The lasso and sparsity levels

MVE440 Statistical learning for big data

Project aim

- Simulate a dataset from the model : $y = X\beta + \varepsilon$
    - Data is simulated by following the instructions on how to simulate data
    - Different sparsity levels are analysed for a couple of different SNR's while fixing a ratio of p/n and vice versa.
    - The data was averaged over different simulations since the data was generated randomly.

- Discuss the effects of:
    - How increasing the sparsity levels affects the result.
        - Specificity
        - Sensitivity
        - Accuracy
    - How does the ratio p/n affect the result?
    - How does the SNR affect the result?

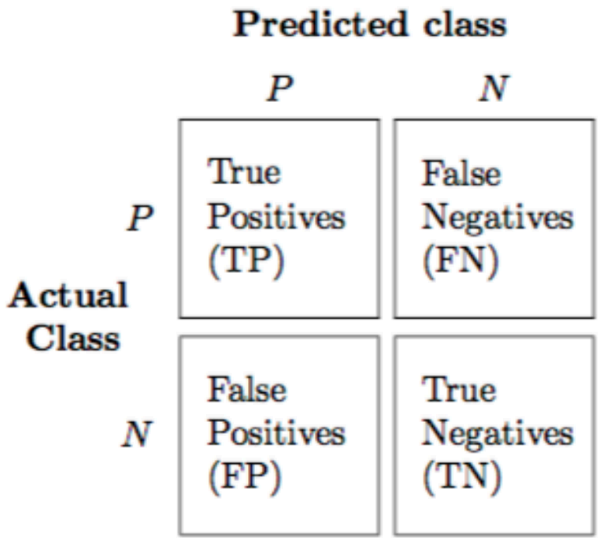
Methods

- Setup:
  - Python
  - Library scikit-learn for machine learning
    - from sklearn.linear_model import Lasso: For a linear model trained with the Lasso (L1 prior as regulariser).
    - from sklearn.linear_model import LassoCV: Lasso linear model with iterative fitting along a regularization path, used to find optimal lambda.
    - from sklearn.metrics import mean_squared_error: Calculates the mean squared error regression loss from the true and predicted values of the model.

- Methods for analysing result:
  - 10-fold cross validation to find optimal hyperparameter lambda
  - Classification metrics
    - Specificity
    - Sensitivity
    - Accuracy
  - Plot of different averaged values for sensitivity, specificity and accuracy against varied values of sparsity level
    - Each presented result is a mean over 10 runs
    - SNR takes the values; 0.5, 1, 5 and 10
    - p/n takes the values; 1.1, 5 and 10
    - The sparsity level is varied between 0.1 and 0.9

Specificity and Sensitivity

- Every time a dataset is simulated for a set of true coefficients, a **confusion matrix** for the recovered coefficients compared to the true coefficients can be constructed.
- **True positives** are coefficients that are non-zero in both the true coefficients and the estimated coefficients. **False positives** are non-zero coefficients in the estimated coefficients but not in the true ones.
- **True negatives** are coefficients that are zero in both the true coefficients and the estimated coefficients. **False negatives** are zero coefficients in the estimated coefficients but not in the true ones.
- Using the confusion matrix e.g. sensitivity and specificity (as well as accuracy) can be computed.



