Variance Reduction with Monte Carlo Estimates of Error Rates in Multivariate Classification

C. Weihs*
Fachbereich Statistik
Universität Dortmund

G. Calzolari
Dipartimento di Statistica
Università degli studi di Firenze

M. C. Röhl
Königstein /Ts.

August 1999

Abstract

In this paper, control variates are proposed to speed up Monte Carlo Simulations to estimate expected error rates in multivariate classification.

KEY WORDS: classification, control variates, error rate, Monte Carlo Simulation, variance reduction

1 Introduction

The aim of this paper is to speed up Monte Carlo Simulations applied to multivariate classification. The most interesting performance measure in classification is the misclassification error.

In the case of given group densities, there are two possibilities to calculate the error rate: either by numerical integration or by Monte Carlo Simulation which is the only feasible method in higher dimensions. In this paper, we focus on the Monte Carlo error estimate. This approach suffers from the variability of the error rates, because the error rate is a random variable by now. Therefore, every principle to reduce this variance is welcome. In the literature various variance reduction techniques are proposed, among those antithetic variables and control variates (see, e.g., [1]). Here,

*D-44221 Dortmund, Germany, Tel. ++49-231-755-4363, e-mail: weihs@amadeus.statistik.uni-dortmund.de

we will concentrate on control variates and demonstrate their variance reduction potential in our special problem.

The paper is organized as follows: In section 2 we will give a brief introduction to multivariate classification. In section 3, we will propose two different control variates which will be studied and compared in section 4 by means of some examples. The paper closes with a conclusion in section 5.

2 Multivariate Classification

Classification deals with the allocation of objects to g predetermined groups $1, 2, \ldots,$ g, say. The aim is to minimize the misclassification error (rate) over all possible future allocations, given the group densities $p_i(x)$ $(i = 1, 2, \ldots, g)$. The minimal error rate is the so–called *Bayes error*.

We measure d features (variables) of the objects we consider important for discrimination between the objects. These can be continuous features (GNP, consumption etc.) or discrete (number of firms, number of inhabitants etc.).

Once the group densities are specified, in order to minimize the error rate we allocate an object with feature vector x to group i, if

$$p_i(x) > p_j(x) \quad (j \neq i). \tag{1}$$

Classification methods often assume the group densities $p_i(x)$ to be normal. Then there are at least two modelling possibilities (see, e.g., [2]):

- Estimate the same covariance matrix for all groups (LDA, linear discriminant analysis) or
- estimate a different covariance matrix for each group (QDA, quadratic discriminant analysis).

Of course, both methods also estimate different mean vectors for each group. In this paper we take QDA as the adequate, and thus standard classification procedure.

Often we additionally want to reduce the dimension from d to $d' = 1$ or 2 to enhance human perception (dimension reduction). The construction of a d'–space with minimal error rate, given the group densities $p_i(x)$ in d–space, can be done by modern optimization techniques, for example *Simulated Annealing* [3]. In each optimization step, a projection space is proposed. Then we determine the group densities (either estimated by means of the projected data or directly derived from the projected densities of the original space) [4], and calculate the error rate in the projection space.

In this paper, we suppose that the projection space is fixed, so that we already have the group densities available. Of course, the following approach can be also applied during optimization at each optimization step.

3 Variance Reduction by Control Variates

3.1 General Ideas

What we have to calculate is the error rate given the group densities. In one dimension, this can easily be done by numerical integration, because we only have to find the intersection points of the different group densities (determined by $p_i(x) = p_j(x)$) and then calculate integrals like

$$\int_a^b p_i(x)\,dx \quad a, b \in \mathbb{R}, \tag{2}$$

where $p_i(x)$ denotes an arbitrary known group density.

But in two or more dimensions, the borderlines between the group densities do not have that simple shapes, even when we assume equal group covariance matrices. Therefore, integration can only be done by means of a grid in two or more dimensions.

Another possibility to calculate the error rate is Monte Carlo Simulation. We generate random realizations from the group densities and allocate them according to our classification rule (1). This approach suffers from the variability of the error rates, because the error rate is a random variable by now.

