Information criteria for structured parameter selection in high dimensional tree and graph models

Maarten Jansen
Université libre de Bruxelles, departments of Compter Science and Mathematics

January 20, 2024

Abstract

Parameter selection in high-dimensional models is typically finetuned in a way that keeps the (relative) number of false positives under control. This is because otherwise the few true positives may be dominated by the many possible false positives. This happens, for instance, when the selection follows from a naive optimisation of an information criterion, such as AIC or Mallows's Cp. It can be argued that the overestimation of the selection comes from the optimisation process itself changing the statistics of the selected variables, in a way that the information criterion no longer reflects the true divergence between the selection and the data generating process. In lasso, the overestimation can also be linked to the shrinkage estimator, which makes the selection too tolerant of false positive selections. For these reasons, this paper works on refined information criteria, carefully balancing false positives and false negatives, for use with estimators without shrinkage. In particular, the paper develops corrected Mallows's Cp criteria for structured selection in trees and graphical models.

Keywords: high-dimensional; sparsity; lasso; variable selection; information criterion

1 Introduction

Whereas statistical inference takes place after the estimation of the parameters in a model, model and parameter selection is a step that in principle precedes and in some cases includes the estimation of the parameters. Statistical inference operates in an asymmetric way w.r.t. the null and alternative hypotheses, thereby reflecting the presumption of innocence. In parameter selection, the decisions whether or not to include parameters in the model are based on symmetric criteria, which can be compared to profiling in a criminal investigation.

In all three steps, parameter selection, estimation and inference, the notion of likelihood may be adopted, although each time in a different role. In inference, the likelihood ratio can be used to test the significance of a larger model against a smaller. An estimation procedure assumes that the model being worked in is correct, or at least it has been selected as the least false model to work in for the statistical inference problem at hand. In this framework, maximising the likelihood function amounts to finding the parameters that make the data fit as well as possible into the proposed model. In a parameter selection, candidate models can be assessed through the capability of modelling or predicting the same sort of data as those that have been observed. Working with maximal likelihood would promote large models, because they would best fit the current observations, including the noise. The best model for predicting these observations, without noise that is, would minimise the *expected* value of the maximum likelihood with respect to the data generating process (DGP, also referred to as ground truth). The expected likelihood is closely related to the Kullback-Leibler (KL) divergence between the proposed model and the DGP. Other, similar divergences between the proposed model and the DGP include the prediction error (PE). PE is basically the KL divergence in a model with additive normal noise, assuming the variance to be known (or easy to estimate outside the parameter selection).

The divergence (i.e., the expected likelihood) depends on the bias and the variance of an estimator in the proposed model. The balance between bias and variance is typically equivalent, at least in expected value, to a closeness-complexity trade-off, formalised in an information criterion. Closeness is typically expressed by the sample likelihood in the model under consideration, while the complexity acts as a penalty compensating for the gap between sample likelihood and the expected likelihood. Akaike's Information Criterion (AIC) [Akaike, 1973] and Mallows's C_p [Mallows, 1973] are prototypes of these information criteria, estimating, respectively, the KL distance and the PE. Other well known criteria include the Bayesian or Schwarz information criterion (BIC) [Schwarz, 1978], Akaike's Final Prediction Error (FPE) [Niedźwiecki and Ciołek, 2017], Exponentially Embedded Families [Kay, 2005], Minimum Description Length (MDL) [Rissanen].

In the last few decades, the interaction between parameter selection and statistical inference has become a major point of attention in statistical research. In one direction, the focus for

which the model is used determines which model performs best [Claeskens and Hjort, 2003]. In the other direction, the parameter selection procedure creates additional uncertainty in the estimators, leading to wider confidence intervals [Berk et al., 2013, van de Geer et al., 2014, Zhang and Zhang, 2014, Lee et al., 2016, Charkhi and Claeskens, 2018].

This paper concentrates on the uncertainty arising from the selection of variables, however not on its effect on the subsequent inference. Instead, it deals with its effect on the statistics of the information criterion used in the process of parameter selection. Indeed, the classical information criteria have been developed for estimating the divergence of a given, fixed model from the DGP. However, if the model under consideration comes from optimising a criterion, that optimisation has interacted with the noise. This interaction affects the statistics of the noise. As a result, the information criterion may not be a good estimator of the divergence after all, and hence, the optimiser of the criterion may point to a suboptimal model in terms of divergence w.r.t. the DGP. The gap between the information criterion and the divergence or error measure can be explained and formalised by the concept of generalised degrees of freedom [Ye, 1998, Hansen and Sokol, 2014, Jansen, 2014].

This paper studies the generalised degrees of freedom in the case of structured parameter selection in trees and graphs. It starts off with a lasso procedure, i.e., a selection and estimation by solving an ℓ_1 regularised least squares problem [Tibshirani, 1996, Chen et al., 1998]. It has been reported [Wainwright, 2009, Tropp, 2006, Zhao and Yu, 2006] that lasso is selection consistent, at least if the DGP is described by a model belonging to the set of models under consideration. Moreover, the parameters in that model are assumed to be sufficiently large, while at the same time the regularisation parameter λ should not be too large, so that for $n \to \infty$ the numbers of false positives and false negatives tend to zero. As these assumptions put quite some restrictions on λ, it is generally impossible to combine minimum prediction errors and selection consistency [Meinshausen and Bühlmann, 2006], which motivates the use of adaptive lasso [Zou, 2006]. In particular, a minimum prediction error choice of λ leaves many false positives resulting in noisy features. This effect can be explained by the lasso shrinkage in two ways. First, the shrinkage reduces the price of false positives in terms of induced variance. As a result, the optimisation of the bias-variance balance is tolerant of their presence in the selection. Second, the shrinkage is responsible for an important bias in the lasso estimator. In order to keep that shrinkage bias under control, the minimum prediction error tends to be achieved for small values of the regularisation λ, which corresponds to large models. Debiasing [Javanmard and Montanari, 2018] or regularisation [Li and Shao, 2015] of the lasso solution leads to minimum prediction error selections with less false positives. Alternative regularisations have also been proposed, reducing the shrinkage in large significant parameters. They include the smoothly clipped absolute deviation (SCAD) [Fan and Li, 2001] and a minimax concave penalty (MCP) [Zhang, 2010]. More recent tree structured sparse estimation has been proposed to be based on deep neural networks [Kim and Chung, 2020].

This paper follows a debiasing approach. It adopts the lasso algorithm, minimising an ℓ_1 regularised sum of squared residuals, for selection purposes only, not for estimation. The built in biased, shrinkage estimation is replaced by a least squares projection onto the selected model. The use of lasso for selection purposes is motivated by the fact that for a given value of λ, the convex optimisation of lasso leads to nearly the same sparsity level as a combinatorially complex ℓ_0 regularisation, leading to the best κ term orthogonal projection [Donoho, 2006]. Undoing the shrinkage after selection has an impact on the bias-variance balance in the KL divergence or in the prediction error, but also on the closeness-complexity balance in the estimation of the error by an information criterion. Indeed, the shrinkage bias occurs mainly at large values of λ, while the variance from the false positives is mainly seen with small values of λ. Undoing the shrinkage takes away the shrinkage bias and increases the impact of false positives on the variance, thus shifting the optimal bias-variance balance towards larger values of λ, meaning smaller models with less variance. Without shrinkage, finding the optimal model size is a more delicate task, because the impact of false positive selections on the variance is no longer tempered by shrinkage. Therefore, a given overestimation of the optimal model size introduces more variance. On the other hand, taking the model too small introduces more bias when there is no shrinkage, because the optimum lies at smaller models, where false negatives occur more frequently.

As for the closeness-complexity balance in the information criterion, a quite remarkable result states that for a linear regression model with normal noise, the number of degrees of freedom of the lasso equals the size of the selected model [Zou et al., 2007, Tibshirani and Taylor, 2012]. It also equals the number of degrees of freedom in a least squares estimation on a fixed model. In other words, with normal errors, the shrinkage bias compensates exactly for the influence of the errors on the optimisation process. This explains why Mallows's C_p for orthogonal projection on a given model has the same form as Stein's Unbiased Risk estimator in soft thresholding [Stein, 1981]. After undoing the shrinkage, the estimation of λ with minimum prediction error or KL divergence requires a tailored information criterion. This is developed in Section 2.

The remainder of the article is organised as follows. Section 2 develops the idea to finetune a shrinkage based sparse selection method (such as lasso) for minimum prediction error when using the selection for orthogonal projection without shrinkage. Section 3 applies the methodology to subtree selection, in particular to regression trees, Section 4 applies the methodology to graphical models representing large multivariate normal random variables.

2 Information criteria for use in parameter selection without shrinkage

Consider the full linear regression model

$$\boldsymbol{Y} = \boldsymbol{\mu} + \sigma \boldsymbol{Z} = \mathbf{X}\boldsymbol{\beta} + \sigma \boldsymbol{Z}, \tag{1}$$

where $\boldsymbol{X}$ is standardised i.i.d. noise, $\boldsymbol{\beta}$ is a sparse m dimensional vector, and $\boldsymbol{Y}$ is a response vector with sample size n. The statement (1) assumes that the DGP is a submodel of the full model. Alternatively, the model in (1) can be considered as a family of approximations to the DGP, from which the member closest to the DGP is defined the least false model. In this definition, the distance between the DGP and the least false model is measured, for instance, by the Kullback-Leibler divergence.

Since in the high-dimensional case m may be larger than n, or otherwise, the design matrix may be fully or nearly singular due to collinearity, an estimator of $\boldsymbol{\beta}$ is searched for in a two steps regularisation procedure. The first step computes a pilot estimator $\breve{\boldsymbol{\beta}}_\lambda$, using a relatively fast selection and estimation procedure. The prototype of a pilot procedure is the lasso, where the estimator is defined by solving the ℓ_1 regularised least squares problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\| + \lambda \|\boldsymbol{\beta}\|_1, \tag{2}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{m} |\beta_j|$. Solving the ℓ_1 regularised least squares problem leads to sample dependent size $\widehat{\kappa}_\lambda$. As an alternative, regularisation can be achieved by fixing the cardinality of the selection, say κ, and from there find the $\widehat{\lambda}_\kappa$ that leads to the best selection S_κ with cardinality κ. The subsequent discussion will therefore consider all quantities as function of κ or indexed by κ instead of λ. Finetuning by κ instead of λ is particularly interesting in selection procedures, other than lasso, that do not rely on an explicit regularisation parameter. An example is developed in Section 3.

The final estimator is then found as $\widehat{\boldsymbol{\beta}}_\kappa = \widetilde{\mathbf{X}}_{\mathrm{S}_\kappa} \boldsymbol{Y}$, where $\widetilde{\mathbf{X}}_{\mathrm{S}}$ is referred to as the analysis matrix, influence matrix or hat matrix, associated with the selection S (for the sake of simplicity in the notations, the subscript κ is omitted in expressions holding for any selection S). A typical choice of the hat matrix $\widetilde{\mathbf{X}}_{\mathrm{S}}$, for a selection S is given by the least squares solution $\widetilde{\mathbf{X}}_{\mathrm{S}} = (\mathbf{X}_{\mathrm{S}}^T \mathbf{X}_{\mathrm{S}})^{-1} \mathbf{X}_{\mathrm{S}}^T$ where $\mathbf{X}_{\mathrm{S}}$ is the submatrix of $\mathbf{X}$ containing all columns $j \in \mathrm{S}$.

