"Sensing-Aware Classification with High-Dimensional Data"

B. Orten, P. Ishwar, W. Karl, V. Saligrama and H. Pien

ICASSP 2011, p. 3700-3703

CISE Technical Report #2011-CA-0019
CISE Technical Report Date: November 30, 2011

Center for Information and Systems Engineering
15 St. Mary's Street, Brookline, MA 02446
617.358.1295

SENSING-AWARE CLASSIFICATION WITH HIGH-DIMENSIONAL DATA

Burkay Orten[†] *Prakash Ishwar*[†] *W. Clem Karl*[†] *Venkatesh Saligrama*[†] *Homer Pien*[‡]

[†] Department of Electrical and Computer Engineering, Boston University, Boston, MA, USA.
[‡]Department of Radiology, Massachusetts General Hospital, Boston, MA, USA.

ABSTRACT

In many applications decisions must be made about the state of an object based on indirect noisy observation of high-dimensional data. An example is the determination of the presence or absence of stroke from tomographic projections. Conventionally, the sensing process is inverted and a classifier is built in the reconstructed domain, which requires complete knowledge of the sensing mechanism. Alternatively, a direct data domain classifier might be constructed, but the constraints imposed by the sensing process are then lost. In this work we study the behavior of a third path we term "sensing-aware classification." Our aim is to contribute to the development of a rigorous theory for such challenging problems. To this end, we consider an abstracted binary classification problem with very high dimensional observations, a restricting sensing configuration, and unknown statistical models of noise and object which must be learned from constrained training data. We analyze the impact of different levels of prior knowledge concerning the sensing mechanism for various classification strategies. In particular we prove that the strategies based on the naive estimation of all model elements results in a classification performance asymptotically no better than guessing whereas sensing-aware, projection-based classification rules attain Bayes-optimal risk. Simulation results are also provided.

Index Terms— Learning, classification, linear discriminant analysis, high-dimensional data, asymptotic analysis

1. INTRODUCTION

In many important applications decisions must be made about the state of an object, such as the determination of the presence or absence of stroke in a brain. Such problems are further complicated when the object of interest is indirectly related to the observed data through a sensing process, as arises in tomographic projection data. A common approach in such situations is to first solve an inverse problem, transforming the observed data, and then to make decisions on the inverted or transformed data. Such data inversion can be extremely costly, requires knowledge of the sensing model, and, if done poorly, can introduce undesired artifacts into the resulting transformed data which can confound subsequent analysis. An alternative approach is to build classifiers directly in the original high-dimensional data domain [1]. But then knowledge of the sensing process is generally treated as unknown hidden structure and often ignored. Such sensing problems also arise in econometrics where the aim is understanding of market risk factors [2], in drug response studies where gene array data can predict whether a group of patients will benefit from drug therapy [3] and in oncology where breast cancer DNA micro-arrays demonstrate certain properties of the underlying tumor [4]. In this work we study a third path we term "sensing-aware classification" that couples observed data with prior information concerning sensing models for data-domain classification. We examine how incorporation of different types of such prior information about the sensing model affects asymptotic classification performance.

Our aim in this work is to contribute to a rigorous analysis and understanding of such challenging problems. To this end, we consider an abstracted binary classification problem with the following elements:

- High dimensional observations
- A restricting sensing configuration
- Unknown statistical models of noise and object
- Estimation of class models from constrained training data
- Fixed signal-to-noise ratio (SNR) observations

We study the effects of different types of prior knowledge concerning the sensing model on classification performance relative to the naive Fisher approach assuming no knowledge and full training of the model from available data samples. We prove that asymptotically the Fisher approach is no better than guessing while incorporation of various forms of prior sensing knowledge can result in classification performance as good as if complete model information were available. In particular, we prove that certain projection-based classifiers can achieve this optimal performance under assumptions on the structure of the sensing mechanism. Also, in our asymptotic

This work was supported in part by the U.S. Department of Homeland Security under award number 2008-ST-061-ED0001, by the U.S. AFOSR under award numbers #FA9550-06-1-0324 and #FA9550-10-1-0458 (subaward A1795) and by the Defense Advanced Research Projects Agency under award number N66001-10-1-2133. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the U.S. Department of Homeland Security, U.S. AFOSR or DARPA.

analysis we are careful to hold the SNR for the problem constant.

