A PRESS statistic for two-block partial least squares regression

Brian McWilliams and Giovanni Montana

Abstract— **Predictive modelling of multivariate data where both the covariates and responses are high-dimensional is becoming an increasingly popular task in many data mining applications. Partial Least Squares (PLS) regression often turns out to be a useful model in these situations since it performs dimensionality reduction by assuming the existence of a small number of latent factors that may explain the linear dependence between input and output. In practice, the number of latent factors to be retained, which controls the complexity of the model and its predictive ability, has to be carefully selected. Typically this is done by cross validating a performance measure, such as the predictive error. Although cross validation works well in many practical settings, it can be computationally expensive. Various extensions to PLS have also been proposed for regularising the PLS solution and performing simultaneous dimensionality reduction and variable selection, but these come at the expense of additional complexity parameters that also need to be tuned by cross-validation. In this paper we derive a computationally efficient alternative to leave-one-out cross validation (LOOCV), a predicted sum of squares (PRESS) statistic for two-block PLS. We show that the PRESS is nearly identical to LOOCV but has the computational expense of only a single PLS model fit. Examples of the PRESS for selecting the number of latent factors and regularisation parameters are provided.**

I. INTRODUCTION

In this work we consider regression settings characterised by an $X \in \mathbb{R}^{n \times p}$ matrix of p covariates and an $Y \in \mathbb{R}^{n \times q}$ matrix of q responses, both observed on n objects. Assuming centred data, the standard regression approach consists of fitting a Multivariate Linear Regression (MLR) model, that is $Y = X\beta + \epsilon$, where $\beta = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y \in \mathbb{R}^{p \times q}$ is a matrix of regression coefficients and $\epsilon \in \mathbb{R}^{n \times q}$ is the matrix of uncorrelated, mean-zero errors. In many situations where the dimensionality of the covariates is very high, the problem of multicollinearity prevents $X^{\mathrm{T}}X$ from being invertible. Furthermore, it can be shown that the columns, $[\beta_1, ..., \beta_q]$, of the matrix of regression coefficients β are in fact the coefficients of the regression of X on the individual response variables, $[y_1, ..., y_q]$, respectively (see, for instance, [1]). This implies that the least squares estimate for β is equivalent to performing q separate multiple regressions. Therefore, the MLR solution contains no information about the correlation between variables in the response.

A common solution to the problems introduced by multicollinearity and multivariate responses involves imposing constraints on β which improves prediction performance by effectively performing dimensionality reduction. Two-block Partial Least Squares (PLS) is a technique which performs simultaneous dimensionality reduction and regression by identifying two sets of latent factors underlying the data which explain the most covariance between covariates and response where the two blocks refer to the setting where both the covariates and the response are multivariate. These latent factors are then used for prediction rather than the full data matrices X and Y. As such, PLS is used in many situations where the number of variables is very large. Two-block PLS is described in detail in section II.

There are several modern applications for which two-block PLS regression is particularly suitable. PLS is widely used in the field of chemometrics where there are many problems which involve predicting the properties of chemicals based on their composition where the latent factors capture the underlying chemical processes [2]. In computational finance, PLS has been used to identify the latent factors which explain the movement of different markets in order to predict stock market returns [3].

Closely related is the field of multi-task learning in which many tasks with univariate output are learned simultaneously. Using information from multiple tasks improves the performance and so can be treated as one large problem with a multivariate response (see, for example [4]). For example, there are similarities in the relevance of a particular web page to a given query depending on geographical location and so learning the predictive relationship between query terms and web pages across multiple locations helps improve relevance of the returned pages [5]. In applications involving web search data, n can be extremely large, in the order of hundreds of thousands.

One critical issues that arises when using PLS regression to attack such problems relates to the selection of the number of latent factors, R, to include in the model. The choice of R is extremely important: if R is too small, important features in the data may not be captured; if R is too large the model will overfit. Furthermore, the number of latent factors can be important in interpreting the results. Recently, several extensions to PLS have been proposed to improve prediction in several settings. When there are many irrelevant variables which do not contribute to the relationship between X and Y (i.e. the underlying model of the data is sparse), we can regularize the latent factors to remove the contribution of noise variables from the regression coefficients. However, better prediction performance comes at the cost of introducing more important parameters such as the degree of sparsity which must be tuned. The success of these and the many other extensions to PLS depends on the careful tuning of these parameters.