Assumption 2.1 *(Projection) It is assumed in this article that $\mathbf{P}_{\mathrm{S}} = \mathbf{X}_{\mathrm{S}} \widetilde{\mathbf{X}}_{\mathrm{S}}$ is a projection, i.e., an idempotent matrix. The projection is not necessarily an orthogonal projection, i.e., $\mathbf{P}_{\mathrm{S}}$ is not necessarily symmetric.*

The procedure includes a finetuning of the regularisation parameter κ, for which the objec-

tive is to minimise the PE of the outcome,

$$\mathrm{PE}(\widehat{\boldsymbol{\beta}}_\kappa) = \frac{1}{n} E\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\kappa\|_2^2.$$

The PE is estimated unbiasedly by the non-studentised version of Mallows's C_p,

$$\Lambda(\widehat{\boldsymbol{\beta}}_\kappa) = \frac{1}{n}\mathrm{SS}_\mathrm{E}(\widehat{\boldsymbol{\beta}}_\kappa) + \frac{2\nu_\kappa}{n}\sigma^2 - \sigma^2, \tag{3}$$

where $\mathrm{SS}_\mathrm{E}(\widehat{\boldsymbol{\beta}}_\kappa) = \|\boldsymbol{e}_\kappa\|_2^2$, with $\boldsymbol{e}_\kappa = \boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_\kappa = \boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\kappa$ the residual vector, and where ν_κ are the generalised degrees of freedom [Ye, 1998], defined and developed by

$$\nu_\kappa = \frac{1}{\sigma^2} E\left[\boldsymbol{\varepsilon}^T(\boldsymbol{\varepsilon} - \boldsymbol{e}_\kappa)\right] = \frac{1}{\sigma^2} E\left[\boldsymbol{\varepsilon}^T(\boldsymbol{Y} - \boldsymbol{\mu} - \boldsymbol{Y} + \widehat{\boldsymbol{\mu}}_\kappa)\right] = \frac{1}{\sigma^2} E\left[\boldsymbol{\varepsilon}^T\widehat{\boldsymbol{\mu}}_\kappa\right]$$

If the nuisance parameter σ^2 is not available or hard to estimate independently from the parameter selection process, then minimisation of (3) can be replaced with a generalised cross validation [Jansen, 2015]

$$\mathrm{GCV}(\widehat{\boldsymbol{\beta}}_\kappa) = \frac{\frac{1}{n}\mathrm{SS}_\mathrm{E}(\widehat{\boldsymbol{\beta}}_\kappa)}{\left(1 - \frac{\nu_\kappa}{n}\right)^2}. \tag{4}$$

If the final estimator were taken to be just the pilot estimator (including the lasso shrinkage, that is), $\widehat{\boldsymbol{\beta}}_\kappa = \breve{\boldsymbol{\beta}}_\kappa$, then the degrees of freedom would be simply $\nu_\kappa = \kappa$, [Zou et al., 2007, Tibshirani and Taylor, 2012]. While the pilot estimator suffers from the above mentioned tolerance of false positives, the minimum GCV shrinkage estimator $\breve{\boldsymbol{\beta}}_{\kappa*}$ can be used to define an estimator of the variance

$$\widehat{\sigma}^2 = \frac{1}{\nu_{\kappa*}}\mathrm{SS}_\mathrm{E}\left(\breve{\boldsymbol{\beta}}_{\kappa*}\right) = \frac{1}{\kappa^*}\mathrm{SS}_\mathrm{E}\left(\breve{\boldsymbol{\beta}}_{\kappa*}\right)$$

for use in the finetuning of the second step of the selection procedure, which is the step leading from the pilot $\breve{\boldsymbol{\beta}}_\kappa$ to the final $\widehat{\boldsymbol{\beta}}_\kappa$. The finetuning in that second step aims at minimising the prediction error or its estimator (3), this time with taking ν_κ to be the degrees of freedom under orthogonal projection without shrinkage.

The degrees of freedom are further developed as

$$\nu_\kappa = \frac{1}{\sigma^2} E\left[\boldsymbol{\varepsilon}^T\mathbf{P}_{\mathrm{S}_\kappa}\boldsymbol{Y}\right] = \frac{1}{\sigma^2} E\left[\sigma\boldsymbol{Z}^T\mathbf{P}_{\mathrm{S}_\kappa}(\boldsymbol{\mu} + \sigma\boldsymbol{Z})\right] = E\left[\|\mathbf{P}_{\mathrm{S}_\kappa}\boldsymbol{Z}\|_2^2\right] + \frac{1}{\sigma}\boldsymbol{\mu}^T E\left[\mathbf{P}_{\mathrm{S}_\kappa}\boldsymbol{Z}\right].$$

If the selection S_κ were independent from the sample, then $E\left[\|\mathbf{P}_{\mathrm{S}_\kappa}\boldsymbol{Z}\|_2^2\right]$ would be equal to $\mathrm{Tr}(\mathbf{P}_{\mathrm{S}_\kappa}) = \kappa$, and $E\left[\mathbf{P}_{\mathrm{S}_\kappa}\boldsymbol{Z}\right]$ would be the zero vector. The interaction between the noise and the selection makes the set S_κ and the vector $\boldsymbol{Z}$ depend from each other, which is precisely the topic of this paper.

Assumption 2.2 *As in Jansen [2014], it is assumed here that the dependence of* S_κ *and* $\boldsymbol{Z}$ *has an effect on the magnitudes of the errors after selection, and not so much on the signs. More precisely, it is assumed that*

$$\nu_\kappa = E\left[\|\mathbf{P}_{\mathrm{S}_\kappa}\boldsymbol{Z}\|_2^2\right] + o\left[\mathrm{PE}(\widehat{\boldsymbol{\beta}}_\kappa)\right] \text{ as } n \to \infty. \tag{5}$$

The offset

$$m_\kappa = (\nu_\kappa - \kappa)\sigma^2/n = E\left[\|\mathbf{P}_{\mathrm{S}_\kappa}\boldsymbol{Z}\|_2^2 - \kappa\right]\sigma^2/n + o\left[\mathrm{PE}(\widehat{\boldsymbol{\beta}}_\kappa)\right] \tag{6}$$

corrects a Mallows's C_p criterion assessing a given, *fixed* model $\Delta(\widehat{\boldsymbol{\beta}}_\kappa) = \frac{1}{n}\mathrm{SS}_\mathrm{E}(\widehat{\boldsymbol{\beta}}_\kappa) + \frac{2\kappa}{n}\sigma^2 - \sigma^2$, for use in parameter selection, as indeed, from (5), $E(\Lambda(\widehat{\boldsymbol{\beta}}_\kappa)) = E(\Delta(\widehat{\boldsymbol{\beta}}_\kappa)) + 2m_\kappa + o\left[\mathrm{PE}(\widehat{\boldsymbol{\beta}}_\kappa)\right]$, keeping in mind that $E\left(\Lambda(\widehat{\boldsymbol{\beta}}_\kappa)\right) = \mathrm{PE}(\widehat{\boldsymbol{\beta}}_\kappa)$. The offset m_κ thus describes the effect of the selection procedure onto the degrees of freedom. Furthermore, the correction can be seen as a double reflection [Jansen, 2014], on both sides of a "mirror" function $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}_\kappa})$, in the sense that

$$m_\kappa = \mathrm{PE}(\widehat{\boldsymbol{\beta}}_\kappa) - \mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}_\kappa}) + o\left[\mathrm{PE}(\widehat{\boldsymbol{\beta}}_\kappa)\right] = \mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}_\kappa}) - E\left(\Delta(\widehat{\boldsymbol{\beta}}_\kappa)\right) + o\left[\mathrm{PE}(\widehat{\boldsymbol{\beta}}_\kappa)\right]. \tag{7}$$

The mirror function $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}_\kappa})$ is given by the prediction error of a least squares estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{O}_\kappa} = \left(\mathbf{X}_{\mathrm{O}_\kappa}^T\mathbf{X}_{\mathrm{O}_\kappa}\right)^{-1}\mathbf{X}_{\mathrm{O}_\kappa}^T\boldsymbol{Y}$, on a selection O_κ made by an oracle observing the response without noise, $\mathbf{X}\boldsymbol{\beta}$.

The evaluation of the mirror correction (6) in practical situations, where no oracle is available, is non-trivial, as it depends on the adopted selection procedure, the size and structure (nested, trees, etc.) of the set of models under consideration, the design matrix $\mathbf{X}$. A bootstrap or any other resampling procedure, for instance, is hard to set up, precisely because m_κ describes the interaction between the noise in the original sample and the selection. Monte-Carlo simulations can be used in the calculation of the subsequent estimation. The estimation developed in this paper starts from the straightforward application of the law of iterated expectations, stating that $m_\kappa = E(\widehat{m}_\kappa)$ where

$$\widehat{m}_\kappa = \left[E\left(\|\mathbf{P}_{\mathrm{S}_\kappa}\boldsymbol{Z}\|_2^2 \middle| \mathrm{S}_\kappa\right) - \kappa\right]\sigma^2/n, \tag{8}$$

which is random, sample dependent through the conditioning on the sample dependent selection S_κ. The central point in the development in the subsequent sections (in particular Section 3.3) will be the understanding of what it means to have observed the event S_κ. In other words, we need to quantify the exact information provided by the selection S_κ at the level of each selected or active parameters β_l, $l \in \mathrm{S}_\kappa$. The issue is related to problems in the domain of post-selection inference. Post-selection inference may proceed in basically two ways. The first, generalistic approach aims at confidence intervals that are valid, regardless of the selection

procedure (possibly limited to a certain class of selection or models) [Berk et al., 2013]. The second approach, related to the discussion in this paper, operates on a conditional, rather than a generalistic level, looking for the conditional distribution of estimators, given a specific selection procedure [Charkhi and Claeskens, 2018]. Our paper studies the effect of the selection on the degrees of freedom and the information criterion, not on the subsequent inference.

The remainder of this paper is devoted to the development and the estimation of the correction m_κ in two graphical models.

3 Finetuning a tree selection

The first application of the proposed selection based information criterion consists in finetuning a backtracking algorithm for best κ-subtree selection. Applications are situated in Best Orthogonal Basis selection [Coifman and Wickerhauser, 1992], classification and regression trees [Breiman et al., 1984], tree structured wavelet regression [Jansen, 2022], and wavelet packets pruning.

3.1 A tree structured model of covariates

The best κ-subtree selection operates on the linear regression model in (1). It replaces a lasso procedure, such as LARS [Efron et al., 2004] or a proximal gradient or subgradient method. The tree structured selection rests on two assumptions, leading to a more specialised selection procedure. The first assumption concerns the set of models to choose from. Denoting a model by the subset of the indices $\mathrm{S} \subset \{1, 2, \ldots, m\}$ corresponding to the nonzero components of the parameter vector $\boldsymbol{\beta}$, the tree structured selection restricts the search to subsets satisfying an imposed hierarchy in a way explained below. The hierarchy is supposed to reflect additional information on the physical nature of the covariates, by stating that a given covariate cannot be part of a model unless at least another specific covariate is selected as well. It thus defines for every component $i \in \{1, 2, \ldots, m\}$, termed a node in this context, a unique parent node $p(i) \in \{0, 1, \ldots, m\}$, where $p(i) = 0$ means that the node has no parent, i.e., it is a root. An important special case is that of the binary tree rooted at node 1, which has $p(i) = \lfloor i/2 \rfloor$, where $\lfloor x \rfloor$ is the floor function of x. The tree selection developed below works on general, not necessarily binary, trees and even in the presence of multiple roots. An extreme example of the latter is the case where $p(i) = 0$ for all nodes i, leading to the situation where there is no hierarchy and all nodes are roots.

Assumption 3.1 *(tree structured selection) A valid selection* S *satisfies the hierarchy in the sense that* $i \in \mathrm{S} \Rightarrow p(i) \in \mathrm{S}$*.*

The second assumption puts restrictions on collinearity.

