Data Using Quadratic Discriminant Function

The DISCRIM Procedure
Classification Results for Calibration Data: SASHELP.IRIS
Resubstitution Results using Quadratic Discriminant Function

Posterior Probability of Membership in Species

Obs	From Species	Classified into Species		Setosa	Versicolor	Virginica
53	**Versicolor**	**Virginica**	*	0.0000	0.3359	0.6641
55	**Versicolor**	**Virginica**	*	0.0000	0.1543	0.8457
103	**Virginica**	**Versicolor**	*	0.0000	0.6050	0.3950

*** Misclassified observation**

Output 36.3.6 *continued*

Discriminant Analysis of Fisher (1936) Iris Data
Using Quadratic Discriminant Function

The DISCRIM Procedure
Classification Summary for Calibration Data: SASHELP.IRIS
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Species

From Species	Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	50
	100.00	0.00	0.00	100.00
Versicolor	0	48	2	50
	0.00	96.00	4.00	100.00
Virginica	0	1	49	50
	0.00	2.00	98.00	100.00
Total	50	49	51	150
	33.33	32.67	34.00	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Species

	Setosa	Versicolor	Virginica	Total
Rate	0.0000	0.0400	0.0200	0.0200
Priors	0.3333	0.3333	0.3333	

Output 36.3.7 Misclassified Observations: Cross Validation

Discriminant Analysis of Fisher (1936) Iris Data
Using Quadratic Discriminant Function

The DISCRIM Procedure
Classification Results for Calibration Data: SASHELP.IRIS
Cross-validation Results using Quadratic Discriminant Function

Posterior Probability of Membership in Species

Obs	From Species	Classified into Species		Setosa	Versicolor	Virginica
52	**Versicolor**	**Virginica**	*	0.0000	0.3134	0.6866
53	**Versicolor**	**Virginica**	*	0.0000	0.1616	0.8384
55	**Versicolor**	**Virginica**	*	0.0000	0.0713	0.9287
103	**Virginica**	**Versicolor**	*	0.0000	0.6632	0.3368

*** Misclassified observation**

Output 36.3.7 *continued*

Discriminant Analysis of Fisher (1936) Iris Data
Using Quadratic Discriminant Function

The DISCRIM Procedure
Classification Summary for Calibration Data: SASHELP.IRIS
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Species

From Species	Setosa	Versicolor	Virginica	Total
Setosa	50 100.00	0 0.00	0 0.00	50 100.00
Versicolor	0 0.00	47 94.00	3 6.00	50 100.00
Virginica	0 0.00	1 2.00	49 98.00	50 100.00
Total	50 33.33	48 32.00	52 34.67	150 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Species

	Setosa	Versicolor	Virginica	Total
Rate	0.0000	0.0600	0.0200	0.0267
Priors	0.3333	0.3333	0.3333	

