MIT Libraries | DSpace@MIT

MIT Open Access Articles

Estimation of signal information content for classification

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation: Fisher, J.W., M. Siracusa, and Kinh Tieu. "Estimation of Signal Information Content for Classification." Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009. DSP/SPE 2009. IEEE 13th. 2009. 353-358.© 2009 Institute of Electrical and Electronics Engineers.

As Published: http://dx.doi.org/10.1109/DSP.2009.4785948

Publisher: Institute of Electrical and Electronics Engineers

Persistent URL: http://hdl.handle.net/1721.1/59304

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

Massachusetts Institute of Technology

ESTIMATION OF SIGNAL INFORMATION CONTENT FOR CLASSIFICATION

John W. Fisher III[1] *Michael Siracusa*[1] *Kinh Tieu*[2]

[1]Massachusetts Institute of Technology*
Cambridge, MA

[2]Mitsubishi Electric Research Laboratories
Cambridge, MA

ABSTRACT

Information measures have long been studied in the context of hypothesis testing leading to variety of bounds on performance based on the information content of a signal or the divergence between distributions. Here we consider the problem of estimation of information content for high-dimensional signals for purposes of classification. Direct estimation of information for high-dimensional signals is generally not tractable therefore we consider an extension to a method first suggested in [1] in which high dimensional signals are mapped to lower dimensional feature spaces yielding lower bounds on information content. We develop an affine-invariant gradient method and examine the utility of the resulting estimates for predicting classification performance empirically.

***Index Terms*—** information measures, mutual information, feature extraction, invariance

1. INTRODUCTION

There is a long history of analysis relating information measures to classification performance in M-ary hypothesis testing. These measures, particularly those of the Ali-Silvey [2] variety (expectations over convex functions of the likelihood ratio), are appealing in the context of machine learning as they lend themselves readily to optimization owing to their smoothness and convexity properties. Basseville [3] surveys their use in signal processing. More recently, Nguyen *et al* [4] have established links between a broad class of Ali-Silvey measures and convex surrogate loss functions (e.g. the exponential and logistic loss functions). In turn, bounds on excess risk have been established [5] for such surrogate loss functions, further establishing the suitability of Ali-Silvey measures as a surrogate criterion for classifier design.

Kullback-Leibler divergence [6], of which mutual information is a special case, is an example of an Ali-Silvey measure. We discuss the use of mutual information (MI) as a criterion for feature extraction, focusing on mappings from high-dimensional measurements to low-dimensional features following a method originally suggested in [1]. There, an estimator utilizing kernel density estimation (KDE) was used to optimize mutual information in a low-dimensional feature space. We discuss two issues. The first is the incorporation of affine invariance into the optimization. The second is the impact of kernel bandwidth on both the quality of the empirical estimate of mutual information as well as the optimization of the mapping parameters. We observe that kernel sizes which are suitable for optimization (i.e. learning the mappings) are different than those that lead to good empirical estimates of the mutual information. The latter is of note owing to the fact that optimal projections cannot, in general, be guaranteed in the nonparametric setting. Consequently, accurate empirical estimates are desired when estimating performance bounds.

*The research of JF and MS was partially supported by the Air Force Office of Scientific Research via Grant FA9550-06-1-0324 and the Army Research Office via grant W911NF-06-1-0076.

2. FANO AND HELLMAN-RAVIV BOUNDS

Fano's inequality [7] is a well known bound relating conditional entropy to the probability of misclassification for M-ary hypothesis testing. Similar to the Cramer-Rao bound for minimum mean-square estimation, Fano's inequality provides a lower bound on what is achievable for classification error and thus motivates mutual information as a design criterion for classifiers. There are various statements of Fano's inequality, all of which exploit the following relationships (see [8] for a derivation)

$$H(C|X) = H(C) - I(C;X) \quad (1)$$
$$= H(E|X) + P_\epsilon H(C|X, E=1) \quad (2)$$

where P_ϵ denotes the probability of misclassification, C is a discrete random variable denoting the class, X is an observed random variable from which C is estimated, the binary random variable E denotes whether a classification error has occurred, $H(\bullet)$ denotes discrete Shannon entropy (we use $h(\bullet)$ for differential entropy), $H(\bullet|\bullet)$ denotes conditional entropy, and $I(\bullet;\bullet)$ denotes the mutual information between two random variables. This relationship separates the uncertainty regarding class label conditioned on an observation X into two terms. The first term is the uncertainty as to whether or not an error has occurred (conditioned on the observation X). The second term is the remaining uncertainty in the class label conditioned on the event that an error has occurred.