For decades, performing K-fold cross validation (CV) on the prediction error has been a popular tool for model

selection in linear models [6], [7]. In PLS regression, model selection has also been commonly performed using Leave-one-out cross validation (LOOCV) [8], [9] and K-fold CV [10]. The CV procedure involves repeated fitting of the model using subsets of the data in the following way. The data is split into K equal sized groups and the parameters are estimated using all but the kth group of observations which is used for testing. The K groups are each left out in turn and the K-fold cross validated error is given by

$$E_{cv} = \frac{1}{K} \sum_{k=1}^{K} (y_k - X_k \beta(k))^2 \ ,$$

where the subscript k denotes only the observations in the kth group are used whereas (k) denotes the estimate of β obtained by leaving out the kth group.

The choice of K is important as it has implications on both the accuracy of the model selection as well as its computational cost [11]. When $K = n$ we obtain leave-one-out cross validation (LOOCV) where the parameters are estimated using all but one observation and evaluated on the remaining observation. LOOCV is a popular choice for model selection as it makes most efficient use of the data available for training. LOOCV also has the property that as the number of samples increases, it is asymptotically equivalent to the Akaikie Information Criterion (AIC) which is a commonly used for model selection in a variety of applications [12]. However LOOCV can be extremely computationally expensive if n or p is large.

The use of these techniques are subject to the same problems as in OLS, however, since PLS is typically used in situations where the data is very high dimensional and the sample size is small, the problems can be amplified. When n is small compared to p and q, constructing K cross-validation sets is wasteful and the resulting model selection becomes less accurate. Similarly, although LOOCV selects the true model with better accuracy under these conditions, the computational cost becomes prohibitive when p, q or n are large since the complexity of PLS is of order $O(n(p^2 q + q^2 p))$ [13]. Problems of performing model selection in PLS are discussed by [14].

It has long been known that for ordinary least squares, the LOOCV of the prediction error has an exact analytic form known as the Predicted Sum of Squares (PRESS) statistic [15]. Using this formulation it is possible to rewrite the LOOCV for each observation as a function of only the residual error and other terms which do not require explicit partitioning of the data. We briefly review this result in Section III-A. However, such a method for computing PRESS efficiently for two-block PLS has not been developed in the literature. In this work we derive a analytic form of PRESS for two-block PLS regression based on the same techniques as PRESS for OLS which we present in section III-B. In section III-C we show that, under mild assumptions, the PRESS is equivalent to LOOCV up to an approximation error of order $O(\sqrt{\frac{\log n}{n}})$. In section IV we illustrate how the PLS PRESS can be used for efficient model selection with

an application to Sparse PLS where the PRESS is used to select the number of latent factors R and the regularization parameter controlling the degree of sparsity in the regression coefficients. Finally, we report on experiments performed using data simulated under the sparse PLS model and show that the PRESS performs almost exactly as LOOCV but at a much smaller computational cost.

II. PLS REGRESSION

PLS is a method for performing simultaneous dimensionality reduction and regression by identifying a few latent factors underlying the data which best explain the covariance between the covariates and the response. Regression is then performed using these lower dimensional latent factors rather than the original data in order to obtain better prediction performance since only the features in the data which are important for prediction are used.

The two-block PLS model assumes X and Y are generated by a small number R, of latent factors [16], [17], [14]

$$X = TP^{\mathrm{T}} + E_x \ , \quad Y = SQ^{\mathrm{T}} + E_y,$$

where the columns of T and $S \in \mathbb{R}^{n \times R}$ are the R latent factors of X and Y and $P \in \mathbb{R}^{p \times R}$ and $Q \in \mathbb{R}^{q \times R}$ are the factor loadings. $E_x \in \mathbb{R}^{n \times p}$ and $E_y \in \mathbb{R}^{n \times q}$ are the matrices of residuals. The latent factors are linear combinations of all the variables and are found so that

