Regression models for high-dimensional data with correlated errors


Department of Statistics, University of Leeds

1 Generalised least squares for PCR and PLSR

Given a generalised linear regression model:

\[ y = X\beta + \varepsilon \]  

where \( y \) is the \( n \times 1 \) response vector; \( X \) is an \( n \times p \) model matrix representing the predictors; and \( \beta \) is a \( p \times 1 \) vector of coefficients to estimate. For mathematical simplicity, it is typical to set the first predictor as the intercept \( \beta_0 \) so that the first column of \( X \) is the \( n \times 1 \) vector of ones. The intercept acts as a sink for the mean effect of included predictors, so one could remove the intercept term from the model by centering response and predictors. Unlike the classical conditions imposed on \( \varepsilon \), we assume more generally that \( \varepsilon \sim N(0, \Sigma_{\varepsilon}) \) where \( \Sigma_{\varepsilon} \) is a positive definite matrix. The variance-covariance matrix could be written in the form \( \sigma_{\varepsilon}^2 \Omega_{\varepsilon} \) so that we can obtain the classical model, \( \sigma_{\varepsilon}^2 I \), as a convenient special case.

The Generalised Least Squares (GLS) method focuses on the efficiency issue which fails in ordinary least squares. Efficient estimation of \( \beta \) in the generalised linear regression model requires knowledge of \( \Omega_{\varepsilon} \). To simplify the model, it is useful to consider cases in which \( \Omega_{\varepsilon} \) is a known, symmetric, positive definite matrix. Since \( \Omega_{\varepsilon} \) is a positive definite symmetric matrix, it can be decomposed into \( \Omega_{\varepsilon} = C\Lambda C' \) where the parts of the decomposition \( C \) and \( \Lambda \) are the characteristic vectors (eigenvectors) and roots (eigenvalues) of \( \Omega_{\varepsilon} \), respectively. Now, let \( G' = C\Lambda^{-1/2} \), so \( \Omega_{\varepsilon}^{-1} = G'G \) where \( G \) is square and nonsingular. Premultiply the model (1) by \( G \) to obtain

\[ Gy = GX\beta + G\varepsilon \quad \text{so} \quad y^* = X^*\beta + \varepsilon^* \]

The variance of \( \varepsilon^* \) is

\[ \text{Var}(\varepsilon^*) = \text{Var}(G\varepsilon) = \sigma_{\varepsilon}^2 G\Omega_{\varepsilon} G' = \sigma_{\varepsilon}^2 I \]

so the Ordinary Least Squares (OLS) regression model can apply to this transformed model. Since \( \Omega_{\varepsilon} \) is known, \( y^* \) and \( X_{n,p}^* \) are observed data. In the classical model, ordinary least squares is efficient. Therefore,

\[ \hat{\beta} = (X'X^*)^{-1}X'^*y^* = (X'G'GX)^{-1}X'G'GY = (X'\Omega_{\varepsilon}^{-1}X)^{-1}X'\Omega_{\varepsilon}^{-1}y \]  

(2)

is the efficient estimator of \( \beta \). This estimator is the generalised least squares estimator of \( \beta \). To use the results of GLS, \( \Omega_{\varepsilon} \) must be known. If \( \Omega_{\varepsilon} \) contains unknown parameters that must be estimated from the data, then generalised least squares is not feasible. But with an unrestricted \( \Omega_{\varepsilon} \), there are vastly too many \( (n(n+1)/2) \) parameters to estimate with only \( n \) observations. Obviously, some structure must be imposed on the model if we are to proceed. Under different assumptions on \( \text{Var}(\varepsilon) \), \( \Omega_{\varepsilon} \) will have a simpler structure with fewer unknown parameters and may be estimated. Substituting an estimated \( \hat{\Omega}_{\varepsilon} \) for \( \Omega_{\varepsilon} \) in (2) yields the feasible generalised least squares (FGLS) estimator. More detailed discussion of the GLS and FGLS theory can also be found in Amemiya (1985), Greene (2000) and Ruud (2000).
For high dimensional modeling, multicollinearity is a frequently occurring phenomena, which causes serious problems with the estimation of $\hat{\beta}$ and associated inferences as well as the interpretation. Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) have great advantages for handling multicollinearity. Before extending GLS to PCR and PLSR, we assume that the original explanatory variables can be summarised by a much smaller group of $k$ uncorrelated components. Consider the model,