978-1-4244-3677-4/09/$25.00 ©2009 IEEE

In the following discussion, minimization of $H(C|X)$ is equivalent to maximization of $I(C;X)$ by equation 1. We use $H(C|X)$ and $I(C;X)$ interchangeably, though, some care must be taken when expressing error bounds in terms of these two measures. In particular, one must make an assumption regarding $H(C)$ (the prior uncertainty in the class variable) in order to give an explicit bound on P_ϵ as a function of $I(C;X)$. The weak form of Fano's inequality

$$H(C|X) \leq 1 + P_\epsilon \log(M-1), \tag{3}$$

with M being the number of classes, substitutes the inequalities

$$H(C|X, E=1) \leq \log(M-1), \text{ and}$$
$$H(E|X) \leq H(P_\epsilon) \leq 1,$$

into equation 2 assuming log base-2. Alternately, the strong Fano bound

$$P_\epsilon \geq \min_P \{P : H(C|X) \leq H(P) + P \log(M-1)\}, \tag{4}$$

is tighter and may be solved numerically. While Equation 3 provides a basis for the use of MI as an optimization criterion when designing M-ary classifiers it is inapplicable to the binary case[1] as when $M = 2$

$$H(C|X, E=1) = 0.$$

The strong Fano bound does apply to the binary case.

Hellman and Raviv [9] give a loose upper bound on the probability of error also specified in terms of conditional entropy

$$P_\epsilon \leq \frac{1}{2} H(C|X) \tag{5}$$

which can also be stated in terms of $I(C;X)$. We refer to this bound as the H-R upper bound. In contrast, to the Fano bounds, the H-R upper bound necessarily assumes the use of the Bayes' optimal classifier. In either case, both the upper and lower bounds motivate $I(C;X)$ as a design criterion.

Figure 1 plots the strong and weak Fano lower bounds as well as the Hellman-Raviv upper bounds as a function of $I(C;X)$ for $M = 4$. What is clear from the figure is that the weak Fano bound is significantly looser than the strong Fano bound (almost everywhere) and particularly for small P_ϵ which is an important regime for classifier design. This disparity grows as M grows large. Consequently, the weak Fano bound is of limited use to compute performance bounds as a function of estimates of information measures.

Fano's inequality has been the motivation for a wide variety of methods exploiting mutual information $I(C;X)$ as a criterion for optimizing decision systems. Previous work includes the design of decision trees [10], feature selection [11, 12] and feature extraction [13, 1]. The distinction between selection and extraction being that the former addresses the selection or ranking of a subset of features from a given list while the latter infers a mapping as a function of all features to a lower-dimensional representation. Many similar feature selection and extraction methods have since been developed differing in the form of probability model or the approximation to conditional entropy. For the remainder of the discussion we will focus exclusively on feature extraction.

[1] a number of references in the machine learning literature erroneously cite Equation 3 as a basis for classifier design in the binary case.

