Computational Statistics (2020) 35:629–646
https://doi.org/10.1007/s00180-019-00912-z

ORIGINAL PAPER

The spectral condition number plot for regularization parameter evaluation

Carel F. W. Peeters[1] · **Mark A. van de Wiel**[1,2] · **Wessel N. van Wieringen**[1,3]

Received: 20 December 2018 / Accepted: 2 July 2019 / Published online: 12 July 2019

Abstract
Many modern statistical applications ask for the estimation of a covariance (or precision) matrix in settings where the number of variables is larger than the number of observations. There exists a broad class of ridge-type estimators that employs regularization to cope with the subsequent singularity of the sample covariance matrix. These estimators depend on a penalty parameter and choosing its value can be hard, in terms of being computationally unfeasible or tenable only for a restricted set of ridge-type estimators. Here we introduce a simple graphical tool, the spectral condition number plot, for informed heuristic penalty parameter assessment. The proposed tool is computationally friendly and can be employed for the full class of ridge-type covariance (precision) estimators.

Keywords Eigenvalues · High-dimensional covariance (precision) estimation · ℓ_2-Penalization · Matrix condition number

1 Introduction

The covariance matrix $\mathbf{\Sigma}$ of a p-dimensional random vector $Y_i^{\mathrm{T}} \equiv [Y_{i1}, \ldots, Y_{ip}] \in \mathbb{R}^p$ is of central importance in many statistical analysis procedures. Estimation of $\mathbf{\Sigma}$ or its

This research was supported by Grant FP7-269553 (EpiRadBio) through the European Community's Seventh Framework Programme (FP7, 2007–2013).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00180-019-00912-z) contains supplementary material, which is available to authorized users.

✉ Carel F. W. Peeters
cf.peeters@amsterdamumc.nl

1 Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, Amsterdam University Medical Centers, Location VUmc, Amsterdam, The Netherlands

2 MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

3 Department of Mathematics, VU University Amsterdam, Amsterdam, The Netherlands

inverse $\boldsymbol{\Sigma}^{-1} \equiv \boldsymbol{\Omega}$ (generally known as the precision matrix) are central to, for example, multivariate regression, time-series analysis, canonical correlation analysis, discriminant analysis, and Gaussian graphical modeling. Let $\mathbf{y}_i^{\mathrm{T}}$ be a realization of Y_i^{T}. It is well-known (Stein 1975) that the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^{\mathrm{T}}$ is a poor estimate of $\boldsymbol{\Sigma}$ when p approaches the sample size n or when $p > n$. When p approaches n, $\hat{\boldsymbol{\Sigma}}$ will tend to become ill-conditioned. When $p > n$, $\hat{\boldsymbol{\Sigma}}$ is singular leaving $\hat{\boldsymbol{\Omega}}$ undefined. However, many contemporary applications in fields such as molecular biology, neuroimaging, and finance, encounter situations of interest where $p > n$.

The estimation of $\boldsymbol{\Sigma}$ can be improved by shrinking the eigenvalues of $\hat{\boldsymbol{\Sigma}}$ towards a central value (e.g., Stein 1975) or by convexly combining $\hat{\boldsymbol{\Sigma}}$ with some well-conditioned target matrix (e.g., Schäfer and Strimmer 2005). These solutions define a class of estimators that can be caught under the umbrella term 'ridge-type covariance estimation'. Such estimators depend on a penalty parameter determining the rate of shrinkage and choosing its value is of prime importance. Available procedures for choosing the penalty have some (situation dependent) disadvantages in the sense that (a) they can be computationally expensive, (b) they can be restricted to special cases within the class of ridge-type estimators, or (c) they are not guaranteed to result in a meaningful penalty-value. There is thus some demand for a generic and computationally friendly procedure on the basis of which one can (i) heuristically select an acceptable (minimal) value for the penalty parameter and (ii) assess if more formal procedures have indeed proposed an acceptable penalty-value. Here such a tool is provided on the basis of the matrix condition number.

The remainder is organized as follows. Section 2 reviews the class of ridge-type estimators of the covariance matrix. In addition, penalty parameter selection is reviewed and an exposé of the matrix condition number is given. The spectral condition number is central to the introduction of the spectral condition number plot in Sect. 3. This graphical display is posited as an exploratory tool, that may function as a fast and simple heuristic in evaluating (a range of) penalty values or in determining a (minimum) value of the penalty parameter when employing ridge estimators of the covariance or precision matrix. We emphasize that it is a generic tool, of use for all ridge-type estimators of either the covariance or precision matrix in situations in which sparsity is not a direct asset. Section 4 illustrates usage of the spectral condition number plot with data from the field of oncogenomics. This illustration exemplifies that this tool may also be of use in situations in which sparsity is indeed ultimately desired, such as in graphical modeling. The Supplementary Material contains an additional illustration regarding high-dimensional factor analysis in which sparsity is not sought after. Section 5 discourses on the software that implements the proposed graphical display. Sections 6 and 7 conclude with discussions.