In order to reduce the Monte Carlo variance of the error rate we introduce control variates (cv). The object of interest is the error rate $error$. We write this in a more complicated but helpful way as

$$error = error_{cv} + (error - error_{cv}) \tag{3}$$

with a new random variable $error_{cv}$. We want to compute the expectation of these error rates

$$\mathrm{E}(error) = \mathrm{E}(error_{cv}) + \mathrm{E}(error - error_{cv}). \tag{4}$$

The idea behind control variates is to choose a random variable $error_{cv}$ so that we can calculate $\mathrm{E}(error_{cv})$ exactly (no variance) and $error$ and $error_{cv}$ are positively correlated. So a sensible way of estimating $\mathrm{E}(error)$ would be

$$\hat{\mathrm{E}}_{cv}(error) = \mathrm{E}(error_{cv}) + \hat{\mathrm{E}}(error - error_{cv}), \tag{5}$$

where the first term on the right hand side has no variance and the second term is computed as the sample mean of Monte-Carlo replicates. Then the variance of the right hand side of (4) is

$$\mathrm{Var}(error - error_{cv})/N, \tag{6}$$

where N is the sample size to determine $error$ and $error_{cv}$, and

$$\begin{aligned}\mathrm{Var}(error - error_{cv}) = {} & \mathrm{Var}(error) + \mathrm{Var}(error_{cv}) \\ & - 2\,\mathrm{Cov}(error, error_{cv}).\end{aligned} \tag{7}$$

Now it becomes clear that a large positive correlation between $error$ and $error_{cv}$ can reduce the variance compared to the "naive" estimator $\hat{\mathrm{E}}_{MC}(error)$, i.e. the sample mean of Monte Carlo replicates of $error$ with variance $\mathrm{Var}(\hat{\mathrm{E}}_{MC}(error)) = \mathrm{Var}(error)/N$. We can even do better. We can use the equation

$$\mathrm{E}(error) = \alpha \mathrm{E}(error_{cv}) + \mathrm{E}(error - \alpha\, error_{cv}) \tag{8}$$

to select that parameter α that minimizes the variance

$$\mathrm{Var}(error - \alpha\, error_{cv}), \tag{9}$$

leading to

$$\alpha = \frac{\mathrm{Cov}(error, error_{cv})}{\mathrm{Var}(error_{cv})} \tag{10}$$

which is almost equal to the correlation coefficient ϱ when $\mathrm{Var}(error) \approx \mathrm{Var}(error_{cv})$ holds. The final result is then

$$\min_{\alpha} \mathrm{Var}(error - \alpha\, error_{cv}) \approx (1 - \varrho^2)\mathrm{Var}(error), \tag{11}$$

i.e. there can always be a gain when $\varrho \neq 0$.

Considering the above arguments, what we look for as a control variate procedure is any classification method which gives results as much as possible correlated with QDA, and for which the exact expected error rate is easily computable. Moreover, one should avoid control variates for which the additional computational effort is that high that overall computation time is increased even in the case of variance reduction.

3.2 Two Specific Control Variates

What is, then, a suitable control variate in our context? We will discuss two possibilities. In both cases the control variate procedure assumes a somewhat simplified problem situation to be true in order to simplify the Monte Carlo procedure. In the first procedure the covariance matrices of the different groups are assumed to be identical. In the second procedure the possibly high dimensional problem is optimally projected to one dimension. Note that we assumed to study problems with normal group densities with individual covariance matrices. Thus, QDA was assumed to be the standard classification method.

1. The first idea is to utilize the error rate computed by LDA as $error_{cv}$ based on the assumption of equal covariance matrices for all the groups. The error rate $error$ is calculated by QDA from N random realizations drawn from the group densities. To get $\hat{\mathrm{E}}_{MC}(error)$ we generate W such error rates and average. Therefore we used $N \times W$ random vectors. Now we take the same random vectors and apply LDA with the same, so-called pooled, covariance matrix for all groups to calculate error rates $error_{cv}$. If Σ_i is the assumed covariance matrix for group i, then $(\sum_{i=1}^{g}(N_i-1)\Sigma_i)/(N-g)$ is the pooled covariance matrix, where N_i is the number of realizations in group i. The differences of the W corresponding estimates $error$ and $error_{cv}$ are used to calculate $\hat{\mathrm{E}}(error - error_{cv})$. At last we calculate $\mathrm{E}(error_{cv})$ in an exact manner (so that we have no variance) by numerical integration based on the densities with pooled covariance matrices. We now have all the ingredients we need for an efficiency comparison with the naive Monte Carlo estimator. The variance of the naive estimator is calculated by the sample of size W of the estimated error rates $error$ and the variance of the control variate estimator by the sample of size W of $(error - error_{cv})$. This approach has the drawback that we have to calculate an exact integral in a projection space which might be two dimensional or of even higher dimension with rather ugly borderlines.

