Thresholded Adaptive Validation: Tuning the Graphical Lasso for Graph Recovery

Mike Laszkiewicz **Asja Fischer** **Johannes Lederer**
Department of Mathematics, Ruhr University Bochum, Germany

Abstract

Many Machine Learning algorithms are formulated as regularized optimization problems, but their performance hinges on a regularization parameter that needs to be calibrated to each application at hand. In this paper, we propose a general calibration scheme for regularized optimization problems and apply it to the graphical lasso, which is a method for Gaussian graphical modeling. The scheme is equipped with theoretical guarantees and motivates a thresholding pipeline that can improve graph recovery. Moreover, requiring at most one line search over the regularization path, the calibration scheme is computationally more efficient than competing schemes that are based on resampling. Finally, we show in simulations that our approach can improve on the graph recovery of other approaches considerably.

1 Introduction

Over the last decades full of technical achievements, we experienced a revolutionary supply of data, confronting us with large-scale data sets. In order to handle and to infer new insights from the appearing wealth of data it presupposes us to put effort into the development of new, scaleable procedures. One approach to address this problem is using graphical models, which proved to serve as an intuitive, easy-understanding visualization of the underlying interaction network that can then be further analyzed. Typical applications for graphical models occur in several modern sciences, including genetics (Dobra et al., 2004), the analysis of brain connectivity networks (Bu & Lederer, 2017), and the investigation of complex financial networks (Denev, 2015). In all of these cases, graphical models can reduce the network of interactions to its relevant parts to lighten the challenge of high-dimensionality. While domain experts can analyze and interpret the structure of interactions across features, we can use this information for more accurate model building. Using a sparse representation of the relevant parts of the model, it is possible to develop more efficient inference algorithms and accelerate sampling from the model. A popular approach to face this challenge is to recover the network from the data using *undirected graphical models*. We call this task *graph recovery*.

An important class of undirected graphical models are *Gaussian graphical models*. There are numerous estimators for Gaussian graphical models, including those that account for high dimensionality, e.g. the graphical lasso (Yuan & Lin, 2007; Banerjee et al., 2007; Friedman et al., 2008), SCAD (Fan et al., 2009), and MCP (Zhang, 2010), which are based on the idea of regularized maximum likelihood estimation, and various other approaches such as neighborhood regression (Meinshausen & Bühlmann, 2006; Sun & Zhang, 2012), TIGER (Liu & Wang, 2017), and SCIO (Liu & Luo, 2015). These estimators reduce the effective dimensionality of the model through a regularization term that is adjusted to the setting at hand with a regularization parameter.

In this paper we generalize the theoretical framework of Chichignoud et al. (2014) and utilize the large body of preliminary theoretical work (Ravikumar et al., 2011) to verify that we can apply this general scheme to the graphical lasso. Important features of the resulting data-driven calibrated estimator are that it comes with a finite sample upper bound on the approximation error and is computationally efficient as it requires at most one graphical lasso solution path. We equip the estimator with a simple, theory-based threshold and observe a significant improvement over other methods in its graph recovery performance.

The rest of the paper is organized as follows. Section 2 briefly reviews Gaussian graphical models and

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

presents the proposed estimator for graph recovery, which we call *thresholded adaptive validation graphical lasso* (thAV). Based on an empirical analysis of toy data sets presented in Section 3, we demonstrate that the thAV outperforms existing methods on the graph recovery tasks. We apply the thAV to real-world data to recover biological networks in Section 4. Finally, we conclude with a discussion and present ideas for future research in Section 5. The supplement contains theoretical results, proofs, and further simulations. The code for our experiments is provided and can further be accessed through our public git repository[1].

2 Thresholded Adaptive Validation for the Graphical Lasso

We begin by giving a brief review of Gaussian graphical models and describing the graphical lasso optimization problem in Section 2.1. In Section 2.2 we apply the adaptive validation (AV) calibration scheme, which was originally proposed by Chichignoud et al. (2014) and which we generalize to general regularized optimization problems in Section A.1 of the supplement, to the graphical lasso. We obtain finite sample results for the ℓ_∞-loss on the off-diagonals of the graphical lasso, and employ these bounds to motivate a thresholded graphical lasso approach.

2.1 Brief Review of Gaussian Graphical Models

An undirected graphical model expresses the conditional dependence structure between components of a multivariate random variable. More precisely, given a high-dimensional random variable $\boldsymbol{z} \in \mathbb{R}^d$, the undirected graphical model depicts for each pair of components z_i, z_j of $\boldsymbol{z}$ if these are independent given the remaining $d-2$ components of $\boldsymbol{z}$ (i.e. $z_i \perp z_j | \boldsymbol{z}_{\setminus\{i,j\}}$). Formally, an undirected graphical model is defined as a pair $(\boldsymbol{z}, \mathcal{G})$, where $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ is a graph with vertices $\mathcal{V} := \{1, ..., d\}$ and edge set $\mathcal{E} := \{(i,j) \in \mathcal{V} \times \mathcal{V} : z_i \not\perp z_j | \boldsymbol{z}_{\setminus\{i,j\}}\}$. It is well-known that in a Gaussian graphical model, i.e. in the case that $\boldsymbol{z} \sim \mathcal{N}_d(\mathbf{0}_d, \Sigma)$, where Σ is the positive definite covariance matrix, we can find an elegant characterization of the conditional dependency structure. It can be seen as a special case of the *Hammersley-Clifford Theorem* (Grimmett, 1973; Besag, 1974; Lauritzen, 1996): for any $i \neq j \in \mathcal{V}$ it holds that

$$z_i \perp z_j | \boldsymbol{z}_{\setminus\{i,j\}} \quad \Leftrightarrow \quad \Theta_{ij} = 0 \ , \tag{1}$$

where $\Theta := \Sigma^{-1}$ is the so called *precision matrix*. Hence, in order to estimate the conditional dependence graph $\mathcal{G}$, one can build on an estimate $\hat{\Theta}$ of the precision matrix Θ and define $\hat{\mathcal{E}} := \{(i,j) \in \mathcal{V} \times \mathcal{V} : \ \hat{\Theta}_{ij} \neq 0\}$.

