Feature Selection using Partial Least Squares Regression and Optimal Experiment Design

Varun K. Nagaraja
Dept. of Computer Science,
University of Maryland, College Park
varun@umiacs.umd.edu

Wael Abd-Almageed
Information Sciences Institute,
University of Southern California
wamageed@isi.edu

***Abstract*—We propose a supervised feature selection technique called the Optimal Loadings, that is based on applying the theory of Optimal Experiment Design (OED) to Partial Least Squares (PLS) regression. We apply the OED criterions to PLS with the goal of selecting an optimal feature subset that minimizes the variance of the regression model and hence minimize its prediction error. We show that the variance of the PLS model can be minimized by employing the OED criterions on the loadings covariance matrix obtained from PLS. We also provide an intuitive viewpoint to the technique by deriving the A-optimality version of the Optimal Loadings criterion using the properties of maximum relevance and minimum redundancy for PLS models. In our experiments we use the D-optimality version of the criterion which maximizes the determinant of the loadings covariance matrix. To overcome the computational challenges in this criterion, we provide an approximate D-optimality criterion along with the theoretical justification.**

I. Introduction

Datasets with a large number of features are prevalent in many fields like Computer Vision, Bioinformatics and Chemometrics. These large datasets pose analytical and computational challenges, and the problem is even worse for high dimensional cases where the number of features is much greater than the number of samples. A feature selection process reduces the dimensionality of the data by identifying a subset of the original features that captures the maximum amount of information from the data. The advantages of feature selection are improving the generalization capability of models, reduce computation time and provide a better understanding of the interaction among features [1].

Among supervised feature selection techniques, ranking by regression coefficients is one of the simplest ways to select features. Partial Least Squares (PLS) [2], [3] is a widely used regression technique for high dimensional datasets. It is extensively used for wavelength selection in Chemometrics and gene selection in Computational Biology [4], [5] as they typically present with high dimensional datasets. The features are usually selected by ranking them according to the value of their PLS regression coefficients or other relevance measures. The caveat of this procedure is that it doesn't jointly look at the features and is susceptible to selecting redundant features. Similar to ℓ_1 and ℓ_2 norm penalized regression techniques, penalized techniques for PLS [6], [7] are one of the approaches to perform feature selection with PLS. The penalized regression techniques enforce sparsity in the regression coefficients along with the minimization of model variance.

The other approach to minimizing the variance of the regression model is to apply the theory of Optimal Experiment Design (OED) [8] and its optimality criterions to PLS regression. The three most commonly used optimality criterions are A-optimality, D-optimality and E-optimality which respectively minimize the trace, determinant and maximum eigenvalue of the covariance matrix of regression coefficients. Optimal Experiment Design has been used for sample selection problems like sensor selection and Active Learning [9]. The optimality criterions are not specific to sample selection and can also be used to measure the optimality of models with different sets of features. Hence we use these criterions with PLS to develop a supervised feature selection technique. We show that an optimal feature subset can be selected by applying these criterions to the loadings covariance matrix obtained from PLS.

We first decompose the prediction error of PLS regression into its bias, variance and noise components. We then apply the OED criterions to the covariance matrix of regression coefficients to derive the A-optimality and D-optimality versions of the Optimal Loadings criterion. We also show that the A-Optimal Loadings criterion can be obtained by explicitly incorporating the property of maximum relevance as maximizing energy content in the loadings matrix. The minimum redundancy property is incorporated as minimizing the condition number of loadings matrix. However, solving the Optimal Loadings criterions is computationally challenging as it is dependent on different PLS models for evaluating different feature subsets. Hence we propose an approximate D-Optimal Loadings criterion that is based on a single loadings covariance matrix obtained with the entire set of features. We also obtain a mathematical relationship between the approximate and the original D-Optimal Loadings criterion and use it to qualitatively justify the approximation.

The advantage of the Optimal Loadings criterions is that the features are evaluated as subsets rather than individual features and hence can simultaneously measure redundancy along with relevance of features. This advantage is clearly evident in our experiments when the number of selected features is small. In our experiments we implement the D-Optimal Loadings criterion that maximizes the determinant of the loadings covariance matrix. Experiments on four datasets indicate that the D-Optimal Loadings criterion performs consistently better than the standard feature selection techniques, in terms of classification accuracies obtained with feature subsets.

II. Related Work

Feature selection techniques can be classified [1] into individual feature ranking methods and feature subset evaluation methods. The individual feature ranking methods use relevance measures to sort the features in a rank order. Fisher Score [10] and ReliefF [11] are two techniques that belong to the ranking methods. Features can also be ranked based on regression coefficients and other informative vectors like Variable Influence on Projection (VIP) [12]. Although these methods have a computational advantage, they fail in the presence of redundant features as the minimum redundancy property needs to be measured by jointly looking at the features. A popular technique that incorporates both the relevance and redundancy properties is the minimum redundancy and maximum relevance (mRMR) framework [13], [14]. It involves an objective function that is based on Information Theoretic measures and uses incremental search techniques to find the feature subsets. The computational challenge in the original mRMR framework is the estimation of mutual information when the number of samples is small and also when the data is continuous. However, a kernel based dependency measure like the Hilbert Schmidt Independence Criterion (HSIC) can be used instead of the mutual information measure. The HSIC has been used as a measure of feature dependence by L.Song et al. [15].

In the presence of high dimensionality, ordinary least squares regression fails due to the singularity of the feature covariance matrix. Hence regularized linear regression, usually with ℓ_1 and/or ℓ_2 penalization [16], [17], is employed to obtain a biased model with smaller variance. Partial Least Squares (PLS) regression [2], [3] is a commonly used technique for handling high dimensional datasets. It provides two viewpoints to the modeling process - as a regression technique and as a feature extraction technique. While it can extract information in a latent space of few dimensions, the sparsity of the features needs to be explicitly incorporated into the PLS formulation for feature selection. In the Sparse PLS of K-A Lê Cao et al. [6], ℓ_1 penalization is applied to the loading vectors in the PLS-SVD formulation to integrate feature selection into the modeling process. The Sparse PLS of H.Chun and S.Keles [7] uses both the ℓ_1 and ℓ_2 penalization like that of Elastic Nets in the PLS formulation.