Output 36.3.8 Output Statistics from Iris Data

Discriminant Analysis of Fisher (1936) Iris Data
Output Discriminant Statistics

Obs	Species	_TYPE_	_NAME_	SepalLength	SepalWidth	PetalLength	PetalWidth
1		N		150.00	150.00	150.00	150.00
2	Setosa	N		50.00	50.00	50.00	50.00
3	Versicolor	N		50.00	50.00	50.00	50.00
4	Virginica	N		50.00	50.00	50.00	50.00
5		MEAN		58.43	30.57	37.58	11.99
6	Setosa	MEAN		50.06	34.28	14.62	2.46
7	Versicolor	MEAN		59.36	27.70	42.60	13.26
8	Virginica	MEAN		65.88	29.74	55.52	20.26
9	Setosa	PRIOR		0.33	0.33	0.33	0.33
10	Versicolor	PRIOR		0.33	0.33	0.33	0.33
11	Virginica	PRIOR		0.33	0.33	0.33	0.33
12	Setosa	CSSCP	SepalLength	608.82	486.16	80.14	50.62
13	Setosa	CSSCP	SepalWidth	486.16	704.08	57.32	45.56
14	Setosa	CSSCP	PetalLength	80.14	57.32	147.78	29.74
15	Setosa	CSSCP	PetalWidth	50.62	45.56	29.74	54.42
16	Versicolor	CSSCP	SepalLength	1305.52	417.40	896.20	273.32
17	Versicolor	CSSCP	SepalWidth	417.40	482.50	405.00	201.90
18	Versicolor	CSSCP	PetalLength	896.20	405.00	1082.00	358.20
19	Versicolor	CSSCP	PetalWidth	273.32	201.90	358.20	191.62
20	Virginica	CSSCP	SepalLength	1981.28	459.44	1486.12	240.56
21	Virginica	CSSCP	SepalWidth	459.44	509.62	349.76	233.38
22	Virginica	CSSCP	PetalLength	1486.12	349.76	1492.48	239.24
23	Virginica	CSSCP	PetalWidth	240.56	233.38	239.24	369.62
24		PSSCP	SepalLength	3895.62	1363.00	2462.46	564.50
25		PSSCP	SepalWidth	1363.00	1696.20	812.08	480.84
26		PSSCP	PetalLength	2462.46	812.08	2722.26	627.18
27		PSSCP	PetalWidth	564.50	480.84	627.18	615.66
28		BSSCP	SepalLength	6321.21	-1995.27	16524.84	7127.93
29		BSSCP	SepalWidth	-1995.27	1134.49	-5723.96	-2293.27
30		BSSCP	PetalLength	16524.84	-5723.96	43710.28	18677.40
31		BSSCP	PetalWidth	7127.93	-2293.27	18677.40	8041.33
32		CSSCP	SepalLength	10216.83	-632.27	18987.30	7692.43
33		CSSCP	SepalWidth	-632.27	2830.69	-4911.88	-1812.43
34		CSSCP	PetalLength	18987.30	-4911.88	46432.54	19304.58
35		CSSCP	PetalWidth	7692.43	-1812.43	19304.58	8656.99
36		RSQUARED		0.62	0.40	0.94	0.93
37	Setosa	COV	SepalLength	12.42	9.92	1.64	1.03
38	Setosa	COV	SepalWidth	9.92	14.37	1.17	0.93
39	Setosa	COV	PetalLength	1.64	1.17	3.02	0.61
40	Setosa	COV	PetalWidth	1.03	0.93	0.61	1.11
41	Versicolor	COV	SepalLength	26.64	8.52	18.29	5.58
42	Versicolor	COV	SepalWidth	8.52	9.85	8.27	4.12
43	Versicolor	COV	PetalLength	18.29	8.27	22.08	7.31
44	Versicolor	COV	PetalWidth	5.58	4.12	7.31	3.91
45	Virginica	COV	SepalLength	40.43	9.38	30.33	4.91
46	Virginica	COV	SepalWidth	9.38	10.40	7.14	4.76

Output 36.3.8 *continued*

Discriminant Analysis of Fisher (1936) Iris Data
Output Discriminant Statistics

Obs	Species	_TYPE_	_NAME_	SepalLength	SepalWidth	PetalLength	PetalWidth
47	Virginica	COV	PetalLength	30.33	7.14	30.46	4.88
48	Virginica	COV	PetalWidth	4.91	4.76	4.88	7.54
49		PCOV	SepalLength	26.50	9.27	16.75	3.84
50		PCOV	SepalWidth	9.27	11.54	5.52	3.27
51		PCOV	PetalLength	16.75	5.52	18.52	4.27
52		PCOV	PetalWidth	3.84	3.27	4.27	4.19
53		BCOV	SepalLength	63.21	-19.95	165.25	71.28
54		BCOV	SepalWidth	-19.95	11.34	-57.24	-22.93
55		BCOV	PetalLength	165.25	-57.24	437.10	186.77
56		BCOV	PetalWidth	71.28	-22.93	186.77	80.41
57		COV	SepalLength	68.57	-4.24	127.43	51.63
58		COV	SepalWidth	-4.24	19.00	-32.97	-12.16
59		COV	PetalLength	127.43	-32.97	311.63	129.56
60		COV	PetalWidth	51.63	-12.16	129.56	58.10
61	Setosa	STD		3.52	3.79	1.74	1.05
62	Versicolor	STD		5.16	3.14	4.70	1.98
63	Virginica	STD		6.36	3.22	5.52	2.75
64		PSTD		5.15	3.40	4.30	2.05
65		BSTD		7.95	3.37	20.91	8.97
66		STD		8.28	4.36	17.65	7.62
67	Setosa	CORR	SepalLength	1.00	0.74	0.27	0.28
68	Setosa	CORR	SepalWidth	0.74	1.00	0.18	0.23
69	Setosa	CORR	PetalLength	0.27	0.18	1.00	0.33
70	Setosa	CORR	PetalWidth	0.28	0.23	0.33	1.00
71	Versicolor	CORR	SepalLength	1.00	0.53	0.75	0.55
72	Versicolor	CORR	SepalWidth	0.53	1.00	0.56	0.66
73	Versicolor	CORR	PetalLength	0.75	0.56	1.00	0.79
74	Versicolor	CORR	PetalWidth	0.55	0.66	0.79	1.00
75	Virginica	CORR	SepalLength	1.00	0.46	0.86	0.28
76	Virginica	CORR	SepalWidth	0.46	1.00	0.40	0.54
77	Virginica	CORR	PetalLength	0.86	0.40	1.00	0.32
78	Virginica	CORR	PetalWidth	0.28	0.54	0.32	1.00
79		PCORR	SepalLength	1.00	0.53	0.76	0.36
80		PCORR	SepalWidth	0.53	1.00	0.38	0.47
81		PCORR	PetalLength	0.76	0.38	1.00	0.48
82		PCORR	PetalWidth	0.36	0.47	0.48	1.00
83		BCORR	SepalLength	1.00	-0.75	0.99	1.00
84		BCORR	SepalWidth	-0.75	1.00	-0.81	-0.76
85		BCORR	PetalLength	0.99	-0.81	1.00	1.00
86		BCORR	PetalWidth	1.00	-0.76	1.00	1.00
87		CORR	SepalLength	1.00	-0.12	0.87	0.82
88		CORR	SepalWidth	-0.12	1.00	-0.43	-0.37
89		CORR	PetalLength	0.87	-0.43	1.00	0.96
90		CORR	PetalWidth	0.82	-0.37	0.96	1.00
91	Setosa	STDMEAN		-1.01	0.85	-1.30	-1.25
92	Versicolor	STDMEAN		0.11	-0.66	0.28	0.17

