

Impact of feature correlations on separation between bivariate normal distributions

Krzysztof Kryszczuk
IBM Zurich Research Laboratory
kkkr@zurich.ibm.com

Andrzej Drygajlo
Swiss Federal Institute of Technology Lausanne (EPFL)
andrzej.drygajlo@epfl.ch

Abstract

The impact of feature correlations on class separation has received limited attention from researchers. Previous reports treat the problem from the viewpoint of multi-classifier fusion and are partially inconsistent in their conclusions. In this paper we show that these ambiguities are the result of incompatible basic assumptions, and that the conclusions from prior art hold only for specific configurations of class-conditional distributions. We show that the impact of feature correlations on class separation between two bivariate normal distributions can be positive or negative, and that it can only be gauged in the context of the parameters of involved marginals. The findings reported in this paper are of importance for the practice of feature extraction, feature selection, and in multi-classifier fusion.

1. Introduction

Classifiers deployed in many applications of pattern recognition use multiple features in order to best separate the classes. It is a well-known fact that individual class-selective power of the features is not the only factor that impacts class separation, and that correlations between the features also play an important role. Recognizing the importance of the topic in the pattern recognition practice, several authors have reported on the impact of feature correlations on class separation. Historically, the predominant notion was that classification features should not be correlated [3].

Recently, the topic has received some attention in the domain of multi-classifier fusion. In [8] Poh and Bengio studied the problem of feature correlation in the

context of biometric authentication. Here, the correlation between the output scores of unimodal classifiers are considered in a multimodal fusion scenario, where the scores become features for a fusion classifier. The authors conclude that positive correlation between the features helps, while negative correlation 'hurts' fusion. The authors assume that the marginal distributions of considered features are normal, and that correlation coefficients are equal for both classes.

In contrast with the results of Poh and Bengio are the results of Koval et al. [5], who argued that using dependent rather than independent, normally distributed features offers prospects of lower classification error rates. The authors illustrate their claim using an example where class separation between two bivariate normal distributions increases as the correlation coefficient changes from 0 (independent case) towards 1. Again, a tacit assumption assigns equal correlation coefficient to both classes.

Several other studies, for instance [9, 10], clearly suggest that it is important to properly account for the existence of correlation between features when designing a classifier, but at the same time they do not shed much light on the actual effect of feature correlation on class separation, nor on the empirical classification errors. The interest in accounting for feature correlations is pragmatically justified: in practice one usually cannot change the actual feature correlations - they can only be observed in the data.

However, understanding of the impact of feature correlations on class separation also has important practical applications. In [6] existence of dependences between class-selective features (baseline classifier scores) and class-independent quality measures is shown to warrant conditional relevance of the latter one, in a biometric application. In fact, in the well-known in machine learning literature *spouse problem* [2] the features are clearly correlated. Understanding the impact of feature correlations on class separation is also of direct relevance to feature generation and feature selection, as

The research reported in this paper was conducted when K. Kryszczuk was with the Swiss Federal Institute of Technology Lausanne (EPFL).

well as for the selection of individual classifiers selected for multi-classifier fusion [7].

Therefore it is important to disambiguate the apparently contradicting results reported in [8] and [5], and doing so is the motivation of this paper. We assume a two-class problem and normal class-conditional feature distributions. In Section 2 we give analytical results for computing class separation in the case when correlation coefficients and covariance matrices are equal for both classes. In Section 3 we show numerical simulation results for more general cases, where these simplifying assumptions are relaxed. Obtained results show that the results of [8] and [5] hold only for specific configurations of distributions and their parameters. We show that given specific conditions, the existence of non-zero feature correlations can either increase or decrease class separation, and we provide formal and intuitive explanations of these findings. In particular, using bivariate normal distributions we show that 1) uncorrelated features do not entail minimal class separation, and 2) the sign of the correlation coefficient does not determine the sign of change in class separation.