In Ordinary Least Squares regression, under uniform noise assumption, the covariance matrix of the regression coefficients is independent of the response variable. This property is used to apply the Optimal Experiment Design [8] to unsupervised feature selection. The Laplacian Score technique [18] is a ranking based algorithm for unsupervised feature selection that has been extended [19] with OED and shown to perform better than the original ranking based algorithm. While both the penalization and the OED approaches have been studied for ordinary least squares regression, only the penalization methods have been tried with PLS. Our work explores the application of the OED criterions to PLS regression.

III. Preliminaries

A. Partial Least Squares

Partial Least Squares is a simultaneous feature extraction and regression technique, well suited for high dimensional problems where the number of samples is much lesser than the number of features ($n \ll p$). The linear PLS model can be expressed as

$$X = TP^\top + X_{res} \tag{1}$$

$$Y = UQ^\top + Y_{res} \tag{2}$$

where $X_{n\times p}$ is the feature matrix, $Y_{n\times q}$ is the matrix of response variables or class labels, $T_{n\times d}$ is called the X-scores, $P_{p\times d}$ is X-loadings, $U_{n\times d}$ is Y-scores, $Q_{q\times d}$ is Y-loadings, X_{res} and Y_{res} are the residuals. The data in X and Y are assumed to be mean-centered. X-scores and Y-scores are the projections of n samples onto a d-dimensional orthogonal subspace. The X-scores are obtained by a linear combination of the variables in X with the weights W^* as shown in Eqn. (3).

$$T = XW^* \tag{3}$$

The inner relation between X-scores and Y-scores is a linear regression model [2] and hence X-scores are called predictors of Y-scores. If B is the regression coefficient for the inner relation between the scores, we can write

$$U = TB \tag{4}$$

Substituting the above Eqn. (4) in Eqn. (2) we get

$$Y = TBQ^\top + Y_{res} \tag{5}$$

$$= T\tilde{B} + Y_{res} \tag{6}$$

where $\tilde{B} = BQ^\top$. The least squares estimate of $\tilde{B}$ is then given by

$$\hat{B} = (T^\top T)^{-1}T^\top Y \tag{7}$$

Hence PLS can be expressed in a linear regression form as,

$$\hat{Y} = T\hat{B} = T(T^\top T)^{-1}T^\top Y \tag{8}$$

For a detailed explanation of the PLS technique, we guide the readers to refer [2], [3].

The two most popular algorithms to obtain the PLS model are NIPALS [3] and SIMPLS [20]. SIMPLS provides weights W^* which can be combined directly with X where as NIPALS provides weights W that act on the residuals Z_a obtained by deflating X at every component a. The relationship between the two is given by [3],

$$W^* = W(P^\top W)^{-1} \tag{9}$$

Here we consider the case of a single response variable $Y_{n\times 1}$ and use the equations from the NIPALS algorithm to obtain the PLS model. However we consider a small variation, where we normalize the scores instead of the loadings. At every iteration for the component a, we have

$$w_a = \frac{Z_a^\top Y}{\sqrt{Y^\top Z_a Z_a^\top Y}} \tag{10}$$

$$t_a = \frac{Z_a w_a}{\sqrt{w_a^\top Z_a^\top Z_a w_a}} \tag{11}$$

$$p_a = Z_a^\top t_a \tag{12}$$

$$Z_{a+1} = Z_a - t_a p_a^\top \tag{13}$$

where $Z_1 = X$. The weights and scores form an orthonormal set i.e. $w_i^\top w_j = 0$ and $t_i^\top t_j = 0$ for $i \neq j$.

B. Notation

Let π denote a subset of feature indices from the set $\{1, 2, 3, \ldots, p\}$ containing exactly k elements. The feature subset matrix X_π is expressed as

$$X_\pi = X_{(n\times p)}\Pi_{(p\times k)} \tag{14}$$

where Π is a column selection matrix that selects k out of p features. Each of the k columns of Π contains a single entry of one at a row indexed by an element in π and zeros elsewhere. Any parameter of a model built with a subset of features is represented by a subscript π.

IV. Optimal Loadings Technique

A. Optimal Experiment Design for PLS

Consider a linear regression model

$$Y = X\beta + \epsilon \tag{15}$$

where $Y_{n\times 1}$ is the response vector, $X_{n\times p}$ is the feature matrix, $\beta_{p\times 1}$ is the regression coefficient vector and $\epsilon_{n\times 1}$ is the noise vector with mean zero and covariance $\sigma^2 I_n$. The noise for different observations are assumed to be independent of each other.

The Partial Least Squares estimate of the regression coefficients can be obtained by substituting for T_π from Eqn. (3) in Eqn. (8).

$$\hat{\beta_\pi} = \Pi W_\pi^* (T_\pi^\top T_\pi)^{-1} T_\pi^\top Y = \Pi W_\pi^* T_\pi^\top Y \tag{16}$$

By substituting for Y from Eqn. (15) in the above Eqn. (16), we find that the mean of the PLS estimate is given by

$$E[\hat{\beta_\pi}] = \Pi W_\pi^* T_\pi^\top X\beta + \Pi W_\pi^* T_\pi^\top E[\epsilon] \tag{17}$$
$$= \Pi W_\pi^* T_\pi^\top X\beta \tag{18}$$

where in Eqn. (17) we have assumed that $\Pi W_\pi^* T_\pi^\top$ and ϵ are negligibly correlated. This is possible when the Signal to Noise Ratio is high and hence the deviation in the PLS model with respect to noise is negligible. The covariance of $\hat{\beta_\pi}$ is given by

$$cov(\hat{\beta_\pi}) = E\left[(\hat{\beta_\pi} - E[\hat{\beta_\pi}])(\hat{\beta_\pi} - E[\hat{\beta_\pi}])^\top\right] \tag{19}$$
$$= E[\hat{\beta_\pi}\hat{\beta_\pi^\top}] - E[\hat{\beta_\pi}]E[\hat{\beta_\pi^\top}] \tag{20}$$
$$= \sigma^2 \Pi W_\pi^* W_\pi^{*\top} \Pi^\top \tag{21}$$

