Revisiting Inter-Genre Similarity

Sturm, Bob L.; Gouyon, Fabien

Published in:
I E E E Signal Processing Letters

DOI (link to publication from Publisher):
10.1109/LSP.2013.2280031

Publication date:
2013

Document Version
Early version, also known as pre-print

Revisiting Inter-Genre Similarity

Bob L. Sturm, *Member, IEEE,* Fabien Gouyon, *Member, IEEE*

Abstract

We revisit the idea of "inter-genre similarity" (IGS) for machine learning in general, and music genre recognition in particular. We show analytically that the probability of error for IGS is higher than naive Bayes classification with zero-one loss (NB). We show empirically that IGS does not perform well, even for data that satisfies all its assumptions.

Index Terms

EDICS: MLSAS-PATT Pattern recognition and classification; AEA-MIR Content-based Processing and Music Information Retrieval

I. Introduction

"Inter-genre similarity" (IGS) [1] proposes a very unique idea for music genre recognition (MGR). Consider we have recordings of classical and jazz music. We know that some instances of these share attributes, e.g., "piano," and some have unique attributes, e.g., "drum kit." Assume that from small pieces of these recordings (e.g., 25 ms), one can detect these attributes. Hence, we expect some recordings of music that uses either of these two genres to overlap; and the more overlap there is between classes, the more difficult it should become to discriminate between them. To counter this, IGS proposes to ignore pieces of recordings having shared attributes, and focus only on those unique to each class.

B. L. Sturm is with the Audio Analysis Lab, AD:MT, Aalborg University Copenhagen, A.C. Meyers Vænge 15, DK-2450 Copenahgen SV, Denmark, (+45) 99407633, fax: (+45) 44651800, e-mail: bst@create.aau.dk. He is supported in part by Independent Postdoc Grant 11-105218 from Det Frie Forskningsråd; and in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. F. Gouyon is with INESC TEC, Porto, Portugal. He is supported by the Media Arts and Technologies project (MAT), co-financed by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF). This publication only reflects the authors views.

More formally, IGS begins by modeling each class using training observations composed of many sub-observations. It then classifies each sub-observation, and relabels those that it misclassifies as being "inter-genre." Finally, IGS builds new models of all classes, and models the inter-genre class using the set of inter-genre sub-observations. Since music signals have long durations, features are typically taken in time-limited and localized ways [2]. Hence, a given piece of music will produce many sub-observations. Assuming independence between sub-observations (common in bags of frames approaches [3]), one can classify a piece of music by maximizing a weighted sum of the log posterior probabilities of the individual sub-observations given the models learned above; however, IGS ignores those sub-observations that are most likely in the inter-genre model. In this way, IGS classifies a piece of music by only considering the sub-observations "unique" to each class, i.e., those well-separated from all classes but one.

Based only on low-level short-time features, IGS has been reported to perform surprisingly well for MGR [1], making it the 7th best of 100 systems tested on the same benchmark dataset [4]. This result, however, contradicts the work in [3], [5], [6], which show clear indications that such low-level short-time features taken in isolation are irrelevant to MGR. In this paper, we seek to resolve this contradiction by analytically and empirically studying IGS to find why it is so successful.

In the next section — our main contribution — we analyze IGS and show it to be an approximation of, and thus inferior to, naive Bayes classification with zero-one loss. In the third section, we empirically test IGS, first using a toy problem that satisfies its assumptions, and then for MGR. We find that, even though its motivations sound convincing, IGS is unreasonable for MGR in particular, and machine learning in general. We make available code to reproduce all the figures and experimental results of this paper: http://imi.aau.dk/~bst.

II. Analysis

We now analyze IGS and find its decision boundary and probability of error for a two-class case that satisfies its assumptions. Its generalization to more classes is direct.

A. Formalization

Consider a training dataset $\mathcal{X} := \{(\mathbf{X}, \omega)_k\}$, where each $\mathbf{X} \in \mathbb{R}^N$ is from one *class* $\omega \in \Omega$, and k is an index of a duple from $\mathcal{X}$. We notate $\mathbf{X}_k$ and ω_k as coming from $(\mathbf{X}, \omega)_k$. We aim to find a function $f : \mathbf{X} \mapsto \Omega$ such that, for some loss function $L : \Omega \times \Omega \mapsto \mathbb{R}$, we minimize the expected loss over all classes. The optimal solution is given by Bayes [7]:

$$f^*(\mathbf{X}) := \arg\min_{\omega \in \Omega} \sum_{\omega_l \in \Omega} L(\omega_l, \omega) P(\omega_l | \mathbf{X}) \tag{1}$$