$$\hat{y} = T_{n,k}q + \xi$$

where $\hat{y}$ is the centered response, the matrix $T_{n,k}$ and vector $q$ denote the components and loadings for PCR or PLSR, and $\xi$ is the matrix of residuals under $k$ components. Because in model (1) an error term $e$ with non-scalar covariance matrix is included, certain additional statistical problems may arise in reference to the residual $\xi$. In such studies, it is important to adjust the effects of the possible confounding error. If the error structure of $e$ can be identified, then we could apply transformation on response and predictors first before employing PCR or PLSR. Unfortunately, the true error structure of $\hat{\xi}$ is usually unknown. In view of the linear model (1) with centered response and predictors, the residual $\xi$ can be written as

$$\hat{\xi} = \hat{y} - \hat{X}\hat{\beta}_P = (\hat{X}\hat{\beta} + e) - \hat{X}\hat{\beta}_P = \hat{X}(\hat{\beta} - \hat{\beta}_P) + e$$

where $\hat{\beta}_P$ corresponds to the regression coefficients for PCR or PLSR. The covariance matrix $\text{Var}(\hat{\xi}) = \text{E}(\xi\xi')$ is given by

$$\text{Var}(\hat{\xi}) = \Sigma_\xi = \sigma_\xi^2\hat{\Omega}_\xi = \text{Var}[\hat{X}(\beta - \hat{\beta}_P)] + \sigma_\xi^2\Omega_\xi$$

Note that the difference between $\text{Var}(\hat{\xi})$ and $\text{Var}(e)$ is due to the confounding error component $\hat{X}(\beta - \hat{\beta}_P)$. In principle, the confounding error represents the unmodelled constituents, and its effect can be either large or small. Without further restrictions or assumptions on this error, we cannot offer analytical results to explain the variance structure of $\xi$. For the desired GLS method to be used in PCR and PLSR, we required that $\hat{\beta}_P \to \beta$, therefore,

$$\lim_{\hat{\beta}_P \to \beta} \text{Var}(\hat{\xi}) = \sigma_\xi^2\Omega_\xi$$

If the regression coefficients $\hat{\beta}_P$ is the same as or very close to the true parameters $\beta$, the disturbance of the confounding error could be ignored and is consequently a special case of model (3) with $\hat{\xi}$ converged to $e$. With further information on $\Omega_\xi$ it is natural to estimate unknown loading $q$ by generalised least squares fitting, i.e. by

$$q_{\text{GLS}} = (T'_{n,k}\Sigma_\xi^{-1}T_{n,k})^{-1}T'_{n,k}\Sigma_\xi^{-1}\hat{y} = (T'_{n,k}\Omega_\xi^{-1}T_{n,k})^{-1}T'_{n,k}\Omega_\xi^{-1}\hat{y}$$

The GLS predictor for $q_{\text{GLS}}$ is also the minimum variance linear unbiased estimator in the generalised regression model. At the same time, it should also be emphasized that the GLS methods is only applied in the second step of the bilinear model (see Luo et al. 2006 for further discussion of the bilinear model).

$$\hat{X} = T_{n,k}P' + \nu$$

$$\hat{y} = T_{n,k}q + \xi$$

By applying the transformation to equation (7), it has no practical effect on the derivation of PCR or PLSR components in equation (6).