The paper is organized as follows. Section 2 gives a detailed description of our abstracted learning problem and provides analytical expressions for the classification performance of the conventional linear discriminant rule. In Section 3 three scenarios corresponding to different levels of prior knowledge on the sensing model are analyzed in more detail and their asymptotic performances are investigated. Numerical simulations provided in Section 4 validate our theoretical findings and present a quantitative comparison of different methods. Section 5 concludes the paper with some final remarks.

2. PROBLEM DEFINITION

We consider the binary classification problem of assigning a label to a (previously unseen) sample $\mathbf{y} \in \mathbb{R}^p$, coming from one of two equally likely classes, which we arbitrarily denote as class 0 and class 1. Under class i we assume that the observation is generated according to the following sensing model:

$$\mathbf{Y}_i = \mathbf{h}X_i + \mathbf{Z}_i, \;\; i \in \{0,1\}, \; X_i \in \mathbb{R}, \mathbf{Y}_i, \mathbf{h}, \mathbf{Z}_i \in \mathbb{R}^p.$$

Here, $\mathbf{Y}_i$ is the observed data, X_i represents an unknown latent object, $\mathbf{Z}_i$ is the additive noise cluttering the observations and $\mathbf{h}$ captures a non-random sensing mechanism which constrains the object contribution to the data.[1] We assume that $X_i \sim \mathcal{N}(m_i, \sigma_x^2)$ and $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_z^2\mathbf{I})$, but these models and/or their parameters are generally unknown a priori. Also, we consider $\mathbf{h} \in \mathbb{R}^p$ to be the first p components of an infinitely long sequence $\mathbf{h}^{full} \in \ell_2$ (i.e. $\mathbf{h}^{full}$ has with finite 2-norm), renormalized so that $||\mathbf{h}|| = ||\mathbf{h}^{full}||$.

We assume that n independent identically distributed (i.i.d.) training samples $\mathbf{y}_{ij}$, $j = 1, .., n$ from each class are available. It then follows that the observed training samples have the following density under each class:

$$p_i(\mathbf{Y}) = \mathcal{N}(\mathbf{h}m_i, \sigma_x^2\mathbf{h}\mathbf{h}^T + \sigma_z^2\mathbf{I}) =: \mathcal{N}(\boldsymbol{\mu}_\mathbf{i}, \boldsymbol{\Sigma}). \quad (1)$$

For convenience we define $d := (m_1 - m_0)$, $m := (m_1 + m_0)/2, \boldsymbol{\Delta} := (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$, $\boldsymbol{\mu} := (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)/2$ and assume w.l.o.g. that $m_1 > m_0$. If $\boldsymbol{\Delta}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known then the minimum probability of error classifier is the Bayes rule:

$$\delta_B(\mathbf{y}) = \mathbf{1}\Big\{\boldsymbol{\Delta}^\mathbf{T}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{y} - \boldsymbol{\mu}) \geq \mathbf{0}\Big\} \quad (2)$$

and the probability of misclassification, $P_M^{\delta_B}$, is given by:

$$P_M^{\delta_B} = Q\left(\frac{1}{2}\sqrt{\boldsymbol{\Delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}}\right) = Q\left(C_1(d, \sigma_x, \sigma_z, \mathbf{h})\right) \quad (3)$$

where $C_1(d, \sigma_x, \sigma_z, \mathbf{h}) = \left(0.5d/(\sigma_x^2 + \sigma_z^2/||\mathbf{h}||^2)^{0.5}\right)$ and $1 - Q(x)$ is the cumulative distribution function of the standard normal.

[1] Upper-case symbols denote random quantities and corresponding lower-case symbols their instantiations. Bold-face symbols are used for matrices whose dimension will be clear from the context.