Output 36.3.8 *continued*

Discriminant Analysis of Fisher (1936) Iris Data
Output Discriminant Statistics

Obs	Species	_TYPE_	_NAME_	SepalLength	SepalWidth	PetalLength	PetalWidth
93	Virginica	STDMEAN		0.90	-0.19	1.02	1.08
94	Setosa	PSTDMEAN		-1.63	1.09	-5.34	-4.66
95	Versicolor	PSTDMEAN		0.18	-0.85	1.17	0.62
96	Virginica	PSTDMEAN		1.45	-0.25	4.17	4.04
97		LNDETERM		8.46	8.46	8.46	8.46
98	Setosa	LNDETERM		5.35	5.35	5.35	5.35
99	Versicolor	LNDETERM		7.55	7.55	7.55	7.55
100	Virginica	LNDETERM		9.49	9.49	9.49	9.49
101	Setosa	QUAD	SepalLength	-0.09	0.06	0.02	0.02
102	Setosa	QUAD	SepalWidth	0.06	-0.08	-0.01	0.01
103	Setosa	QUAD	PetalLength	0.02	-0.01	-0.19	0.09
104	Setosa	QUAD	PetalWidth	0.02	0.01	0.09	-0.53
105	Setosa	QUAD	_LINEAR_	4.46	-0.76	3.36	-3.13
106	Setosa	QUAD	_CONST_	-121.83	-121.83	-121.83	-121.83
107	Versicolor	QUAD	SepalLength	-0.05	0.02	0.04	-0.03
108	Versicolor	QUAD	SepalWidth	0.02	-0.10	-0.01	0.10
109	Versicolor	QUAD	PetalLength	0.04	-0.01	-0.10	0.13
110	Versicolor	QUAD	PetalWidth	-0.03	0.10	0.13	-0.44
111	Versicolor	QUAD	_LINEAR_	1.80	1.60	0.33	-1.47
112	Versicolor	QUAD	_CONST_	-76.55	-76.55	-76.55	-76.55
113	Virginica	QUAD	SepalLength	-0.05	0.02	0.05	-0.01
114	Virginica	QUAD	SepalWidth	0.02	-0.08	-0.01	0.04
115	Virginica	QUAD	PetalLength	0.05	-0.01	-0.07	0.01
116	Virginica	QUAD	PetalWidth	-0.01	0.04	0.01	-0.10
117	Virginica	QUAD	_LINEAR_	0.74	1.32	0.62	0.97
118	Virginica	QUAD	_CONST_	-75.82	-75.82	-75.82	-75.82

Example 36.4: Linear Discriminant Analysis of Remote-Sensing Data on Crops

In this example, the remote-sensing data are used. In this data set, the observations are grouped into five crops: clover, corn, cotton, soybeans, and sugar beets. Four measures called x1 through x4 make up the descriptive variables.