2. A second possibility to determine the error rate $error$ is to use another control variate: the error rate of an "optimal" one dimensional projection. This can be obtained by the largest eigenvalue and the corresponding eigenvector of QDA in the original space or by direct minimization of the error rate. We do the same as in 1 to obtain $\hat{\mathrm{E}}_{MC}(error)$. But then we project the random vectors onto the optimally discriminating direction taking into account the different covariance structures and build the differences of corresponding error estimates to compute $\hat{\mathrm{E}}(error - error_{cv})$. Now, the exact calculation of $\mathrm{E}(error_{cv})$ is simply a one dimensional integration with clearcut intersection points. This speeds up the procedure compared to 1. To construct the optimally discriminating one dimensional projection we follow an idea in [5] where it was proposed to project on the first eigenvector v_1 of MM^T, where

$$M = (\mu_g - \mu_1, ..., \mu_2 - \mu_1, \Sigma_g - \Sigma_1, ..., \Sigma_2 - \Sigma_1) \tag{12}$$

where the μ_i are the group means and the Σ_i are (again) the group covariance matrices, i = 1, ..., g. The projected means, variances and feature vectors then have the form: $\mu_i^* = v_1^T \mu_i$, $\sigma_i = v_1^T \Sigma_i v_1$ and $x^* = v_1^T x$.

In order to represent adequate control variates the additional computation time of procedures 1 and 2 have to be small relative to the computation time of naive Monte Carlo. That this is the case not considering the computation of the exact expected error rates should be clear by the following arguments.

- Naive Monte Carlo estimates the means and the covariance matrices of the groups, and evaluates the corresponding estimated group densities for each observation.

- Procedure 1 additionally needs to compute the mean of the estimated covariance matrices of the groups, and to evaluate group densities for each observation corresponding to the pooled covariance matrix in each group.
- Procedure 2 additionally computes the 'difference matrix' M, its first eigenvector v_1, and the corresponding projections of the group means and covariance matrices, and evaluates the corresponding 1D normal densities in each projected observation.

Therefore, in procedures 1 and 2 the 'preparation' of the density evaluation does not depend on the number of observations, resulting in a much smaller additional 'preparation time' than the preparation time for naive Monte Carlo for big numbers of observations. Moreover, in procedure 2 also the additional density evaluations are much quicker than the density evaluations in naive Monte Carlo, since they are in 1D.

In procedure 1 the exact expected error rates have to be calculated numerically, in general. For the exact expected error rates in procedure 2, however, an analytic formula can be derived, even. This will be done in the next subsection. In section 4 we will demonstrate the differences between procedures 1 and 2 by some examples.

3.3 Exact expected error rates

In procedure 2 exact expected error rates have to be calculated for univariate normal projected distributions. In this case a general formula for the exact expected error rate could be given depending on the intersection points of the univariate normal densities corresponding to the projected group means and variances. In order to illustrate the idea, let us discuss the 2 and 3 groups cases. Moreover, let us assume equal a-priori probabilities $1/g$ for all g groups for simplicity. In the simulations in the following sections, we also will discuss these cases only.

In the case of 2 groups let the intersection point of the two normal densities be x_{12}. Then, obviously, the exact expected error rate corresponding to these densities is (cp. figure 1)

$$\mathrm{E}(error_{cv}) = ((1 - \Phi_1(x_{12}) + \Phi_2(x_{12}))/2) \tag{13}$$

where Φ_1 is the normal distribution with mean to the left of x_{12}, and Φ_2 the distribution with mean to the right.

