Exercises from 'Regression and SVD'

1. Show that for Vr defined in the SVD of phi, VrVrT is a projection matrix and that tr(Vr VrT) = r .

$(V_r V_r^T)^2 = V_r V_r^T$

We know $\phi^T\phi = U_r\Lambda_r U_r^T$ => $(U_r^T)^{-1}\Lambda_r^{-1}U_r^{-1} = (\phi^T\phi)^{-1}$

We also know $\phi = V_r\Lambda_r^{1/2}U_r^T$

=> $V_r = \phi(U_r^T)^{-1}\Lambda_r^{-1/2}$

=> $V_r^T = \Lambda_r^{-1/2}(U_r)^{-1}\phi^T$

Because $(\Lambda_r^{-1/2})^T = \Lambda_r^{-1/2}$ and $((U_r^T)^{-1})^T = (U_r)^{-1}$

$$(V_r V_r^T)^2 = V_r V_r^T V_r V_r^T$$

$V_r V_r^T V_r V_r^T = V_r(\Lambda_r^{-1/2}(U_r)^{-1}\phi^T)(\phi(U_r^T)^{-1}\Lambda_r^{-1/2})V_r^T$

$= V_r(\Lambda_r)^{-1/2}(U_r)^{-1}(\phi^T\phi)(U_r^T)^{-1}(\Lambda_r)^{-1/2}V_r^T$

$= V_r(\Lambda_r)^{-1/2}(U_r)^{-1}U_r\Lambda_r U_r^T(U_r^T)^{-1}(\Lambda_r)^{-1/2}V_r^T$

$= V_r(\Lambda_r)^{-1/2}\Lambda_r(\Lambda_r)^{-1/2}V_r^T$

$= V_r V_r^T$

Show the trace $= V_r V_r^T$ = r

$tr(V_r V_r^T) = tr(V_r^T V_r)$

$tr(V_r^T V_r) = tr(\Lambda_r^{-1/2}(U_r)^{-1}\phi^T\phi(U_r^T)^{-1}\Lambda_r^{-1/2})$

$= tr(\Lambda_r^{-1/2}(U_r)^{-1}U_r\Lambda_r U_r^T(U_r^T)^{-1}\Lambda_r^{-1/2})$

$= tr(\Lambda_r^{-1/2}\Lambda_r\Lambda_r^{-1/2})$

$= tr(I_{(r)}) = r$

2. Weighted least squares: Consider the following small vari- ation of the regression problem. The observations are not ho- moscedastic, that is var(tn) changes with n. Mathematically, the linear regression model now writes as t|X ~ ((phi)w, (qD)-1) where D is a known precision matrix (typically diagonal in practice). Show that the normal equations in this case become:

$N(t|x,\phi w,(qD)^{-1}) = 1/(2\pi)^{D/2}(qD)^{1/2}e^{(-1/2(t-\phi w)^T qD(t-\phi w))}$

$log(\ N(t|x,\phi w,(qD)^{-1}) = log(1/(2\pi)^{D/2}) + (1/2)log(qD) - (1/2)(t-\phi w)^T qD(t-\phi w))$

$= log(1/(2\pi)^{D/2}) + (1/2)log(qD) - (1/2)(q)(t^T Dt - t^T D\phi w - w^T\phi^T Dt + w^T\phi^T D\phi w)$

Differentiating with respect to w and equating to 0. (-1/2) and (q) get cancelled off.

$0 = -t^T D\phi - t^T D^T\phi + w^T(\phi^T D\phi + (\phi^T D\phi)^T)$

$2t^T D^T\phi = w^T(2\phi^T D\phi)$ because $D^T = D$ and $(\phi^T D\phi)^T = (\phi^T D^T(\phi^T)^T) = \phi^T D\phi$