In the first PROC DISCRIM statement, the DISCRIM procedure uses normal-theory methods (METHOD=NORMAL) assuming equal variances (POOL=YES) in five crops. The PRIORS statement, PRIORS PROP, sets the prior probabilities proportional to the sample sizes. The LIST option lists the resubstitution classification results for each observation (Output 36.4.2). The CROSSVALIDATE option displays cross validation error-rate estimates (Output 36.4.3). The OUTSTAT= option stores the calibration information in a new data set to classify future observations. A second PROC DISCRIM statement uses this calibration information to classify a test data set. Note that the values of the identification variable, xvalues, are obtained by rereading the x1 through x4 fields in the data lines as a single character variable. The following statements produce Output 36.4.1 through Output 36.4.3:

```
title 'Discriminant Analysis of Remote Sensing Data on Five Crops';

data crops;
   input Crop $ 1-10 x1-x4 xvalues $ 11-21;
   datalines;
Corn      16 27 31 33
Corn      15 23 30 30
Corn      16 27 27 26
Corn      18 20 25 23
Corn      15 15 31 32
Corn      15 32 32 15
Corn      12 15 16 73
Soybeans  20 23 23 25
Soybeans  24 24 25 32
Soybeans  21 25 23 24
Soybeans  27 45 24 12
Soybeans  12 13 15 42
Soybeans  22 32 31 43
Cotton    31 32 33 34
Cotton    29 24 26 28
Cotton    34 32 28 45
Cotton    26 25 23 24
Cotton    53 48 75 26
Cotton    34 35 25 78
Sugarbeets22 23 25 42
Sugarbeets25 25 24 26
Sugarbeets34 25 16 52
Sugarbeets54 23 21 54
Sugarbeets25 43 32 15
Sugarbeets26 54  2 54
Clover    12 45 32 54
Clover    24 58 25 34
Clover    87 54 61 21
Clover    51 31 31 16
Clover    96 48 54 62
Clover    31 31 11 11
Clover    56 13 13 71
Clover    32 13 27 32
Clover    36 26 54 32
Clover    53 08 06 54
Clover    32 32 62 16
;

title2 'Using the Linear Discriminant Function';

proc discrim data=crops outstat=cropstat method=normal pool=yes
   list crossvalidate;
   class Crop;
   priors prop;
   id xvalues;
   var x1-x4;
run;
```

Output 36.4.1 Linear Discriminant Function on Crop Data

Discriminant Analysis of Remote Sensing Data on Five Crops Using the Linear Discriminant Function

The DISCRIM Procedure

Total Sample Size	36	**DF Total**	35
Variables	4	**DF Within Classes**	31
Classes	5	**DF Between Classes**	4

Number of Observations Read	36
Number of Observations Used	36

Class Level Information

Crop	Variable Name	Frequency	Weight	Proportion	Prior Probability
Clover	Clover	11	11.0000	0.305556	0.305556
Corn	Corn	7	7.0000	0.194444	0.194444
Cotton	Cotton	6	6.0000	0.166667	0.166667
Soybeans	Soybeans	6	6.0000	0.166667	0.166667
Sugarbeets	Sugarbeets	6	6.0000	0.166667	0.166667

Pooled Covariance Matrix Information

Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
4	21.30189

Discriminant Analysis of Remote Sensing Data on Five Crops Using the Linear Discriminant Function

The DISCRIM Procedure

Generalized Squared Distance to Crop

From Crop	Clover	Corn	Cotton	Soybeans	Sugarbeets
Clover	2.37125	7.52830	4.44969	6.16665	5.07262
Corn	6.62433	3.27522	5.46798	4.31383	6.47395
Cotton	3.23741	5.15968	3.58352	5.01819	4.87908
Soybeans	4.95438	4.00552	5.01819	3.58352	4.65998
Sugarbeets	3.86034	6.16564	4.87908	4.65998	3.58352

Linear Discriminant Function for Crop

Variable	Clover	Corn	Cotton	Soybeans	Sugarbeets
Constant	-10.98457	-7.72070	-11.46537	-7.28260	-9.80179
x1	0.08907	-0.04180	0.02462	0.0000369	0.04245
x2	0.17379	0.11970	0.17596	0.15896	0.20988
x3	0.11899	0.16511	0.15880	0.10622	0.06540
x4	0.15637	0.16768	0.18362	0.14133	0.16408

Output 36.4.2 Misclassified Observations: Resubstitution

Discriminant Analysis of Remote Sensing Data on Five Crops Using the Linear Discriminant Function

The DISCRIM Procedure
Classification Results for Calibration Data: WORK.CROPS
Resubstitution Results using Linear Discriminant Function