2 Eigenstructure regularization and the condition number

2.1 Ridge-type shrinkage estimation

Regularization of the covariance matrix goes back to Stein (1975, 1986), who proposed shrinking the sample eigenvalues towards some central value. This work spurred a

large body of literature (see, e.g., Haff 1980, 1991; Yang and Berger 1994; Won et al. 2013; Chi and Lange 2014). Of late, two encompassing forms of what is referred to as 'ridge-type covariance estimation' have emerged.

The first form considers convex combinations of $\hat{\mathbf{\Sigma}}$ and some positive definite (p.d.) target matrix $\mathbf{T}$ (Devlin et al. 1975; Ledoit and Wolf 2003, 2004a, b; Schäfer and Strimmer 2005; Fisher and Sun 2011):

$$\hat{\mathbf{\Sigma}}^{\mathrm{I}}(\lambda_{\mathrm{I}}) = (1 - \lambda_{\mathrm{I}})\hat{\mathbf{\Sigma}} + \lambda_{\mathrm{I}}\mathbf{T}, \tag{1}$$

with penalty parameter $\lambda_{\mathrm{I}} \in (0, 1]$. Such an estimator can be motivated from the Steinian bias-variance tradeoff as it seeks to balance the high-variance, low-bias matrix $\hat{\mathbf{\Sigma}}$ with the lower-variance, higher-bias matrix $\mathbf{T}$. The second form is of more recent vintage and considers the ad-hoc estimator (Warton 2008; Yuan and Chan 2008; Ha and Sun 2014):

$$\hat{\mathbf{\Sigma}}^{\mathrm{II}}(\lambda_{\mathrm{II}}) = \hat{\mathbf{\Sigma}} + \lambda_{\mathrm{II}}\mathbf{I}_p, \tag{2}$$

with $\lambda_{\mathrm{II}} \in (0, \infty)$. This second form is motivated, much like how ridge regression was introduced by Hoerl and Kennard (1970), as an ad-hoc modification of $\hat{\mathbf{\Sigma}}$ in order to deal with singularity in the high-dimensional setting.

van Wieringen and Peeters (2016) show that both (1) and (2) can be justified as penalized maximum likelihood estimators (cf. Warton 2008). However, neither (1) nor (2) utilizes a proper ℓ_2-penalty in that perspective. Starting from the proper ridge-type ℓ_2-penalty $\frac{\lambda_a}{2}\mathrm{tr}\left[(\mathbf{\Omega} - \mathbf{T})^{\mathrm{T}}(\mathbf{\Omega} - \mathbf{T})\right]$, van Wieringen and Peeters (2016) derive an alternative estimator:

$$\hat{\mathbf{\Sigma}}^{a}(\lambda_a) = \left[\lambda_a\mathbf{I}_p + \frac{1}{4}(\hat{\mathbf{\Sigma}} - \lambda_a\mathbf{T})^2\right]^{1/2} + \frac{1}{2}(\hat{\mathbf{\Sigma}} - \lambda_a\mathbf{T}), \tag{3}$$

with $\lambda_a \in (0, \infty)$. van Wieringen and Peeters (2016) show that, when considering a p.d. $\mathbf{T}$, the estimator (3) is an alternative to (1) with superior behavior in terms of risk. When considering the non-p.d. choice $\mathbf{T} = \mathbf{0}$, they show that (3) acts as an alternative to (2), again with superior behavior.

Clearly, one may obtain ridge estimators of the precision matrix by considering the inverses of (1), (2), and (3). For comparisons of these estimators see Lin and Perlman (1985), Daniels and Kass (2001) and van Wieringen and Peeters (2016). For expositions of other penalized covariance and precision estimators we confine by referring to Pourahmadi (2013).

2.2 Penalty parameter selection

The choice of penalty-value is crucial to the aforementioned estimators. Let λ denote a generic penalty. When choosing λ too small, an ill-conditioned estimate may ensue when $p > n$ (see Sect. 2.3). When choosing λ too large, relevant data signal may be lost. Many options for choosing λ are available. The ridge estimators, in contrast to ℓ_1-regularized estimators of the covariance or precision matrix (e.g., Friedman et al. 2008; Bien and Tibshirani 2011), do not generally produce sparse estimates. This implies that

model-selection-consistent methods (such as usage of the BIC), are not appropriate. Rather, for ℓ_2-type estimators, it is more appropriate to seek loss efficiency.