2. Bivariate Gaussian case, $\rho_A = \rho_B = \rho$

Let us analyse how correlation between features impacts class separation between classes A and B for bivariate normal class-dependent jointly distributed random variables X and Y , whose instances we denote as x and y , respectively. For that we assume that class-conditional marginal distributions of $p(x|A)$, $p(x|B)$, $p(y|A)$ and $p(y|B)$ are normal, $p(f|\omega) = \mathcal{N}(\mu_{f,\omega}, \sigma_{f,\omega}^2)$, where $f \in \{x, y\}$, $\omega \in \{A, B\}$, and $\mu_{f,\omega}$ and $\sigma_{f,\omega}^2$ are respective means and variances. Without a loss of generality let us place the means of the $p(x, y|A)$ at the origin of the respective axes, hence $\mu_{x,B} = \mu_x$ and $\mu_{y,B} = \mu_y$ become distances between respective means. Let us introduce simplifying assumptions: $\sigma_{y,A}^2 = \sigma_{y,B}^2 = \sigma_y^2$ and $\sigma_{x,A}^2 = \sigma_{x,B}^2 = \sigma_x^2$. Assume the Pearson's correlation coefficient [1] between X and Y to be the same for both classes A and B , $\rho_A = \rho_B = \rho$. Bivariate conditional distributions of $p(x, y|A)$ and $p(x, y|B)$ whose marginal distributions are normal are given by [4]:

$$\begin{aligned} p(x, y|A) &= \frac{e^{\left(-\frac{1}{2-2\rho_A^2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2xy}{\sigma_x\sigma_y} \right)\right)}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_A^2}} \\ p(x, y|B) &= \frac{e^{\left(-\frac{1}{2-2\rho_B^2} \left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2(x-\mu_x)y}{\sigma_x\sigma_y} \right)\right)}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_B^2}} \end{aligned} \quad (1)$$

In Eq. (1) the correlation coefficients are constrained to satisfy $-1 < \rho_\omega < 1$ in order to avoid a degenerate case where $x \propto y$ [1]. We are interested how the class separation, and consequently the expected minimal classification error¹ depend on the correlation coefficients ρ_A and ρ_B between y and x .

Consider the Kullback-Leibler divergence between two distributions, $u(\mathbf{x})$ and $v(\mathbf{x})$ [1]:

$$D_{KL}(u(\mathbf{x}), v(\mathbf{x})) = \int_{-\infty}^{\infty} v(\mathbf{x}) \ln \frac{v(\mathbf{x})}{u(\mathbf{x})} d\mathbf{x} \quad (2)$$

For multivariate Gaussian distributions Eq. (2) becomes [5]:

$$\begin{aligned} D_{KL}(p(\mathbf{e}|A), p(\mathbf{e}|B)) &= \\ \ln \frac{|\Sigma_A|}{|\Sigma_B|} + \text{tr}(\Sigma_A^{-1}\Sigma_B) + (\mu_A - \mu_B)^T \Sigma_B^{-1}(\mu_A - \mu_B), \end{aligned} \quad (3)$$

where Σ_A and Σ_B are covariance matrices of A and B and $\mu_A = [\mu_{x,A} \mu_{y,A}]^T$, $\mu_B = [\mu_{x,B} \mu_{y,B}]^T$. In general case divergence is asymmetric, $D_{KL}(A, B) \neq D_{KL}(B, A)$ but in our case $\Sigma_A = \Sigma_B = \Sigma$. In this situation Eq. (4) becomes:

$$\begin{aligned} D_{KL}(p(\mathbf{e}|A), p(\mathbf{e}|B)) &= D_{KL}(p(\mathbf{e}|B), p(\mathbf{e}|A)) = \\ &= (\mu_B)^T \Sigma^{-1}(\mu_A) + \beta, \end{aligned} \quad (4)$$

where

$$\beta = \ln \frac{|\Sigma|}{|\Sigma|} + \text{tr}(\Sigma^{-1}\Sigma) = \text{const.}$$

First let us represent Eq. (4) as an explicit function of ρ :

$$D_{KL}(\rho) = \frac{1}{1-\rho^2} \left(\frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_y^2}{\sigma_y^2} - \frac{2\rho\mu_x\mu_y}{\sigma_x\sigma_y} \right)$$