where $P(\omega|\mathbf{X})$ is the posterior of ω. Now consider $\mathbf{X}$ to be a super-vector (*observation*) of N vectors (*sub-observations*) $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}$. With a zero-one loss, and assuming all sub-observations are independent and identically distributed, $f^*(\mathbf{X})$ in (1) reduces to [7]:

$$f_{NB}(\mathbf{X}) := \arg\max_{\omega\in\Omega} \sum_{i=1}^{N} \log\left[p(\mathbf{x}^{(i)}|\omega)P(\omega)\right] \tag{2}$$

where $p(\mathbf{x}^{(i)}|\omega)$ is the conditional density of the ith sub-observation, and $P(\omega)$ is the prior of ω. This is known as naive Bayes classification with zero-one loss (NB). When we do not know the true conditional densities and priors, we must estimate them from $\mathcal{X}$ by building models of the classes.

The principle of IGS is to classify an observation based only on the sub-observations likely to be unique to a single class. First, IGS uses $\mathcal{X}$ to build models of sub-observations from all classes. It then builds a set of sub-observations based on whether each would be misclassified by the Bayes criterion:

$$\mathcal{X}_{IGS} := \left\{(\mathbf{x}^{(i)}, \omega)_n \in \mathcal{X} : i \in [1, N], p(\mathbf{x}_n^{(i)}|\omega_n)P(\omega_n) \leq \max_{\omega\in\Omega\setminus\omega_n} p(\mathbf{x}_n^{(i)}|\omega)P(\omega)\right\} \tag{3}$$

by a slight abuse of notation to say that $(\mathbf{x}^{(i)}, \omega)_n$ is the duple of the ith sub-observation and its label of $(\mathbf{X}, \omega)_n$. $\mathcal{X}_{IGS}$ is thus all sub-observations in $\mathcal{X}$ that have a higher posterior in classes different from their true ones, given the models constructed initially. IGS then models $\mathcal{X}_{IGS}$ as if it is a new class, the label of which we notate Λ. Finally, IGS builds new models of the remaining sub-observations in each class. This produces new conditional densities $\{\hat{p}(\mathbf{x}|\omega) : \omega \in \{\Omega, \Lambda\}\}$ and new priors $\{\hat{P}(\omega) : \omega \in \{\Omega, \Lambda\}\}$, where $\hat{p}(\mathbf{x}|\Lambda)$ is the conditional density of the a sub-observation in Λ.

The IGS classification rule is only a slight alteration of (2):

$$g(\mathbf{X}) := \arg\max_{\omega\in\Omega} \gamma \sum_{i=1}^{N} \beta_i \log\left[\hat{p}(\mathbf{x}^{(i)}|\omega)\hat{P}(\omega)\right] \tag{4}$$

where the weights $\{\beta_i : i = [1, N]\}$ and γ are defined by

$$\beta_i := \begin{cases} 1, & \hat{p}(\mathbf{x}^{(i)}|\omega)\hat{P}(\omega) > \hat{p}(\mathbf{x}^{(i)}|\Lambda)\hat{P}(\Lambda) \\ 0, & \text{else} \end{cases} \tag{5}$$

$$\gamma := \begin{cases} 0, & \forall i\, \beta_i = 0 \\ 1/\sum_{j=1}^{N} \beta_j, & \text{else.} \end{cases} \tag{6}$$

There are a few significant differences between (4) and the decision function presented in [1]. First, while $P(\Lambda)$ is not explicitly defined, our personal communication reveals they assume $\hat{P}(\Lambda) = \hat{P}(\omega) = 1/|\Omega|$. Second, [1] does not define the condition for when all $\beta_i = 0$. We assume that IGS then picks randomly from Ω. (This never occurs in our experiments in Section III-B.)