A generic strategy for determining an optimal value for λ that can be used for any ridge-type estimator of Sect. 2.1 is k-fold cross-validation (of the likelihood function). Asymptotically, such an approach can be explained in terms of minimizing Kullback-Leibler divergence. Unfortunately, this strategy is computationally prohibitive for large p and/or large n. Lian (2011) and Vujačić et al. (2015) propose approximate solutions to the leave-one-out cross-validation score. While these approximations imply gains in computational efficiency, they are not guaranteed to propose a reasonable optimal value for λ.

Ledoit and Wolf (2004b) propose a strategy to determine analytically an optimal value for λ under a modified Frobenius loss for the estimator (1) under certain choices of $\mathbf{T}$ (cf. Fisher and Sun 2011). This optimal value requires information on the unknown population matrix $\boldsymbol{\Sigma}$. The optimal penalty-value thus needs to be approximated with some estimate of $\boldsymbol{\Sigma}$. Ledoit and Wolf (2004b) utilize an n-consistent estimator while Schäfer and Strimmer (2005) use the unbiased estimate $\frac{n}{n-1}\hat{\boldsymbol{\Sigma}}$. In practice, this may result in overshrinkage (Daniels and Kass 2001) or even negative penalty-values (Schäfer and Strimmer 2005).

Given the concerns stated above, there is some demand for a generic and computationally friendly tool for usage in the following situations of interest:

(i) When one wants to speedily determine a (minimal) value for λ for which $\hat{\boldsymbol{\Sigma}}(\lambda)$ is well-conditioned;
(ii) When one wants to determine speedily whether an optimal λ proposed by some other (formal) procedure indeed leads to a well-conditioned estimate $\hat{\boldsymbol{\Sigma}}(\lambda)$;
(iii) When one wants to determine speedily a reasonable minimal value for λ for usage in a search-grid (for an optimal such value) by other, optimization-based, procedures.

In Sect. 3 we propose such a tool based on the spectral condition number.

2.3 Spectral condition number

The estimators from Sect. 2.1 are p.d. when their penalty-values are strictly positive. However, they are not necessarily well-conditioned for any strictly positive penalty-value when $p \gtrsim n$, especially when the penalty takes a value in the lower range. We seek to quantify the condition of the estimators w.r.t. a given penalty-value. To this end we utilize a condition number (Von Neumann and Goldstine 1947; Turing 1948).

Consider a nonsingular matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ as well as the matrix norm $\|\cdot\|$. The condition number w.r.t. matrix inversion can be defined as (Higham 1995)

$$\text{cond}(\mathbf{A}) := \lim_{\epsilon\to 0^+} \sup_{\|\delta\mathbf{A}\|\leq\epsilon\|\mathbf{A}\|} \frac{\|(\mathbf{A}+\delta\mathbf{A})^{-1}-\mathbf{A}^{-1}\|}{\epsilon\|\mathbf{A}^{-1}\|}, \tag{4}$$

indicating the sensitivity of inversion of $\mathbf{A}$ w.r.t. small perturbations $\delta\mathbf{A}$. When the norm in (4) is induced by a vector norm the condition number is characterized as (Higham 1995):

$$\text{cond}(\mathbf{A}) = \mathcal{C}(\mathbf{A}) := \|\mathbf{A}\|\|\mathbf{A}^{-1}\|. \tag{5}$$

For singular matrices $\mathcal{C}(\mathbf{A})$ would equal ∞. Hence, a high condition number is indicative of near-singularity and quantifies an ill-conditioned matrix. Indeed, the inverse of the condition number gives the relative distance of $\mathbf{A}$ to the set of singular matrices $\mathcal{S}$ (Demmel 1987):

$$\text{dist}(\mathbf{A}, \mathcal{S}) = \inf_{\mathbf{S} \in \mathcal{S}} \frac{\|\mathbf{A} - \mathbf{S}\|}{\|\mathbf{A}\|} = \frac{1}{\mathcal{C}(\mathbf{A})}.$$

Another interpretation can be found in error propagation. A high condition number implies severe loss of accuracy or large propagation of error when performing matrix inversion under finite precision arithmetic. One can expect to loose at least $\lfloor \log_{10} \mathcal{C}(\mathbf{A}) \rfloor$ digits of accuracy in computing the inverse of $\mathbf{A}$ (e.g., Chapter 8 and Section 6.4 of, respectively, Cheney and Kincaid 2008; Gentle 2007). In terms of error propagation, $\mathcal{C}(\mathbf{A})$ is also a reasonable sensitivity measure for linear systems $\mathbf{Ax} = \mathbf{b}$ (Higham 1995).