Posterior Probability of Membership in Crop

xvalues	From Crop	Classified into Crop		Clover	Corn	Cotton	Soybeans	Sugarbeets
16 27 31 33	Corn	Corn		0.0894	0.4054	0.1763	0.2392	0.0897
15 23 30 30	Corn	Corn		0.0769	0.4558	0.1421	0.2530	0.0722
16 27 27 26	Corn	Corn		0.0982	0.3422	0.1365	0.3073	0.1157
18 20 25 23	Corn	Corn		0.1052	0.3634	0.1078	0.3281	0.0955
15 15 31 32	Corn	Corn		0.0588	0.5754	0.1173	0.2087	0.0398
15 32 32 15	Corn	Soybeans	*	0.0972	0.3278	0.1318	0.3420	0.1011
12 15 16 73	Corn	Corn		0.0454	0.5238	0.1849	0.1376	0.1083
20 23 23 25	Soybeans	Soybeans		0.1330	0.2804	0.1176	0.3305	0.1385
24 24 25 32	Soybeans	Soybeans		0.1768	0.2483	0.1586	0.2660	0.1502
21 25 23 24	Soybeans	Soybeans		0.1481	0.2431	0.1200	0.3318	0.1570
27 45 24 12	Soybeans	Sugarbeets	*	0.2357	0.0547	0.1016	0.2721	0.3359
12 13 15 42	Soybeans	Corn	*	0.0549	0.4749	0.0920	0.2768	0.1013
22 32 31 43	Soybeans	Cotton	*	0.1474	0.2606	0.2624	0.1848	0.1448
31 32 33 34	Cotton	Clover	*	0.2815	0.1518	0.2377	0.1767	0.1523
29 24 26 28	Cotton	Soybeans	*	0.2521	0.1842	0.1529	0.2549	0.1559
34 32 28 45	Cotton	Clover	*	0.3125	0.1023	0.2404	0.1357	0.2091
26 25 23 24	Cotton	Soybeans	*	0.2121	0.1809	0.1245	0.3045	0.1780
53 48 75 26	Cotton	Clover	*	0.4837	0.0391	0.4384	0.0223	0.0166
34 35 25 78	Cotton	Cotton		0.2256	0.0794	0.3810	0.0592	0.2548
22 23 25 42	Sugarbeets	Corn	*	0.1421	0.3066	0.1901	0.2231	0.1381
25 25 24 26	Sugarbeets	Soybeans	*	0.1969	0.2050	0.1354	0.2960	0.1667
34 25 16 52	Sugarbeets	Sugarbeets		0.2928	0.0871	0.1665	0.1479	0.3056
54 23 21 54	Sugarbeets	Clover	*	0.6215	0.0194	0.1250	0.0496	0.1845
25 43 32 15	Sugarbeets	Soybeans	*	0.2258	0.1135	0.1646	0.2770	0.2191
26 54 2 54	Sugarbeets	Sugarbeets		0.0850	0.0081	0.0521	0.0661	0.7887
12 45 32 54	Clover	Cotton	*	0.0693	0.2663	0.3394	0.1460	0.1789
24 58 25 34	Clover	Sugarbeets	*	0.1647	0.0376	0.1680	0.1452	0.4845
87 54 61 21	Clover	Clover		0.9328	0.0003	0.0478	0.0025	0.0165
51 31 31 16	Clover	Clover		0.6642	0.0205	0.0872	0.0959	0.1322
96 48 54 62	Clover	Clover		0.9215	0.0002	0.0604	0.0007	0.0173
31 31 11 11	Clover	Sugarbeets	*	0.2525	0.0402	0.0473	0.3012	0.3588
56 13 13 71	Clover	Clover		0.6132	0.0212	0.1226	0.0408	0.2023
32 13 27 32	Clover	Clover		0.2669	0.2616	0.1512	0.2260	0.0943
36 26 54 32	Clover	Cotton	*	0.2650	0.2645	0.3495	0.0918	0.0292
53 08 06 54	Clover	Clover		0.5914	0.0237	0.0676	0.0781	0.2392
32 32 62 16	Clover	Cotton	*	0.2163	0.3180	0.3327	0.1125	0.0206

*** Misclassified observation**

Output 36.4.2 *continued*

Discriminant Analysis of Remote Sensing Data on Five Crops Using the Linear Discriminant Function

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CROPS
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into Crop