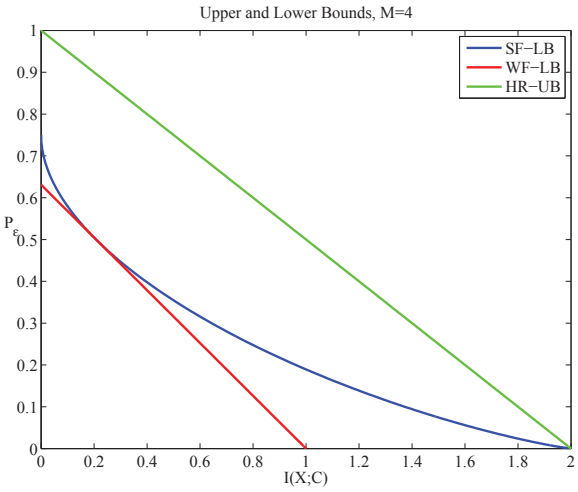

Fig. 1. Weak Fano lower bound (red), strong Fano lower bound (blue), H-R upper bound (green) on P_ϵ for $M = 4$.

Fig. 2. Directed graph from class label C to estimate $\hat{C}$ with embedded feature extraction.

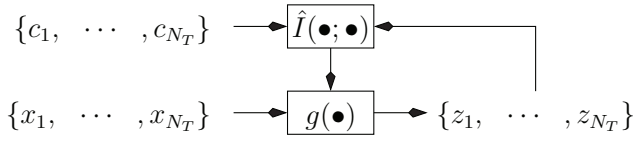

Fig. 3. Block diagram for learning information preserving mappings: Labeled samples x_i are passed through a parameterized mapping $g(\bullet)$ resulting in lower-dimensional features z_i which are then combined with class labels c_i to compute a perturbation to $g(\bullet)$ so as to increase mutual information.

3. INFORMATION PRESERVING MAPPINGS

The Markov chain of Figure 2 depicts the statistical model of a classification system with feature extraction embedded in the process. Here $X \in \Re^D$ represents a high-dimensional measurement while $Z \in \Re^L$ is a mapping of the measurement to a lower-dimensional space. The last term, $\hat{C}(Z)$ is the classifier. Ideally, one would like to design the parameters of the mapping $g : \Re^D \rightarrow \Re^L$ so as to minimize $P_\epsilon = \Pr\left\{C \neq \hat{C}(Z)\right\}$, however, as suggested in [1, 14] and motivated by the preceding, a surrogate criterion is to maximize $I(C;Z)$, thus optimizing bounds on P_ϵ. Note that under this surrogate criterion one can optimize properties of the features *without* explicitly constructing a classifier.

Ideally, one would like Z to be a sufficient statistic. One measure of sufficiency is the K-L divergence

$$D\left(p\left(C|X\right)||p\left(C|Z\right)\right) = I(C;X) - I(C;Z)$$

which captures information loss as a consequence of feature extraction. This is a form of the well known data processing inequality[8]. It can be shown that $I(C;X) \geq I(C;Z)$ with equality if and only if Z is sufficient for C.

3.1. Empirical Estimation and Optimization of Mutual Information

We define the following sets of high-dimensional samples and their labels as

$$\mathcal{X} = \{x_1, \cdots, x_N\} \quad \mathcal{C} = \{c_1, \cdots, c_N\}$$

where $x_i \sim p(X|C = c_i)$ and $c_i \sim P(C)$. The set of features is denoted

$$\begin{aligned}\mathcal{Z} &= \{z_1, \cdots, z_N\} \\ &= \{g(x_1;G), \cdots, g(x_N;G)\} = g(\mathcal{X};G)\end{aligned}$$

where G are the parameters of a function that we wish to optimize. Finally, we denote

$$\mathcal{X}_j = \{x_i | c_i = j\} \quad \mathcal{Z}_j = \{z_i | c_i = j\}$$

as the *subset* of samples (or features) grouped by common label. The resubstitution estimate of $I(C;Z)$ is

$$\begin{aligned}&\hat{I}\left(\mathcal{C}, \mathcal{Z}, \Theta\left(\mathcal{Z}\right)\right) = \\ &\qquad \frac{1}{N}\sum_{z_i \in \mathcal{Z}} \log\left(\frac{\hat{p}\left(z_i | c_i; \Theta_{c_i}\left(\mathcal{Z}_{c_i}\right)\right)}{\hat{p}\left(z_i; \Theta\left(\mathcal{Z}\right)\right)}\right)\end{aligned} \tag{6}$$