Note that the Bayes rule requires both the knowledge of the sensing model and the model parameters. As previously discussed these are generally unknown in advance and the classification rule needs to be designed using only the available training data. In the next section we will consider classification rules with varying levels of prior knowledge about the sensing process and analyze their asymptotic performance as for fixed SNR. Specifically, σ_x, σ_z, d and $||\mathbf{h}|| = ||\mathbf{h}^{full}||$ are held fixed as $p, n \to \infty$ so that the difficulty level of the problem remains constant.

3. ANALYSIS OF SENSING-AWARE CLASSIFICATION RULES

In this section we consider three scenarios representing varying levels of prior knowledge about the sensing model. For each scenario we propose a classification rule $\hat{\delta}$ whose structure is tuned to the prior knowledge and whose parameters are estimated from the training data. As $p, n \to \infty$ we analyze the asymptotic expected misclassification error probability $\mathbb{E}[P_M^{\hat{\delta}}(\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})]$ where the expectation is over the randomness in the training samples.

Scenario 1: "Nothing known"

In this scenario we assume that there is no knowledge of the sensing process and only partial information is available about the distribution of the observations. Specifically, we assume that $\mathbf{Y}_i \sim \mathcal{N}(\boldsymbol{\mu} + (2i-1)\boldsymbol{\Delta}/2, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ is known, $\boldsymbol{\Delta}$ and $\boldsymbol{\Sigma}$ are unknown. Since the optimum classification rule for the known parameter case is given by (2), a natural idea is to use this rule in our current scenario where the known parameters $\boldsymbol{\Delta}$ and $\boldsymbol{\Sigma}$ are replaced with their estimated counterparts. The resulting *empirical* Fisher rule and its misclassification probability are given by:

$$\delta_F(\mathbf{y}; \hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\Sigma}}) = \mathbf{1}\Big\{\hat{\boldsymbol{\Delta}}^\mathbf{T}\hat{\boldsymbol{\Sigma}}^{-\mathbf{1}}(\mathbf{y} - \boldsymbol{\mu}) \geq \mathbf{0}\Big\} \quad (4)$$

$$P_M^{\delta_F} = Q\left(\frac{\hat{\boldsymbol{\Delta}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Delta}}}{2(\hat{\boldsymbol{\Delta}}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Delta}})^{\frac{1}{2}}}\right) =: Q\left(\Psi_\Sigma(\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\Sigma}})\right) \quad (5)$$

where the ML estimates of $\boldsymbol{\Delta}$ and $\boldsymbol{\Sigma}$ are computed as

$$\hat{\boldsymbol{\Delta}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{y}_{1j} - \frac{1}{n}\sum_{j=1}^{n}\mathbf{y}_{0j} \quad (6)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{2(n-1)}\sum_{i=0}^{1}\sum_{j=1}^{n}(\mathbf{y}_{ij} - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_{ij} - \hat{\boldsymbol{\mu}}_i)^T \quad (7)$$

and $\hat{\boldsymbol{\mu}}_1 = \boldsymbol{\mu} + \hat{\boldsymbol{\Delta}}/2$ and $\hat{\boldsymbol{\mu}}_0 = \boldsymbol{\mu} - \hat{\boldsymbol{\Delta}}/2$. Our main result for this scenario is that the Fisher rule for an arbitrary $\boldsymbol{\Sigma}$ performs no better than random guessing when dimensionality increases faster than the number of training examples as demonstrated in the following theorem.

