

Beyond Lasso: A Survey of Nonconvex Regularization in Gaussian Graphical Models

Donald R. Williams
University of California, Davis

Studying complex relations in multivariate datasets is a common task in psychological science. Recently, the Gaussian graphical model has emerged as an increasingly popular model for characterizing the conditional dependence structure of random variables. Although the graphical lasso (ℓ_1 -regularization) is the most well-known estimator across the sciences, it has several drawbacks that make it less than ideal for model selection. There are now alternative forms of regularization that were developed specifically to overcome issues inherent to the ℓ_1 -penalty. To date, this information has not been synthesized. This paper provides a comprehensive survey of nonconvex regularization that spans from the smoothly clipped absolute deviation penalty to continuous approximations of the ℓ_0 -penalty (i.e., best subset) for directly estimating the inverse covariance matrix. A common thread shared by these penalties is that they all enjoy the oracle properties, that is, they perform as though the *true* generating model were known in advance. To ensure their theoretical properties are general, I conducted extensive numerical experiments that indicated superior and more than competitive performance when compared to glasso and non-regularized model selection, respectively, all the while being computationally feasible for many variables. In addition, the important topics of tuning parameter selection and statistical inference in regularized models are reviewed. The penalties are employed to estimate the dependence structure of post-traumatic stress disorder symptoms. The discussion includes several ideas for future research, including a plethora of information to facilitate their study. I have implemented the methods in the R package **GGMncv**.

Keywords: Gaussian graphical model, nonconvex regularization, lasso, oracle properties, one-step estimator

Studying complex relations in multivariate datasets is a common task in psychological science. Cognitive neuroscientists model brain connectivity with the goal of unearthing functional and structural associations between cortical regions (Ortiz, Munilla, Álvarez-Illán, Górriz, & Ramírez, 2015). In clinical psychology, researchers wish to better understand the intricate web of symptom interrelations that underlie mental health disorders (Borsboom, Cramer, Schmittmann, Epskamp, & Waldorp, 2011; McNally, 2016). To this end, Gaussian graphical modeling (GGM) has emerged as an oft-used tool in the chest of research psychologists. The basic idea is to characterize multivariate relations by learning the conditional dependence structure, that is, the partial correlation “network” (Epskamp & Fried, 2018; Epskamp, Waldorp, Mottus, & Borsboom, 2018). The cortical regions or symptoms are “nodes” and the featured connections linking nodes are “edges” that graphically represent the

conditional dependence structure.

Inherent to Gaussian graphical modeling is the problem of *covariance* selection (Dempster, 1972), in that the off-diagonal elements of the inverse covariance (precision) matrix capture the dependence structure, that, when standardized and signed reversed, yield partial correlation coefficients. Even with a moderate number of variables, this can pose computational challenges, because, with p variables, there are $\frac{1}{2} \cdot p(p-1)$ potential edges.¹ Hence, the model space of possible conditional dependence graphs is $2^{\frac{1}{2}p(p-1)}$ (Mohammadi & Wit, 2015a). As a result, the GGM has served as a methodological testing grounds of sorts, where algorithms are continuously developed with the goals of accurate model selection and computational feasibility (for but a few of the possibilities see Fan, Liao, & Liu, 2016).

Perhaps the most widely known method is the graphical lasso (a.k.a., “least absolute shrinkage and selection operator,” Friedman, Hastie, & Tibshirani, 2008; Tibshirani, 1996; Witten, Friedman, & Simon, 2011), which adds a penalty to

DRW was supported by a National Science Foundation Graduate Research Fellowship under Grant No. 1650042. R code to reproduce the simulations is provided [online](#).

¹On the other hand, in the ubiquitous case of model selection in regression there are only p potential coefficients (the number of possible models is 2^p).

the estimation function that is based on the sum of the absolute values of the edges (i.e., ℓ_1 -regularization). The effect of this penalty is to push smaller estimates to exactly zero, thereby achieving covariance selection. Although psychological data is rarely high-dimensional ($n < p$, Williams & Rast, 2019; Williams, Rhemtulla, Wysocki, & Rast, 2019), which indicates estimation accuracy is not (usually) a problem (Williams & Rodriguez, 2020), it is important to note that glasso and related forms of regularization are still computationally attractive. This is because best subset model selection with, say, the maximum likelihood estimator (MLE), requires an intensive combinatorial search—this quickly becomes an insurmountable challenge for covariance selection in particular. This computational advantage has contributed to the widespread use of ℓ_1 -regularization in the sciences, including psychology, where it emerged as the *de facto* default in clinical applications (Epskamp & Fried, 2018).

However, in the statistics literature, it has long been known that the ℓ -penalty has several drawbacks, including that it is less than ideal for model selection. In fact, Mazumder, Friedman, and Hastie (2011) explicitly noted the following:

...greedier methods like subset regression and the nonconvex methods we discuss here achieve sparser models than the lasso for the same or better prediction accuracy, and enjoy superior variable-selection properties (p. 1125).

This quote captures that ℓ_1 -regularization is known to have an inflated false positive rate (Figure 5 in Williams et al., 2019) and will not necessarily converge on the *true* model (Figure 3 in Williams & Rast, 2019). Indeed, it has also been suggested to “interpret the second ‘s’ in lasso as screening rather than selection” (p. 278, Tibshirani, 2011). In regression, this is due to a rather strong assumption known as the irrepresentable condition (Zhao & Yu, 2006) and mutual incoherence (Wainwright, 2009). This is unlikely to hold with many variables, unless the ground truth is extremely sparse (see Table 1 in Zhao & Yu, 2006), which is not typically the case in psychological applications (see Table 2 in Wysocki & Rhemtulla, 2019). Note that an analogous assumption similarly applies in GGMs (see Equation 28 in Ravikumar, Wainwright, Raskutti, & Yu, 2011). These limitations form the primary impetus for nearly two decades of ongoing research exploring alternative forms of regularization. Unfortunately, these alternatives have yet to find their way into the psychological literature. For example, lasso was recently described for model selection in regression (McNeish, 2015), all the while penalties known to overcome the limitations of lasso have been readily available to research psychologists for decades (e.g., Fan & Li, 2001). Bringing the psychological literature up to date with the current statistical literature is thus long overdue.

In addition to model selection consistency² (or lack thereof), there are other important limitations of ℓ_1 -regularization. In particular, it does not enjoy the so-called oracle properties:

1. The *true* model is selected, with a probability tending to 1, as the sample size becomes large.
2. The parameters are asymptotically normal and they have the same asymptotic variance as non-regularized estimation.

In other words, “the penalized likelihood estimators work as well as if the correct submodel were known in advance” (p. 1348, Fan & Li, 2001). Other desirable properties include *unbiasedness*, which runs contrary to the common, albeit naive view, that trading in variance for bias is necessarily desirable. This is because removing the bias is required to make statistical inference (Javanmard & Montanari, 2013) and it can also improve predictive accuracy (p. 16, Hastie, Tibshirani, & Wainwright, 2015). And note that the bias of lasso does not necessarily dissipate as $n \rightarrow \infty$, that is, “in general they [non-zero estimates] are not consistent” (p. 91, Friedman, Hastie, & Tibshirani, 2001). There are a variety of approaches to *debias* the estimates, such as re-estimating the model with only the selected predictors using ordinary least squares (i.e., lasso + OLS, Belloni & Chernozhukov, 2013), the desparsified estimator of Javanmard and Montanari (2013), and the relaxed lasso (Meinshausen, 2007). An alternative for bias mitigation has been investigating nonconvex penalization. To this end, Fan and Li (2001) introduced the smoothly clipped absolute deviation (SCAD) penalty that enjoys the oracle properties and it reduces bias to the large edges (those that are likely to be nonzero). Likewise, another popular choice for improved model selection is the minimax concave penalty of C. H. Zhang (2010).

Although SCAD and MCP are the most well-known nonconvex penalties, there are several others that have been developed over the years. And there is reason to suspect that these may have competitive performance. For example, while best subset selection (a.k.a, the ℓ_0 penalty) is computationally intensive, as it requires fitting every possible model (but see Hazimeh & Mazumder, 2020), it is often considered the gold standard for model selection. This is because the penalty is applied directly to the number of parameters, that is, the ℓ_0 (pseudo) norm. Hence, forms of regularization that approximate the ℓ_0 penalty are highly desirable. This was explicitly noted in Zhao and Yu (2006), that interestingly enough also proved the irrepresentable condition:

Therefore, to get universal consistency, we need to reduce the amount of shrinkage on the parameters that are away from zero and regularize in a

²This refers to selecting the true model as $n \rightarrow \infty$

more similar fashion as ℓ_0 -penalty. However, as a consequence, this breaks the convexity of the lasso penalty, therefore more sophisticated algorithms are needed for solving the minimization problems (p. 2554).

There are now a variety of nonconvex penalties, including algorithms for their estimation, that were specifically developed with those goals in mind. To date, however, this information has not been synthesized in the psychological literature.

This paper includes three major contributions. First, nonconvex penalties, spanning from the SCAD and MCP to those that approximate the ℓ_0 -penalty, are thoroughly reviewed with the most common psychological applications in mind (i.e., assuming there are sufficiently more observations than variables or low-dimensional data). Although these penalties have purported benefits compared to ℓ_1 -regularization, they have not been compared with data common to psychology. Hence, the second contribution is a simulation study characterizing their performance for model selection. This includes determining whether they offer advantages compared to non-regularized methods. In contrast to convex problems, nonconvex optimization presents several challenges. To this end, the final major contribution is a review of computational approaches for implementing nonconvex penalties in Gaussian graphical models. To my knowledge, no software exists for this purpose, so I have implemented the methodology in the R package **GGMncv**.

Additionally, there are important aspects of regularization that have not been discussed in the psychological literature. This includes turning insensitive regularization and approaches for making statistical inference which is not straightforward in regularized models. These topics are also discussed, with the hope that nonconvex penalties are used judiciously in substantive applications.

The paper is organized as follows. The Gaussian graphical model is first introduced. The next section includes a review of nonconvex penalties. The following section reviews approaches for selecting the tuning parameter for nonconvex penalties. A simulation study is then provided, where various nonconvex penalties are compared to non-regularized estimation and ℓ_1 -regularization with simulation, with the goal of determining their suitability for psychological applications. The following section reviews challenges and approaches for making statistical inference. The discussion includes several ideas for further research into the properties of nonconvex regularization.

The Gaussian Graphical Model

A Brief Note on Generality

In this work, I assume that the data are continuous and normally distributed, that is, multivariate Gaussian. Accord-

ingly, I rely heavily upon the Pearson partial correlation coefficient to keep the exposition manageable. This does not limit the generality of this work, in that all ideas can seamlessly be applied to polychoric (Pearson, 1900), Spearman's rank (S. Kim, 2015), the so-called Gaussian rank estimator (i.e., based on Van Der Waerden scores, see references in Boudt, Cornelissen, Croux, & Boudt, 2012), and Kendall's tau based partial correlations (Johnson, 1979), each of which are commonly used in the Gaussian graphical modeling literature (Hoff, 2007; Liu, Han, Yuan, Lafferty, & Wasserman, 2012; Mohammadi & Wit, 2015a). This far-reaching applicability is due to requiring only an estimate of the covariance matrix.

Model Formulation

For multivariate normal data (Baba, Shibata, & Sibuya, 2004; Baba & Sibuya, 2005), a GGM captures conditional relations that are typically visualized to infer the underlying dependence structure (i.e., the partial correlation “network”; Højsgaard, Edwards, & Lauritzen, 2012; Lauritzen, 1996). There is an undirected graph that is denoted $G = (V, E)$, consists of a vertex $V = \{1, \dots, p\}$ and an edge set $E \subset V \times V$. The former refers to “nodes” that are, say, symptoms in a psychopathology scale, whereas E is the estimated network structure. Let $\mathbf{y} = (y_1, \dots, y_p)'$ be a random vector indexed by the graph's vertices that is assumed to follow a multivariate normal distribution, $\mathbf{y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with the mean vector $\boldsymbol{\mu} = (0_1, \dots, 0_p)'$ and a $p \times p$ positive definite covariance matrix $\boldsymbol{\Sigma}$. In the following, we use \mathbf{Y} to denote the $n \times p$ data matrix, where each row corresponds to the observations from the i th individual $i = \{1, \dots, n\}$.

The undirected graph is obtained by determining which off-diagonal elements in the precision matrix, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, are nonzero. That is, $(i, j) \in E$ when nodes i and j are determined to be conditionally dependent and set to zero otherwise. Note that standardizing $\boldsymbol{\Theta}$ and reversing the sign yields partial correlations, that is,

$$\rho_{ijz} = \frac{-\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}, \quad (1)$$

where z contains the nodes conditioned on. Note that it is possible to determine E by using Bayesian (Williams & Mulder, 2020) or frequentist hypothesis testing for each partial correlation (Drton & Perlman, 2005). When using nonconvex or ℓ_1 -regularization (that is convex), however, the standard approach is to work with the inverse correlation matrix, \mathbf{R}^{-1} , that is,

$$\boldsymbol{\Theta} = \mathbf{R}^{-1} = (d\boldsymbol{\Sigma}d)^{-1}, \quad (2)$$

where d is a diagonal matrix with $d_{ii} = 1/\sqrt{\sigma_{ii}}$. This is done to ensure that the parameters are on the same scale. In what follows, the off-diagonal elements of $\boldsymbol{\Theta}$ are denoted as θ .

Nonconvex Regularized Estimation

To estimate the conditional dependence structure for a class of nonconvex penalties,³ the penalized likelihood is defined as

$$l(\mathbf{R}^{-1}) = \log \det \mathbf{R}^{-1} - \text{tr}(\mathbf{R}\mathbf{R}^{-1}) - \sum_{i \neq j} p_{\lambda_{ij}, \gamma}(|\mathbf{R}_{ij}^{-1}|), \quad (3)$$

where \mathbf{R} is the sample correlation matrix and $p_{\lambda, \gamma}(\cdot)$ is a generic penalty function (Gao, Pu, Wu, & Xu, 2009), with the regularization parameter, λ , that controls the amount of penalization, such that there are a range of possibilities, depending on λ and γ . The additional parameter γ influences the shape of the penalty (Breheny & Huang, 2011). One point of emphasis is the subscripts on λ_{ij} that effectively allows each parameter to be differentially regularized. It is this feature that differentiates nonconvex from ℓ_1 -regularization. Indeed, by replacing $p_{\lambda, \gamma}(\cdot)$ with $p_{\lambda}(\cdot)$ in Equation 3, this results in the ℓ_1 penalized likelihood. Assuming a common value for λ is less than ideal because it is agnostic towards the signal (or lack thereof) in the data. Accordingly, “the lasso penalty increases linearly in the magnitude of its argument. As a result, it produces substantial biases in the estimates for large regression coefficients” (p. 524, Fan, Feng, & Wu, 2009). As described in Zhao and Yu (p. 2554, 2006), this bias (or “over-shrinkage of the nonzeros”) contributes to the issues of ℓ_1 -regularization for model selection. Nonconvex penalization can reduce this bias, in that, as described below, the penalty is permitted to be a function of θ_{ij} . This can allow values likely to be nonzero the chance to escape regularization, whereas, depending on the penalty function, values likely to be zero are penalized harshly.

