Biometrika (2010), **97**, 1, *pp*. 254–259
doi: 10.1093/biomet/asp084
Advance Access publication 28 January 2010

The maximal data piling direction for discrimination

BY JEONGYOUN AHN

Department of Statistics, University of Georgia, Athens, Georgia 30602, U.S.A.
jyahn@stat.uga.edu

AND J. S. MARRON

Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.
marron@email.unc.edu

SUMMARY

We study a discriminant direction vector that generally exists only in high-dimension, low sample size settings. Projections of data onto this direction vector take on only two distinct values, one for each class. There exist infinitely many such directions in the subspace generated by the data; but the maximal data piling vector has the longest distance between the projections. This paper investigates mathematical properties and classification performance of this discrimination method.

Some key words: Classification; Fisher's linear discrimination; High dimension, low sample size; Maximal data piling; Support vector machine.

1. INTRODUCTION

Projection of high-dimension, low sample size data onto low-dimensional subspaces is often useful to understand structure in datasets, as effectively used in Marron et al. (2007). In binary discrimination, an interesting projection is to the normal direction vector of a separating hyperplane between the two classes. Assume that we have a dataset in $\mathcal{R}^d$ with sample size N and $d \geqslant N - 1$. Also assume that the data are not degenerate, in the sense of generating a subspace of dimension N. Then there exist direction vectors onto which the data project to only two distinct values, one for each class. This paper studies a direction vector which is optimal among these, in the sense that it maximizes the distance between the projected values. We call it the maximal data piling direction vector since the data are piled on two points while maximizing the distance in between.

Data piling was first discussed by Marron et al. (2007). They observed that the support vector machine classifier (Vapnik, 1998) yielded substantial data piling when it was applied to high-dimension, low sample size data; in particular, many data vectors lie on the margin boundary, i.e. become support vectors. Maximal data piling can be regarded as an extreme version of support vector machines since it has N support vectors all of which are exactly on the margin boundary. The margin obtained from support vector machines is always larger than or equal to that from maximal data piling, since support vector machines maximize the margin subject to $t_i(v^{\mathrm{T}} z_i + b) \geqslant 1$ and maximal data piling effectively maximizes the margin subject to a stricter constraint $t_i(v^{\mathrm{T}} z_i + b) = 1$ $(i = 1, \ldots, N)$, where $t_i = \pm 1$ is the class label, z_i is the input vector and $f(z) = v^{\mathrm{T}} z + b = 0$ is a separating hyperplane between the two classes.

Let us illustrate data piling with a real high-dimension, low sample size dataset. The leukemia microarray gene expression dataset in Golub et al. (1999) has two classes of cancer type, ALL and AML, and 7129 genes. Both training and testing datasets are available, with sample sizes 38 and 34, respectively. Figure 1 shows the projections of the data onto the normal direction vector of the support vector machine separating