In order to find characteristic points of $D_{KL}(\rho)$ we compute the 1st and the 2nd derivatives of $D_{KL}(\rho)$:

$$\begin{aligned} \frac{d}{d\rho} D_{KL}(\rho) &= \frac{\rho}{(1-\rho^2)^2} \left(\frac{2\mu_x^2}{\sigma_x^2} + \frac{2\mu_y^2}{\sigma_y^2} - \frac{4\rho\mu_x\mu_y}{\sigma_x\sigma_y} \right) - \\ &\quad - \frac{2\mu_x\mu_y}{\sigma_x\sigma_y(1-\rho^2)}, \end{aligned} \quad (5)$$

¹assuming equal misclassification cost for both classes.

$$\frac{d^2}{d\rho^2}D_{KL}(\rho) = \left(\frac{2\mu_x^2}{\sigma_x^2} + \frac{2\mu_y^2}{\sigma_y^2} - \frac{4\rho\mu_x\mu_y}{\sigma_x\sigma_y} \right) \left(\frac{1}{(1-\rho^2)^2} + \frac{4\rho^2}{(1-\rho^2)^2} \right) - \frac{8\rho\mu_x\mu_y}{\sigma_x\sigma_y(1-\rho^2)}. \quad (6)$$

In order to find the extrema of $D_{KL}(\rho)$ we need to find ρ for which the first derivative of $D_{KL}(\rho)$ given by 5 is equal zero.

$$\frac{d}{d\rho}D_{KL}(\rho) = 0$$

$$\frac{\rho}{(1-\rho^2)} \left(\frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_y^2}{\sigma_y^2} - \frac{2\rho\mu_x\mu_y}{\sigma_x\sigma_y} \right) - \frac{\mu_x\mu_y}{\sigma_x\sigma_y} = 0 \quad (7)$$

$$(\mu_x\sigma_y - \rho\mu_y\sigma_x)(\mu_y\sigma_x - \rho\mu_x\sigma_y) = 0.$$

Solution of the above equation yields

$$\rho_1 = \frac{\mu_y\sigma_x}{\mu_x\sigma_y}, \rho_2 = \frac{\mu_x\sigma_y}{\mu_y\sigma_x}. \quad (8)$$

Note that since by definition $-1 < \rho < 1$ then $|\rho_1| < 1 \Leftrightarrow |\rho_2| > 1$ and Eq. (4) has only one valid solution. Let us assume that $|\rho_1| < 1$. At this point Eq. (6) evaluates to

$$\frac{d}{d\rho}D_{KL}(\rho_1) = \frac{2\sigma_y^2\mu_x^4}{(\sigma_y^2\mu_x^2 - \sigma_x^2\mu_y^2)\sigma_x^2}. \quad (9)$$

Since $-1 < \rho_1 < 1$ then necessarily

$$\sigma_y^2\mu_x^2 - \sigma_x^2\mu_y^2 > 0 \Rightarrow \frac{d}{d\rho}D_{KL}(\rho_1) > 0, \quad (10)$$

which indicates that $D_{KL}(\rho)$ has a minimum and the separation between classes A and B is minimal at ρ_1 . Should $|\rho_2| < 1$ be assumed then ρ_2 would be the one and only one valid solution that minimizes D_{KL} .

For most distributions, whose variances are non-zero, feature independence results in minimal $D_{KL}(\rho)$ only for $\mu_x = 0$ or $\mu_y = 0$. However, if neither $\mu_x = 0$ nor $\mu_y = 0$ then correlation between the features will either increase or decrease class separation in respect to the independent case. This situation is shown in Figure 1(a), where $\Sigma_A = \Sigma_B$, $\rho_A = \rho_B = \rho$, $\mu_x \neq 0$ and $0 \leq \mu_y \leq 3$. In this example, we show the impact of the correlation coefficient $\rho = \rho_A = \rho_B$ and of μ_y on the class overlap, measured by the Matusita distance

$$E = \int_y \int_x \sqrt{p(x, y|A)p(x, y|B)} dx dy. \quad (11)$$

The maxima of $E(\rho)$ reached for each assumed μ_y are connected with a bold black line. The example clearly illustrates the theoretical results from this section, and supports our hypotheses worded in Section 1. In particular, for $\mu_y = 0$ the shape of $E(\rho)$ supports the findings from [5], but not for $\mu_y > 0$. Also, the findings reported in [8] hold only for a specific case where $\mu_x = \mu_y = 1$, but not for other values of μ_y .