For a new sample (x, y) such that $y = x^\top\beta + e$ and $\hat{y} = x^\top\hat{\beta}_\pi$, the mean squared prediction error of PLS can be decomposed into its bias, variance and noise components.

$$E[(y - \hat{y})^2] \tag{22}$$
$$= x^\top E[(\beta - \hat{\beta}_\pi)(\beta - \hat{\beta}_\pi)^\top]x + \sigma^2 \tag{23}$$
$$= Bias^2 + x^\top\left(\sigma^2 \Pi W_\pi^* W_\pi^{*\top} \Pi^\top\right)x + \sigma^2 \tag{24}$$

where

$$Bias^2 = x^\top(I_p - \Pi W_\pi^* T_\pi^\top X)\beta\beta^\top(I_p - X^\top T_\pi W_\pi^{*\top}\Pi^\top)x \tag{25}$$

Since the squared prediction error is directly proportional to $cov(\hat{\beta_\pi})$, the prediction error can be minimized by minimizing the covariance of PLS regression coefficients. Also, in high dimensional datasets, reducing the model variance helps avoid overfitting to the data. The theory of Optimal Experiment Design proposes to minimize this covariance by optimizing the eigenvalues of $\Pi W_\pi^* W_\pi^{*\top}\Pi^\top$ through various criterions.

Lemma 1. *The matrices $W_\pi^* W_\pi^{*\top}$ and $(P_\pi P_\pi^\top)^\dagger$ have the same non-zero eigenvalues, where $\dagger$ represents the Moore-Penrose inverse.*

Proof: By substituting for W_π^* from Eqn. (9), we get

$$\text{eigval}(W_\pi^* W_\pi^{*\top}) \tag{26}$$
$$= \text{eigval}\left[W_\pi (P_\pi^\top W_\pi)^{-1}(P_\pi^\top W_\pi)^{-1})^\top W_\pi^\top\right] \tag{27}$$
$$= \text{eigval}\left[W_\pi W_\pi^\top (P_\pi P_\pi^\top)^\dagger W_\pi W_\pi^\top\right] \tag{28}$$
$$= \text{eigval}\left[(P_\pi P_\pi^\top)^\dagger\right] \tag{29}$$

where eigval() refers to the eigenvalues of a matrix. Eqn. (28) can be regarded as a similarity transformation since W_π is orthonormal. The rank of these matrices is equal to the number of latent components (d) extracted. ■

Using the above Lemma 1 and applying the A-optimality criterion to the covariance matrix of PLS coefficients in Eqn. (21) we get,

$$\arg\min_{\Pi}\ \text{trace}\left[\Pi(P_\pi P_\pi^\top)^\dagger\Pi^\top\right] \tag{30}$$

We can drop the pre and post multiplication by Π as it is only padding zeros to change the size of the matrix, $(P_\pi P_\pi^\top)^\dagger$, from $k \times k$ to $p \times p$. For a fixed number of selected features, k, the A-optimal criterion can be rewritten as

Definition 1 (A-Optimal Loadings criterion)**.** *The A-optimality version of Optimal Loadings criterion is given by*

$$\arg\min_{\Pi}\ \text{trace}\left[(P_\pi P_\pi^\top)^\dagger\right] \tag{31}$$

We could also apply the D-optimality or E-optimality criterion which minimize the determinant or the maximum eigenvalue respectively, instead of the trace in Eqn. (31). Among these optimality criterions, the D-optimality criterion is the most popular due to availability of off-the-shelf algorithms in convex optimization toolboxes and row exchange algorithms. It also simplifies the determinant minimization of an inverse to maximizing the determinant of the matrix itself. The D-optimality version of the criterion (31) is given by

$$\arg\min_{\Pi}\ \det{}^\dagger\left[(P_\pi P_\pi^\top)^\dagger\right] \tag{32}$$

which is equivalent to

Definition 2 (D-Optimal Loadings criterion)**.** *The D-optimality version of Optimal Loadings criterion is given by*

$$\arg\max_{\Pi}\ \det{}^\dagger\left(P_\pi P_\pi^\top\right) \tag{33}$$

where $\det{}^\dagger()$ *represents pseudo-determinant which is a product of non-zero eigenvalues of the matrix.*

The actual determinant is substituted by a pseudo determinant as the criterion involves a rank deficient matrix.

B. PLS models with Maximum Relevance and Minimum Redundancy

The A-Optimal Loadings criterion (31) can also be obtained by applying the requirements of maximum relevance and minimum redundancy for feature subsets. The following derivation provides an intuitive viewpoint to the same criterion that is obtained from the theory of Optimal Experiment Design.