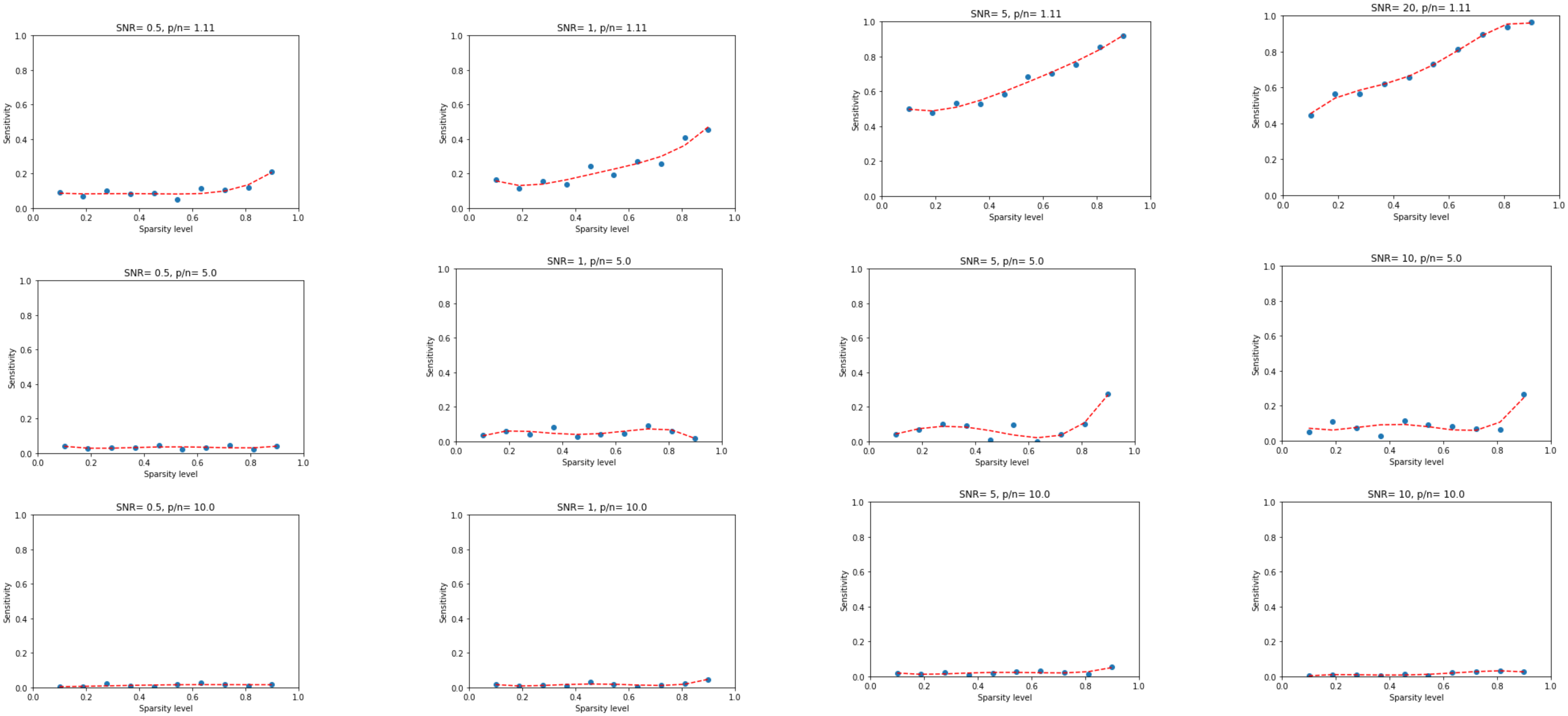
- **Sensitivity=TP/(TP+FN)**
- Measures the proportion of actual positives that are correctly identified as such.

- **Specificity=TN/(TN+FP)**
- Measures the proportion of actual negatives that are correctly identified as such.

- **Accuracy=TP+TN/(TP+TN+FP+FN)**
- Presents the percentage of the predictions that are correct.

