

# Transductive De-Noising and Dimensionality Reduction using Total Bregman Regression

Sreangsu Acharyya  
University Of Texas at Austin  
Electrical Engineering

## Abstract

Our goal on one hand is to use labels or other forms of ground truth data to guide the tasks of de-noising and dimensionality reduction and balance the objectives of better prediction and better data summarization, on the other hand it is to explicitly model the noise in the feature values. We use a generalization of  $L_2$  loss, on which PCA and K-Means are based, to the Bregman family which, as a consequence widens the applicability of the proposed algorithms to cases where the data may be constrained to lie on sets of integers or sets of labels rather than  $\mathbb{R}^d$  as in PCA or K-means formulations. This makes it possible to handle different prediction tasks such as classification and regression in a unified way. Two tasks are formulated (i) Transductive Total Bregman Regression (ii) Transductive Bregman PCA.

## 1 Introduction

Real world data typically need to be preprocessed by denoising and dimensionality reduction prior to regression or classification. We generalize the widely used dimensionality reduction technique: Principal Component Analysis, to a Transductive setting using Bregman loss function [1]. Popular preprocessing practices of PCA and K-means are not theoretically guided by the end goal of better prediction. The objective here, similar to that of LDA such as Fisher projections is to do (semi) supervised preprocessing. Fisher Projection assumes that, data originates from separate equivariance Gaussians, that leads to linear class decision boundary. What is gained by the formulation proposed here, are four things (i) removal of the requirement for a generative model of the features, as here we work directly with the posterior distribution, (ii) a generalization to the wider model of exponential family of which Gaussian is a special case (iii) applicability to case where labels are available on a subset of the data (iv) balancing quality of prediction and data reconstruction, which may serve as a regularization as demonstrated in the information bottleneck model [12]

**1.1 Related work** Dimensionality reduction literature has primarily focused on unsupervised approaches, among well known exceptions are the classical Fisher discriminant analysis and its kernel variant [2], Mixture Discriminant analysis (MDA) [6] etc. Very recently, independent of our work, an approach for supervised dimensionality reduction has been proposed [11] which improves on the MDA model and discusses generalization to mixtures of exponential family distributions. Though there is an overlap in the use of exponential family models, their approach is quite different from ours. Like previous approaches they follow a generative model for the features, but use maximum conditional likelihood (MCL) algorithm to induce the discriminatory projections. Our approach is more direct and computationally simpler in that we work with a discriminatory model <sup>1</sup> and build on logistic regression and EM.

## 2 Squared Euclidean Loss

**2.1 Principal Component Analysis** considers approximating a set of points  $X_i \in \mathbf{R}^d$ , by their projection on a low-dimensional subspace computed as  $\hat{X}_i = \mathbf{U}\mathbf{U}^\dagger X_i$ , where  $U$  has as its  $k$  columns orthonormal basis vectors  $u_i$  of the low-d subspace. Minimizing the reconstruction error can be formulated as  $\text{Min}_U \sum_i \|X_i - \hat{X}_i\|^2$  s.t.  $\hat{X}_i = \mathbf{U}\mathbf{U}^\dagger X_i$ . This error term can be motivated by the assumption that data-points  $X_i \in \mathbf{R}^d$  were generated independently from as many Gaussian distributions with their means (or equivalently the de-noised data) lying on some low-d subspace.

## 2.2 Total Least Squares Regression [5]

Linear least squares regression is suitable when output of a linear system  $Y_i = \beta^\dagger X_i$  corresponding to precisely known input  $X_i$ , is corrupted by an additive 0-mean Gaussian noise. Often the in-

<sup>1</sup>we do model the joint distribution of the output  $Y$  and input  $X$ , but through parametric assumptions on  $P(Y|X)$  rather than on  $P(X|Y)$  that requires the handling of the partition function

puts cannot be set accurately (in experiments) or correctly observed. TLS formulation accounts for noise both in input and output, it can be posed as  $\text{Min}_{\beta, X_i} \sum \left( \|\hat{Y}_i - \beta^\dagger X_i\|^2 + \|X_i - \hat{X}_i\|^2 \right)$  where  $\hat{Y}_i$  and  $\hat{X}_i$  are the corrupted observed output and input to the system. The cost can be minimized globally in closed form. TLS has an “errors in variable” generative model, where  $\hat{Y}_i|X_i \sim N(\beta^\dagger X_i, I)$  and  $\hat{X}_i \sim N(X_i, I)$ .