The reconstruction error in a feature extraction technique measures the difference between the original energy content in all the features and the amount captured by the latent components. While it is our goal to obtain features that best explain a response variable, the structure in data should also be preserved. By substituting for p_a from Eqn. (12) in Eqn. (13), we get

$$Z_{a+1} = [I - t_a t_a^\top] Z_a = [I - \sum_{i=1}^{a} t_i t_i^\top] X \tag{34}$$

The reconstruction error can also be viewed as the residuals that cannot be explained by the PLS model. Hence we can use Eqn. (34) to express the error in a form similar to that of reconstruction error for PCA.

$$error^2 = ||X_{res}||_2^2 = \left|\left|X - TT^\top X\right|\right|_2^2 \tag{35}$$
$$= \text{trace}\left[X^\top X - X^\top TT^\top X\right] \tag{36}$$
$$= \text{trace}\left[X^\top X\right] - \text{trace}\left[PP^\top\right] \tag{37}$$

where Eqn. (37) is obtained by substituting for X from Eqn. (1) in the second term and making use of the fact that the scores T are orthogonal to the residuals X_{res}. The reconstruction error reduces with increase in the number of components extracted. But for a fixed number of components d, the error is minimum when the $\text{trace}\left[PP^\top\right]$ is maximum. Therefore we start by defining the feature selection criterion as

$$\arg\max_{\Pi} \ \text{trace}\left[P_\pi P_\pi^\top\right] \tag{38}$$

It should be noted that the reconstruction error in itself is not considered in criterion (38). This criterion tries to select the feature subset that contains the maximum energy content (measured by Frobenius norm) in the PLS model after feature selection.

The criterion (38) is also directly proportional to covariance between the features X and the response variable Y. This can be seen by substituting for p_π from Eqn. (12) in criterion (38) and then expanding up to w_π in Eqn. (10). We get

$$\text{trace}[P_\pi P_\pi^\top] = \sum_{a=1}^{d} \frac{Y^\top \left(Z_a Z_a^\top\right)^3 Y}{Y^\top \left(Z_a Z_a^\top\right)^2 Y} \tag{39}$$

Since PLS extracts components such that the covariance between features and response variable and the covariance between features itself are simultaneously maximized, the criterion (38) simultaneously satisfies the relevance property towards the response variable and the latent information in features.

However, the trace criterion (38) does not measure the redundancy property and hence we incorporate condition number of $P_\pi^\top$ to measure the linear dependence of columns/features. Since we want to minimize the condition number, the criterion (38) can be rewritten as

$$\arg\max_{\Pi} \left(\frac{\text{trace}\left[P_\pi P_\pi^\top\right]}{\left(\kappa\left(P_\pi^\top\right)\right)^2} \right) \tag{40}$$

The condition number in Frobenius norm is defined as

$$\left(\kappa\left(P_\pi^\top\right)\right)^2 = \text{trace}\left[P_\pi P_\pi^\top\right] . \text{trace}\left[(P_\pi P_\pi^\top)^\dagger\right] \tag{41}$$

We now substitute for κ in criterion (40) to obtain

$$\arg\min_{\Pi} \ \text{trace}\left[(P_\pi P_\pi^\top)^\dagger\right] \tag{42}$$

This is the same A-Optimal Loadings criterion (31) obtained earlier by applying the Optimal Experiment Design to Partial Least Squares regression.

C. Approximation for the D-Optimal Loadings criterion

In our experiments we choose to implement the D-optimality version of Optimal Loadings criterion as it simplifies the minimization of determinant of inverse matrix to the maximization of determinant itself. The availability of off-the-shelf algorithms for determinant maximization is another advantage of using the D-optimality criterion.

The loadings in criterion (33) is dependent on π and is infeasible to construct a PLS model every time a subset of features is to be evaluated. This would defeat the purpose of a feature selection technique. Hence we try to express the criterion in terms of loadings obtained with all features. From Eqn. (1), we have

$$X_\pi^\top X_\pi = P_\pi P_\pi^\top + X_{res\,\pi}^\top X_{res\,\pi} \tag{43}$$
$$\Pi^\top X^\top X \Pi = \Pi^\top P P^\top \Pi + \Pi^\top X_{res}^\top X_{res} \Pi \tag{44}$$

The right hand terms of the above Eqns. (43) and (44) can be equated (Eqn. (14)) to obtain

$$P_\pi P_\pi^\top = \Pi^\top P P^\top \Pi + \Delta_\pi \tag{45}$$

where Δ_π is a symmetric matrix given by

$$\Delta_\pi = \left[\Pi^\top X_{res}^\top X_{res} \Pi - X_{res\,\pi}^\top X_{res\,\pi}\right] \tag{46}$$

Since we use the D-optimality criterion for feature selection, we discuss the relationship between the determinants of $P_\pi P_\pi^\top$ and $\Pi^\top P P^\top \Pi$. The singularity of these matrices presents difficulties in quantifying their behavior. Therefore we obtain the relationship between the determinants of regularized matrices $(P_\pi P_\pi^\top + I)$ and $(\Pi^\top P P^\top \Pi + I)$.

Theorem 1. *The relationship between the determinants of $(\Pi^\top P P^\top \Pi + I)$ and $(P_\pi P_\pi^\top + I)$ is given by,*

$$\det(P_\pi P_\pi^\top + I) = \det(M + \Lambda M \Sigma^{-1}) \ \det(\Pi^\top P P^\top \Pi + I) \tag{47}$$

where M is a unitary matrix, Λ is a diagonal matrix of real eigenvalues of Δ_π and Σ is a diagonal matrix of positive eigenvalues of $(\Pi^\top P P^\top \Pi + I)$.

Proof: Let the two symmetric, positive semi-definite matrices $P_\pi P_\pi^\top$ and $\Pi^\top P P^\top \Pi$, each be of rank d and size $k \times k$, with the relationship between them as

$$P_\pi P_\pi^\top = \Pi^\top P P^\top \Pi + \Delta_\pi \tag{48}$$

where Δ_π is a symmetric matrix given by

$$\Delta_\pi = \left[\Pi^\top X_{res}^\top X_{res}\Pi - X_{res\,\pi}^\top X_{res\,\pi}\right] \tag{49}$$

We make use of Sherman-Morrison-Woodbury formula [21] for expressing the determinant of sum of matrices.