Given n samples $\boldsymbol{z}^{(1)}, \dots, \boldsymbol{z}^{(n)}$ drawn independently from $\mathcal{N}_d(\mathbf{0}_d, \Sigma)$, an evident approach to estimate Θ is to employ maximum likelihood estimation. But it is well-known that its performance suffers in the high-dimensional setting where $n \approx d$ or even $n < d$, and that it does not exist in the latter setting (Wainwright, 2019). A typical approach to overcome the burdens that come with high-dimensionality is to assume a sparsity structure on the target, that is, to assume Θ to have many zero-entries. This does not only improve theoretical guarantees but also makes the conditional dependence graph more interpretable. Moreover, imposing a sparsity structure is in accordance with the scientific beliefs in typical application areas in which graphical models are being used (Thieffry et al., 1998; Jeong et al., 2001). The probably most-frequently used sparsity encouraging estimation procedure for Gaussian graphical models is the *graphical lasso* (Yuan & Lin, 2007)

$$\hat{\Theta}_r = \underset{\Omega \in \mathcal{S}_{\mathrm{d}}^+}{\operatorname{argmin}} \Big\{ \operatorname{tr}\Big[\frac{1}{n} \sum_{i=1}^{n} \big(\boldsymbol{z}^{(i)}\big)^\top \boldsymbol{z}^{(i)} \Omega \Big] - \log\left[\det[\Omega]\right] + r \|\Omega\|_{1,\mathrm{off}} \Big\} \ , \tag{2}$$

where $\mathcal{S}_{\mathrm{d}}^+$ is the set of positive definite and symmetric matrices in $\mathbb{R}^{d \times d}$, tr denotes the trace, r is a problem-dependent regularization parameter, and $\|\Omega\|_{1,\mathrm{off}} := \sum_{i \neq j} |\Omega_{ij}|$ denotes the ℓ_1-norm of $\Omega \in \mathcal{S}_{\mathrm{d}}^+$ on its off-diagonal. Of course, the performance of the estimator hinges on the choice of r, and while general theoretical results for the graphical lasso exist (e.g. those presented by Zhuang & Lederer (2018)), to the best of our knowledge there are none that allow for a sophisticated choice of r for graph recovery tasks that occur in practice.

2.2 Thresholded Adaptive Validation

In this section, we transfer the AV technique proposed by Chichignoud et al. (2014) for the lasso to the graphical lasso. As can be seen from our derivation in Section A.1 of the supplement, the technique can be applied to tune any general regularized optimization problem over a set $\mathbb{S}$ of the form

$$\hat{\Theta}_r \in \underset{\Omega \in \mathbb{S}}{\operatorname{argmin}} \big\{ f(\boldsymbol{Z}, \Omega) + r h(\Omega) \big\} \ , \tag{3}$$

where r is a real-valued regularization parameter, f is a function measuring the fit of the estimator Ω given the observed data $\boldsymbol{Z}$, and h is some regularization function depending only on Ω. Comparing (2) with (3) shows

[1] `https://github.com/MikeLasz/thav.glasso`

that the graphical lasso belongs to this class of regularized optimization problems. We can therefore apply the calibration scheme proposed by Chichignoud et al. (2014) and which we generalized in the supplement to obtain the following definition:

Definition 1 (AV)**.** *Let $\mathcal{R}$ be a finite and nonempty set of regularization parameters. Then, the adaptive validation (AV) calibration scheme selects the regularization parameter according to*

$$\hat{r} := \min\big\{r \in \mathcal{R} :\ \ell\big(\hat{\Theta}_{r'}, \hat{\Theta}_{r''}\big) \leq C(r' + r'') \quad \forall\, r', r'' \in \mathcal{R} \cap [r, \infty)\big\}\ ,$$

where $\ell : \mathcal{S}_d^+ \times \mathcal{S}_d^+ \to \mathbb{R}$ is the ℓ_∞-distance on the off-diagonals, $C \in \mathbb{R}$ is a constant (specified in the following), and $\hat{\Theta}_{r'}$, $\hat{\Theta}_{r''}$ are the graphical lasso estimators (2) *using regularization parameter r' and r'', respectively. We call $\hat{\Theta}_{\hat{r}}$ (resulting from inserting $\hat{r}$ into* (2)*) the AV estimator.*

The constant C stems from an assumption, which the theory of Chichignoud et al. (2014) relies on, namely that there exists this constant and a class of events $(\mathcal{T}_r)_{r\in\mathcal{R}}$, which are increasing in r, such that conditioned on $\mathcal{T}_r$ it holds

$$\ell(\Theta, \hat{\Theta}_r) \leq Cr\ . \tag{4}$$

Particularly, we only require the existence of the set of events $(\mathcal{T}_r)_{r\in\mathcal{R}}$ for some fixed C and do not need to have access to it. As demonstrated by Theorem 8 in Section A.2 of the supplement, we can show based on the investigations of Ravikumar et al. (2011) that this assumption holds true for the graphical lasso[2]. Motivated by this, the smallest regularization parameter r_δ^* that enables us to apply (4) with probability $1-\delta$, for some $\delta \in (0, 1]$, can be seen as a natural candidate for r:

$$r_\delta^* := \operatorname*{argmin}_{r\in\mathcal{R}} \big\{\mathbb{P}\big(\mathcal{T}_r\big) \geq 1 - \delta\big\}\ .$$

However, r_δ^* is inaccessible in practice, since we usually cannot measure $\mathbb{P}(\mathcal{T}_r)$. Nonetheless, the AV estimator $\hat{\Theta}_{\hat{r}}$ also results in a good approximation of the precision matrix as guaranteed by the following theorem, which is based on the generalization of the Theorem 3 of Chichignoud et al. (2014) [3] applied onto the graphical lasso.