From Crop	Clover	Corn	Cotton	Soybeans	Sugarbeets	Total
Clover	6 54.55	0 0.00	3 27.27	0 0.00	2 18.18	11 100.00
Corn	0 0.00	6 85.71	0 0.00	1 14.29	0 0.00	7 100.00
Cotton	3 50.00	0 0.00	1 16.67	2 33.33	0 0.00	6 100.00
Soybeans	0 0.00	1 16.67	1 16.67	3 50.00	1 16.67	6 100.00
Sugarbeets	1 16.67	1 16.67	0 0.00	2 33.33	2 33.33	6 100.00
Total	10 27.78	8 22.22	5 13.89	8 22.22	5 13.89	36 100.00
Priors	0.30556	0.19444	0.16667	0.16667	0.16667	

Error Count Estimates for Crop

	Clover	Corn	Cotton	Soybeans	Sugarbeets	Total
Rate	0.4545	0.1429	0.8333	0.5000	0.6667	0.5000
Priors	0.3056	0.1944	0.1667	0.1667	0.1667	

Output 36.4.3 Misclassified Observations: Cross Validation

Discriminant Analysis of Remote Sensing Data on Five Crops Using the Linear Discriminant Function

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CROPS
Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into Crop						
From Crop	**Clover**	**Corn**	**Cotton**	**Soybeans**	**Sugarbeets**	**Total**
Clover	4 36.36	3 27.27	1 9.09	0 0.00	3 27.27	11 100.00
Corn	0 0.00	4 57.14	1 14.29	2 28.57	0 0.00	7 100.00
Cotton	3 50.00	0 0.00	0 0.00	2 33.33	1 16.67	6 100.00
Soybeans	0 0.00	1 16.67	1 16.67	3 50.00	1 16.67	6 100.00
Sugarbeets	2 33.33	1 16.67	0 0.00	2 33.33	1 16.67	6 100.00
Total	9 25.00	9 25.00	3 8.33	9 25.00	6 16.67	36 100.00
Priors	0.30556	0.19444	0.16667	0.16667	0.16667	

Error Count Estimates for Crop						
	Clover	**Corn**	**Cotton**	**Soybeans**	**Sugarbeets**	**Total**
Rate	0.6364	0.4286	1.0000	0.5000	0.8333	0.6667
Priors	0.3056	0.1944	0.1667	0.1667	0.1667	

Next, you can use the calibration information stored in the `Cropstat` data set to classify a test data set. The TESTLIST option lists the classification results for each observation in the test data set. The following statements produce Output 36.4.4 and Output 36.4.5:

```
data test;
   input Crop $ 1-10 x1-x4 xvalues $ 11-21;
   datalines;
Corn      16 27 31 33
Soybeans  21 25 23 24
Cotton    29 24 26 28
Sugarbeets54 23 21 54
Clover    32 32 62 16
;

title2 'Classification of Test Data';

proc discrim data=cropstat testdata=test testout=tout testlist;
   class Crop;
   testid xvalues;
   var x1-x4;
run;

proc print data=tout;
```

```
   title 'Discriminant Analysis of Remote Sensing Data on Five Crops';
   title2 'Output Classification Results of Test Data';
run;
```

Output 36.4.4 Classification of Test Data

Discriminant Analysis of Remote Sensing Data on Five Crops
Classification of Test Data

The DISCRIM Procedure
Classification Results for Test Data: WORK.TEST
Classification Results using Linear Discriminant Function

			Posterior Probability of Membership in Crop				
xvalues	**From Crop**	**Classified into Crop**	**Clover**	**Corn**	**Cotton**	**Soybeans**	**Sugarbeets**
16 27 31 33	**Corn**	**Corn**	0.0894	0.4054	0.1763	0.2392	0.0897
21 25 23 24	**Soybeans**	**Soybeans**	0.1481	0.2431	0.1200	0.3318	0.1570
29 24 26 28	**Cotton**	**Soybeans** *	0.2521	0.1842	0.1529	0.2549	0.1559
54 23 21 54	**Sugarbeets**	**Clover** *	0.6215	0.0194	0.1250	0.0496	0.1845
32 32 62 16	**Clover**	**Cotton** *	0.2163	0.3180	0.3327	0.1125	0.0206

*** Misclassified observation**

Discriminant Analysis of Remote Sensing Data on Five Crops
Classification of Test Data

The DISCRIM Procedure
Classification Summary for Test Data: WORK.TEST
Classification Summary using Linear Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	5
Number of Observations Used	5