where the parameters of class-conditional density estimates

$$\Theta = \{\Theta_1, \cdots, \Theta_M\},$$

are functions of $\{\mathcal{Z}_1, \cdots, \mathcal{Z}_M\}$, respectively. Equation 6 can also be expressed as an explicit function of $\mathcal{X}$ and G by substituting $z_i = g(x_i;G)$ and $\mathcal{Z} = g(\mathcal{X};G)$ as appropriate. We will drop the explicit dependence when it is clear from the context. Additionally, we assume that the estimate of $\hat{p}\left(z\right)$ is marginally-consistent

$$\hat{p}\left(z;\Theta\right) = \sum_{j=1}^{M} \pi_j \hat{p}\left(z|j;\Theta_j\right) \quad ; \; \pi_j = \frac{|\mathcal{Z}_j|}{N}$$

where we assume that the prior probabilities of each class π_j are reflected in the samples[2]. Under this assumption, Equation 6 decomposes into two terms

$$\begin{aligned}&\hat{I}\left(\mathcal{C}, \mathcal{Z}, \Theta\left(\mathcal{Z}\right)\right) = \\ &\quad \frac{1}{N}\sum_{z_i \in \mathcal{Z}} \log\left(\frac{1}{\hat{p}\left(z_i;\Theta\left(\mathcal{Z}\right)\right)}\right) - \\ &\quad \sum_{j=1}^{M}\sum_{z_i \in \mathcal{Z}_j} \pi_j \log\left(\frac{1}{\hat{p}\left(z_i|j;\Theta_j\left(\mathcal{Z}_j\right)\right)}\right)\end{aligned} \tag{7}$$

$$= \hat{h}\left(\mathcal{C}, \mathcal{Z}, \Theta\left(\mathcal{Z}\right)\right) - \sum_{j=1}^{M} \pi_j \hat{h}\left(\mathcal{C}_j, \mathcal{Z}_j, \Theta\left(\mathcal{Z}_j\right)\right). \tag{8}$$

where the first term is the resubstitution estimate of the entropy of $p\left(Z\right)$ and the second term is the resubstitution estimate of the conditional entropy of $p\left(Z|C\right)$. Following 8, we will derive a gradient update for $\hat{h}$ which can be linearly combined to compute a gradient update for $\hat{I}$.

3.2. Gradient Updates

In Equation 6, dependence on G comes through two terms – the samples over which the estimate is evaluated and the parameters. For the remainder of the discussion we will assume that $\hat{p}\left(\bullet\right)$ is a KDE, in which case the parameters are comprised of a kernel size and the set of samples

$$\Theta_j = \{\mathcal{Z}_j, \sigma_j\} = \{g\left(\mathcal{X}_j;G\right), \sigma_j\}$$

and

$$\begin{aligned}\hat{p}\left(z|j;\Theta_j\right) &= \frac{1}{N_{\Theta_j}}\sum_{\theta_i \in \Theta_j} k\left(z - \theta_i, \sigma_j\right), \\ &= \frac{1}{|\mathcal{Z}_j|}\sum_{z_i \in \mathcal{Z}_j} k\left(z - z_i, \sigma_j\right), \\ &= \frac{1}{|\mathcal{X}_j|}\sum_{x_i \in \mathcal{X}_j} k\left(g(x;G) - g(x_i;G), \sigma_j\right)\end{aligned}$$

where $|\Theta_j| = N_{\Theta_j} + 1$.

Expanding the resubstitution estimate of the entropy of $p\left(Z|C\right)$ yields

$$\hat{h}\left(\mathcal{C}_j, \mathcal{Z}_j, \Theta_j\right) = -\frac{1}{|\mathcal{Z}_j|}\sum_{z_i \in \mathcal{Z}_j} \log \hat{p}\left(z_i|j;\Theta_j\right) \tag{9}$$

[2] If this is not the case, i.e. $\pi_j \neq |\mathcal{Z}_j|/N$, there is a straightforward modification.

