Networks with more nodes than observations

Lourens Waldorp

University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 NP, the Netherlands

Abstract

In networks the interest is often to obtain an estimate of the connections between the nodes. To exclude the possibility of missing connections and relations, a large set of variables is selected to include in a network analysis. It may then be the case that more nodes than observations are available in the dataset at hand. In classical statistical approaches an estimate of the connections cannot be obtained and this cannot be resolved. However, recent advances in estimation suggest that even in the situation where we expect only a few connections per node, but have many possible connections (high-dimensional), where overparameterisation is possible, it should be possible to obtain good estimates of the connections in a network. Here we show that indeed with particular model selection procedures reasonable estimates of which nodes are connected are possible. We consider the well-known ridge estimate accompanied by model selection procedures and compare these with the Lasso to select which nodes are connected. The issue with overparameterised models is that the variance of the estimates is inflated leading to unduly high mean squared error, the main tool in model selection. It turns out that adequate network estimates can be obtained whenever, implicitly or explicitly, the possible variance inflation is counteracted and overparameterisation is limited (not too many parameters in linear models).

Key words: model selection, double descent, bias-variance trade-off, mean squared error

1 Introduction

Often we are concerned that the number of observations in a study is insufficient to warrant proper analysis using networks (Epskamp et al., 2018). In particular, a network contains many possible connections and therefore many parameters to be estimated. For instance, with 20 nodes, there are already 190 parameters for the edges in a network. In the classical statistical framework we would need about 1000 observations to reliably estimate these parameters.

Email address: waldorp@uva.nl (Lourens Waldorp).

Recent advances have brought several estimation techniques (e.g., Lasso and ridge) that allow for the so-called high-dimensional situation, where we have more parameters than observations (overparameterisation). In the case we discuss here, the comparison is about the number of nodes in a network and the number of observations. We refer to a scenario as overparameterised or high-dimensional when the number of possible edges to a node is larger than the number of observations but the true number of edges to a node is relatively small. This means that we require only a few edges (that are actually true) but the possible number of edges to search through is larger than the number of observations.

In this paper we will focus on the ridge estimate that is designed for the high-dimensional situation. Often the least absolute shrinkage and selection operator (Lasso, Tibshirani, 1996) is used for high-dimensional situations because the Lasso simultaneously estimates and selects edges to a node. However, the Lasso does not have a closed form solution and is discontinuous, which makes standard inference problematic (Bühlmann et al., 2014). In contrast, the ridge estimator does have a closed solution and is continuous. This makes the ridge estimate more amenable to analysis. The ridge estimate is constructed by adding independent dimensions to the predictors so that a unique solution can be obtained (Hoerl and Kennard, 1970; Rao, 1990). The solution is biased, but can relatively easily be used for inference (Bühlmann et al., 2013). The ridge estimate can also be used for model selection (Akaike and Bayes information criteria and minimum description length) to obtain a reliable estimate of the neighbours of each node in a network. Here we focus on reliable model selection for high-dimensional linear models that may be correctly specified or misspecified.

Interestingly, recent advances in machine learning have shown that overparametrisation (i.e., the situation where we have far more parameters than observations) can be dealt with in a reasonable manner (Hastie et al., 2019; Bartlett et al., 2020; Dwivedi et al., 2020). From these investigations it has appeared that in linear models the true connections of a node can still be identified, sometimes even in the misspecified case (e.g., a true non-linear model estimated by a linear model). By implicitly or explicitly taking into account the high-dimensional space, it is possible to obtain reasonable estimates of the true underlying model. We show that with a cross-validated ridge parameter or with minimum description length (in combination with the ridge estimate), reasonable estimates of the correct set of connected nodes can be obtained as long as complexity of the parameter space (the number of parameters in linear models) is not too high. If, however, the complexity of the parameter space is extremely high, then for misspecified models, the true connections to a node cannot be correctly identified.

We start with the Gaussian graphical model in Section 2. Then in Section 3 both the standard case with more observations than parameters and the high-dimensional case are discussed, where we mostly focus on the ridge estimate. In Section 4 we discuss mean squared error, and its accompanying decomposition

in bias (squared) and variance. We provide some ideas that are frequently used in model selection. Then in Section 5 we turn to model selection. And in Section 6 we include several simulations to determine the performance of standard and less standard model selection criteria. Many of the details of the methods have been deferred to the Appendix.

2 Gaussian Graphical model

The Gaussian graphical model (GGM) is a Gaussian (normal) multivariate distribution associated with a set of nodes and edges in a network. A node j in the network represents a Gaussian random variable X_j, and an edge (i, j) between two nodes i and j in the network represents a conditional dependence relation between the two variables X_i and X_j given all other variables X_k with $k \neq i, j$.

The main characterisation of a GGM are the partial correlations (or partial covariances). That is why a GGM is sometimes referred to as a partial correlation network. This characterisation is because whenever a partial correlation is 0, then the two variables are conditionally independent (Lauritzen, 1996, Proposition 5.2) and so no edge is present in the network; when a partial correlation is $\neq 0$, an edge is present in the network. So, to construct a network we need to determine the 0 partial correlations.

The partial correlations can be determined from regressions. This is called neighbourhood selection or nodewise selection (Meinshausen and Bühlmann, 2006; Bühlmann et al., 2014). This is because the regression coefficient is proportional to the partial covariance. Hence, if the partial covariance is 0, then the regression coefficient must also be 0 (see Appendix A for details).

In nodewise selection, each node in turn is the dependent variable and the remaining nodes are the independent or predictor variables. We call variable X_i the dependent variable and denote it by y and the remaining nodes are X_j for $j \neq i$. In the scenarios we discuss we will be selecting a set of nodes that are relevant to predicting Y. We let J denote the index set for the predictors used in predicting Y (which equals X_i) with a subset of the remaining nodes. The set J never contains the dependent variable $Y = X_i$ (no self loops), so that J is a subset of $\{0, 1, \ldots, i-1, i+1, \ldots, p\}$. Then we consider predictions of Y for model J defined by

$$\hat{Y}_J = \sum_{j \in J} X_j \beta_j \tag{1}$$

The true set of predictors is indexed by S, and we call $\hat{Y}_S = \sum_{j \in S} X_j \beta_j$ the true prediction. This implies that only for those nodes $j \in S$ is $\beta_j \neq 0$ and for all other nodes $\beta_j = 0$.

For each model (subset) J we can obtain an estimate of the coefficients corresponding to the variables x_j such that $j \in J$. By estimating β_j we obtain information on the neighbourhood of node i: for any $\beta_j \neq 0$ we draw the edge

(i, j); and if $\beta_j = 0$, then there is no edge between i and j. We therefore require estimates of the β_j, which we discuss next.

3 Estimation

By using the regressions for the GGM, we can consider one regression, i.e. one neighbourhood, of a node and use this as representative for all nodes. This works because each node is similar to the others, when the assumptions are the same for each node. Since we are interested in the high-dimensional scenario ($p > n$), we will need to adjust standard methods to be able to compute the estimate in linear regression. Here we discuss two cases. First, the scenario with sufficient number of observations for traditional least squares estimation ($p < n$). And, second, the high-dimensional scenario with $p > n$, which requires some additional constraint to make the estimate unique.

3.1 Estimation when $p < n$

We begin with the usual situation where we have more observations than parameters (coefficients). This is referred to as the low-dimensional scenario $p < n$, since we have more observations n than parameters p. In linear regression the standard way to obtain an estimate $\hat{\beta}_j^{LS}$ is obtained by using least squares (see, e.g., Searle, 1971; Magnus and Neudecker, 1999; Bilodeau and Brenner, 1999). In least squares an $\hat{\beta}_j$ for $j \neq i$ is obtained by minimising the squared residuals (see Appendix B). If there is no collinearity (i.e., the correlation between predictors is bounded away from 1), then in the $p < n$ scenario $\hat{\beta}$ is unbiased (if the linear model is approximately correct) and has minimal variance (Rao, 1990; Bilodeau and Brenner, 1999; Seber and Lee, 2012, and see Appendix B for more details).

3.2 Estimation when $p > n$

Because we have more parameters than observations when $p > n$, we can no longer use the standard least squares approach to obtain a unique solution (Rao, 1990, see also Appendix C). To obtain a unique solution a constraint on the parameters can be imposed. One such solution is the least absolute shrinkage and selection operator (Lasso, Tibshirani, 1996), where the sum of absolute parameters is constrained to be small. The Lasso not only estimates the coefficients but also selects them by shrinking small coefficients to exactly 0 (see Appendix C). Precisely because of the selection property the Lasso has been very popular (van Borkulo et al., 2014; Epskamp and Fried, 2018). However, disadvantages of the Lasso are that

(1) low false positive rate guarantees require thresholds which yield low sensitivity (Wainwright, 2009; Haslbeck and Waldorp, 2020), and

(2) the Lasso has no closed form (no single formula like for least squares) and is discontinuous (Hastie et al., 2015).

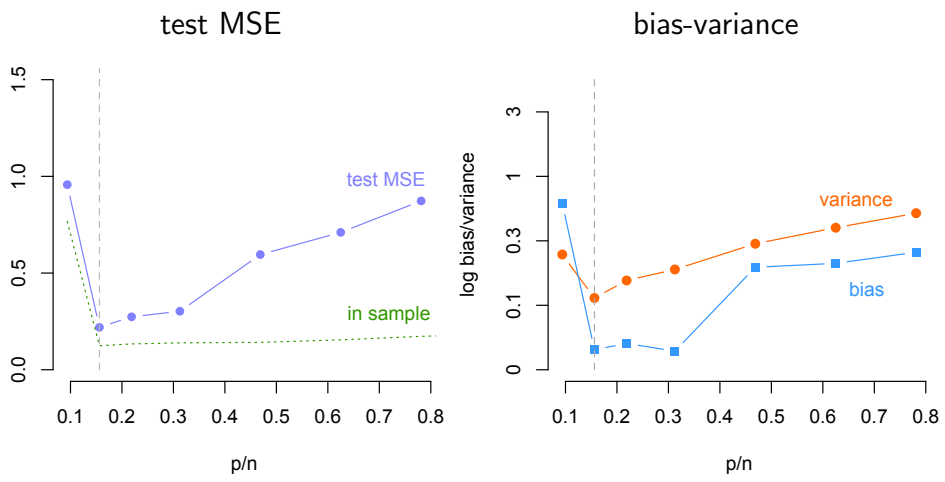
Especially (2) leads to problems for inference. Several workarounds to obtain confidence intervals and p-values have been obtained though (Bühlmann et al., 2013; van de Geer et al., 2014; Lockhart et al., 2014; Dezeure et al., 2015). For our analysis here, we require a closed form and continuous solution, as we will see shortly. We therefore turn to another type of constraint on the parameters to make the solution unique.

Here we choose to impose the constraint that the sum of squared parameters $\hat{\beta}_j^2$ is smaller than some preset value c. The ridge estimate, introduced by Hoerl and Kennard (1970), requires a tuning (penalty) parameter α to be set. This parameter can be obtained by k-fold cross-validation. The ridge estimator has a Bayesian interpretation (Gruber, 1990), where α denotes common variance for the normally distributed prior distribution with independent parameters. The ridge estimator is biased (i.e., the expected value of the estimate is not the true estimator), like the Lasso, but the solution is in closed form and is continuous. Hence, inference on the ridge estimator directly is possible (but see Bühlmann et al., 2013, for improvements) and, as we will see shortly, the ridge estimator leads to shrinkage useful for model selection and linked to a geometric interpretation in high-dimensional models. In Appendix C we provide some details of the ridge estimator.

4 Model selection, test error, and the bias-variance trade-off

Model selection is usually considered as a trade-off between model fit and generalisation to other datasets. Rather than fitting exactly all datapoints of a particular dataset (in sample fit), the aim is to find regularity that will allow the model to describe other datasets well (Myung and Pitt, 1998; Grünwald et al., 2005; Claeskens and Hjort, 2008). How close the model follows the dataset is referred to as bias and how well the model generalises to other datasets is linked to the variance (Hastie et al., 2001). Exact definitions of the bias and variance can be found in Appendix D.