$$\det(P_\pi P_\pi^\top + I) = \det(\Pi^\top PP^\top\Pi + I + \Delta_\pi) \tag{50}$$
$$= \det(\Pi^\top PP^\top\Pi + I + U\Lambda U^\top) \tag{51}$$
$$= \det(I + \Lambda U^\top(\Pi^\top PP^\top\Pi + I)^{-1}U)\ \det(\Pi^\top PP^\top\Pi + I) \tag{52}$$
$$= \det(I + \Lambda U^\top V\Sigma^{-1}V^\top U)\ \det(\Pi^\top PP^\top\Pi + I) \tag{53}$$
$$= \det(I + \Lambda M\Sigma^{-1}M^\top)\ \det(\Pi^\top PP^\top\Pi + I) \tag{54}$$
$$= \det(M + \Lambda M\Sigma^{-1})\ \det(\Pi^\top PP^\top\Pi + I) \tag{55}$$

where we have applied Eigen-decomposition on Δ_π and $(\Pi^\top PP^\top\Pi + I)$. Σ and Λ are diagonal matrices containing non-negative eigenvalues (σ) of $(\Pi^\top PP^\top\Pi + I)$ and real eigenvalues (λ) of Δ_π, respectively. M is a unitary matrix obtained as a product of two other unitary matrices U and V. ■

The two determinants are highly correlated when the condition number of $(M + \Lambda M\Sigma^{-1})$ is small. The condition number of a matrix measures the asymptotic worst case of the amount of perturbation that can be produced by the matrix when multiplied with other matrices. The eigenvalues in Σ and Λ are indicators of the energy content in structured data and noise respectively, where noise is any structure that cannot be explained by the first d components of the PLS model. The theoretical and empirical observations (found in the supplementary material) suggest that the condition number is small when the variance in noise is low and levels of noise are far away from that of structure in data. Therefore under the assumption of high Signal to Noise Ratio, we can ignore Δ_π and substitute for $P_\pi P_\pi^\top$ from Eqn. (45) in criterion (33). The approximate feature selection criterion is given by,

$$\arg\max_{\Pi}\ \det{}^\dagger(\Pi^\top PP^\top\Pi) \tag{56}$$

The number of components in $\Pi^\top PP^\top\Pi$ and $P_\pi P_\pi^\top$ must be equal to compare the information between the two matrices. The number of components in PLS regression determines the bias and variance of the model. It is usually chosen such that the cross-validation error of PLS regression is minimum.

The experiments and discussion in the following sections use the D-optimality criterion for feature selection. D-optimal designs are usually generated by employing row exchange algorithms [22], [23]. These algorithms add or delete rows, starting from a non-singular set, in order to increase the determinant. The algorithm iterates until the increment in determinant becomes lesser than some fixed threshold or the number of iterations reach a maximum value. However, it is not guaranteed that the iterations will converge to the global maximum value. One of the first exchange algorithms was developed by V.V.Fedorov and several modifications have been proposed to improve the computational performance [23]. The traditional D-optimal experiment design differs from the feature selection problem, as it allows duplicate samples. Hence the standard exchange algorithms need to be tweaked to avoid duplicates for feature selection.

Since the D-optimality criterion involves maximization of the determinant, it can also be treated as a convex optimization problem [24]. The integer constraints $\pi_i \in \{0, 1\}$ need to be relaxed to $\pi_i \in [0, 1]$.

$$\text{minimize} \quad -\log\det\left[\sum_{i=1}^{p}\pi_i P_i P_i^\top\right] \tag{57}$$

$$\text{subject to} \quad \sum_{i=1}^{p}\pi_i = k \tag{58}$$

$$0 \le \pi_i \le 1,\ i = 1, \dots p \tag{59}$$

It can be seen that the solution to the original problem in Criterion (56) is a feasible solution to the above relaxed problem. Usually we obtain a discrete solution by considering the k largest values of π_i, which can lead to a sub-optimal solution to the original problem. The $\log\det$ criterion is an objective function available with popular SDP solvers [25]. One of the disadvantages of the convex optimization methods is that they store the entire convex hull of features, which is difficult to handle for large loadings matrix due to memory restrictions.

V. Analysis of the relationship between $P_\pi P_\pi^\top$ and $\Pi^\top PP^\top\Pi$

We can obtain an upper bound for the relationship (47) by finding the largest singular value of $\det(M + \Lambda M\Sigma^{-1})$. The spectral norm measures the largest singular value of a matrix. Using few of the properties of norms we can write

$$||M + \Lambda M\Sigma^{-1}||_2 \le ||M||_2 + ||\Lambda||_2||M||_2||\Sigma^{-1}||_2 \tag{60}$$
$$= 1 + \frac{\lambda_{max}}{\sigma_{min}} \tag{61}$$

where $\lambda_{max} = \max_i|\lambda_i|$ and $\sigma_{min} = \min_i\sigma_i$. Therefore the upper bound for relationship (47) is given by,

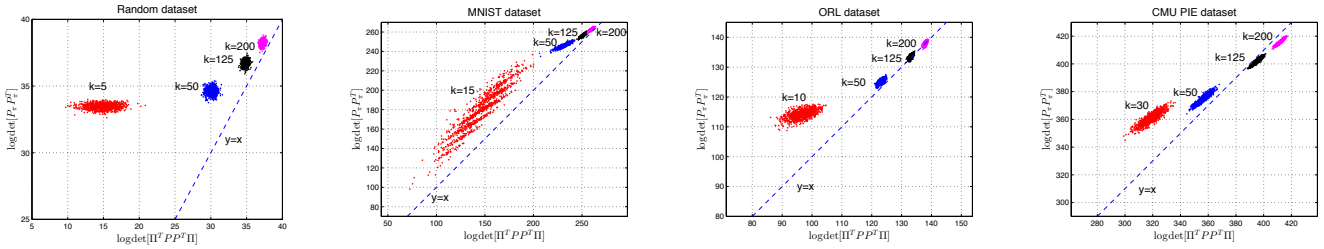
$$\det(P_\pi P_\pi^\top + I) \le \left(1 + \frac{\lambda_{max}}{\sigma_{min}}\right)^k \det(\Pi^\top PP^\top\Pi + I) \tag{62}$$