Theorem 1. *Let $\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\Sigma}}$ be as defined in (6) and (7) respectively. If $(p/n) \to \infty$, then $\mathbb{E}_{\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\Sigma}}}\left[Q\left(\Psi_\Sigma(\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\Sigma}})\right)\right] \to \frac{1}{2}$.*

Proof sketch: Bickel and Levina proved this result for the case of $\mathbf{\Sigma} = \mathbf{I}$ in [5]. Observe that for $\mathbf{\Sigma} = \mathbf{I}$:

$$0.5 \overset{(i)}{\geq} \mathbb{E}\left[Q\left(\Psi_I(\hat{\mathbf{\Delta}}, \hat{\mathbf{\Sigma}})\right)\right] \overset{(ii)}{\geq} Q\left(\mathbb{E}\left[\Psi_I(\hat{\mathbf{\Delta}}, \hat{\mathbf{\Sigma}})\right]\right) \quad (8)$$

where (i) is because the maximum value of the $Q(x)$ for non-negative arguments is 0.5 and (ii) is due to Jensen's inequality. In [5] it is shown that as $(p/n) \to \infty, \mathbb{E}\left[\Psi_I(\hat{\mathbf{\Delta}}, \hat{\mathbf{\Sigma}})\right] \to 0$ which implies that $Q\left(\mathbb{E}\left[\Psi_I(\hat{\mathbf{\Delta}}, \hat{\mathbf{\Sigma}})\right]\right) \to 1/2$ by the dominated convergence theorem. For any general non-singular $\mathbf{\Sigma}$, $\Psi_{\mathbf{\Sigma}}(\hat{\mathbf{\Delta}}, \hat{\mathbf{\Sigma}}) = \Psi_I(\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Delta}}, \mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2}) \to 0$ as shown above and the proof follows. ■

Scenario 2: Both sensing structure and $\mathbf{h}$ *known*

In this scenario, the observations are known to be generated according to the sensing model $\mathbf{Y}_i = \mathbf{h}X_i + \mathbf{Z}_i$ with $X_i \sim \mathcal{N}(m_i, \sigma_x^2)$ and $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_z^2\mathbf{I})$ but σ_x and σ_z are unknown. This represents the most informative scenario in that one not only knows the linear sensing structure but also has access to $\mathbf{h}$. In this case, variability is controlled through the random object X and $\mathbf{h}$ merely serves as a direction in $\mathbb{R}^p$ where the one dimensional data is lying on. The effect of the added noise and dimensionality can be reduced dramatically if we project the data samples onto this known direction and carry out the classification on the projected space. Before analyzing the performance of the projected classifier, we can derive the empirical Fisher rule after projection onto any given direction $\mathbf{r}$ and its misclassification probability as:

$$\delta_r(\mathbf{y}; \hat{\mathbf{\Delta}}, \hat{\boldsymbol{\mu}}) = \mathbf{1}\Big\{ \operatorname{sign}(\hat{\mathbf{\Delta}}^T\mathbf{r})\mathbf{r}^T(\mathbf{y} - \hat{\boldsymbol{\mu}})) \geq 0\Big\} \quad (9)$$

$$P_M^{\delta_r} = \frac{1}{2}\sum_{i=0}^{1} Q\left((-1)^i \frac{\operatorname{sign}(\hat{\mathbf{\Delta}}^T\mathbf{r})\mathbf{r}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_i)}{\sqrt{\mathbf{r}^T\mathbf{\Sigma}\mathbf{r}}}\right) \quad (10)$$

The main result for this scenario is that if the direction $\mathbf{r}$ is chosen to be the unit vector in the direction of $\mathbf{h}$, we can attain the Bayes optimal performance asymptotically.

Theorem 2. *If* $\mathbf{r} = \mathbf{h}/||\mathbf{h}||$, *then* $P_M^{\delta_r} \xrightarrow{p} Q(C_1(d, \sigma_x, \sigma_z, \mathbf{h}))$

Proof sketch: Define $C_2(\mathbf{h}) := \sqrt{\sigma_x^2\mathbf{h}\mathbf{h}^T + \sigma_n^2}$, and $\tilde{\mathbf{h}} := \frac{\mathbf{h}}{||\mathbf{h}||}$. Then

$$P_M^{\delta_h} = \frac{1}{2}\sum_{i=0}^{1} Q\left((-1)^i \operatorname{sign}\left(\frac{\hat{\mathbf{\Delta}}^T\tilde{\mathbf{h}}}{C_2(\mathbf{h})}\right)\frac{\tilde{\mathbf{h}}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_i)}{C_2(\mathbf{h})}\right) \quad (11)$$