We can specify (5) with regard to a particular norm. We have special interest in the ℓ_2-condition number $\mathcal{C}_2(\mathbf{A})$, usually called the *spectral condition number* for its relation to the spectral decomposition. When $\mathbf{A}$ is a symmetric p.d. matrix, it is well-known that (Von Neumann and Goldstine 1947; Gentle 2007)

$$\mathcal{C}_2(\mathbf{A}) = \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2 = \frac{d(\mathbf{A})_1}{d(\mathbf{A})_p},$$

where $d(\mathbf{A})_1 \geq \ldots \geq d(\mathbf{A})_p$ are the eigenvalues of $\mathbf{A}$. We can connect the machinery of ridge-type regularization to this spectral condition number.

Let $\mathbf{V}\mathbf{D}(\hat{\mathbf{\Sigma}})\mathbf{V}^{\mathrm{T}}$ be the spectral decomposition of $\hat{\mathbf{\Sigma}}$ with $\mathbf{D}(\hat{\mathbf{\Sigma}})$ denoting a diagonal matrix with the eigenvalues of $\hat{\mathbf{\Sigma}}$ on the diagonal and where $\mathbf{V}$ denotes the matrix that contains the corresponding eigenvectors as columns. Note that the orthogonality of $\mathbf{V}$ implies $\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_p$. This decomposition can then be used to show that, like the Stein estimator (Stein 1975), estimate (2) is rotation equivariant: $\hat{\mathbf{\Sigma}}^{\mathrm{II}}(\lambda_{\mathrm{II}}) = \mathbf{V}\mathbf{D}(\hat{\mathbf{\Sigma}})\mathbf{V}^{\mathrm{T}} + \lambda_{\mathrm{II}}\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{V}[\mathbf{D}(\hat{\mathbf{\Sigma}}) + \lambda_{\mathrm{II}}\mathbf{I}_p]\mathbf{V}^{\mathrm{T}}$. That is, the estimator leaves the eigenvectors of $\hat{\mathbf{\Sigma}}$ intact and thus solely performs shrinkage on the eigenvalues. When choosing $\mathbf{T} = \varphi\mathbf{I}_p$ with $\varphi \in [0, \infty)$, the estimator (3) also is of the rotation equivariant form, as we may then write:

$$\hat{\mathbf{\Sigma}}^{a}(\lambda_a) = \mathbf{V}\left\{\left[\lambda_a\mathbf{I}_p + \frac{1}{4}[\mathbf{D}(\hat{\mathbf{\Sigma}}) - \lambda_a\varphi\mathbf{I}_p]^2\right]^{1/2} + \frac{1}{2}[\mathbf{D}(\hat{\mathbf{\Sigma}}) - \lambda_a\varphi\mathbf{I}_p]\right\}\mathbf{V}^{\mathrm{T}}. \quad (6)$$

An analogous decomposition can be given for the estimator (1) when choosing $\mathbf{T} = \mu\mathbf{I}_p$, with $\mu \in (0, \infty)$. The effect of the shrinkage estimators can then, using the rotation equivariance property, also be explained in terms of restricting the spectral condition number. For example, using (6), we have:

$$1 \leq \frac{\sqrt{\lambda_a + [d(\hat{\mathbf{\Sigma}})_1 - \lambda_a\varphi]^2/4} + [d(\hat{\mathbf{\Sigma}})_1 - \lambda_a\varphi]/2}{\sqrt{\lambda_a + [d(\hat{\mathbf{\Sigma}})_p - \lambda_a\varphi]^2/4} + [d(\hat{\mathbf{\Sigma}})_p - \lambda_a\varphi]/2} < \frac{d(\hat{\mathbf{\Sigma}})_1}{d(\hat{\mathbf{\Sigma}})_p} \leq \infty.$$

Similar statements can be made regarding all rotation equivariant versions of the estimators discussed in Sect. 2.1. Similar statements can also be made when considering a target $\mathbf{T}$ for estimators (1) and (3) that is not of the form $\alpha\mathbf{I}_p$ whenever this target is well-conditioned (i.e., has a lower condition number than $\hat{\mathbf{\Sigma}}$).