To determine the lower bound, we will need to find the smallest singular value of $\det(M + \Lambda M\Sigma^{-1})$. However the only safe bound that can be obtained is that the smallest singular value is greater than zero as the determinants on both sides of the relationship need to be positive. Nevertheless, a qualitative discussion can be provided by estimating the smallest singular value by a lower bound for the norms of the columns. We will first express the matrix $M + \Lambda M\Sigma^{-1}$ as,

$$M + \Lambda M\Sigma^{-1} = \tag{63}$$
$$\begin{pmatrix} (1+\frac{\lambda_1}{\sigma_1})m_{11} & (1+\frac{\lambda_1}{\sigma_2})m_{12} & \dots & (1+\frac{\lambda_1}{\sigma_k})m_{1k} \\ (1+\frac{\lambda_2}{\sigma_1})m_{21} & (1+\frac{\lambda_2}{\sigma_2})m_{22} & \dots & (1+\frac{\lambda_2}{\sigma_k})m_{2k} \\ \vdots & & \ddots & \\ (1+\frac{\lambda_k}{\sigma_1})m_{k1} & (1+\frac{\lambda_k}{\sigma_2})m_{k2} & \dots & (1+\frac{\lambda_k}{\sigma_k})m_{kk} \end{pmatrix} \tag{64}$$

The norm of a column is given by

$$\beta_i = \sqrt{\sum_{j=1}^{k}\left(1 + \frac{\lambda_j}{\sigma_i}\right)^2 m_{ji}^2} \tag{65}$$



(a) Random data (b) MNIST dataset (c) ORL dataset (d) CMU PIE dataset

Fig. 1. Relationship between the original criterion $\log\det[P_\pi P_\pi^\top]$ and the approximate criterion $\log\det[\Pi^\top PP^\top\Pi]$, that are obtained by applying PLS for varying number of features, k, in a subset π. The approximate and original criterions are positively correlated for the real datasets. Hence, by maximizing the approximate criterion we are not too far away from the maximum of the original criterion.

Since λ_j can be negative, the lower bound is dependent on ratio between λ_j and σ_i. Therefore we just let l be the column that minimizes the column norms and calculate a $\lambda_{min,l}$ such that

$$\lambda_{min,l} = \arg\min_{\lambda_j} \left|1 + \frac{\lambda_j}{\sigma_l}\right| \tag{66}$$

Then a lower bound for the norms of columns is given by

$$\beta_i \geq \left|1 + \frac{\lambda_{min,l}}{\sigma_l}\right| \tag{67}$$

and an approximate lower bound on the determinant can be written as

$$\det(P_\pi P_\pi^\top + I) \gtrsim \left|1 + \frac{\lambda_{min,l}}{\sigma_l}\right|^k \det(\Pi^\top PP^\top\Pi + I) \tag{68}$$

Combining the two bounds in (62) and (68) we get

$$\left|1 + \frac{\lambda_{min,l}}{\sigma_l}\right|^k \lesssim \frac{\det(P_\pi P_\pi^\top + I)}{\det(\Pi^\top PP^\top\Pi + I)} \leq \left(1 + \frac{\lambda_{max}}{\sigma_{min}}\right)^k \tag{69}$$

In most practical situations the bounds in inequality (69) are much tighter. The number of non-zero eigenvalues of Δ_π are usually few and hence the exponential factor of k is also low.

Figure 1 shows a quantitative relationship between the log determinants of $P_\pi P_\pi^\top$ and $\Pi^\top PP^\top\Pi$ for random data and three of the datasets (ORL, MNIST, CMU PIE) used in our experiments. The random data is of size 100×1000. The point clouds are generated by observing the determinant values for randomly selected subsets of size k. We can see that for the MNIST (Figure 1(b)), ORL (Figure 1(c)) and CMU PIE (Figure 1(d)) datasets, the point clouds are very narrow and the behavior of two matrices are almost positively correlated.

We can use the bounds in inequality (69) to qualitatively describe the situations when the point clouds in Figure 1 will be narrow so that the determinants are positively correlated. The point clouds are narrower when the ratio between the bounds is close to one. Minimizing this ratio is equivalent to minimizing the condition number of the matrix $(W+\Lambda W\Sigma^{-1})$ which is the ratio of its largest singular value to its smallest singular value. The largest singular value of $(W + \Lambda W\Sigma^{-1})$ is $1 + \lambda_{max}$ since the minimum of σ_i is one. The condition number is then given by,

$$\kappa \simeq \frac{1 + \lambda_{max}}{\left|1 + \frac{\lambda_{min,l}}{\sigma_l}\right|} \tag{70}$$

The eigenvalues (λ) of Δ_π are usually few large positive values coming mostly due to $\Pi^\top X_{res}^\top X_{res}\Pi$ and few negative values coming due to $-(X_{res\,\pi}^\top X_{res\,\pi})$. These eigenvalues are indicators of the information content in the noise, where noise is any structure that cannot be explained by the first d components of the PLS model. The eigenvalues in Σ are indicators of information content in the structured data.

When all the λ_j are positive, the approximate lower bound in inequality (67) is $\left(1 + \frac{\lambda_{min}}{\sigma_{max}}\right)$ where $\lambda_{min} = \min_j \lambda_j$. In this case, the condition number is low when the ratio between λ_{max} and λ_{min} is low i.e. the variance in noise is low. When there are negative eigenvalues, $\lambda_{min,l}$ is satisfied by an eigenvalue whose absolute value is close to σ_l. In such a situation, the condition number is low when λ_i are farther from σ_i i.e. the levels of noise and structured data are separated. Therefore the approximation gets better as the variance in noise gets lower and noise levels are farther away from that of structured data. In many real datasets, linear regression can provide good models and hence in such situations by maximizing $\det(\Pi^\top PP^\top\Pi+I)$, we are not too far away from the maximum of $\det(P_\pi P_\pi^\top + I)$.