Connection to Bayesian Regularization

Before proceeding to describe various nonconvex penalties, it is worth highlighting an interesting connection to the current Bayesian literature. The most recent regularization approaches, such as the horseshoe prior distribution (Y. Li, Craig, & Bhadra, 2017; Piironen & Vehtari, 2017), also aim to reduce the bias incurred by the Bayesian lasso (Khondker, Zhu, Chu, Lin, & Ibrahim, 2013; T. Park & Casella, 2008; H. Wang, 2012). The basic idea is eloquently stated in Carvalho, Polson, and Scott (2009):

The horseshoe prior has two interesting features that make it particularly useful...Its flat, Cauchy-like tails allow strong signals to remain large (that is, un-shrunk) a posteriori. Yet its infinitely tall spike at the origin provides severe shrinkage for the zero elements (p. 74).

These are the very advantages of using nonconvex penalties, especially those that approximate the ℓ_0 (pseudo) norm,

under a penalized likelihood framework. Although a comparison with Bayesian methods is beyond the scope of this paper, it suffices to note that the above goals are readily accomplished with nonconvex regularization, with the added benefit of avoiding the cumbersome nature of MCMC algorithms. I refer interested readers to Van Erp, Oberski, and Mulder (2019) for an overview of Bayesian regularization (including the horseshoe prior distribution).

Nonconvex Penalties

In this section, several nonconvex penalties are introduced. This cannot be an exhaustive review, as there are numerous penalties that are very similar to one another. In these cases, the reader is provided with relevant references and the most common penalty is presented. Furthermore, although there are nonconvex penalties that have three tuning parameters (see section 2.1 in D. Kim, Lee, & Kwon, 2018), this review focuses on those penalties with one additional tuning parameter (denoted γ in Equation 3). Finally, the relevant literature is faithfully summarized but it should be noted that it remains unclear whether their theoretical properties generalize to a variety of situations (this is fully investigated in Section Numerical Experiments).

Smoothly Clipped Absolute Deviation

The lasso was introduced in Tibshirani (1996). Not long after, five years to be exact, it was conjectured in Fan and Li (2001) that ℓ_1 -regularization does not enjoy the oracle property. The theoretical results in Fan and Li (2001) demonstrated that the ℓ_1 -penalty can be consistent for model selection and consistently estimate the parameters, but it cannot satisfy both properties *simultaneously* (see Theorem 2 and Remark 1 in Fan & Li, 2001). For the former, this requires that $\sqrt{n}\lambda \rightarrow \infty$, whereas, for the latter, root- n consistency requires that $\lambda = O(n^{-1/2})$ (p. 1353, Fan & Li, 2001). To achieve both properties simultaneously, Fan and Li (2001) proposed the SCAD penalty defined on $[0, \infty]$, that is,

$$p_{\lambda, \gamma}(\theta) \begin{cases} \lambda|\theta| & \text{if } |\theta| \leq \lambda \\ \frac{2\gamma\lambda|\theta| - \theta^2 - \lambda^2}{2(\gamma - 1)} & \text{if } \lambda < |\theta| \leq \gamma\lambda \\ \frac{\lambda^2(\gamma + 1)}{2} & \text{if } |\theta| > \gamma\lambda \end{cases} \quad (4)$$

for $\lambda \geq 0$ and $\gamma > 2$. Note that the penalty is clearly a function of θ (Figure 1). On the other hand, the ℓ_1 -penalty function is simply $\lambda|\theta|$. Although the formulation in Equation 4 opens the door for various possible combinations of λ and γ , setting $\gamma = 3.7$ is a common choice that was motivated by minimizing Bayesian risk (p. 1351, Fan & Li,

³This general class includes those nonconvex penalties with two tuning parameters.

2001). Furthermore, it was shown that selecting γ with, say, cross-validation, did not provide notable benefits.

In addition to the oracle properties, [Fan and Li \(2001\)](#) outlined three more desirable characteristics for a penalty, each of which is satisfied by the SCAD:

1. *Unbiasedness*: The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias.
2. *Sparsity*: The resulting estimator...automatically sets small estimated coefficients to zero to reduce model complexity.
3. *Continuity*: The resulting estimator is continuous...to avoid instability in model prediction (p. 1349, [Fan & Li, 2001](#)).

These are important to acknowledge because they have served as benchmarks for developing nonconvex penalties over the years. Indeed, all of the following penalties have these three characteristics, in addition to the oracle properties.

Minimax Concave Penalty

Another popular nonconvex penalty is the MCP that was introduced in ([C. H. Zhang, 2010](#)). The penalty function, again defined on $[0, \infty]$, is written as

$$p_{\lambda, \gamma}(\theta) \begin{cases} \lambda\theta - \frac{\theta^2}{2\gamma} & \text{if } \theta \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & \text{if } \theta > \gamma\lambda \end{cases} \quad (5)$$

for $\lambda \geq 0$ and $\gamma > 1$. It is common to set γ to 2.0, although selecting it is also a possibility. The MCP closely resembles the SCAD, that is, “Both penalties begin by applying the same rate of penalization as the lasso, and reduce that rate to 0 as θ gets further away from zero” (p. 237, [Breheny & Huang, 2011](#)). To make sense of this, it is informative to consider the derivative of Equation 5, that is,

$$p'_{\lambda, \gamma}(\theta) \begin{cases} \lambda - \frac{|\theta|}{\gamma} & \text{if } \theta \leq \gamma\lambda \\ 0 & \text{if } |\theta| > \gamma\lambda \end{cases} \quad (6)$$

Hence, when $|\theta|$ is greater than $\gamma\lambda$, the rate of penalization is zero. Note one key distinction between the SCAD and MCP is when the penalty is relaxed for increasingly larger values. This is illustrated in Figure 1.

In [C. H. Zhang \(2010\)](#), it was proven that the MCP does not require the irrerepresentable condition, which is *necessary* for consistent model selection for the ℓ_1 -penalty, although it rarely holds in practice. This theoretical result has a direct bearing on psychological applications in particular. At its crux, the irrerepresentable condition states that the important

and unimportant predictors cannot be correlated (at least not too much, i.e., the total “irrelevant” covariance is below 1). However, in fields such as clinical psychology, it is common to estimate the dependence structure of psychometric scales that, by construction, contain highly correlated variables. Although beyond the scope of this work, it is possible to evaluate this assumption in GGMs (i.e., Equation 28 in [Ravikumar et al., 2011](#)). This was formally done in [Heinävaara, Leppä-aho, Corander, and Honkela \(p. 104 of 2016\)](#) for gene regulatory “networks,” and I too have evaluated it and found the assumption essentially never holds, unless there is only, say, three or four variables, which is very uncommon (Table 1 in [Wysocki & Rhemtulla, 2019](#)). The failure of glasso for these kinds of data was recently highlighted in [Williams and Rast \(2019\)](#) and [Williams et al. \(2019\)](#). An open question, to be addressed below, is whether the MCP (and the other penalties) holds up across a gamut of situations.

Approximating the ℓ_0 -Penalty

Recall in the introduction that it was noted that a desirable penalty would closely mimic the behavior of best subset selection. Although this is not accomplished by either SCAD or MCP, there are a variety of nonconvex penalties that were designed specifically to approximate the ℓ_0 -penalty, that is,

$$p_{\lambda}(\theta) = \lambda I\{\theta \neq 0\} = \begin{cases} \lambda & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0. \end{cases} \quad (7)$$

This penalty is ideal because the “estimators are penalized according to the number of nonzero parameter estimates; thus, model complexity is penalized in a straightforward and intuitive way” (p. 929, [Dicker, Huang, & Lin, 2013](#)). For the goal of model selection, there are two main drawbacks of best subset selection. The first is computational, because it is a so-called *NP*-hard problem that requires exploring all possible combinatorial subsets. This is perhaps not an insurmountable challenge for normal data, given multiple regression can be used to estimate the dependence structure (“neighborhood selection” [Meinshausen & Bühlmann, 2006](#)), but directly estimating Θ is necessary for ordinal data with few categories in particular ([Rhemtulla, Brosseau-Liard, & Savalei, 2012](#)). Second, regarding the three properties described above, the solution is discontinuous rather than exhibiting continuity. This can translate into the estimator being relatively unstable (p. 2354 in [Breiman, 1996](#)).

Notice that these limitations have little to do with the overall performance for model selection, as it has long been known that (non-regularized) best subsets with, say, the Bayesian information criterion, will provide consistent model selection in the familiar setting of a fixed number of variables and an increasing sample size (e.g., [Casella, Girón, Martínez, & Moreno, 2009](#)). This setting is relevant to psychology, for example, when using a psychometric scale, p is

indeed fixed and the objective is to have an adequate sample size, or “power,” to detect the majority of edges.⁴ Hence, the key advantage of employing regularization in these situations for model selection is computational in nature. This also applies to the following penalties, where the idea is to have the performance of best subset selection, all the while providing a computationally efficient and continuous solution.

Seamless ℓ_0 Penalty. For multiple regression, [Dicker et al. \(2013\)](#) introduced the seamless ℓ_0 -penalty (SELO). This can be written as

$$p_{\lambda,\gamma}(\theta) = \frac{\lambda}{\log(2)} \log\left(\frac{|\theta|}{|\theta| + \gamma} + 1\right), \quad (8)$$

for $\gamma > 0$. Note that when $\gamma \rightarrow 0$, say, $\gamma = 0.01$, this approximates $p_{\lambda}(\theta) \approx \lambda I\{r \neq 0\}$ in Equation (7). Hence, γ provides the desired property of continuity that overcomes this issue of best subsets. This can be seen in Figure 1, where the SELO and ℓ_0 penalties are juxtaposed. Importantly, the performance of SELO does not depend heavily on γ (Table 2 in [Dicker et al., 2013](#)). As a result, it is common to set $\gamma = 0.01$. The SELO has proven quite influential, for example, it has since been extended to generalized linear ([Z. Li, Wang, & Lin, 2012](#)), Cox regression ([Y. Shi, Xu, Cao, & Jiao, 2019](#)), multivariate panel ([H. Zhang, Sun, & Wang, 2013](#)), proportional hazard ([Yongxiu, Yuling, Yueyong, & Yanyan, 2018](#)), and quantile regression models ([Ciuperca, 2015](#)). Furthermore, [Y.-y. Shi, Cao, Yu, and Jiao \(2018\)](#) generalized SELO (GSELO) in that the original formulation in Equation (8) was shown to belong to a more general family of penalties (see Table 1 in [Y.-y. Shi et al., 2018](#)).

Atan Penalty. In [Y. Wang and Zhu \(2016\)](#), an arctangent penalty was introduced (ATAN). The ATAN penalty function is defined on $[0, \infty)$ and it is written as

$$p_{\lambda,\gamma}(|r|) = \lambda\left(\gamma + \frac{2}{\pi}\right) \arctan\left(\frac{|\theta|}{\gamma}\right), \quad (9)$$

for $\gamma > 0$. [Y. Wang and Zhu \(2016\)](#) suggested fixing γ to 0.01. A key feature of Equation (9) is that there is a direct relationship between the ℓ_0 and ℓ_1 penalties. That is, as $\gamma \rightarrow \infty$, this results in ℓ_1 -regularization. On the other hand, as $\gamma \rightarrow 0$, this approximates the ℓ_0 -penalty. This differentiates the ATAN from the SELO penalty, as γ in the SELO does not govern a continuous transition from the ℓ_1 to ℓ_0 penalties. This is illustrated in Figure 1.

Illustrative Penalty Functions

Figure 1 includes each penalty function for several choices of γ ($\lambda = 1$). In each panel, as a reference point, the ℓ_0 and ℓ_1 penalties are also included. First note the discontinuity of the ℓ_0 -penalty, that is, it has an “all-or-nothing nature” ([Van Derwerken, 2011](#)). This aggressiveness can produce instability ([Breiman, 1996](#)). In contrast, although the ℓ_1 -penalty function is continuous, it increases linearly with the size of θ .

This results in nonvanishing bias, even with the sample size is large. Intuitively, this is the opposite of what we would like to be the case, in that, ideally, large effects should not be penalized harshly. This is one useful aspect of nonconvex regularization that can be seen for each penalty in Figure 1.

For the SCAD and MCP, it is clear that they resemble the ℓ_1 -penalty for small values of θ , but the penalties taper for larger values of θ . The difference is “the way that they make the transition” (p. 10, [Breheny & Huang, 2011](#)), with the MCP immediately parting ways from ℓ_1 -regularization. Recall that the SELO and ATAN penalties approximate the ℓ_0 -penalty. For both, as $\gamma \rightarrow 0$, the penalty function becomes more similar to best subset selection. However, it is also clear that the penalty is continuous which satisfies the continuity property described in [Fan and Li \(2001\)](#). Of note is the ATAN penalty, where, as γ increases from zero, it gradually moves from the ℓ_0 to the ℓ_1 -penalty. I find this property lends itself to an intuitive understanding of approximating the ℓ_0 penalty (Equation 1).

Illustrative Regularization Paths

Figure 2 includes regularization paths for each penalty. These were obtained from 20 post-traumatic stress disorder symptoms ([Armour et al., 2017](#)). This illustrates the partial correlations across a range of λ values. The nature of each penalty function can be seen in the paths. For example, for ℓ_1 -regularization, large values are heavily penalized, which is expected given the linear increase in the penalty function. Moreover, the regularization happened immediately (with small λ values), and, at some point in the path, all relations were eventually “pushed” to zero.

The nonconvex penalties have much different paths. Notice that large values escaped regularization early in the path, especially for the SELO and ATAN penalties. On the other hand, small values were quickly pushed to zero (e.g., for the ATAN penalty). Again, this is advantageous because it can produce nearly unbiased estimates, which is a desirable feature for parameter estimation ([Fan & Li, 2001](#)) and prediction ([Hastie et al., 2015](#)), in addition to improving model selection (e.g., when the irrerepresentable condition does not hold, [Zhao & Yu, 2006](#)). Second, several relations were never “pushed” to zero, even for very large λ values. This is due to the additional tuning parameter, γ , which ensures that values likely to be nonzero can “escape” penalization. In other words, signal in the data is incorporated into the estimator.

⁴In general, p is typically much smaller than n in the social-behavioral sciences. The ℓ_1 -penalty, on the other hand, is often studied, including rather elaborate proofs therein, when $n < p$. This entails strong assumptions (e.g., sparsity) over and above the customary assumptions (e.g., independent errors).