**2.3 Supervised PCA** The dimensionality of real life data are frequently reduced using PCA. The reduced variables  $X_i$  obtained from the observed values  $\hat{X}_i$  are then used for the prediction of  $Y_i$ , often, in a linear model. It is desirable to choose the directions of projections such that the accuracy of the regression is minimally diminished. The regression loss is  $\|\hat{Y}_i - \beta^\dagger \mathbf{U}\mathbf{U}^\dagger \hat{X}_i\|^2$  and the dimensionality reduction (reconstruction) loss is  $\|\hat{X}_i - \mathbf{U}\mathbf{U}^\dagger \hat{X}_i\|^2$ . The objective function is posed as a summation of these terms:  $\text{Min}_{\mathbf{U}, \beta} \|\hat{Y}_i - \beta^\dagger X_i\|^2 + \|X_i - \hat{X}_i\|^2$  s.t.  $X_i = \mathbf{U}\mathbf{U}^\dagger \hat{X}_i$ . We provide an algorithm for minimizing a generalization of the above cost functions where the squared Euclidean distances are replaced by other Bregman divergences.

### 3 Bregman Divergence

Bregman Divergences, special cases of which are squared loss and KL divergence are a generalized notion of distance that need not satisfy triangle inequality or symmetry. Similar to the squared loss, a Bregman projection can be defined and they obey a generalization of the Pythagoras theorem discussed below. Here we examine how they are related to Maximum likelihood of exponential family of distributions.

Let  $\phi : S \mapsto \mathbb{R}$  be a strictly convex function defined on a convex set  $S \subseteq \mathbb{R}^d$ , such that  $\phi$  is differentiable on  $\text{Int}(S)$ , the interior of  $S$  [10].

**DEFINITION 1. Bregman divergence**  $D_\phi : S \times \text{Int}(S) \mapsto [0, \infty)$  is defined as  $D_\phi(X, Y) = \phi(X) - \phi(Y) - (X - Y)^\dagger \nabla \phi(Y)$ , where  $\nabla \phi$  is gradient of  $\phi$ .

**LEMMA 1.**  $D_\phi(X, Y) \geq 0 \quad \forall X, Y \in \text{Int}(\text{Dom}(\phi))$  and  $D_\phi(X, Y) = 0$  iff  $X = Y$

**DEFINITION 2. The Bregman Projection** of a point  $Y \in \text{Int}(S)$  onto a closed convex set  $\Delta$  is defined as  $\text{Proj}_\Delta(Y) = \text{Argmin}_{X \in \Delta \cap \text{Int}(S)} D_\phi(X, Y)$ .

**THEOREM 1.** [7] Given a Bregman Divergence  $D_\phi : S \times \text{Int}(S) \mapsto [0, \infty)$ , a hyper plane  $\Delta \in \mathbb{R}^d$  with  $\Delta \cap \text{Int}(S) \neq \emptyset$ , and points  $Y \in \text{Int}(S)$  and  $X \in \Delta$

$$D_\phi(X, Y) = D_\phi(X, \text{Proj}_\Delta(Y)) + D_\phi(\text{Proj}_\Delta(Y), Y)$$

**COROLLARY 1.** Given the Bregman Divergence  $D_\phi$ , and a convex set  $\Gamma \subset \Delta \subset \text{Dom}(\phi)$

$$\text{Argmin}_{X \in \Gamma} D_\phi(X, Y) = \text{Proj}_\Gamma(\text{Proj}_\Delta(Y))$$

**DEFINITION 3.** An exponential family distribution is of the form  $p(Y|\theta) = e^{\{\theta Y - A(\theta)\}} h(Y)^2$  where  $Y \in \mathbb{R}$  constitutes the sufficient statistic and  $\theta \in S \subset \mathbb{R}$  is the canonical parameter.