The classical idea of model selection is illustrated in Figure 1. The situation is a linear regression model as described in Section 2 for the GGM. The regression coefficients are estimated by ridge regression for an increasing number of predictors that we put in the model with a subset of the data called the training set (we describe the details of the simulation in Section 6). In Figure 1(b) we see the bias (squared) of the test set (independent data not used for estimation) based on the coefficients of the training set as a function of p/n, the ratio of the number of parameters in the model and the number of observations (in the training set). The correct model is at $p/n \approx 0.15$ (the dashed vertical line). We see that the bias decreases sharply until the correct value is obtained and then increases again, and similarly for the variance. The bias and variance together make up the prediction error (test mean squared

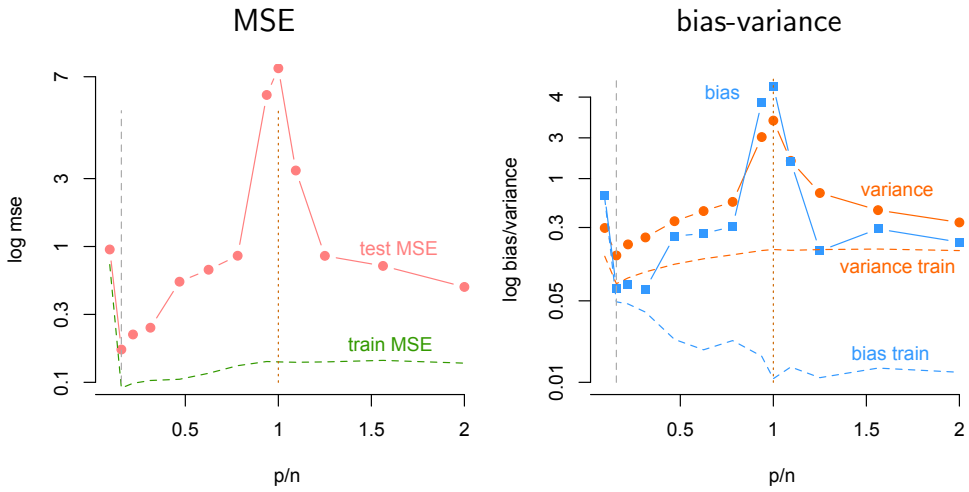


(a) (b)

Fig. 1. In (a) the test and train mean squared error (MSE) for a true linear model as a function of p/n, the ratio of the number of parameters and the number of observations. The true model is at $p/n \approx 0.15$. In (b) are the squared bias (squares) and variance (circles) for the test set.

error or excess risk). This is shown in Figure 1(a). Here we see that at the point ≈ 0.15 (the correct model), where the test mean squared error (test MSE) is lowest, the corresponding model should be selected. The MSE and its relation to the bias and variance is discussed in Appendix D. This is the classic case discussed in for instance Hastie et al. (2001, Chapter 7) and Grünwald et al. (2005, Chapter 1 and Chapter 2). The test MSE is used instead of the MSE based on the training (estimation) data (seen as the dashed green line in Figure 1(a)) because the training MSE is overly optimistic and so remains low (see Efron, 1986; Hastie et al., 2001, and Appendix D). This trade-off between bias and variance implies that the complexity of the model (number of parameters in linear models) cannot be too high because then the model tends to overfit, i.e., fit only the training data and does not generalise well to other data.

In recent years it has appeared that this classical scenario is not all there is to it. There are situations where a model can over fit and still generalise well. This phenomenon (sometimes called benign overfitting, Bartlett et al., 2020) is seen in Figure 2(a). The linear regression coefficients are estimated using the ridge estimator where the number of parameters (predictors) is increased such that the ratio p/n (for the training set) ranges between 0.1 to 2. At $p/n = 0.1$ we have 10 observations per parameter and at $p/n = 2$ we have 0.5 observations per parameter. The classical case is seen up to $p/n = 1$ (i.e., $p = n$), the interpolation point. From that point on, the model is over-parameterised ($p > n$) and has very low training MSE (green dashed line in Figure 2(a)). In Figure 2(b), we see that, as expected, the bias decreases after the interpolation point. Unexpectedly, the variance starts to decrease as well,



(a) (b)

Fig. 2. In (a) the test and train mean squared error (MSE) for a true linear model as a function of p/n, the ratio of the number of parameters and the number of observations (in the training set). The true model with 5 predictors is at $p/n \approx 0.15$, indicated by the vertical grey dashed line. In (b) is the bias (squares) and variance (circles) for the test set and for the training set (dashed lines).

suggesting that the model will generalise well to other datasets. But this seems paradoxical since overparametrisation (and hence overfitting) would suggest that noise is being modeled. But the decreasing variance implies that the over-parameterised model is capable of capturing regularities in other datasets, i.e., it will generalise well.

This phenomenon is often referred to as the double descent (Belkin et al., 2019), since there is a second descent after the interpolation point ($p = n$). The double descent has been investigated thoroughly (Hastie et al., 2019; Bartlett et al., 2020, 2021) with some results as follows. We assume for these results that the all random variables are normally distributed and have independent identically distributed predictors.

(a) In linear regression when the model is approximately correct and p/n is large (overparameterised), the global minimum of the MSE is obtained at the correct model.

(b) If the model is misspecified and p/n is large (overparameterised), then the global minimum of the MSE *could* be obtained at an overparameterised (incorrect) model.

The first result (a) implies that we are still able to correctly identify the approximately correct model when correctly specified. And the second result (b) implies that when the model is not correctly specified, it might be the case that in model selection an overparameterised model will do better in terms of prediction (MSE). This last result explains why in deep learning (and in machine learning) astounding results have been obtained by hugely overparameterised models. Examples are object and speech recognition and

traffic sign classification (see, e.g., Goodfellow et al., 2016). In Appendix D we provide some more details on the mean squared error, including using cross-validation which removes the peak in Figure 2 (see Figure D.1).

Here we focus on the issue that when $p > n$ and we want to select the neighbourhood of a node in a network, then we want to be able to identify the correct set of nodes (i.e., the correct model). The problem can be considered as an issue induced by high-dimensional space. In the case where $p > n$ (overparameterisation) we obtain zero training error, even with a linear model. This is because the complexity of the model is so vast, that any data point can be accounted for. Conceptually, the volume of the parameter space more or less explodes with increasing dimension and, hence, dominates the regression (see Appendix E and Appendix F). In order to obtain a correct model in such high-dimensional situations, this increased model complexity needs to be taken into account either explicitly or implicitly. It will turn out that the tuning parameter α of the ridge regression acts as a way to keep the test variance low, and that model selection can explicitly take model complexity into account in an appropriate way for high-dimensional models.

We explain the effect of the tuning parameter on the variance (and bias) and the relation to model complexity in detail in Appendix F. Here we give a summary of those results. The tuning parameter α can be expressed as the inverse of the signal-to-noise ratio (SNR). The SNR is defined as the ratio of the Euclidean length of the parameters ($||\beta||_2^2$, see Appendix C) and the noise variance σ_e^2 ($\alpha = 1/\text{SNR}$ and $\text{SNR} = ||\beta||_2^2/\sigma_e^2$). The Euclidean length is a representation of the volume for the variables β. So, if we restrict $||\beta||_2^2$ to be no larger than c (ridge regression), then we are restricting the volume of the possible values that the parameters in β can take. This implies that when we increase the volume ($||\beta||_2^2$) we reduce the tuning parameter α, and hence penalise the regression less. This can be interpreted as follows: imposing a higher α in ridge regression will lead to decreased variance (shown in Appendix F). This also implies that the peak at the interpolation point in Figure 2 is an artifact of applying a suboptimal tuning parameter (see Figure D.1, where in the middle row the peak has disappeared upon careful selection of α). This is verified in a simulation where the α was obtained with cross-validation (see Appendix D). The α was increased there from about 0.07 before the interpolation point, to about 7 after the interpolation point. Summarising, we can say the following.

(a) We can increase α directly, reducing the test variance, and hence obtaining reasonable test MSE. This leads indirectly to a lower SNR (because $\alpha = 1/\text{SNR}$); or

(b) we can constrain the size of $||\beta||_2$ in the ridge estimator, therefore, decreasing the model complexity (volume) of the model. This will reduce SNR and hence, increase α, leading to a decrease in the test variance. We do this by constraining the ridge estimator with small c such that $||\beta||_2 \leq c$.

In both cases (a) and (b) we are either directly or indirectly constraining the

total signal $||\beta||_2$. In model selection this can be done by either increasing α, or equivalently, decreasing c in the ridge constraint $||\beta||_2 \leq c$, or by incorporating the complexity (volume) of the model (the n-dimensional ball $\mathbb{B}^n(||\beta||_2)$, see Appendix F). In the next section we will see that different model selection procedures impose different constraints to reduce the impact of the large parameter volume.

5 Model selection procedures

Many model selection techniques exist, like C_p Mallows, the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and minimum description length (MDL), to name a few. We focus here on the AIC, BIC and MDL because these criteria appear to be susceptible or sensitive (to some extent) to overparameterisation. We are particularly interested in situation (a) described above where the linear model is approximately correct (because we are considering the GGM) and hence, even though we have $p > n$ we are still able to select the correct model. We will also consider situation (b) with misspecified models.

Model selection criteria have different origins and have therefore different derivations. There are excellent books that describe these derivations well. For instance, the AIC and BIC (and other model selection criteria) are described in Claeskens and Hjort (2008) and the MDL is described extensively in the book by Grünwald (2007). Here we conceptually describe what the model selection criteria do and refer to the Appendix (G, H, I) for more details on the different model selection criteria.

A regression model is represented by a distribution for the residuals and, for the GGM using linear regressions, is identified with a set of parameters indexed by the set J. Increasing the number of included variables in a linear regression is hence identified with a different distribution. Model selection is therefore concerned with distinguishing probability distributions. The so-called Kullbeck-Leibler divergence (KL) is often used to distinguish distributions, where a KL of 0 indicates no difference between distributions, and a value larger than 0 indicates different distributions (see Appendix G). Both the AIC and MDL can be considered as minimising the KL of a hypothesised distribution (our linear regression models) and some generative (unknown) model (Grunwald, 2000; Myung et al., 2006). The AIC estimates the KL and corrects the incurred bias of doing so (see Appendix G). The MDL, on the other hand, applies the idea of encoding the data and hypothesised model to obtain a reasonably approximating model (see Appendix H and I).

As a consequence of the possible overparametrisation ($p > n$), variance is expected to be inflated. There are two ways in which the inflation of variance can be counteracted to obtain adequate model selection. The first is the strength of regularisation in the ridge estimate (Appendix C) and the second is including the model complexity of the regression (Appendix F). Hoerl and

Kennard (1970) showed that the ridge parameter reduces the test variance and increases the bias (see Appendix F). Hence, model selection obtaining minimal MSE results in finding a balance between bias and variance. Therefore, a carefully selected ridge parameter may lead to optimal model selection performance. We will examine both a standard version of the AIC and one where the ridge parameter was obtained by k-fold cross-validation, referred to as the AIC-CV. Instead of using cross-validation to counteract variance inflation, model complexity can be taken into account. The MDL does this by considering the volume of the model space. In linear regression this can be represented by the Fisher information and the p-dimensional ball with radius the Euclidean length of the parameters β (see Appendix F). By including exactly this model complexity term in the MDL, it will be possible to counteract the variance inflation and hence obtain accurate model selection. We examine three versions of the MDL that incorporate this kind of model complexity, referred to as MDL, which is considered the standard and includes the model volume term (Grünwald, 2007), MDL-S, which is equivalent to the BIC and does not include a term for model volume (Schwartz, 1978; Grünwald et al., 2005), and MDL-opt, where the regularisation parameter is minimised (Dwivedi et al., 2020). See the Appendix for more details on each of these model selection criteria.

6 Simulations

In order to determine the behaviour of the model selection criteria in realistic high-dimensional scenarios, we perform simulations. We fix the true model at $d = 5$ non-zero coefficients and the number of total observations is fixed at $n = 40$. We vary the number of coefficients to be estimated from $p = 3$ to 64. So we obtain a ratio of the number of parameters and the number of observations of $p/n = 0.1$ to 2. Data are generated either according to a linear model (correct model specification) or a sigmoid function (model misspecification). The regression coefficients are estimated by ridge regression with a training set of $n_{\text{train}} = 32$ (80% of the total sample size $n = 40$). The test set is of size $n_{\text{test}} = 8$. We consider the case where we know the true model is linear and we estimate the linear model (case (a) in Section 4), and we misspecify the model because the true model is sigmoidal and we use a linear approximation (case (b) in Section 4). More details about the simulations are given in Appendix J.