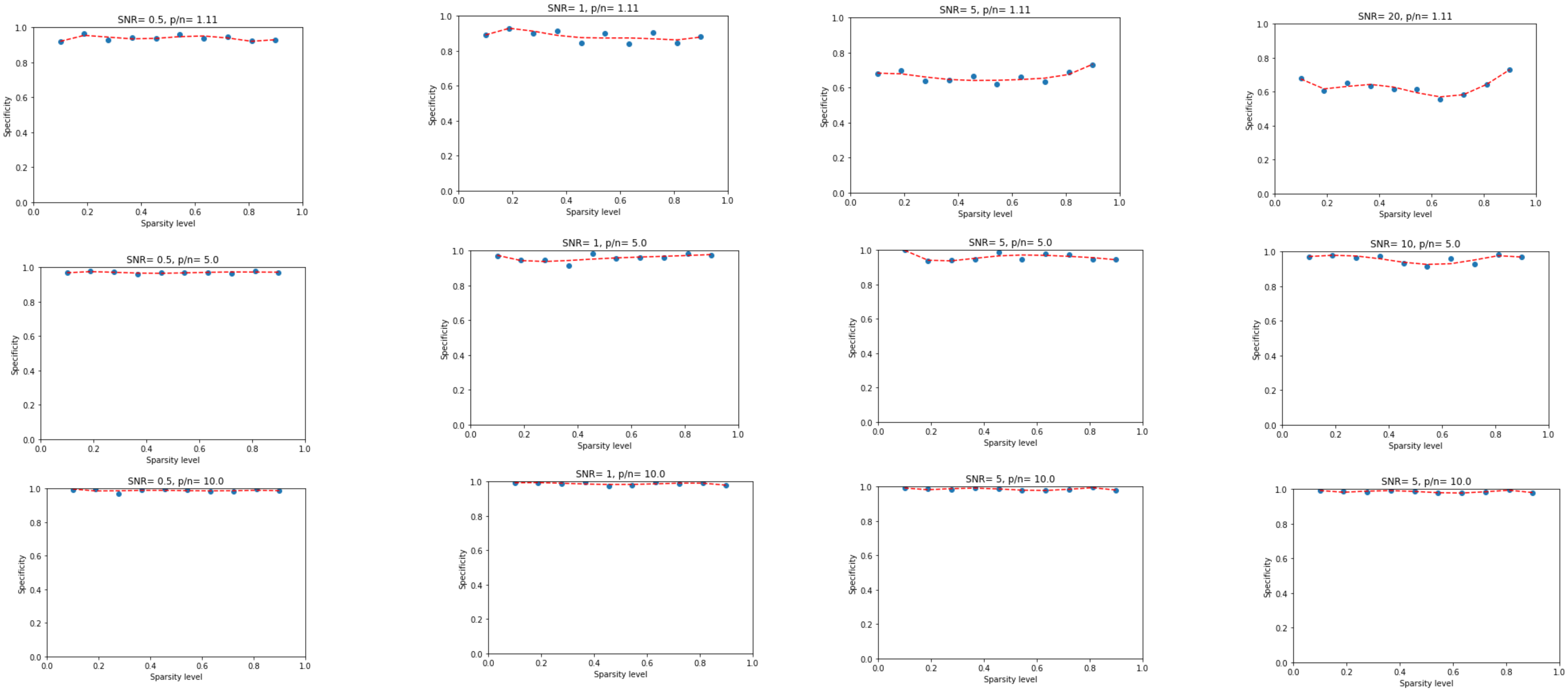
Result sensitivity

- The ratio p/n increases from the top to the bottom and SNR increases from the left to the right