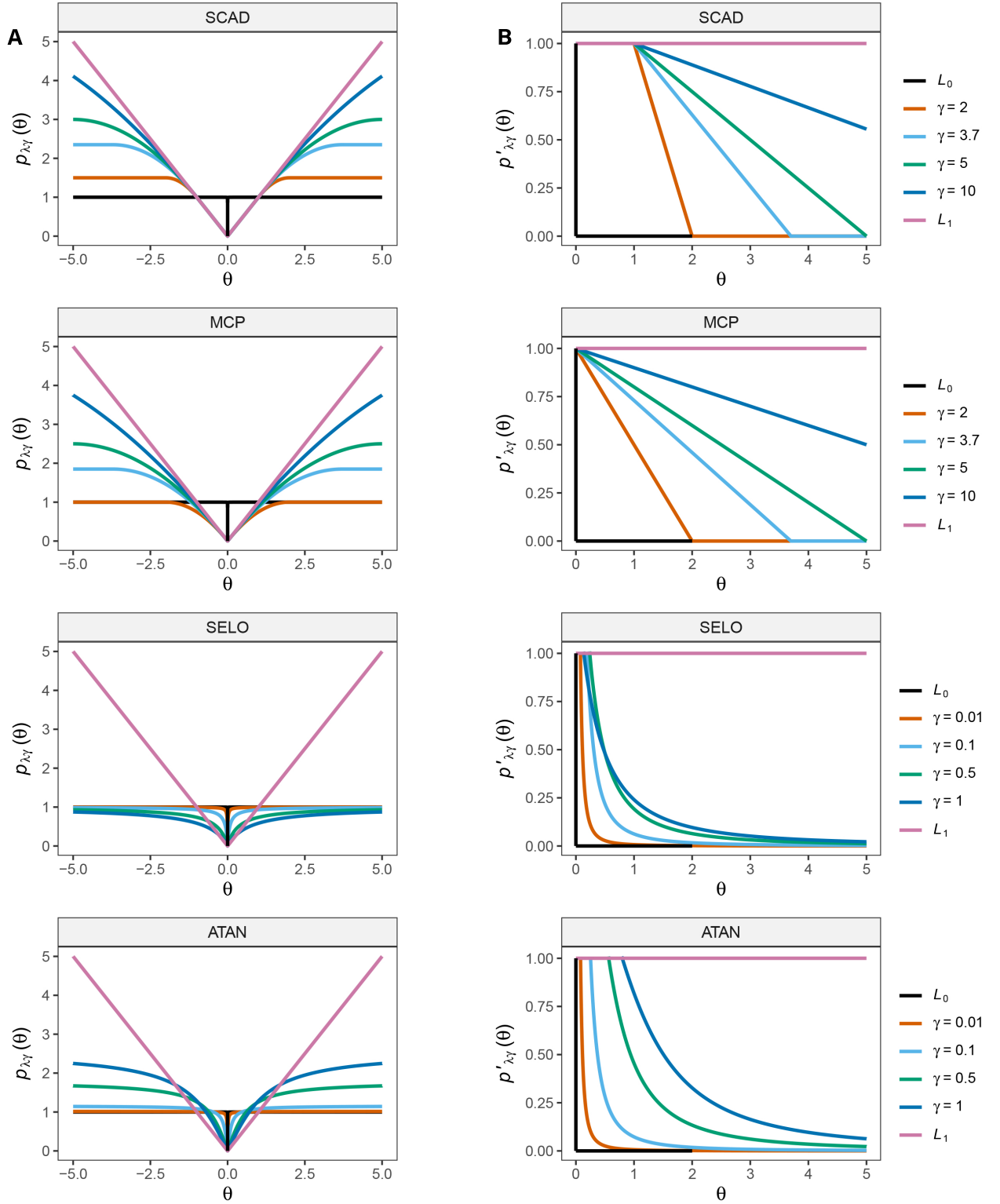


Figure 1. Panel A includes example penalty functions $p_{\lambda\gamma}(\theta)$. The ℓ_1 penalty increases linearly with the size of θ . In nonconvex regularization, the penalty diminishes for larger values (a desirable feature). Panel B includes the respective derivatives $p'_{\lambda\gamma}(\theta)$. The SELO and ATAN penalty derivatives approach that of the ℓ_0 penalty (i.e., best subset) as $\gamma \rightarrow 0$, whereas the SCAD and MCP begin with same rate as lasso but then taper off to zero (notice that the MCP immediately departs ways from the ℓ_1 penalty).

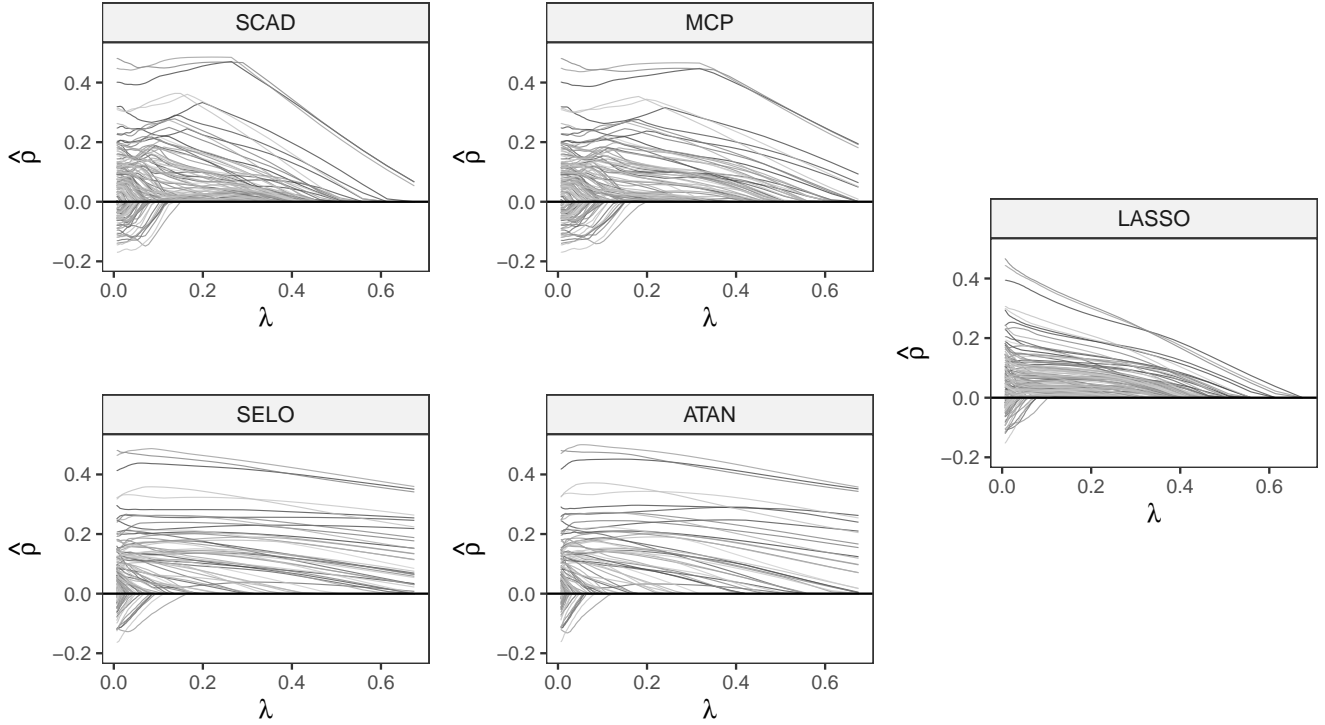


Figure 2. Example regularization paths for partial correlations estimated from 20 PTSD symptoms (Armour, Fried, Deserno, Tsai, & Pietrzak, 2017). The one-step estimator in Algorithm 2 was used for the nonconvex penalties, with γ fixed to 3.7 (SCAD), 2 (MCP), 0.1 (SELO), and 0.1 (ATAN). Note that early on (small λ values), when using the SELO and ATAN penalties in particular, the larger values “escaped” regularization. Conversely, the linearity of the ℓ_1 penalty is readily apparent and large values are penalized early in the path.

Computational Details

The success of lasso is in large part due to computational feasibility (not because it is superior than alternatives). This allows for scaling to *ultra* high-dimensional data, such as genetic regulatory networks, that often include more than 1,000 variables (and thus potentially millions of edges). For extensions to glasso, say, jointly estimating several GGMs, the number of variables investigated are very rare in psychological applications ($p = 500$ and $p = 1000$, Danaher, Wang, & Witten, 2014). The computational advantage is due to the convexity of the ℓ_1 objective function. This allows for developing highly efficient algorithms (i.e., via convex optimization). The first major advancement was the Least Angle Regression (LARS) that was introduced with the goal of providing a “less greedy version of traditional forward selection methods” (p. 204, Efron, Hastie, Johnstone, Tibshirani, & others, 2004). As a bonus, it was shown that a simple modification to the LARS algorithm could be used to estimate lasso regression, with the benefit of being much faster than alternatives (Efron et al., 2004). More recent advances are coordinate-descent algorithms that are under the hood of the

popular R packages **glmnet** (Friedman, Hastie, & Tibshirani, 2010) and **glasso** (Friedman et al., 2008).

On the other hand, a primary limitation of the nonconvex penalties is computational in nature. In my view, they are too readily dismissed for this reason “[nonconvex objective functions] are interesting but seem difficult, as even the computation of the global solution is infeasible in general” (see the rejoinder in Lockhart, Taylor, Tibshirani, & Tibshirani, 2014). This was also noted in Bühlmann and Meier (2008):

The SCAD-penalty function has been often criticized from a computational point of view as it corresponds to a nonconvex objective function which is difficult to minimize... (p. 1534).

This applies more generally to all nonconvex penalties. The underlying issue is that convergence of a convex function ensures a unique solution (i.e., the global minimum), whereas there is no such guarantee for nonconvex functions. This problem is exacerbated in high-dimensional data, although it should be noted that these situations are uncommon in the psychological literature (Table 2 in Wysocki & Rhemtulla, 2019). A pragmatic argument to this “issue” was put forth

Algorithm 1: The local linear approximation (LLA) algorithm

1. Initialize $\hat{\Theta} = \hat{\Theta}^{(\text{initial})}$ and compute the adaptive weight (i.e., penalty function derivative)

$$\hat{\mathbf{W}}^{(0)} = (\hat{w}_{ij}^{(0)}) = p'_{\lambda_{ij}, \gamma}(|\hat{\theta}_{ij}^{(0)}|)$$

2. For $k = 1, 2, \dots$ repeat until convergence

- (2.a) Obtain $\hat{\Theta}^{(k)}$ by maximizing the following penalized likelihood (Fan et al., 2009)

$$\hat{\Theta}^{(k)} = \arg \max_{\Theta} \log \det \Theta - \text{tr}(\Sigma \Theta) - \sum_{i=1}^p \sum_{j=1}^p \hat{w}_{ij}^{(k-1)} \cdot |\theta_{i,j}|$$

- (2.b) Update the weight matrix $\hat{\mathbf{W}}^{(k)}$

$$\hat{\theta}_{ij}^{(k)} = p'_{\lambda_{ij}, \gamma}(|\hat{\theta}_{ij}^{(k)}|)$$

Note. The initial value is the maximum likelihood estimate.

in Fan, Xue, and Zou (2014), that is, “It is perfectly fine that the computed local solution is not the global minimizer, as long as it has the desired properties.” Moreover, this was also proved for precision matrices (section 3.3 in Fan et al., 2014). Furthermore, as described below, it turns out that there are accurate approximations to various nonconvex penalties that retain their oracle properties.

Local Linear Approximation. Zou and Li (2008) is a seminal work for computing the nonconvex penalized likelihood. The first advancement was an iterative algorithm based on local linear approximation (LLA). This overcame two major drawbacks of previous approaches, such as the local quadratic approximation (LQA) described for computing the SCAD penalty (section 3.3 in Fan & Li, 2001). First, the LQA requires the deletion (based on some threshold) of small relations because ridge regression is used to solve the problem, thereby providing a fully connected graph (i.e., no variable selection) (p. 1517 Zou & Li, 2008). Note that this is distinct from ℓ_1 -regularization, where small values are often deleted after the model is selected by default (e.g., Haslbeck & Waldorp, 2015), due to many false positive selections.⁵ On the other hand, the LLA automatically provides a sparse solution. The second advantage of LLA is computational, that is, Fan et al. (2009) demonstrated that employing nonconvex penalties in GGMs can be recast in terms of iteratively weighted ℓ_1 -regularization. The importance of this cannot be understated: at each step of the algorithm, coordinate-descent algorithms (e.g., those implemented in **glasso**) can accommodate nonconvex penalties.

The LLA is provided in Algorithm 1. The modification to the ℓ_1 penalized likelihood is slight, in that the derivative of a given penalty function is used as the regularization parameter (i.e., λ_{ij} in Equation 3). At each step, the derivative is computed for each element of the glasso estimator, $p'_{\lambda, \gamma}(|\hat{\theta}|^{(k-1)})$, which then serves as $\lambda_{ij}^{(k)}$ for the k th step of the algorithm.

This is repeated until convergence.

One-Step Estimator. The second major contribution of Zou and Li (2008) was describing one-step estimators. A drawback of a fully iterative LLA (Algorithm 1) is that reaching convergence can require a large number of steps (iterations), each of which requires solving Equation 3 and computing the derivative of the penalty function. The one-step estimator, as the name implies, requires computing each only once. This is accurate “provided that the initial estimates are good enough” (p. 1511, Zou & Li, 2008). This accuracy of this approach is due to computing the derivative of $\hat{\theta}$, which is estimated from the data and it determines the amount of regularization for each relation. In high-dimensional settings ($n < p$), the question of a “good enough” initial value is challenging, given the maximum likelihood estimator (MLE) is not available.⁶ In this case, it is common to use the customary ridge estimator (p. 1423, Zou, 2006) or marginal ordinary least squares estimates (M. Wang, Wang, & Wang, 2014). In low-dimensional settings, however, using the MLE is recommended for the initial estimate of Θ . Hence, in the most common psychological applications, it is straightforward to compute $p'_{\lambda, \gamma}(\hat{\theta})$. A one-step algorithm is provided in Algorithm 2.

Software

The vast majority of software that implements nonconvex penalties is for regression modeling, including the MCP

⁵The false positives inherent to the ℓ_1 -penalty are typically small in magnitude. This is fully described in Meinshausen and Yu (p. 267, 2009). It is nonetheless important to note, by itself (no post-processing), lasso has limitations for model selection.

⁶The MLE also becomes unstable when $p \rightarrow n$, but, typically, n (observations) is sufficiently larger than p (nodes) in psychological applications.

Algorithm 2: One-step estimator algorithm

1. Initialize $\hat{\Theta}^{(\text{initial})}$ and compute the adaptive weight (i.e., penalty function derivative)

$$\hat{\mathbf{W}} = (\hat{w}_{ij}) = p'_{\lambda_{ij}, \gamma}(|\hat{\theta}_{ij}|)$$

2. Obtain $\hat{\Theta}$ by maximizing the following penalized likelihood (Fan et al., 2009)

$$\hat{\Theta} = \log \det \Theta - \text{tr}(\Sigma \Theta) - \sum_{i=1}^p \sum_{j=1}^p \hat{w}_{ij} \cdot |\theta_{i,j}|$$

3. Use $\hat{\Theta}$ as the final estimate

Note. The initial value is the maximum likelihood estimate.

and SCAD in the popular R package **ncvreg** (Breheny & Huang, 2011). Furthermore, the recent R package **ncpen** implements a total seven nonconvex penalties for regression models (D. Kim et al., 2018). To my knowledge, there are no R packages that specifically feature nonconvex penalties for GGMs. A honorable mention is **pgraph** that includes the SCAD (using LLA), but note that the package was built for a newly developed projection-based method (Fan, Feng, & Xia, 2020).