6.1 Results

Figure 3 shows the proportion of correct ($p = 5$) coefficients in different scenarios. The top row shows for signal-to-noise ratio (SNR, i.e., the variance of the coefficients divided by the variance of the noise, see Appendix J) 1 and 2 the case where we know the linear model is correct and we only need to discover how many coefficients are non-zero. The MDL and AIC-CV (where

the ridge parameter is obtained with cross-validation) show equal performance when SNR is 1, but when SNR is 2, MDL is increased to about 0.9 while the AIC-CV performs similarly to SNR is 1. The poor performance of the AIC (without obtaining the ridge parameter with cross validation) and the MDL-S (equivalent to the BIC) is because of the high-dimensional case. This can be deduced from the bottom row of Figure 3, where we only take into account dimensions up to 20 (before the interpolation point of $p/n = 1$). It can be seen that the proportion correct is adequate for the MDL-S and AIC. So, without additional penalty for the high-dimensional case, the AIC and MDL-S (BIC) do not perform adequately. Figure J.2 in Appendix J additionally shows that at high dimensions the AIC and MDL-S penalties cannot weigh up to the excellent fit of the model; the values drop sharply so that at high dimensions models with a large number of parameters are selected. The MDL hardly overfits (see Figure J.1 in Appendix J) while the AIC-CV tends to overfit somewhat. It is also clear from the results in Figure 3 that the Lasso performs generally adequately.

It is of little consequence for the proportion of correct models selected whether the predictors are correlated or not. In Appendix J, Figure J.5, we show the accuracy with two types of correlated predictors, indicating that performance is relatively similar to the uncorrelated situation.

When the model is misspecified (generated with a sigmoidal and estimated with a linear model), as shown in the middle row of Figure 3, the MDL and AIC-CV perform best (identifying the non-zero coefficients) depending on a high and low SNR. The performance is on average a little worse, but not unreasonably so. However, when the model is misspecified and the number of parameters (but not the true dimension, $p = 5$ is still true) in the linear model is increased to a hugely overparameterised model, then model selection can go wrong. Figure J.4 shows that model selection deteriorates to extremely poor performance in terms of the correct number of connections due to misspecification and allowing for large overparameterisation. In Figure J.3 in Appendix J it can be seen that when the model is misspecified it is possible that the global minimum of the MSE is at a highly overparameterised model (in this case $p = 320$, so that $p/n = 10$). Model selection in such cases of high overparameterisation will fail since the smallest MSE is at the incorrect, highly overparameterised model. When the model is correctly specified (linear in this case), then the global minimum is at the correct model ($p = 5$), as can be seen in Figure J.3. These results correspond to those of Hastie et al. (2019).

7 Conclusion and discussion

We considered estimation and selection of edges in a Gaussian graphical model when the number of nodes exceeds the number of observations. This is an issue because regular estimation techniques like least squares cannot work in such situations. We used the ridge regression estimate to solve this issue, leading

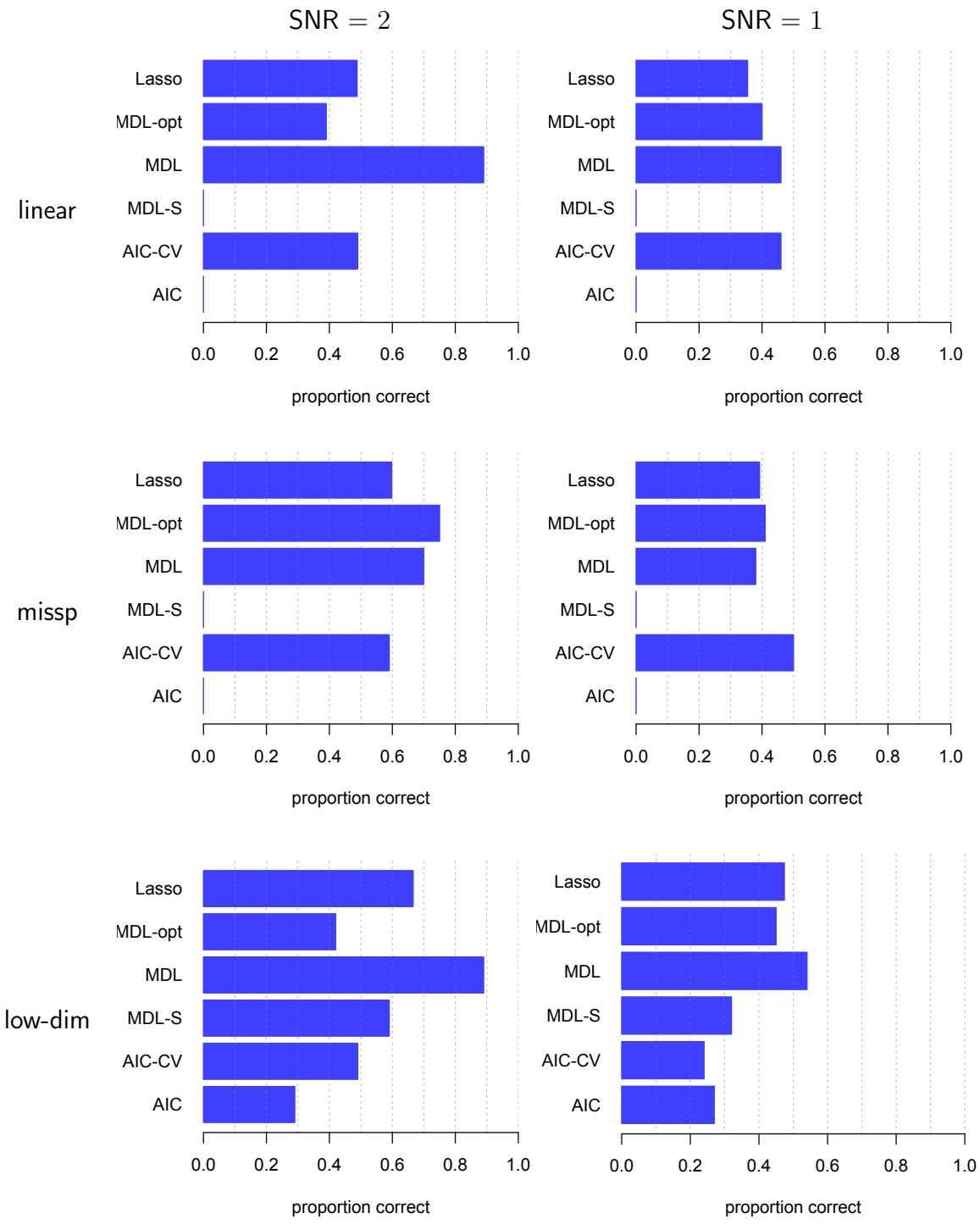


Fig. 3. Proportion of correct decisions (identifying the $p = 5$ non-zero coefficients) of each of the model selection procedures. In the left column the data were generated with an SNR of 2 and in the right column an SNR of 1. In the top row the linear model generated the data (true model) and in the middle row the true model is an sigmoid function (while a linear model was used to estimate the coefficients). In the bottom row only the results up to $p = 20$ were incorporated, corresponding to the normal, low-dimensional situation.

to a biased but closed form estimate. With this estimate we could investigate a range of possible scenarios with many observations to parameters up until only half an observation per parameter (overparameterised).

We next applied several model selection criteria to ascertain whether they could deal with the overparameterised case. Minimum description length appears to be the best choice. The reason is that it explicitly takes into account the volume of the parameter space, which reduces as the number of parameters is increased, and hence, yields a larger penalty. For reasonable results with fewer observations than parameters it is required that the signal-to-noise ratio be reasonable.

In the case where the model is misspecified (here we investigated a true sigmoidal estimated by a linear model), model selection might be said to fail in terms of the dimension (number of connected nodes) if the model is highly overparameterised (here 320 parameters instead of 64). Model selection would fail in those highly overparameterised situations because the test MSE is lower at high values of overparameterisation. Since we considered the ridge estimator we were able to discover that such highly overparameterised models are able to generalise to other samples because of the ridge parameter. The ridge parameter could be interpreted as the inverse of the signal-to-noise-ratio. In low signal-to-noise-ratio situations the ridge parameter will be high, leading to a larger ridge penalty. This penalty was then seen to reduce the parameter variance and, hence, implies that the generalisation can still be reasonable. This also (partly) explains why some highly overparameterised machine learning techniques are able to predict well, mimicking regularised estimation by restricting the class of functions in empirical risk minimisation (Hastie et al., 2019).

References

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Ben-Israel, A. and Greville, T. (1974). *Generalized inverses: theory and applications*. New York: John Wiley and Sons.

Bilodeau, M. and Brenner, D. (1999). *Theory of multivariate statistics*. New York: Springer-Verlag.

Bühlmann, P. et al. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242.

Bühlmann, P., Kalisch, M., and Meier, L. (2014). High-dimensional statistics

with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Cheema, P. and Sugiyama, M. (2020). Double descent risk and volume saturation effects: A geometric perspective. *arXiv preprint arXiv:2006.04366*.

Claeskens, G. and Hjort, N. (2008). *Model selection and model averaging*. Cambridge University Press, Cambridge.

Cover, T. and Thomas, J. (2006). *Elements of information theory*. Wiley and Sons, 2nd edition.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558.

Dwivedi, R., Singh, C., Yu, B., and Wainwright, M. J. (2020). Revisiting complexity and the bias-variance tradeoff. *arXiv preprint arXiv:2006.10189*.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American statistical Association*, 81(394):461–470.

Epskamp, S., Borsboom, D., and Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50(1):195–212.

Epskamp, S. and Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological methods*, 23(4):617.

Giraud, C. (2014). *Introduction to high-dimensional statistics*, volume 138. CRC Press.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Gruber, M. (1990). *Regression Estimators: A Comparative Study*. Academic Press, Boston.

Grunwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, 44:133–152.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.

Grünwald, P. D., Myung, I. J., and Pitt, M. A. (2005). *Advances in minimum description length: Theory and applications*. MIT press.

Hansen, M. H. and Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774.

Haslbeck, J. M. and Waldorp, L. J. (2020). mgm: Structure estimation for time-varying mixed graphical models in high-dimensional data. *Journal of Statistical Software*, 93(8).

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Jones, B. and West, M. (2005). Covariance decomposition in undirected gaussian graphical models. *Biometrika*, 92(4):79–786.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of mathematical statistics*, 22:79–86.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press, Oxford.

Lockhart, R., Taylor, J., Tibshirani, R. J., Tibshirani, R., et al. (2014). A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468.

Magnus, J. R. and Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and economics, revised edition*. Chichester: John Wiley & Sons.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488).

Myung, I. and Pitt, M. (1998). Issues in selecting mathematical models of cognition. In Grainger, J. and Jacobs, A., editors, *Localist connectionist approaches to human cognition*, pages 327–355. Lawrence Erlbaum Associates.

Myung, J. I., Navarro, D. J., and Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, 50(2):167–179.

Pötscher, B. M. and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *Journal of Multivariate Analysis*, 100(9):2065–2082.

Prasolov, V. V. (1994). *Problems and theorems in linear algebra*, volume 134. American Mathematical Soc.

Rao, C. (1990). *Linear statistical inference and its applications*. John Wiley and Sons, second edition edition.

Rao, C. and Toutenberg, H. (1999). *Linear models: least squares and alternatives*. Springer.

Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics*, 14:1080–1100.

Schott, J. R. (1997). *Matrix analysis for statistics*. New York: John Wiley & Sons.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Searle, S. (1971). *Linear Models*. John Wiley and Sons, New York.

Seber, G. A. and Lee, A. J. (2012). *Linear regression analysis*, volume 936. John Wiley & Sons.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., and Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific reports*, 4.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Appendix

A Gaussian graphical model

The Gaussian graphical model (GGM) is characterised by the correspondence between a network of p nodes, labeled $j = 1, 2, \ldots, p$ and a set of random variables $X_1, X_2, \ldots, X_p$ that are jointly Gaussian (multivariate normally) distributed (Koller and Friedman, 2009). The edges in a GGM correspond to conditional dependence between nodes (given all reamining variables). The Gaussian distribution is completely described by its means μ_j and covariances σ_{ij}. The inverse of the covariance matrix $\Sigma = (\sigma_{ij}, i, j = 1, 2, \ldots, p)$, denoted by $\Sigma^{-1} = \Theta$, contain the partial covariances. The partial covariance θ_{ij} is defined as the covariance between the residuals of nodes i and j when all other variables (not i and j) have been regressed out (Lauritzen, 1996). In a GGM a partial covariance of 0 is equal to conditional independence. We show this by example.