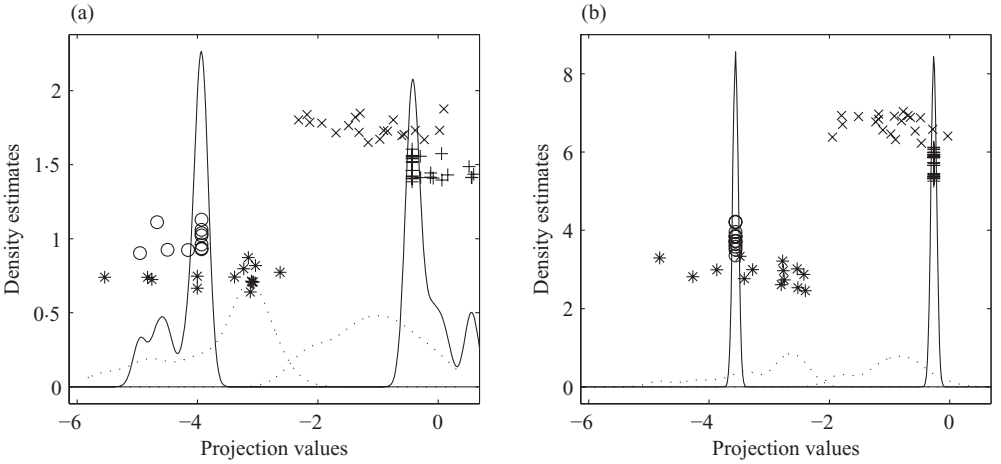

Fig. 1. Projections of the leukemia microarray data onto the support vector machine (a) and the maximal data piling (b) directions. ALL training (+), AML training (o), ALL testing (x), and AML testing (*) data are plotted with random y-coordinates. Kernel density estimates are drawn to show the distributions of the projections for each group. Solid curves are the density estimates based on the training data and dotted curves are for the testing data. (a) shows substantial data piling and (b) shows complete data piling.

hyperplane in (a) and onto the maximal data piling direction vector in (b). Random y-coordinates are used for visual separation of the data.

Figure 1(a) shows that many training data points are support vectors on the margin boundary. Figure 1(b) shows that the training data vectors are piled completely at two points. The projected testing data are perfectly separable in both panels, which means zero test error for both classification methods.

In general, as discussed in Marron et al. (2007), data piling may not be desirable because it overfits the data by incorporating noise artifacts in order to obtain data piling. However, we will see that this seemingly artificial direction vector can work surprisingly well for some underlying distributional settings, and for some real data examples.

2. Maximal data piling direction vector

2·1. *Definition*

Suppose that we have a Class $+1$ sample $x_1, \ldots, x_m$ and a Class -1 sample $y_1, \ldots, y_n$ in $\mathcal{R}^d$. Assume that $d \geqslant N - 1 = m + n - 1$. Let X and Y be data matrices for each class with data vectors as columns. Let X_c and Y_c denote their centred versions and let $C = [X_c, Y_c]$ be the horizontal concatenation of the two matrices. Let $w = \bar{x} - \bar{y}$ denote the group mean difference vector.

The maximal data piling vector v_{M} can be obtained from the optimization problem of finding v that maximizes the difference between the projected class means $(v^{\mathrm{T}}w)^2$ subject to the data piling constraint $C^{\mathrm{T}}v = 0$ which insists that the projection of each data point onto v is the same as its class mean and the normalization constraint $v^{\mathrm{T}}v = 1$. The symmetric projection matrix on to the orthogonal complement of the column space of C is $Q = I_d - CC^{\dagger}$, where $A^{\dagger}$ is the Moore–Penrose generalized inverse of A. By page 245 in Searle (1982), the solution to the piling constraint is $v = Qr$, where r is an arbitrary vector in $\mathcal{R}^d$. Then it is straightforward to see that the optimal v maximizing the objective function $(v^{\mathrm{T}}w)^2$ is

$$v_{\mathrm{M}} \propto Qw. \tag{1}$$

Geometrically, the maximal data piling vector v_{M} has an explicit interpretation in the data space and is uniquely defined within the subspace generated by the data. It lies within the $(N-1)$-dimensional subspace generated by the globally centred data vectors, while being orthogonal to the $(N-2)$-dimensional subspace generated by the class-wise centred data vectors. Because of its explicit geometrical interpretation, maximal

data piling can give useful information on the structure of high-dimensional data. For example, a small piling distance suggests some possibly mislabelled data points. Also, since the distance between the piling sites is a natural measure of the distance between two groups of high-dimension, low sample size data, maximal data piling can easily be incorporated into any multivariate method for which proximity between groups of observations is of interest, for example clustering methods.