Since

$$\frac{\hat{\mathbf{\Delta}}^T\tilde{\mathbf{h}}}{C_2(\mathbf{h})} \sim \mathcal{N}(2C_1(d, \sigma_x, \sigma_z, \mathbf{h}), \frac{2}{n}) \xrightarrow{p} 2C_1(d, \sigma_x, \sigma_z, \mathbf{h})$$

$$\frac{\tilde{\mathbf{h}}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)}{C_2(\mathbf{h})} \sim \mathcal{N}(C_1(d, \sigma_x, \sigma_z, \mathbf{h}), \frac{1}{2n}) \xrightarrow{p} C_1(d, \sigma_x, \sigma_z, \mathbf{h})$$

and $C_1 > 0$, the product $\operatorname{sign}(\hat{\mathbf{\Delta}}^T\tilde{\mathbf{h}}/C_2(\mathbf{h}))(\tilde{\mathbf{h}}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_i)/C_2(\mathbf{h}))$ converges in probability to $C_1(d, \sigma_x, \sigma_z, \mathbf{h})$ by Slutsky's theorem. Finally, since Q is a continuous function we conclude that $P_M^{\delta_r} \xrightarrow{p} Q\left(C_1(d, \sigma_x, \sigma_z, \mathbf{h})\right)$. ■

Observe that this asymptotic result requires only $n \to \infty$ and does not depend on p, i.e., the projected classifier will perform as good as the optimal Bayes rule irrespective of how p scales with n if we have enough training samples.

Scenario 3: Sensing structure known $\mathbf{h}$ *unknown*

The prior information available in this scenario is similar to the previous one with the exception of $\mathbf{h}$ being unknown. We have seen that if $\mathbf{h}$ is known, classification on the projected data is asymptotically as good as the Bayes classifier. However, if the only available knowledge is the linear sensing structure and $\mathbf{h}$ is not given, it will be clear that the data essentially lies along a line whose direction is unknown. Then a reasonable approach is to estimate $\mathbf{h}$ using the training samples. Since $\mathbf{\Delta} = \mathbf{h}d$ is along the same direction as the $\mathbf{h}$ itself, an immediate idea is to use $\hat{\mathbf{\Delta}}$ as the estimate of this unknown direction. As in the second scenario, projecting samples onto an estimate of $\mathbf{\Delta}$ allows one to have a simpler classification rule which does not require an estimate of the covariance matrix and the new classification rule using $\mathbf{r} = \hat{\mathbf{\Delta}}$ in (9) is:

$$\delta_{\hat{\Delta}}(\mathbf{y}; \hat{\mathbf{\Delta}}, \hat{\boldsymbol{\mu}}) = \mathbf{1}\Big\{\hat{\mathbf{\Delta}}^{\mathbf{T}}(\mathbf{y} - \hat{\boldsymbol{\mu}}) \geq \mathbf{0}\Big\} \quad (12)$$

One might expect the rule that is classifying samples that are projected onto $\hat{\mathbf{\Delta}}$ to perform quite well. However, as we will demonstrate below this method performs poorly since the sample mean $\hat{\mathbf{\Delta}}$, which is the maximum likelihood estimator, has a high variance that goes to infinity as $p \to \infty$. We can overcome this problem by replacing the sample mean with a regularized estimate of $\mathbf{\Delta}$ as shown below.

Theorem 3. *Let* $\hat{\mathbf{\Delta}}$ *be the sample mean and* $\delta_{\hat{\Delta}}(\mathbf{y}; \hat{\mathbf{\Delta}}, \hat{\boldsymbol{\mu}})$ *be the classification rule as defined above.*

a) If $\frac{p}{n} \to \infty$ *and* $\frac{p}{n^2} \to 0$ *then* $P_M^{\delta_{\hat{\Delta}}} \xrightarrow{p} \frac{1}{2}$.

b) Define the truncating estimator as $\hat{\mathbf{\Delta}}_t = W\hat{\mathbf{\Delta}}$ *where* $\mathbf{W} := diag(\mathbf{w})$, $w_i = \mathbf{1}(i \leq t)$. *If* $t \to \infty, \frac{t}{n} \to 0$ *then* $P_M^{\delta_{\hat{\Delta}_t}} \xrightarrow{p} P_M^{\delta}$.