Consider three Gaussian variables with mean 0 and covariance matrix Σ. We write the density of the three variables as f_{123} and the conditional density as $f_{13|2}$ for the joint density of variables X_1 and X_3 given X_2. We will want to come to the conclusion that the product of the conditional distributions $f_{1|2}f_{3|2}$ is the same as the conditional distribution $f_{13|2}$. To do this we consider an example with covariance and inverse covariance matrix, respectively,

$$\Sigma = \begin{pmatrix} 3/4 & -1/2 & 1/4 \\ -1/2 & 1 & -1/2 \\ 1/4 & -1/2 & 3/4 \end{pmatrix} \quad \text{and} \quad \Theta = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

Note the 0 at $\theta_{13} = \theta_{31}$ such that variables X_1 and X_3 are conditionally independent given X_2. In terms of the distribution this implies the following. So this implies that we have the density $f_{13|2}$. We now compute this from the density of thew corresponding conditional Gaussian distribution. The distribution with Θ is

$$f_{123}(x) = \frac{1}{\sqrt{(2\pi)^3}}\sqrt{\det(\Theta)}\exp\left(-\tfrac{1}{2}x^\top \Theta x\right)$$

where the term in the exponential is

$$x^\top \Theta x = 2x_1^2 + 2x_2^2 + 2x_3^2 + 2x_1x_2 + 2x_2x_3$$

Note that because $\theta_{13} = 0$ there is no term $2x_1x_3$ and that is why we can rewrite the density. If we take the density of X_1 and X_2, using the first two rows and columns of Θ and calling that Θ_{12}, then we obtain

$$f_{12}(x_1, x_2) = \frac{1}{2\pi}\det(\Theta_{12})\exp\left(-\tfrac{1}{2}(2x_1^2 + 2x_1x_2 + 2x_2^2)\right)$$

And similarly for the density of X_2 and X_3 with Θ_{23} the second and third rows and columns of Θ, gives

$$f_{23}(x_2, x_3) = \frac{1}{2\pi}\det(\Theta_{23})\exp\left(-\tfrac{1}{2}(2x_2^2 + 2x_2x_3 + 2x_3^2)\right)$$

The density of only X_2, ignoring the other two variables is

$$f_2(x_2) = \frac{1}{\sqrt{2\pi}}\Theta_{22}\exp\left(-\tfrac{1}{2}2x_2^2\right)$$

Then we find (after some algebra) that

$$\frac{f_{12}f_{23}}{f_2} = f_{123} \quad \iff \quad \frac{f_{12}f_{23}}{f_2 f_2} = f_{1|2}f_{3|2} = f_{13|2}$$

And we see that if a conditional covariance $\theta_{ij} = 0$, then there is a conditional independence. This is true only for the multivariate normal distribution.

Hence, the GGM is particularly attractive for graphical models because whenever the partial correlation between variables X_i and X_j is 0, then X_i and X_j are conditionally independent given all other variables (Lauritzen, 1996, Proposition 5.2). Hence, an edge in the GGM is present only if the partial correlation is non-zero. The partial correlation is a function of the inverse covariance matrix and has elements θ_{ij}. The partial correlation between X_i and X_j is defined by (Koller and Friedman, 2009; Epskamp and Fried, 2018)

$$\rho_{ij|\star} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$$

where $\star$ indicates all remaining variables in the network except i and j. If $\rho_{ij|\star} = 0$, then no edge should be present between nodes i and j in the network. A GGM can be interpreted as a network where a correlation corresponds to a (set of) path(s) between two nodes (Jones and West, 2005).

We can determine whether a partial correlation also by considering the regression coefficients β_{ij} from the regression

$$X_i = \sum_{j \neq i} X_j \beta_{ij} + e_i$$

where the sum is over all nodes j that are not i. The reason for the correspondence between the partial correlation and the regression coefficient is that (Lauritzen, 1996)

$$\beta_{ij} = -\frac{\theta_{ij}}{\theta_{ii}}$$

Hence, whenever $\beta_{ij} = 0$ then also $\theta_{ij} = 0$ and vice versa. We can, therefore, consider each node in turn and determine the non-zero coefficients of the regressions. Because we have both β_{ij} and β_{ji} we will have to decide to use the *and*-rule, where both β_{ij} and β_{ji} are non-zero to include the edge (i, j), or the *or*-rule, where either β_{ij} or β_{ji} is non-zero to include the edge (i, j) (Meinshausen and Bühlmann, 2006).

B Estimation when $p < n$

In the GGM each node i is the dependent variable in turn, and so the remaining $p - 1$ variables are the predictors. For convenience, we choose $X_i = Y$ to be the dependent variable and we write p instead of $p - 1$ for convenience. We also ignore the i from β_{ij} and simply write β_j, for convenience of notation.

To obtain the estimates of β_j from the linear regression, we minimise the least squares function

$$\text{LS}(J) = \sum_{i=1}^{n} (Y_i - \hat{Y}_J)^2 \quad \text{and} \quad \hat{Y}_J = \sum_{j \in J} X_j \beta_j$$

We can rewrite the model in matrix algebra, with $Y = (Y_1, Y_2, \ldots, Y_n)$ the n-dimensional vector, $X = (X_1, X_2, \ldots, X_p)$ the $n \times p$-dimensional matrix and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, $Y = X\beta$. The least squares function can be written in terms of the Euclidean distance function $||a||_2 = \sqrt{a_1^2 + a_2^2 + \cdots a_p^2}$. The least squares function is then $\text{LS}(J) = ||Y - X\beta_J||_2^2$. Minimisation is achieved from the normal equations

$$X^\top X \beta = X^\top Y \tag{B.1}$$

Similar to numbers on the real line $\mathbb{R}$, we require that we can invert $X^\top X$ so that we obtain our estimate. For numbers on the real line we get for $xb = c$,

that we can invert x so that $b = c/x$. Similarly, the inverse $(X^\top X)^{-1}$ has the property that $(X^\top X)^{-1}(X^\top X) = I$, where the identity matrix I is such that $IX = X$. Then our least squares (LS) estimate is

$$\hat{\beta}^{LS} = (X^\top X)^{-1} X^\top Y$$

It is clear from this derivation that we need the inverse $(X^\top X)^{-1}$. This assumes that the rank (i.e., the number of linearly independent vectors in X) is at least p. A necessary (but not sufficient) condition is that $p < n$. For sufficiency we also require that there is no collinearity (i.e., the eignevalues of $X^\top X$ are all positive). Given these assumptions, in the linear model the estimate $\hat{\beta}^{LS}$ is unbiased (i.e. $\mathbb{E}(\hat{\beta}^{LS}) = \beta$) and has smallest variance (i.e., $\text{var}(\hat{\beta}^{LS}) \leq \text{var}(\tilde{\beta})$ for any other estimate $\tilde{\beta}$). Good references for these results are, for instance, Rao (1990) and Seber and Lee (2012).

C Estimation when $p > n$

From Appendix B we know that in order to obtain an estimate $\hat{\beta}$ we require the inverse of $X^\top X$ to exist. If $p > n$ then this is not the case. As a consequence, there is no unique solution (Schott, 1997). We can understand this by considering the eigenvalues (spectrum) of $X^\top X$, denoted by δ_j. If the $p \times p$ matrix $X^\top X$ has full rank p, then it has p eigenvalues $\delta_j > 0$ (Schott, 1997). If $X^\top X$ does not have full rank (because in our case $p > n$), but has rank k, say, then there are $p - k$ eigenvalues equal to 0. This refers to the idea that a part of the space spanned by $X^\top X$ projects on the null-space (i.e., the space such that $X^\top X u = 0$ for any vector $u \neq 0$). Hence, there is no unique solution when $X^\top X$ is $p > n$.

One way to obtain a unique solution is to impose constraints on the vector β. One of the constraints is to impose an upper bound on the sum of absolute values, i.e., $\sum_j |\beta_j| \leq c$, for some $c > 0$. Then the problem

$$\min_\beta ||Y - X\beta||_2^2 \quad \text{such that} \quad \sum_{j \neq i} |\beta_j| \leq c$$

is called the least absolute shrinkage and selection operator, abbreviated by Lasso (Tibshirani, 1996). The Lasso has several attractive properties, like selecting the non-zero coefficients and consistency (i.e., converging with n to the true value), given certain strong assumptions (see, e.g., Bühlmann and van de Geer, 2011; Hastie et al., 2015). The Lasso has no closed form solution and is not amenable to inference (Pötscher and Leeb, 2009; Bühlmann et al., 2014). However, several workarounds have been obtained, like the debiased Lasso (van de Geer et al., 2014) and the multi sample-split (Meinshausen et al., 2009; Dezeure et al., 2015).

Another constraint that can be imposed on the vector β is to upper bound

the sum of squares of the parameters. This leads to the problem

$$\min_{\beta} ||Y - X\beta||_2^2 \quad \text{such that} \quad \sum_{j \neq i} \beta_j^2 \leq c$$

The solution to this minimisation problem is given by

$$\hat{\beta} = (X^\top X + \alpha I_p)^{-1} X^\top Y$$

where $\alpha > 0$ is some constant. This is called the ridge estimator (Hoerl and Kennard, 1970). It depends on setting the value α. This solution is equivalent to jointly minimising the least squares function and the sum of squared parameters (Rao and Toutenberg, 1999). Hence we obtain the solution such that $||\hat{\beta}||_2^2 = c$ (Hastie et al., 2001).

The benefits of the ridge regression over the Lasso are (Bühlmann et al., 2013)

(a) the ridge regression has a closed form solution
(b) the ridge estimator is continuous and so allows a sampling distribution and the calculation of standard errors

Property (b) is especially appealing in practice because it allows for direct inference with the ridge estimate (Bühlmann et al., 2013). The estimate has a bias $(X^\top X + \alpha I_p)^{-1} X^\top X - I$, which shows that if $\alpha \to 0$, and there is no collinearity, then the bias is negligible. The variance can be computed from which the standard errors are obtained for confidence intervals and p-values (Hoerl and Kennard, 1970; Gruber, 1990). In Bühlmann et al. (2013) a strategy is proposed to reduce the bias and obtain more reliable confidence intervals and p-values.

D Excess risk and mean squared error

Let $\hat{y}_S$ be the true prediction based on the true variable based on the true set S of predictors, i.e.

$$\hat{Y}_S = \sum_{j \in S} X_j \beta_j \tag{D.1}$$

The data are generated according to an additive model where errors e are added to $\hat{y}_S$, so that

$$Y = \hat{Y}_S + e = \sum_{j \in S} X_j \beta_j + e \tag{D.2}$$

where the errors e are independent and identically normally distributed with mean 0 and variance σ_e^2.

We define the mean squared error (MSE) as a quantity that represents the expected loss incurred by estimating with model J instead of the true model

S. MSE, also known as excess risk or test MSE, is defined in terms of a so-called test set, an independent point (or set of points) not encountered before in the so-called training set (Hastie et al., 2001). Let (X_0, Y_0) be a test data point independent of, but from the same distribution as, all other (X_i, Y_i). With the training data the ridge estimate $\hat{\beta}_J$ with model J is obtained. Then the excess risk (out of sample prediction risk, Hastie et al., 2019) with respect to the true parameter β is defined as

$$R(J) = \mathbb{E}\left[(Y_0 - X_0^\top \hat{\beta}_J)^2 - (Y_0 - X_0^\top \beta)^2\right] = \mathbb{E}\left(\hat{Y}_J - \hat{Y}_S\right)^2 \tag{D.3}$$

where the expectation is with respect to Y_0 and X_0 and the training X is conditioned on. Similar to Hastie et al. (2019) and Bartlett et al. (2021) we obtain the excess risk (again conditioning on the training X)

$$R(J) = \mathbb{E}\left[(Y_0 - X_0^\top \beta + X_0^\top(\beta + \hat{\beta}_J))^2\right] - \mathbb{E}\left[(Y_0 - X_0^\top \beta)^2\right]$$

so that

$$R(J) = \mathbb{E}\left(X_0^\top(\beta - \hat{\beta}_J)\right)^2 = B(J) + V(J) \tag{D.4}$$

where

$$B(J) = (\beta - \mathbb{E}\hat{\beta}_J)^\top \Sigma (\beta - \mathbb{E}\hat{\beta}_J) \qquad V(J) = \mathbb{E}(\hat{\beta}_J - \mathbb{E}\hat{\beta}_J)^\top \Sigma (\hat{\beta}_J - \mathbb{E}\hat{\beta}_J) \tag{D.5}$$

where $B(J)$ is called the squared bias and $V(J)$ is called the variance, and $\Sigma = \mathbb{E}(X_0 X_0^\top)$, which is the same as the covariance matrix of the predictors $\mathbb{E}(X^\top X)$. This is the famous variance and squared bias decomposition (Hastie et al., 2001). The classical idea is that reducing bias will increase variance and vice versa, thus balancing the MSE. Hence, a good model will minimise the MSE (excess risk).