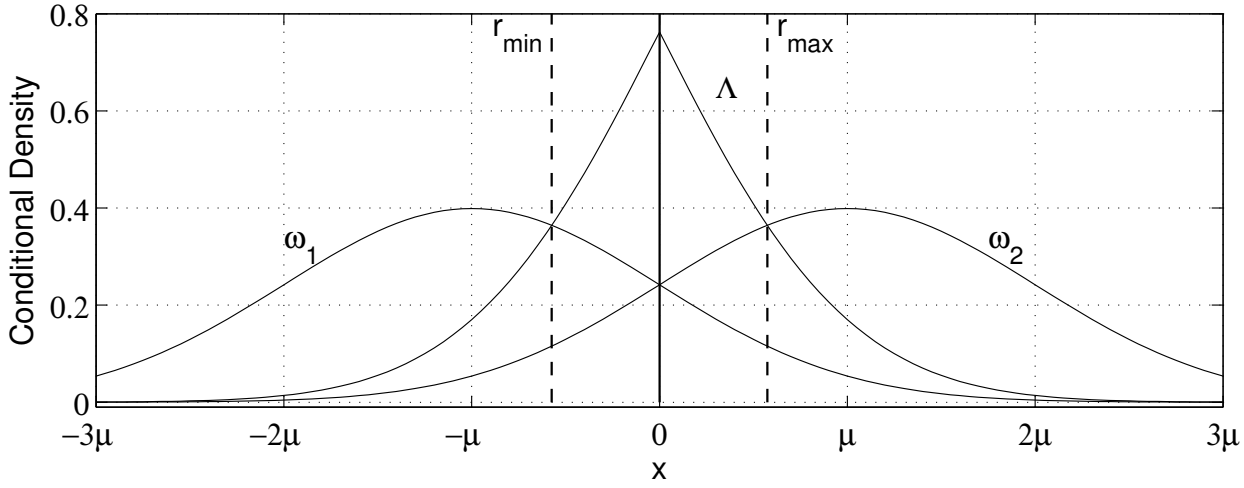

Fig. 1. Conditional densities of classes and shared attributes. Bayes decision boundary is solid vertical line. IGS guesses when between dashed lines.

B. Decision boundary

We now analyze the decision boundaries of IGS and compare it to that for NB. Consider a two-class scenario where sub-observations of each are independently and identically distributed Gaussian. The two classes have variance $\sigma_1 = \sigma_2 = 1$, and means, $-\mu_1 = \mu_2 = \mu \geq 0$. We choose from $\{\omega_1, \omega_2\}$ using N sub-observations drawn from one of the distributions. Since we know the true model of a sub-observation, we know for each class the true model of the N-dimensional observation: a multivariate Gaussian with covariance $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_N$ ($N \times N$ identity matrix), and mean $-\mathbf{m}_1 = \mathbf{m}_2 = \mu\mathbf{1}$ (length-N vector of ones). To simplify the analysis, we rotate the observation space such that the means have a non-zero component only in the first dimension. Thus, the conditional distributions of the N-dimensional rotated observations are

$$p(\mathbf{v}|\omega_1) = \frac{1}{(2\pi)^{N/2}} e^{-\|\mathbf{v}+\sqrt{N}\mu\mathbf{e}_1\|^2/2} \tag{7}$$

$$p(\mathbf{v}|\omega_2) = \frac{1}{(2\pi)^{N/2}} e^{-\|\mathbf{v}-\sqrt{N}\mu\mathbf{e}_1\|^2/2} \tag{8}$$

where $\mathbf{e}_1$ has 1 in its first row, and zeros everywhere else. The decision criterion for NB with zero-one loss is given by

$$\mathbf{e}_1^T\mathbf{v} \underset{\omega_1}{\overset{\omega_2}{\gtrless}} B = \frac{\log\left[P(\omega_1)/P(\omega_2)\right]}{2\mu\sqrt{N}}. \tag{9}$$

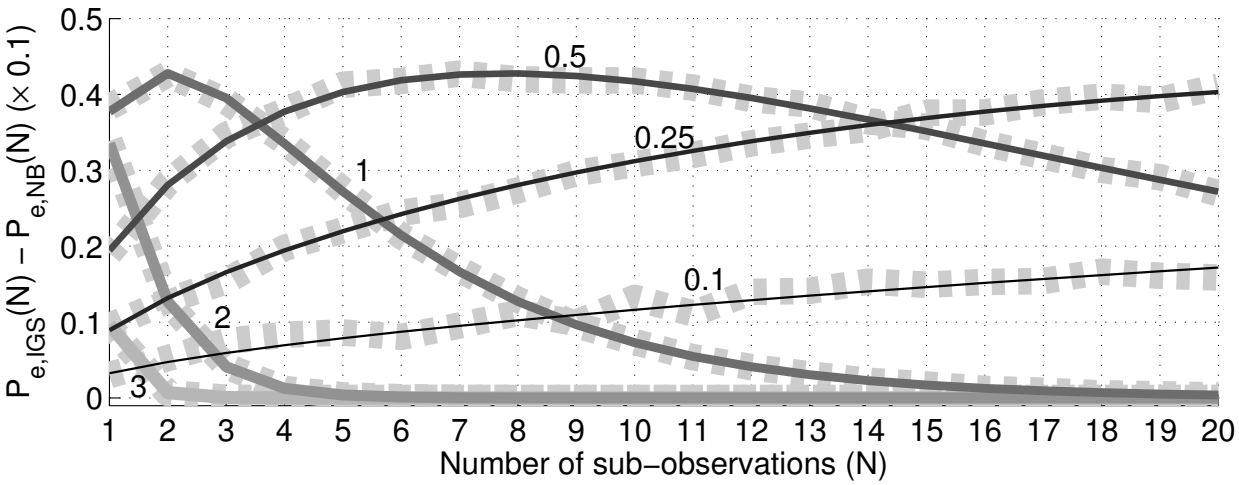

Fig. 2. For the scenario in Section II-B, difference in probability of error for IGS (4) and NB (2) for five μ (labeled), as a function of the number of sub-observations N. Thick dashed lines are of empirical estimations.