Proof sketch: a) The proof is parallel to that proof of Theorem 2. The probability of error for $\delta_{\hat{\Delta}}(\mathbf{y}; \hat{\mathbf{\Delta}}, \hat{\boldsymbol{\mu}})$ is:

$$P_M^{\delta_{\hat{\Delta}}} = \frac{1}{2}\sum_{i=0}^{1} Q\left((-1)^i \frac{\hat{\mathbf{\Delta}}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_i)}{\sqrt{\hat{\mathbf{\Delta}}^T\mathbf{\Sigma}\hat{\mathbf{\Delta}}}}\right) \quad (13)$$

For the numerator term notice that $\mathbb{E}[\hat{\mathbf{\Delta}}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_i)] = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||^2/2$ and $\operatorname{var}(\hat{\mathbf{\Delta}}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_i)) = O(p)/n^2 + O(||\mathbf{h}||^4)/n$. Since the variance of the numerator term goes to 0, under the specified asymptotic conditions it converges in the mean square sense to its expected value. A similar analysis for the denominator shows that $\mathbb{E}[(\hat{\mathbf{\Delta}}^T\mathbf{\Sigma}\hat{\mathbf{\Delta}})^{1/2}] \to \infty$ and $\operatorname{var}((\hat{\mathbf{\Delta}}^T\mathbf{\Sigma}\hat{\mathbf{\Delta}})^{1/2}) \to 0$. However $1/(\hat{\mathbf{\Delta}}^T\mathbf{\Sigma}\hat{\mathbf{\Delta}})^{1/2} \xrightarrow{p} 0$ as $p, n \to \infty$. Since the numerator and denominator both converge in probability, by Slutsky's and Continuous Mapping theorems it follows that $P_M^{\delta_{\hat{\Delta}}} \xrightarrow{p} Q(0) = 1/2$.

b) Although the sample mean is an unbiased estimator, its variance scales with dimensionality p making it a poor estimator of $\mathbf{h}$ in our asymptotic setup. We can obtain a regularized estimate of $\mathbf{\Delta}$ by truncating $\hat{\mathbf{\Delta}}$ to $\mathbf{W}\hat{\mathbf{\Delta}}$. Then the bias-variance decomposition of $\hat{\mathbf{\Delta}}_t = \mathbf{W}\hat{\mathbf{\Delta}}$ reveals that:

$$\begin{aligned}\mathbb{E}[||\hat{\mathbf{\Delta}}_t - \mathbf{\Delta}||^2] &= \mathbf{\Delta}^T(\mathbf{I}-\mathbf{W})\mathbf{\Delta} + \mathrm{tr}\left(\frac{2}{n}\mathbf{W}\mathbf{\Sigma}\mathbf{W}\right) \\ &= d\sum_{k=1}^{p}(1-w_i)^2h_i^2 + \frac{2}{n}\sum_{k=1}^{p}w_i^2\sigma_{ii}^2 \\ &= d\sum_{k=t+1}^{p}h_i^2 + \frac{2}{n}\sum_{k=1}^{t}\sigma_{ii}^2\end{aligned}$$

As $t \to \infty$, the bias term goes to 0 (since $\mathbf{h}^{full} \in \ell_2$) whereas the $\frac{t}{n} \to 0$ condition ensures a vanishing variance, i.e., $\hat{\mathbf{\Delta}}_t$ is an asymptotically unbiased and consistent estimator of $\mathbf{\Delta}$. Since $\hat{\mathbf{\Delta}}_t \xrightarrow{p} \mathbf{\Delta}$, using convergence theorems, $P_M^{\delta_{\hat{\Delta}_t}} \xrightarrow{p} P_M^{\delta}$. ■