**LEMMA 2.** The set  $S$  is convex and log cumulant  $A(\theta)$  is a strictly convex function.  $\mu = \mathbb{E}[Y|\theta] = \nabla A(\theta)$  and  $\sigma = \mathbb{E}[YY^\dagger|\theta] = \nabla^2 A(\theta)$ . Log likelihood  $l(\theta) = \log(p(Y|\theta))$  is strictly concave and  $\theta^* = \text{Argmin}_\theta -\log(p(Y|\theta))$  satisfies  $\nabla A(\theta^*) = Y$ .

Compare the log likelihood achieved at  $\theta$  to the best that could have been achieved for fixed  $Y$ . The difference is  $l(\theta^*) - l(\theta) = D_A(\theta, \theta^*)$  a function of  $\theta$  only, thus maximizing the likelihood is equivalent to minimizing the divergence  $D_A(\theta, \nabla^{-1} A(Y))$ .

**3.1 Canonical Generalized Linear Model** In canonical generalized linear models (GLM) [9],  $\theta_i$  is assumed to be a linear function of  $X_i$  of the form  $\theta_i = \beta^\dagger X_i$ , following which the maximum likelihood reduces to a Bregman Linear Regression problem:  $\beta^* = \text{Argmin}_\beta D_A(\beta^\dagger X, \theta^*)$ . Lemma 2 leads to  $\mathbb{E}[y_i|X_i] = \nabla A(\beta^\dagger X_i)$  as the predicted value of  $Y$  for a given  $X$ . The function  $\nabla A()$  is the inverse of what is known as the “Canonical Link Function”, in GLM literature.

The Iteratively Re-weighted Least Squares (IRLS) technique fits GLMs through Fisher scoring, which in the canonical case is the Newton-Raphson update  $\beta^{t+1} = \beta^t - [\nabla^2 l]^{-1} [\nabla l]$ . The log likelihood  $l = \sum l_j$  where  $l_j = Y_j \theta_j - A_j(\theta_j)$  and  $\theta_j = \beta^\dagger X_j$  and

$$\nabla l = [Y_j - \mu_j(\beta^t)] X_j; \quad \nabla^2 l = -X_j(\mathbf{h}) \Sigma_j(\beta^t) X_j(\mathbf{k})$$

### 4 Total Bregman Regression

Total Bregman Regression (TBR) is posed as a generalization of TLS. Here the output  $\hat{Y}_i$  is distributed according to some exponential family distribution with mean  $Y$ . For instance if  $Y_i$  takes values in two class labels 0 and 1, the choice could be a binomial distribution, leading to the standard logistic regression. For simplicity we assume that the the observed inputs  $\hat{X}_i$  have a Gaussian distribution about the mean at  $X_i$ , but any other exponential family can also be handled. Given  $\hat{Y}_i \sim e^{\{\beta^\dagger X_i \hat{Y}_i - A(\beta^\dagger X_i)\}}$ ,  $\hat{X}_i \sim e^{\{X_i^\dagger \hat{X}_i - \sum_j X_i^2(j)\}}$

<sup>2</sup>in the rest of the paper we will ignore the term  $h(Y)$  as it is independent of the parameters we wish to estimate

and  $\phi(X_i) = \sum_j X_i^2(j)$ <sup>3</sup> the TBR may be posed as

$$(4.1) \text{Min}_{\beta, X_i} \sum \left( D_A(\beta^\dagger X_i, \nabla^{-1} A(Y_i)) + \alpha D_\phi(X_i, \hat{X}_i) \right)$$

where  $\alpha$  is a positive weight that accounts for the dispersion parameter of the exponential distribution (such as variance of a Gaussian distribution) and weights the reconstruction error cost with respect to prediction cost.