Using this boundary, we find the distribution of the IGS class:

$$p(\mathbf{v}|\Lambda) = \begin{cases} \alpha p(\mathbf{v}|\omega_1)P(\omega_1), \mathbf{e}_1^T\mathbf{v} > B \\ \alpha p(\mathbf{v}|\omega_2)P(\omega_2), \mathbf{e}_1^T\mathbf{v} \leq B \end{cases} \tag{10}$$

where, with $\Phi(x)$ the cumulative distribution function of a Normal random variable, the normalization factor is

$$\alpha^{-1}(N) = P(\omega_1)\left(1 - \Phi(B + \mu\sqrt{N})\right) + P(\omega_2)\Phi(B - \mu\sqrt{N}). \tag{11}$$

Figure 1 shows our scenario for the sub-observations, assuming $P(\omega_1) = P(\omega_2) = P(\Lambda)$ as in [1]. In this case, $B = 0$.

By the weights (5), IGS ignores a sub-observation if its posterior in Λ (10) is greater than in an element of Ω, i.e., (7) or (8). Solving for where $p(\mathbf{v}|\omega_1)P(\omega_1) = p(\mathbf{v}|\Lambda)P(\Lambda)$ and $p(\mathbf{v}|\omega_2)P(\omega_2) = p(\mathbf{v}|\Lambda)P(\Lambda)$, we find IGS picks a class randomly if $\mathbf{e}_1^T\mathbf{v} \in \mathcal{R}(N)$, where

$$\mathcal{R}(N) := [r_{\min}(N), r_{\max}(N)] = B + \frac{\log[\alpha(N)P(\omega_1)P(\Lambda)/P(\omega_2)]}{2\mu\sqrt{N}}\,[-1, 1]\,. \tag{12}$$

Figure 1 shows this region for our scenario.

Assume $P(\omega_1) = P(\omega_2)$. The derivation of the following are given in the appendix. The width of $\mathcal{R}(N)$ as $\mu \to 0^+$

$$\lim_{\mu\to 0^+} r_{\max}(N) - r_{\min}(N) = \frac{2}{\sqrt{2\pi}}. \tag{13}$$

For $\mu \neq 0$, the width of $\mathcal{R}(N)$ diverges in high dimensions

$$\lim_{N\to\infty} r_{\max}(N) - r_{\min}(N) = \infty. \tag{14}$$

Furthermore, for $\mu \neq 0$, the width of $\mathcal{R}(N)$ is strictly increasing with N, i.e., $\mathcal{R}(N+1) > \mathcal{R}(N)$. Hence, the width of $\mathcal{R}(N)$ is never smaller than $2/\sqrt{2\pi}$.

It is important to note that, while the divergence of the width of $\mathcal{R}(N)$ is clearly to be expected when the separation between the modes of the sub-observation distributions of each class $2\mu > 0$, we do not have the liberty of setting $\mu = 1/\sqrt{N}$ to hold it constant. In other words, we cannot control the distributions of the sub-observations. The only parameter we can control is N, the number of sub-observations we make. Thus, while the NB decision boundary remains infinitesimally thin as $N \to \infty$, that of IGS becomes infinitely thick.

C. Probability of error

We now find the probability of error for IGS in the scenario above, and compare it to that for NB. As before, we work in the rotated observation space, and assume $P(\omega_1) = P(\omega_2) = P(\Lambda)$, and thus $B = 0$. For NB, the probability of error is

$$P_{\text{e,NB}}(N) = P\left[\mathbf{e}_1^T\mathbf{v} \geq 0 \middle| \omega_1\right]\frac{1}{2} + P\left[\mathbf{e}_1^T\mathbf{v} \leq 0 \middle| \omega_2\right]\frac{1}{2} = \Phi(-\mu\sqrt{N}). \tag{15}$$