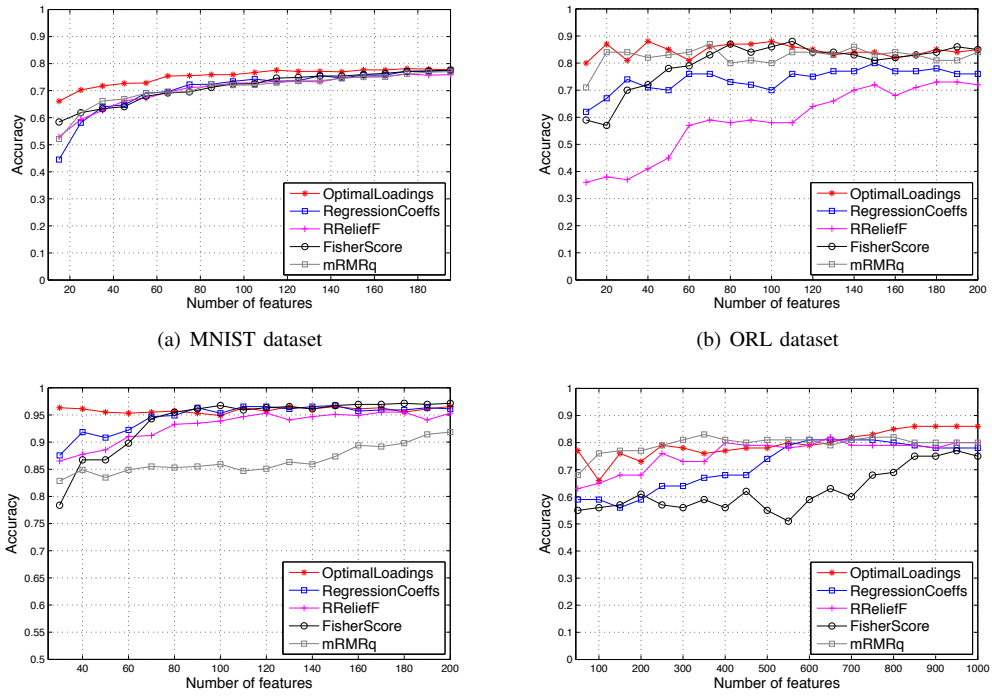
VI. Experiments and Results

To evaluate the performance of our feature selection criterion, we test it in a classification framework where feature selection is treated as a preprocessing filter that produces the indices, π, of the selected feature subset. The feature subset is then used to obtain low dimensional subspaces using PLS. The classification is performed using a Linear Discriminant Classifier in the low-dimensional projection subspace. In a cross-validation setting, the test data is separated from training data for both feature selection and classifier training. This experimental setup is used in order to avoid any overoptimistic performance results obtained when evaluating feature selection using the entire data, as reported in [26].

A. Datasets

The experiments are performed on four datasets - two of them are face image datasets, one is a handwritten digit dataset and the last one is a mass-spectrometric dataset of cancerous and normal tissues. For all of the three image datasets, pixel values are used as features and no feature extraction is performed. The first dataset is a subset[1] [27] of the MNIST handwritten digits. This dataset contains 200

[1] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html



(a) MNIST dataset

(b) ORL dataset

(c) CMU PIE dataset

(d) Arcene dataset

Fig. 2. Classification performance with feature subsets: The D-Optimal Loadings criterion performs better than others on the MNIST and the CMU PIE datasets and performs equally well with the mRMR technique on the ORL and the Arcene datasets. It also shows a consistent performance especially when the number of selected features is small.

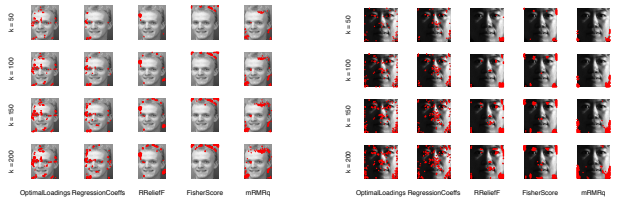
images each for 10 different digit classes, producing a dataset of size 2000×784. The second one is a subset of the AT&T ORL face image database. The dataset consists of face images for 10 subjects with 10 images for each subject with pose variations, which produces a dataset of size 100×10304. The third dataset is a subset of the CMU PIE database that contains face images of 10 different people in a fixed frontal pose (Pose 27) with light and illumination changes. There are 49 images per person, hence producing a dataset of size 490×4096. The fourth is the Arcene dataset from the NIPS Feature Selection Challenge. It contains training and validation sets each of size 100×10000. There are two classes in this dataset.

B. Comparison with other Feature Selection Techniques

We evaluate the performance of D-Optimal Loadings criterion along with other supervised feature selection techniques such as ranking by regression coefficients, Fisher Score [10], RRelief-F [11] and mRMR [14]. For the D-Optimal Loadings criterion, the number of components is chosen based on the minimization of cross validation error of PLS regression and the determinant maximization is performed using a tweaked version of the row exchange algorithm available in MATLAB Statistics Toolbox. The same PLS model is used to obtain the regression coefficients and the top features are selected based on the absolute value of their coefficient. We use the regression version of Relief-F as it showed better performance than the classification version. In RRelief-F, the neighborhood and number of samples for quality estimation are set to 10 and 100 respectively. Finally for the mRMR technique we use the Mutual Information Quotient scheme since it is shown to perform better than the MI Difference scheme. Here we do not discretize the data any further.

We compare the performance using classification accuracies obtained using a Linear Discriminant Classifier. We prefer to use a simple linear classifier so as to avoid tuning the new parameters introduced by nonlinear classifiers. Since the number of selected features can be greater than the number of samples, the classifier is trained in a PLS subspace to avoid over-fitting. The feature subset is used to construct a subspace whose dimensions are again selected based on least cross-validation error for PLS regression. This happens to be same as that used for the D-Optimal Loadings criterion. Given the number of components as d, the experiments are conducted for varying sizes of the feature subset. During the test phase, we select the feature subset from test data, find projections using weights from training phase and then classify using the trained model.