Number of Observations and Percent Classified into Crop						
From Crop	**Clover**	**Corn**	**Cotton**	**Soybeans**	**Sugarbeets**	**Total**
Clover	0 0.00	0 0.00	1 100.00	0 0.00	0 0.00	1 100.00
Corn	0 0.00	1 100.00	0 0.00	0 0.00	0 0.00	1 100.00
Cotton	0 0.00	0 0.00	0 0.00	1 100.00	0 0.00	1 100.00
Soybeans	0 0.00	0 0.00	0 0.00	1 100.00	0 0.00	1 100.00
Sugarbeets	1 100.00	0 0.00	0 0.00	0 0.00	0 0.00	1 100.00
Total	1 20.00	1 20.00	1 20.00	2 40.00	0 0.00	5 100.00
Priors	0.30556	0.19444	0.16667	0.16667	0.16667	

Error Count Estimates for Crop						
	Clover	**Corn**	**Cotton**	**Soybeans**	**Sugarbeets**	**Total**
Rate	1.0000	0.0000	1.0000	0.0000	1.0000	0.6389
Priors	0.3056	0.1944	0.1667	0.1667	0.1667	

Output 36.4.5 Output Data Set of the Classification Results for Test Data

Discriminant Analysis of Remote Sensing Data on Five Crops
Output Classification Results of Test Data

Obs	Crop	x1	x2	x3	x4	xvalues	Clover	Corn	Cotton	Soybeans	Sugarbeets	_INTO_
1	Corn	16	27	31	33	16 27 31 33	0.08935	0.40543	0.17632	0.23918	0.08972	Corn
2	Soybeans	21	25	23	24	21 25 23 24	0.14811	0.24308	0.11999	0.33184	0.15698	Soybeans
3	Cotton	29	24	26	28	29 24 26 28	0.25213	0.18420	0.15294	0.25486	0.15588	Soybeans
4	Sugarbeets	54	23	21	54	54 23 21 54	0.62150	0.01937	0.12498	0.04962	0.18452	Clover
5	Clover	32	32	62	16	32 32 62 16	0.21633	0.31799	0.33266	0.11246	0.02056	Cotton

In this next example, PROC DISCRIM uses normal-theory methods (METHOD=NORMAL) assuming unequal variances (POOL=NO) for the remote-sensing data. The PRIORS statement, PRIORS PROP, sets the prior probabilities proportional to the sample sizes. The CROSSVALIDATE option displays cross validation error-rate estimates. Note that the total error count estimate by cross validation (0.5556) is much larger than the total error count estimate by resubstitution (0.1111). The following statements produce Output 36.4.6:

```
title2 'Using Quadratic Discriminant Function';

proc discrim data=crops method=normal pool=no crossvalidate;
   class Crop;
   priors prop;
   id xvalues;
   var x1-x4;
run;
```

Output 36.4.6 Quadratic Discriminant Function on Crop Data

Discriminant Analysis of Remote Sensing Data on Five Crops
Using Quadratic Discriminant Function

The DISCRIM Procedure

Total Sample Size	36	**DF Total**	35
Variables	4	**DF Within Classes**	31
Classes	5	**DF Between Classes**	4

Number of Observations Read	36
Number of Observations Used	36

Class Level Information

Crop	Variable Name	Frequency	Weight	Proportion	Prior Probability
Clover	Clover	11	11.0000	0.305556	0.305556
Corn	Corn	7	7.0000	0.194444	0.194444
Cotton	Cotton	6	6.0000	0.166667	0.166667
Soybeans	Soybeans	6	6.0000	0.166667	0.166667
Sugarbeets	Sugarbeets	6	6.0000	0.166667	0.166667

Output 36.4.6 *continued*

Within Covariance Matrix Information		
Crop	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
Clover	4	23.64618
Corn	4	11.13472
Cotton	4	13.23569
Soybeans	4	12.45263
Sugarbeets	4	17.76293

Discriminant Analysis of Remote Sensing Data on Five Crops Using Quadratic Discriminant Function

The DISCRIM Procedure

Generalized Squared Distance to Crop					
From Crop	Clover	Corn	Cotton	Soybeans	Sugarbeets
Clover	26.01743	1320	104.18297	194.10546	31.40816
Corn	27.73809	14.40994	150.50763	38.36252	25.55421
Cotton	26.38544	588.86232	16.81921	52.03266	37.15560
Soybeans	27.07134	46.42131	41.01631	16.03615	23.15920
Sugarbeets	26.80188	332.11563	43.98280	107.95676	21.34645

Discriminant Analysis of Remote Sensing Data on Five Crops Using Quadratic Discriminant Function