For IGS, the probability of error is

$$\begin{aligned} P_{\text{e,IGS}}(N) = {} & P\left[\mathbf{e}_1^T\mathbf{v} \geq r_{\max}(N) \middle| \omega_1\right]\frac{1}{2} + P\left[\mathbf{e}_1^T\mathbf{v} \leq r_{\min}(N) \middle| \omega_2\right]\frac{1}{2} \\ & + P\left[\mathbf{e}_1^T\mathbf{v} \in \mathcal{R}(N) | \omega_1\right]\frac{1}{4} + P\left[\mathbf{e}_1^T\mathbf{v} \in \mathcal{R}(N) | \omega_2\right]\frac{1}{4} \end{aligned} \tag{16}$$

where the last terms come from the choice to have IGS pick a class randomly if $\mathbf{e}_1^T\mathbf{v} \in \mathcal{R}(N)$, and that $P[\mathbf{e}_1^T\mathbf{v} \geq r_{\max}(N)|\omega_1] = P[\mathbf{e}_1^T\mathbf{v} \geq r_{\min}(N)|\omega_2]$, and $P[\mathbf{e}_1^T\mathbf{v} \in \mathcal{R}(N)|\omega_1] = P[\mathbf{e}_1^T\mathbf{v} \in \mathcal{R}(N)|\omega_2]$. The above simplifies to

$$P_{\text{e,IGS}}(N) = \frac{1}{2}\Phi\left(r_{\min}(N) - \mu\sqrt{N}\right) + \frac{1}{2}\Phi\left(r_{\max}(N) - \mu\sqrt{N}\right). \tag{17}$$

Figure 2 shows the difference between these errors for several μ as we consider more sub-observations. We see that while NB always outperforms IGS, the difference between the two can become smaller as N increases. As the modes of the two class distributions become very close, IGS requires many more observations than NB to perform just as well. That the two perform similarly in high dimensions (dependent upon μ) makes sense when we compute the probability of an N-dimensional observation falling in $\mathcal{R}(N)$ as N goes to infinity $\lim_{N\to\infty} P[\forall i = [1, N](x^{(i)} \in \mathcal{R}(N))] = \lim_{N\to\infty} p^N = 0$ since $p = P[x^{(1)} \in \mathcal{R}(N)] < 1$ for any μ and N.

D. Summary

When we use N sub-observations, the decision boundary for NB (2) is a single $(N-1)$-dimensional hyperplane. For IGS, it is a pair of these hyperplanes, the separation between which increases unbounded as we use more sub-observations. For the two-class multivariate Gaussian scenario, we show this separation is never less than $2/\sqrt{2\pi}$, and increases monotonically as N increases. We see for this scenario the probability of error of IGS is always inferior to that of NB. This is satisfying since IGS (4) is only NB (2), but using fewer sub-observations. While both make the assumption of independence between sub-observation, NB (2) is optimal by definition (it minimizes zero-one loss), and so anything different must have a higher expected loss, and thus probability of error.

For more classes than two, our analysis helps to visualize what happens. In the observation space, there exists between each pair of classes a decision boundary: infinitesimally thin for NB, and with a non-zero thickness for IGS. These combine to form a partition of the space; but in the case of IGS, the partition boundaries have a volume while for NB it has zero volume. As N grows, this volume grows unbounded for IGS, but remains zero for NB. Finally, since NB minimizes the probability of error in the multiclass case [7], a change to its criteria necessarily produces a higher error. We show this to be the case in the following simulations.

III. Simulations

We now test and compare NB and IGS, first for an ideal dataset satisfying the assumptions of IGS, and then for the same MGR dataset used in [1].

A. An ideal dataset

Figure 3 shows realizations of a dataset of two-dimensional sub-observations from four classes. The true model of each class is a mixture of two Gaussians: one with zero mean and identity covariance, and the other with non-zero mean and correlated dimensions. Let the probability (mixture weight) of a sub-observation coming from the central Gaussian be $p_0 > 0$. Clearly, all sub-observations of a class are independently and identically distributed, thus satisfying a key assumption of NB and IGS. In left and center of Fig. 3, $p_0 = 0.5$ for all classes, while for the right plot $p_0 = 0.98$. We sample 1000 observations from each class, where each observation has $N = 2$ sub-observations. As in [1], we assume equal priors for all classes and the shared attributes class. Using stratified 2-fold cross-validation, we estimate (by expectation maximization) a Gaussian mixture model (GMM) of order two and full covariances, for each class. Using these models, IGS constructs the same kind of GMM for the shared attributes in (3), and