Assumption 3.2 *(frame condition) Let the m columns of the design matrix be ℓ_2 normalised, i.e., the diagonal of $\mathbf{X}^T\mathbf{X}$ is the identity matrix. Then we assume the existence of a matrix $\widetilde{\mathbf{X}}^T$ with the same size $n \times m$ as that of $\mathbf{X}$, and of two positive constants γ and Γ, independent from n and m so that for any $\boldsymbol{\mu} \in \mathbb{R}^n$,*

$$\frac{\gamma}{m}\|\widetilde{\mathbf{X}}\boldsymbol{\mu}\|_2^2 \leqslant \frac{1}{n}\|\boldsymbol{\mu}\|_2^2 \leqslant \frac{\Gamma}{m}\|\widetilde{\mathbf{X}}\boldsymbol{\mu}\|_2^2,$$

while $\mathbf{X}\widetilde{\mathbf{X}}$ is m/n times the $n \times n$ identity matrix. For each selection S, the estimation is supposed to be constructed by composing the hat matrix $\widetilde{\mathbf{X}}_\mathrm{S}$ from the columns of $\widetilde{\mathbf{X}}$ corresponding to the elements in S.

In this setting, the hat matrix $\widetilde{\mathbf{X}}_\mathrm{S}$ is not constructed after but *before* selection, from the full hat matrix $\widetilde{\mathbf{X}}$, whose columns depend on all columns of $\mathbf{X}$, not only on the selected ones in $\mathbf{X}_\mathrm{S}$. More precisely, we have $\widetilde{\mathbf{X}}_\mathrm{S} = (n/m)\mathbf{D}_\mathrm{S}\widetilde{\mathbf{X}}$. With k the cardinality of S, the matrix $\mathbf{D}_\mathrm{S}$ is a $k \times n$ diagonal selection with elements $D_{\mathrm{S};ij} = 1$ if the ith element of the sorted sequence from S equals j, for $i = 1, 2, \ldots, k$.

An example of the matrix $\widetilde{\mathbf{X}}$ is the pseudo-inverse (or Moore-Penrose inverse), while the smallest and largest singular values of $\mathbf{X}$ can be associated to $\sqrt{n\gamma/m}$ and $\sqrt{n\Gamma/m}$ respectively. When $n = m$, such as in a fast (decimated) wavelet transform, the matrix $\widetilde{\mathbf{X}}$ corresponds to the forward (analysis) data transform while the matrix $\mathbf{X}$ represents the reconstruction. In non-orthogonal transforms, the forward transform $\widetilde{\mathbf{X}}$ does not coincide with the pseudo-inverse of the reconstruction.

The vector $\boldsymbol{\beta} = (n/m)\widetilde{\mathbf{X}}\boldsymbol{\mu}$ is a valid model, in the sense that $\mathbf{X}\boldsymbol{\beta}$ predicts the observations from the DGP exactly. In the presence of noise, $\widetilde{\boldsymbol{Y}} = (n/m)\widetilde{\mathbf{X}}\boldsymbol{Y}$ is an unbiased vector of pseudo-observations of $\boldsymbol{\beta}$. The sparse estimator $\widehat{\boldsymbol{\beta}}_\mathrm{S}$ is then given by diagonal selection of the pseudo-observations, $\widehat{\boldsymbol{\beta}}_\mathrm{S} = \mathbf{D}_\mathrm{S}\widetilde{\boldsymbol{Y}}$.

Because of Assumption 3.2, the prediction error of an estimator $\widehat{\boldsymbol{\beta}}_\mathrm{S}$ is bounded by

$$\frac{\gamma m^2}{n^2}\mathrm{R}(\widehat{\boldsymbol{\beta}}_\mathrm{S}) \leqslant \mathrm{PE}(\widehat{\boldsymbol{\beta}}_\mathrm{S}) \leqslant \frac{\Gamma m^2}{n^2}\mathrm{R}(\widehat{\boldsymbol{\beta}}_\mathrm{S}),$$

where $\mathrm{R}(\widehat{\boldsymbol{\beta}}_\mathrm{S})$ is the risk or prediction error on the selected pseudo-observations,

$$\mathrm{R}(\widehat{\boldsymbol{\beta}}_\mathrm{S}) = \frac{1}{m}E\|\widehat{\boldsymbol{\beta}}_\mathrm{S} - \boldsymbol{\beta}\|_2^2 \tag{9}$$

The prediction error $\mathrm{R}(\widehat{\boldsymbol{\beta}}_\mathrm{S})$ can be estimated by an information criterion as in (3), defined on the pseudo-observations. We have

$$\mathrm{SS}_\mathrm{E}(\widehat{\boldsymbol{\beta}}_\mathrm{S}) = \|\widetilde{\boldsymbol{Y}} - \mathbf{D}_\mathrm{S}\widetilde{\boldsymbol{Y}}\|_2^2 = \|\widetilde{\boldsymbol{Y}}\| - \sum_{\ell \in \mathrm{S}} \widetilde{Y}_\ell^2. \tag{10}$$

With $\widetilde{\boldsymbol{Z}} = \widetilde{\mathbf{X}}\boldsymbol{Z}$, the degrees of freedom are given by $\nu_{\mathrm{S}} = E\left[\|\mathbf{D}_{\mathrm{S}}\widetilde{\boldsymbol{Z}}\|_2^2\right] = E\left[\sum_{\ell\in\mathrm{S}}\widetilde{Z}_\ell^2\right]$, taking a possible sample dependence of the selection S into account.

3.2 Best κ-subtree selection

The best κ-term model S_κ is supposed to be the subtree that maximises the amount of information accumulated in its κ elements, under the constraint of the hierarchy imposed by the parent function $p(i)$. The amount of information is quantified by an accumulative mass function $M(\mathrm{S};\widetilde{\boldsymbol{Y}})$, where

$$M(\mathrm{S};\boldsymbol{x}) = \sum_{i\in\mathrm{S}} M_i(x_i). \tag{11}$$

The mass function is accumulative in the sense that if $\mathrm{S}_1 \cap \mathrm{S}_2 = \emptyset$, then $M(\mathrm{S}_1 \cup \mathrm{S}_2;\boldsymbol{x}) = M(\mathrm{S}_1;\boldsymbol{x}) + M(\mathrm{S}_2;\boldsymbol{x})$. The elementary mass functions $M_i(x_i)$ are supposed to be convex with absolute minimum at zero. Most often these functions will be taken to be symmetric, leading to non-decreasing function of $|x_i|$. Simple examples include the choices $M_i(x_i) = |x_i|$ or $M_i(x_i) = |x_i|^2$, the latter being the default choice in the subsequent discussion. The function M_i may depend on i, allowing us to attach more importance to some nodes, regardless of their values. For instance, nodes close to the root may be a priori more important that nodes further away.

The backtracking algorithm for searching S_κ finds the solutions for all κ at once. It extends weakest link or cost complexity pruning techniques to trees whose subsequent best κ-subtrees (for increasing κ, that is) cannot be guaranteed to be nested. It proceeds as follows

1. Take as input a vector $\boldsymbol{x}$ in n nodes on which lies a hierarchy, defined by a function $p(i)$ that maps each node onto its parent. The objective is to find for each κ the selection S_κ maximising $M(\mathrm{S};\boldsymbol{x})$, defined in (11) among all subsets S with κ elements and satisfying the hierarchy.

2. **If** the set of nodes i with $p(i) = 0$ has more than one element, then introduce a "superroot" as a parent to all these nodes. The result of this step is a rooted tree. Define for each node i, including superroot $i = 0$, the set of its children $\mathsf{C}_i = \{j | p(j) = i\}$. Let A denote the set of active nodes, i.e., nodes open for further processing, initialising $\mathsf{A} \leftarrow \{1, 2, \ldots, n\}$. Initialise $j \leftarrow 0$, and the depth $d \leftarrow 0$.

3. **While** the set A is not empty,

   (a) First *descend* into the tree, as far as possible
   While $\mathsf{C}_j \cap \mathsf{A}$ is not empty:
   - Set $j \leftarrow \min(\mathsf{C}_j \cap \mathsf{A})$.
   - Set $d \leftarrow d + 1$.

- Initialise or re-initialise the best 1 subtree, rooted at the current node at depth d, $\mathrm{S}_{1,d} \leftarrow \{j\}$.

(b) We are now at a node j that has no active children. This node is *merged* with its parent, $i = p(j)$. Find the best combinations of subtrees of parent and child, thus incorporating the subtrees of j into those of i.
For $\kappa = 1, 2, \ldots$

- Find $\ell = \arg \max_{l=1,2,\ldots,\kappa} \left[M(\mathrm{S}_{l,d-1}; \boldsymbol{x}) + M(\mathrm{S}_{\kappa-l,d}; \boldsymbol{x}) \right]$.
- Set $\mathrm{S}_{\kappa,d-1} = \mathrm{S}_{\ell,d-1} \cup \mathrm{S}_{\kappa-\ell,d}$.

(c) Take j out of the active set A. Then go one up one level, i.e., $d \leftarrow d-1$, and $j \leftarrow i$. If the new j still has children within the active set, a new descent will take place along one of these children. Otherwise, j will be merged with its parent, and so, until all nodes have been visited and merged.

4. Set $\mathrm{S}_\kappa \leftarrow \mathrm{S}_{\kappa,0}$.

Once the selections S_κ have been found, finetuning amounts to choosing κ according to the criterion obtained by substitution of (10) into (3). The calculation of the degrees of freedom in (3) requires some further attention, as discussed Section 3.3.

3.3 The effect of the selection on the degrees of freedom

Further development of $\widehat{m}_\kappa$ in (8) is based on symmetric mass functions $M_i(u)$. If $M_i(u)$ is symmetric, then there exists a random threshold $\widehat{\theta}_{\kappa;i}$, depending on all $\widetilde{Y}_j$ with $j \neq i$, so that

$$i \in \mathrm{S}_\kappa \Leftrightarrow |\widetilde{Y}_i| \geqslant \widehat{\theta}_{\kappa;i}.$$

From this it follows that

$$\widehat{m}_\kappa = (\sigma^2/n) \sum_{i \in \mathrm{S}_\kappa} \left\{ E\left[\widetilde{Z}_i^2 \middle| \widetilde{Y}_i^2 > \widehat{\theta}_{\kappa;i}^2 \right] - 1 \right\},$$

which, according to the main result in Marquis and Jansen [2022], can be well approximated by replacing the sparse vector $\boldsymbol{\beta}$ in the model of the pseudo-observations $\widetilde{\boldsymbol{Y}}$ by the zero vector. More precisely, it holds that

$$E\left[\widetilde{Z}_i^2 \middle| \widetilde{Y}_i^2 > \widehat{\theta}_{\kappa;i}^2 \right] = E\left[\widetilde{Z}_i^2 \middle| \widetilde{Z}_i^2 > \widehat{\theta}_{\kappa;i}^2 \right] + o(\mathrm{R}(\kappa)). \tag{12}$$

Unfortunately, to the best of the author's knowledge, there seems to be no fast way to find or even approximate the functions that map the pseudo-observations $\widetilde{\boldsymbol{Y}}$ onto the thresholds $\widehat{\theta}_{\kappa;i}$.

As an alternative, the expected values $E(\widehat{m}_\kappa)$ can be approximated numerically quite well by running Monte Carlo simulations with $\boldsymbol{\beta} = \mathbf{0}$ and computer generated, pseudo-random vectors $\boldsymbol{Z}$ on top of the tree structure specified by the parent function $p(i)$. The numerical simulation thus adopts the same approximation as (12).