Example 1 For clarification consider the following toy example. Assume we have a sample covariance matrix $\hat{\mathbf{\Sigma}}$ whose largest eigenvalue $d(\hat{\mathbf{\Sigma}})_1 = 3$. Additionally assume that $\hat{\mathbf{\Sigma}}$ is estimated in a situation where $p > n$ so that $d(\hat{\mathbf{\Sigma}})_p = 0$ and, hence, $d(\hat{\mathbf{\Sigma}})_1/d(\hat{\mathbf{\Sigma}})_p = 3/0 = \infty$ (under the IEEE computing Standard for Floating-Point Arithmetic; IEEE Computer Society 2008). Say we are interested in regularization using the estimator (3) using a scalar target matrix with $\varphi = 2$. Even using a very small penalty of $\lambda_a = 1 \times 10^{-1}$ it is then quickly verified that $d[\hat{\mathbf{\Sigma}}^a(\lambda_a)]_1/d[\hat{\mathbf{\Sigma}}^a(\lambda_a)]_p = 300{,}003 < d(\hat{\mathbf{\Sigma}})_1/d(\hat{\mathbf{\Sigma}})_p$. Under a large penalty such as $\lambda_a = 10{,}000$ we find that $d[\hat{\mathbf{\Sigma}}^a(\lambda_a)]_1/d[\hat{\mathbf{\Sigma}}^a(\lambda_a)]_p = 1.00015$. Indeed, van Wieringen and Peeters (2016) have shown that, in this rotation equivariant setting, $d[\hat{\mathbf{\Sigma}}^a(\lambda_a)]_j \to 1/\varphi$ as $\lambda_a \to \infty$ for all j. Hence, $d[\hat{\mathbf{\Sigma}}^a(\lambda_a)]_1/d[\hat{\mathbf{\Sigma}}^a(\lambda_a)]_p \to 1$ as the penalty value grows very large. Section 1 of the Supplementary Material visualizes this behavior (in the given setting) for various scenarios of interest: $\varphi < 1/d(\hat{\mathbf{\Sigma}})_1$, $1/d(\hat{\mathbf{\Sigma}})_1 < \varphi < 1$, $\varphi = 1$, $1 < \varphi < d(\hat{\mathbf{\Sigma}})_1$, and $\varphi > d(\hat{\mathbf{\Sigma}})_1$. □

Let $\hat{\mathbf{\Sigma}}(\lambda)$ denote a generic ridge-type estimator of the covariance matrix under generic penalty λ. We thus quantify the conditioning of $\hat{\mathbf{\Sigma}}(\lambda)$ for given λ (and possibly a given $\mathbf{T}$) w.r.t. perturbations $\lambda + \delta\lambda$ with

$$\mathcal{C}_2[\hat{\mathbf{\Sigma}}(\lambda)] = \|\hat{\mathbf{\Sigma}}(\lambda)\|_2\|\hat{\mathbf{\Omega}}(\lambda)\|_2 = \frac{d[\hat{\mathbf{\Sigma}}(\lambda)]_1}{d[\hat{\mathbf{\Sigma}}(\lambda)]_p}. \tag{7}$$

A useful property is that $\mathcal{C}_2[\hat{\mathbf{\Sigma}}(\lambda)] = \mathcal{C}_2[\hat{\mathbf{\Omega}}(\lambda)]$, i.e., knowing the condition of the covariance matrix implies knowing the condition of the precision matrix (so essential in contemporary topics such as graphical modeling). The condition number (7) can be used to construct a simple and computationally friendly visual tool for penalty parameter evaluation.

3 The spectral condition number plot

3.1 The basic plot

As one may appreciate from the exposition in Sect. 2.3, when $\hat{\mathbf{\Sigma}}(\lambda)$ moves away from near-singularity, small increments in the penalty λ can cause dramatic changes in $\mathcal{C}_2[\hat{\mathbf{\Sigma}}(\lambda)]$. One can expect that at some point along the domain of λ, its value will be large enough for $\mathcal{C}_2[\hat{\mathbf{\Sigma}}(\lambda)]$ to stabilize (in some relative sense). We will cast these expectations regarding the behavior in a (loose) definition:

Heuristic Definition. Let $\hat{\mathbf{\Sigma}}(\lambda)$ denote a generic ridge-type estimator of the covariance matrix under generic fixed penalty λ. In addition, let $\Delta\lambda$ indicate a real perturbation in