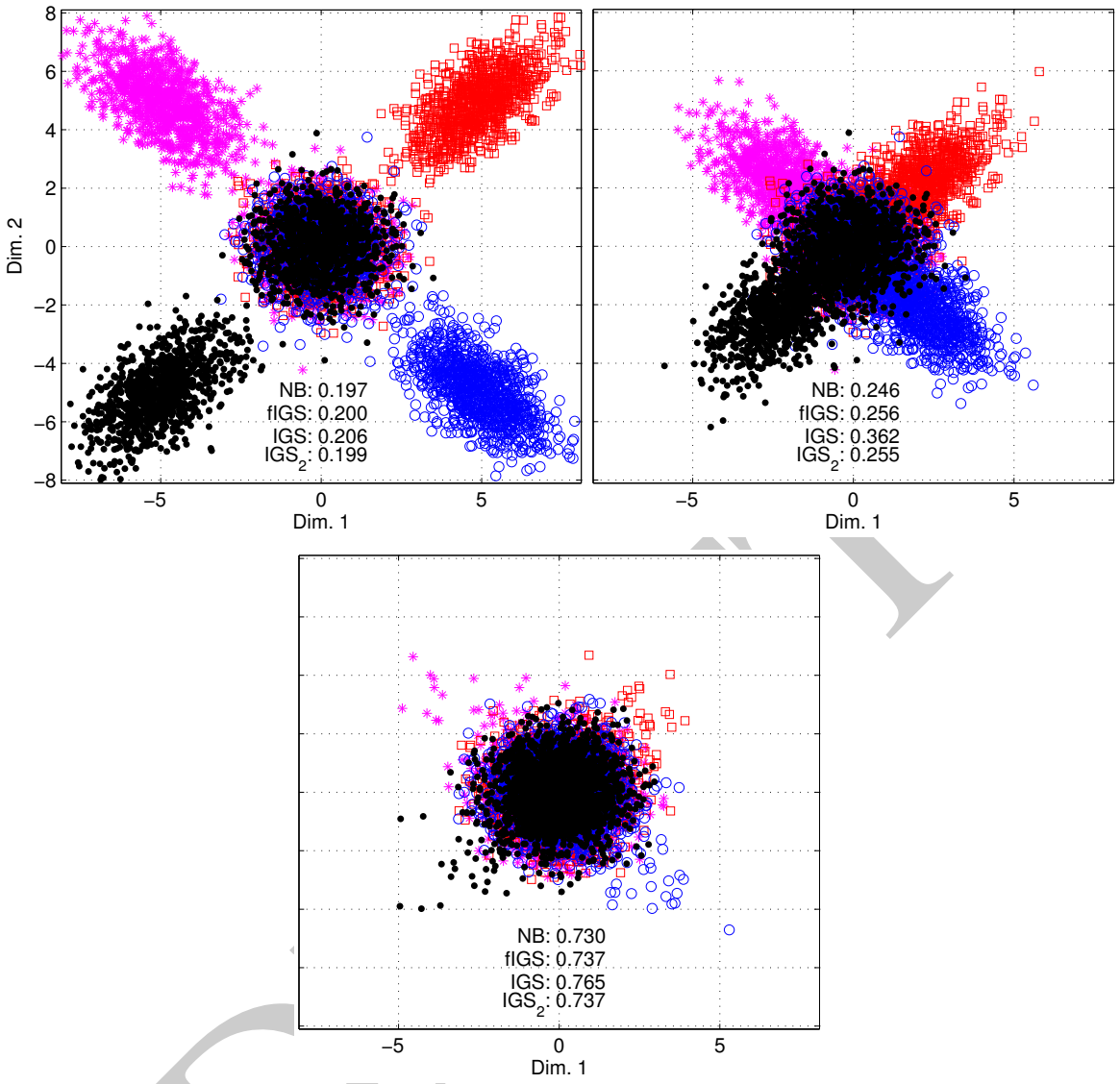

Fig. 3. Three datasets ideal for IGS, and their empirical classification error rates. Gaussians in mixtures: equal weights (left and center), and unequal weights (right, probability of a sub-observation coming from the shared region is 0.98). Observations are made of 2 sub-observations sampled from a distribution, classified by: "NB" (2); "fIGS" (2) using the model parameters estimated without the sub-observations in (3); "IGS" (4); "IGS2" alternation of (4) described in Section III-A.

estimates a new GMM for each class. In IGS_2, we revise the condition in (5): $\exists\omega \in \Omega[\hat{p}(\mathbf{x}^{(i)}|\omega)\hat{P}(\omega) > \hat{p}(\mathbf{x}^{(i)}|\Lambda)\hat{P}(\Lambda)]$ to see its effect.