The training bias $B(J)$ is 0 for the linear model, when $p < n$, and the linear model is correct (Hastie et al., 2019). However, for the ridge estimator $\hat{\beta}$ in Appendix C the bias is non-zero (Hoerl and Kennard, 1970). It turns out that when the in-sample (training) residuals are 0 (i.e., overparameterised), then still the variance on the test set need not be large (Bartlett et al., 2021, see also Appendix F).

Figure D.1 shows for three different estimators the test MSE (left column) and the bias and variance (right column). For the ridge estimator we see the peak at the interpolation point ($p = n$). The peak disappears for carefully selected ridge parameter α using k-fold cross-validation. The Lasso also does not show such a peak but has larger MSE,

The interpolation point at $p = n$ is the point where it is possible that each datapoint is captured by the model in the training set, also in the linear case. We see this in Figure D.2(a), where the residuals of the training (estimation) data as a function of the predicted values $\hat{Y}$ are plotted. Here, the true number

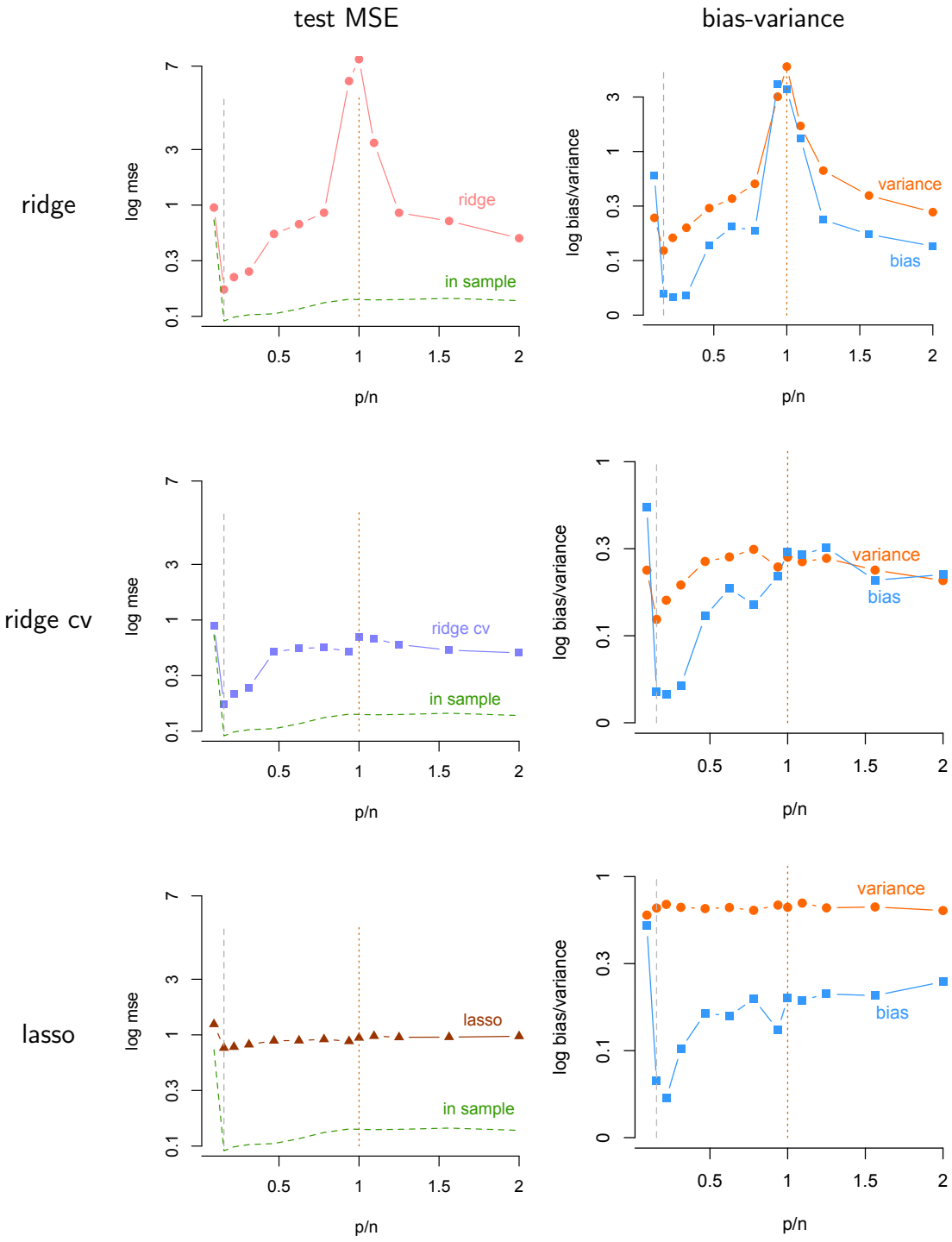
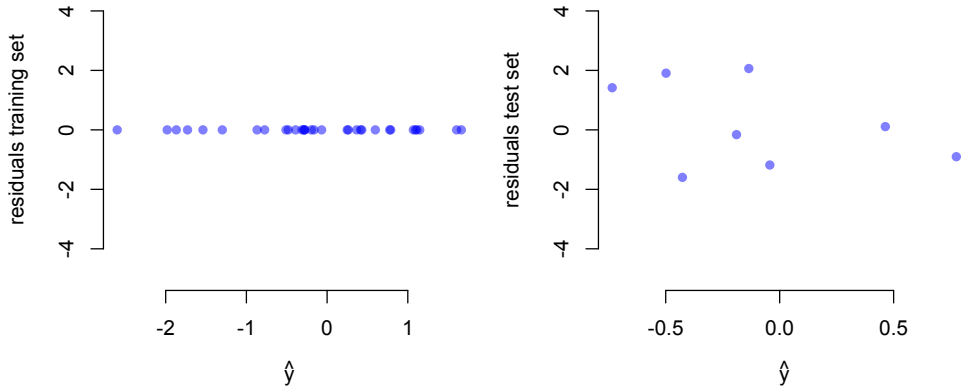


Fig. D.1.

of non-zero coefficients is 5 but in the overparameterised case we use $p = 64$. As seen, the residuals are 0 and so $\hat{Y}_S = \sum_{j \in S} X_j \beta$ is the same as $\hat{Y}_J = \sum_{j \in J} X_j \hat{\beta}_J$, where $\hat{\beta}$ is the ridge estimate for model J and model J contains all 64 predictors. Because the observations are predicted exactly, it was assumed



(a) (b)

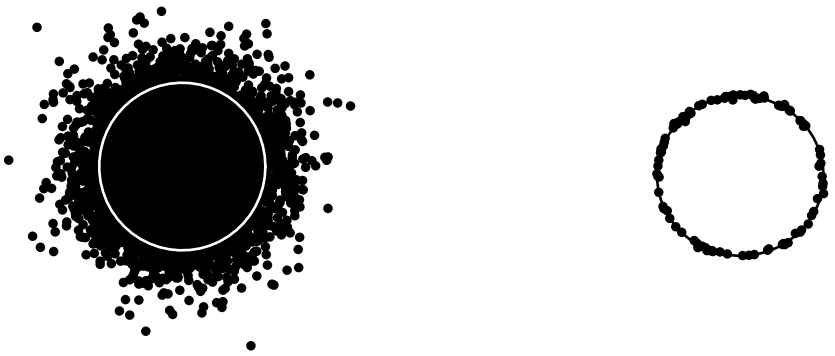
Fig. D.2. Residuals of the training set (a), showing that the model with 64 predictors (while correct is 5) is interpolating the data, fitting exactly each point. The residuals of the test set (b) remain relatively small, indicating that the variance need not explode at interpolation.

that the variance of the residuals on a test set would be large. Figure D.2(b) shows that this is not the case. There is some variance in the residuals of the test set but not too much. The noise that ends up in the predictors is apparently hidden in the unimportant directions of the predictors (Bartlett et al., 2020).

E High-dimensional space

For some intuition on high-dimensional spaces, the isotropic Gaussian, i.e., with mean 0 and variances of 1 and covariances of 0, is convenient. Let X be a Gaussian random variable of dimension p with mean 0 and isotropic covariance matrix $\mathbb{E}(XX^\top) = \Sigma$ of dimensions $p \times p$. We will consider an inequality that shows that with high probability, for large dimension p most of the draws from an isotropic distribution are at $\sqrt{p}$, the radius of the spheroid of the distribution. This is shown in Figure E.1(a) and (b). In Figure E.1(a) we see a set of observations from a two-dimensional Gaussian isotropic distribution. Observations are distributed randomly in the circle. A similar set of observations from an isotropic Gaussian distribution with dimension $p = 20$ is seen in Figure E.1(b). Most observations are near the surface of the circle (this is a projection on a two-dimensional space).

We can make this clear by considering the probability of an observation falling within a certain bound of the origin (Vershynin, 2018, Chapter 3). Suppose we are interested in the length (norm) of the vector X in p dimensions, i.e., we want to know the probability of $||X||_2 = \sqrt{X_1^2 + X_2^2 + \cdots X_p^2}$ being



(a) (b)

Fig. E.1. In (a) is a two-dimensional Gaussian distribution, showing that there is no preference of direction, as expected from an isotropic random variable. In (b) is a high-dimensional analogous version (a projection on a two-dimensional space). Here most of the observations from the isotropic Gaussian distribution is around the surface at $\sqrt{p}$.

smaller than $\sqrt{p}$. It is intuitive that the length of X is indeed $\sqrt{p}$ because

$$\mathbb{E}||X||_2^2 = \mathbb{E}\sum_{j=1}^{p} X_j^2 = \sum_{j=1}^{p} \mathbb{E}X_j^2 = p$$

So, we should expect many of the values of the square $||X||_2^2$ to be around $\sqrt{p}$. Suppose we are considering the probability of X being near 0, i.e., $||X||_2 \approx 0$. Then we find that

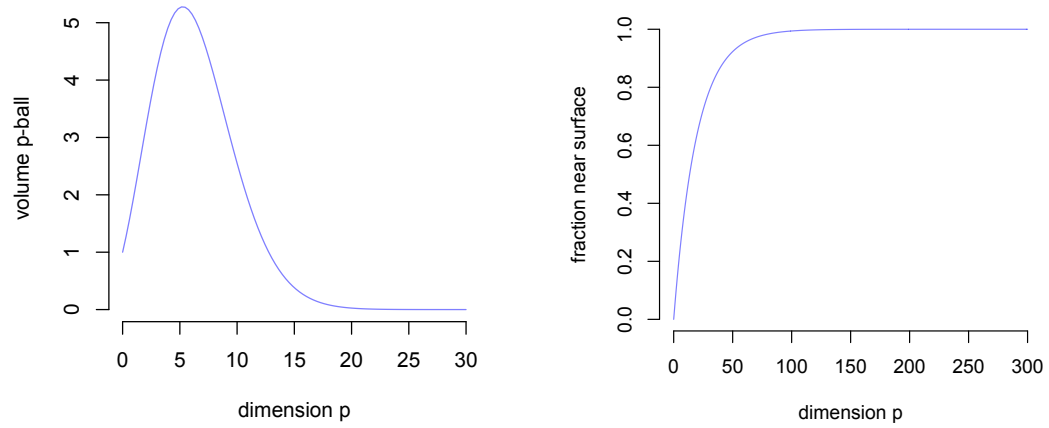
$$\mathbb{P}(|||X||_2 - \sqrt{p}| \leq t) \approx \mathbb{P}(\sqrt{p} \leq t)$$

So, for small values t, say 1 or 2, the probability can only be high if p is small.