3. Bivariate Gaussian case, $\rho_A \neq \rho_B$

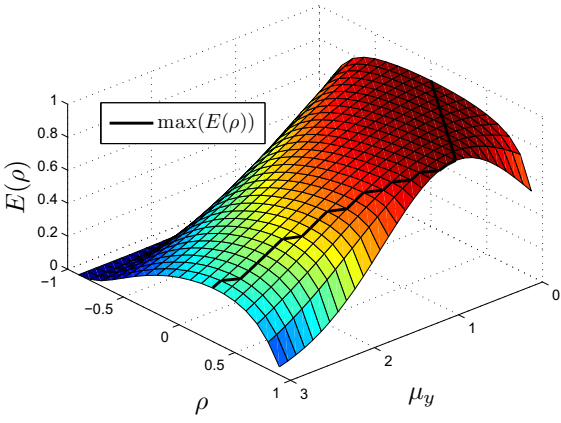
Let us now focus on the class overlap given by Eq. (11) for $\rho_A \neq \rho_B$. The detailed analysis of the conditions which must be met for to reach an extremum for arbitrary covariance configuration and when $\rho_A \neq \rho_B$ is symbolically complex beyond the frames of this paper and therefore we decide to skip it, and revert to numerical simulations instead. Given the number of free parameters of arbitrary bivariate normal distributions, we are forced to limit the number of simulations to a few examples, necessary to substantiate our claims worded in Section 1. In these examples, shown in Figures 1(b) and (c), the error measure $E(\rho_A, \rho_B)$ is plotted against the correlation coefficients ρ_A and ρ_B for fixed parameters of the marginals $p(x, y|A)$, $p(x, y|B)$.

The examples shown in Figures 1(b) and (c) show that the function $E(\rho_A, \rho_B)$ is not always concave and can have either an extremum or an inflection point, depending on the actual values of ρ_A , ρ_B and on the parameters of the marginals. In Figure 1(b), $E(\rho_A, \rho_B)$ reaches a maximum for non-zero ρ_A and ρ_B , thus for dependent rather than independent features. In Figure 1(c) we show an example where for $\sigma_{y,B}^2 \gg \sigma_{y,A}^2$ and $\rho_A = \rho_B = 0$ (uncorrelated features), the computed class overlap reaches not a maximum but an inflection point. Maximum of $E(\rho_A, \rho_B)$ is reached for $|\rho_A| \rightarrow 1$.

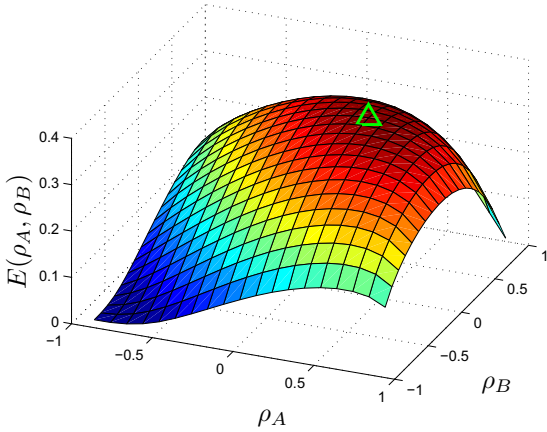
4 Conclusions

In this paper we have disambiguated the apparent conflict in reported impact of correlation between features on class separation, and consequently on the expected classification error, for bivariate normal distributions. We have shown that previously reported results hold only for specific configurations of marginal distributions and cannot be extended to arbitrary distributions. We have substantiated our claims analytically and using numerical simulations.