There are some insightful formulas that are equivalent to (1). Let Z_c denote the centred version of the data matrix $Z = [X, Y]$. The following formula for the maximal data piling direction,

$$v_{\mathrm{M}} \propto \left(Z_c Z_c^{\mathrm{T}}\right)^{\dagger} w, \tag{2}$$

is a simple replacement of the pooled sample covariance matrix in Fisher's discriminant vector (Fisher, 1936) by the global sample covariance matrix. When $d \leqslant N-2$, (2) is equivalent to Fisher's discriminant vector, which does not have the piling property. When $d \geqslant N-1$, (2) is equivalent to (1) and yields complete data piling. While Fisher's method is invariant under affine transformation of data when $d \leqslant N-2$, (2) is not. However, we can view maximal data piling as an appropriate high-dimension, low sample size version of Fisher's linear discrimination in the sense that it yields data projections with zero within-class scatter and maximized between-classes scatter.

The equivalence between the seemingly unrelated formulas (1) and (2) can be understood in a unified framework using the following argument. We double-rotate the matrix Z_c so that it takes the form

$$Z^* = U Z_c V^{\mathrm{T}} = \begin{pmatrix} F & s_1 & 0 \\ 0 & s_0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

where F is an $(N-2)$-dimensional square matrix and nonsingular part of C, s_1 is $(N-2)$-vector, s_0 is a scalar and $[s_1^{\mathrm{T}}, s_0]^{\mathrm{T}}$ is the rotated w. The key properties of the orthogonal matrices U and V are as follows: as for the left rotation matrix U, the first $q = N-2$ columns span the column space of C and the first $q+1$ columns span the column space of Z_c; as for the right rotation matrix V, the first q columns span the row space of C, the $(q+1)$th column is proportional to $[n1_m^{\mathrm{T}}, -m1_n^{\mathrm{T}}]^{\mathrm{T}}$, and the last column is proportional to 1_N.

Let w^* denote the rotated mean difference vector Uw, which is proportional to the $(N-1)$th column of Z^*. Then the objective function $(v^{\mathrm{T}}w)^2$ equals $(v^{*\mathrm{T}}w^*)^2$, where $v^* = Uv$. The piling constraint $C^{\mathrm{T}}v = 0$ implies that the first $(N-2)$ components of v^* are zeros. Hence the optimal v takes the simple form $v^* \propto [0_{N-2}^{\mathrm{T}}, 1, 0_{d-N-1}^{\mathrm{T}}]^{\mathrm{T}}$, where 0_n is an n-vector of zeros. It is also straightforward to see that the formula (2) implies the same form in the rotated data space.

The above formula (2) can be simplified to $v_{\mathrm{M}} \propto Z_c^{\mathrm{T}\dagger} t$, where $t = [1_m^{\mathrm{T}}, -1_n^{\mathrm{T}}]^{\mathrm{T}}$ is the N-vector of class labels. Then projections of all the Class ± 1 data vectors onto v_{M} are $\{\pm 1 - (m-n)/N - \bar{z}^{\mathrm{T}} v_{\mathrm{M}}\}/\|Z_c^{\mathrm{T}\dagger} t\|$, where $\bar{z}$ is the overall mean vector and $\|\cdot\|$ is the L_2 norm. Thus the distance between the piling sites is $2/\|Z_c^{\mathrm{T}\dagger} t\|$. Any vector of the form $v = (Z-G)^{\mathrm{T}\dagger} t$, where $G = g1_N^{\mathrm{T}}$ and g is any vector in $\mathcal{R}^d$, will yield complete data piling. However, the maximal data piling vector, with $g = \bar{z}$, has the largest distance between the piling sites.

For the classification task, we use the sign of $f(z) = v_{\mathrm{M}}^{\mathrm{T}} z + b$ as the assigned class label for a new data point z. We used $b = -v_{\mathrm{M}}^{\mathrm{T}}(m\bar{x} + n\bar{y})/N$ for the examples in the present paper, but it can also be determined by considering sampling probabilities.

2·2. *Multiclass data piling*

In general when there are $K > 2$ classes, complete data piling means that the projections of the whole dataset are piled onto only K distinct points in a $(K-1)$-dimensional subspace. The maximal data piling subspace, denoted by $\mathcal{S}_{\mathrm{M}}$, has the projection values with the largest in-between distances.