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CROPS
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Crop						
From Crop	Clover	Corn	Cotton	Soybeans	Sugarbeets	Total
Clover	9 81.82	0 0.00	0 0.00	0 0.00	2 18.18	11 100.00
Corn	0 0.00	7 100.00	0 0.00	0 0.00	0 0.00	7 100.00
Cotton	0 0.00	0 0.00	6 100.00	0 0.00	0 0.00	6 100.00
Soybeans	0 0.00	0 0.00	0 0.00	6 100.00	0 0.00	6 100.00
Sugarbeets	0 0.00	0 0.00	1 16.67	1 16.67	4 66.67	6 100.00
Total	9 25.00	7 19.44	7 19.44	7 19.44	6 16.67	36 100.00
Priors	0.30556	0.19444	0.16667	0.16667	0.16667	

Error Count Estimates for Crop						
	Clover	Corn	Cotton	Soybeans	Sugarbeets	Total
Rate	0.1818	0.0000	0.0000	0.0000	0.3333	0.1111
Priors	0.3056	0.1944	0.1667	0.1667	0.1667	

Output 36.4.6 *continued*

Discriminant Analysis of Remote Sensing Data on Five Crops
Using Quadratic Discriminant Function

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CROPS
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Crop

From Crop	Clover	Corn	Cotton	Soybeans	Sugarbeets	Total
Clover	9 81.82	0 0.00	0 0.00	0 0.00	2 18.18	11 100.00
Corn	3 42.86	2 28.57	0 0.00	0 0.00	2 28.57	7 100.00
Cotton	3 50.00	0 0.00	2 33.33	0 0.00	1 16.67	6 100.00
Soybeans	3 50.00	0 0.00	0 0.00	2 33.33	1 16.67	6 100.00
Sugarbeets	3 50.00	0 0.00	1 16.67	1 16.67	1 16.67	6 100.00
Total	21 58.33	2 5.56	3 8.33	3 8.33	7 19.44	36 100.00
Priors	0.30556	0.19444	0.16667	0.16667	0.16667	

Error Count Estimates for Crop

	Clover	Corn	Cotton	Soybeans	Sugarbeets	Total
Rate	0.1818	0.7143	0.6667	0.6667	0.8333	0.5556
Priors	0.3056	0.1944	0.1667	0.1667	0.1667	

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. 2nd ed. New York: John Wiley & Sons.

Epanechnikov, V. A. (1969). "Nonparametric Estimation of a Multivariate Probability Density." *Theory of Probability and Its Applications* 14:153–158.

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7:179–188.

Friedman, J. H., Bentley, J. L., and Finkel, R. A. (1977). "An Algorithm for Finding Best Matches in Logarithmic Expected Time." *ACM Transactions on Mathematical Software* 3:209–226.

Fukunaga, K., and Kessel, D. L. (1973). "Nonparametric Bayes Error Estimation Using Unclassified Samples." *IEEE Transactions on Information Theory* 19:434–440.

Glick, N. (1978). "Additive Estimators for Probabilities of Correct Classification." *Pattern Recognition* 10:211–222.

Hand, D. J. (1982). *Kernel Discriminant Analysis*. New York: Research Studies Press.

Hora, S. C., and Wilcox, J. B. (1982). “Estimation of Error Rates in Several-Population Discriminant Analysis.” *Journal of Marketing Research* 19:57–61.

Kendall, M. G., Stuart, A., and Ord, J. K. (1983). *The Advanced Theory of Statistics*. 4th ed. Vol. 3. New York: Macmillan.

Kshirsagar, A. M. (1972). *Multivariate Analysis*. New York: Marcel Dekker.

Lachenbruch, P. A., and Mickey, M. R. (1968). “Estimation of Error Rates in Discriminant Analysis.” *Technometrics* 10:1–10.

Lawley, D. N. (1959). “Tests of Significance in Canonical Analysis.” *Biometrika* 46:59–66.

Morrison, D. F. (1976). *Multivariate Statistical Methods*. 2nd ed. New York: McGraw-Hill.

Parzen, E. (1962). “On Estimation of a Probability Density Function and Mode.” *Annals of Mathematical Statistics* 33:1065–1076.

Perlman, M. D. (1980). “Unbiasedness of the Likelihood Ratio Tests for Equality of Several Covariance Matrices and Equality of Several Multivariate Normal Populations.” *Annals of Statistics* 8:247–263.

Puranen, J. (1917). “Fish Catch data set (1917).” Journal of Statistics Education Data Archive. Accessed May 22, 2009. `http://www.amstat.org/publications/jse/datasets/fishcatch.txt`.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. 2nd ed. New York: John Wiley & Sons.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rosenblatt, M. (1956). “Remarks on Some Nonparametric Estimates of a Density Function.” *Annals of Mathematical Statistics* 27:832–837.

Subject Index

Syntax Index