By the chain rule

$$\nabla_G \hat{h} = \left(\nabla_{\mathcal{Z}} \hat{h} + \nabla_\Theta \hat{h} \nabla_{\mathcal{Z}} \Theta \right) \nabla_G \mathcal{Z}. \quad (10)$$

where we have dropped the explicit dependence on j which denotes the set over which the gradient is being evaluated. Conditional terms are computed over subsets with common labels (i.e. $\mathcal{Z}_j$) while the unconditional terms are computed over the full set of samples $\mathcal{Z}$. Taking each term separately, the gradient of the entropy estimate with respect to a set of samples $\mathcal{Z}$ is

$$\nabla_{\mathcal{Z}} \hat{h} = \left[\nabla_{z_1} \hat{h}, \ldots, \nabla_{z_{N_z}} \hat{h} \right] \quad (11)$$

$$\nabla_{z_k} \hat{h} = -\frac{1}{N_z} \frac{1}{\hat{p}\left(z_k; \Theta\right)} \nabla_{z_k} \hat{p}\left(z_k; \Theta\right) \quad (12)$$

$$\nabla_{z_k} \hat{p}\left(z_k; \Theta\right) = \frac{1}{N_\Theta} \sum_{j=1}^{N_\Theta} k^{'}\left(z_k - \theta_j; \sigma\right) \quad (13)$$

where $N_z = N_\theta = N$ or $N_z = N_\theta = |\mathcal{Z}_j|$ depending on whether an unconditional or conditional entropy gradient is being evaluated and $k^{'}\left(\bullet; \sigma\right)$ is the derivative of the kernel $k\left(\bullet; \sigma\right)$.

The gradient with respect to the parameters Θ for a KDE is:

$$\nabla_\Theta \hat{h} = \left[\nabla_{\theta_1} \hat{h} \ldots \nabla_{\theta_N} \hat{h}, \nabla_\sigma \hat{h} \right] \quad (14)$$

$$\nabla_{\theta_k} \hat{h} = -\frac{1}{N_z} \sum_{i=1}^{N_z} \frac{1}{\hat{p}\left(z_i; \Theta\right)} \nabla_{\Theta_k} \hat{p}\left(z_i; \Theta\right) \quad (15)$$

where

$$\nabla_{\theta_k} \hat{p}\left(z_i; \Theta\right) = -\frac{1}{N_\Theta} k^{'}\left(z_i - \theta_k; \sigma\right) \quad (16)$$

For fixed σ

$$\nabla_{\mathcal{Z}} \Theta = \begin{bmatrix} \mathcal{I} & 0 \\ 0 & 0 \end{bmatrix} \quad (17)$$

For simplicity we will assume a linear mapping, that is $z_i = g\left(\mathcal{X}_i; G\right) = G^T \mathcal{X}_i$ and consequently

$$\nabla_G \mathcal{Z} = \mathcal{X}^T \quad (18)$$

other mappings are possible. Combined, equations 12, 13, 15, 16, 17, and 18 comprise the terms for computing $\nabla_{\hat{h}} G$. Consequently,

$$\nabla_G \hat{I} = \nabla_G \hat{h}\left(\mathcal{C}, \mathcal{Z}, \Theta\right) - \sum_{j=1}^{M} \pi_j \nabla_G \hat{h}\left(\mathcal{C}_j, \mathcal{Z}_j, \Theta_j\right). \quad (19)$$

3.3. Affine Invariance

Ali-Silvey measures are invariant to invertible transformations of the random variables; however, nonparametric estimators of these measures are generally not invariant to such transformations. Here we show how invariance can be incorporated directly into the gradient calculations. The alternative is to rescale both the kernel size and the data throughout the optimization for which there is no analytic approach.