Let $X_1, \ldots, X_K$ be the $d \times n_k$ data matrices for Classes $1, \ldots, K$. The subspace $\mathcal{S}_{\mathrm{M}}$ can be constructed from maximal data piling direction vectors. Let $v_{\mathrm{M}}^{(k)}$ be the maximal data piling direction vector for the binary discrimination of Class k versus the rest. Then it can be shown that $\mathcal{S}_{\mathrm{M}}$ is spanned by any $K-1$ vectors out of $v_{\mathrm{M}}^{(1)}, \ldots, v_{\mathrm{M}}^{(K)}$ and also $v_{\mathrm{M}}^{(k)} = \sum_{j \neq k}^{K} a_j v_{\mathrm{M}}^{(j)}$, where $a_j = n_j(N-n_j)\{n_k(N-n_k)\}^{-1}$. For

classification, we obtain $f_k(x) = v_k^{\mathrm{T}} x + b_k$ $(k = 1, \ldots, K)$, and then assign a new data point to the ℓth class if $f_\ell = \max_{1 \leqslant k \leqslant K} f_k$.

3. Data examples

3·1. *Simulation*

For a wide range of dimensions $d = 2, 10, \ldots, 500, 1000$, $m = n = 25$ Gaussian samples were generated from two classes which are different only in their population means. We considered the compound symmetry structure for underlying covariance matrices $\Sigma_{d,\rho} = (1 - \rho) I_d + \rho 1_d 1_d^{\mathrm{T}}$, where $\rho = 0, 0{\cdot}25, 0{\cdot}50, 0{\cdot}75$. The population mean for the Class $+1(-1)$ was $[+\mu_0(-\mu_0), 0, \ldots, 0]^{\mathrm{T}}$, where the constant μ_0 was determined so that the Mahalanobis distance between the two classes is maintained at 4·4. For each run, misclassification rates were evaluated with an independent testing dataset of size 10 000.

In Fig. 2, maximal data piling is compared with some simple prototype methods such as the nearest centroid method, Fisher's linear discrimination, and the naïve Bayes method in the left panels and also with some complex methods such as support vector machines, distance weighted discrimination (Marron et al., 2007), and regularized logistic regression (le Cessie & van Houwelingen, 1992) in the right panels. For Fisher's method, the Moore–Penrose generalized inverse was used for $d \geqslant 50$. The tuning parameters of the three complex methods were chosen using an independent tuning dataset of the same size as the training dataset. The logarithm of the mean misclassification rates with error bars from 100 replications are shown in the figure.

As pointed out in § 2·1, maximal data piling and Fisher's method are identical when $d \leqslant N - 2$ if one uses the formula (2), and they show very poor performance when $d = N$. From the practical point of view, this fact strongly discourages the application of Fisher's method when the dimension is about the same as the sample size. A theoretical explanation of this can be found in Deev (1970), where it was shown that the misclassification error of Fisher's method converges to $1/2$ as N increases and $d \approx N$, due to the accumulation of errors in estimating the unknown inverse covariance matrix. As Bickel & Levina (2004) explained, Fisher's method performs poorly as d tends to infinity, regardless of the value of ρ.

The classification performance of maximal data piling is always by far the best in the left panels for higher dimensions except when $\rho = 0$, where the nearest centroid method is Bayes optimal. In the right panels which compare with much more sophisticated methods, maximal data piling shows surprisingly competitive performance for higher dimensions, especially when ρ is large.

When $\rho = 0$, all methods except Fisher's showed similar performance for large d. This can be explained by the geometric representation of high-dimension, low sample size data by Hall et al. (2005) and Ahn et al. (2007). They found that as d increases, under some conditions, the geometrical structure of the data becomes rigid and the data vectors asymptotically form a regular simplex. In binary discrimination, the data from each class form two simplices, thus any reasonable discrimination method will find the separating hyperplane with the normal vector bisecting the two simplices, which is exactly what maximal data piling does. When ρ is large, the strong correlations among variables make the data more concentrated and behave lower-dimensionally, thus regularization by maximal data piling is essential.