In the case of 3 groups let the distribution indices again be chosen so that a lower index indicates a lower mean. We are now interested in the relative location of the intersection points of the 3 densities. The error rate of the leftmost group 1 is determined by the first intersection on the right hand side with one of the densities of the other groups. For the rightmost group 3 the same is true for the densities on the left. The error rate of the middle group 2 is, correspondingly, determined by

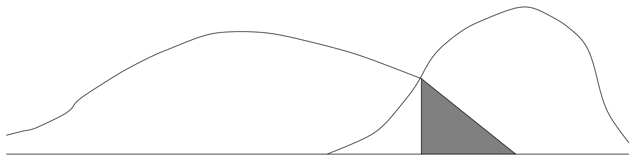

Figure 1: The error rate of the left group is gray shaded.

the first intersections points of its density with the other densities on the left and on the right. For simplicity let us now assume that the relevant intersection points are x_{12}, determining the error of group 1 and the 'left part' of the error of group 2, and x_{23}, determining the error of group 3 and the 'right part' of the error of group 2. This then leads to the following formula for the exact error rate corresponding to the 3 groups:

$$\mathrm{E}(error_{cv}) = ((1 - \Phi_1(x_{12})) + (\Phi_2(x_{12}) + (1 - \Phi_2(x_{23})) + \Phi_3(x_{23}))/3) \qquad (14)$$

As an example consider 3 groups with group means $\mu_1 = -3$, $\mu_2 = -2$, and $\mu_3 = 0$, and with standard deviations $\sigma_1 = 2.037$, $\sigma_2 = 0.9406$, and $\sigma_3 = 1$. These parameters lead to intersection points $x_{12} = -3.17$ and $x_{23} = -1$, as well as an exact error rate $\mathrm{E}(error_{cv}) = 31.45\%$.

For procedure 1 we only succeeded to find a general analytic formula for the exact expected error rate in the case of 2 groups. Procedure 1 assumes equal covariance matrices for all groups. This covariance matrix is estimated by the pooled covariance matrix $\hat{\Sigma} = (\hat{\Sigma}_1 + \hat{\Sigma}_2)/2$, where $\hat{\Sigma}_i$ is the estimated covariance matrix of group i, $i = 1, 2$. For normal group distributions with means μ_1 and μ_2 and a common covariance matrix Σ one can show (see [6], p. 12) that the exact error rate is

$$\mathrm{E}(error_{cv}) = \Phi(-0.5\delta_{12}) \quad \text{where} \quad \delta_{12} = \sqrt{(\mu_2 - \mu_1)^T \Sigma^{-1} (\mu_2 - \mu_1)} \qquad (15)$$

and Φ is the distribution function of the standard normal.

4 Simulations

4.1 Known densities

In this subsection we assume that the group densities are fully known so that parameter estimation is superfluous. This means in particular that QDA as well as LDA is carried out with the correct densities. In this way the outcome does not depend on the goodness of parameter estimation. In the next subsection, we will discuss the case when density parameters have to be estimated.

In all examples sample size N = 100 is used for each group. Also, W = 100 is used. In order to be independent of the drawn random vectors, this experiment was repeated V = 100 times and the means of the mean error rates and the corresponding standard deviations as well as the correlation coefficients will be reported in what follows.

First Simulation:
First we compare procedures 1 and 2 using two groups with the following parameters of normal distributions:

$$\mu_1 = (0,0)' \quad \text{and} \quad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{16}$$

as well as

$$\mu_2 = (2,0)' \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \tag{17}$$

The true expected error is approximately 14.97% calculated by exact integration to be able to judge the following results.

By means of the naive Monte Carlo estimator we obtain

$$\hat{\mathrm{E}}_{MC}(error) = (15.00 \pm 2.53)\% \tag{18}$$

where 2.53% is the estimated standard deviation of $\hat{\mathrm{E}}_{MC}(error)$. Obviously, the bias is negligible.

With procedure 2 one obtains

$$\hat{\mathrm{E}}(error - error_{cv}) = (-0.92 \pm 1.65)\% \tag{19}$$

and $\mathrm{E}(error_{cv})$ equals 15.87% (exact integration). Summing up for the right hand side of equation (5) we arrive at $(14.95 \pm 1.65)\%$. This expression shows a distinctly lower variance than (18).The mean estimated correlation coefficient ist $\varrho = 0.79$. The lowest standard deviation we can get by (8) is therefore 1.55%. This corresponds to a variance reduction of more than 60% in relation to the naive Monte Carlo.