3.4 Application to regression trees

Let $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ represents uncorrelated, zero mean noise, while the vector $\boldsymbol{\mu}$ is modelled to consist of constant segments, separated by change points from a set $\mathrm{CP} \subset \{i + 1/2, i = 1, 2, \ldots, n\}$, meaning that $\mu_i = \mu_{i+1}$ unless $i + 1/2 \in \mathrm{CP}$. The objective is to identify the set CP. With $\widehat{\mathrm{CP}}$ the estimated set of change points, the elements of the vector $\boldsymbol{\mu}$ can be estimated by

$$\widehat{\mu}_i = \overline{Y}_{I(i)} = \frac{1}{|I(i)|} \sum_{\ell \in I(i)} Y_\ell,$$

where $I(i)$ is the set $\{l, l+1, \ldots, r\}$ so that $i \in I(i)$, and $l - 1/2, r + 1/2 \in \widehat{\mathrm{CP}}$, while $\{\ell + 1/2, \ell \in I(i)\} \cap \widehat{\mathrm{CP}} = \emptyset$.

The search for candidate change points proceeds through a greedy tree construction, starting with $I_{0,0} = \{1, 2, \ldots, n\}$ and $n_0 = 1$. Then for $j = 0, 1, 2, \ldots$ and for $\ell = 0, \ldots, n_j - 1$, find a partitioning of the set $I_{j,\ell} = I_{j+1,2\ell} \cup I_{j+1,2\ell+1}$, as long as $I_{j,\ell}$ has at least two elements. The partitioning is defined by a new candidate change point $t_{j,\ell} + 1/2$, fixing $I_{j+1,2\ell} = \{i \in I_{j,\ell}; i < t_{j,\ell} + 1/2\}$ and $I_{j+1,2\ell+1}$ as its complement. The value of $t_{j,\ell}$ is chosen to maximise a contrast function $c_{j,\ell} = c(\boldsymbol{Y}, I_{j,\ell}, I_{j+1,2\ell})$, in which the argument $I_{j+1,2\ell}$ depends on $t_{j,\ell}$. It is clear that the set of all $I_{j,\ell}$ constitute a binary tree rooted at $I_{0,0}$, referred to as the refinement tree. When the refinement is pursued until all leaves $I_{j,\ell}$ are singletons, then the tree has $2n - 1$ nodes in total, root and leaves included.

The contrast $c_{j,\ell}$ may or may not coincide with the absolute or squared value of the offset or detail

$$d_{j,\ell} = \frac{\overline{Y}_{j+1,2\ell+1} - \overline{Y}_{j+1,2\ell}}{\sqrt{\frac{1}{n_{j+1,2\ell+1}} + \frac{1}{n_{j+1,2\ell}}}},$$

where $n_{j,\ell}$ is the cardinality of $I_{j,\ell}$ and $\overline{Y}_{j,\ell}$ the average value of $\boldsymbol{Y}$ on $I_{j,\ell}$. The $n - 1$ detail coefficients $d_{j,\ell}$ can each be associated with the corresponding internal node $I_{j,\ell}$ in the refinement tree. The detail coefficients are completed by a single overall normalised average value $s_{0,0} = \sqrt{n} \cdot \overline{Y}$. Within a given refinement tree, the mapping of the vector $\boldsymbol{Y}$ onto the vector of details and overall average is an orthogonal transform. This transform is known as the data-adaptive, unbalanced, orthogonal Haar-wavelet transform (AUHT) [Girardi and Sweldens, 1997]. Selection of $\widehat{\mathrm{CP}}_\kappa$ then amounts to the selection of the best κ-subtree T_κ of the refinement tree.

The quality of T_κ can be measured in the prediction error $\mathrm{PE}(\boldsymbol{\mu}_\kappa) = \frac{1}{n}E\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2$ and estimated by $\Lambda(\boldsymbol{\mu}_\kappa) = \frac{1}{n}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{Y}\|_2^2 + \frac{2\nu_\kappa}{n}\sigma^2 - \sigma^2$. Thanks to the orthogonality of the AUHT, it holds that

$$\Lambda(\boldsymbol{\mu}_\kappa) = \frac{1}{n}\sum_{(j,\ell)\notin T_\kappa} d_{j,\ell}^2 + \frac{2\nu_\kappa}{n}\sigma^2 - \sigma^2.$$

The degrees of freedom ν_κ can be approximated using (6), which amounts to

$$\nu_\kappa \approx E\left[\|\mathbf{P}_{\mathrm{S}_\kappa}\boldsymbol{Z}\|_2^2\right]. \tag{13}$$

The Monte Carlo calculation proceeds by running the tree selection procedure on a pseudo-random vector $\boldsymbol{Z}$.

3.5 A simulation study with change points under Poisson noise

If the noise in the model $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\eta}$ is uncorrelated but heteroscedastic, then the AUHT vector $\boldsymbol{d}$ will be heteroscedastic as well. Let $\widetilde{\mathbf{W}}$ denote the data-adaptive matrix that maps $\boldsymbol{Y}$ onto $\boldsymbol{d}$ and define $\boldsymbol{v} = \widetilde{\mathbf{W}}\boldsymbol{\mu}$, then a large value of $d_{j,\ell}$ may be due to large variance or to a large absolute expected value. The distinction between these two cases is quantified by taking a standardised prediction error,

$$\mathrm{PE}(\widehat{\boldsymbol{v}}_\kappa) = \frac{1}{n}\sum_{j,\ell}\left(\frac{\widehat{v}_{j,\ell} - v_{j,\ell}}{\sigma_{j,\ell}}\right)^2,$$

where $\sigma_{j,\ell}^2 = \mathrm{var}(d_{j,\ell})$. This prediction error is estimated by

$$\Lambda(\boldsymbol{v}_\kappa) = \frac{1}{n}\sum_{(j,\ell)\notin T_\kappa}\frac{d_{j,\ell}^2}{\sigma_{j,\ell}^2} + \frac{2\nu_\kappa}{n} - 1. \tag{14}$$

We apply the procedure to Poisson distributed observations with intermittent intensities, as illustrated in Figure 1. The sample size is $n = 4000$. The intensity curve, depicted in solid black line, is taken from the well known 'blocks' test function [Donoho and Johnstone, 1994], vertically translated by adding 3.5, in order to create comparable settings as in the simulation study in Jansen [2007].

The variances $\sigma_{j,\ell}^2$ are estimated as the diagonal elements of the estimated covariance matrix

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \widetilde{\mathbf{W}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}\widetilde{\mathbf{W}}^T,$$

where the matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}$ is obtained as the diagonal matrix whose elements are pilot estimators $\widehat{\mu}_{0;i}$ of $\mu_i = \mathrm{var}(Y_i)$. The pilot estimator is obtained from an unstructured, threshold based

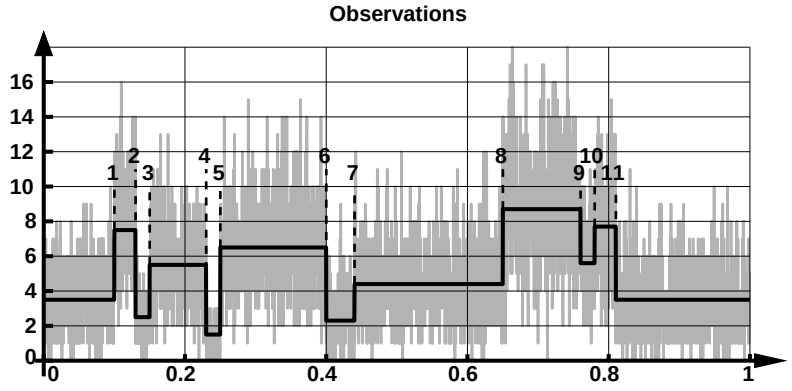


Figure 1: Poisson intensity curve (black line) along with $n = 4000$ corresponding pseudo-random Poisson observations. The objective is to retrieve the change points (discontinuities) in the black curve. The intensity curve is obtained by adding the value 3.5 to the values of the well known 'blocks' test function [Donoho and Johnstone, 1994]. The intensity curve has $n_1 = 11$ change points, numbered 1 to 11 for further reference.

selection $\widehat{\boldsymbol{\mu}}_0 = \widetilde{\mathbf{W}}^{-1}\widehat{\boldsymbol{v}}_0$, where

$$\widehat{v}_{0;j,k} = \mathrm{ST}(d_{j,k}/\sigma_{00,j,k}, \lambda) \cdot \sigma_{00,j,k}.$$

Here $\mathrm{ST}(x, \lambda) = \mathrm{sign}(x) \cdot (|x| - \lambda) \cdot I(|x| > \lambda)$ is the soft-threshold function, in which $I(|x| > \lambda)$ is the indicator function, i.e., $I(|x| > \lambda) = 1 \Leftrightarrow |x| > \lambda$ and $I(|x| > \lambda) = 0$ otherwise. The threshold in the pilot estimator is selected by a GCV or C_p criterion as well. The pre-pilot estimator $\sigma_{00,j,k}$ is obtained as a diagonal element in the unbiased estimator of the covariance matrix $\widetilde{\mathbf{W}}\mathrm{diag}(\boldsymbol{Y})\widetilde{\mathbf{W}}^T$. The reason for not taking this unbiased estimator in the tree structured selection and estimation lies in the large variance of the unbiased estimator, adding fluctuations to the standardised data, thus falsely suggesting the presence of change points.

The Figures 2 and 3 depict two reconstructions of the Poisson intensity curve in Figure 1. Both reconstructions operate on an AUHT based regression tree, in which the contrast function, used in the AUHT refinement is given by

$$c_{j,\ell} = \frac{\overline{Y}_{j+1,2\ell+1} - \overline{Y}_{j+1,2\ell}}{\sqrt{\left(\frac{1}{n_{j+1,2\ell+1}} + \frac{1}{n_{j+1,2\ell}}\right)^q}},$$

where $q = 1$ would lead to $c_{j,\ell} = d_{j,\ell}$. Higher values of q promote balanced refinements, i.e., splitting an set of points $I_{j,\ell}$ near its midpoint. It is found empirically that $q = 2$ leads to better results that $q = 1$, although the issue requires a closer look in further research.

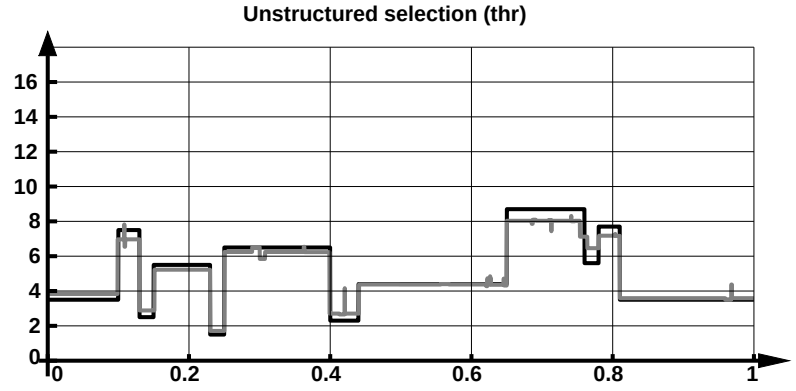


Figure 2: Reconstruction of the Poisson intensity curve (black line) in Figure 1 using soft-thresholding applied to the AUHT coefficients of the observations (grey line in Figure 1). The threshold is finetuned by minimisation of the Generalised Cross Validation score, defined in (4).