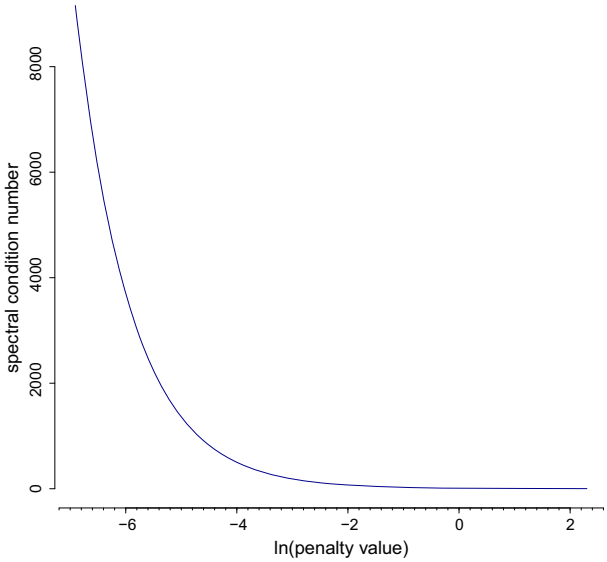

Fig. 1 First example of a spectral condition number plot. The ridge estimator used is given in (2). R code to generate the data and produce the plot can be found in Section 4 of the Supplementary Material. A reasonable minimal value for the penalty parameter can be found at the x-axis at approximately -3. The exponent of this number signifies a minimal value of the penalty for which the estimate $\hat{\mathbf{\Sigma}}(\lambda)$ is well-conditioned according to the Heuristic Definition. The condition number $\mathcal{C}_2[\hat{\mathbf{\Sigma}}^{\mathrm{II}}(\exp(-3))] \approx 184.95$

λ as opposed to the theoretical perturbation $\delta\lambda$. We will term the estimate $\hat{\mathbf{\Sigma}}(\lambda)$ *well-conditioned* when small increments $\Delta\lambda$ in λ translate to (relatively) small changes in $\mathcal{C}_2[\hat{\mathbf{\Sigma}}(\lambda + \Delta\lambda)]$ vis-à-vis $\mathcal{C}_2[\hat{\mathbf{\Sigma}}(\lambda)]$.

From experience, when considering ridge-type estimation of $\mathbf{\Sigma}$ or its inverse in $p > n$ situations, the point of relative stabilization can be characterized by a leveling-off of the acceleration along the curve when plotting the condition number $\mathcal{C}_2[\hat{\mathbf{\Sigma}}(\lambda)]$ against the (chosen) domain of λ. Consider Fig. 1, which is the first example of what we call the *spectral condition number plot*.

Figure 1 indeed plots (7) against the natural logarithm of λ. As should be clear from Sect. 2.3, the spectral condition number displays (mostly) decreasing concave upward behavior in the feasible domain of the penalty parameter with a vertical asymptote at $\ln(0)$ and a horizontal asymptote at $\mathcal{C}_2(\mathbf{T})$ (which amounts to 1 in case of a strictly positive scalar target). The logarithm is used on the x-axis as, especially for estimators (2) and (3), it is more natural to consider orders of magnitude for λ. In addition, usage of the logarithm 'decompresses' the lower domain of λ, which enhances the visualization of the point of relative stabilization, as it is in the lower domain of the penalty parameter where ill-conditioning usually ensues when $p > n$. Figure 1 uses simulated data (see Section 4 of the Supplementary Material) with $p = 100$ and $n = 25$. The estimator used is the ad-hoc ridge-type estimator given by (2). One can observe relative stabilization of the spectral condition number—in the sense of the Heuristic Definition—at approximately $\exp(-3) \approx .05$. This value can be taken as a

reasonable (minimal) value for the penalty parameter. The spectral condition number plot can be a simple visual tool of interest in the situations sketched in Sect. 2.2, as will be illustrated in Sect. 4.

3.2 Interpretational aids

The basic spectral condition number plot can be amended with interpretational aids. The software (see Sect. 5) can add two such aids to form a panel of plots. These aids support the heuristic choice for a penalty-value and provide additional information on the basic plot.

The first aid is the visualization of $\lfloor \log_{10} \mathcal{C}_2[\hat{\boldsymbol{\Sigma}}(\lambda)] \rfloor$ against the domain of λ. As stated in Sect. 2.3, $\lfloor \log_{10} \mathcal{C}_2[\hat{\boldsymbol{\Sigma}}(\lambda)] \rfloor$ provides an estimate of the digits of accuracy one can expect to loose (on top of the digit loss due to inherent numerical imprecision) in operations based on $\hat{\boldsymbol{\Sigma}}(\lambda)$. Note that this estimate is dependent on the norm. This aid can, for example, support choosing a (miminal) penalty-value on the basis of the error propagation (in terms of approximate loss in digits of accuracy) one finds acceptable. Figure 2 gives an example.