$$\mathrm{Cov}(T, S)^2 = \max_{U,V} \mathrm{Cov}(XU, YV)^2,$$

where $U \in \mathbb{R}^{p \times R}$ and $V \in \mathbb{R}^{q \times R}$ are found by computing the singular value decomposition (SVD) of $M = X^{\mathrm{T}} Y$ as $M = UGV^{\mathrm{T}}$ where U and V are orthogonal matrices and G is a diagonal matrix of singular values. There exists an "inner-model" relationship between the latent factors

$$S = TD + H, \tag{1}$$

where $D \in \mathbb{R}^{R \times R}$ is a diagonal matrix and H is a matrix of residuals so that the relationship between X and Y can be written as a regression problem

$$\begin{aligned} Y &= TDQ^{\mathrm{T}} + (HQ^{\mathrm{T}} + E_y) \\ &= X\beta + E_y^*, \end{aligned}$$

where $\beta = UDQ^{\mathrm{T}}$. Since U and V are computed using the first R singular vectors of the SVD of M, the R PLS directions are orthogonal and so the latent factors can be estimated independently.

III. PRESS STATISTIC

A. The PRESS statistic for OLS regression

In the least squares framework, leave one out cross validation is written as

$$E_{LOOCV} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i \beta^{OLS}(i) \right)^2,$$

where β^{OLS} are the OLS regression coefficients. The subscript i denotes the ith observation whereas (i) denotes the

estimate of β using the observations $(1, ..., i-1, i+1, ..., n)$. Given a solution $\beta(j)$, $\beta(i)$, $j \neq i$ is estimated by adding the jth observation and removing the ith observation. Estimating β requires computing the inverse covariance matrix, $P = (X^TX)^{-1}$ which is computationally expensive. However, since each $\beta(i)$ is different from β by only one sample, we can easily compute $P(i) = (X(i)^TX(i))^{-1}$ from P using the Morrison-Sherman-Woodbury theorem without the need to perform another matrix inversion [15]:

$$\left(X(i)^TX(i)\right)^{-1} = \left(X^TX - x_i^Tx_i\right)^{-1}$$
$$= P + \frac{Px_ix_i^TP}{1-h_i},$$

where $h_i = x_i^TPx_i$. This allows the leave-one-out estimate, $\hat{\beta}(i)$ to be written as a function of $\hat{\beta}$ in the following way, without the need to explicitly remove any observations

$$\hat{\beta}(i) = \left(X(i)^TX(i)\right)^{-1} X(i)^Ty(i)$$
$$= \left(\hat{\beta} - \frac{(y_i - x_i\beta)Px_i}{1-h_i}\right).$$

Finally, the ith LOOCV residual can simply be written as a function of the ith residual error and does not require any explicit permutation of the data, as follows

$$e(i) = \frac{e_i}{1-h_i}.$$

In the next section we derive a similar formulation for the PRESS for PLS.

B. A PRESS statistic for two-block PLS

As described in section II, estimating the PLS regression coefficients involves first estimating the inner relationship between latent factors and then the outer relationship between S and Y. Both of these steps are the result of lower dimensional least squares regression problems. Using the same techniques as in section III-A, we can derive a similar expression for the PLS PRESS statistic in two steps. However, in PLS the latent factors are estimated by projecting the data onto the subspace spanned by U and V respectively which are found using the SVD of $M = UGV^T$.

Our developments rely on a mild assumption: provided the sample size n is sufficiently large, any estimate of the latent factors obtained using $n-1$ samples is *close enough* to the estimate obtained using n data points. In terms of the PLS model, this assumption relates to the difference between $g_r(M_n)$, the rth singular value of M_n (which has been estimated using n rows of X and Y) and $g_r(M_{n-1})$, estimated using $n-1$ rows.

Formally, we assume that the following inequality

$$|g_r(M_n) - g_r(M_{n-1})| \leq \epsilon \qquad (2)$$

holds for $1 \leq r \leq R$ where the approximation error, ϵ is arbitrarily small. Since the rank r approximation error of the SVD is given by g_{r+1}, if the difference between the pairs of singular values is small, it implies the difference between the corresponding pairs of singular vectors is also

small. In other words, within the LOOCV iterations, it is not necessary to recompute the SVD of $X(i)^TY(i)$. We show that this assumption holds in Section III-C.