To address this dearth of software, I developed the R package **GGMncv** that includes all nonconvex penalties described in this work. There are additional penalties as well (see the Appendix). Furthermore, both the LLA (Algorithm 1) and the one-step estimator (Algorithm 2) are implemented, regularization paths can be visualized (e.g., Figure 2), there are a variety of approaches for selecting the tuning parameter, and a recent method for making statistical inference in regularized GGMs is available (Section Statistical Inference). Together, **GGMncv** is a comprehensive toolbox that was specifically developed for psychological science.⁷

Tuning Parameter Selection

Because these conconvex penalties have two tuning parameters, their selection is especially important to consider. First, the regularization parameter, λ , determines the amount of penalization, with $\lambda = 0$ providing the MLE of Θ and $\lambda \rightarrow \infty$ an empty graph. Second, γ governs the shape of the penalty function, which then has a direct bearing on the rate of penalization (Figure 2). Whereas there is a large literature on choosing λ ,⁸ including with customary information criterion for model selection (e.g., Foygel & Drton, 2010), cross-validation (Feng & Yu, 2013), and the stability approach to regularization selection Liu, Roeder, and Wasserman (2010), there is less information about selecting γ (besides that it is commonly fixed). From a computational perspective, it would be advantageous to reduce the number of models fitted, which can be achieved by fixing either λ or γ . In what

follows, recent approaches that avoid data-driven selection are described, in addition to the customary approach of using an information criterion.

Information Criteria

It is commonplace to select a model according to minimizing (or maximizing) some information criterion. In the context of regularization, this is different than using non-regularized estimation. This difference arises from the latter directly applying a penalty to the MLE, whereas, for the former, the penalized likelihood is first solved with, say, a constraint on the sum of edge weights (ℓ_1 -regularization). Hence, if the estimator itself is not consistent (e.g., if the irrepresentable condition is not satisfied), then the theoretical properties of a given information criterion will not necessarily apply (p. 1042, Y. Kim, Kwon, & Choi, 2012). This was noted in Bühlmann and Van De Geer (2011) “The BIC criterion is described which, however, has no theoretical justification for variable selection with the lasso” (p. 17). Accordingly, any approach based on *only* the solution path of the ℓ_1 -penalty will need to satisfy the necessary assumptions (Ravikumar et al., 2011; Zhao & Yu, 2006). Nonetheless, it is common practice to use an information criterion with ℓ_1 -regularization in psychology, which I also adopt in the numerical experiments.

On the other hand, there is a theoretical justification for using, say, BIC, in combination with nonconvex penalties. That is, when λ is selected with BIC, the desirable property of model selection consistency is not compromised. This is due to the consistency of the solution path being proven for the MCP (C. H. Zhang, 2010) and SCAD Y. Kim and Kwon (2012). Gao, Pu, Wu, and Xu (2012) extended those results to GGMs, that demonstrated the consistency of using

⁷Of course, **GGMncv** can also be used for high-dimensional data.

⁸For an overview of these approaches, interested readers are referred to Wysocki and Rhemtulla (2019).

BIC with the SCAD penalty. Similar results are also provided for the SELO (Dicker et al., 2013) and ATAN penalties (Y. Wang & Zhu, 2016). Notably, these are all for low-dimensional data, that is, when n is sufficiently larger than p . Of course, this also requires many assumptions that are beyond the scope of this work. Their generality is investigated with numerical experiments.

Generalized Information Criteria. All information criteria can be understood as penalizing the likelihood, with each differing in the severity of the penalty. Following Y. Kim et al. (2012), it is thus informative to consider a generic formulation for negative 2 times the log-likelihood, that is,

$$-2l_n(\hat{\Theta}) = -2\left[\frac{n}{2}\log\det\hat{\Theta} - \text{tr}(\hat{\mathbf{S}}\hat{\Theta})\right] \quad (10)$$

where $\hat{\Theta}$ is the estimated precision matrix (e.g., for a given λ and γ) and $\hat{\mathbf{S}}$ is the sample-based covariance matrix. With Equation 10 in hand, it is straightforward to define a variety of GICs, of which the well-known Akaike Information criterion and Bayesian Information criterion are special cases.

Akaike Information Criterion. There are at least two justifications for AIC model selection: (1) expected out-of-sample prediction, since AIC and leave-one-out cross validation (LOO-CV) are asymptotically equivalent (Y. Zhang & Yang, 2015); and (2) minimizing Kullback-Leibler (KL) divergence from the target and approximating data generating model (Burnham & Anderson, 2004). AIC is computed as

$$\text{AIC} = -2l_n(\hat{\Theta}) + 2 \cdot |\mathbf{E}|, \quad (11)$$

where $|\mathbf{E}|$ refers to the cardinality of the edge set, that is, the number of edges. The properties of AIC are well-known, including that it is not consistent for model selection. This also applies to model selection with LOO-CV (Gronau & Wagenmakers, 2018). This results in a relatively high false positive rate that does not diminish with more data. In fact, Dziak, Coffman, Lanza, Li, and Jeremiin (2019) pointed out that AIC has a direct correspondence to significance testing with a liberal error rate. However, it would be a mistake to think model selection consistency is the only goal worth pursuing, for example, in regression, “for any model selection criterion to be consistent, it must behave suboptimally for estimating the regression function” (p. 937, Yang, 2005). The latter matters most for prediction (or estimating the regression function), whereas consistent model selection limits spurious associations. Hence, when prediction is of interest, while also being tolerant of spurious associations, AIC can be used as a computationally efficient approximation to LOO-CV, given they are asymptotically equivalent (Stone, 1977).

Bayesian Information Criterion. The Bayesian information criterion also has two primary justifications: (1) expected out-of-sample prediction, since minimizing BIC is

equivalent to leave- v -out cross-validation, where $v = n[1 - 1/(\log(n) - 1)]$ (Shao, 1997); and (2) minimizing BIC approximates selecting the most probable model, assuming the *true* model is in the candidate set (Raferty, 1995). When the *true* model is not in the set, this is referred to as the *m*-open setting (Bernardo & Smith, 2001), wherein selecting based on BIC minimizes the KL-divergence from the *true* model (Yao, Vehtari, Simpson, & Gelman, 2017). In GGMs, BIC is computed as

$$\text{BIC} = -2l_n(\hat{\Theta}) + \log(n) \cdot |\mathbf{E}|, \quad (12)$$

where $|\mathbf{E}|$ is the number of edges. Note that, compared to AIC, the likelihood is penalized more harshly, that is, $\log(n)$ instead of 2, which produces a model including fewer edges. It might be tempting to think, given the known model selection consistency of BIC, that this would also apply to using it in concert with the ℓ_1 -penalty. This is not the case, necessarily, as lasso still requires strong assumptions over and above, say, maximum likelihood, to converge on the *true* model. This does not apply to non-regularized estimation with the ℓ_0 -penalty or nonconvex penalties, assuming that p does not increase with n , that is, the standard asymptotic setting ($p/n \rightarrow 0$). An open question is whether approximations to the ℓ_0 maintain this property in situations known to be challenging for lasso (Section Numerical Experiments).

Alternative Criteria. There are numerous additional criteria. For example, BIC is typically studied in fixed and moderate p settings. However, in high-dimensional data, or when p grows with n , BIC will start to over select. In these situations, it is common to use the *extended* BIC that adds $4 \cdot |\mathbf{E}| \cdot \gamma \cdot \log(p)$ to Equation 12 (Chen & Chen, 2008; Foygel & Drton, 2010). There are many other options, including the Risk Inflation Criterion ($|\mathbf{E}| \cdot 2 \cdot \log(p)$, Foster & George, 1994), the modified RIC ($2 \cdot (\log(p) + \log(\log(p)))$, Foster & George, 1994), and the GIC₃ of Y. Kim et al. (2012) which is consistent when the error distribution is heavy tailed to name but a few.

I refer interested readers to Y. Kim et al. (2012) for a thorough discussion of the theory underlying GICs (section 5 includes 6 GICs in total). A key takeaway is that “the range of consistent model selection criteria is rather large, and it is not clear which one is better with finite samples” (p. 1048 Y. Kim et al., 2012). It was noted, however, that the GICs can be understood in reference to AIC and BIC, in that, for a sparse model, a GIC can be selected with a larger penalty added to Equation 10.

Tuning Insensitive Estimation

Given the sheer number of possibilities for selecting the tuning parameters, this raises the question of whether it is possible to side-step this issue altogether. To this end, there

has been considerable work investigating optimal regularization without the need for, say, selecting λ with an information criterion. For lasso, the foremost challenge is that λ is directly related to the residual variance, σ^2 (or the “noise” more generally), which is unknown in reality. In [Belloni, Chernozhukov, and Wang \(2011\)](#), however, it was shown that $\lambda = \sqrt{\log(p)/n}$ is the theoretically consistent regularization parameter for the square root lasso which decouples the dependence of λ on σ^2 . Similar results were provided for the scaled lasso in [\(Sun & Zhang, 2012, 2013\)](#), where a “universal” regularization parameter was introduced. Although these findings were initially for regression, they have since been extended to graphical models. For example, the tuning insensitive graph estimation and regression (TIGER) method [\(X. Li, Zhao, Yuan, & Liu, 2015; Liu & Wang, 2017\)](#), it was suggested for the glasso estimator (p. 1218, [Janková & van de Geer, 2015](#)), and it serves as the default in the popular R package **SILGGM** [\(R. Zhang, Ren, & Chen, 2018\)](#). In simulation studies, a routine finding is that data-dependent tuning parameter selection does not always seem necessary, which allows for avoiding the computational intensive nature of selecting λ .

For the nonconvex penalties discussed in this work, there is an additional tuning parameter, γ , that could also be selected. This opens the door to searching for some combination of λ and γ . To my knowledge, there are no formal proofs for an optimal value of γ . Here to, however, selecting γ , as opposed to setting it to some value, is not always necessary [\(Dicker et al., 2013; Fan & Li, 2001\)](#). Of course, given that the properties of nonconvex penalties are less studied than lasso, it could very well be that both λ and γ should be selected in data common to psychology.

Numerical Experiments

A routine finding is that nonconvex penalties regularly outperform ℓ_1 -regularization, for both the goal of model selection and prediction. Furthermore, it has also been shown that the approximate ℓ_0 penalties can outperform the more common nonconvex penalties, including SCAD and MCP. From my perspective, this is not too surprising because it is hard to imagine a newly proposed method being published in the statistics literature if it did not offer some benefit. However, it is commonplace in statistics papers to only include small-scale simulations in which the conditions mirror the methods’ assumptions, which are over and above customary assumptions of, say, best subset selection with BIC. As an example, although lasso can be consistent for model selection, this requires assumptions that do not seem to hold in psychological applications. In [Williams et al. \(2019\)](#), [Williams and Rast \(2019\)](#), and [Wysocki and Rhemtulla \(2019\)](#), each of which used data representative of the “network” literature, glasso was shown to have an inflated false positive rate (e.g., $> 25\%$). Moreover, in [Williams et al. \(2019\)](#) and

[Williams and Rast \(2019\)](#), there was no discernible benefit compared to non-regularized estimation. Accordingly, before unleashing nonconvex penalization upon psychological science, there are several loose ends that must be addressed.

Open Questions

1. To date, most of the work on nonconvex penalties has been in regression models. In [Fan et al. \(2009\)](#), the one-step estimator for SCAD was used to estimate GGMs, where performance for model selection and classification accuracy was superior than ℓ_1 -regularization. However, the performance of the MCP, SELO, and ATAN penalties has not been investigated in GGMs.
2. In [Williams et al. \(2019\)](#), it was demonstrated that ℓ_1 -regularization fails (i.e., a high false positive rate) when the *true* network is densely connected, which is exacerbated by larger edge sizes (see Figure 6 in [Williams et al., 2019](#)). Therefore, given psychological networks are far from sparse, it is important to establish that these nonconvex penalties overcome those issues.
3. To my knowledge, tuning insensitivity for λ has only been used for nonconvex regularization in [Shen, Chen, Gu, and So \(2016\)](#). It would be advantageous to avoid data-driven model selection, either for λ or γ , which is now commonplace in the regularization literature. Hence, the adequacy of setting $\lambda_{ij} = \sqrt{\log(p)/n}$ must be thoroughly investigated with nonconvex regularization.
4. Finally, it is important to demonstrate that nonconvex regularization offers benefits compared to non-regularized estimation. In this way, the literature does not become inundated with superfluous methodology.

These questions are answered via simulation, that, in addition to the nonconvex penalties, includes the customary glasso estimator and a non-regularized approach. The models are selected with BIC (Equation 12). Employing the ℓ_0 penalty, which requires exploring all possible models, is too computationally demanding for even a modest number of variables. Accordingly, the included non-regularized approach is forward selection (Non-reg FS). Note that forward selection is considered a “locally optimal” version of best subsets (p. 3, [Hastie, Tibshirani, & Tibshirani, 2017](#)). The R package **mixGGM** was used to estimate the non-regularized model [\(Fop, Murphy, & Scrucca, 2019\)](#) and **GGMncv** for the regularized estimators [\(Williams, 2020a\)](#). Note that **GGMncv** implements Algorithms 1 and 2 with **glassoFast** [\(Sustik, Calderhead, & Clavel, 2018\)](#).

Computational Time

An important advantage of the introduced methods is computational, even for a moderate number of nodes, where non-regularized methods can quickly become cumbersome. To demonstrate this point, I computed the computational time averaged across 10 simulation trials. Because the non-convex penalties should be similar in timing, only the ATAN penalty was included in this simulation. The time was computed for the LLA (Algorithm 1) and one-step estimator (Algorithm 2), for both fixed λ and selecting λ with the Bayesian information criterion (γ was fixed to 0.01). The *true* network structures were generated with the R package **BDgraph** using the default settings (i.e., a random network and 30% connectivity, Mohammadi & Wit, 2015b). The sample size was set to $n = 500$ and $p = \{5, 10, \dots, 30\}$.

Results. Table 1 includes the results. The computational advantage of the regularized models is apparent. For example, with $p = 30$, the regularized methods were all under half of a second, whereas the non-regularized method took nearly 40 seconds. It is important to note that the non-regularized method (Non-reg FS) is not best subset selection (only a “locally optimal” version), which would take potentially hours for $p = 30$. In this light, the computational advantages of penalties that approximate the ℓ_0 are clear, in that the timing is similar to glasso.

Besides estimating the dependence structure, this difference in time is also important because it eases the burden of resampling. It is common to compute inclusion “probabilities” with the non-parametric bootstrap. The basic idea is to compute the proportion of bootstrap samples for which each edge was selected (Figure 6.4 in Hastie et al., 2015). Drawing 1,000 bootstrap samples with Non-reg FS would take several hours (with $p = 30$). Conversely, the regularized methods would take a matter of seconds.

Synthetic Partial Correlations

Recall Williams et al. (2019) demonstrated that the false positive rate (FPR) of glasso is a function of the number of nodes, sparsity, and edges size: the FPR increases with more connections, which is further exacerbated by larger edge sizes and more nodes (Figure 6 in Williams et al., 2019). Accordingly, in this simulation, I follow the exact simulation procedure in Williams et al. (2019) with the goal of determining whether these nonconvex penalties overcome the known issues of ℓ_1 -regularization. I investigated two sample sizes ($n = 500$ and 5000), two values for $p = \{10 \text{ and } 20\}$, and 5 sparsity levels $\eta = \{90\%, 80\%, \dots, 50\%\}$. The magnitudes of the partial correlations were varied by adjusting the degrees of freedom of the G -Wishart distribution (Mohammadi & Wit, 2015b), corresponding to 90 % of the edge falling within ± 0.40 and ± 0.25 , for a given sparsity level. These distributions were approximately normal, such that smaller values were sampled more often than larger values. The network structure was again random and all generating matrices were positive definite. For each trial, the simulation procedure followed these steps:

1. Generate data from a $p \times p$ dimensional G -Wishart distributed precision matrix $\Theta \sim W_G(df, I_p)$, where I_p is an identity matrix (Mohammadi & Wit, 2015a).
2. Estimate the conditional dependence structure with glasso, SCAD, MCP, SELO, ATAN, and Non-reg FS.
3. Compute the FPR (1 - specificity).