The reconstruction in Figure 2 adopts simple, unstructured soft-thresholding on the AUHT coefficients, where the threshold is chosen by minimisation of the GCV expression in (4). The degrees of freedom in a soft-threshold scheme are given by $\nu_\kappa = \kappa$ [Zou et al., 2007, Tibshirani and Taylor, 2012, Jansen, 2015]. The reconstruction in Figure 3 is obtained from the

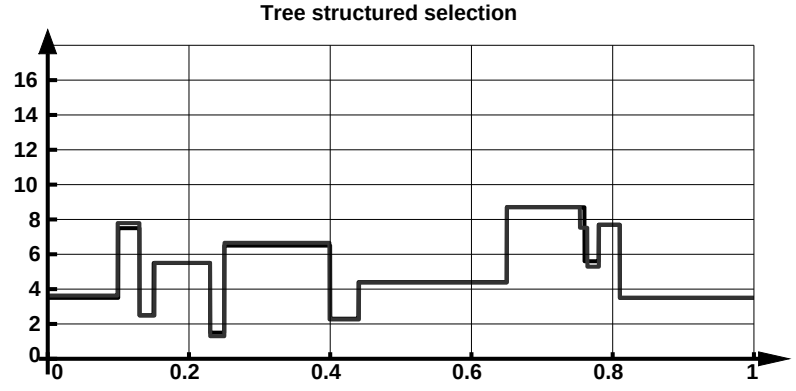


Figure 3: Reconstruction of the Poisson intensity curve (black line) in Figure 1 using tree structured selection of the AUHT coefficients of the observations (grey line in Figure 1).

best κ-subtree selection of the AUHT coefficients. The value of κ is found by minimisation of $\Lambda(\boldsymbol{v}_\kappa)$ in (14). This minimisation is visualised in Figure 4, which compares several criteria for the evaluation of a best κ-subtree. Leave-half-out Cross Validation, marked as CV in the

figure and reported as an appropriated option for applying cross validation in the context of variable selection [Yang, 2007] clearly minimises at too large subtrees (even beyond the range depicted in the figure, not to mention the fluctuations in the curve). Naive use of Mallows's C_p (with $\nu_\kappa = \kappa$) leads to a similar behaviour. The plots of CV and C_p in Figure 4 are represen-

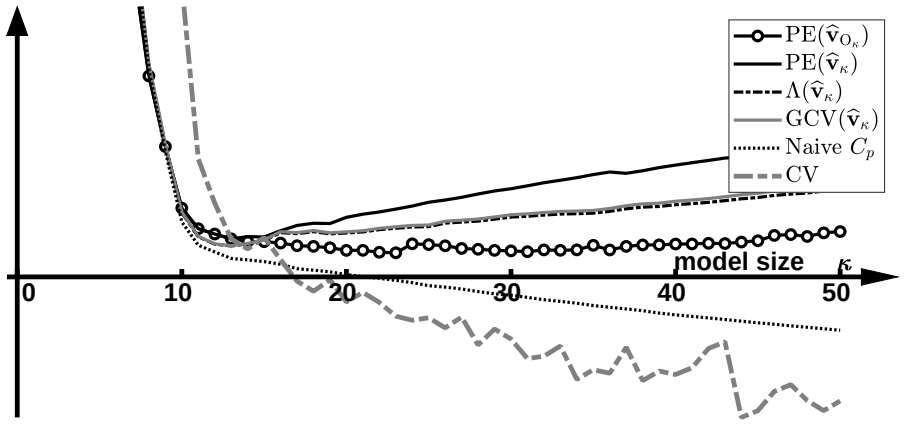


Figure 4: Curve of the estimated prediction error $\Lambda(\boldsymbol{v}_\kappa)$ as a function of the size (number of nodes) κ of the selected subtree. For comparison, the figures also depicts the true prediction error, $\mathrm{PE}(\widehat{\boldsymbol{v}}_\kappa)$, and the GCV alternative for the estimated curve, using (4). Alternatives not using the mirror correction include the naive implementation of Mallows's C_p with $\nu_\kappa = \kappa$ and classical leave-half-out cross validation (CV). These two methods do not come close to identifying the correct optimal κ, leading to overestimated subtrees. The curve of C_p as a function of κ can be seen to be the reflection of the $\Lambda(\boldsymbol{v}_\kappa)$ curve w.r.t. mirror, i.e., the oracular PE-curve, $\mathrm{PE}(\widehat{\boldsymbol{v}}_{\mathrm{O}_\kappa})$.

tative for the essential problem that affects any parameter selection (not just tree structured approaches) based on information criteria: right after the initial, straightforward selection of the most prominent covariates, characterised by a steep drop of the criterion's value, the selection procedure enters a more critical phase, in which it has to distinguish among more questionable candidates. In this phase, any information criterion will encounter insignificant candidates with an accidentally high score. This high score comes from the fact that the false positive covariate carries more noise than an arbitrary insignificant covariate. This discrepancy is described by the mirror correction: a false positive covariate may appear to be the best candidate for selection, whereas in reality it induces more noise than the acceptance of a random candidate.

The minimisation of the $\Lambda(\widehat{v}_\kappa)$ as a function of the subtree size κ leads to the subtree depicted in Figure 5, which in its turn gives rise to the reconstruction in Figure 3. The selected subtree contains the root of the AUHT refinement tree, which represents to overall average value. It also contains twelve nodes corresponding to refinements in the construction of the

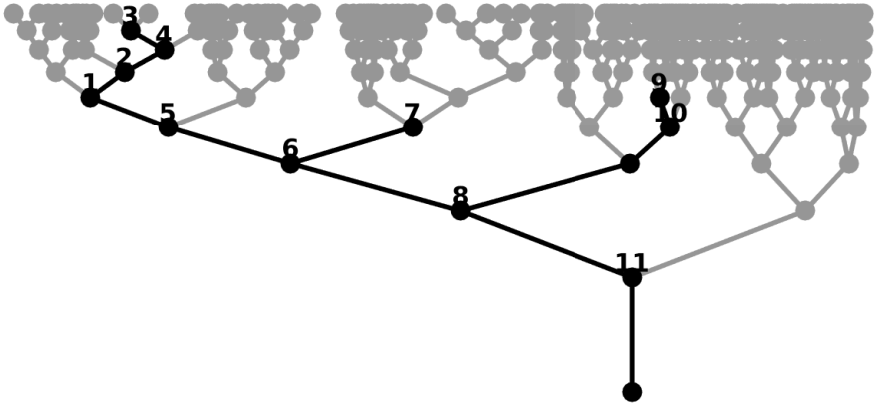


Figure 5: Subtree selected by minimising $\Lambda(\widehat{v}_\kappa)$ in Figure 4, leading to the reconstruction $\widehat{\boldsymbol{\mu}}$ in Figure 3. The first ten levels of the full AUHT tree are depicted in background grey.

regression tree. Eleven of these twelve refinements can be associated to a real change point in the intensity curve, as can be seen from the nodes being marked with the corresponding change point number in Figure 1. One refinement does not correspond to a real change point, making it a false positive. As the corresponding node fathers two real change point nodes, this false positive is due to the construction of the AUHT, not to the tree structured selection algorithm, nor to the finetuning of that selection based on the minimisation of the information criterion. Also note that none of the eleven real change points were missed.

Figure 6 and Table 1 summarises a simulation study for 200 realisations with the same blocks signal intensity curve. The Figure plots the positions of the true positives across the simulation runs, leaving a gap whenever the change point was missed in a simulation runs (these gaps occurring mainly on the curve of change point number 10). The plot reveals, not surprisingly, that the false negative probability (i.e., the gap probability), as well as the variance of the estimated location (i.e., the fluctuation of the curve) of a change point depends on the height of the change and on the range on both sides of the change point (i.e., the distance to the nearest change point on the left and the right).

3.6 A simulation study with varying model sizes

The importance of using modified information criteria in the selection of sparse subtrees can also be illustrated by looking at the size of the selected subtree as a function of the true model size. To this end, the subsequent simulation study reconsiders the full tree structured model depicted in grey in Figure 5. For the purpose of this simulation study, the link with

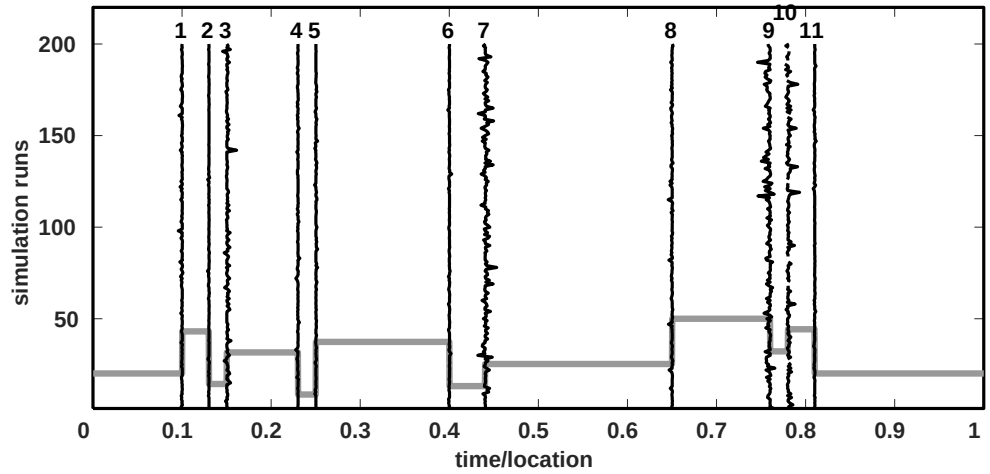


Figure 6: Simulation study of 200 runs, plotting for each change point the locations of its estimator throughout the 200 runs.

		Proportion of missing change points		
		0	1	total
Proportion of false positives	0	30.0	1.5	31.5
	1	33.5	3.0	36.5
	2	19.0	0	19.0
	3	7.0	0	7.0
	4	4.5	0	4.5
	5	1.5	0	1.5
	total	95.5	3.4	100

Table 1: Percentages of reconstructions subdivided according to number of missing change points and number of false positives.

change point analysis in Poisson data is omitted. Instead, for each of the true model sizes $\kappa = \{1, 2, \ldots, 200\}$, the study constructs, at random, a κ-subtree S_κ of size κ. All nodes $i \in \mathrm{S}_\kappa$ receive, independently from each other, a value μ_i from a Laplace distribution, meaning that these values of μ_i are independent realisations from a random variable with density function $f_\mu(m) = (\ell/2)\exp(-\ell|m|)$, where hyperparameter ℓ equals $1/5$ throughout the simulation study. The other nodes, $i \notin \mathrm{S}_\kappa$ have a value $\mu_i = 0$. To all values of μ_i, with i in and outside S_κ alike, zero mean, independent, homoscedastic, normal noise σZ_i is added, in the normal signal-plus-noise model $\boldsymbol{Y} = \boldsymbol{\mu} + \sigma\boldsymbol{Z}$. The objective is to estimate the subtree S_κ from the observations $\boldsymbol{Y}$, using the information that the set of nonzeros in $\boldsymbol{\mu}$ constitutes a subtree. The experiment is

repeated a hundred times on the same κ subtree, each time with newly generated values of $\boldsymbol{\mu}$ and $\boldsymbol{Z}$.