This assumption (2) implies that the ith PLS prediction error is $e(i) = y_i - x_iUD(i)Q(i)^T$. Since the PLS inner model coefficient, D in Eq (1) is estimated using least squares, we can derive an efficient update formula for $D(i)$ as a function of D using the Morrison-Sherman-Woodbury theorem

$$D(i) = \left(T(i)^TT(i)\right)^{-1} T(i)^TS(i)$$
$$= \left(D - \frac{(s_i - t_iD)P_tt_i}{1-h_t}\right),$$

where $P_t = (T^TT)^{-1}$ is the inverse covariance matrix of the $X-$latent factors and $h_{t,i} = t_i^TP_tt_i$ is analogous to the OLS hat-matrix. Similarly, the $Y-$loading Q is also obtained using least squares and we can derive a similar update formula as follows

$$Q(i) = \left(Q - \frac{(y_i - s_iQ)P_ss_i}{1-h_s}\right),$$

where $P_s = (S^TS)^{-1}$ and $h_{s,i} = s_i^TP_ss_i$. Under the assumption (2), U and V are fixed and so these recursive relationships allow us to formulate the ith PLS prediction error, $e(i)$ in terms of the ith PLS regression residual error e_i in the following way for one latent factor:

$$e^{(r)}(i) = y_i - x_i\sum_{r=1}^{R}U^{(r)}D^{(r)}(i)Q^{(r)}(i)$$
$$= \sum_{r=1}^{R}e_i^{(r)}\left[\frac{1+a}{(1-h_s)} + \frac{-a-b}{(1-h_t)(1-h_s)}\right],$$

where the following identities from the PLS model have been used: $s_iQ^T = y_i - E_{y,i}$ $t_i = x_iU$, $s_i = x_iUD$ and $\beta = x_iUDQ^T$. where: $a = 1 - E_y/e_i$ and $b = y_iH_iP_ss_i/e_i$.

Since each of the PLS latent factors is estimated independently, for $R > 1$ the overall leave one out error is simply the sum of the individual errors

$$e(i) = \sum_{r=1}^{R}e^{(r)}(i) - y_i(R-1).$$

Finally, the overall PRESS statistic is computed as

$$E_{PRESS} = \frac{1}{n}\sum_{i=1}^{n}\|e(i)\|_2^2, \qquad (3)$$

where $\|\cdot\|_2^2$ is the squared Euclidean norm.

C. A bound on the approximation error

The assumption, (2) is key to developing an efficient version of the PRESS for PLS. It ensures that the SVD need only be computed once with all the data instead of n times in a leave-one-out fashion which results in a computational saving of $O(n(p^2q + q^2p))$ [13] which becomes prohibitively expensive when both p and q are large. From a conceptual point of view, this assumption implicitly states that underlying the data there are true latent factors which contribute most

of the covariance between X and Y. Therefore, removing a single observation should not greatly affect the estimation of the latent factors. In this section we formalise this assumption by introducing a theorem which places an upper bound on error ϵ.

In presenting the theorem, we first rely on two Lemmas. The first result, from details an upper bound on the difference between the expected value of a covariance matrix, of a random vector x and its sample covariance matrix using n samples.

Lemma 1 (adapted from [18]): Let x be a random vector in $\mathbb{R}^p$. Assume for normalization that $\|\mathbb{E}x^\mathrm{T}x\|_2 \leq 1$. Let $x_1, ..., x_n$ be independent realizations of x. Let

$$m_n = C\sqrt{\frac{\log n}{n}}A$$

where C is a constant and $A = \|x\|_2$. If $m < 1$ then

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}x_i^\mathrm{T}x_i - \mathbb{E}x^\mathrm{T}x\right\|_2 \leq m_n$$

The second Lemma is a result from matrix perturbation theory which details an upper bound on the maximum difference between the singular values of a matrix M and a perturbation matrix $M + E$.

Lemma 2 (adapted from [19]): For $M, E \in \mathbb{R}^{n \times p}$,

$$\max_i |g_i(M + E) - g_i(M)| \leq \|E\|_2$$

We are now able to state our result.