A simulation study with different ρs for each class, not shown here to save space, also achieved similar results. The performance of maximal data piling was comparable to the complex methods except regularized logistic regression, which performed significantly better than others.

3·2. *Text classification*

We used the text classification data from the Text Retrieval Conference (Zhao & Karypis, 2005), available at www-users.cs.umn.edu/∼karypis/cluto. We selected three datasets with relatively small dimensionalities for ease of computation. The classes with size less than 30 were deleted in order to maintain enough data for five-fold crossvalidation. After this preprocessing, the selected three datasets Tr23, Tr41, and Tr45 have three, seven, and eight classes, with dimensionalities 5832, 7454, and 8261, and with sample sizes 172, 825, and 658, respectively. We divided each dataset into training and testing datasets of the same

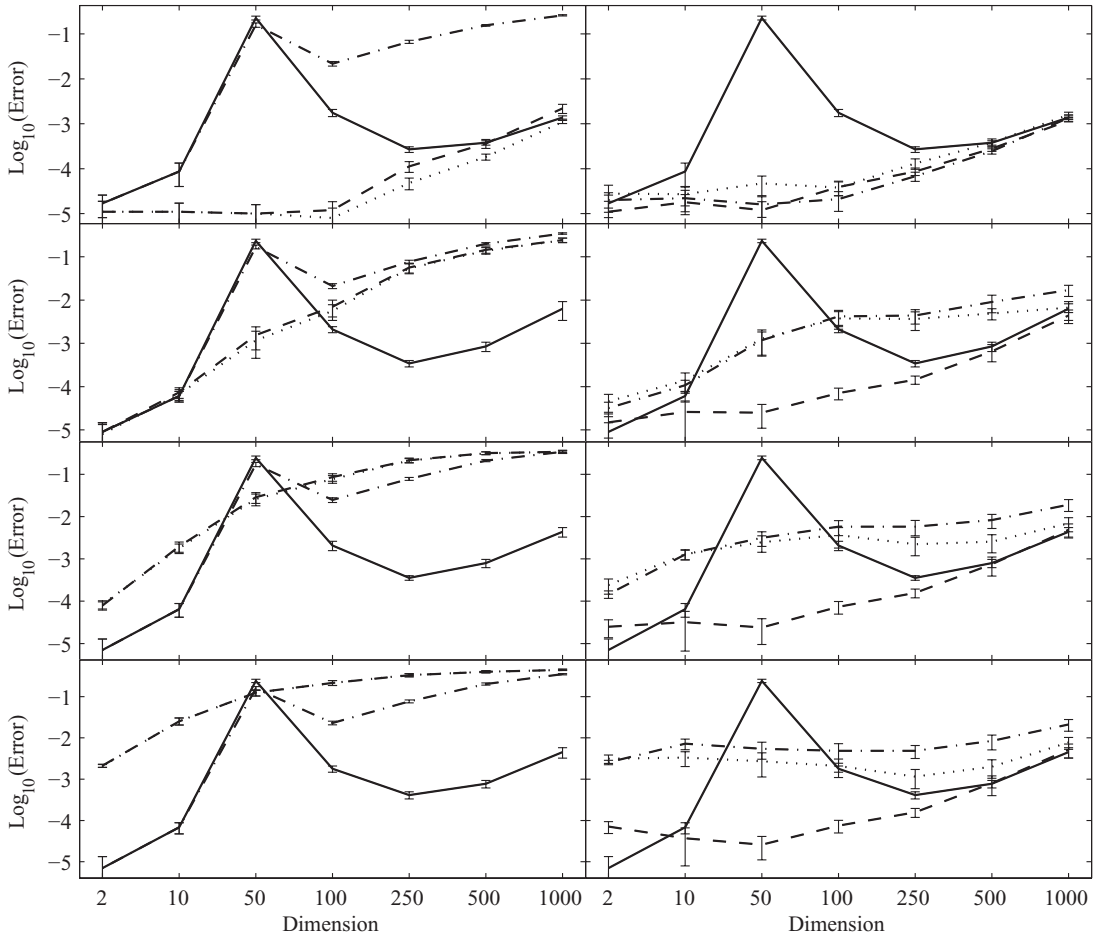

Fig. 2. Result of the simulated data for $\rho = 0, 0{\cdot}25, 0{\cdot}50, 0{\cdot}75$ from the first to the fourth row, respectively. Logarithms of mean misclassification rates with error bars (±standard error) are displayed. Left panels compare maximal data piling (solid) with nearest centroid (dotted), Fisher's linear discriminant (dot-dashed), and naïve Bayes (dashed). Right panels compare maximal data piling (solid) with support vector machines (dotted), distance weighted discrimination (dot-dashed), and regularized logistic regression (dashed).

Table 1. *The mean error rates (standard error) of text classification data*

Dataset	MDP	SVM	DWD	RLR
Tr23	0·2113 (0·0151)	0·1520 (0·0100)	0·1478 (0·0086)	0·1441 (0·0192)
Tr41	0·0819 (0·0065)	0·0503 (0·0039)	0·0432 (0·0046)	0·0579 (0·0038)
Tr45	0·1183 (0·0042)	0·1047 (0·0038)	0·0754 (0·0027)	0·0815 (0·0019)

MDP, maximal data piling; SVM, support vector machines; DWD, distance weighted discrimination (Marron et al., 2007); RLR, regularized logistic regression (le Cessie & van Houwelingen, 1992).

size, which was replicated 10 times. Five-fold crossvalidation was used for hyperparameter tuning for all methods except maximal data piling.