$$t^T D^T\phi = w^T\phi^T D\phi$$

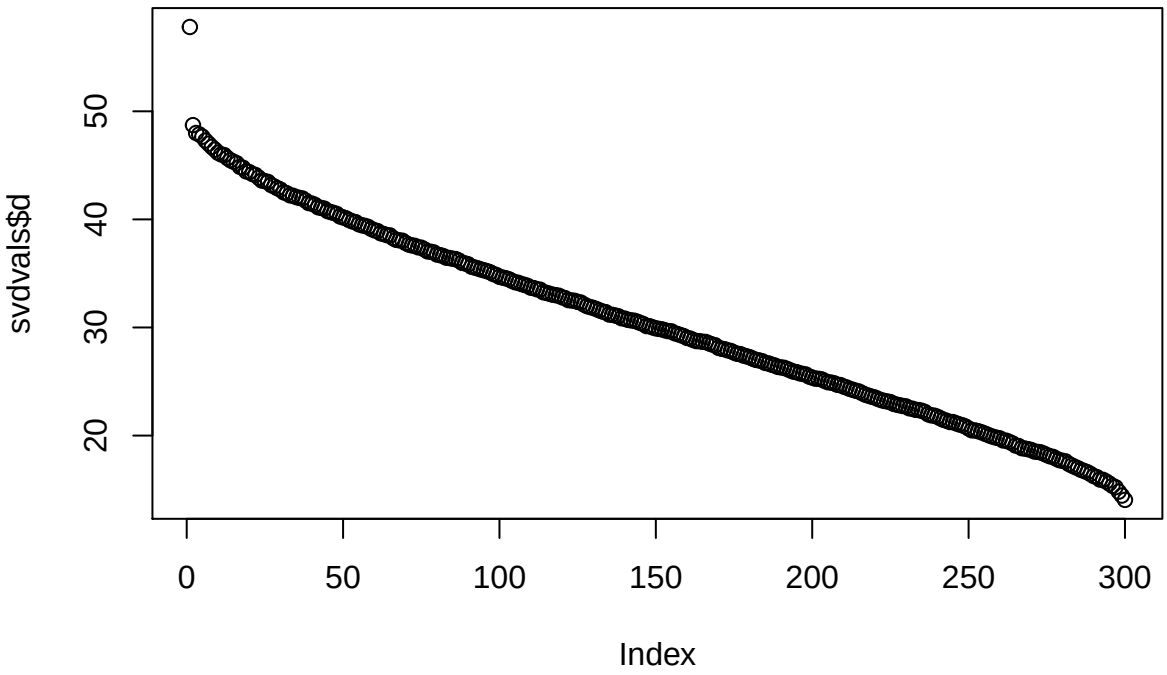
Taking Transpose on both sides:

$$\phi^T D(t^T)^T = \phi^T D^T (\phi^T)^T (w^T)^T$$

$$\phi^T Dt = \phi^T D\phi w$$

3. Extract the first 300 rows and 1001 columns from the synthetic_regression.txt dataset; hence you have 300 replica- tions and 1000 input variables.

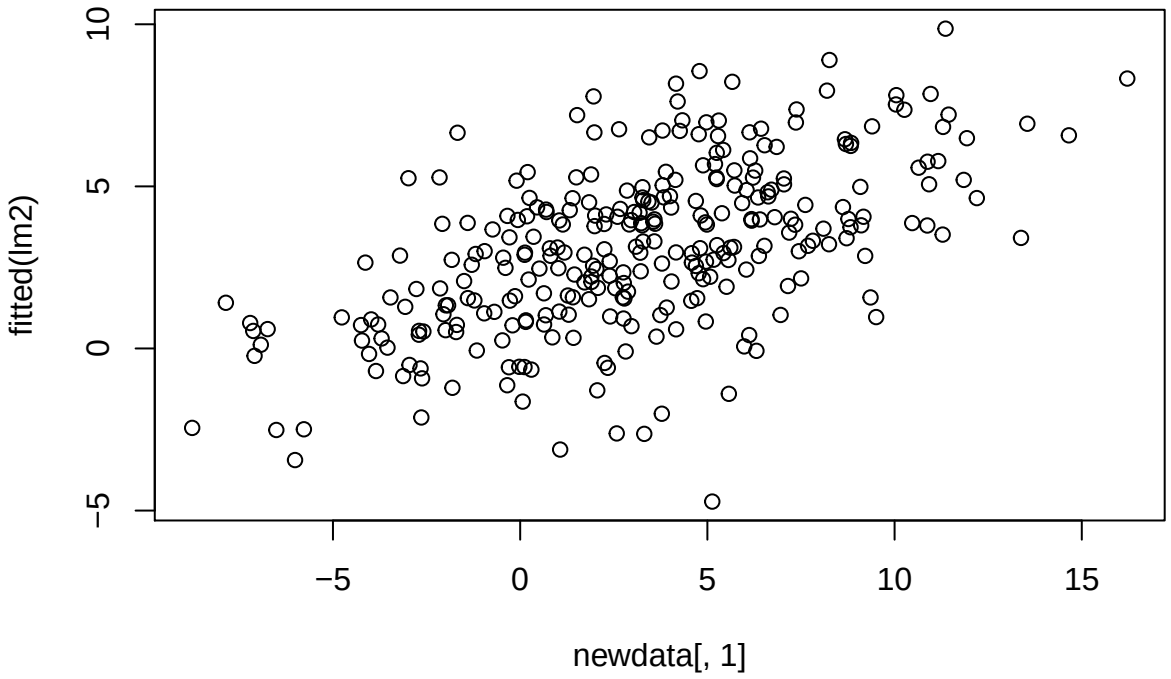
3.1 Plot the non-zero singular values of the input matrix.



3.2 Determine r, the rank of the input matrix.

300

3.3 Do a principal component regression based on the largest 30 eigenvalues of the input matrix and by including an intercept term (hence 31 features in total). Plot fitted vs observed values according to this model. Compare the fit of this model to the one you obtained in a previous exercise using the first (in order of appearance) 30 input variables and the intercept. Is it better or worse? Explain in at most 3 lines your findings and explanation.



The fit of the PCA model is slightly worse than the fit of the linear model.

The reason is the linear model minimizes the error with respect to the observed data whereas the PCA minimizes the error with respect to the model itself. So the linear model is a better fit for the observed data, where it is considered in the fitting, whereas the PCA is a better fit of the included features themselves.