Theorem 1: Let M_n be the sample covariance matrix of $X_n \in \mathbb{R}^{n \times p}$ and M_{n-1} be the sample covariance matrix of $X_{n-1} \in \mathbb{R}^{(n-1) \times p}$. $g_r(M)$ is the rth ordered singular value of M then

$$\max_i |g_i(M_n) - g_i(M_{n-1})| \leq m_n - m_{n-1}$$

Proof: To prove this theorem, we first establish a bound on the error between the sample covariance matrices M_n and M_{n-1} by applying Lemma 1 with n and $n-1$ samples to obtain the following inequalities

$$\|M - M_n\|_2 \leq m_n \quad (4)$$
$$\|M - M_{n-1}\|_2 \leq m_{n-1} \quad (5)$$

subtracting Eq (4) from Eq (5) and applying Minkowski's inequality we arrive at an expression for the difference between terms M_n and M_{n-1} as follows

$$\|M - M_{n-1}\|_2 - \|M_n - A_{n-1}\|_2 \leq m_n - m_{n-1}$$
$$\|M_n - M_{n-1}\|_2 \leq m_n - m_{n-1} \quad (6)$$

We now relate this result to the difference between computing the SVD of M_n and the SVD of M_{n-1} by recognizing that M_n is obtained as a result of perturbing M_{n-1} by $M_1 = x_1^\mathrm{T}x_1$ where $x_1 \in \mathbb{R}^{1 \times p}$ is the single observation missing from M_n. Using Lemma 2 and Eq (6) we obtain

$$\max_i |g_i(M_{n-1} + M_1) - g_i(M_{n-1})| \leq \|M_n - M_{n-1}\|_2$$
$$\leq m_n - m_{n-1} \quad (7)$$

which proves the theorem. ∎

This theorem details an upper bound on the maximum difference between pairs of ordered singular values of the covariance matrix of M estimated with all n observations and the covariance matrix estimated with $n-1$ observations. Since A and the constant C do not depend on n and so are the same for m_n and m_{n-1}. Therefore, the value of the error term defined by the bound decreases as $O(\sqrt{\frac{\log n}{n}})$.

IV. MODEL SELECTION IN SPARSE PLS

Although the dimensionality reduction inherent in PLS is often able to extract important features from the data, in some situations where p and q are very large, there may be many irrelevant and noisy variables which do not contribute to the predictive relationship between X and Y. Therefore, for the purposes of obtaining a good prediction and for interpreting the resulting regression coefficients, is it important to determine exactly which variables are the important ones and to construct a regression model using only those variables. Sparse PLS regression has found many applications in various areas, including genomics [9], computational finance [14] and machine translation [20].

A Sparse PLS algorithm [14] can be obtained by rewriting the SVD, $M = UGV^\mathrm{T}$ as a LASSO penalized regression problem

$$\min_{\tilde{u},\tilde{v}} \left\|M - \tilde{u}\tilde{v}^\mathrm{T}\right\|_2^2 + \gamma\|\tilde{u}\|_1 \quad \text{s.t.} \quad \|\tilde{v}\|_2 = 1 , \quad (8)$$

where $\tilde{u}$ and $\tilde{v} \in \mathbb{R}^{p \times 1}$ are the estimates of $u^{(1)}$ and $v^{(1)}$, the first left and right singular vectors respectively. As such, they are restricted to be vectors with unit norm so that a unique solution may be obtained. The amount of sparsity in the solution is controlled by γ. If γ is large enough, it will force some variables to be exactly zero. The problem of Eq. (8) can be solved in an iterative fashion by first setting $\tilde{u} = u^{(1)}$ and $\tilde{v} = v^{(1)}$ as before. The Lasso penalty can then be applied as a component-wise soft thresholding operation on the elements of $\tilde{u}$ (see, for instance, [21]). The sparse $\tilde{u}$ are found by applying the threshold component-wise as follows:

$$\tilde{u}^* = \mathrm{sgn}\,(Hv)\,(|Hv| - \gamma)_+$$
$$\tilde{v}^* = H\tilde{u}^*/\|H\tilde{u}^*\|_2 .$$