Here we present the case for linear mappings as in the previous development. The standard $(k+1)$th gradient ascent update to G be expressed

$$G^{(k+1)} = G^{(k)} + \Delta_G$$

where $\Delta_G \propto \nabla_G \hat{I}$. However, we wish to define a gradient update which *cannot* be explained as an invertible transformation in the low-dimensional feature space. Such changes in the feature space do not reflect a *real* change in expected information. The condition which ensures that the perturbation is affine invariant is

$$\Delta_G^T G^{(k)} = 0, \quad (20)$$

that is, the columns of $G^{(k)}$ are orthogonal to the columns of Δ_G. This results in a constrained optimization

$$\max_{\Delta_G} \nabla_G \hat{I}^T \Delta_G$$
$$\text{such that } \Delta_G^T \Delta_G = \mu$$
$$\text{and } \Delta_G^T G^{(k)} \Delta_G = 0$$

the solution to which is

$$\Delta_G = \mu (I - G^{(k)} ({G^{(k)}}^T G^{(k)})^{-1} {G^{(k)}}^T) \nabla_G \hat{I}$$

That is, ΔG is proportional to the projection of $\nabla_G \hat{I}$ onto the orthogonal subspace of $G^{(k)}$.

4. EMPIRICAL RESULTS

We present experiments which illustrate various aspects of the approach. We are primarily concerned with two issues. The first issue is the sensitivity of the optimization to kernel size. We find that one can use an overly smooth kernel (i.e. one that gives a poor estimate of mutual information) and yet still leads to an informative projection. The second issue is to what degree the empirical estimate of $I(C; Z)$ is useful for performance prediction. This essentially amounts to how close the resubstitution estimate of mutual information in the feature space is to the actual information. Not surprisingly, the estimator is sensitive to kernel size. However, the empirical implications are that one can use a wide range of kernel sizes to find the projection and then use a leave-one-out kernel size to estimate information content.

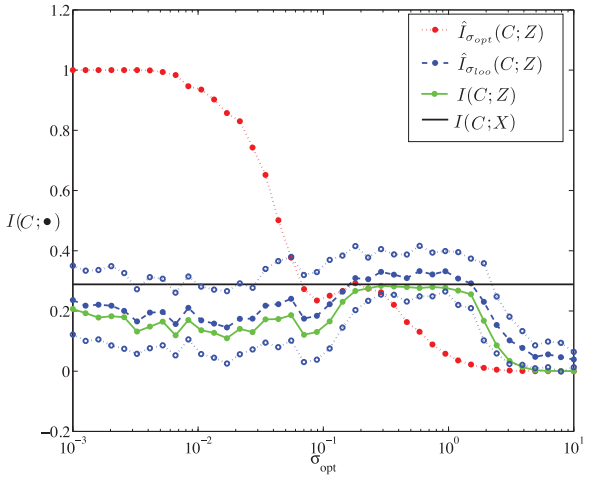

Fig. 4. For each kernel size 50 samples were generated with 100 points each ($N = 100$). The average MI for each sampling as measured using the optimization kernel size, σ_{opt}, a leave-one-out estimate σ_{loo} and the known distribution are shown in red, blue and green respectively. The dotted blue lines above and below $\hat{I}_{\sigma_{loo}}$ represent +/- the variance using that estimate.

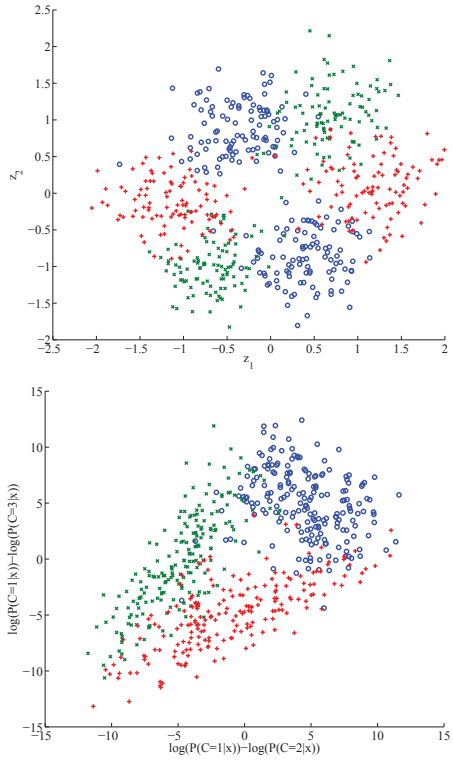

Fig. 5. Mapping of 4-D measurement space to 2-D feature space (top) as compared to log-likelihood space (bottom). Intrinsic dimension of class information is 2-D.