Theorem 2 (Finite Sample Bound for the AV)**.** *Suppose that $\hat{r}$ is the regularization parameter selected by the AV calibration scheme and C is the constant from* (4)*. Then, for any $\delta \in (0, 1]$, it holds that*

$$\hat{r} \leq r_\delta^* \quad \text{and} \quad \ell\big(\Theta, \hat{\Theta}_{\hat{r}}\big) \leq 3Cr_\delta^* \tag{5}$$

with probability at least $1 - \delta$.

For simplicity of notation, let us denote the AV estimator by $\hat{\Theta} := \hat{\Theta}_{\hat{r}}$ from here on. The finite sample upper bound (5) immediately implies that it holds with probability $1 - \delta$ that

1. for any zero entry $\Theta_{ij} = 0$ of the true precision matrix the corresponding entry $\hat{\Theta}_{ij}$ of the AV estimate satisfies $|\hat{\Theta}_{ij}| \in [0, 3Cr_\delta^*]$;
2. for any significant non-zero entry Θ_{ij} with $|\Theta_{ij}| > (3+\lambda)Cr_\delta^*$, for some constant λ, the corresponding entry $\hat{\Theta}_{ij}$ of the AV estimate is also non-zero with $|\hat{\Theta}_{ij}| > \lambda Cr_\delta^*$.

These observations suggest a strategy for efficient graph recovery: by including all edges (i, j) to the edge set that satisfy $|\hat{\Theta}_{ij}| > \lambda Cr_\delta^*$, we make sure that we recover all significant entries (see 2.). Pursuing this strategy, we can also shrink the interval, in which AV missclassifies zero entries to $[\lambda Cr_\delta^*, 3Cr_\delta^*]$ (see 1.). However, as r_δ^* is inaccessible, we propose to replace it in the selection strategy by the AV regularization parameter, leading to the thresholded estimator defined in the following.

Definition 3 (thAV)**.** *Let $\hat{\Theta}$ be the AV estimator. Then, we define the thresholded adaptive validation graphical lasso (thAV) estimator by*

$$\big(\hat{\Theta}^t\big)_{ij} := \big(\hat{\Theta}_{ij}\mathbb{1}_{\{|\hat{\Theta}_{ij}|>t\}}\big)_{ij}\ , \tag{6}$$

where $t := \lambda C\hat{r}$ is the threshold, $\lambda \in (0, 3]$, and $\mathbb{1}_A$ is the indicator function over a set A. The resulting estimated edge set is then

$$\begin{aligned}\hat{\mathcal{E}} :&= \big\{(i, j) \in \mathcal{V} \times \mathcal{V} :\ \hat{\Theta}_{ij}^t \neq 0\big\} \\ &= \big\{(i, j) \in \mathcal{V} \times \mathcal{V} :\ |\hat{\Theta}_{ij}| > \lambda C\hat{r}\big\}\ .\end{aligned}$$

As we know from Theorem 2 that $\hat{r} \leq r_\delta^*$ with probability $1 - \delta$, we can use the above observations to derive the following corollary that guarantees outstanding graph recovery properties of the thAV:

Corollary 4 (Finite Sample Graph Recovery)**.** *Let $\hat{\Theta}$ be the AV estimator and $\hat{\Theta}^t$ the thAV estimator with $t = \lambda C\hat{r}$, where C is the constant from Assumption* (4) *and $\lambda \in (0, 3]$. Then, it holds with probability $1-\delta$ that*

1. *for all $(i, j) \in \mathcal{V}$ such that $\Theta_{ij} = 0$ it is $|\hat{\Theta}_{ij}| \in [0,\, 3Cr_\delta^*]$ and therefore*

$$(i, j) \in \hat{\mathcal{E}} \iff |\hat{\Theta}_{ij}| \in (\lambda C\hat{r},\, 3Cr_\delta^*]\ .$$

[2] Note that even though we obtain a class of events $\big(\mathcal{T}_r\big)_{r\in\mathcal{R}}$ building on a similar interpretation as Chichignoud et al. (2014), we have to resort to a more involved primal-dual-witness construction to prove the validity of this upper bound.

[3] The generalized version we derived corresponds to Theorem 3 in Section A.1 of the supplement.

2. *for all* $(i,j) \in \mathcal{V}$ *such that* $|\Theta_{ij}| > (3+\lambda)Cr^*_\delta$ *it is* $(i,j) \in \hat{\mathcal{E}}$.

The proof of Corollary 4 can be found in Section A.3 of the supplement. As far as we know, there is no other theoretical result so far that justifies a specific choice for a threshold in a thresholded version of the graphical lasso. Moreover, the corollary offers a theoretical ground for balancing the tradeoff between false positive and false negative rate: while maintaining finite-sample guarantees, we can regulate λ according to our needs to decrease the false negative rate (part 2) at the cost of increasing the interval $(\lambda C\hat{r}, 3Cr^*_\delta]$ in which thAV missclassifies negatives (part 1).

Importantly, the thAV also comes with notable computational benefits, since the computations in Definition 1 only require at most 1 solution path. Using the *glasso* R package (Friedman et al., 2008; Witten et al., 2011), we can efficiently compute the thAV as described by Algorithm 1 in Section B of the supplement.