Typically, selecting the optimal sparsity parameter, λ involves evaluating the CV error over a grid of points in $[\lambda_{min}, \lambda_{max}]$. However, since the behaviour of the CV error as a function of λ is not well understood this often requires specifying a large number of points which exacerbates the already large computational burden of performing cross validation. In such cases, model selection is often performed using problem-specific knowledge and intuition. Problems of performing model selection in sparse PLS are discussed by [14]. Since the PLS PRESS is computationally efficient for large n, p and q we can evaluate the Sparse PLS model over a large grid of values of λ and quickly compute the PRESS to

determine the best cross-validated degree of sparsity. Some experimental results are presented in the following Section.

V. EXPERIMENTS

In this section we report on the performance of the PRESS in experiments on simulated data where the predictive relationship between the covariates and the response is dependent on latent factors according to the PLS model in Eq (1) where $E_x \in \mathbb{R}^{n \times p}$ and $E_y \in \mathbb{R}^{n \times q}$ are matrices of i.i.d random noise simulated from a normal distribution, $N(0, 1)$. For a fixed value of R, n and $p = q$ we simulate R pairs of latent factors t and s of length n from a bivariate normal distribution in descending order of covariance with means drawn from a uniform distribution. Secondly we simulate R separate pairs of loading vectors u and v of length p and q respectively from a uniform distribution, $U(0, 1)$. In order to ensure the contribution of each latent factor is orthogonal, we orthogonalise the vectors $U = [u_1, ..., u_R]$ with respect to each other using the QR decomposition, similarly for $V = [v_1, ..., v_R]$.

To test the performance of the PRESS for selecting the number of latent factors, we perform a Monte Carlo simulation where for each iteration we draw R as an integer from $U(2, 8)$ so that the true number of latent factors in the simulated data is constantly changing. We measure the sensitivity of model selection using the PRESS and LOOCV by comparing the number of latent factors which minimizes these quantities minimum with the true value of R.

To test the performance of the PRESS for sparse PLS, we use the same simulation setting except now we fix $R = 1$ and induce sparsity into the latent factor loadings for X, u_r. Now $\|u_r\|_0 = p/j$ which implies that only p/j of the p elements in u_r are non-zero and thus contribute to the predictive relationship between X and Y. By altering j, we change how many of the variables in X are useful for predicting Y.

We perform a Monte Carlo simulation whereby for each iteration we randomize the true number of important variables in X by drawing j from $U(1, 2)$ so that up to half of the variables in X may be unimportant. We evaluate sparse PLS over a grid of 100 values of γ which span the parameter space. We measure the sensitivity of model selection using the PRESS as compared to LOOCV by comparing the selected variables with the truly important variables.

Table I reports on the ratio between the sensitivity achieved by the PRESS, π_{PRESS} and the sensitivity of LOOCV, π_{LOOCV} for both of these settings averaged over 200 trials for different values of n, p and q. When selecting R, PRESS and LOOCV achieve almost exactly the same sensitivity for all values with only very small discrepancies when $n < p$. When selecting γ the error between PRESS and LOOCV is noticeable when p and q are large compared with n. The standard error of the sensitivity ratio (in parenthesis) also increases when p and q are large.

Figure 1 compares the computational time using PRESS and LOOCV for selecting γ as a function of n for different

| n | p, q | π_{PRESS}/π_{LOOCV} | | | |
		Selecting R		Selecting γ	
50	100	1.0	(0.303)	0.92	(0.228)
	500	0.98	(0.390)	0.73	(0.344)
	1000	0.95	(0.351)	0.70	(0.333)
100	100	1.0	(0.141)	0.99	(0.164)
	500	0.99	(0.412)	0.91	(0.190)
	1000	0.99	(0.389)	0.90	(0.203)
200	100	1.0	(0.0)	1.0	(0.071)
	500	1.0	(0.095)	0.95	(0.055)
	1000	1.0	(0.332)	0.91	(0.158)
300	100	1.0	(0.0)	1.0	(0.014)
	500	1.0	(0.1)	1.0	(0.055)
	1000	1.0	(0.127)	0.96	(0.11)
500	100	1.0	(0.0)	1.0	(0.0)
	500	1.0	(0.095)	1.0	(0.0)
	1000	1.0	(0.12)	1.0	(0.017)