The mean test misclassification rates with standard error are shown in Table 1. Some methods are better than others for some datasets, although not significantly. Even though the maximal data piling shows the weakest performance in terms of the means, it is within the margin of error for Tr23 and Tr45. This result suggests some serious correlations in the datasets, which makes good sense since the variables are frequencies of words in a document.

ACKNOWLEDGEMENT

This research was partly supported by the National Science Foundation, U.S.A. The authors are thankful to Junyong Park for helpful suggestions and to the associate editor for helpful comments that have significantly improved the paper.

REFERENCES

AHN, J., MARRON, J. S., MULLER, K. E. & CHI, Y. Y. (2007). High dimension, low sample size geometric representation holds under mild conditions. *Biometrika* **3**, 760–6.

BICKEL, P. & LEVINA, E. (2004). Some theory for Fisher's linear discriminant function, 'naïve Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10**, 989–1010.

DEEV, A. D. (1970). Representation of the discriminant analysis statistics and asymptotic expansions when the space dimension is comparable with sample size. *Rep. Acad. Sci. USSR* **195**, 756–62.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–88.

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D. & LANDER, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–7.

HALL, P., MARRON, J. S. & NEEMAN, A. (2005). Geometric representation of high dimension low sample size data. *J. R. Statist. Soc.* B **67**, 427–44.

LE CESSIE, S. & VAN HOUWELINGEN, J. C. (1992). Ridge estimators in logistic regression. *Appl. Statist.* **41**, 191–201.

MARRON, J. S., TODD, M. J. & AHN, J. (2007). Distance-weighted discrimination. *J. Am. Statist. Assoc.* **102**, 1267–71.

SEARLE, S. R. (1982). *Matrix Algebra Useful for Statistics*. New York: Wiley.

VAPNIK, V. N. (1998). *Statistical Learning Theory*. New York: Wiley.

ZHAO, Y. & KARYPIS, G. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining Know. Disc.* **10**, 141–68.

[*Received January* 2008. *Revised September* 2009]