Finally, note that even though there exist theoretical bounds justifying (4) (see Section A.2 of the supplement), they are usually too loose or bounded to restrictions that are hard to interpret and violated in practice. Thus, what we observe in practice is usually not (4), but rather a more robust "quantiled version" of it:

$$\ell_{1-\alpha}(\Theta, \hat{\Theta}) \leq Cr \ ,$$

where $\ell_{1-\alpha}(\Theta, \hat{\Theta})$ defines the $1-\alpha$ quantile of the set of absolute differences $\{|\Theta_{ij} - \hat{\Theta}_{ij}|\}_{ij}$. Under this assumption, one could derive the same theory for the quantiled version of the loss, i.e. by replacing all ℓ by $\ell_{1-\alpha}$ in this section (compare with the general theory in Section A.1 of the supplement). However, in practice the results are very similar and our method is computationally less expensive.

3 Simulation study

In this section we compare the thAV to various other commonly used methods to estimate a Gaussian graphical model, which are the StARS (Liu et al., 2010), the scaled lasso (Sun & Zhang, 2012), the TIGER (Liu & Wang, 2017), the regularized score matching estimator (rSME) tuned via eBIC (Lin et al., 2016), and the SCIO[4] tuned via CV and via the Bregman-criterion (Liu & Luo, 2015)[5]. We sample synthetic data from a Gaussian distribution $\mathcal{N}_d(\mathbf{0}_d, \Theta^{-1})$, whereby we adopt a similar precision matrix generation procedure from Caballe et al. (2015) for sampling random and scale-free graphs. A detailed description of the based generation process can be found in Section C.1 of the supplement. We scale the data such that it is centered and has empirically unit variance.

If not stated differently, we use $t = C\hat{r}$ and $C = 0.7$ in the following. We define the set of possible regularization parameters to be $\mathcal{R} := \{0.05 + i(r_{\max} - 0.05)/40 : i \in \{1, ..., 40\}\}$, where $r_{\max} := \max_{i \neq j} |\hat{\Sigma}_{ij}|$ is the largest off-diagonal entry in absolute value of the empirical covariance matrix[6]. To enhance the robustness of our algorithm, we scale the graphical lasso estimators in Definition 1 such that they have unit diagonal entries. If not specified, the results of all experiments are averaged over 25 iterations and standard deviations are shown in parenthesis. Besides the experiments presented in this paper, we present additional investigations and repeat all experiment in various settings in Section C.3 of the supplement. We provide the code for all experiments with the submission, which can further be accessed through the public git repository.

Performance in F_1-score Table 1 shows the performance of the different methods for the graph recovery task. The performance is evaluated based on precision, recall, and the resulting F_1-score, which are defined in Section C.3 of the supplement. The proposed thAV estimator not only clearly outperforms the baseline methods but it also has a noticeable advantage over the oracle graphical lasso estimator, which is the (non-thresholded) graphical lasso estimator that achieves maximal F_1-score among all regularization parameters. This implies that it is mandatory to apply thresholding on top of regularized optimization to obtain good graph recovery results with the graphical lasso. Remarkably, in estimating a random graph, we observe that the thAV always achieves a recall of above 0.9 while maintaining good precision. This is in stark contrast to the other methods, which seem to overestimate the graphs resulting in a high recall but comparably low precision. Moreover, the results indicate that scale-free graphs in general are much harder to estimate than random graphs. As it has already been reported in other works (see Liu & Ihler (2011); Tang et al. (2015) and references therein), the graphical lasso is not able to provide a good estimation of a scale-free graph because its regularization does not impose any preference for identifying hub-like structures. Nevertheless, thAV remains superior to the other methods in terms of reaching the highest F_1-Score in most cases. Again, thAV can find a good balance between precision and

[4]The regularized score matching estimator (rSME) and the SCIO estimator solve the same optimization problem in the Gaussian setting.

[5]We have also evaluated RIC, which is the default graphical lasso calibration scheme in the *huge* R package (Zhao et al., 2012). However, we decided to exclude it from our simulation study due to bad results, computational instability, and a lack of theory.

[6]This is the smallest regularization parameter that estimates an empty graph.

Table 1: Graph recovery performance for varying graphs and sample size. The bold numbers indicate the best score in each setting.