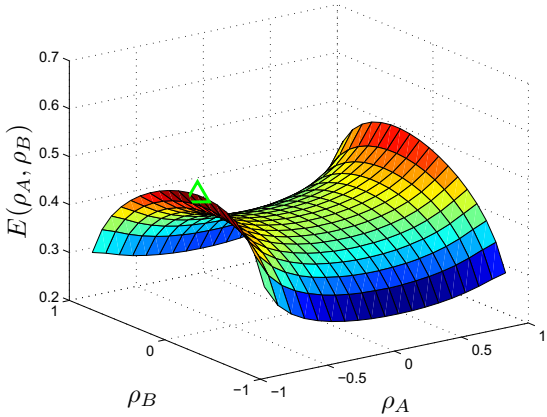
Our results carry important practical implications for the design of feature selection and feature extraction algorithms. First, the results reported here make it evident that feature correlations must be taken into account together with other distribution parameters and cannot be



(a) $\sigma_{x,A}^2 = \sigma_{y,A}^2 = 1, \sigma_{x,B}^2 = \sigma_{y,B}^2 = 1, \mu_x = 1, 0 \leq \mu_y \leq 3$



(b) $\sigma_{x,A}^2 = 1, \sigma_{x,B}^2 = 1, \mu_x = 1, \mu_y = 10$



(c) $\sigma_{x,A}^2 = \sigma_{y,A}^2 = 1, \sigma_{x,B}^2 = \sigma_{y,B}^2 = 5, \mu_x = 1, \mu_y = 0$

Figure 1. (a) impact of correlation coefficient ρ and distance between means μ_y on distribution overlap measure $E(\rho)$, (b) impact of correlation coefficients ρ_A, ρ_B on the location of stationary points of $E(\rho_A, \rho_B)$. All graphs drawn for two bivariate normal distributions.

used by themselves to gauge the usefulness of a given feature set. Second, when building feature extraction algorithms, correlations between features can be often pre-designed, in particular when feature correlations are effects of causal relationships between features. In such situations, designing correlated features can be beneficial for classification performance. An example of such application can be extraction of quality measures for biometric classification. Finally, reported findings are directly relevant to multi-classifier fusion, where scores of individual classifiers can be considered features to the fusion algorithm.

References

- [1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, New York, 2nd edition, 2001.
- [2] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, editors. *Feature Extraction, Foundations and Applications*. Studies in Fuzziness and Soft Computing. Springer Verlag, 2006.
- [3] M. A. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato, Hamilton, New Zealand, 1999.
- [4] J. F. Kenney and E. S. Keeping. *Mathematics of Statistics*. Van Nostrand, Princeton, NJ, 1951.
- [5] O. Koval, S. Voloshynovskiy, and T. Pun. Error exponent analysis of person identification based on fusion of dependent/independent modalities. In *Proc. of SPIE Photonics West, Electronic Imaging 2006, Multimedia Content Analysis, Management, and Retrieval 2006 (EI122)*, 2006.
- [6] K. Kryszczuk and A. Drygajlo. Improving classification with class-independent quality measures: Q-stack in face verification. In *Proc. of the 2nd International Conference on Biometric ICB'07*, Seoul, Korea, 2007.
- [7] J. Meynet and J.-P. Thiran. Information theoretic combination of classifiers with application to AdaBoost. In M. Haindl, J. Kittler, and F. Roli, editors, *Multiple Classifier Systems, 7th International Workshop, LNCS 4472*, pages 171–179, Prague, Czech Republic, 2007.
- [8] N. Poh and S. Bengio. How do correlation and variance of base-experts affect fusion in biometric authentication tasks? *IEEE Transactions on Signal Processing*, 53(11):4384–4396, November 2005.
- [9] F. Roli, G. Fumera, and J. Kittler. Fixed and trained combiners for fusion of imbalanced pattern classifiers. In *Proc. of the Intl. Conf. on Information Fusion*, pages 278–284, Annapolis, MD, USA, 2002.
- [10] O. Ushmaev and S. Novikov. Biometric fusion: Robust approach. In *Proc. of the 2nd Workshop on Multimodal User Authentication*, Toulouse, France, 2006.