Consider a two class problem with univariate Gaussian class conditional distributions, embedded with an extra dimension of independent Gaussian noise ($M = 2$, $D = 2$). In this case a linear projection to one dimension is sufficient for optimal classification ($L = 1$). For each kernel size in the logarithmic range $[10^{-3}, 10^{1}]$, the optimization produces a projection G, from which we record mutual information in three different ways: 1) using the specified kernel size $\hat{I}_{\sigma_{opt}}(C;Z)$, 2) using a size which optimizes the leave-one-out likelihood of the projected data $\hat{I}_{\sigma_{loo}}(C;Z)$ *upon completion of the optimization using a fixed kernel size*, and 3) using the known distribution in the learned feature space $I(C;Z)$. Note that the affine invariant gradient of the previous section allows the use of a fixed kernel size throughout the optimization. This eliminates the computationally burdensome step of adapting the kernel size throughout the optimization. Figure 4 shows the impact of kernel size on these values.

The optimization succeeds for sizes in the range $[.2, 1.5]$, where the mutual information $I(C;Z)$ is close to $I(C;X)$. Moreover, while $\hat{I}_{\sigma_{opt}}(C;Z)$ varies significantly as a function of size, $\hat{I}_{\sigma_{loo}}(C;Z)$ is close to $I(C;Z)$. In general, smaller sizes overestimate the mutual information while larger sizes underestimate. Informative projections can be found over a range of sizes with leave-one-out kernel size providing an accurate estimate of mutual information as a final step.

Next we show some simple figures that illustrate the nature of the learned projections. It is well known that the log-likelihood ratio is a sufficient statistic for classification. Figure 5(top) shows a learned 2D feature space of 2D data embedded in 4D with two extra noise dimensions. Figure 5(bottom) is the corresponding log likelihood ratio space. As be seen both preserve class information (though we shall quantify this), but in different ways. The difference being that the feature space reflects the multi-modal nature of the data, which is a point of comparison only.

In Figure 6 the 2D learned feature and log likelihood ratio spaces are shown for 1D data embedded in 4D (i.e. there are 3 extra noise dimensions). While the log-likelihood ratio is a sufficient statistic, it is not minimal, as can be seen by the degeneracy of the figure. The feature space is also not minimal, however, one can see in this simple example that optimization over a 1D feature space would produce a near minimal sufficient statistic.

Finally, figure 7 shows the results of multiple trials in which class separability (i.e. P_ϵ) is varied. In these experiments the feature space matches the intrinsic dimension of the class information. Blue marks indicate the P_ϵ and $I(C;X)$ of the known model. Red marks indicate the P_ϵ and true $I(C;Z)$ obtained from the learned projection over multiple trials. Finally, green marks indicate P_ϵ and *estimated* $I(C;Z)$ from the resubstitution estimate. As can be seen for low P_ϵ accurate bounds and informative features are learned while for higher P_ϵ estimates of information content have an optimistic