We expect for well-separable classes (μ large) an error rate around $3p_0/4$ when classifying individual sub-observations by NB (2). For observations (made of two sub-observations), the error rate will improve since with probability $1 - p_0^2$ at least one of the sub-observations comes from a region that is separable

Algo \ GMM	8	16	24	32	48
IGS [1]	66.36	70.18	72.58	74.96	76.73
IGS	15.40	15.50	16.10	15.60	16.20
IGS_2	49.50	50.80	51.10	53.90	51.70
fIGS [1]	53.89	55.51	56.12	56.38	57.48
fIGS	49.40	50.30	50.70	52.90	51.30
NB	51.10	52.60	54.00	54.40	53.20

TABLE I

MEAN ACCURACIES OF OUR EXPERIMENTS AND THOSE IN [1] (GRAY). THE LAST ROW SHOWS PERCENT OF ALL SUB-OBSERVATIONS IN (3).

from those of the other classes. We see this for each case as in Fig. 3. Using NB (2), but with the models estimated from all sub-observations except those in (3) — "flat" IGS (fIGS) in [1] — we find error rates higher than NB using the initial models. Using IGS (4) results in even higher error rates.

B. Music genre dataset

We now test the accuracy of IGS in reproducing genre labels [8]. We reproduce the experiments in [1] as closely as possible: we use the *GTZAN* benchmark dataset [4], [9]; stratified 2-fold cross-validation; sub-observations consisting of 13 Mel-frequency cepstral coefficients (MFCCs) [10], computed from 25 ms windows with a hop of 10 ms; GMM models of varying orders (8 to 48), with diagonal covariance matrices; equal priors; and observations composed of 300 consecutive sub-observations (which is 3 s duration). We use the same folds for all algorithms. We set the maximum number of iterations for expectation maximization (`gmdistribution.fit` in MATLAB) to 500, and use a small regularization constant (10^{-8}) to avoid condition problems in covariance estimation.

Table I shows that IGS performs significantly worse than NB, and than reported in [1]. (Private correspondence with the authors of [1] reveals that they cannot reproduce their results.) Though it is not clear whether in [1] training uses all sub-observations from each 30 s excerpt in the training set, or only ones from a randomly selected 3 s portion, we find no significant difference between these using GMMs of order 8. Table I also shows performance does not necessarily increase as the order of the model grows — which agrees with the fact that the estimation of the parameters of a distribution depends on the number of observations and their dimensionality. As the observation dimension increases, the performance of a system does not necessarily improve unless the amount of data grows.

IV. Conclusions

Our analysis of IGS shows that it approximates NB with zero-one loss. It must then have a higher probability of error. Our experimental work supports this finding, and shows IGS for MGR behaves quite poorly — which is expected if low-level short-time features taken in isolation are not representative of music genre [3], [5], [6]. Furthermore, our results show a trend expected from parameter estimation: estimating more parameters without more data does not necessarily improve the performance of the system. The bottom line here is that all observations are informative; the extra overhead in IGS does not provide any benefit to naive Bayes classification.

Acknowledgments

We wish to thank U. Bağci and E. Erzin for their open and helpful discussion of various aspects of this work; and the associate editor and anonymous reviewers for their many helpful suggestions.

References

[1] U. Bağci and E. Erzin, "Automatic classification of musical genres using inter-genre similarity," *IEEE Signal Proc. Letters*, vol. 14, no. 8, pp. 521–524, Aug. 2007.

[2] G. Peeters, B. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *J. Acoust. Soc. Amer.*, vol. 130, no. 5, pp. 2902–2916, 2011.

[3] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?" *J. Neg. Results Speech Audio Sci.*, vol. 1, no. 1, 2004.

[4] B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," *http://arxiv.org/abs/1306.1461*, 2013.

[5] J. Jensen, M. Christensen, D. Ellis, and S. Jensen, "Quantitative analysis of a common audio similarity measure," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 693–703, May 2009.

[6] G. Marques, T. Langlois, F. Gouyon, M. Lopes, and M. Sordo, "Short-term feature space and music genre classification," *J. New Music Research*, vol. 40, no. 2, pp. 127–137, 2011.

[7] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Amsterdam, The Netherlands: Academic Press, Elsevier, 2009.

[8] B. L. Sturm, "Classification accuracy is not enough: On the evaluation of music genre recognition systems," *J. Intell. Info. Systems (in press)*, 2013.

[9] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, July 2002.

[10] M. Slaney, "Auditory toolbox," Interval Research Corporation, Tech. Rep., 1998.

APPENDIX

We now state and prove three theorems describing the behavior of $\mathcal{R}(N)$ (12) as a function of the number of sub-observations N for the scenario in Section II-B. As Bagci and Erzin [1] assume $P(\omega_1) = P(\omega_2) = P(\Lambda)$, and thus $B = 0$.

Theorem 1: The limiting width of $\mathcal{R}(N)$ is (13).