4. SIMULATIONS

In this section we illustrate our theoretical findings using some simulations. In order to capture the effect of growing dimensionality, we consider a sequence of problems which are "equally difficult" in the sense that the Bayes risk P_M^{δ} is kept constant while p is increasing where the goal is to understand the influence of p on the misclassification probability. In our simulations, for $i = 1, ..., p,\ h_i = (0.95)^i, n = \lfloor p^{0.5} \rfloor, \sigma_x = 1, \sigma_n = 2$, and $t = \lfloor n^{0.7} \rfloor$. We varied p from 30 to 1500 and renormalized $\mathbf{h}$ so that $||\mathbf{h}||$ is held fixed at the value $||\mathbf{h}|| = 2$ which ensures that $P_M^{\delta} = 0.1$ for all p. Plots of the misclassification probability versus dimensionality p for different scenarios are shown in Fig. 1. Observe that the Fisher rule's performance decays rapidly as p increases. At the other end of the spectrum is the rule where classification is performed on the samples which are projected onto $||\mathbf{h}||$. As expected this rule's performance is very close to the Bayes risk for almost all p. The unregularized (i.e., ML) estimate-based rule of scenario 3 performs better than the Fisher rule even though the misclassification probability of both methods goes to 0.5 as $p \to \infty$. Finally, regularized (i.e., truncation) estimate-based rule of scenario 3 quickly dominates the unregularized rule for large p and approaches the Bayes risk, consistent with our theoretical findings.

5. CONCLUSIONS

In this paper we considered a sensing-aware approach to classification involving high-dimensional observations coupled with constrained training data. We showed that the popular Fisher rule which ignores the sensing structure asymptotically performs no better than random guessing. Then we demonstrated that complete knowledge of the sensing model (i.e., knowledge of $\mathbf{h}$) can be used to asymptotically obtain optimal Bayes performance in the presence of unknown object and noise models. We also investigated two methods which attempt to incorporate varying degrees of prior sensing model information to estimate $\mathbf{h}$. We proved that using the obvious sample mean estimator fails due to its high variance. Finally, we identified a class of sensing models for which a truncation based regularization approach again yields asymptotically optimal classification performance.

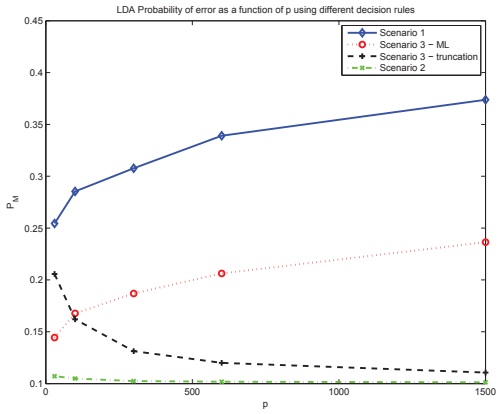

Fig. 1. Performance comparison of 4 classification rules representing different levels of prior knowledge on the sensing process.

6. REFERENCES

[1] A. Sayeed, M. Petrou, N. Spyrou, A. Kadyrov, and T. Spinks, "Diagnostic features of alzheimer's disease extracted from pet sinograms," *Physics in Medicine and Biology*, vol. 47, no. 1, pp. 137, 2002.

[2] H. Lustig, N. Roussanov, and A. Verdelhan, "Common risk factors in currency markets," in *Finance Intl. Meeting*, Paris, 2008, AFFI-EUROFIDAI.

[3] K. M. Borgwardt, S. V. N. Vishwanathan, and H. Kriegel, "Class prediction from time series gene expression profiles using dynamical systems kernels," in *Pacific Symp. of Biocomputing, Maui, HI*, 2006, pp. 547–558.

[4] C. Carvalho, J. Chang, J. Lucas, J. R. Nevins, and Q. Wang, "High-dimensional sparse factor models and latent factor regression," Tech. Rep., Duke Unv, 2005.

[5] P. J. Bickel and E. Levina, "Some theory for fisher's linear discriminant function, "naive bayes", and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, no. 6, pp. 989–1010, 2004.