We have described how GLS combines with PCR and PLSR to handle high dimensional data in case where $\Omega_\xi$ is known. However, in most circumstances, $\Omega_\xi$ is unknown and must be estimated. Since an unrestricted $\Omega_\xi$ contains $n(n + 1)/2$ parameters, it is clear that some restriction will have to be placed on the structure of $\Omega_\xi$ in order to proceed with the estimation. For the next two sections, we will examine autocorrelation in the model error.
2 Serial correlation

The simplest way to parameterise autocorrelated regression errors is the first-order autoregression or AR(1) process,

\[ \xi_t = \rho \xi_{t-1} + u_t \]

where the random error \( u_t \) are assumed to be Gaussian white noise, \( N(0, \sigma^2_u) \). Under this model, \( \text{Cor}(\xi_t, \xi_{t-s}) = \rho^s \) and \( \text{Cov}(\xi_t, \xi_{t-s}) = \sigma^2_\xi \rho^s \) where \( \sigma^2_\xi = \sigma^2_u / (1 - \rho^2) \).

Examining the fitted residuals from PCR or PLSR can reveal the underlying error structure. The partial autocorrelation function (PACF) is one of the efficient test statistics for locating autocorrelation in the residuals. In general, the partial correlation at lag \( s \) is the autocorrelation between \( \xi_t \) and \( \xi_{t-s} \) that is not accounted for by lags 1 through \( s-1 \). The partial correlogram, the graph for PACF, is a commonly used tool for identifying the order of an autoregressive model. After finding the appropriate order for the autoregressive model, it is possible to derive the error variance for \( \xi \).

3 Simulation study

To understand better the methods described above, we performed a simulation study where the underlying error structures are known. We compare the different proposed methods on a high-dimensional data set \( X \) of size \( n = 120 \) and \( p = 60 \). The data set was produced from the specified multivariate normal distribution \( N_p(0, \Sigma_X) \) with zero mean and unit variance for comparison purposes. The covariance matrix \( \Sigma_X \) is block diagonal with 3 equal size blocks, high correlations fixed within each block at \( 0.9, 0.8, 0.7 \) to induce the multicollinearity problem deliberately. To account for the adverse effect of correlated error, the most frequently analysed AR(1) process \( \varepsilon \) was added as an error term to the simulated model. The simulated model with 5 variables from each correlation block was built as follows,

\[ y = [x_1 : x_5] + [x_{21} : x_{25}] + [x_{41} : x_{45}] + \varepsilon \quad \text{where} \quad \varepsilon_t = 0.9\varepsilon_{t-1} + u_t \]

Here the expression \([x_1 : x_5]\) represents \( \sum_{i=1}^5 x_i \) and \( u_t \) denotes the Gaussian white noise with standard deviation of 2. According to simulation settings, we know the true predictors and the correlated error structure. The goal here is to determine how good the extended PCR and PLSR (GLS-PCR, GLS-PLSR) approaches are at detecting and recovering the true model under conditions of correlated error. The first 120 observations were used to create a training data set and the remaining 80 observations considered as a validation set. The performance of the considered methods is evaluated by means of Root Mean Square Error (RMSE) in the validation data set. The simulations were repeated 100 times to obtain reliable results.

4 Results and discussions

The prediction of the models on the test set are summarised in figure 1. We observe comparable performance of all three GLS regression techniques. In terms of RMSE, Ordinary Least Squares (OLS) which had the largest mean value showed that the validation data are badly fitted. The reason is that the OLS regression does not explicitly take into account the problems of multicollinearity and correlated error. Although the results achieved with GLS regression were superior to OLS, GLS poorly estimating the AR coefficients due to the influence of multicollinearity. Both PCR and PLSR can handle multicollinearity in the data, thus a large reduction of RMSE values were found for both techniques. We can see that the PLSR fitted slightly better
Figure 1: Comparison between the predictive performance of different regression methods

than PCR. This is probably due to the fact that the response variable is involved in the construction of PLSR components, contrary to the PCR components that only depend on the $x$-variables. Using GLS in PCR and PLSR the RMSE decreased drastically and resulted in average values close to the standard deviation of the simulated white noise. This further improvement reflected the fact that the underlying model and correlated error structure were appropriately estimated. Compared with GLS-PCR, the RMSE values are slightly in favour of GLS-PLSR.

References