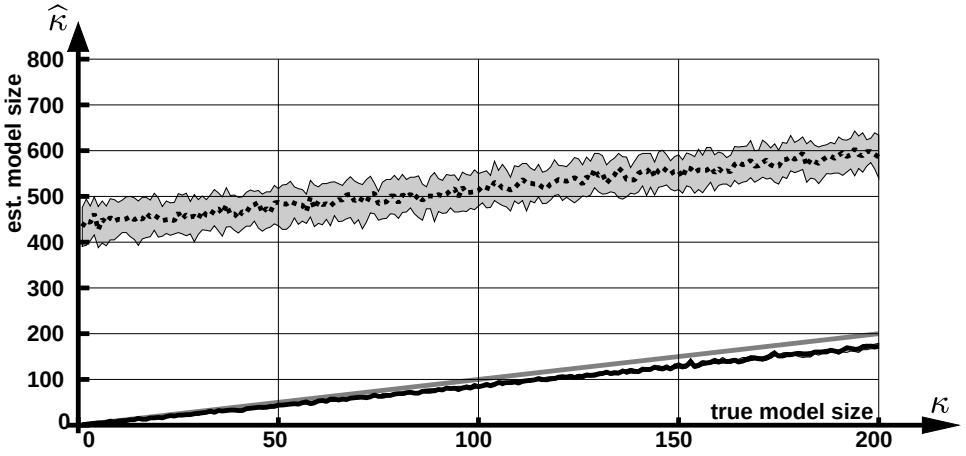


Figure 7: Median sizes (thick lines) along with lower and upper quartiles of the Mallows's C_p selected trees as a function of the size of the true subtrees in hundred replicates within the full tree structure represented depicted in grey in Figure 5. The solid line depicts the median sizes of the criterion proposed in this article. The corresponding quartiles are hardly visible. The dotted line depicts the median sizes of the classical criterion, against a clearly visible shaded background of the interquartile ranges. The grey, thick line is the identity function. The plots confirm that classical criterion is unsuited for use in a context of sparsity.

Figure 7 depicts the sizes of the selected trees, $\mathrm{S}_{\widehat{\kappa}}$ as a function of the true sizes of the randomly chosen subtrees S_{κ}. The selection of $\mathrm{S}_{\widehat{\kappa}}$ proceeds by minimising the C_p criterium in (3) or (14) for the estimation the the prediction error. The plots in the figure compare the classical C_p criterium with the newly proposed version in this paper. The dotted line in Figure 7 is the median curve of outcomes when using the classical criterion, given by (14) with $\nu_{\kappa} = \kappa$. The median is taken over the hundred replicates of the signal-plus-noise model with given S_{κ}. The shaded band represents the interquartile range, i.e., the lower and upper empirical quartiles of the hundred replicates. It is clear that the estimated trees, $\mathrm{S}_{\widehat{\kappa}}$, unacceptably overestimate the true subtree S_{κ}. Using the C_p criterium (14) with ν_{κ} as in (13) leads to the median curve in solid line and the associated interquartile range, nearly too narrow to be visible behind the thick median line. This confirms that the variance of the estimated tree size is much smaller. Moreover, the median curve is close to the identity curve, suggesting that the modified criterion succeeds in finding most of the true subtree. The modified criterion tends to slightly underestimate the true subtree, meaning that it tends to miss a small proportion of significant nodes. This is explained by the presence of small nonzero values of μ_i near the leaves of the true subtree. Such

values are likely to be missed by any selection procedure when the noise relatively large.

3.7 Comparison with modified information criteria for finite samples

Modified information criteria have also been proposed in the finite sample setting, typically when the number of candidate covariates reaches 10% of the sample size [Broersen and Wensink, 1998, Broersen, 2000, Broersen and de Waele, 2004, Mariani et al., 2015, Stoica and Selen, 2004], i.e., when $m > n/10$, using the notations of this paper. In the simulation studies of Sections 3.5 and 3.6, we have $m = n$. Although this meets the conditions $m > n/10$, there is a fundamental difference with the finite sampling setting. Indeed, in the finite sampling setting, the set of covariates is relatively large compared to the sample size, because the sample size is far from infinity. As a result, the asymptotic results supporting the well known criteria such as AIC, are to be amended for finite samples. On the other hand, in the sparse, high-dimensional setting of this paper, both m and n are supposed to be large, i.e., can be thought to grow to infinity. Whereas in the case of finite sampling, the true model size κ typically remains a constant along with m, a sparsity model often assumes that κ grows slowly with n. Further distinctions are made based on the ratio κ/n when both tend to infinity.

The simulation in Section 3.6 is illustrative for the difference in settings. The plot in Figure 7 has small values of κ on the left side. In a finite sample size experiment, the small ration κ/n would correspond to the case of sample size growing to infinity, hence to the classical case. In the high-dimensional setting, it reflects growing sparsity. It can be seen in Figure 7 that small values κ lead to the largest discrepancies between the classical and the newly proposed information criteria, thus corresponding to the situation where the correction of the classical approach is most needed.

4 Sparse graphical model selection

4.1 Model and estimation

In the second application, the information, i.e., the parameters to be selected, estimated, and inferred, are not situated at the nodes of a tree but rather in the edges of a graphical model. The graphical model represents the concentration or precision matrix of a large multivariate normal random variable, $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The concentration matrix is the inverse of the covariance matrix, $\mathbf{K} = \boldsymbol{\Sigma}^{-1}$, assuming the regularity of that matrix. The concentration matrix describes the conditional dependencies between the components of $\boldsymbol{X}$. Indeed, let X_c with $c \in \{1, 2, \ldots, m\}$ be one of the components of $\boldsymbol{X}$ and denote by c' the complementary set of indices, i.e., c' is $\{1, 2, \ldots, m\}$ without c. Furthermore, let Y_c denote the observation of X_c conditioned on the

other values $\boldsymbol{X}_{c'}$, i.e., $Y_c = X_c|\boldsymbol{X}_{c'}$. Then it holds that

$$Y_c \sim N(-\mathbf{K}_{c,c}^{-1}\mathbf{K}_{c,c'}(\boldsymbol{X}_{c'} - \boldsymbol{\mu}_{c'}), \mathbf{K}_{cc}^{-1}). \tag{15}$$

In the subsequent discussion, we assume that $\boldsymbol{\mu}$, which is of no interest in the question of conditional dependencies, is known to be zero. Learning the concentration matrix can be identified as a so-called nodewise regression problem [Meinshausen and Bühlmann, 2006, Zhou et al., 2011] $Y_c = \boldsymbol{X}_{c'}^T\boldsymbol{\beta}_c + \sigma_c Z_c$, where $\sigma_c = \mathbf{K}_{cc}^{-1/2}$ while $Z_c = \mathbf{K}_{cc}^{1/2}(Y_c - \boldsymbol{X}_{c'}^T\boldsymbol{\beta}_c)$ is a standard normally distributed random variable and $\boldsymbol{\beta}_c^T = -\mathbf{K}_{c,c}^{-1}\mathbf{K}_{c,c'}$. Repeated observations $\boldsymbol{X}_i$, $i = 1, 2, \ldots n$, of the m-variate $\boldsymbol{X}$ define for each component c a $n \times (m-1)$ design matrix $\mathbf{X}_{c'}^T$, each row corresponding to one observation in the regression model.

With a sample size n larger than m, it can be hoped that the sample covariance matrix

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^T$$

has full rank, so its inverse $\widehat{\mathbf{K}} = \widehat{\boldsymbol{\Sigma}}^{-1}$ may serve as an estimator of the concentration matrix. In any case, only a non-singular $\widehat{\mathbf{K}}$ can be the maximum likelihood estimator of $\mathbf{K}$ in the multivariate normal model. Indeed, the log-likelihood is given by

$$\log \mathrm{L}(\mathbf{K}) = \sum_{i=1}^{n}\left[\frac{1}{2}\log\det\mathbf{K} - \frac{m}{2}\log(2\pi) - \frac{1}{2}\boldsymbol{X}_i^T\mathbf{K}\boldsymbol{X}_i\right],$$

which is unbounded if $\det\mathbf{K} = 0$. As

$$\sum_{i=1}^{n}\boldsymbol{X}_i^T\mathbf{K}\boldsymbol{X}_i = \sum_{i=1}^{n}\mathrm{Tr}\left(\boldsymbol{X}_i^T\mathbf{K}\boldsymbol{X}_i\right) = \sum_{i=1}^{n}\mathrm{Tr}\left(\mathbf{K}\boldsymbol{X}_i\boldsymbol{X}_i^T\right) = \mathrm{Tr}\left(\mathbf{K}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T\right),$$

the log-likelihood can be written as $\log \mathrm{L}(\mathbf{K}) = \frac{n}{2}\left[\log\det\mathbf{K} - \mathrm{Tr}\left(\mathbf{K}\widehat{\boldsymbol{\Sigma}}\right) - m\log(2\pi)\right]$. A local maximum is reached in $\widehat{\mathbf{K}}_{\mathrm{ML}}$ if $\nabla\log\mathrm{L}\left(\widehat{\mathbf{K}}_{\mathrm{ML}}\right)$ equals the zero matrix. As $\nabla\log\det\mathbf{K} = \mathbf{K}^{-T}$ and $\nabla\mathrm{Tr}\left(\mathbf{K}\widehat{\boldsymbol{\Sigma}}\right) = \widehat{\boldsymbol{\Sigma}}^T$, this can be verified to develop as $\widehat{\mathbf{K}}_{\mathrm{ML}}\widehat{\boldsymbol{\Sigma}} = \mathbf{I}$. The solution provided by nodewise regression [Meinshausen and Bühlmann, 2006] satisfies this equation if $\widehat{\boldsymbol{\Sigma}}$ is non-singular, but not otherwise. Nodewise regression does not impose symmetric concentration matrices explicitly, although symmetry of $\widehat{\mathbf{K}}_{\mathrm{ML}}$ follows automatically whenever $\widehat{\mathbf{K}}_{\mathrm{ML}}$ is indeed the maximum likelihood estimator. In practice, the computation of $\widehat{\mathbf{K}}_{\mathrm{ML}}$ is often unstable, even when n is larger than m. The graphical lasso [Banerjee et al., 2008, Friedman et al., 2008, Mazumder and Hastie, 2012, Sojoudi, 2016] obtains a regularised estimator $\widehat{\mathbf{K}}_\lambda$ by the maximisation

$$\widehat{\mathbf{K}}_\lambda = \arg\max_{\mathbf{K}}\left[\log\det\mathbf{K} - \mathrm{Tr}\left(\mathbf{K}\widehat{\boldsymbol{\Sigma}}\right) - \lambda\|\mathbf{K}\|_1\right],$$

where $\|\mathbf{K}\|_1 = \mathrm{Tr}\left(\mathbf{K}\mathrm{sign}(\mathbf{K})^T\right)$ stands in this case for the sum of the absolute values of all elements in $\mathbf{K}$ (hence not for the classical induced ℓ_1 matrix norm). Solvers for this constrained optimisation problem have been proposed based on an iterative sequence of lasso solvers, where each iteration step works on one column of $\widehat{\mathbf{\Sigma}}_\lambda$, keeping the other columns constant in that iteration step. The iterative solving increases the computational complexity, compared to nodewise regression [Meinshausen and Bühlmann, 2006], which is an issue if we want to equip the solver with a finetuning of the regularisation. In applications involving big data, the nodewise regression procedure is easy to implement on parallel computers. Moreover, while in a simple lasso problem the link between the regularisation parameter λ and the size κ of the active set is easy to establish, this problem is nontrivial in the framework of graphical models for sparse concentration matrices. For these reasons, we adopt the direct solver of Meinshausen and Bühlmann [2006] as selection method. Once the set of nonzeros in the concentration matrix has been selected, estimation within this set takes place according to a constrained maximum likelihood principle, as outlined in the following section.