Another way to consider this is by looking at the volume of a p-dimensional ball. The volume of the p-dimensional ball in Euclidean space $\mathbb{R}^p$ with radius r is (Giraud, 2014, Section 1.2)

$$\mathbb{B}^p(r) = \frac{\pi^{p/2}}{\Gamma(p/2+1)} r^p \approx \left(\frac{2\pi e r^2}{p}\right)^{p/2} \frac{1}{p\pi} \quad \text{for large } p \tag{E.1}$$

where Γ is the Gamma function $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ and $\alpha > 0$. Figure E.2(a) shows that the volume of the ball is close to 0 already when the dimension is about 20. Moreover, Figure E.2(b) shows that most of the points from the Gaussian isotropic distribution will be near the surface of the p-dimensional ball $\mathbb{B}^p(\sqrt{p})$. The figure shows the fraction of the points outside an inner p-dimensional ball of radius $0.95r$, so very close to the ball of radius r. The fraction increases to 1 exponentially fast (Giraud, 2014). Already at $p = 30$



(a) (b)

Fig. E.2. In (a) we see that, indeed, the volume of a p-dimensional ball with radius 1 in high dimensions gets smaller and smaller, already are $p = 20$ is the volume close to 0. In (b) is the consequence of this, where we see that the fraction of observations from an isotropic distribution near the surface increases exponentially to 1.

the fraction in the last 5% of the ball is nearly 80%. This shows that most of the points will be near the surface of the p-dimensional ball.

F Geometry and metric entropy of Gaussian process

We discuss some properties of the linear model in terms of geometry to explicate its complexity. This will connect to the model and overparameterisation to the reduced test variance.

We have the linear model $X\beta$ where X is an $n \times p$ matrix and β is a p-dimensional vector. Complexity of a model can be described by the volume of the space it describes. If the volume is large then the bias will be small but the test variance will be high. On the other hand, a low volume corresponds to a larger bias and a low test variance (Cheema and Sugiyama, 2020; Grünwald, 2007; Wainwright, 2019). The volume of the linear model is

$$\text{vol}(X\beta) = \sqrt{\det(XX^\top)}\text{vol}(\beta)$$

and $\text{vol}(\beta) = \mathbb{B}^n(\sqrt{n}||\beta||_2)$, the n-dimensional ball with radius $\sqrt{n}||\beta||_2$. We therefore see that model complexity becomes high if $||\beta||_2$ becomes high. We must therefore constrain $||\beta||_2$ in order to reduce model complexity. This is what we do with ridge regression. But the parameter α must be large enough so that the volume does not grow too large.

Compared to the volume of the noise $\mathbb{B}(\sqrt{n\sigma^2})$, we obtain the ratio

$$\frac{\text{vol}(X\beta)}{\text{vol}(e)} = \sqrt{\det(XX^\top)}\frac{\mathbb{B}^n(\sqrt{n}||\beta||_2)}{\mathbb{B}^n(\sqrt{n\sigma_e^2})}$$

An n-dimensional ball $\mathbb{B}^n(r)$ with radius r can be written in terms of the n-dimensional ball with radius 1: $\mathbb{B}^n(1)r^n$. Hence we find for the ratio of volumes $\text{vol}(X\beta)/\text{vol}(e)$

$$\sqrt{\det(XX^\top)}\left(\frac{||\beta||_2^2}{\sigma_e^2}\right)^{n/2} = \sqrt{\det(\alpha^{-1}XX^\top)}$$

where $\alpha^{-1} = ||\beta||_2^2/\sigma_e^2$ is the signal-to-noise-ratio (SNR).

It turns out that when $p > n$ we obtain $X^\top(XX^\top)Y$ as the so-called minimum norm estimate $\hat{\beta}^{MN}$ of β (Ben-Israel and Greville, 1974; Bartlett et al., 2021). The minimum-norm solution is related to the ridge estimator as follows

$$\lim_{\alpha\to 0^+}(X^\top X + \alpha I)^{-1}X^\top = X^\top(XX^\top)^{-1}$$

Intuitively, the minimum-norm solution is obtained for the smallest α that makes the augmented inverse of $X^\top X$ possible. It is therefore reasonable to substitute for $XX^\top$ the ridge version $X^\top X + \alpha I$, where α is small. Hence, we obtain the volume ratio $\text{vol}((X^\top, \sqrt{\alpha}I)^\top\beta)/\text{vol}(e)$ with the ridge version (Cheema and Sugiyama, 2020)

$$\sqrt{\det(\alpha^{-1}XX^\top + I_n)} = \sqrt{\det(X^\top X + \alpha I_p)}$$

by the Weinstein–Aronszajn identity. Because $\alpha = 1/\text{SNR} = \sigma_e^2/||\beta||_2^2$, we obtain small α, close to the minimum-norm estimate, if we have high signal $||\beta||_2^2$.

To show the impact of the geometry on the variance and the bias squared (and hence the MSE), we will rewrite the variance and bias squared from Appendix D. The variance can further be reduced to a simpler version if we assume that $\Sigma = \sigma_\xi^2 I$ (i.e., is isotropic). We then obtain the test variance

$$V(J) = \sigma_\xi^2\sigma_e^2\text{tr}(W(\alpha)^{-2}X^\top X)$$

where $W(\alpha) = X^\top X + \alpha I$. Noting that the eigenvalue decomposition for $W(\alpha)$ is $U\Lambda(\alpha)U^\top$ with diagonal matrix $\Lambda(\alpha)$ with elements $\lambda_j + \alpha$, where λ_j is an eigenvalue of $X^\top X$. The squared inverse $W(\alpha)^{-2}$ is then $U\Lambda(\alpha)^{-2}U^\top$ with diagonal matrix $\Lambda(\alpha)^{-2}$ with elements $1/(\lambda_j + \alpha)^2$. Then we obtain for the variance

$$V(J) = \sigma_\xi^2\sigma_e^2\sum_{j=1}^{n}\frac{\lambda_j}{(\lambda_j + \alpha)^2}$$

Hence, by increasing α we see that we are lowering the variance $V(J)$, and hence the test MSE. Additionally, we can observe that if the contribution of the eigenvalues λ_{d+1} up to λ_p, for some $d < p$, is small, then the variance will not increase, and hence, prediction may still be accurate. This argument is made precise in Bartlett et al. (2020, 2021).

Similar to the test variance $V(J)$ the bias $B(J)$ can also be reduced to a more amenable form when we assume that $\Sigma = \sigma_\xi^2 I$. With this assumption we obtain

$$B(J) = \sigma_\xi^2 \text{tr}((I - W(\alpha))^{-1} X^\top X)^2 \beta\beta^\top) \tag{F.1}$$

Because $W(\alpha)$ is non-singular (given α large enough) and $W(\alpha)$ and $\beta\beta^\top$ are symmetric, we can diagonalise both $W(\alpha)$ and $\beta\beta^\top$ simultaneously (Prasolov, 1994, Section 20). Letting λ_j be the eigenvalues of $X^\top X$ and let γ_1 be the single positive eigenvalue of $\beta\beta^\top$ (all other eigenvalues are 0 because it is a rank 1 matrix). Then we obtain for the bias

$$B(J) = \sigma_\xi^2 \frac{\gamma_1 \alpha}{(\lambda_1 + \alpha)^2} + \sigma_\xi^2 \sum_{j=2}^{n} \frac{\alpha}{(\lambda_j + \alpha)^2} \tag{F.2}$$

This shows a slightly different behaviour for the bias (squared) than the test variance $V(J)$: the bias *could* be increased by increasing α. Since $\alpha = \sigma_e^2 / ||\beta||_2^2$, we see that by constraining $||\beta||_2$ we could increase the bias.

We can now conclude from this analysis that there are two views on how to manage the test variance.

- (a) We can increase α directly, reducing the test variance $V(J)$, and hence obtaining reasonable test MSE. This leads indirectly to a lower SNR (because $\alpha = 1/\text{SNR}$); or
- (b) we can constrain the size of $||\beta||_2$ in the ridge estimator, therefore, decreasing the model complexity (volume) of the model. This will reduce SNR and hence, increase α, leading to a decrease in the test variance $V(J)$. We do this by constraining the ridge estimator with small c such that $||\beta||_2 \leq c$.

In both cases (a) and (b) we are either directly or indirectly constraining the total signal $||\beta||_2$. In model selection this can be done by either increasing α, equivalently, decreasing c in the ridge constraint $||\beta||_2 \leq c$, or by incorporating the complexity (volume) of the model (the n-dimensional ball $\mathbb{B}^n(||\beta||_2)$).

From the variance equation $V(J)$ we also see that if the last $n - k$, say, eigenvalues are together (summed up) κ, then the variance $V(J)$ is dominated by the sum of the first k eigenvalues $\lambda_j/(\lambda_j + \alpha)^2$. This implies that adding more predictors (making p larger) will not affect the variance $V(J)$ much, which is stabilised by the ridge parameter α.

G Akaike information criterion

For the derivation of the Akaike information criterion (AIC), we follow Claeskens and Hjort (2008). The AIC can be derived from the Kullbeck-Leibler divergence (KL). The KL is a way to quantify the overlap between two distributions. The KL is 0 if the distributions are the same (almost surely) and is > 0 if they are not the same. Hence, the AIC aims to minimise the KL across the different models.

Let f and g be two densities, where we designate g as the true underlying distribution and we use f as an approximation. In practice we estimate a parameter $\hat{\beta}$ that we plug-in f, denoted by $f(\cdot, \hat{\beta})$. The KL is for f and g defined as (Kullback and Leibler, 1951; Cover and Thomas, 2006)

$$\mathrm{KL}(f, g) = \mathbb{E}_{Y|x} \log g(Y \mid x) - \mathbb{E}_{Y|x} \log f(Y \mid x, \hat{\beta})$$

where $\mathbb{E}_{Y|x}$ indicates the expectation over Y conditional on x (the values of the predictors in the nodewise GGM). Across different models the term $\mathbb{E}_{Y|x} \log g(Y \mid x)$ is fixed and so we concentrate on the second term. So, maximising the second term of the KL equals minimising the KL itself. The expectation of this second term is

$$L = \mathbb{E}_X \left(\mathbb{E}_{Y|x} \log f(Y \mid x, \hat{\beta}) \right)$$

The AIC tries to estimate this expectation and then select the model with the highest value (and hence the minimal value of KL). This expectation is estimated by the log-likelihood $\frac{1}{n} L(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} f(Y_i \mid x_i, \hat{\beta})$. And it can then be shown that this estimate is biased and needs to be corrected, since

$$\mathbb{E}(\tfrac{1}{n} L(\hat{\beta}) - L) \approx \frac{p}{n} \quad \text{and, hence} \quad \frac{1}{n}(-2L(\hat{\beta}) + 2p)$$

Discarding the $1/n$ leads to the standard AIC.

Here, for the GGM we used the Gaussian distribution for f. Then (Seber and Lee, 2012)

$$\sum_{i=1}^{n} \log f(Y_i \mid x_i, \beta) = \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(Y_i - x_i^\top \beta)^2 - \frac{n}{2} \log \sigma^2 - \frac{n}{2} \log 2\pi$$

Plugging in our ridge estimate $\hat{\beta}_J$ for model J (see Appendix C), multiplying by -2 and ignoring constants, we obtain the AIC

$$\mathrm{AIC}(J) = n \log \hat{\sigma}_J + 2p_J$$

where p_J is the number of parameters for model J and

$$\hat{\sigma}_J^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \hat{\beta}_J)^2$$

is plugged in for σ^2. The standard AIC has a fixed (independent of the data) value of the ridge parameter α. The AIC-CV has a ridge parameter α that has been obtained with k-fold cross validation.