TABLE I

COMPARING THE RATIO OF THE SENSITIVITY, π (THE PROPORTION OF TIMES THE CORRECT MODEL IS CHOSEN), π_{PRESS}/π_{LOOCV} WHEN SELECTING R AND γ AS A FUNCTION OF n, p AND q. THE VALUE IN PARENTHESIS IS THE MONTE CARLO STANDARD ERROR. FOR SUFFICIENTLY LARGE n, BOTH METHODS ACHIEVE THE SAME SENSITIVITY. FOR SMALLER VALUES OF n THE PERFORMANCE OF PRESS IS ONLY MARGINALLY WORSE.

values of p and q. We report on computational timings using a 2.0GHz Intel Core2Duo with 4GB of RAM. Relative timing is measured using the `tic`, `toc` function in Matlab v7.8.0. It can be seen that increase in computation time for LOOCV is quadratic as a function of n and the number of variables. In comparison, the increase in computation time for PRESS is linear in these quantities and very small relative to LOOCV. Because of this, it becomes computationally prohibitive to perform LOOCV with a greater number of observations or variables than we have presented.

Figure 2 reports on the approximation error between LOOCV and PRESS for $p = q = 100$ as a function of n. As the number of samples increases, it can be seen that the error decreases as $O(\sqrt{\frac{\log(n)}{n}})$.

In the simulations we have focussed on situations where the response is multivariate. However, the case where the response is univariate ($q = 1$) also commonly occurs. In this situation, the latent factor for Y collapses to a scalar and so we would expect the error between LOOCV and PRESS to be smaller.

VI. CONCLUSIONS

We have showed that in order to obtain good prediction in high dimensional settings using PLS, a computationally efficient and accurate model selection method is needed to tune the many crucial parameters. In this work we have derived an analytic form for the PRESS statistic for two-block PLS regression and we have proved that under mild assumptions, the PRESS is equivalent to LOOCV.

We have also presented an example application where the PRESS can be used to tune additional parameters which improve prediction and interpretability of results in the case
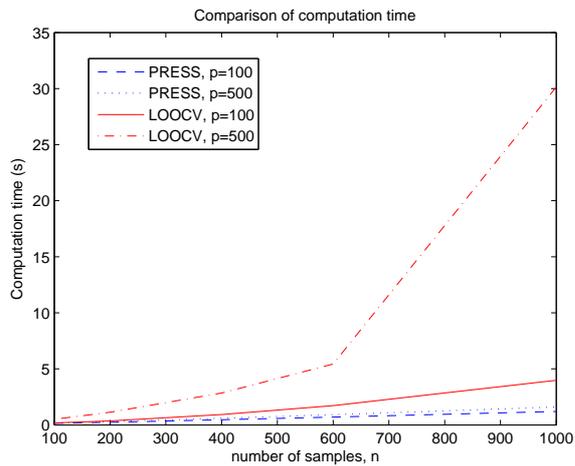
Fig. 1. Comparison of computational timing between PRESS and LOOCV. For fixed values of p and q, computational time (in seconds) for PRESS is linear in the number of observations, n whereas for LOOCV, the computational time increases as a function of n^2. PRESS is also linear in the number of variables whereas LOOCV is quadratic in this quantity.
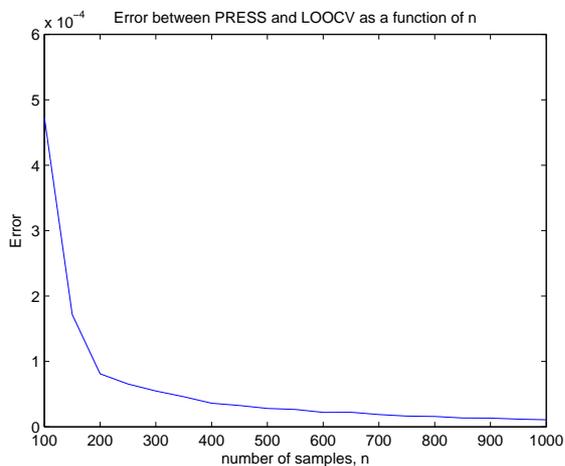

Fig. 2. The approximation error between LOOCV and PRESS. For a fixed p and $q = 100$, the approximation is small. As the number of samples, n increases, the error between LOOCV and PRESS decreases as $O(\sqrt{\frac{\log(n)}{n}})$. See table V for further results.

of sparse PLS. We have showed through simulations that using the PRESS to tune the number of latent factors, R and the sparsity parameter γ, performs almost identically to LOOCV which is the method most commonly used in the literature at a far lower computational cost. When the number of samples is large, LOOCV and PRESS perform identically.