We provide an iterative coordinate ascent algorithm following a random initialization of  $\beta$  in figure 1. TBR has similarities with its corresponding Bregman PCA problem which was solved by Collins, Dasgupta, et al [3]. Since both the sub-problems in the figure are convex in their arguments each of them can be globally maximized. One should note that this does not by itself ensure global optima as the cost is not convex in both the arguments together. Note the  $\beta$  step is standard IRLS as practiced in GLM, while  $X$  step requires a minor modification to IRLS.

**initialize**  $\beta^1$  and **set**  $t = 1$

**Repeat** till convergence

**X step:** **set**  $X_i^t$  to  $\text{Argmin}_{X_i} D_A(\beta^{t\dagger} X_i, \nabla^{-1} A(Y_i)) + \alpha D_\phi(X_i, \hat{X}_i)$  using IRLS with  $j = \{1, 2\}$ ,  $\mathbf{X}_1 = \beta^t$ ,  $\mathbf{X}_2 = I$ ,  $\mathbf{Y}_1 = \nabla^{-1}(Y_i)$  and  $\mathbf{Y}_2 = \hat{X}_i$ ,  $\mathbf{A}_1 = A$ ,  $\mathbf{A}_2 = \phi$

**$\beta$  step:** **set**  $\beta^{t+1} = \text{Argmin}_\beta D_A(\beta^\dagger X_i^t, \nabla^{-1}(Y_i))$  using IRLS with  $j = \{1, n\}$ ,  $\mathbf{X}_j = X_j$ ,  $\mathbf{Y}_j = \nabla^{-1}(Y_j)$  and  $\mathbf{A}_j = A$

**return**  $\beta^{t+1}, X_i$

Figure 1: Total Bregman Linear Regression Algorithm

**4.1 Transductive Prediction** The TBR output  $Y_i$  is generated from the de-noised input  $X_i$  which is not available for test data. We describe an Expectation Maximization (EM) based method of using TBR for prediction. The unlabeled “test” data are transductively de-noised and then used for prediction with the learnt regressors  $\beta$ . One important observation to make in this missing label setting is that EM method can be used in Bregman Linear regression with guarantees of global maxima of the likelihood function in absence of feature noise.

LEMMA 3. *The Bregman Linear Regression cost function is convex in  $\beta$  and  $Y$  together.*

<sup>3</sup>everywhere in the paper we use the notation  $\phi(X_i)$  as an abbreviation for  $\sum_j X_i^2(j)$

We treat the missing labels on test points as a hidden variable. For a complete log-likelihood of the the data  $\text{Log}P(X, Y|\beta)$  of which  $Y$  is hidden and  $X$  known random variables and  $\beta$  the parameters to be estimated, EM proposes an iterative update of  $\beta$  as a maximization of the Expected complete likelihood as follows [4]:  $\theta_{t+1} = \text{Argmax}_{\mathbb{E}_{Y|X, \theta_t}} \text{Log}P(X, Y)$ . The conditional expectation is to be computed over the distribution of the hidden variable conditioned on the available data (input  $X$ ). The update ensures a local maxima of the marginal loglikelihood  $\text{Log}(\sum_Y P(X, Y))$ . The conditional expectation is the output of the Bregman Regression function itself as a result of which no extra computation is necessary for the E-step.

**initialize**  $\beta^1$  and **set**  $t = 1$

**Repeat** till convergence

**E** estimate unknown  $\mathbb{E}Y_k^t$  conditioned on  $X_k^t$  and  $\beta^t$

**M** aximize by  $\beta^{t+1} = \text{TBR}(\beta^t, \mathbb{E}[\mathbf{Y}^t|\mathbf{X}^t], \mathbf{X}^t)$

**Return**  $\beta^{t+1}, \mathbf{X}^{t+1}, \mathbb{E}[\mathbf{Y}^{t+1}|\mathbf{X}^t]$

Figure 2: Transductive Prediction

## 5 Supervised and Transductive Bregman PCA

Here the supervised PCA is generalized to generative models belonging to an exponential family. The difference from the TBR model is that, now the input  $X$  is not only perturbed, but also constrained to lie on some unknown subspace. For  $\mathbf{U}\mathbf{U}^\dagger$  a projection matrix the Supervised Bregman PCA can be posed as

$$(5.2) \text{Min}_{\beta, \mathbf{U}} \sum \left( D_A(\beta^\dagger X_i, \nabla^{-1} A(Y_i)) + \alpha D_\phi(X_i, \hat{X}_i) \right) \text{ s.t. } X = \mathbf{U}\mathbf{U}^\dagger \hat{X}$$