Moreover, with procedure 1 one obtains

$$\hat{\mathrm{E}}(error - error_{cv}) = (-0.02 \pm 0.61)\% \tag{20}$$

and $\mathrm{E}(error_{cv})$ equals 15.08% (exact integration). Summing up for the right hand side of equation (5) we arrive at $(15.06 \pm 0.61)\%$. This expression shows an even much lower variance than with procedure 2. This indicates that LDA is a very adequate method for this example. Indeed, the mean estimated correlation coefficient is $\varrho = 0.97$.

Second Simulation:

Now we compare procedures 1 and 2 by an example with three different groups which do not separate that nicely as in the previous simulation. In addition to the

Sim	naiveMC	proc1	proc2	var1	var2	cor1	cor2	min1	min2	mvar1	mvar2
1	2.53	0.61	1.65	94%	57%	0.97	0.79	0.61	1.55	**94%**	62%
2	2.55	2.31	1.96	18%	41%	0.59	0.70	2.06	1.82	35%	**49%**
3	2.55	1.93	2.45	43%	8%	0.74	0.43	1.72	2.30	**55%**	9%

Table 1: Monte Carlo standard deviations, variance reductions, and correlations for known densities

two groups in the first simulation we use a third group with the following parameters of a normal distribution:

$$\mu_3 = (3,0)' \quad \text{and} \quad \Sigma_3 = \begin{pmatrix} 2 & -0.3 \\ -0.3 & 2.4 \end{pmatrix}. \tag{21}$$

The true expected error rate is approximately 28.44%.

The results of naive Monte Carlo and the two control variate procedures are summarized in Table 1. Note that 'Sim' indicates the simulation number, 'naiveMC' the estimated mean standard deviation of the naive Monte Carlo, 'proc1' and 'proc2' the corresponding standard deviations of the control variate procedures, 'var1' and 'var2' the corresponding percentages of variance reduction, 'cor1' and 'cor2' the mean correlation coefficients, 'min1' and 'min2' the corresponding minimal standard deviations of the control variate procedures, and 'mvar1' and 'mvar2' the corresponding maximum percentages of variance reduction.

Analysing Table 1 note particularly that for simulation 2 procedure 2 leads to a bigger variance reduction than procedure 1, but that the maximum variance reduction is nevertheless smaller than for simulation 1 since the univariate projected constellation of the groups is more complicated in this example. The bad performance of procedure 1 indicates that in this example the covariance matrices of the different groups can not be assumed to be approximately equal.

Third Simulation:

Up to now, the examples were mainly one dimensional in that the groups were shifted in the first component only. Since this might lead to an overoptimistic judgement of procedure 2, the third example is the same as the second, but with

$$\mu_3 = (1,1)' \tag{22}$$

i.e. with a mean shifted in both directions.

The corresponding Monte Carlo results can also be found in Table 1. Note in particular that now again procedure 1 is very adequate, but procedure 2, unfortunately, does not lead to a substantial variance reduction, and might thus even cause an increase in computer time. The problems of procedure 2 also become clear noting that the exact expected error rate is 45% for this procedure in contrast to a true expected error rate of around 33%.

As a preliminary conclusion this indicates that procedure 2 is useful probably only if a 1-dimensional projection does not alter the problem too much. Naturally, a similar statement is true for procedure 1, but the assumption of equal, or at least similar, covariance structures is, most of the time, not that much problematic. That the amount of variance reduction depends on the 'similarity' of the control variate to the true error rate could have already been deduced from equation (11). On the other hand, for procedure 1 the exact expected error rate is not easily computable, in general, and procedure 2 is much quicker.

4.2 Estimated densities

Since in practice the distributions of the grouped observations are not known, the implementation of the control variate procedures has somewhat to be adapted. For the purpose of this paper we nevertheless assume normal distributions for convenience so that only the corresponding distribution parameters have to be estimated.