The results were averaged across 50 simulation trials.

Results. Figure 3 includes the results. For glasso, first notice the inflated FPR, as the graph becomes densely connected, which is also affected by edge size. For example, with $p = 20$ and $n = 500$, the FPR was less than 0.10 with 90% sparsity but nearly 0.60 with 50% sparsity for edges ± 0.40 . This reproduces the results in Williams et al. (2019).

Table 1
Computing time in seconds.

Method	Number of nodes (p)					
	5	10	15	20	25	30
ATAN ^{λ fixed} _{one-step}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)
ATAN ^{λ fixed} _{LLA}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)
ATAN ^{λ selected} _{one-step}	0.04 (0.01)	0.04 (0.01)	0.05 (0.01)	0.07 (0.02)	0.08 (0.02)	0.42 (0.70)
ATAN ^{λ selected} _{LLA}	0.05 (0.01)	0.06 (0.01)	0.07 (0.02)	0.10 (0.01)	0.14 (0.03)	0.46 (0.49)
glasso ^{λ fixed}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)
glasso ^{λ selected}	0.05 (0.03)	0.04 (0.01)	0.05 (0.01)	0.08 (0.02)	0.10 (0.02)	0.34 (0.46)
Non-reg FS	6.91 (0.20)	7.21 (0.32)	7.96 (0.55)	11.80 (1.87)	22.10 (4.62)	39.60 (10.4)

Note. The results are rounded to the second decimal place.

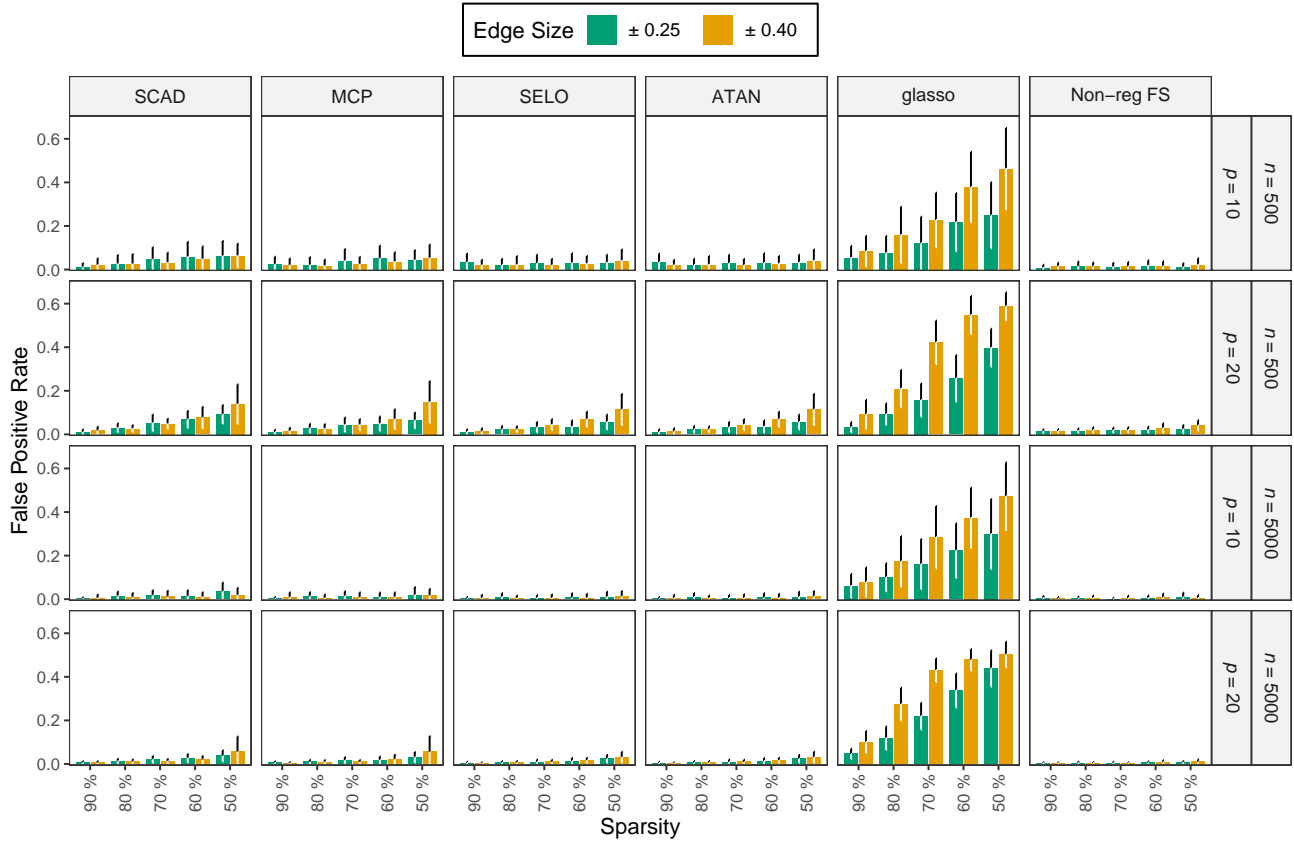


Figure 3. Simulation results (Section [Synthetic Partial Correlations](#)). The idea here was to verify that the nonconvex penalties overcome the issues inherent to glasso that were demonstrated in [Williams et al. \(2019\)](#). Namely, that the false positive rate increases with connectivity (i.e., less sparse), the distribution of edge sizes, and the number nodes (p). Although not completely immune, the nonconvex penalties are clearly less sensitive to those factors, with the added bonus of the false positive rate diminishing to negligible levels with more data. The \pm denotes that 90 % of the partial correlations were in that range. Sparsity is the probability an edge was zero for a given network. Sparsity decreases when moving from the left to right. The error bars denote one standard deviation.

It is apparent that each nonconvex penalty is less susceptible to those factors, although there was a slight increase in the FPR. At worst, for example, was the MCP that went from a FPR of nearly zero (90% sparsity) to roughly 0.10 (50% sparsity). The continuous approximations (SELO and ATAN) to the ℓ_0 -penalty fared even better, that is, they were more resistant to the assumed sparsity. A noticeable increase, although still relatively low compared to glasso, was observed for the most densely connected graphs with $n = 500$. Importantly, the FPR diminished in the larger sample size ($n = 5,000$) for all nonconvex penalties (less than 0.05 for all sparsity levels). This indicates that contrary to glasso, in which, with even $n = 5,000$, the FPR is still very high ($> .40$) and thus far from converging upon the *true* model (even though BIC was used), these nonconvex penalties will apparently converge upon the ground *truth* across a range of situations (which is expected when using BIC).

When compared to Non-reg FS, the SELO and ATAN penalties provided competitive performance, whereas the SCAD and MCP had a higher FPR. Notably, the comparable FPR of the SELO and ATAN penalties was achieved in a fraction of the time (Table 1). Hence, when computational feasibility is important, for example, when computing inclusion “probabilities” and/or there is a large number of nodes, the SELO and ATAN are the most promising when the goal is to have the lowest FPR (or conversely to have the highest specificity).

Empirical Partial Correlations

In my experience, generating conditional dependence structures that reflect those observed in psychology is a challenging task. This is due to requiring a positive definite matrix, while also having a dense network with a plausible distribution of edge sizes. To this end, following [Ep-](#)

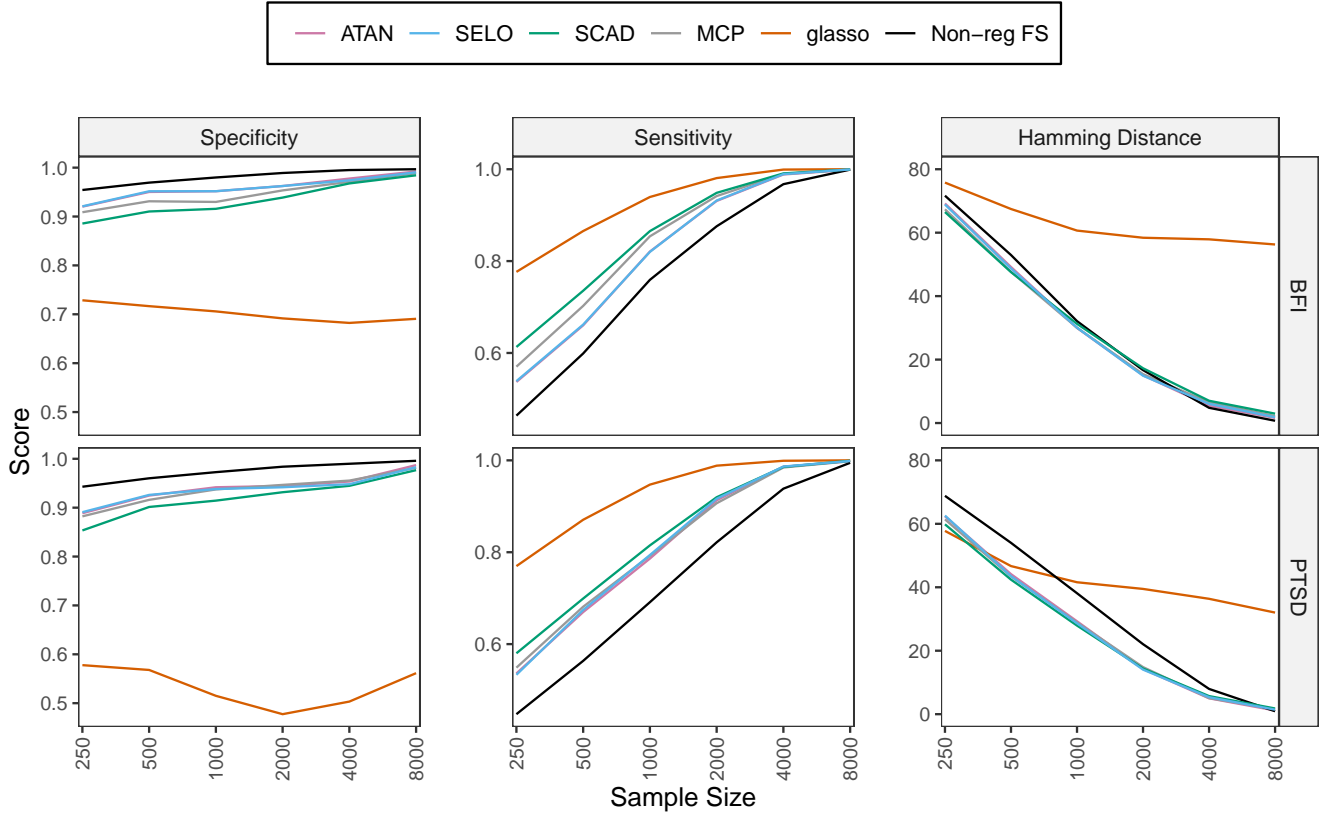


Figure 4. Simulation results (Section Empirical Partial Correlations). Note that specificity is 1 - the false positive rate and sensitivity 1 - the false negative rate (i.e., analogous to “power”). Hamming distance is the number of position in which the *true* and estimated adjacency matrices differ, with a score of zero denoting that the *true* structure was recovered. There are two key takeaways: (1) all of the nonconvex penalties converged upon the generating model, similar to the non-regularized method, but glasso did not; (2) the nonconvex penalties appear to strike a balance between glasso (very low specificity and the highest sensitivity) and the non-regularized estimator (the highest specificity and the lowest sensitivity). This is reflected in the Hamming distance scores, in that the nonconvex penalties were at worse comparable with Non-reg FS.

skamp (2016), I have found it useful to examine performance for partial correlation matrices estimated from data. Thus, in this section, network structures are derived from 20 PTSD symptoms (Armour et al., 2017) and 25 items measuring personality (BFI Revelle, 2019). The *true* network structure was obtained by first estimating the partial correlation matrix and then setting absolute values less than 0.05 to zero. Note that this produced reasonably dense graphs (PTSD: 38% sparsity; BFI: 61% sparsity) which provides a litmus test of sorts for these nonconvex penalties (i.e., Figure 3) For each of the 50 simulation trials, specificity (SPC), sensitivity (SN, analogous to “power”), and Hamming distance (HD, the number of positions in the true and estimated adjacency matrices that differ) were computed for $n = \{250, 500, 1,000, 2,000, 4,000, 8,000\}$. For the nonconvex penalties, γ was set to the recommended default (SCAD: 3.7; MCP: 2; SELO: 0.01; ATAN: 0.01).

Results. Figure 4 includes these results. As a reference point, first note that glasso had the lowest specificity, that is, it had the highest FPR that *always* exceeded 0.25. As expected, given the results in Figure 3, specificity was lowest for the PTSD “network” (due to less sparsity). Of note, sensitivity was the highest for glasso. In other words, rather than limiting spurious associations, employing the ℓ_1 -penalty emphasizes discovery. This was also noted in regularized structural equation models, “lasso kept more variables in the model (more Type I and fewer Type II errors)” (p. 72, Jacobucci, Brandmaier, & Kievit, 2019). However, the HD scores indicated that glasso was, by far, the furthest away from the true adjacency matrix (it did not approach the *true* model).