Expressing  $\beta^\dagger \mathbf{U}\mathbf{U}^\dagger \hat{X}$  as a dot-product of the vectors  $\mathbf{U}^\dagger \beta = \gamma$  and  $\mathbf{U}^\dagger \hat{X} = V$  in the subspace spanned by columns of  $\mathbf{U}$ . the objective function can be posed as

$$\text{Min}_{\gamma, V_i, \mathbf{U}} \sum_i D_F \left( \begin{bmatrix} \gamma^\dagger \\ \mathbf{U} \end{bmatrix} V_i; g \left( \begin{bmatrix} \hat{Y}_i \\ \hat{X}_i \end{bmatrix} \right) \right)$$

Where  $F(Z_i) = A(\theta_i) + \phi(X_i)$  and  $G$  is the Legendre dual of  $F$  and  $g = \nabla G$ . Similar to the Total Bregman Regression case, a coordinate ascent algorithm based on alternate minimization has been posed in figure 3. The  $\gamma$  step is standard IRLS and does not require further elaboration. The  $\mathbf{U}$  step may be implemented as  $k$  least squares problem with different right hand sides as in  $\mathbf{V}^\dagger \mathbf{U}^\dagger = \mathbf{X}^\dagger$ , where  $k$  is the number of components sought. The  $V$  step consists of a multidimensional

```

initialize  $\gamma^1, \mathbf{U}^1, \mathbf{V}^1$  and set  $t = 1$ 
Repeat till convergence
   $\begin{pmatrix} \gamma^\dagger \\ \mathbf{U} \end{pmatrix}$  step: set :
    •  $\gamma^{t+1} = \text{Argmin}_\gamma \sum_i D_A(\gamma^\dagger V_i^t, \nabla^{-1} A(Y_i))$ 
      and
    •  $\mathbf{U}^{t+1} = \text{Argmin}_{\mathbf{U}} \sum_i D_\phi(X_i, \mathbf{U} V_i^t)$  using IRLS.
  V step:  $V_i^{t+1} = \sum_i \text{Argmin}_{V_i} D_A(\gamma^{t+1 \dagger} V_i, \nabla^{-1}(Y_i)) + \alpha D_\phi(X_i, \mathbf{U}^{t+1} V_i)$ 
return  $\mathbf{U}$ 

```

Figure 3: Supervised Bregman PCA

regression [8] problem where the same  $V_i^{t+1}$  is used to fit two regression problems.

The above steps in the algorithm can be simplified to an approximate closed form solution through the use of Pythagoras theorem for Bregman divergences and the using the solutions  $\beta^*$  and  $\hat{X}_i^*$  obtained for the TBR problem. TBR constitutes a relaxation of a constraint of the supervised Bregman PCA problem, where  $X$  is no longer constrained to be the projection  $\mathbf{U}\mathbf{U}^\dagger \hat{X}$ . The TBR solution  $\beta^*$  and  $\mathbf{X}^*$  is evaluated first, and then projected according to the Bregman divergence  $D_F()$  onto the space spanned by  $\mathbf{U}$ . First we show that no local optimality is lost as a result of this projection.

LEMMA 4. For  $X_i$  in the range of matrix  $\mathbf{U}\mathbf{U}^\dagger$

$$\begin{aligned} & \text{Argmin}_{X_i} \sum_i D_F \left( \begin{bmatrix} \beta^\dagger \\ \mathbf{I} \end{bmatrix} X_i; g \left( \begin{bmatrix} \hat{Y}_i \\ \hat{X}_i \end{bmatrix} \right) \right) \\ & = \text{Argmin}_{X_i} \sum_i D_F \left( \begin{bmatrix} \beta^\dagger \\ \mathbf{I} \end{bmatrix} X_i; \begin{bmatrix} \beta^{*\dagger} \\ I \end{bmatrix} X_i^* \right) \end{aligned}$$