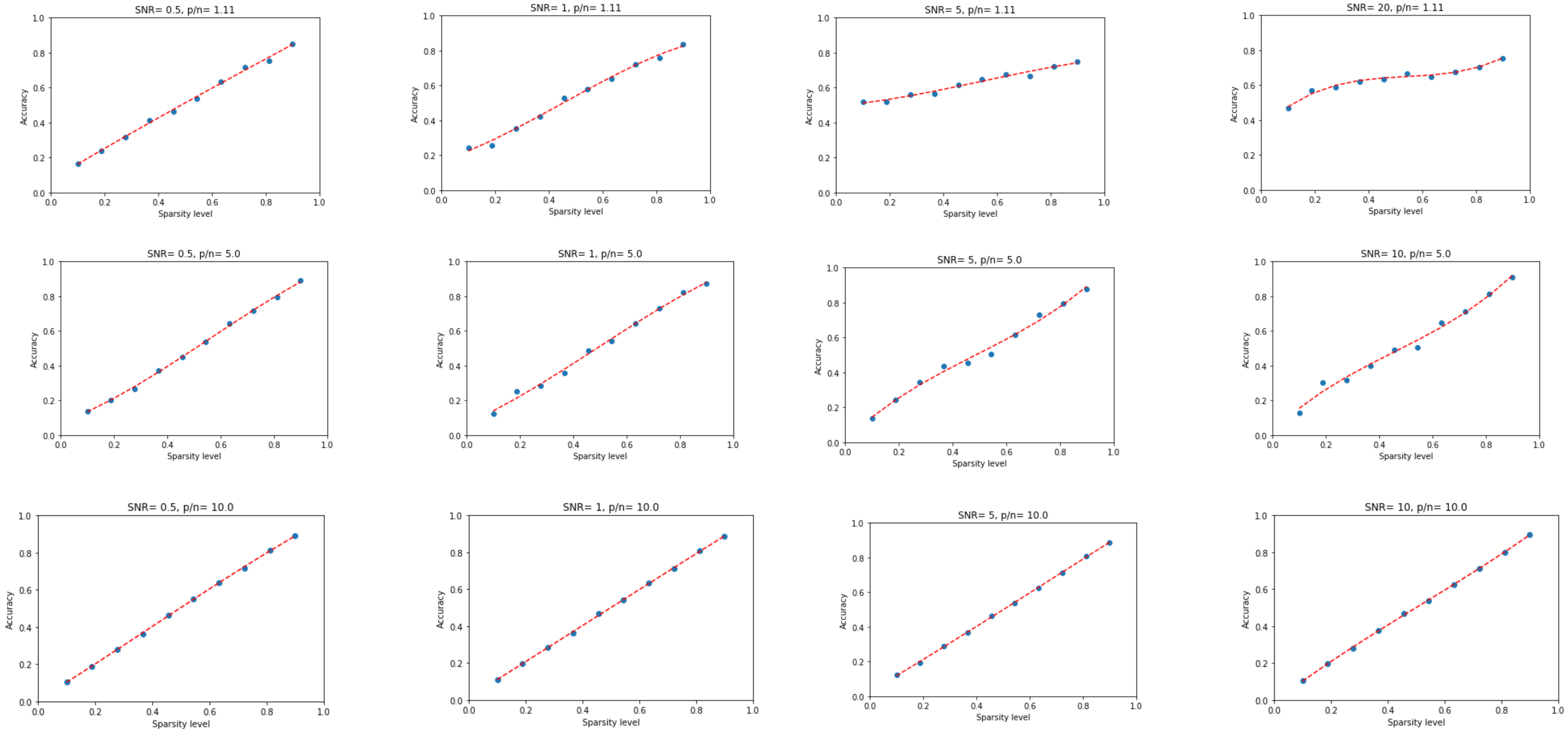
Result specificity

- The ratio p/n increases from the top to the bottom and SNR increases from the left to the right



Result accuracy

- The ratio p/n increases from the top to the bottom and SNR increases from the left to the right



Summary of results

Sensitivity

- As the p/n ratio increases, the sensitivity generally decreases.
- At a high p/n ratio, neither the SNR or the sparsity level have an impact on the sensitivity
- At a decreasing p/n ratio however, the sensitivity increases with increasing SNR and sparsity level.

Specificity

- For the specificity, the opposite occurs.
  - With increasing p/n-ratio, the specificity increases.
  - At a high p/n-ratio, the specificity is almost equal to 1, regardless of sparsity or SNR.
- With a decreasing p/n-ratio, the specificity decreases with increasing SNR.
- Variations in the sparsity does not seem to have a large impact on the specificity as for example the impact from the ratio and SNR (when the ratio is low).

Accuracy

- A general trend that can be observed is that the accuracy increases for increasing sparsity levels, regardless of the ratio and SNR.
- The accuracy can be observed to be approximately linearly dependent of the sparsity level for high p/n ratios.
  - This linear trend doesn’t however hold for for low p/n-rations with increasing SNRs.

Discussion

- Accuracy is mainly used as a final measure of how well the lasso regression worked, i.e. the percentage of the predictions that are correctly predicted.

- For data with low SNR (more noise than structure in the data) and increasing p/n-ratio (i.e. more features than observations):
  - The accuracy is linearly correlated to the sparsity level, the sensitivity is approximately 0 and specificity approximately 1.
  - This indicates that the resulting feature selection from the model predicts most of the features (non-zero coefficients) as non relevant and setting them to zero.
  - This indicated that lasso-regression is not as good at predicting the non-zero (relevant) coefficients.

- An increase in the SNR and decrease in the p/n-ratio, result in an increased sensitivity.
  - With more structure than noise in the data, it is easier to accurately predict the non-zero coefficients.
  - The model gets better in predicting non-zero coefficients, i.e. predicting relevant features and penalizing them.
  - For the same parameter setup, the specificity decreases.

- The sparsity level has the most impact at a high SNR level and low p/n-ratio i.e. when the lasso-regression method performs best. An increase in sparsity leads to an increase in sensitivity at this setup.

Conclusion

- The lasso method generally works better for a higher sparsity level
- Increasing the p/n ratio requires a sparse true model in order to obtain a good performance from the lasso method. This is due to the increasing complexity of predictions for increasing feature dimensions.
  - If the p/n ratio is small the lasso performed well, due to that we have a lot of information (a lot of data) which makes it easier for the lasso method to predict which features are relevant and which are not.
- The lasso-regression method performs best when the SNR is high and the p/n-ratio is low.
  - An increase in sparsity improves the performance due to the fact that lasso regression results in sparse solutions
  - Sparse data results in the lasso method being more selective when choosing which coefficients to set to zero and thereby resulting in a more accurate model.
- At low SNR and high p/n-ratio, the lasso predicts most of the coefficient to be zero.
  - Might be due to the fact that noisy data with many features results in complex models that are hard to predict.
  - If we have dense data (low sparsity) and a large ratio, lasso will put features that might be relevant to zero leading to poor predictions, due to lack of information.
- To conclude, the sparsity level has a great impact on the performance since the lasso method seemed to perform better in sparse data.