All four nonconvex penalties, on the other hand, did not have any of those issues. For example, with more data, specificity for each approached a perfect score. This is expected, given BIC was used to select the model, which parallels the

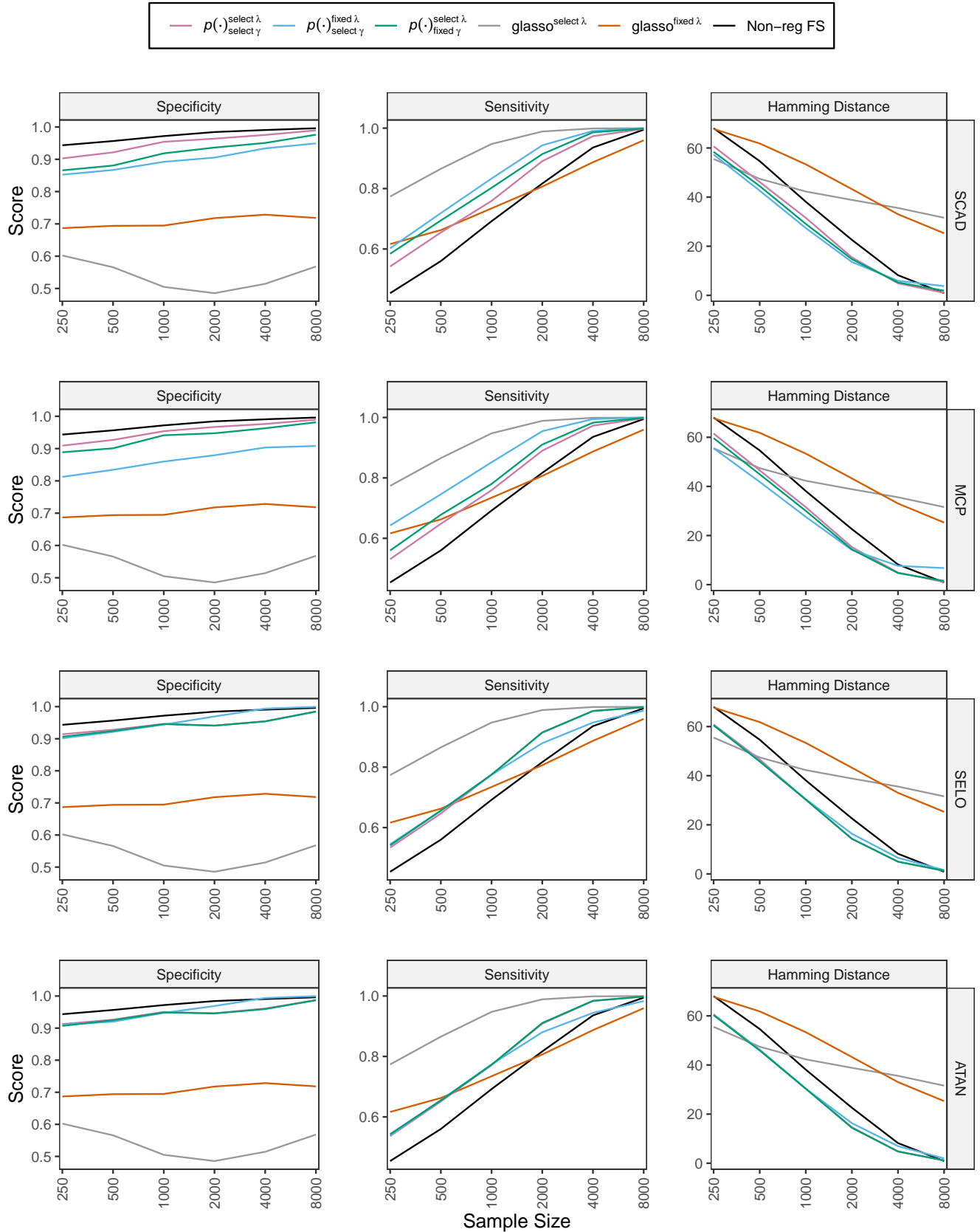


Figure 5. Simulation results (Section [Tuning Parameter Selection](#)). $p(\cdot)$ is a generic penalty function, with, say, $p(\cdot)^{\text{select } \lambda, \text{select } \gamma}$, denoting that both λ and γ were selected. The non-regularized approach and glasso are included in each panel. These results indicate that selecting both tuning parameters is unnecessary, with the SELO and ATAN penalties especially immune to which are selected.

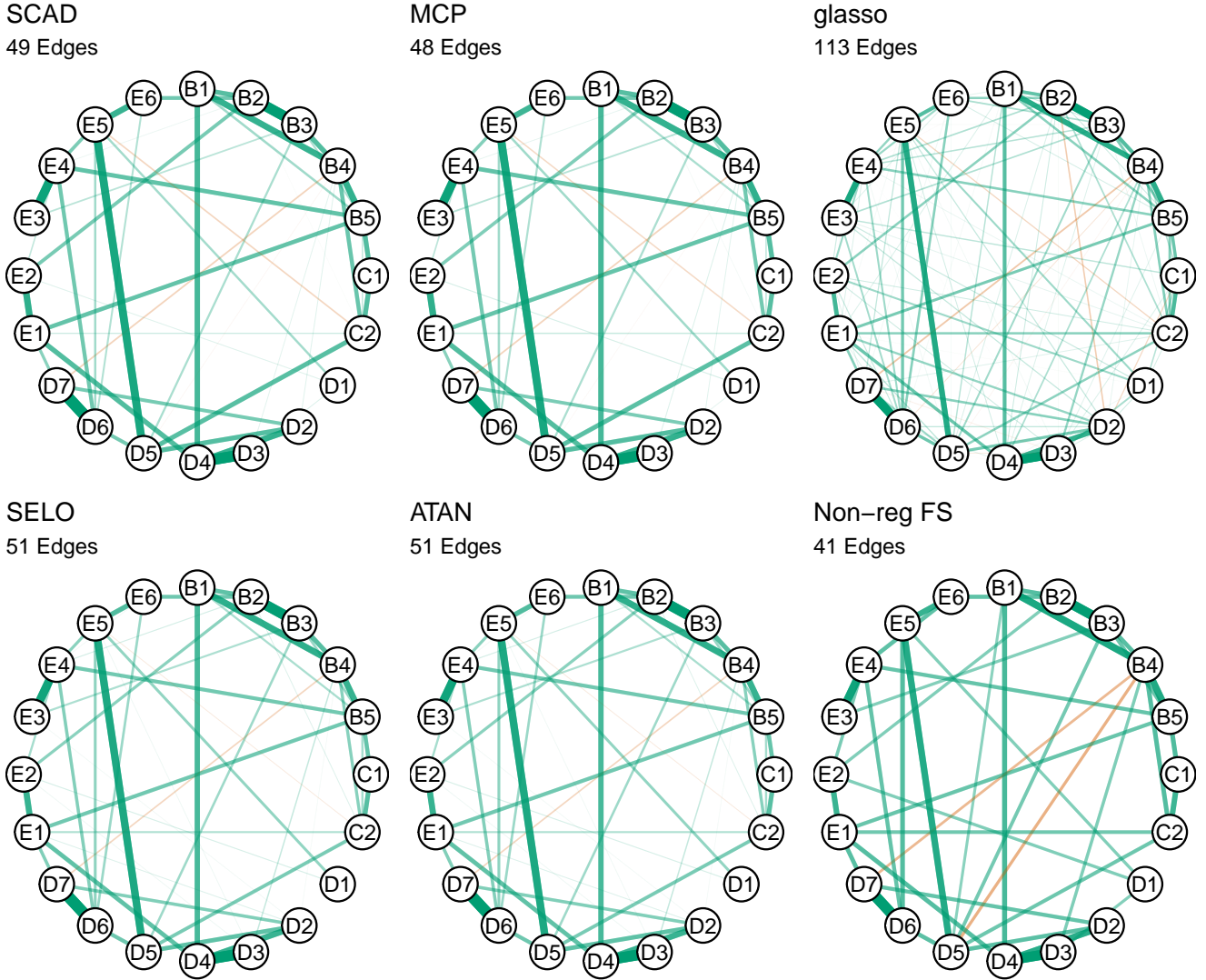


Figure 6. The estimated conditional dependence structure of 20 PTSD symptoms (see Table B1).

proof provided in [Gao et al. \(2012\)](#). If the goal is to reduce spurious associations, then the ATAN and SELO penalties have clear benefits, as the FPR was the lowest in each dataset. Interestingly, ATAN and SELO provided almost identical scores. Conversely, SCAD and MCP had higher sensitivity than the ATAN and SELO penalties. The HD scores, however, indicated that each nonconvex penalty was quite similar to one another, with each reducing to zero (indicating convergence upon the generating structure).

Recall from Section [Open Questions](#), I emphasized that it was important to demonstrate that these methods have a benefit compared to customary statistical approaches in psychological science. In addition to computational (Table 1), these results revealed competitive performance with Non-reg FS. For example, although they had lower but still high specificity, sensitivity was also higher. This translated into,

at worse, comparable scores for HD. Note that specificity could be improved further by employing an alternative GIC. Together, these nonconvex penalties provide a computationally efficient solution without sacrificing the performance of non-regularized model selection.

Tuning Parameter Selection

Thus far, all simulations have selected λ with BIC and fixed γ to the recommended values. It could be that selecting either one or possibly both can improve model selection. I investigated this possibility by selecting both λ and γ (denoted $p(\cdot)_{\text{select}}^{\text{select } \lambda, \text{select } \gamma}$), fixing λ (i.e., $\sqrt{\log(p)/n}$) and selecting γ (denoted $p(\cdot)_{\text{fixed } \lambda}^{\text{select } \gamma}$), as well as selecting λ and fixing γ (denoted $p(\cdot)_{\text{fixed } \gamma}^{\text{select } \lambda}$). As a reference point, glasso was also included with λ fixed and selected, in addition to the non-regularized

method (Non-reg FS). The PTSD partial correlations were again used as the *true* network structure, which ensures performance is evaluated in a dense setting. The results were again averaged over 50 simulation trials.

Results. Figure 5 includes these results. The underwhelming performance of glasso was again striking, given each way of selecting the model with nonconvex penalties provided superior performance. Interestingly, selecting λ for glasso resulted in lower specificity than fixing it to $\sqrt{\log(p)/n}$, although both, in my opinion, leave much to be desired. This indicates that the theoretically optimal regularization parameter does not sidestep the underlying assumptions of the ℓ_1 -penalty.

For SCAD and MCP, it appears that the choice of which tuning parameter to select is a compromise between specificity and sensitivity. For example, $p(\cdot)_{\text{select}}^{\text{select } \lambda}$ had the highest specificity but the lowest sensitivity. However, note the HD scores were very similar. Of course, this does not suggest that both γ and λ need to be selected to ensure the FPR is low, as an alternative GIC can be employed that applies a harsher penalty to the likelihood. The results of SELO and ATAN were very interesting, in that they were essentially immune to which tuning parameter was selected. Together, this indicates that not much is gained (or lost) by searching for some combination of γ and λ , and for the SELO and ATAN penalties in particular, it does not seem to matter which is selected.

Application

In this section, the nonconvex penalties are used to estimate the conditional dependence structure of 20 PTSD symptoms ($n = 221$, [Armour et al., 2017](#)). All models were selected with BIC. For the nonconvex penalties, γ was selected and λ was set to $\sqrt{\log(p)/n}$.

Results

The estimated conditional dependence structures are provided in Figure 5. The “networks” parallel the findings from the experiments, namely that glasso included the most edges, Non-reg FS had the fewest, and the nonconvex penalties were similar to Non-reg FS but with a few more edges (but less than half the number of glasso). The nonconvex penalties were remarkably consistent with one another (differ at most by 3 edges), with SELO and ATAN providing nearly identical edge weights.

A Brief Note on Interpretation. I emphasize that these graphs are merely the *estimated* dependence structure. In reference to the simulation results, sensitivity is low for all methods with a small sample size (besides glasso). This translates into missing quite a bit. Furthermore, just because, say, the ATAN penalty converges on the *true* model in simulation, this does not suggest it has nicely identified the most

important edges. To rule out meaningful relations, that is evidence for conditional independence, this requires a confidence interval⁹. In other words, no connection in the graph does not suggest evidence for the null, and, further, at this point, we cannot support the claim that the “thickest” lines (i.e., the largest edges) are greater than any of the relations pushed to zero. This glimpse into the dependence structure, although limited, nonetheless provides what are likely to be nonzero relations. This information can be harnessed to improve future inquiries.

Statistical Inference

I would be remiss to not discuss the important topic of statistical inference. Because nonconvex penalties have desirable properties, it might be tempting to think that these approaches lead to rich inference. This would be a mistake. In my experience, researchers *always* want to do more than detect nonzero relations in GGMs. For example, to determine which edges are the strongest or to rule relations out of the “network” (i.e., conditional independence), each of which requires more than merely mining data. In other words, statistical inference still requires a p -value or confidence interval, neither of which is straightforward to obtain after data-driven model selection ([Leeb & Pötscher, 2005](#)) or from regularized estimators ([Javanmard & Montanari, 2015](#)). This is due to difficulties obtaining a valid standard error. One approach might be to employ the bootstrap. However, those intervals will not have the correct coverage probabilities. The issue was stated in [Bühlmann, Kalisch, and Meier \(2014\)](#):

The (limiting) distribution of such a sparse estimator is non-Gaussian with point mass at zero, and this is the reason why standard bootstrap or subsampling techniques do not provide valid confidence regions or p -values. Thus, we have to use other approaches to quantify uncertainty (p. 7).

Moreover, even for those values far from zero, the nonvanishing bias and variance reducing properties of ℓ_1 penalization also present challenges: the bootstrapped sampling distribution will often be too narrow and pulled towards zero, thereby missing the *true* value excessively. Hence, the standard error and resulting “confidence intervals” are not very meaningful (see p. 18 in [Goeman, Meijer, & Chaturvedi, 2018](#)).

Although nonconvex penalties mitigate bias to some degree, their sparsity inducing properties are still omnipresent. Therefore, attempting to construct confidence intervals with a bootstrap strategy should be avoided (see for example [Leeb & Pötscher, 2008](#); [Pötscher & Leeb, 2009](#)). Developing

⁹Bayesian hypothesis testing can also be used to gain evidence for the null hypothesis (see for example [Williams & Mulder, 2020](#))

methods to make statistical inference from regularized estimators, possibly after model selection, is an active area of research in the statistics literature.

Model Selection Bias

Although model selection is a customary practice in psychology, what is less appreciated (or not considered at all) is that “[after model selection] such inference enjoys none of the guarantees that classical statistical theory provides for tests and confidence intervals” (p. 802, [Berk, Brown, Buja, Zhang, & Zhao, 2013](#)). This applies to both regularized and non-regularized estimation. A critical assumption of both a p -values and confidence interval is that the model is fixed (pre-data), that is, it was not determined from the data (post-data). When selecting a model, however, this creates the problem of *conditional* (on the selected model) inference (see Section 2.1 in [Devezer, Navarro, Vandekerckhove, & Buzbas, 2020](#)). To make sense of this, consider that edges will be selected when they are large, which necessarily distorts the sampling distribution of the estimator (i.e., resulting in a truncated normal distribution, p. 5 in [Meir & Drton, 2017](#)). As a result, corrections are needed to make post-selection inference. Such examples include adjusting p -values (e.g. [Berk et al., 2013](#)) and, for lasso, “characterizing the distribution of post-selection estimators conditioned on the selection event” (p.1 in [Lee, Sun, Sun, & Taylor, 2016](#)). For an overview of this topic, I refer interested readers to [Taylor and Tibshirani \(2015\)](#).

Debiased Estimators

Avoiding model selection circumvents the issue of conditional inference. The remaining issue for making statistical inference is overcoming the inherent bias and sparsity of the estimator. To this end, related work focuses specifically on debiasing ℓ_1 -regularized estimates ([Janková & van de Geer, 2015](#); [Javanmard & Montanari, 2013, 2015](#)). Note that this is sometimes referred to as *depsarsifying* ([Van De Geer, Bühlmann, Ritov, & Dezeure, 2014](#)). The basic idea is to remove the bias and sparsity, after which confidence intervals can be constructed. This was first worked out for regression, and, more recently, it was extended to glasso in [Janková and van de Geer \(2015\)](#). The debiased glasso estimator, $\hat{\mathbf{T}}$, is defined as

$$\hat{\mathbf{T}} = 2\hat{\mathbf{\Theta}} - \hat{\mathbf{\Theta}}\hat{\mathbf{R}}\hat{\mathbf{\Theta}} \quad (13)$$

with $\hat{\mathbf{\Theta}}$ the glasso estimator of the precision matrix and $\hat{\mathbf{R}}$ the sample-based correlation matrix. As the name implies, this step removes the zeros and attempts to correct the bias from the ℓ_1 -penalty. The asymptotic variance is then given as

$$\text{Var}[\hat{\mathbf{T}}] = \text{diag}(\hat{\mathbf{T}})\text{diag}(\hat{\mathbf{T}})' + \hat{\mathbf{T}}^2 \quad (14)$$

where $\text{diag}(\hat{\mathbf{T}})$ is a diagonal matrix with $\text{diag}(t_{11}, \dots, t_{pp})$. This readily allows for computing p -values and confidence intervals for each off-diagonal element of the desparsified estimator. The simulation results in [Janková and van de Geer \(2015\)](#) showed that the desparsified confidence intervals provided advantages compared to maximum likelihood, but it should be noted that the smallest number of nodes considered was 80, with merely $n = 250$. This is not representative of the psychological [Janková and van de Geer](#) (e.g., Table 2 in [2015](#)). Accordingly, the advantage compared to simply using significance testing with, say, non-regularized partial correlations, is not entirely clear ([Drton & Perlman, 2004, 2005](#); [Williams & Rast, 2019](#)).