*Proof.* follows from Corrolary 1

Having removed the link function  $g$  out of the way we approximation the cost with second order Taylor expansion about  $\hat{X}^*$ . This approximation is somewhat countered by the fact that the cost function also penalizes the distance of  $X$  from  $\hat{X}$ , forcing  $X$  to lie sufficiently near, so that the approximation error is small. Using  $D_A(\theta, \theta^*) \simeq (\theta - \theta^*)^\dagger \nabla^2 A(\theta) (\theta - \theta^*)$  we have the approximate cost function as

$$X^{*\dagger} (I - \mathbf{U}\mathbf{U}^\dagger) \left[ \beta^{*\dagger} \nabla^2 A(\beta^{*\dagger} X^*) \beta^{*\dagger} + 2\alpha I \right] (I - \mathbf{U}\mathbf{U}^\dagger) X^*$$

which is optimized by the eigenvectors of

$X^* \left[ \beta^{*\dagger} \nabla^2 A(\beta^{*\dagger} X^*) \beta^{*\dagger} + 2\alpha I \right] X^{*\dagger}$  thus providing us with a closed form solution.

It is instructive to compare this with the standard PCA solution, which is the principal eigenvector of  $\hat{\mathbf{X}}\hat{\mathbf{X}}^\dagger$ , for the case  $\nabla^2 A(\beta^{*\dagger} X^*) = \mathbf{I}$  we can see that the supervised principal components are rotated towards the regression coefficients  $\beta^*$  which is the direction in which the ‘‘classes’’ are ‘‘separated’’. The quality of the approximation depends on the weight  $\alpha$ , higher its value, better is the approximation.

## 6 Experimental Results and Comments

The wide applicability of the formulation especially to the case beyond Least Squares regression is shown on two tasks that LS regression does not handle naturally, (i) classification and (ii) integer regression. Supervised Bregman PCA (SBPCA) is compared both with Fisher Linear Discriminant (FLD) when applicable (i.e. for classification tasks) as well as standard PCA. For PCA, the predictor was learnt using Bregman regression on the dimensionality reduced data. For FLD, we used its natural Gaussian generative model for classification. We only report 10 fold cross-validation case.

Experimental results show in figure 5, prediction accuracy on the test set overlaid with 1 sigma error bars. SBPPCA quickly achieves higher accuracy over PCA, it also outperforms FLD by capitalizing, what we believe, on its discriminative modeling. FLD does beat SBPCA at low dimensionality and the point of intersection depends on how we balance prediction accuracy with reconstruction error.

The data sets used are summarized in the following table. To test our algorithm on the pendigits data set, we tested our algorithm on 3 comparatively difficult binary problems generated from the original data set. The Pima Indian diabetes dataset is not linearly separable therefore not amenable to linear methods unless combined with kernels, SVM decision trees achieve classification accuracy of 77%, similarly SVM decision trees achieved an accuracy of 96% on Wisconsin Breast cancer data set. The objective in abalone data set is to predict the age in years of abalone shells. UCI repository does not report any past study of integer regression but report one study where the ages were discretized into 3 classes and obtained a classification accuracy of 32% which is close to random.

Figure 4 shows visual evidence of how TBR distorts the input vectors to provide better classification accuracy, it was observed that after distortion, within class variance decreases

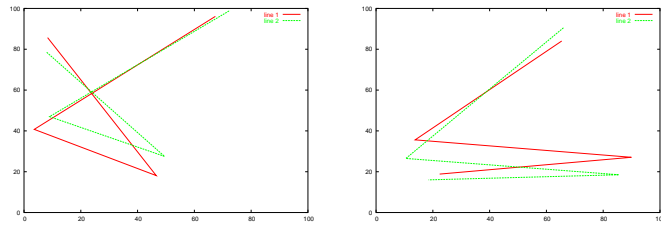


Figure 4: Examples of transductive denoising observed in Pen56 data set.