H Minimum description length

The minimum description length (MDL) is derived from the idea that compression (encoding) of data can be translated into probabilities (Grünwald et al., 2005). MDL is concerned with the length of (prefix or invertible) optimal codes which can be separated into a part where the data are encoded given the model and a part where the model is encoded. The length of the data can be encoded by the log likelihood (Hansen and Yu, 2001), i.e. $\text{length}(Y, x) = -\log f(Y \mid x, \hat{\beta})$, where f is the conditional density of Y given the predictors x (see Appendix G). Additionally, the model is encoded, called the complexity, which leads to a specific formulation depending on the type of encoding (Grünwald, 2007). For normally distributed random variables the MDL can be approximated by (Myung et al., 2006)

$$\text{MDL}(J) = \underbrace{\frac{n}{2}\log\hat{\sigma}_J^2}_{\text{data code length}} + \underbrace{\frac{1}{2}\text{tr}(S_J)\log n + \log\int_B \sqrt{\det(F_J)}d\beta}_{\text{model code length}} + o(1)$$

where $\text{tr}(S_J)$ is the trace of matrix S_J representing the effective number of parameters for model J, and F_J is the Fisher information matrix for model J. For the case $n > p$ and linear regression we obtain $F_J = X^\top X/\sigma^2$. The matrix S_J is used to determine the number of effective parameters. In the case $p < n$ we have that $S_J = X(X^\top X)^{-1}X^\top$, and the trace of this matrix is p (this is the same as the rank in this case). However, for the case $p > n$ we find that $X^\top X$ is singular, implying that $\det(X^\top X) = 0$. This is because the rank of $X^\top X$ is $\min\{n, p\}$, and because $n < p$, we have that the rank in this case is n, while the dimensions are $p \times p$. Thus, we must have that there are $p - n$ dimensions that are not represented by $X^\top X$ (i.e., $p - n$ vectors project onto the kernel or null space). We can use ridge regression to alleviate the problem by using instead of $X^\top X$ the matrix from ridge regression $X^\top X + \alpha I$ (Hastie et al., 2001). We then find that $S_J = X(X^\top X + \alpha I)^{-1}X^\top$. This gives the general equation for the number of effective parameters (Hastie et al., 2001, Section 7.6)

$$df(\alpha) = \text{tr}(S_J) = \text{tr}X(X^\top X + \alpha I)^{-1}X^\top = \sum_{j=1}^{\min\{p,n\}} \frac{\lambda_j}{\lambda_j + \alpha}$$

where λ_j are the eigenvalues of $X^\top X$.

To approximate the integral term we make use of the fact that in linear regression the Fisher information is independent of the parameter β_j, and

hence the integral represents the open p-dimensional ball $\mathbb{B}^p(||\hat{\beta}||_2)$. The p-dimensional ball is centered at $||\hat{\beta}||_2$ because in the ridge solution we obtain the solution such that $\alpha(||\beta||_2^2 - c) = 0$ (see Appendix C). We then obtain the MDL (Cheema and Sugiyama, 2020)

$$\log \int_B \sqrt{\det(F)} d\beta \approx \log \det(X^\top X + \alpha I) + \log \mathbb{B}^p(||\beta||_2)/\sigma^p$$

The p-dimensional ball in the last term, $\log \mathbb{B}^p(||\beta||_2)$, was investigated in Appendix E. The log of this term will become large, and hence apply a stronger penalty with increasing dimension. This is why increasing the dimension can have positive effects on the generalisation error (variance in the test set).

By the volume of the p-dimensional ball in (E.1) we can therefore approximate the MDL by

$$\text{MDL} = \frac{n}{2} \log \hat{\sigma}^2 + \frac{1}{2} \text{tr}(\hat{S}) \log n + \log \det(X^\top X + \alpha I) + \log V_p(||\hat{\beta}||_2)/\sigma^p$$

This version of MDL is used in the simulations and is referred to as MDL.

The complexity of MDL has also been used without the integral term (Rissanen, 1986; Grunwald, 2000). This approximation is similar to the Bayesian information criterion (BIC) introduced by Schwartz (1978). Hence, we call this version MDL-S.

I Minimum description length: optimised

The minimum description length (MDL) principle has been extended to the high-dimensional setting of $n < p$ in another way than in Appendix H. The principle of approximating the code length for the data given the model and the code length for the model is the same, but in order to account for the high-dimensions (such that $d/n \to \gamma < \infty$), the effective degrees of freedom are used in combination with an optimal α for the effective degrees of freedom (Dwivedi et al., 2020). The MDL is then obtained by minimising a function for α that gives the smallest MDL value. The function to be optimised over α is

$$\text{MDL-opt} = \frac{n}{2} \log \hat{\sigma}^2 + \frac{1}{2\hat{\sigma}^2} ||\hat{\beta}||_2^2 + \sum_{j=1}^{\min\{n,d\}} \log \left(1 + \frac{\lambda_j}{\alpha}\right)$$

where λ_j are the eigenvalues of $X^\top X$ and $\hat{\beta}$ is the ridge estimator with value α. The second term is included here because of the Bayesian interpretation of MDL in Dwivedi et al. (2020), where this term is entered as a prior for β. This version of MDL is referred to as MDL-opt.

J Simulation details

In the case where the true model is linear we have $f(X) = X\beta$. In the misspecified case we used instead of the linear model the sigmoid function $f(X) = 1/(1 + \exp(X\beta))$. Again we keep $d = 5$ non-zero coefficients fixed. The coefficients β were normalised so that the L_2 norm was 1, i.e., $||\beta||_2 = 1$.

To generate data Y we added Gaussian uncorrelated noise e to the model, so that $Y = f(X)+e$. The standard deviation of the noise was set to 0.5 (SNR of 2) and 1 (SNR of 1). The signal-to-noise ratio (SNR) is defined as $||\beta||_2^2/\sigma_e^2$, the ratio of the variance of the signal (coefficients) and the noise variance. Since we fix $||\beta||_2^2 = 1$, the variance of the noise determines the SNR, set to either 1 or 2. The number of generated observations is fixed at $n = 40$.

For estimation we selected a fraction of 0.80 for training (estimation) of parameters and the remaining 0.20 for testing. We fix $n = 40$ observations in total, and so obtain $n_{\text{train}} = 32$ and $n_{\text{test}} = 8$. We fix the dimension of the true model at $d = 5$ and vary the number of parameters p from 3 to 64. Hence we obtain the ratio p/n for training data from 0.09 to 2, where the last setting means we have two parameters per observation. With fixed $d = 5$ we have that the correct ratio p/n is at $5/32 = 0.15625$.

The ridge estimate is either obtained with parameter $\alpha = 0.0001$, or the ridge estimate is obtained with α estimated with 10-fold cross-validation, with the α with the smallest MSE. The Lasso is obtained with parameter α from 10-fold cross-validation. Estimation was performed with the `R` package `glmnet`.

The measure we used to evaluate the model selection criteria was the proportion of correct decisions, i.e., the correct set of coefficients divided by the number of runs (which was $R = 100$).

$$\text{proportion correct} := \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}\{\hat{S} = S\}$$

where $\hat{S}$ is the set of indices for the estimated non-zero coefficients (support) and S is the true support. Figure J.1 shows the results for all models J (all values p) in the linear case (true model known) with SNR is 2. It shows that the MDL and AIC-CV mostly obtain the correct model (at the grey dashed line at $p/n \approx 0.15$), but the AIC-CV tends to overfit somewhat. The MDL-S and AIC overfit vastly due to the low penalty of only the number of parameters.

For some more detail on the performance of the model selection criteria we plot in Figure J.2 the individual performance (blue) and average performance (red) for each of the methods and include the proportion correct. This is the linear model with SNR is 2. It is clear that MDL, MDL-opt and AIC-CV show that they are robust against high-dimensional scenarios, while MDL-S (BIC) and AIC are not.

For the misspecified model (Figure J.3(b)) we see that when p/n is about 10, then the test MSE becomes lower (0.0844) than when p/n is at the correct value ≈ 0.15 and the test MSE is 0.0877. This is not the case when the model

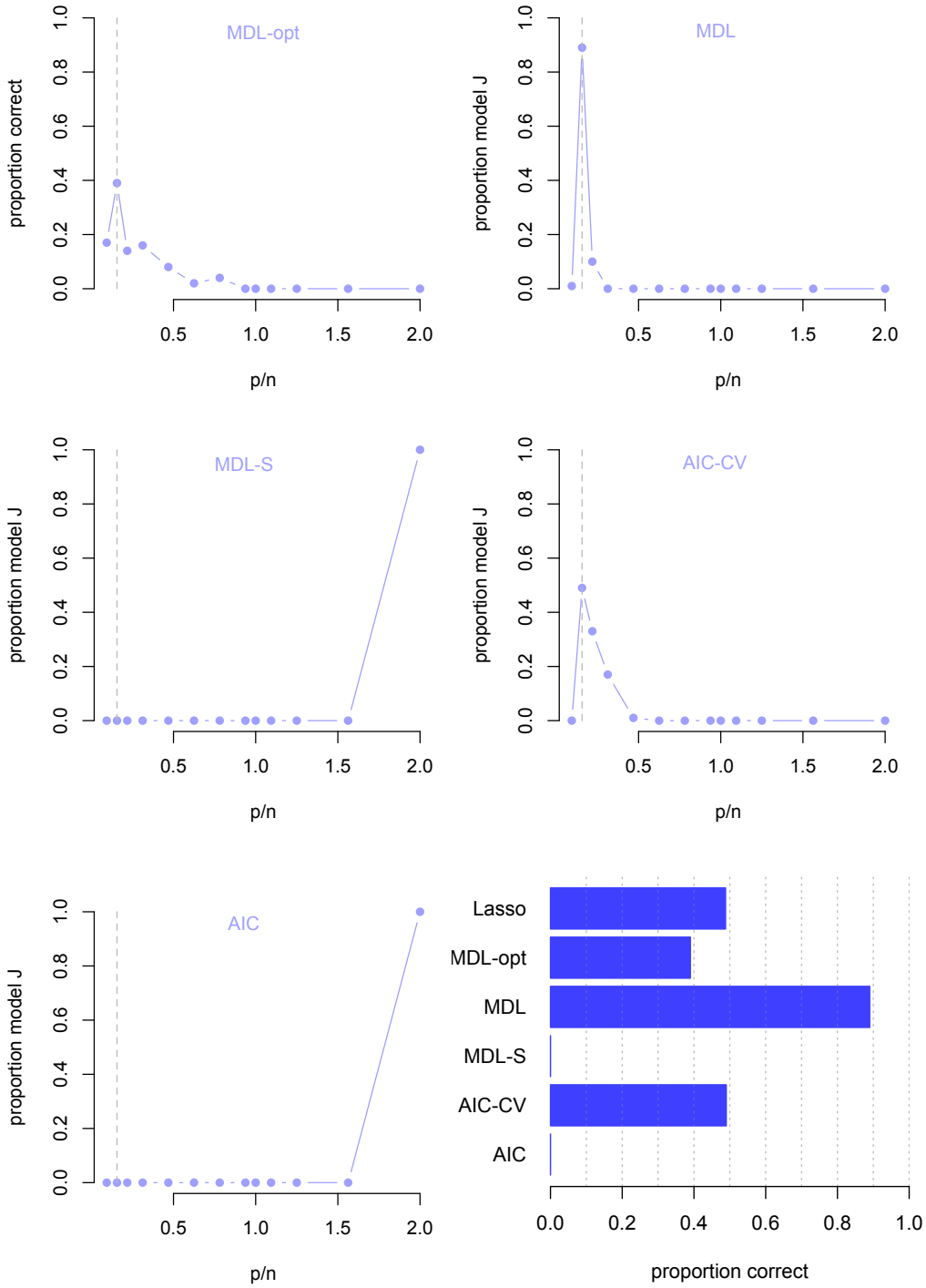


Fig. J.1. Proportion of selected model J for each of the methods. The grey dashed line at $p/n \approx 0.15$ represents the correct model. Left of this line represents underfit and right from this line represents overfit. In the right bottom panel is the proportion correct for each of the methods, as displayed in Figure 3 in the main text.