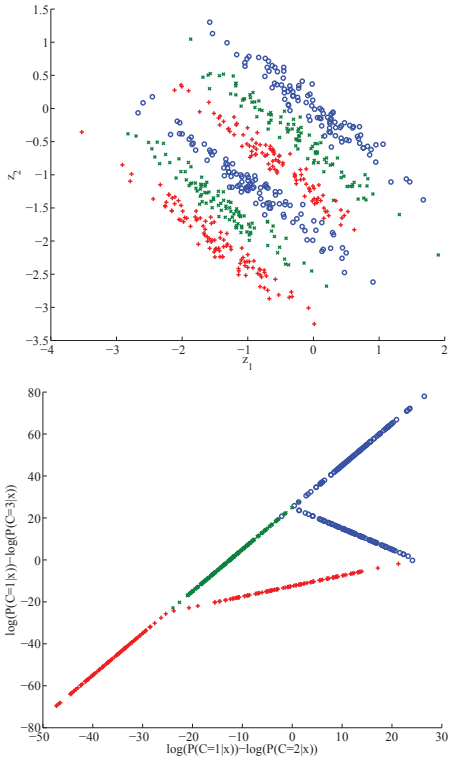

Fig. 6. Mapping of 4-D measurement space to 2-D feature space(top) as compared to log-likelihood space (bottom). Intrinsic dimension of class information is 1-D.

bias with one clear outlier. These points are overlain on the known Fano and H-R bounds for a $M = 3$. Excepting one outlier, all estimates of $I(C;Z)$ give accurate bounds on P_ϵ.

5. DISCUSSION

We extended a method for extracting informative features first suggested by [1]. Affine invariance precludes the computationally burdensome step of adjusting kernel size thoughout the optimization. Empirical results indicate that useful projections may be *learned* using overly smooth kernels, though some other means (e.g. leave-one-out cross-validation) was needed to accurately esitimate actual information content. Furthermore, in our experiments, estimated information content gave accurate performance bounds.

6. REFERENCES

[1] John W. Fisher III and Jose C. Principe, "A methodology for information theoretic feature extraction," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, A. Stuberud, Ed., 1998.

[2] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

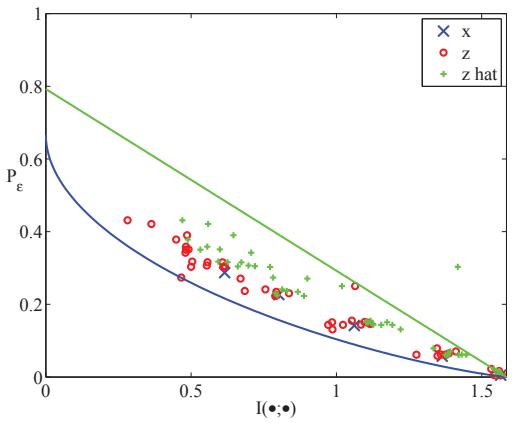

Fig. 7. Comparison of P_ϵ of learned features to known model as well as $\hat{I}(C;Z)$ to true $I(C;Z)$.

[3] Michele Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349–369, Dec 1989.

[4] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "On surrogate loss functions and f-divergences," *Annals of Statistics*, to appear.

[5] Peter L. Bartlett, Michael I. Jordan, and Jon D. Mcauliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, 2003.

[6] Solomon Kullback, *Information Theory and Statistics*, John Wiley and Sons, New York, 1959.

[7] R.M. Fano, "Class notes for transmission of information, course 6.574," MIT, Cambridge, MA, 1952.

[8] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., New York, 1991.

[9] M. Hellman and J. Raviv, "Probability of error, equivocation, and the chernoff bound," *Information Theory, IEEE Transactions on*, vol. 16, no. 4, pp. 368–372, Jul 1970.

[10] I.K. Sethi, "Entropy nets: from decision trees to neural networks," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1605–1613, Oct 1990.

[11] C.H. Chen, "Theoretical comparison of a class of feature selection criteria in pattern recognition," *Computers, IEEE Transactions on*, vol. C-20, no. 9, pp. 1054–1056, Sept. 1971.

[12] R. Battiti, "Using the mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, pp. 537–550, 1994.

[13] K.D. Bollacker and J. Ghosh, "Mutual information feature extractors for neural classifiers," *Neural Networks, 1996., IEEE International Conference on*, vol. 3, pp. 1528–1533 vol.3, Jun 1996.

[14] John W. Fisher III and Alan S. Willsky, "Information theoretic feature extraction for atr (invited paper)," in *Proceedings of 34th Asilomar Conference on Signals, Systems, and Computers*, Fred J. Taylor, Ed., Pacific Grove, CA, October 1999.