We found that the cross validation error of PLS regression stabilizes at around 10, 15, 30 and 20 components for the ORL, MNIST, CMU PIE and Arcene datasets respectively. Using these number of components, we perform a 20-fold cross-validation experiment for the ORL dataset and 10 fold cross-validation for MNIST and CMU PIE datasets. A larger number of folds is used for the ORL dataset due to smaller number of samples. For the Arcene dataset, the validation set is used as the test set and entire training set is used for training. Figure 2 shows the classification accuracies obtained with D-Optimal Loadings, Regression coefficients, Fisher score, Relief-F and mRMR for the four datasets. The D-Optimal Loadings criterion outperforms other techniques on the MNIST and the CMU PIE datasets and performs equally well with the mRMR technique on the ORL and the Arcene datasets. The D-Optimal Loadings technique can very well handle the situation when the number of selected features is small. We



(a) Sample ORL image (b) Sample CMU PIE image

Fig. 3. Feature points selected by D-Optimal Loadings, Regression Coefficients, Relief-F, Fisher Score and mRMR techniques. The features selected by D-Optimal Loadings are well distributed across the significant regions of the image unlike others that tend to get clustered or lie in noisy regions.

see that the Fisher score and Relief-F are generally worse performing for smaller number of features since they do not handle redundancy among features. In Figure 3 the feature points selected by the five techniques are shown overlaid on sample images from two of the datasets. The features selected by D-Optimal Loadings are well distributed across the significant regions of the image unlike others that tend to get clustered or lie in the noisy regions.

VII. Conclusion

Our work explores the application of the theory Optimal Experiment Design (OED) to Partial Least Squares (PLS) regression. We use the OED to derive the A-Optimal Loadings and D-Optimal Loadings feature selection criterions with the goal of minimizing the variance of the PLS regression model. We specifically use an approximation of the D-Optimal Loadings criterion that maximizes the determinant of loadings covariance matrix to select an optimal feature subset. The availability of off-the-shelf row exchange algorithms and convex optimization methods for determinant maximization hastens the feature selection stage in a pattern analysis problem. One of the important characteristics of the Optimal Loadings criterions is that they are based on optimization of eigenvalues which is necessarily evaluated at a subset level. We also provide insight into the technique by deriving the A-Optimal Loadings criterion by using just the properties of maximum relevance and minimum redundancy for feature subsets. The results from our experiments with four datasets indicate that the D-Optimal Loadings criterion selects better feature subsets when compared to other techniques such as mRMR and Relief-F. Apart from classification accuracies, the locations of feature points on these images also indicate that it selects non-redundant features from the significant regions of the image.

References

[1] I. Guyon and A. Elisseefi, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, no. 7-8, pp. 1157–1182, Oct. 2003.

[2] P. Geladi, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, no. 1, pp. 1–17, 1986.

[3] S. Wold, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, Oct. 2001.

[4] D. Nguyen and D. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39–50, 2002.

[5] A.-L. Boulesteix and K. Strimmer, "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data." *Briefings in Bioinformatics*, vol. 8, no. 1, pp. 32–44, Jan. 2007.

[6] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse, "A sparse PLS for variable selection when integrating omics data." *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, p. Article 35, Jan. 2008.

[7] H. Chun and S. Keles, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection." *Journal of the Royal Statistical Society. Series B, Statistical methodology*, vol. 72, no. 1, pp. 3–25, Jan. 2010.

[8] F. Pukelsheim, *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006, vol. 50.

[9] X. He, "Laplacian Regularized D-optimal Design for active learning and its application to image retrieval." *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 254–63, Jan. 2010.

[10] R. Duda, P. Hart, and D. Stork, "Pattern Classification and Scene Analysis 2nd ed." 1995.

[11] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine learning*, vol. 53, no. 1, pp. 23–69, 2003.

[12] R. F. Teófilo, J. a. P. a. Martins, and M. M. C. Ferreira, "Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression," *Journal of Chemometrics*, vol. 23, no. 1, pp. 32–48, Jan. 2009.

[13] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, pp. 1–8, 2005.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–38, Aug. 2005.

[15] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1393–1434, 2012.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[17] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, Apr. 2005.

[18] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in Neural Information Processing Systems*, vol. 18, 2006, p. 507.

[19] X. He, M. Ji, C. Zhang, and H. Bao, "A Variance Minimization Criterion to Feature Selection using Laplacian Regularization." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 2013–2025, Mar. 2011.

[20] S. De Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251–263, 1993.

[21] F. Giannessi, P. Pardalos, and T. Rapcsak, "Optimization Theory: Recent Developments from Matrahaza," pp. 124–125, 2002.

[22] R. C. S. John and N. R. Draper, "D-Optimality for Regression Designs: A Review," *Technometrics*, vol. 17, no. 1, p. 15, Feb. 1975.

[23] R. Cook, "A comparison of algorithms for constructing exact D-optimal designs," *Technometrics*, vol. 22, no. 3, pp. 315–324, 1980.

[24] L. Vandenberghe, S. Boyd, and S.-P. Wu, "Determinant Maximization with Linear Matrix Inequality Constraints," *SIAM Journal on Matrix Analysis and Applications*, vol. 19, no. 2, p. 499, 1998.

[25] R. Tütüncü, K. Toh, and M. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical programming*, vol. 95, no. 2, pp. 189–217, 2003.

[26] P. Smialowski, D. Frishman, and S. Kramer, "Pitfalls of supervised feature selection." *Bioinformatics (Oxford, England)*, vol. 26, no. 3, pp. 440–3, Feb. 2010.

[27] D. Cai, X. He, and Y. Hu, "Learning a spatially smooth subspace for face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2007.