	random			scale-free		
	F_1	Precision	Recall	F_1	Precision	Recall
$n = 300$, $d = 200$						
oracle	0.70 (0.13)	0.60 (0.16)	0.89 (0.03)	0.37 (0.12)	0.37 (0.22)	0.63 (0.23)
StARS	0.59 (0.14)	0.44 (0.13)	0.93 (0.09)	0.29 (0.13)	0.20 (0.10)	0.65 (0.12)
scaled lasso	0.68 (0.02)	0.52 (0.03)	0.98 (0.01)	0.40 (0.07)	0.26 (0.05)	0.84 (0.07)
TIGER	0.47 (0.09)	0.31 (0.08)	**0.99** (0.01)	0.34 (0.07)	0.21 (0.05)	**0.87** (0.07)
rSME (eBIC)	0.64 (0.17)	0.49 (0.16)	0.98 (0.01)	0.47 (0.23)	0.42 (0.24)	0.74 (0.14)
scio (CV)	0.19 (0.36)	0.23 (0.41)	0.17 (0.34)	0.15 (0.19)	0.44 (0.50)	0.09 (0.12)
scio (Bregman)	0.13 (0.17)	0.11 (0.22)	0.96 (0.16)	0.24 (0.18)	**0.51** (0.44)	0.57 (0.37)
thAV	**0.91** (0.03)	**0.90** (0.04)	0.93 (0.05)	**0.54** (0.13)	0.48 (0.19)	0.70 (0.13)
$n = 200$, $d = 300$						
oracle	0.70 (0.10)	0.63 (0.14)	0.81 (0.03)	0.29 (0.07)	0.25 (0.15)	0.47 (0.13)
StARS	0.54 (0.11)	0.39 (0.11)	0.93 (0.03)	0.25 (0.07)	0.17 (0.06)	0.54 (0.10)
scaled lasso	0.65 (0.03)	0.49 (0.02)	0.94 (0.02)	**0.30** (0.04)	0.20 (0.03)	0.59 (0.08)
TIGER	0.45 (0.08)	0.30 (0.07)	0.96 (0.02)	0.25 (0.05)	0.15 (0.04)	0.67 (0.09)
rSME (eBIC)	0.02 (0.00)	0.01 (0.00)	**0.99** (0.01)	0.01 (0.00)	0.01 (0.00)	**0.95** (0.02)
scio (CV)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.06 (0.09)	0.36 (0.49)	0.03 (0.05)
scio (Bregman)	0.26 (0.29)	0.27 (0.35)	0.86 (0.21)	0.16 (0.12)	**0.57** (0.47)	0.36 (0.27)
thAV	**0.79** (0.09)	**0.73** (0.15)	0.90 (0.04)	0.28 (0.11)	0.21 (0.14)	0.60 (0.11)
$n = 400$, $d = 200$						
oracle	0.74 (0.13)	0.65 (0.17)	0.91 (0.03)	0.42 (0.10)	0.42 (0.22)	0.67 (0.26)
StARS	0.63 (0.13)	0.48 (0.14)	0.96 (0.03)	0.34 (0.12)	0.23 (0.09)	0.70 (0.14)
scaled lasso	0.70 (0.03)	0.54 (0.03)	0.99 (0.01)	0.44 (0.05)	0.29 (0.04)	0.90 (0.07)
TIGER	0.48 (0.09)	0.32 (0.08)	0.99 (0.01)	0.33 (0.05)	0.20 (0.04)	**0.93** (0.06)
rSME (eBIC)	0.56 (0.20)	0.41 (0.18)	**1.00** (0.00)	0.54 (0.13)	0.43 (0.15)	0.82 (0.11)
scio (CV)	0.48 (0.45)	0.52 (0.47)	0.48 (0.46)	0.14 (0.25)	0.28 (0.45)	0.10 (0.19)
scio (Bregman)	0.16 (0.16)	0.12 (0.22)	0.97 (0.13)	0.21 (0.18)	0.35 (0.40)	0.71 (0.36)
thAV	**0.93** (0.04)	**0.92** (0.06)	0.95 (0.04)	**0.63** (0.13)	**0.59** (0.19)	0.75 (0.14)

recall, whereas methods such as StARS and TIGER are overestimating the graph, which results in comparably low precision.

In Table 1, it appears that the recovery performance drops with an increment of d, which makes sense since the number of parameters increases quadratically with d. However, in our next experiments (Table 2), in which we investigate the thAV in the setting $d \in \{600, \dots, 1000\}$ and set $n = 500$, we observe that this is surprisingly not the case when enough data, but still $n < d$, is available. The F_1-score for a random graph remains stable across all d at an impressive value of 0.96. In the case of a scale-free graph, the performance decays slowly, while maintaining a good trade-off between precision and recall. Note that the support recovery in the case $d = 1000$ involves about 500 000 parameters.

Moreover, the careful reader will actually realize that the proposed calibration scheme can also be employed to tune the rSME estimator. Because of the generality of our results from Section A.1 we can employ existing results from Lin et al. (2016) to verify the validity of Assumption (4) for the rSME. As it is shown in Section C.4 of the supplement, the rSME calibrated with the thAV approach performs comparably to the thAV graphical lasso. This is an important observation as the rSME can be applied to estimate the conditional dependency structure of a pairwise interaction model, which is a broader model class than the class of Gaussian graphical models. Hence, the calibration technique and the underlying theory can naturally be extended to the non-Gaussian setting.

Furthermore, we repeat the empirical study with the modified graphical lasso proposed by Liu & Ihler (2011), which we calibrate and clip via the thAV technique. This estimator employs a power law regularization that encourages the appearance of nodes with a high degree and are thus better suited for scale-free graphs. The experimental study is shown in Section C.4 of the supplement.

Dependence on C If we increase the constant C in the AV calibration scheme, we decrease $\hat{r}$ (see Proposition 9 of the supplement) and therefore employ less

Table 2: Graph recovery performance of the thAV with fixed sample size $n = 500$ for large-scale examples. The results are based on 20 iterations.

	RANDOM			SCALE-FREE		
	F_1	PRECISION	RECALL	F_1	PRECISION	RECALL
$d = 600$	0.96 (0.03)	0.97 (0.02)	0.94 (0.05)	0.43 (0.17)	0.45 (0.29)	0.61 (0.17)
$d = 700$	0.96 (0.02)	0.98 (0.02)	0.93 (0.04)	0.42 (0.16)	0.43 (0.25)	0.53 (0.13)
$d = 800$	0.95 (0.02)	0.97 (0.03)	0.93 (0.04)	0.40 (0.18)	0.45 (0.29)	0.52 (0.16)
$d = 900$	0.96 (0.02)	0.97 (0.04)	0.95 (0.03)	0.33 (0.17)	0.33 (0.26)	0.54 (0.16)
$d = 1000$	0.96 (0.02)	0.98 (0.01)	0.93 (0.03)	0.34 (0.17)	0.36 (0.27)	0.46 (0.15)

regularization. Hence, the AV estimator is inherently related to the choice of C. We plot the performance of different AV estimators with varying thresholds in Figure 1 and make two crucial observations. First, we see significant dissimilarities in the performance of the unthresholded AV estimators: because the calibrated regularization parameter ranges from $\hat{r} = 0.23$ in the case $C = 0.5$ to $\hat{r} = 0.09$ in the case $C = 0.8$, the F_1-score drops from approximately 0.70 to 0.35. Thus, the AV estimator's performance heavily depends on the choice of C. But after thresholding, and this is the second observation, the thresholded AV estimators' performance curves become very similar and reach almost the same peak. We can observe the same behaviour in the other settings, as it is shown in Section C.3 of the supplement. Importantly, we also show in the supplement that we do not observe a similar performance peak if we threshold the unregularized optimization problem (setting $r = 0$ in (2)). Hence, neither regularization via regularized optimization is sufficient for graph recovery, see the performance of the oracle estimator in Table 1, nor does unregularized thresholding yield to good results. Therefore we claim that it is necessary to apply both types of regularizations, as it is done by thAV and which additionally encourages stability in C.