4.2 Estimation of the nonzero elements in the concentration matrix

Let $\mathrm{S}_{\mathrm{NW}} \subset \{1, 2, \ldots, m\} \times \{1, 2, \ldots, m\}$ be the selection by the lasso in the nodewise regression framework. A pair $(i, j) \in \mathrm{S}_{\mathrm{NW}}$ means that the corresponding entry in the estimated concentration matrix is nonzero. The selection proceeds row by row by application of lasso to the vectors $\boldsymbol{\beta}_c^T = -\mathbf{K}_{c,c}^{-1}\mathbf{K}_{c,c'}$ with $c \in \{1, 2, \ldots, m\}$ in the linear models (15). The selection is finetuned by optimisation of the criterion

$$\widehat{\Lambda}_c(\widehat{\boldsymbol{\beta}}_{c\kappa}) = \frac{1}{n}\mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{c\kappa}) + \frac{2\kappa}{n}\sigma^2 + 2\widehat{m}_\kappa - \sigma^2,$$

i.e., by filling in (6) into (3), and estimating the correction as in (8). As before, a pilot estimator can be used to deal with the nuisance parameter σ^2. The nodewise lasso offers no automatic symmetry in the selection. The subsequent discussion assumes the selection S_{NW} to be symmetrised by keeping the pair (i, j) in S_{NW} if and only if (j, i) was also selected.

Once S_{NW} has been identified, the shrinkage estimator of the lasso is replaced by the constrained maximum likelihood estimator $\widehat{\mathbf{K}}_{\mathrm{ML,S}}$, maximising $\log \mathrm{L}(\mathbf{K})$ under the condition that $\widehat{K}_{\mathrm{ML,S};ij} = 0$ unless $(i, j) \in \mathrm{S}_{\mathrm{NW}}$. Introducing a matrix of Lagrange multipliers $\widehat{\mathbf{\Lambda}}_{\mathrm{S}}$ with $\widehat{\Lambda}_{\mathrm{S};ij} = 0$ whenever $(i, j) \in \mathrm{S}_{\mathrm{NW}}$ (for the corresponding element of $\widehat{\mathbf{K}}_{\mathrm{ML,S}}$ is unconstrained), the constrained maximum likelihood problem is given by

$$\widehat{\mathbf{K}}_{\mathrm{ML,S}} = \arg\max_{\mathbf{K}} \left[\log\det\mathbf{K} - \mathrm{Tr}\left(\mathbf{K}\widehat{\mathbf{\Sigma}}\right) - \mathrm{Tr}\left(\mathbf{K}\widehat{\mathbf{\Lambda}}_{\mathrm{S}}\right)\right].$$

Taking the derivatives w.r.t. the elements in $\mathbf{K}$ leads to $\mathbf{K}^{-T} - \widehat{\mathbf{\Sigma}}^T - \widehat{\mathbf{\Lambda}}_{\mathrm{S}}^T = \mathbf{0}$. The condition that $\widehat{\Lambda}_{\mathrm{S};ij} = 0$ for $(i, j) \in \mathrm{S}_{\mathrm{NW}}$ then means that the gradient $\mathbf{\Lambda}_{\mathrm{S}}^T = \nabla \log \mathrm{L}(\mathbf{K}) = \mathbf{K}^{-T} - \widehat{\mathbf{\Sigma}}^T$ must

have zero entries in all $(i,j) \in \mathrm{S_{NW}}$.

Unless $\mathrm{S_{NW}}$ contains all possible pairs (i,j), nodewise regression cannot possibly find this constrained maximum likelihood solution. Indeed, Let $(i,j) \notin \mathrm{S_{NW}}$, then $\widehat{K}_{ij} = \widehat{K}_{ji}$ is imposed to be zero. This affects all other $\widehat{K}_{ik}$ and $\widehat{K}_{jk}$ at rows i and j of $\widehat{\mathbf{K}}$. Nodewise regression at row $c = k$, unaware of the zero $\widehat{K}_{ij}$, will find values $\widehat{K}_{ki}$ and $\widehat{K}_{kj}$ as if there is no constraint on $\widehat{K}_{ij}$. The resulting estimated concentration matrix cannot be symmetric (even though the selection $\mathrm{S_{NW}}$ is symmetric), nor can it maximise the likelihood. Let $\widehat{\mathbf{K}}_{\mathrm{NW,S}}$ be the outcome of constrained nodewise regression, then a symmetrised version of it, for instance $(\widehat{\mathbf{K}}_{\mathrm{NW,S}} + \widehat{\mathbf{K}}^T_{\mathrm{NW,S}})/2$, can be used as initial value in an iterative search for $\widehat{\mathbf{K}}_{\mathrm{ML,S}}$. An iterative search can be implemented by projection of the gradient $\mathbf{\Lambda}^T_{\mathrm{S}}$ onto the space of admissible descends, i.e., replacing the partial derivative w.r.t. K_{ij} by zero if $(i,j) \notin \mathrm{S_{NW}}$. The iteration stops as soon as the other partial derivatives, those w.r.t. elements in $\mathrm{S_{NW}}$ are zero.

Let $\mathbf{\Lambda}^T_{\mathrm{S},0}$ denote the search direction, then the iteration step updates the current solution $\mathbf{K}$ to $\mathbf{K} + \omega\mathbf{\Lambda}^T_{\mathrm{S},0}$, where ω maximises the function $g(\omega) = \log \mathrm{L}\left(\mathbf{K} + \omega\mathbf{\Lambda}^T_{\mathrm{S},0}\right)$. Taking the derivative yields

$$g'(\omega) = \mathrm{Tr}\left[\left(\mathbf{I} + \omega\mathbf{K}^{-1}\mathbf{\Lambda}^T_{\mathrm{S},0}\right)\mathbf{K}^{-1}\mathbf{\Lambda}^T_{\mathrm{S},0}\right] - \mathrm{Tr}\left[\widehat{\mathbf{\Sigma}}\mathbf{\Lambda}^T_{\mathrm{S},0}\right].$$

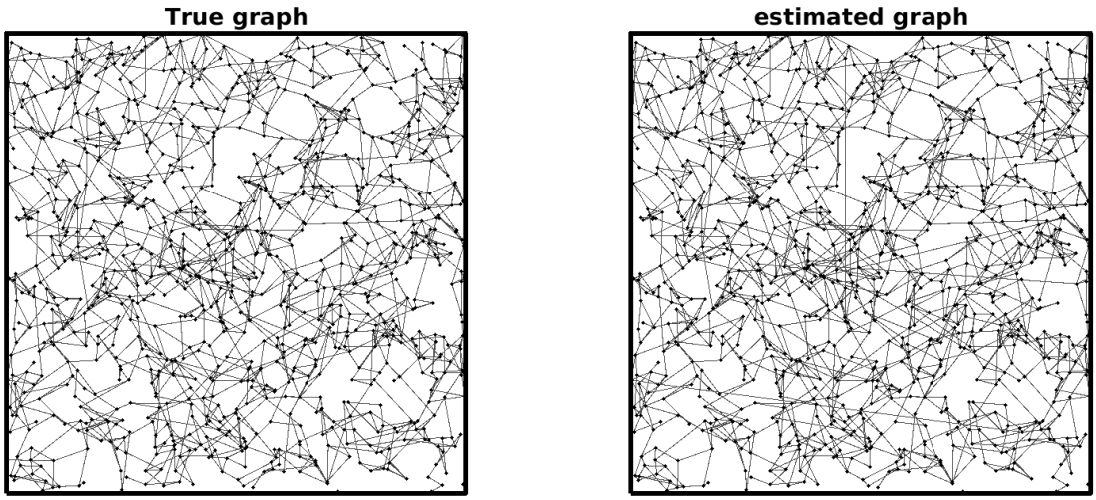
With $\boldsymbol{\gamma}$ denoting the vector of eigenvalues of $\mathbf{K}^{-1}\mathbf{\Lambda}^T_{\mathrm{S},0}$, this is

$$g'(\omega) = \sum_{i=1}^{m} \frac{\gamma_i}{1 + \omega\gamma_i} - \mathrm{Tr}\left[\widehat{\mathbf{\Sigma}}\mathbf{\Lambda}^T_{\mathrm{S},0}\right],$$

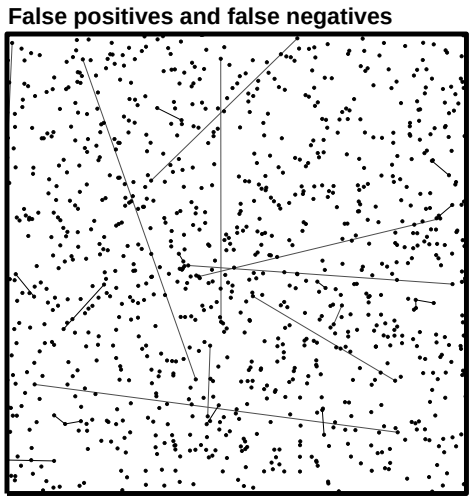
whose zero can be found numerically.

4.3 A short simulation study

Before proceeding to a real data analysis, a short simulation study reveals some understanding in the working of the proposed sparse selection method. Figure 8 displays the setting of the simulation study of Meinshausen and Bühlmann [2006, p.1448], along with the outcome of the proposed refined Mallows's C_p criterion in nodewise regression. The results in the figure were obtained from a sample of $n = 600$ independent observations of a m-variate normal random vector $\boldsymbol{X}$, with $m = 1000$. For the sake of the graphical representation, the components of $\boldsymbol{X}$ are associated with random nodes in a 2D scatter plot, depicted in the Figure. The positions of the nodes have no physical meaning, except for their role in the generation of the concentration matrix. The concentration matrix $\mathbf{K}$ is sparse in the sense that at most $n_e = 4$ nonzero off-diagonals appear on each row and each column. Nonzero K_{ij} are selected at random, but with probabilities roughly inversely proportional to the distances between the nodes corresponding to the ith and jth components of $\boldsymbol{X}$. The graph representing $\mathbf{K}$ will therefore show short edges between adjacent nodes, as can be seen in Figure 8(a). Initially, the values of the nonzero

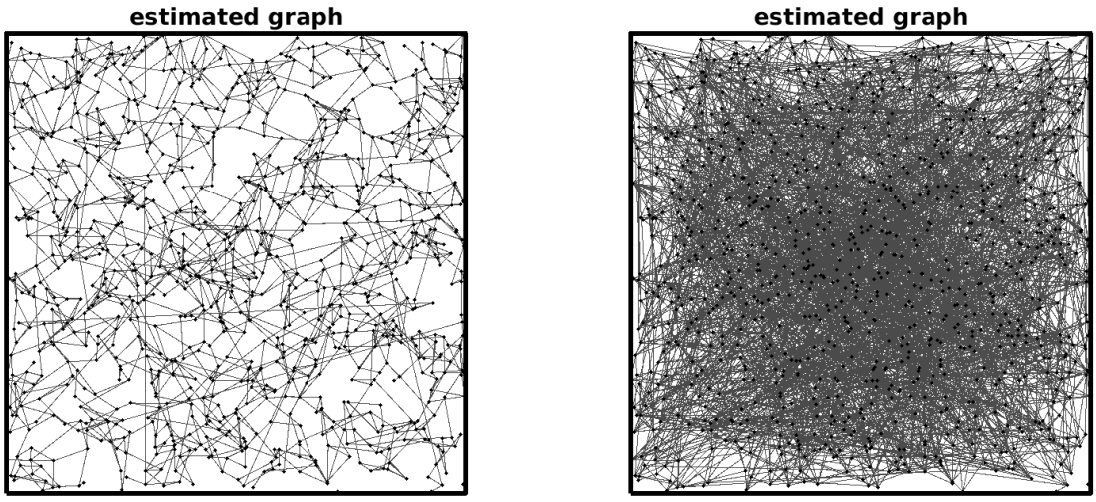


(a) (b)



(c)

Figure 8: Simulation of a nodewise regression on a sample of $n = 600$ observations of m-variate normal random vector, with $m = 1000$, as in Meinshausen and Bühlmann [2006]. (a) Graph representing the sparse concentration matrix $\mathbf{K}$. (b) Estimation of the graph based on the proposed refined Mallows's C_p criterion. (c) False positives and false negatives (type I and type II errors). The false positives are the long edges, connecting components i and j far from each other in the scatter plot. (The estimation method is not aware of the distances in the scatter plot.)