Although we have showed that the analytic PRESS statistic for two-block PLS regression is an important contribution and can be easily applied to the many settings where parameters must be tuned accurately, there are still opportunities for further work. Another such possibility is to construct an objective function in terms of γ, the regularization parameter, so that the optimal degree of sparsity can be tuned automatically.

REFERENCES

[1] A. J. Izenman, *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* Springer Texts in Statistics, 2008, ch. 6, pp. 107–190.

[2] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, Oct. 2001. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0169743901001551

[3] C.-B. Cengiz and H. Herwartz, "Modeling stock index returns by means of partial least-squares methods: An out-of-sample analysis for three stock markets," *Applied Stochastic Models in Business and Industry*, no. November 2009, pp. n/a–n/a, 2010. [Online]. Available: http://doi.wiley.com/10.1002/asmb.826

[4] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, Jan. 2008. [Online]. Available: http://www.springerlink.com/index/10.1007/s10994-007-5040-8

[5] O. Chapelle, K. Weinberger, S. Louis, and B. Tseng, "Multi-Task Learning for Boosting with Application to Web Search Ranking," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington D.C., 2010.

[6] M. Stone, "Cross-validation and multinomial prediction," *Biometrika*, vol. 61, no. 3, pp. 509–515, 1974. [Online]. Available: http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/61.3.509

[7] G. Wahba, "Practical approximate solutions to linear operator equations when the data are noisy," *SIAM Journal on Numerical Analysis*, vol. 14, no. 4, pp. 651–667, 1977.

[8] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, "The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses," *SIAM journal on Scientific Computing*, vol. 5, no. 3, pp. 735–743, 1984.

[9] K.-A. L. Cao, D. Rossouw, C. Robert-Granié, and P. Besse, "A sparse PLS for variable selection when integrating omics data." *Stat Appl Genet Mol Biol*, vol. 7, p. Article 35, 2008. [Online]. Available: http://dx.doi.org/10.2202/1544-6115.1390

[10] H. Chun and S. Kele, "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. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20107611

[11] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, 1974. [Online]. Available: http://www.jstor.org/stable/2984809

[12] ——, "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 44–47, 1977. [Online]. Available: http://www.jstor.org/stable/2984877

[13] G. Golub and C. F. Van Loan, *Matrix Computations.* The Johns Hopkins University Press, 1996.

[14] B. McWilliams and G. Montana, "Sparse partial least squares for on-line variable selection in multivariate data streams." *Statistical Analysis and Data Mining*, no. to appear, 2010.

[15] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, 1st ed. New York, New York, USA: Wiley, 1980.

[16] R. Rosipal and N. Krämer, "Overview and Recent Advances in Partial Least Squares," in *Subspace, Latent Structure and Feature Selection*. Springer, 2006, pp. 34–51.

[17] J. Wegelin, "A Survey of Partial Least Squares (PLS) Methods, with Emphasis on the Two-Block Case," University of Washington, Tech. Rep., 2000.

[18] M. Rudelson and R. Vershynin, "Sampling from large matrices," *Journal of the ACM*, vol. 54, no. 4, pp. 21–es, Jul. 2007. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1255443.1255449

[19] G. Stewart and J.-g. Sun, *Matrix Perturbation Theory*, 1st ed. San Diego, CA: Academic Press,, 1990.

[20] D. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis," 2009. [Online]. Available: http://arxiv.org/pdf/0908.2724

[21] J. Friedman, E. Hastie, H. Höfling, and R. Tibshirani, "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, vol. 1, pp. 302–332, 2007.