Summary

I hope this section highlighted that using nonconvex regularization does not preclude statistical rigor: simply throwing data at the algorithms in **GGMncv** does not nicely sort the important from the unimportant. To reemphasize, just because a value was set to zero, this does not provide evidence for the null hypothesis and there is not necessarily a difference between an edge that was detected and a relation that was “pushed” to zero ([Gelman & Stern, 2006](#)). These inferences would require a valid measure of uncertainty.

Nonconvex regularization can be used for the explicit goal of data mining, which can provide the first glimpse into a dependence structure. To answer richer questions, this requires a valid measure of uncertainty, which, to date, appears to be more easily obtained with non-regularized estimation (e.g., bootstrapping partial correlations that allows for comparing relations within and between GGMs). Hence, when there is a more refined research question, say, comparing edges or ruling out edges, researchers should forgo model selection and nonconvex regularization, using instead non-regularized methods that readily provide an accurate sampling distribution.

Discussion

This paper surveyed nonconvex penalties for estimating Gaussian graphical models in psychological science. Additional topics included tuning parameter selection and the long-standing challenges of making statistical inference from regularized estimators. All of the penalties were shown to be superior to glasso and more than competitive when compared to non-regularized based model selection.

In addition to a comprehensive review, this work was also the first to thoroughly compare these penalties in a variety of settings. The idea here was to ensure that they overcome the known issues of ℓ_1 -regularization that were demonstrated in [Williams et al. \(2019\)](#) and [Williams and Rast \(2019\)](#), in addition to ensuring they offer benefits compared to non-regularized estimation. Here it emerged that each penalty does in fact overcome the issues inherent to lasso for model

selection. The benefit compared to non-regularized estimation was primarily computational in nature. This was not completely unexpected, given that, in the ubiquitous case of low-dimensional data, we already have excellent methods at the ready. This computational advantage has practical utility when there are many nodes or when resampling is needed to compute edge inclusion “probabilities” (see for example Abram et al., 2016; Bunea et al., 2011)

Although this work was not focused on finding a “winner,” in my opinion the ATAN penalty is the most promising. This is because the additional tuning parameter, γ , strikes a balance between the ℓ_0 and ℓ_1 penalties, which lends itself to more easily understanding the penalty function (Figure 1). And the ATAN was apparently immune to which tuning parameter was selected (Figure 5). Of course, each form of nonconvex regularization is clearly superior than the ℓ_1 -penalty for model selection, which indicates they are all viable options for exploratory data analysis.

Future Directions

This review is not meant to be the final but the first word on nonconvex regularization in psychological science. Accordingly, this work provides a solid foundation from which future inquiry can be built. I offer the following suggestions for further study of nonconvex penalization.

First, I did not look at estimation accuracy such as mean squared error of the edges. This decision was made in part to keep the results manageable. This can be investigated in the future, although, from my experience, I do not think regularization will provide notable benefits, assuming the p/n ratio is sufficiently small (e.g., $p = 20$ and $n = 200$). One possibility is to take advantage of the computational efficiency of the one-step estimator (Algorithm 2), and then, once a model is selected, estimating the MLE given those constraints (pp. 631 - 634 in Hastie, Tibshirani, & Friedman, 2008). This is similar in spirit to Foygel and Drton (2010), that did this with glasso, but in this case the solution path is obtained from nonconvex regularization. Hence, even if nonconvex regularization does not improve accuracy compared to the MLE, the computational advantage is still omnipresent.

Second, I did not consider predictive accuracy or the possibility of overfitting, which is often assumed to be an advantage of regularization (but see Williams & Rodriguez, 2020). The predictive advantage of regularization depends on the signal-to-noise ratio (SNR; $\frac{R^2}{1-R^2}$). In noisy settings, with hardly variance explained, the *relaxed* lasso is advantageous (Hastie et al., 2017), but this dissipates with more signal and when $p/n \rightarrow 0$. On the other hand, nonconvex penalties (including the ℓ_0) are known to be less than ideal in low SNR situations. In the “network” literature, where nodes are items from psychometric scales, the SNR should not be terribly low, Williams and Rodriguez (2020) showed that non-regularized methods were at least competitive with

lasso for out-of-sample prediction. Accordingly, it would be informative to look at a range of SNR values, especially those for which model recovery is perhaps a bridge too far.

Third, ordinal data is commonly encountered in the social-behavioral sciences. To this end, exploring these penalties with polychoric partial correlations is an important future direction. The nonconvex penalties depend on the initial estimates, which is not typically an issue in the most common situations, assuming it is “good enough” (p. 1511, Zou & Li, 2008). However, polychoric partials will typically have more sampling variance (e.g., Williams, 2020b), which suggests that the initial estimate for computing the derivative will also have more variance than Gaussian data. Here shrinkage estimators should be investigated (e.g., Ledoit & Wolf, 2004; Van Wieringen & Peeters, 2016). I anticipate that these findings will generalize, with the potential challenge of overcoming lower specificity for small sample sizes in particular (due to variability in the initial estimate).

Forth, there are additional forms of regularization not considered in this work, such as, for example, the elastic net that can strike a balance between ℓ_1 (lasso) and ℓ_2 (ridge) penalties (Zou & Hastie, 2005). It should be noted that the elastic net idea (e.g., to better accommodate multicollinearity) can be used with the SCAD penalty (Becker, Toedt, Lichter, & Benner, 2011; Zeng & Xie, 2014). Furthermore, the bridge regression includes a broad class of possibilities (Frank & Friedman, 1993), typically defined as ℓ_q ($q > 0$), that includes lasso ($q = 1$) and ridge ($q = 2$) as special cases. Notice that ℓ_q also includes an approximation to the ℓ_0 penalty ($q \rightarrow 0$, p. 2 in Frommlet & Nuel, 2016). The LQA and LLA algorithms for (approximately) solving the ℓ_q penalized likelihood were compared in (C. Park & Yoon, 2011). These options could be explored in the future (the ℓ_q is implemented in GGMncv).

Conclusion

I end with a note of caution. I encourage researchers to resist over reliance upon nonconvex regularization, especially when the end game is inference. It has its place, for example, to gain the first glimpse into a dependence structure or to formulate hypotheses to then test with inferential statistics. With this in mind, nonconvex regularization is a viable option for crude inquiries.

References

- Abram, S. V., Helwig, N. E., Moodie, C. A., DeYoung, C. G., MacDonald III, A. W., & Waller, N. G. (2016). Bootstrap enhanced penalized regression for variable selection with neuroimaging data. *Frontiers in neuroscience*, 10, 344.
- Armour, C., Fried, E. I., Deserno, M. K., Tsai, J., & Pietrzak, R. H. (2017). A network analysis of DSM-5 posttraumatic stress disorder symptoms and correlates in U.S. military veterans.

- Journal of Anxiety Disorders*, 45(May 2013), 49–59. doi: 10.1016/j.janxdis.2016.11.008
- Baba, K., Shibata, R., & Sibuya, M. (2004). Partial Correlation and Conditional Correlation as Measures of Conditional Independence. *Australian & New Zealand Journal of Statistics*, 46(4), 657–664. doi: 10.1111/j.1467-842X.2004.00360.x
- Baba, K., & Sibuya, M. (2005). Equivalence of Partial and Conditional Correlation Coefficients. *J. Japan Statist. Soc.*, 35(1), 1–19.
- Becker, N., Toedt, G., Lichter, P., & Benner, A. (2011). Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data. *BMC bioinformatics*, 12(1), 138.
- Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), 521–547. doi: 10.3150/11-BEJ410
- Belloni, A., Chernozhukov, V., & Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4), 791–806. doi: 10.1093/biomet/asr043
- Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, 41(2), 802–837. doi: 10.1214/12-AOS1077
- Bernardo, J. M., & Smith, A. F. M. (2001). *Bayesian theory*. IOP Publishing.
- Borsboom, D., Cramer, A. O. J., Schmittmann, V. D., Epskamp, S., & Waldorp, L. J. (2011). The Small World of Psychopathology. *PLoS ONE*, 6(11), e27407. doi: 10.1371/journal.pone.0027407
- Boudt, K., Cornelissen, J., Croux, C., & Boudt, K. (2012). The Gaussian rank correlation estimator: robustness properties. *Stat Comput*, 22(2), 471–483. doi: 10.1007/s11222-011-9237-0
- Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, 71(3), 731–740.
- Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5(1), 232–253. doi: 10.1214/10-AOAS388
- Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6), 2350–2383.
- Bühlmann, P., Kalisch, M., & Meier, L. (2014). High-Dimensional Statistics with a View Toward Applications in Biology. *Annual Review of Statistics and Its Application*, 1(1), 255–278. doi: 10.1146/annurev-statistics-022513-115545
- Bühlmann, P., & Meier, L. (2008). Discussion: One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4), 1534–1541. doi: 10.1214/07-AOS0316A
- Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.
- Bunea, F., She, Y., Ombao, H., Gongvatana, A., Devlin, K., & Cohen, R. (2011). Penalized least squares regression methods and applications to neuroimaging. *NeuroImage*, 55(4), 1519–1527.
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33(2), 261–304. doi: 10.1177/0049124104268644
- Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling Sparsity via the Horseshoe. In *Journal of machine learning research wcp* (Vol. 5, pp. 73–80).
- Casella, G., Girón, F. J., Martínez, M. L., & Moreno, E. (2009). Consistency of bayesian procedures for variable selection. *Annals of Statistics*, 37(3), 1207–1228. doi: 10.1214/08-AOS606
- Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3), 759–771. doi: 10.1093/biomet/asn034
- Cheng, L., Zeng, P., Zhu, Y., & others. (2017). BS-SIM: An effective variable selection method for high-dimensional single index model. *Electronic Journal of Statistics*, 11(2), 3522–3548.
- Ciuperca, G. (2015). Model selection in high-dimensional quantile regression with seamless L0 penalty. *Statistics & Probability Letters*, 107, 313–323.
- Danaher, P., Wang, P., & Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76(2), 373–397. doi: 10.1111/rssb.12033
- Dempster, A. (1972, 3). Covariance Selection. *Biometrics*, 28(1), 157–175. doi: 10.2307/2528966
- Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020). The case for formal methodology in scientific reform. *bioRxiv*.
- Dicker, L., Huang, B., & Lin, X. (2013). Variable selection and estimation with the seamless-L0 penalty. *Statistica Sinica*, 23(2), 929–962. doi: 10.5705/ss.2011.074
- Drton, M., & Perlman, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika*, 91(3), 591–602. doi: 10.1093/biomet/91.3.591
- Drton, M., & Perlman, M. D. (2005). Multiple Testing and Error Control in Gaussian Graphical Model Selection. , 22(3), 430–449. doi: 10.1214/088342307000000113
- Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jeremiin, L. S. (2019). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, 00(March), 1–13. doi: 10.1093/bib/bbz016
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., & others. (2004). Least angle regression. *The Annals of statistics*, 32(2), 407–499.
- Epskamp, S. (2016). Regularized Gaussian Psychological Networks: Brief Report on the Performance of Extended BIC Model Selection. *arXiv*, 1–6.
- Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, 23(4), 617–634. doi: 10.1037/met0000167
- Epskamp, S., Waldorp, L. J., Mottus, R., & Borsboom, D. (2018). The Gaussian Graphical Model in Cross-Sectional and Time-Series Data. *Multivariate Behavioral Research*, 53(4), 453–480. doi: 10.1080/00273171.2018.1454823
- Fan, J., Feng, Y., & Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The annals of applied statistics*, 3(2), 521.

- Fan, J., Feng, Y., & Xia, L. (2020). A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models. *Journal of Econometrics*.
- Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456), 1348–1360. doi: 10.1198/016214501753382273
- Fan, J., Liao, Y., & Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econometrics Journal*, 19(1), C1–C32. doi: 10.1111/ectj.12061
- Fan, J., Xue, L., & Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, 42(3), 819–849. doi: 10.1214/13-AOS1198
- Feng, Y., & Yu, Y. (2013). Consistent cross-validation for tuning parameter selection in high-dimensional variable selection. *arXiv preprint arXiv:1308.5390*, 1(2.1).
- Fop, M., Murphy, T. B., & Scrucca, L. (2019). Model-based clustering with sparse covariance matrices. *Statistics and Computing*, 29(4), 791–819.
- Foster, D. P., & George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 1947–1975.
- Foygel, R., & Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. *Advances in Neural Information Processing Systems*, 604–612.
- Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2), 109–135.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. New York: Springer series in statistics.
- Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441. doi: 10.1093/biostatistics/kxm045
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 1.
- Frommlet, F., & Nuel, G. (2016). An adaptive ridge procedure for l0 regularization. *PloS one*, 11(2), e0148620.
- Gao, X., Pu, D. Q., Wu, Y., & Xu, H. (2009). Tuning parameter selection for penalized likelihood estimation of inverse covariance matrix. *arXiv*. doi: 10.5705/ss.2009.210
- Gao, X., Pu, D. Q., Wu, Y., & Xu, H. (2012). Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statistica Sinica*, 1123–1146.
- Gasso, G., Rakotomamonjy, A., & Canu, S. (2009). Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Transactions on Signal Processing*, 57(12), 4686–4698.
- Gelman, A., & Stern, H. (2006). The difference between significant and not significant is not itself statistically significant. *The American Statistician*, 60(4), 328–331.
- Goeman, J., Meijer, R., & Chaturvedi, N. (2018). L1 and L2 penalized regression models. *Vignette R Package Penalized*. URL <http://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf>.
- Gronau, Q. F., & Wagenmakers, E.-J. (2018). Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection. *PsyArXiv*. doi: 10.17605/OSF.IO/AT7CX
- Haslbeck, J. M. B., & Waldorp, L. J. (2015). MGM: Estimating Time-Varying Mixed Graphical Models in High-Dimensional Data.
- Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.
- Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso. *arXiv*.
- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton: CRC Press. doi: 10.1201/b18401-1
- Hazimeh, H., & Mazumder, R. (2020). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5), 1517–1537. doi: 10.1287/opre.2019.1919
- Heinävaara, O., Leppä-aho, J., Corander, J., & Honkela, A. (2016). On the inconsistency of L1-penalised sparse precision matrix estimation. *BMC Bioinformatics*, 17(Suppl 16). doi: 10.1186/s12859-016-1309-x
- Hoff, P. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1), 265–283. doi: 10.1214/07-aoas107
- Højsgaard, S., Edwards, D., & Lauritzen, S. (2012). *Graphical Models with R*. doi: 10.1007/978-1-4614-2299-0
- Jacobucci, R., Brandmaier, A. M., & Kievit, R. A. (2019). A Practical Guide to Variable Selection in Structural Equation Modeling by Using Regularized Multiple-Indicators, Multiple-Causes Models. *Advances in Methods and Practices in Psychological Science*, 2(1), 55–76. doi: 10.1177/2515245919826527
- Janková, J., & van de Geer, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9, 1205–1229. doi: 10.1214/15-EJS1031
- Javanmard, A., & Montanari, A. (2013). Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. , 1–21.
- Javanmard, A., & Montanari, A. (2015). De-biasing the Lasso: Optimal Sample Size for Gaussian Designs. *arXiv*.
- Johnson, N. S. (1979). Nonnull Properties of Kendall's Partial Rank Correlation Coefficient. *Biometrika*, 66(2), 333. doi: 10.2307/2335667
- Khondker, Z. S., Zhu, H., Chu, H., Lin, W., & Ibrahim, J. G. (2013, 4). The Bayesian Covariance Lasso. *Statistics and its interface*, 6(2), 243–259. doi: 10.1007/s11103-011-9767-z .Plastid
- Kim, D., Lee, S., & Kwon, S. (2018). A unified algorithm for the non-convex penalized estimation: The ncpen package. *arxiv*.
- Kim, S. (2015). ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications*, 22, 665–674. doi: 10.5351/CSAM.2015.22 .6.665
- Kim, Y., & Kwon, S. (2012). Global optimality of nonconvex penalized estimators. *Biometrika*, 99(2), 315–325.
- Kim, Y., Kwon, S., & Choi, H. (2012). Consistent Model Selection Criteria on High Dimensions. *Journal of Machine Learning Research*, 13, 1037–1057.
- Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon

