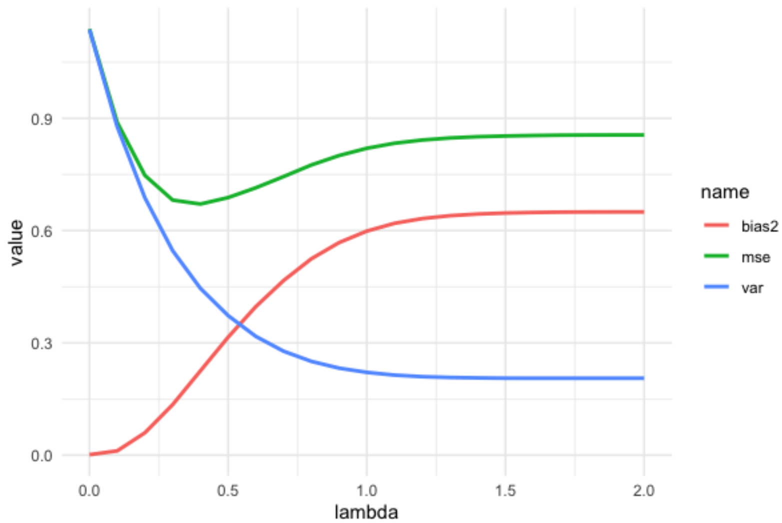
Assignment solution key

1d) The squared bias, variance and MSE should look like something like this as a function of λ:



Note that this image was generated from a different problem, so it only indicates the general shape of these curves and it's not numerically accurate.

The results illustrate the usual bias-variance tradeoff. When $\lambda = 0$, the lasso(=OLS) regression gives unbiased predictions, but the variance of the predictions is very large, because the model is estimated with 20 training observations only. So, we can benefit a lot by increasing the penalty λ a bit, which causes a small increase in bias but radically reduces the variance. The optimal value of λ is where the MSE reaches its minimum. After a certain point increasing λ further is not beneficial because the drop in the variance slows down and stops.

Hence, even in this simple setup ridge regression with a carefully chosen positive λ can outperform OLS ($\lambda = 0$).

2a) The training sample RSS (MSE) steadily decreases. This is because as s increases the constraint becomes looser (equivalently, the penalty for non-zero coefficients decreases) and hence a smaller minimum RSS can be reached over the training sample. (When the constraint is so loose that it doesn't bind any more, the solution is the same as OLS.)

2b) The training sample RSS (MSE) has a U-shape. This is again due to the usual bias-variance tradoff. If s is small, the fitted model is an underfit – the coefficients on the predictors are shrunk toward zero or exactly zero and the model is too simple. So, the predictions for new data points have large bias and small variance. As the constraint loosens (s increases) the shrinkage also loosens and more coefficients need to be estimated. Both changes contribute to an increase in the variance of the predictions while bias decreases because the model is more flexible (and eventually it becomes too flexible, i.e., an overfit). The net effect is the usual U-shape.

2c) Steadily increases (see above)

2d) Steadily decreases (see above)

2e) Unchanged.

3d) If you use the first 500 observations as the training sample and the last 500 as the test sample, the four models give the following test mean squared prediction errors:

DENSE DGP	$N_{tr} = 75$ MSPE	$N_{tr} = 150$ MSPE	$N_{tr} = 500$ MSPE
OLS			4.334
PCA (k=1)			15.006
PCA (k=5)			12.924
PCA (k=10)			8.458

These numbers are very similar to those on the slides.

3e) For your convenience, I am reproducing the table from the slides below:

DENSE DGP	$N_{tr} = 75$ MSPE	$N_{tr} = 150$ MSPE	$N_{tr} = 500$ MSPE
OLS	12.9	6.0	4.5
PCA (k=1)	14.9	14.7	14.6
PCA (k=5)	13.6	13.0	12.7
PCA (k=10)	9.3	8.5	8.0

Discussion. In this 'dense' model all 50 predictors (features) are relevant, so in general what we expect is that if we replace these 50 features with a few principal components, the mean squared prediction error will increase. (The PCA models will be an underfit with large bias.) However, there is also a gain. The smaller models will have fewer coefficients to estimate, which can reduce the variance of the predictions. This gain can be particularly large when the sample size is small. This is precisely what we see in the $N_{tr} = 75$ column: the PCA models with $k = 1$ and $k = 5$ have higher MSPE than OLS because the increase in bias is too large and the reduction in variance is not sufficient to compensate for this. But the PCA model with $k = 10$ hits a sweet spot – it already includes enough principal components not to have too large a bias, but still has relatively few coefficients to estimate, so the variance is lower than for the full model. This is how it can beat OLS.

For $N_{tr} = 500$ OLS gives unbiased predictions with small variance because with this many training sample observations all 50 coefficients can be estimated reasonably precisely. So, there is nothing to gain from compressing the data – the reduced models have much higher bias and not much smaller variance.