Name	Dim	Class	Points
Pendigits 5 & 6	16	2	2110
Pendigits 2 & 7	16	2	2285
Pendigits 4 & 9	16	2	2199
Pima Indian Diabetes	8	2	767
Wisconsin Breast Cancer	34	2	196
Abalone	8	Int	4177

## 7 Conclusion

In this study we generalized the Total Least Squares to (i) Bregman divergences and (ii) a transductive setting. The former widened the applicability of the regression algorithm to other tasks such as classification, integer regression etc. The latter allowed denoising of unlabeled data so that they can predict the output better. Next we formulated a dimensionality reduction technique, that directly optimizes the goal of better prediction using the dimension reduced variables. It performed better than PCA or Fisher's projection provided the classes were linearly separable and the parametric discriminative assumptions were appropriate.

## References

- [1] I. Dhillon, A. Banerjee, S. Merugu and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research (JMLR)* (2005), 2005.
- [2] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.
- [3] Michael Collins, S. Dasgupta, and Robert E. Schapire. A generalization of principal components analysis to the exponential family. In *NIPS*, pages 617–624, 2001.
- [4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 34:1–38, 1977.
- [5] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press and North Oxford Academic, Baltimore, MD, USA and Oxford, England, 1993.
- [6] T. Hastie and R. Tibshirani. Discriminant analysis by

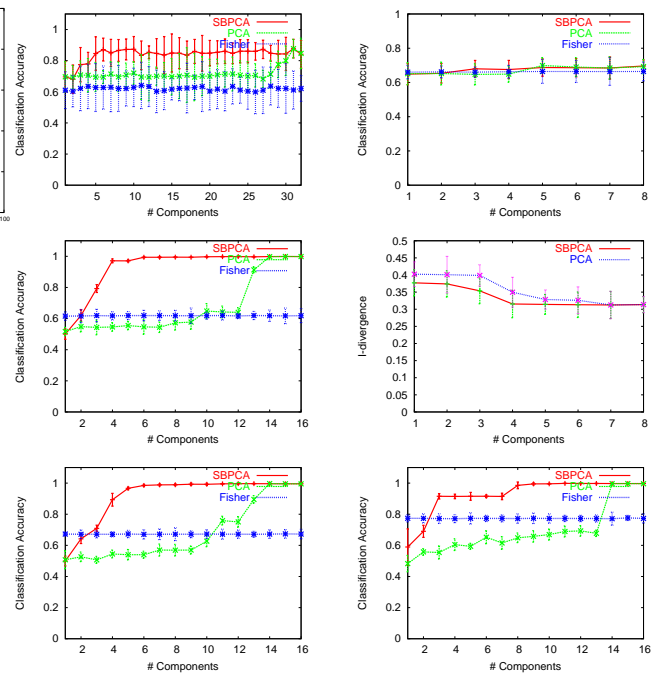


Figure 5: Top: Dimensionality Reduction using SBPCA, FLD and PCA with increasing number of components on Breast Cancer and Diabetes dataset. SBPCA yields better classification accuracy using fewer number of components. On diabetes data set which is not linearly separable, all the algorithms perform comparably. Mid: Corresponding learning curves for Pen56 and Abalone. For Abalone, the Y axis plots both I-divergence (*Bregman divergence corresponding to Poisson distribution*) as well as absolute deviation between the actual and predicted integer age of the shells. Bottom: Learning curves for pen27 and pen49

- Gaussian mixtures. *Journal of the Royal Statistical Society series B*, 58:158–176, 1996.
- [7] Mark Herbster and Manfred K. Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1:281–309, 2001.
- [8] Jyrki Kivinen and Manfred K. Warmuth. Relative loss bounds for multidimensional regression problems. *Machine Learning*, 45(3):301–329, 2001.
- [9] J.A. Nelder P. McCullagh. *Generalized Linear Models*. Chapman and Hall, New York, 1983.
- [10] R. Tyrrell Rockafellar. *Convex analysis*. Princeton, N.J., Princeton University Press, 1970.
- [11] Alon Orlitsky Sajama. Supervised dimensionality reduction using mixture models. In *ICML*, 2005.
- [12] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.