- Press.
- Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2), 365–411. doi: 10.1016/S0047-259X(03)00096-4
- Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3), 907–927. doi: 10.1214/15-AOS1371
- Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1), 21–59. doi: 10.1017/S0266466605050036
- Leeb, H., & Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges estimator. *Journal of Econometrics*, 142(1), 201–211.
- Li, X., Zhao, T., Yuan, X., & Liu, H. (2015). The flare Package for High Dimensional Linear Regression and Precision Matrix Estimation in R *. *Journal of Machine Learning Research*, 16, 553–557.
- Li, Y., Craig, B. A., & Bhadra, A. (2017). The Graphical Horseshoe Estimator for Inverse Covariance Matrices.
- Li, Z., Wang, S., & Lin, X. (2012). Variable selection and estimation in generalized linear models with the seamless L0 penalty. *Canadian Journal of Statistics*, 40(4), 745–769.
- Liu, H., Han, F., Yuan, M., Lafferty, J., & Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4), 2293–2326. doi: 10.1214/12-aos1037
- Liu, H., Roeder, K., & Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in neural information processing systems* (pp. 1432–1440).
- Liu, H., & Wang, L. (2017). TIGER: A tuning-insensitive approach for optimally estimating gaussian graphical models. *Electronic Journal of Statistics*, 11(1), 241–294. doi: 10.1214/16-EJS1195
- Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *Annals of Statistics*, 42(2), 413–468. doi: 10.1214/13-AOS1175
- Lv, J., Fan, Y., & others. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, 37(6A), 3498–3528.
- Mazumder, R., Friedman, J. H., & Hastie, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495), 1125–1138. doi: 10.1198/jasa.2011.tm09738
- McNally, R. J. (2016). *Can network analysis transform psychopathology?* (Vol. 86). Elsevier Ltd. doi: 10.1016/j.brat.2016.06.006
- McNeish, D. M. (2015). Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. *Multivariate Behavioral Research*, 50(5), 471–484. doi: 10.1080/00273171.2015.1036965
- Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics and Data Analysis*, 52(1), 374–393. doi: 10.1016/j.csda.2006.12.019
- Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3), 1436–1462. doi: 10.1214/009053606000000281
- Meinshausen, N., & Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The annals of statistics*, 37(1), 246–270.
- Meir, A., & Drton, M. (2017). Tractable Post-Selection Maximum Likelihood Inference for the Lasso. *arXiv*, 1–58.
- Mohammadi, A., & Wit, E. C. (2015a). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1), 109–138. doi: 10.1214/14-BA889
- Mohammadi, A., & Wit, E. C. (2015b). BDgraph: An R Package for Bayesian Structure Learning in Graphical Models. doi: 10.1359/JBMR.0301229
- Nikolova, M. (2000). Local strong homogeneity of a regularized estimator. *SIAM Journal on Applied Mathematics*, 61(2), 633–658.
- Ortiz, A., Munilla, J., Álvarez-Illán, I., Górriz, J. M., & Ramírez, J. (2015). Exploratory graphical models of functional and structural connectivity patterns for Alzheimers disease diagnosis. *Frontiers in Computational Neuroscience*, 9(November), 1–18. doi: 10.3389/fncom.2015.00132
- Park, C., & Yoon, Y. J. (2011). Bridge regression: adaptivity and group selection. *Journal of Statistical Planning and Inference*, 141(11), 3506–3519.
- Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482), 681–686. doi: 10.1198/016214508000000337
- Pearson, K. (1900). I. Mathematical contributions to the theory of evolution.VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 195(262-273), 1–47.
- Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2), 5018–5051.
- Pötscher, B. M., & Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, 100(9), 2065–2082.
- Raferty, A. (1995). Bayesian model selection in social research. *Sociological methodology*, 111–163.
- Ravikumar, P., Wainwright, M. J., Raskutti, G., & Yu, B. (2011). High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5(January 2010), 935–980. doi: 10.1214/11-EJS631
- Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois. Retrieved from <https://cran.r-project.org/package=psych>
- Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373. doi: 10.1037/a0029315
- Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7, 221–264.
- Shen, X., Chen, L., Gu, Y., & So, H.-C. (2016). Square-root lasso with nonconvex regularization: An admm approach. *IEEE*

- Signal Processing Letters*, 23(7), 934–938.
- Shi, Y., Xu, D., Cao, Y., & Jiao, Y. (2019). Variable Selection via Generalized SELO-Penalized Cox Regression Models. *Journal of Systems Science and Complexity*, 32(2), 709–736.
- Shi, Y.-Y., Cao, Y.-X., Jiao, Y.-L., & Liu, Y.-Y. (2014). SICA for Coxs proportional hazards model with a diverging number of parameters. *Acta Mathematicae Applicatae Sinica, English Series*, 30(4), 887–902.
- Shi, Y.-y., Cao, Y.-x., Yu, J.-c., & Jiao, Y.-l. (2018). Variable selection via generalized SELO-penalized linear regression models. *Applied Mathematics-A Journal of Chinese Universities*, 33(2), 145–162.
- Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 44–47. doi: 10.1111/j.2517-6161.1977.tb01603.x
- Sun, T., & Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4), 879–898.
- Sun, T., & Zhang, C.-H. (2013). Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14(1), 3385–3418.
- Sustik, M. A., Calderhead, B., & Clavel, J. (2018). *glassoFast: Fast Graphical LASSO*. Retrieved from <https://cran.r-project.org/package=glassoFast>
- Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25), 7629–7634.
- Tibshirani, R. (1996). *Regression Selection and Shrinkage via the Lasso* (Vol. 58) (No. 1). doi: 10.2307/2346178
- Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73(3), 273–282. doi: 10.1111/j.1467-9868.2011.00771.x
- Van De Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3), 1166–1202. doi: 10.1214/14-AOS1221
- VanDerwerken, D. N. (2011). Variable Selection and Parameter Estimation Using a Continuous and Differentiable Approximation to the L0 Penalty Function.
- Van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89, 31–50.
- Van Wieringen, W. N., & Peeters, C. F. (2016). Ridge estimation of inverse covariance matrices from high-dimensional data. *Computational Statistics and Data Analysis*, 103, 284–303. doi: 10.1016/j.csda.2016.05.012
- Wainwright, M. J. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using L1-constrained quadratic programming. *IEEE Transactions on Information Theory*, 55(5), 2183–2202.
- Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4), 867–886. doi: 10.1214/12-BA729
- Wang, M., Wang, X., & Wang, X. (2014). A note on the one-step estimator for ultrahigh dimensionality. *Journal of Computational and Applied Mathematics*, 260, 91–98.
- Wang, Y., Fan, Q., & Zhu, L. (2018). Variable selection and estimation using a continuous approximation to the L_0 penalty. *Annals of the Institute of Statistical Mathematics*, 70(1), 191–214.
- Wang, Y., & Zhu, L. (2016). Variable Selection and Parameter Estimation with the Atan Regularization Method. *Journal of Probability and Statistics*, 2016, 1–12. doi: 10.1155/2016/6495417
- Williams, D. R. (2020a). *GGMncv: Gaussian Graphical Models with Non-Convex Penalties*.
- Williams, D. R. (2020b). Learning to Live with Sampling Variability: Expected Replicability in Partial Correlation Networks. *PsyArXiv*. doi: 10.31234/OSF.IO/FB4SA
- Williams, D. R., & Mulder, J. (2020). Bayesian hypothesis testing for Gaussian graphical models: Conditional independence and order constraints. *Journal of Mathematical Psychology*, 99, 102441.
- Williams, D. R., & Rast, P. (2019). Back to the basics: Re-thinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology*. doi: 10.1111/bmsp.12173
- Williams, D. R., Rhemtulla, M., Wysocki, A. C., & Rast, P. (2019). On Nonregularized Estimation of Psychological Networks. *Multivariate Behavioral Research*, 54(5), 1–23. doi: 10.1080/00273171.2019.1575716
- Williams, D. R., & Rodriguez, J. E. (2020). Why Overfitting is Not (Usually) a Problem in Partial Correlation Networks. *PsyArXiv*. doi: 10.31234/OSF.IO/8PR9B
- Witten, D. M., Friedman, J. H., & Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4), 892–900. doi: 10.1198/jcgs.2011.11051a
- Wysocki, A. C., & Rhemtulla, M. (2019). On Penalty Parameter Selection for Estimating Network Models. *Multivariate Behavioral Research*, 1–15. doi: 10.1080/00273171.2019.1672516
- Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92(4).
- Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2017). Using stacking to average Bayesian predictive distributions. , 1–78. doi: 10.1214/17-BA1091
- Yongxiu, C., Yuling, J., Yueyong, S., & Yanyan, L. (2018). Penalized variable selection procedure for Cox proportional hazards model via seamless- L_0 penalty. *SCIENTIA SINICA Mathematica*, 48(5), 643.
- Zeng, L., & Xie, J. (2014). Group variable selection via SCAD-L2. *Statistics*, 48(1), 49–66.
- Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2), 894–942. doi: 10.1214/09-AOS729
- Zhang, H., Sun, J., & Wang, D. (2013). Variable selection and estimation for multivariate panel count data via the seamless- L_0 penalty. *Canadian Journal of Statistics*, 41(2), 368–385.
- Zhang, R., Ren, Z., & Chen, W. (2018). SILGGM: An extensive R package for efficient statistical inference in large-scale gene networks. *PLoS Computational Biology*, 14, e1006369. doi: 10.1371/journal.pcbi.1006369

- Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1), 95–112. doi: 10.1016/j.jeconom.2015.02.006
- Zhao, P., & Yu, B. (2006). On Model Selection Consistency of Lasso. *The Journal of Machine Learning Research*, 7, 2541–2563. doi: 10.1109/TIT.2006.883611
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418–1429.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society*, 67(2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x
- Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4), 1509–1533. doi: 10.1214/009053607000000802

Appendix A Additional Penalties

The following penalties are also implemented in the R package **GGMncv**. They were not included in the paper in an effort to keep the exposition manageable. However, they too are expected to overcome the various issues of ℓ_1 -regularization and also provide a computationally efficient alternative to non-regularized model selection.

Smooth Integration of Counting and Absolute Deviation

The smooth integration of counting and absolute deviation penalty (SICA) was introduced in (Equation 9 in [Lv, Fan, & others, 2009](#)). The basic idea was to modify the transformed ℓ_1 penalty of [Nikolova \(2000\)](#), such that it provides a family of penalties. The SICA penalty, noting that $\theta \in (0, \infty)$, is defined as

$$p_{\lambda,\gamma}(|\theta|) = \lambda \left(\frac{|\theta|(\gamma + 1)}{|\theta| + \gamma} \right) \quad (15)$$

for $\gamma \in [0, \infty)$. Figure 1 in [Lv et al. \(2009\)](#) visualizes the penalty function. The resulting family of penalties “has the interpretation of a smooth homotopy between ℓ_0 and ℓ_1 penalties” (p. 3502 in [Lv et al., 2009](#)). Note that this is also the idea behind the SELO and ATAN penalties. Originally introduced for penalized least squares, the SICA has since been extended to Cox’s proportional hazard model ([Y.-Y. Shi, Cao, Jiao, & Liu, 2014](#)) and the single index model ([Cheng, Zeng, Zhu, & others, 2017](#)).

Exponential Penalty

There is an interesting line of research exploring “exponential” penalties. To my knowledge, there are three such penalties that turn out to be quite similar to one another. To save space, the most recent is described here. Termed the EXP penalty, [Y. Wang, Fan, and Zhu \(2018\)](#) proposed the following penalty

$$p_{\lambda,\gamma}(|\theta|) = \lambda \left(1 - \exp^{-\frac{|\theta|}{\gamma}} \right) \quad (16)$$

for $\gamma > 0$. Note that when $\gamma \rightarrow 0$, say, $\gamma = 0.01$, this also results in $p_{\lambda,\gamma}(\theta) \approx \lambda I\{\theta \neq 0\}$, that is, a continuous approximation to the ℓ_0 -penalty. Similar penalties can be found in [Van Derwerken](#) (p. 11, 2011) and [Breheny](#) (Equation 6 therein, 2015), with the latter for group lasso.

Log Penalty

The log penalty extends the elastic net family to accommodate nonconvex penalties ([Mazumder et al., 2011](#)). For example, the elastic net combines ℓ_1 and ℓ_2 regularization, whereas this generalization ranges between ℓ_0 (best subset) and ℓ_1 (lasso). The log penalty is defined as

$$p_{\lambda,\gamma}(|\theta|) = \frac{\lambda}{\log(\gamma + 1)} \log(\gamma\theta + 1) \quad (17)$$

for $\gamma > 0$, with $\gamma \rightarrow 0+$ approaching lasso and $\gamma \rightarrow \infty$ approximating the ℓ_0 -penalty. The log penalty was studied extensively with a one-step estimator in [Zou and Li \(2008\)](#). Note that alternative algorithms with better convergence properties are described in [Gasso, Rakotomamonjy, and Canu \(2009\)](#).

Appendix B Node Descriptions

Table B1
Node descriptions

Node	Symptom
B1	Intrusive Thoughts
B2	Nightmares
B3	Flashbacks
B4	Emotional cue reactivity
B5	Psychological cue reactivity
C1	Avoidance of thoughts
C2	Avoidance of reminders
D1	Trauma-related amnesia
D2	Negative beliefs
D3	Blame of self or others
D4	Negative trauma-related emotions
D5	Loss of interest
D6	Detachment
D7	Restricted affect
E1	Irritability/anger
E2	Self-destructive/reckless behavior
E3	Hypervigilance
E4	Exaggerated startle response
E5	Difficulty concentrating
E6	Sleep disturbance