Moreover, since the densities have to be estimated from the observations, the true expected misclassification rate has to be estimated by means of a resampling method in order to avoid overoptimism. As the resampling method we use leave-one-out cross validation. I.e. we preliminarily eliminate one observation, estimate the densities from the remaining observations, and predict the class of the eliminated observation by means of these estimated densities. This is done for each observation.

This causes two problems. First, the extra loop for resampling leads to such a big computational effort that the number of replicates of the whole Monte Carlo experiment is reduced to $V = 10$ for simulation 1 and to $V = 5$ otherwise. Second, the exact expectations in procedures 1 and 2 should not have to be computed for each resampled sample. Thus, we propose to compute the exact expectation from the "observed" sample only, and use this value for all resampled samples also. Moreover, for the purpose of the simulations for this paper we decided to use the exact expectations from the densities used to generate the observations in order to reduce computational effort. Finally, we used the same example densities as in the preceding section to be able to judge the extra variance caused by parameter estimation.

The results of the Monte Carlo simulations can be found in Table 2. Note the increase of variance and the very small correlation ϱ with procedure 2 in the third simulation. The optimal standard deviation reachable by (8) in this case would thus be 2.53, which is only very slightly lower than 2.57 with the naive Monte Carlo. Since, nevertheless, the results are very similar to the results of the simulations with known densities, the conclusions from the last subsection appear to be valid also in the case of density parameters to be estimated.

Sim	naiveMC	proc1	proc2	var1	var2	cor1	cor2	min1	min2	mvar1	mvar2
1	2.60	0.88	1.66	89%	59%	0.94	0.87	0.88	1.28	**89%**	76%
2	2.53	2.38	2.07	12%	33%	0.59	0.68	2.04	1.86	35%	**54%**
3	2.57	2.00	3.29	39%	-	0.73	0.17	1.76	2.53	**47%**	3%

Table 2: Monte Carlo standard deviations, variance reductions, and correlations for estimated densities

5 Conclusion

In this paper it was shown that the variance of the misclassification error rate estimated by Monte Carlo Simulation can be substantially reduced by means of control variates. The amount of variance reduction depends on the 'similarity' of the control variate to the true error rate. What one, thus, has to look for to construct an adequate control variate is a classification method with two properties: an error rate highly correlated to the true error rate, and an exact expected error rate which can be calculated easily. In other words, the method should be simple enough to be able to calculate the exact expected error rate easily, but also sophisticated enough to mimic the dependence of the true error rate on the data structure.

The main result of this paper can be stated as follows: Since it is relatively easy to compute its exact error rate, the error rate of normal group density approximations in the optimal 1D projection might be recommended as a control variate as long as the 1D projection sufficiently represents the high dimensional problem. This should be tested on the basis of the whole dataset.

Acknowledgment

The idea for this paper originated from a research visit of Prof. Calzolari in Dortmund in summer 1998. The paper was intensively worked on during a research visit of Prof. Weihs in Florence in spring 1999. We thank the Dipartimento di Statistica of the Università degli studi di Firenze for its kind support. This work has also been supported by the Collaborative Research Centre "Reduction of Complexity in Multivariate Data Structures" (SFB 475) of the German Research Foundation (DFG). Moreover, we would like to thank cand.stat. T. Hothorn for proposing the analytic formulas for the exact expected error rates and for programming in R.

References

[1] R. Y. Rubinstein and B. Melamed, *Modern Simulation and Modeling*, John Wiley & Sons, 89-97 (1998).

[2] G. J. McLachlan, *Discriminant Analysis and Statistical Recognition*, John Wiley & Sons (1992).

[3] I. O. Bohachevsky, M. E. Johnson, M. L. Stein, *Function Optimization*, Technometrics, **28**, 3, 209-217 (1986).

[4] M. C. Röhl and C. Weihs, *Optimal vs. Classical Linear Dimension Reduction*, in: W. Gaul, H. Locarek-Junge (eds.), Classification in the Information Age, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 252-259 (1999).

[5] D. M. Young, V. R. Marco and P. L. Odell, *Quadratic Discrimination: Some Results on Optimal Low-Dimensional Representation*, Journal of Statistical Planning and Inference, 17, 307-319 (1987).

[6] J. Läuter, *Stabile multivariate Verfahren* Mathematische Lehrbücher und Monographien 81, Akademie-Verlag, Berlin (1992).