To further investigate the stability of the thAV in C, we consider pairs of thAV estimators resulting from different choices of C, which we call $\hat{\Theta}^{t'}_{C'}$ and $\hat{\Theta}^{t'}_{C''}$. Table 3 reports the differences between these estimators by calculating $F_1(\hat{\Theta}^{t'}_{C'}, \hat{\Theta}^{t'}_{C''})$ for a random graph. We do not only achieve a high $F_1(\Theta, \hat{\Theta}^t_C)$ for any C, but also the different estimates are all very similar, i.e. $F_1(\hat{\Theta}^t_{C'}, \hat{\Theta}^t_{C''})$ is always above 0.80. Therefore, we can confirm that the recovered graphs are stable in the choice of C.

Varying graph density In this last experiment, we put emphasis on the adaptability of thAV on the density of the graph, i.e. the proportion of edges to the number of nodes in the graph. The graph's density of a random graph is controlled via p, the probability that a pair of nodes is being connected. Details can be found in Section C.1 of the supplement. We observe from Table 4 that the F_1-score of the thAV estimator remains stable across all densities, whereas the other estimators tend to perform better for dense graphs. This is no surprise since we have seen in the previous experiments that the other estimators tend to overestimate the presence of edges in a graph as indicated by a high recall but low precision. In all investigated settings, the thAV estimator outperforms the competing estimation procedures considerably.

Table 3: Similarity $F_1(\hat{\Theta}^t_{C'}, \hat{\Theta}^t_{C''})$ for different choices of C (C' and C'') for a random graph with $d = 200$ using $n = 300$ samples. The performance scores $F_1(\Theta, \hat{\Theta}^t_C)$ are 0.80 (0.06), 0.88 (0.04), 0.91 (0.05), 0.85 (0.10) for C in 0.5, 0.6, 0.7, 0.8, respectively.

C	0.6	0.7	0.8
0.5	0.88 (0.05)	0.82 (0.08)	0.81 (0.10)
0.6	1	0.93 (0.04)	0.84 (0.09)
0.7	-	1	0.92 (0.05)
0.8	-	-	1

4 Applications

Graphical model recovery plays a big role in understanding biological networks. In this section, we apply our procedure on 2 open-source data sets to obtain sparse and interpretable graph structures.

Recovering a Microbial Network It is believed that the human microbiome plays a fundamental role in human health. Thus, the American Gut Project (McDonald et al., 2018) was launched to pave the way to find associations among the microbiome, but also to confirm associations between the microbiome and other aspects of human health, like psychiatric stability. In this example, we estimate the microbial network to enhance the understanding of the roles and the relations between the microbes. Since microbial datasets come

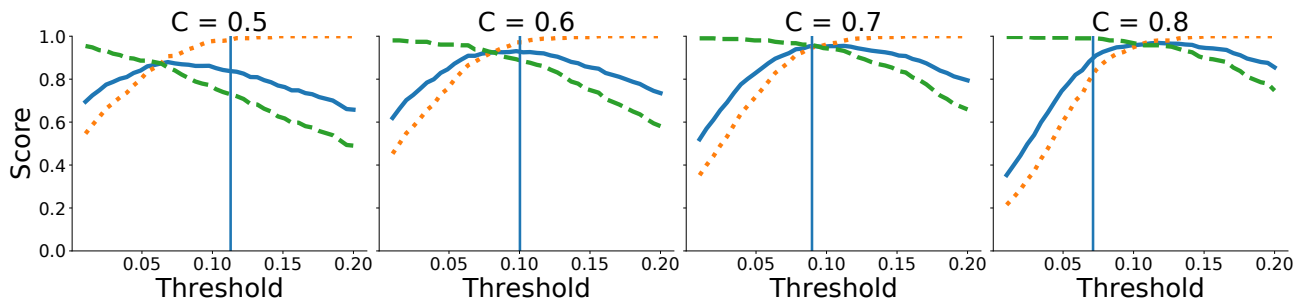


Figure 1: The F_1-score (blue, solid), precision (yellow, dotted), and recall (green, dashed) of a thresholded AV estimator for a random graph with $d = 200$ based on $n = 300$ samples for $C \in \{0.5, 0.6, 0.7, 0.8\}$ with varying thresholds. The resulting AV regularization parameters are 0.23, 0.17, 0.13, 0.09, respectively. The vertical line depicts the proposed threshold $t = C\hat{r}$ corresponding to the thAV estimate.

with some technical problems, it is vital to preprocess the data. We refer to the work of Kurtz et al. (2015) and Yoon et al. (2019) for details about the problems and a suitable preprocessing routine for microbial data. We use the preprocessed amgut2.filt.phy data, which is included in the *SpiecEasi* package (Kurtz et al., 2015). The data set measures the abundance of microbial operational taxonomic units (OTUs) and consists of $n = 296$ samples and $d = 138$ different OTUs. We employ the nonparanormal transformation (Liu et al., 2009) and estimate the microbial network using the thAV. The thAV estimator returns a very sparse graph that identifies various clusters, see Section D of the supplement. However, the estimator includes no interactions between different classes of microbes. To get insight about interactions across classes of microbes, we reduced the truncation parameter λ to 0.5. Note that the results of Corollary 4 are valid for each $\lambda \in (0, 3]$. The resulting graph is depicted in Figure 2(a).