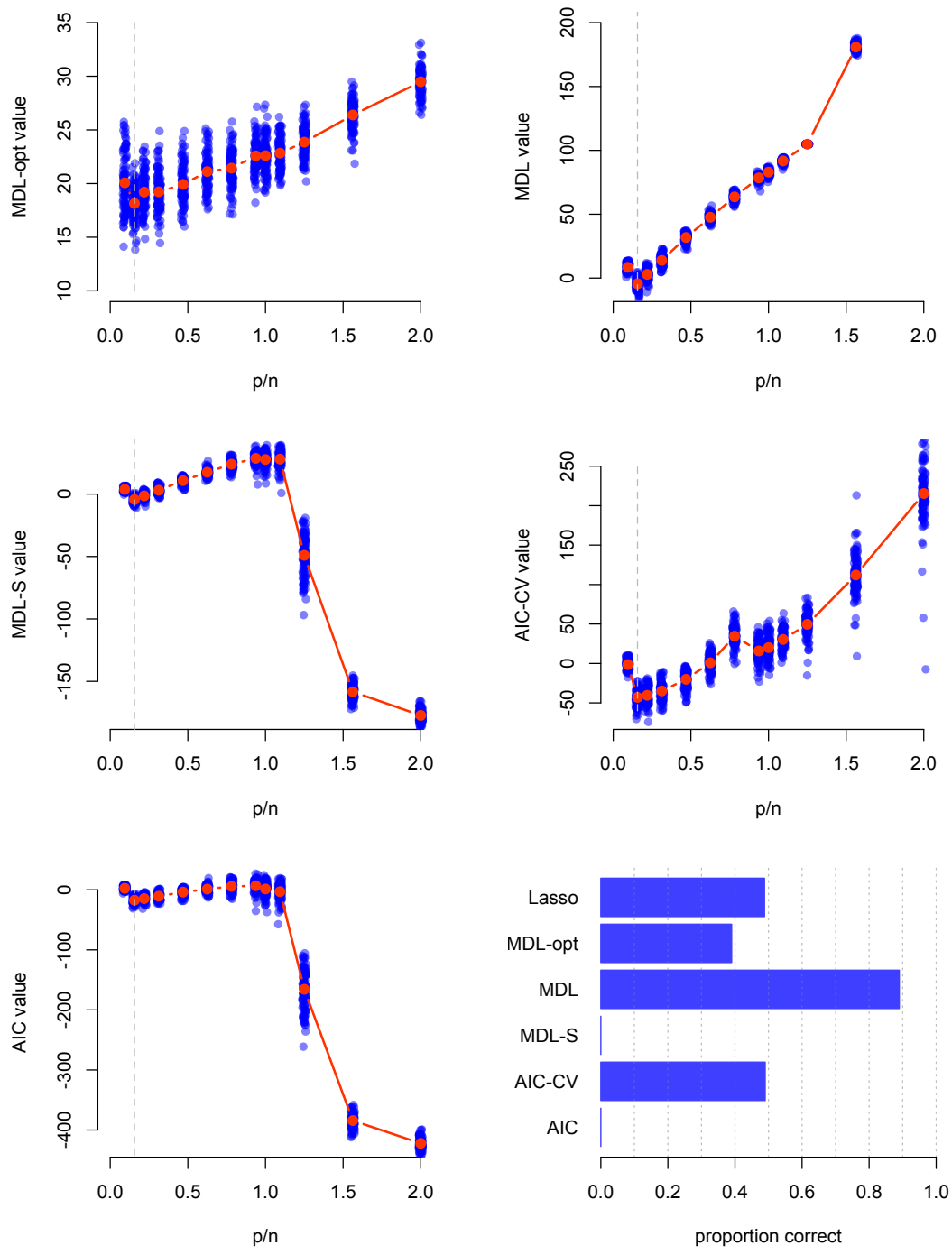


Fig. J.2. Values of the methods for selecting the number of connected nodes of a single node for 100 runs (blue) and the average (red). The data had a signal-to-noise ratio of 2. The true model had 5 connections and is indicated by the vertical dashed gray line at $p/n \approx 0.15$, where the lowest value should occur. In the bottom right panel is the proportion correct selections for each of the methods.

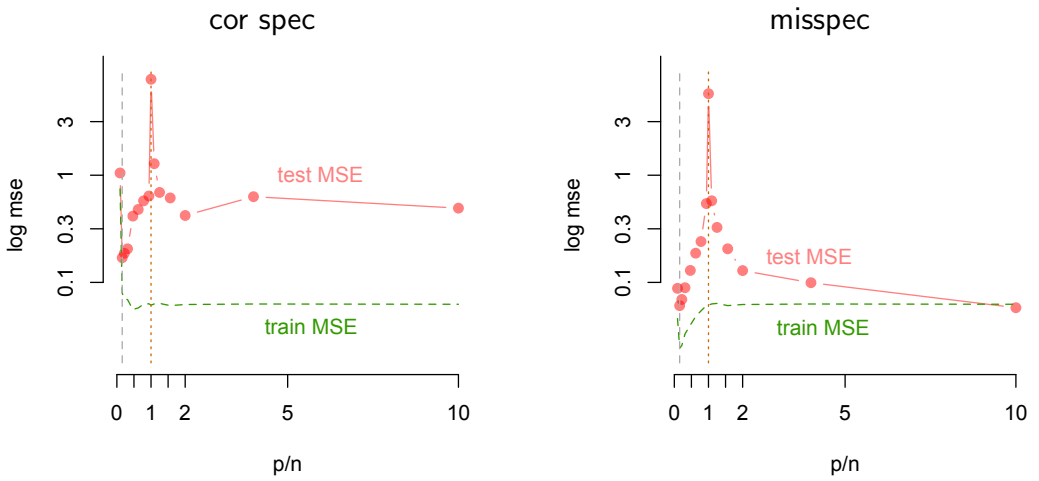


Fig. J.3. MSE for correctly specified (left) and misspecified model (right) with p predictors up to 320, so that p/n ranges between 0.1 and 10. For the true linear model (left) the global minimum of the MSE remains the correct model with $p = 5$, approximately at $p/n \approx 0.15$. For the misspecified model, however, we see that the global minimum has now shifted from the correct model with $p = 5$ and MSE 0.0877 to the model with $p = 320$ and MSE 0.0844.

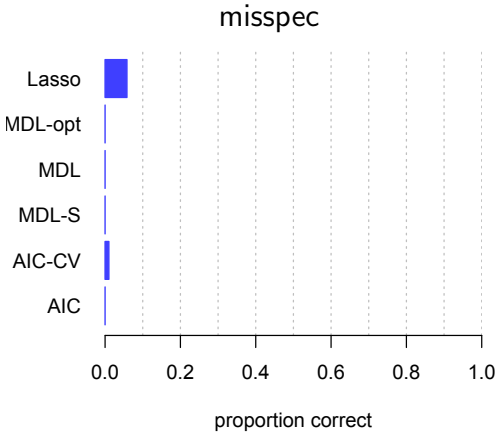


Fig. J.4. Proportion correctly identified number of predictors when the model is misspecified with p predictors up to 320, so that p/n ranges between 0.1 and 10. The SNR is 2.

is correctly specified as linear (see Figure J.3(a)).

We also investigated two scenarios where the covariance matrix Σ of the predictors is not proportional to the identity matrix, i.e., the predictors were correlated. In Figure J.5(a) are the proportions correct for Σ had equal correlations of 0.4 for all predictors. It shows, in line with the results of Hastie et al. (2019), that this scenario seems to help correct recovery. On the other hand, random correlations across Σ, i.e., correlations uniform between 0.1 and 0.8 between 20% of the predictors, resulted in somewhat poorer overall per-

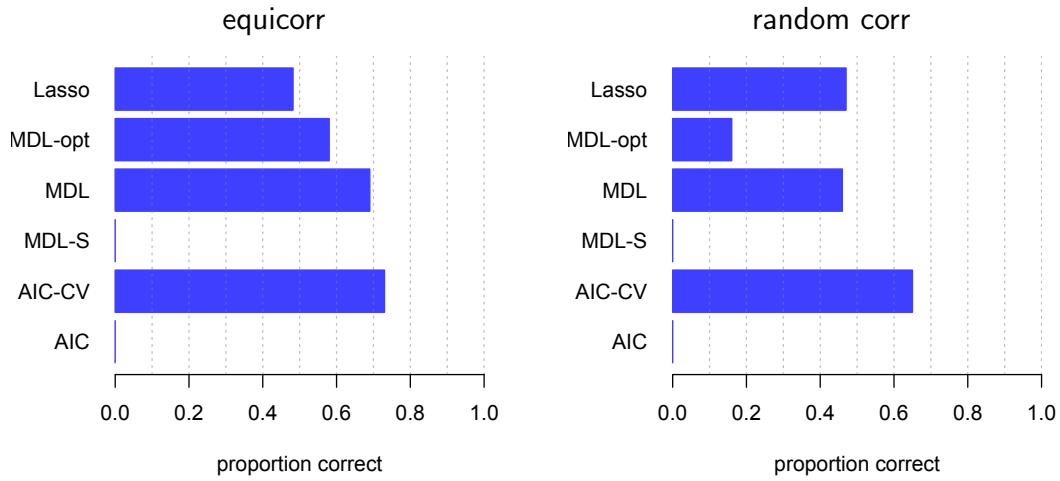


Fig. J.5. Proportions correct of the different model selection procedures when the predictors all were correlated 0.4 (left) or where the correlations were random between 0.1 and 0.8 for 20% of the predictors (right). In both settings the SNR was 2.

formance, except for the AIC-CV.

References

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Ben-Israel, A. and Greville, T. (1974). *Generalized inverses: theory and applications*. New York: John Wiley and Sons.

Bilodeau, M. and Brenner, D. (1999). *Theory of multivariate statistics*. New York: Springer-Verlag.

Bühlmann, P. et al. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242.

Bühlmann, P., Kalisch, M., and Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Cheema, P. and Sugiyama, M. (2020). Double descent risk and volume saturation effects: A geometric perspective. *arXiv preprint arXiv:2006.04366*.

Claeskens, G. and Hjort, N. (2008). *Model selection and model averaging*. Cambridge University Press, Cambridge.

Cover, T. and Thomas, J. (2006). *Elements of information theory*. Wiley and Sons, 2nd edition.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558.

Dwivedi, R., Singh, C., Yu, B., and Wainwright, M. J. (2020). Revisiting complexity and the bias-variance tradeoff. *arXiv preprint arXiv:2006.10189*.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American statistical Association*, 81(394):461–470.

Epskamp, S., Borsboom, D., and Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50(1):195–212.

Epskamp, S. and Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological methods*, 23(4):617.

Giraud, C. (2014). *Introduction to high-dimensional statistics*, volume 138. CRC Press.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Gruber, M. (1990). *Regression Estimators: A Comparative Study*. Academic Press, Boston.

Grunwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, 44:133–152.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.

Grünwald, P. D., Myung, I. J., and Pitt, M. A. (2005). *Advances in minimum description length: Theory and applications*. MIT press.

Hansen, M. H. and Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774.

Haslbeck, J. M. and Waldorp, L. J. (2020). mgm: Structure estimation for time-varying mixed graphical models in high-dimensional data. *Journal of Statistical Software*, 93(8).

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Jones, B. and West, M. (2005). Covariance decomposition in undirected gaussian graphical models. *Biometrika*, 92(4):79–786.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of mathematical statistics*, 22:79–86.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press, Oxford.

Lockhart, R., Taylor, J., Tibshirani, R. J., Tibshirani, R., et al. (2014). A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468.

Magnus, J. R. and Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and economics, revised edition*. Chichester: John Wiley & Sons.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488).

Myung, I. and Pitt, M. (1998). Issues in selecting mathematical models of cognition. In Grainger, J. and Jacobs, A., editors, *Localist connectionist approaches to human cognition*, pages 327–355. Lawrence Erlbaum Associates.

Myung, J. I., Navarro, D. J., and Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, 50(2):167–179.

Pötscher, B. M. and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *Journal of Multivariate Analysis*, 100(9):2065–2082.

Prasolov, V. V. (1994). *Problems and theorems in linear algebra*, volume 134. American Mathematical Soc.

Rao, C. (1990). *Linear statistical inference and its applications*. John Wiley and Sons, second edition edition.

Rao, C. and Toutenberg, H. (1999). *Linear models: least squares and alternatives*. Springer.

Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics*, 14:1080–1100.

Schott, J. R. (1997). *Matrix analysis for statistics*. New York: John Wiley & Sons.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Searle, S. (1971). *Linear Models*. John Wiley and Sons, New York.

Seber, G. A. and Lee, A. J. (2012). *Linear regression analysis*, volume 936. John Wiley & Sons.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., and Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific reports*, 4.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymp-

totically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.