Proof: Starting from (12)

$$\lim_{\mu\to 0^+} r_{\max}(N) - r_{\min}(N) = \lim_{\mu\to 0^+} -\frac{\log \alpha^{-1}(N)/P(\Lambda)}{\mu\sqrt{N}} \tag{18}$$

where we have used the fact that $\log f = -\log f^{-1}$ to make the succeeding easier. Using L'Hôpital's rule, we find the partial derivatives of the numerator and denominator with respect to μ for N fixed. That of the numerator is

$$-\frac{\partial}{\partial\mu}\log \alpha^{-1}(N)/P(\Lambda) = -\frac{1}{\alpha^{-1}(N)}\frac{\partial}{\partial\mu}\alpha^{-1}(N) = -\frac{2P(\Lambda)}{\alpha^{-1}(N)}\frac{\partial}{\partial\mu}\Phi(-\mu\sqrt{N}). \tag{19}$$

By the Leibniz integral rule, we know

$$\frac{\partial}{\partial\mu}\Phi(-\mu\sqrt{N}) = \sqrt{\frac{N}{2\pi}}e^{-(\mu\sqrt{N})^2/2} \tag{20}$$

and thus by substitution

$$-\frac{\partial}{\partial\mu}\log \alpha^{-1}(N)/P(\Lambda) = \sqrt{\frac{N}{2\pi}}\frac{e^{-(\mu\sqrt{N})^2/2}}{\Phi(-\mu\sqrt{N})}. \tag{21}$$

Since the partial derivative of the denominator with respect to μ is simply $\sqrt{N}$, we arrive at (13) by setting $\mu = 0$. ■

Theorem 2: For $\mu \neq 0$, the width of $\mathcal{R}(N)$ diverges as N increases.

Proof: Consider the limit of $r_{\max}(N)$ as $N \to \infty$:

$$\begin{aligned}\lim_{N\to\infty} r_{\max}(N) &= \lim_{N\to\infty} -\frac{\log \alpha^{-1}(N)/P(\Lambda)}{2\mu\sqrt{N}} = \lim_{N\to\infty} -\frac{\log \alpha^{-1}(N)}{2\mu\sqrt{N}} \\ &= \lim_{N\to\infty} -\frac{1}{2\mu\sqrt{N}}\log\Big[P(\Lambda)\left(1-\Phi(\mu\sqrt{N})\right) + P(\Lambda)\Phi(-\mu\sqrt{N})\Big] \\ &= \lim_{N\to\infty} -\frac{\log\Phi(-\mu\sqrt{N})}{2\mu\sqrt{N}}\end{aligned} \tag{22}$$

since $1 - \Phi(\mu\sqrt{N}) = \Phi(-\mu\sqrt{N})$. Since the numerator involves the logarithm of a number less than 1, and which shrinks as N grows, it approaches $-\infty$ much faster than that the denominator approaches ∞. Hence, $\lim_{N\to\infty} r_{\max}(N) = \infty$. In the same way, we find $\lim_{N\to\infty} r_{\min}(N) = -\infty$. ■

Theorem 3: For $\mu \neq 0$, the width of $\mathcal{R}(N)$ is strictly increasing with N.

Proof: Consider taking the ratio of the two widths

$$\frac{r_{\max}(N+1)-r_{\min}(N+1)}{r_{\max}(N)-r_{\min}(N)}=\frac{\sqrt{N}}{\sqrt{N+1}}\frac{\log\alpha^{-1}(N+1)/P(\Lambda)}{\log\alpha^{-1}(N)/P(\Lambda)}=\frac{\sqrt{N}}{\sqrt{N+1}}\frac{\log 2\Phi(-\mu\sqrt{N+1})}{\log 2\Phi(-\mu\sqrt{N})} \tag{23}$$

The first term is monotonic increasing, and the smallest it can be is $1/\sqrt{2}$ for $N=1$. Without loss of generality, consider $\mu>0$ (otherwise, switch class labels), and thus $\Phi(-\mu\sqrt{N+1})<\Phi(-\mu\sqrt{N})$. The second term is then also monotonic increasing. Here we must determine if

$$\frac{\log 2\Phi(-\mu\sqrt{2})}{\log 2\Phi(-\mu)}\geq\sqrt{2}. \tag{24}$$

The left hand side monotonically decreases as μ shrinks. So, by L'Hôpital's rule and the Leibniz integral rule, we find

$$\lim_{\mu\to 0^+}\frac{\log 2\Phi(-\mu\sqrt{2})}{\log 2\Phi(-\mu)}=\lim_{\mu\to 0^+}\sqrt{2}\frac{\Phi(-\mu)}{\Phi(-\mu\sqrt{2})}\frac{e^{-\mu^2}}{e^{-\mu^2/2}}=\sqrt{2}. \tag{25}$$

The completes the proof. ■