Let $\mathfrak{C}_{\ln(\lambda)\mapsto\mathcal{C}_2[\hat{\boldsymbol{\Sigma}}(\lambda)]}$ denote the curvature (of the basic plot) that maps the natural logarithm of the penalty-value to the condition number of the regularized precision matrix. We seek to approximate the second-order derivative of this curvature (the acceleration) at given penalty-values in the domain $[\lambda_{\min}, \lambda_{\max}]$. The software (see Sect. 5) requires the specification of $\lambda_{\min}$ and $\lambda_{\max}$, as well as the number of steps one wants to take along the domain $[\lambda_{\min}, \lambda_{\max}]$. Say we take $s = 1, \ldots, S$ steps, such that $\lambda_1 = \lambda_{\min}, \lambda_2, \ldots, \lambda_{S-1}, \lambda_S = \lambda_{\max}$. The implementation takes steps that are log-equidistant, hence $\ln(\lambda_s) - \ln(\lambda_{s-1}) = [\ln(\lambda_{\max}) - \ln(\lambda_{\min})]/(S-1) \equiv \tau$ for all $s = 2, \ldots, S$. The central finite difference approximation to the second-order derivative (see e.g., LeVeque 2007) of $\mathfrak{C}_{\ln(\lambda)\mapsto\mathcal{C}_2[\hat{\boldsymbol{\Sigma}}(\lambda)]}$ at $\ln(\lambda_s)$ then takes the following form:

$$\mathfrak{C}''_{\ln(\lambda_s)\mapsto\mathcal{C}_2[\hat{\boldsymbol{\Sigma}}(\lambda_s)]} \approx \frac{\mathcal{C}_2[\hat{\boldsymbol{\Sigma}}(\lambda_{s+1})] - 2\mathcal{C}_2[\hat{\boldsymbol{\Sigma}}(\lambda_s)] + \mathcal{C}_2[\hat{\boldsymbol{\Sigma}}(\lambda_{s-1})]}{\tau^2},$$

which is available for $s = 2, \ldots, S-1$. The second visual aid thus plots $\mathfrak{C}''_{\ln(\lambda)\mapsto\mathcal{C}_2[\hat{\boldsymbol{\Sigma}}(\lambda)]}$ against the feasible domain of λ (see Fig. 2 for an example). The behavior of the condition number along the domain of the penalty-value is to always decrease. This decreasing behavior is not consistently concave upward for (1) (under general non-scalar targets) and (3), as there may be parts of the domain where the behavior is concave downward. This aid may then more clearly indicate inflection points in the regularization path of the condition number. In addition, this aid may put perspective on the point of relative stabilization when the y-axis of the basic plot represents a very wide range.

3.3 Choosing $\lambda_{\min}$ and $\lambda_{\max}$ for plotting

As stated, the software requires the specification of $\lambda_{\min}$ and $\lambda_{\max}$. Practical choices for these arguments depend on the type of ridge-estimator one wants to use. Estimator