Recovering a Gene Network In pharmacology, the vitamin *riboflavin* is industrially produced using diverse microorganisms. Being able to fully understand the bacterias' genome, biologists promise to further optimize the *riboflavin* production. The *riboflavin* data set is provided by the DSM in Switzerland and contains $n = 71$ samples and $d = 4088$ gen expressions. The R package *hdi* (Dezeure et al., 2015) provides this data set in its implementation. We compare the results of the thAV (see Figure 2(b)) with those of Bühlmann et al. (2014), who analyzed the same data set using a neighborhood regression approach[7], and observe that the thAV returns a much sparser graph with more cluster-like structures. This does not only increase the interpretability of the graph but also imposes some tight connections between several genes within these clusters.

[7]Similar to Bühlmann et al. (2014), we shrink the data set by only considering the 100 genes with the highest empirical variance and scale the data using the nonparanormal transformation.

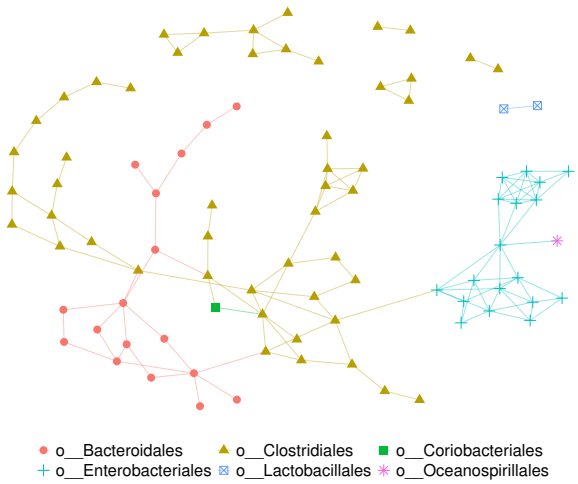
5 Discussion and Conclusion

Graphical models are a very popular framework for co-occurrence networks, and the graphical lasso is one of the most standard estimators in this framework. In this paper, we generalize the theoretical framework of Chichignoud et al. (2014) for deriving a calibration scheme for the lasso and successfully transfer it to calibrate the graphical lasso. However, our empirical study reveals that graphical lasso estimation itself is not sufficient for effective support recovery, so an additional thresholding step becomes necessary. Our resulting calibration method comes with a finite sample result that allows us to derive a corollary suggesting how to choose a theoretically founded threshold in such a thresholded graphical lasso approach. The resulting estimator, which we call thresholded adaptive validation (thAV) estimator, provides a simple and fast implementation with finite sample guarantees on the recovery performance. To the best of our knowledge, this is the first thresholding methodology for the graphical lasso that comes with a practical, theory-based threshold. Moreover, the thAV clearly outmatches existing graph recovery methods in our empirical analysis, showing both, a high recall but also a high precision in most settings. Other methods, which do not come with a practical threshold, tend to overestimate the graph. Thus, we would recommend the thAV as the method of choice for applications requiring an interpretable and sparse graph structure.

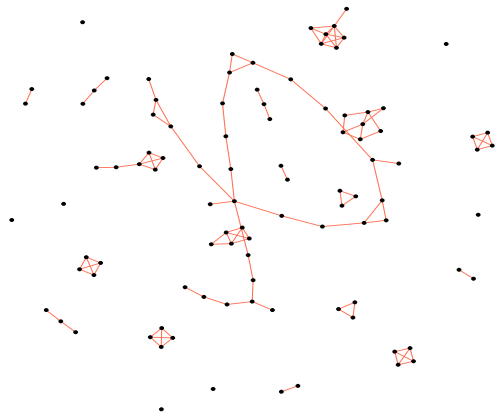
One shortcoming of the proposed procedure is that we replace the tuning parameter r by a quantity C, and we even introduce λ, which defines our threshold. However, regarding λ, we derive a finite sample result for every $\lambda \in (0, 3]$. And secondly, the correspondence between C and the AV regularization parameter

Table 4: F_1-score if we change the connection probability to $p \in \{2/d, 4/d\}$ in a random graph in various settings. The previous simulations employed $p = 3/d$. The bold numbers indicate the best score in each setting.

	$2/d$	$4/d$
$n = 300$, $d = 200$		
StARS	0.55 (0.14)	0.57 (0.13)
scaled lasso	0.60 (0.03)	0.73 (0.02)
TIGER	0.39 (0.09)	0.53 (0.08)
rSME (eBIC)	0.51 (0.27)	0.55 (0.23)
SCIO (CV)	0.33 (0.40)	0.22 (0.39)
SCIO (Bregman)	0.23 (0.29)	0.22 (0.23)
thAV	**0.91** (0.05)	**0.89** (0.04)
$n = 200$, $d = 300$		
StARS	0.49 (0.11)	0.52 (0.09)
scaled lasso	0.58 (0.03)	0.68 (0.02)
TIGER	0.35 (0.09)	0.48 (0.10)
rSME (eBIC)	0.01 (0.00)	0.03 (0.00)
SCIO (CV)	0.14 (0.26)	0.20 (0.31)
SCIO (Bregman)	0.31 (0.36)	0.33 (0.31)
thAV	**0.87** (0.03)	**0.73** (0.11)
$n = 400$, $d = 200$		
StARS	0.61 (0.16)	0.61 (0.12)
scaled lasso	0.60 (0.03)	0.74 (0.02)
TIGER	0.38 (0.07)	0.52 (0.06)
rSME (eBIC)	0.55 (0.22)	0.54 (0.14)
SCIO (CV)	0.21 (0.39)	0.35 (0.44)
SCIO (Bregman)	0.14 (0.20)	0.15 (0.01)
thAV	**0.94** (0.04)	**0.93** (0.03)



(a)



(b)

Figure 2: The thAV based on the American Gut Data is shown in Figure 2(a). To avoid too large graphics, we exclude isolated vertices. The color and the shape of a node imply the biological cluster of each OTU. Figure 2(b) depicts the thAV applied on the *ribloflavin* data.

leads to a threshold that regulates the impact of C on the thAV, resulting in an estimator that is stable in C. In contrast, the calibration via the regularization parameter r is not equipped with such a stability property. The introduction of new hyperparameters can also not be seen as a disadvantage in comparison to related methods, which replace the regularization parameter by other hyperparameters as well. For instance, the StARS calibration scheme for the graphical lasso introduces new parameters N and b, which define the number of N subsamples of size b, and the parameter β, which restricts the instability. TIGER introduces a new regularization parameter ξ and the authors (Liu & Wang, 2017) argue that the final problem is regularization-parameter-insensitive.