(a) (b)

Figure 9: (a) Lasso shrinkage estimation using the oracular information that the number of edges in each node is bounded by four. (b) Nodewise regression using Lasso shrinkage implemented by LARS with Mallows's C_p as stopping criterion.

off-diagonals in $\mathbf{K}$ are set to $(1/n_e - 0.005) = 0.245$ to guarantee positive definiteness of $\mathbf{K}$. After inversion of the initial matrix $\mathbf{K}$, the resulting covariance matrix is standardised by left- and right multiplication with a diagonal matrix so that the diagonal elements of the covariance matrix (the variances, that is) are all one. The concentration matrix is rescaled accordingly.

As illustrated in Figures 8(b) and 8(c), the proposed method cannot eliminate all false positives (type I errors), nor does it prevent the occurrence of false negatives (type II errors). Instead, it succeeds in finding a delicate balance between the two objectives. Theoretic results quantifying these findings are interesting topics of further research. At first sight, the result in Meinshausen and Bühlmann [2006], depicted in Figure 9(a), is superior, were it not for the upperbound of four on the number of edges in each vertex, used throughout the nodewise regression. In some applications, knowledge of such an upperbound may be available, whereas in other applications, this information should be considered as oracular. In absence of this information, LARS equipped with the classical C_p based stopping criterion (having κ instead of ν_κ as penalty in (3) would lead to a massive overestimation of the true model, as depicted in Figure 9(b). Alternatively, a selection with focus on the false discovery rate may be too conservative, leading to an uncontrolled number of false negatives.

4.4 A real data example

The symmetrised nodewise regression approach with the proposed refined Mallows's C_p is now applied to the gene expression measurements reported in Spira et al. [2007] and analysed in Danaher et al. [2014] to illustrate the graphical lasso across multiple populations. The data are available from the Gene Expression Omnibus [Barrett et al., 2005], `https://www.ncbi.nlm.nih.gov/geo/`, accession code GDS2771.

The observations come from two populations: $n_1 = 90$ individuals belong to the control group, while $n_2 = 97$ patients have been diagnosed with lung cancer. The objective is to investigate whether the diagnosis explains differences in covariance structure between the gene expressions. Just like in Danaher et al. [2014], the genes in the upper 20% quantile of the expression variances are taken out from further analysis, because covariances among these genes are supposed to be dominated by noise. The expression measurements of the remaining $m = 17827$ genes (out of 22283 originally) are studentised within each population.

Let $\boldsymbol{X}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ denote the gene expressions in the control group and $\boldsymbol{X}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ those of the patients, then the symmetrised nodewise regression with the proposed refined Mallows's C_p criterion (with the same settings as in the simulation study) selects $\widehat{\kappa}_1^* = 16198$ and $\widehat{\kappa}_2^* = 13134$ nonzero concentration values $\widehat{K}_{ij} = \widehat{K}_{ji}$, with $1 \leqslant i < j \leqslant m$. In a full model of $m(m-1)/2$ concentration parameters, these selections account for and 0.0102% and 0.0083% of nonzeros respectively. Motivated by application specific, practical considerations, the selections in Danaher et al. [2014, Sections 6 and 8] are even sparser. It may be interesting, however, to allow a wider (yet still very sparse) selection in a first stage, as illustrated by Figure 10. The Figure depicts the ordered magnitudes of the selected off-diagonal elements of the concentration matrix (the elements that are represented by an edge in the graphical model). The values are compared with those obtained by applying the same estimation to n Monte-Carlo observations from a vector of independent normal random variables. The simulated variables have the same variances as the observed ones, in the sense that the variances in the simulation are taken from the diagonal of the sample covariance matrix. With diagonal covariance and concentration matrices, the graphical model of the simulated data is know to have no edges. As the refined Mallows's C_p criterion of this paper reduces false positive selections, it can be expected that the size of the selected set is much smaller in the simulated data than in the observed control and patient data. Therefore, for the sake of comparison, the selection on the simulated data uses the classical definition of Mallows's C_p, by taking $\nu_\kappa = \kappa$ in (3), leading to a vast selection of all false positives. Comparing the magnitudes of these false positives with the selected values in the observed data reveals that they are much smaller and flatter when sorted. This suggests that the selected values in the two populations do have some significance.

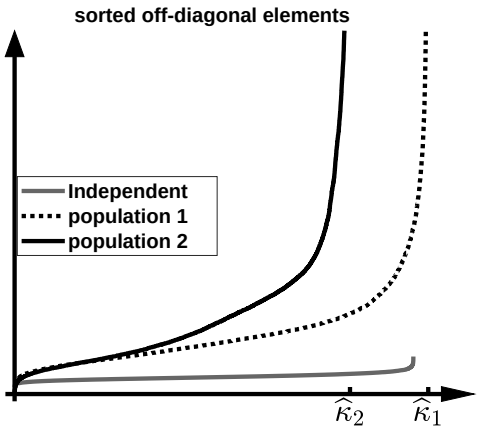


Figure 10: Sorted magnitudes of the selected off-diagonal entries of the concentration matrix. Comparison with (falsely) selected off-diagonal entries in a simulated vector of independent random variables.

5 Conclusion

Parameter selection in high dimensional data is often monitored by practical, application driven considerations or by methods explicitly controlling the false positives to at least some degree. Information criteria, such as Mallows's C_p, but also AIC and others, are often found to be too tolerant of false positives. This paper has explored the use of more a refined Mallows's C_p criterion in high dimensional graphical and tree models. The classical definition of Mallows's C_p, designed for assessment of a *fixed* model, works well for finetuning a lasso shrinkage selection. As lasso shrinkage is tolerant of the presence of false positives, this finetuning leads to largely overestimated models. In contrast to this, the refined criterion developed for graphs and trees in this paper, focuses on *finetuning* selection for estimation *without shrinkage*. This way, the refined criterion, carefully balancing false positive and false negative selections, proves to be interesting in applications where the avoidance of both false positives and false negatives are important objectives.

6 Software

Software for the reproduction of the figures is available in the Matlab package *ThreshLab*, available on `https://maarten.jansen.web.ulb.be/software/threshlab.html`. After installation, type `help makefigsdoftreesandgraphs` to get started.

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csáki, editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W.-C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar. NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res.*, 33:D562–D566, 2005.

R. Berk, L. Brown, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees (CART)*. Wadsworth, Monterey, CA, USA, 1984.

P.M.T. Broersen. Finite sample criteria for autoregressive order selection. *IEEE Transactions on Signal Processing*, 48(12):3550–3558, 2000. doi: $10.1109/78.887047$.

P.M.T. Broersen and S. de Waele. Finite sample properties of arma order selection. *IEEE Transactions on Instrumentation and Measurement*, 53(3):645–651, 2004. doi: $10.1109/\mathrm{TIM}.2004.827058$.

P.M.T. Broersen and H.E. Wensink. Autoregressive model order selection by a finite sample estimator for the kullback-leibler discrepancy. *IEEE Transactions on Signal Processing*, 46 (7):2058–2061, 1998. doi: $10.1109/78.700984$.

A. Charkhi and G. Claeskens. Asymptotic postselection inference for the Akaike information criterion. *Biometrika*, 105(3):645–664, 2018.

S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. on Scientific Computing*, 20(1):33–61, 1998.

G. Claeskens and N. Hjort. The focused information criterion. *J. Amer. Statist. Assoc.*, 98: 900–916, 2003.

R. R. Coifman and M. V. Wickerhauser. Entropy based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2):713–718, 1992.

P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*, 76(2): 373–397, 2014.

D. L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Comm. on Pure and Applied Mathematics*, 59: 797–829, 2006.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

B. Efron, T. J. Hastie, I. M. Johnstone, and R. J. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. with discussion.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. American Statistical Association*, 96(456):1348–1360, 12 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

M. Girardi and W. Sweldens. A new class of unbalanced Haar wavelets that form an unconditional basis for L_p on general measure spaces. *J. of Fourier Analysis and Applications*, 3(4): 457–474, 1997.

N. R. Hansen and A. Sokol. Degrees of freedom for nonlinear leastsquares estimation. Preprint, available as arXiv:1402.2997, 2014.

M. Jansen. Multiscale change point analysis in poisson count data. *Chemometrics and Intelligent Laboratory Systems*, 85(2):159–169, February 2007.

M. Jansen. Information criteria for variable selection under sparsity. *Biometrika*, 101(1):37–55, 2014.

M. Jansen. Generalized cross validation in variable selection with and without shrinkage. *Journal of Statistical Planning and Inference*, 159:90–104, 2015.

M. Jansen. *Wavelets from a statistical perspective*. CRC Press, first edition, 2022.

A. Javanmard and A. Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.

S. Kay. Exponentially embedded families - new approaches to model order estimation. *IEEE Transactions on Aerospace and Electronic Systems*, 41(1):333–345, 2005.

K.-S. Kim and S.-Y. Chung. Tree search network for sparse estimation. *Digital Signal Processing*, 100, 2020.

J. D. Lee, D. L. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Q. Li and J. Shao. Regularizing lasso: a consistent variable selection method. *Statistica Sinica*, 25:975–992, 2015.

C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

A. Mariani, A. Giorgetti, and M. Chiani. Model order selection based on information theoretic criteria: Design of the penalty. *IEEE Transactions on Signal Processing*, 63(11):2779–2789, 2015. doi: $10.1109/\mathrm{TSP}.2015.2414900$.

B. Marquis and M. Jansen. Information criteria bias correction for group selection. *Statistical Papers*, To Appear:–, 2022.

R. Mazumder and T. Hastie. The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125–2149, 2012.

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

M. Niedźwiecki and M. Ciołek. Akaike's final prediction error criterion revisited. In *40th International Conference on Telecommunications and Signal Processing (TSP), Barcelona, Spain*, pages 237–242, 2017. doi: $10.1109/\mathrm{TSP}.2017.8075977$.

J. Rissanen. Modeling by the shortest data description. *Automatica*.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

S. Sojoudi. Equivalence of graphical lasso and thresholding for sparse graphs. *Journal of Machine Learning Research*, 17:1–21, 2016.

A. Spira, J. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y. Dumas, P. Calner, P. Sebastiani, S. Sridhar, J. Beamis, C. Lamb, T. Anderson, N. Gerry, J. Keane, M. Lenburg, and J. Brody. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*, 13(3):361–366, 2007.

C. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

P. Stoica and Y. Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004. doi: $10.1109/\mathrm{MSP}.2004.1311138$.

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

R. J. Tibshirani and J. E. Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, March 2006.

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

M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55 (5):2183–2202, May 2009.

Y. Yang. Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35:2450–2473, 2007.

J. Ye. On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.*, 93:120–131, 1998.

C. H. Zhang. Nearly unbiased variable selection under the minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series B*, 76:217–242, 2014.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

S. Zhou, Ph. Rütimann, M. Xu, and P. Bühlmann. High-dimensional covariance estimation based on gaussian graphical models. *Journal of Machine Learning Research*, 12:2975–3026, 2011.

H. Zou. The adaptive lasso and its oracle properties. *J. American Statistical Association*, 101: 1418–1429, 2006.

H. Zou, T. J. Hastie, and R. J. Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.