On the other hand, a major advantage of this calibration scheme is its generality. While we focus on the graphical lasso in this paper, the derived theoretical framework can also serve as a foundation for other thresholding approaches, which are not limited to Gaussian graphical modeling. For instance, the proposed method has the potential to effectively tune the rSME, which is a graph recovery method for the pairwise interaction model. Using the same primal dual-witness technique, the authors (Lin et al., 2016) prove the assumption on which our theoretical framework is based (see (4)). Hence, we can derive the same theory for these types of estimators using the adaptive validation technique. We note that many empirical results regarding the rSME are rather limited to ROC curves, which conceal the regularization parameter selection. Moreover, we can also employ the framework for methods that aim to recover the support of specific types of graph topologies, such as particularly scale-free graphs (Liu & Ihler, 2011).

For future work, we hope to apply the proposed gen-

eral framework to calibrate regularized optimization problems and equip these estimators with finite sample theoretical guarantees. These might include other applications of sparse precision matrix estimation such as high-dimensional discriminant analysis and portfolio allocation (see Fan et al. (2016) and references therein), but also applications beyond the scope of sparse precision matrix estimation.

Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC- 2092 CASA – 390781972. We thank Yannick Düren, Fang Xie, Mahsa Taheri, Shih-Ting Huang, Ute Krämer, Björn Pietzenuk, and Lara Syllwasschy for their insightful comments. Finally, we also thank the anonymous reviewers for their careful reading and their useful suggestions.

References

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation. 2007. arXiv:0707.0704.

Besag, J. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 36(2):192–225, 1974.

Bu, Y. and Lederer, J. Integrating additional knowledge into estimation of graphical models. 2017. arXiv:1704.02739.

Bühlmann, P., Kalisch, M., and Meier, L. High-dimensional statistics with a view toward applications in biology. *Annu. Rev. Stat. Appl.*, 1(1):255–278, 2014.

Caballe, A., Bochkina, N., and Mayer, C. Selection of the regularization parameter in graphical models using network characteristics. 2015. arXiv:1509.05326.

Chichignoud, M., Lederer, J., and Wainwright, M. A practical scheme and fast algorithm to tune the lasso with optimality guarantees. 2014. arXiv:1410.0247.

Denev, A. *Probabilistic graphical models: A new way of thinking in financial modelling*. Risk Books, 2015.

Dezeure, R., Buehlmann, P., Meier, L., and Meinshausen, N. High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statist. Sci.*, pp. 533–558, 2015.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.*, 90(1):196–212, 2004.

Fan, J., Feng, Y., and Wu, Y. Network exploration via the adaptive lasso and scad penalties. *Ann. Appl. Stat.*, 3(2):521–541, 2009.

Fan, J., Liao, Y., and Liu, H. An overview of the estimation of large covariance and precision matrices. *Econom. J.*, 19(1):C1–C32, 2016.

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

Grimmett, G. R. A theorem about random fields. *Bull. Lond. Math. Soc.*, 5:81–84, 1973.

Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.*, 11(5):e1004226, 2015.

Lauritzen, S. L. *Graphical models*, volume 17 of *Oxford statistical science series*. Clarendon, 1996.

Lin, L., Drton, M., and Shojaie, A. Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.*, 10(1):806, 2016.

Liu, H. and Wang, L. Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. *Electron. J. Stat.*, 11(1):241–294, 2017.

Liu, H., Lafferty, J., and Wasserman, L. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

Liu, H., Roeder, K., and Wasserman, L. Stability approach to regularization selection (stars) for high dimensional graphical models. *Adv. Neural Inf. Process. Syst.*, 24 2:1432–1440, 2010.

Liu, Q. and Ihler, A. Learning scale free networks by reweighted l1 regularization. *Proc. Mach. Learn. Res.*, 15:40–48, 2011.

Liu, W. and Luo, X. Fast and adaptive sparse precision matrix estimation in high dimensions. *J. Multivariate Anal.*, 135:153 – 162, 2015.

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., and Ackermann, e. a. American gut: An open platform for citizen science microbiome research. *mSystems*, 3(3):e00031–18, 2018.

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

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by

minimizing l1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980, 2011.

Sun, T. and Zhang, C.-H. Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.*, 14, 02 2012.

Tang, Q., Sun, S., and Xu, J. Learning scale-free networks by dynamic node specific degree prior. *ICML*, 37:2247–2255, 2015.

Thieffry, D., Huerta, A. M., Pérez-Rueda, E., and Collado-Vides, J. From specific gene regulation to genomic networks: A global analysis of transcriptional regulation in escherichia coli. *BioEssays*, 20 (5):433–440, 1998.

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

Witten, D. M., Friedman, J. H., and Simon, N. New insights and faster computations for the graphical lasso. *J. Comput. Graph. Statist.*, 20(4):892–900, 2011.

Yoon, G., Gaynanova, I., and Müller, C. L. Microbial networks in spring - semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Front. Genet.*, 10:516, 2019.

Yuan, M. and Lin, Y. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1): 19–35, 2007.

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. The huge package for high-dimensional undirected graph estimation in r. *J. Mach. Learn. Res.*, 13:1059–1062, 2012.

Zhuang, R. and Lederer, J. Maximum regularized likelihood estimators: A general prediction theory and applications. *Stat*, 7(1):e186, 2018.