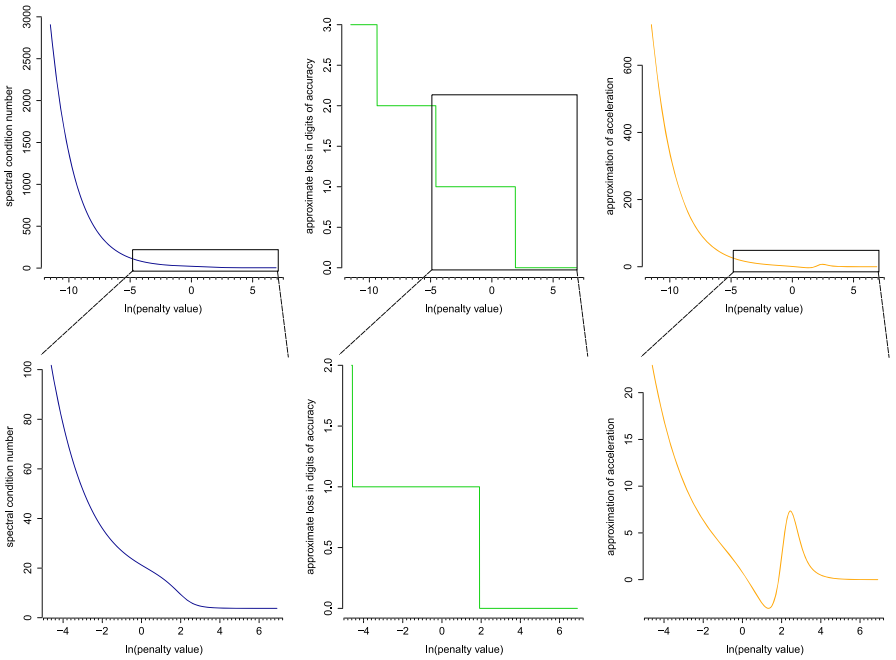

Fig. 2 The spectral condition number plot with interpretational aids. The data are the same as for Fig. 1 (see Section 4 of the Supplementary Material). The ridge estimator used is given in (3) with target $\mathbf{T} = (\hat{\boldsymbol{\Sigma}} \circ \mathbf{I}_p)^{-1}$. This estimator exhibits nonlinear shrinkage. The left-hand panels give the basic spectral condition number plot. The middle and right-hand panels exemplify the interpretational aids to the basic plot: the approximate loss in digits of accuracy (middle panel) and the approximation of the acceleration along the curve in the basic plot (right-hand panel). The top panels give the basic condition number plot and its interpretational aids for the domain $\lambda_a \in [1 \times 10^{-5}, 1000]$. The bottom panels zoom in on the boxed areas. The interpretational aids can support the selection of a (minimal) penalty-value and may provide additional information on the basic plot. For example, say we are interested in choosing (approximately) the minimal value for the penalty for which the error propagation (in terms of approximate loss in digits of accuracy) is at most 1. From the middle panels we see that we should then choose the penalty-value to be $\exp(-4.2)$. From the right-hand panels we may infer that the regularization path of the condition number displays decreasing concave downward behavior for penalty-values between approximately $\exp(.2)$ and $\exp(1.8)$

(1) has a natural upper-bound to its domain for the penalty-value: 1. It is then natural to set $\lambda_{\max} = 1$ for this estimator. One needs to choose $\lambda_{\min}$ strictly positive, but small such that it is indicative of ill-conditioning when present. A practical choice that abides these wishes is 1×10^{-5}. Hence, when using estimator (1) in producing the condition number plot we recommend to set $\lambda_{\mathrm{I}} \in [1 \times 10^{-5}, 1]$

Estimators (2) and (3) do not have a natural upper-bound to the domain for their penalty-values. Hence, here one also needs to choose $\lambda_{\max}$ in such a way that the domain of well-conditionedness is represented. We can do this by realizing that estimators (1) and (3) behave, when they have the same p.d. target matrix, similarly at the boundaries of the penalty domain when mapping λ_{I} and λ_a to the same scale (van Wieringen and Peeters 2016). This mapping may be obtained as $\lambda_{\mathrm{I}} = 1 - 1/(\lambda_a + 1)$. When $\lambda_a = 1 \times 10^{-5}$ then $\lambda_{\mathrm{I}} \approx 1 \times 10^{-5}$. When $\lambda_a = 20$ then $\lambda_{\mathrm{I}} \approx .95$, implying almost full shrinkage towards the target in the latter. Hence, for plotting one may

choose $\lambda_a \in [1 \times 10^{-5}, 20]$. As (2) behaves similarly to (3) at the boundaries of the penalty parameter domain when the latter has the null matrix as the target, it is also a good practical choice to set $\lambda_{\text{II}} \in [1 \times 10^{-5}, 20]$. Note that most illustrations abide by these recommendations.

Software implementing the spectral condition number plot is discussed in Sect. 5. The following section illustrates, using oncogenomics data, the various uses of the spectral condition number plot with regard to covariance or precision matrix regularization. Section 2 of the Supplementary Material contains a second data example to further illustrate usage of the condition number plot. Section 4 of the Supplementary Material contains all R code with which these illustrations can be reproduced (including querying the data).

4 Illustration

4.1 Context and data

Various histological variants of kidney cancer are designated with the amalgamation 'renal cell carcinoma' (RCC). Chromophobe RCC (ChRCC) is a rather rare and predominantly sporadic histological variant, accounting for 4-6% of RCC cases (Stec et al. 2009). ChRCC originates in the distal convoluted tubule (Shuch et al. 2015), a portion of the nephron (the basic structural unit of the kidney) that serves to maintain electrolyte balance (Subramanya and Ellison 2014). Often, histological variants of cancer have a distinct pathogenesis contingent upon the deregulation of certain molecular pathways. A pathway can be thought of as a collection of molecular features that work interdependently to regulate some biochemical function. Recent evidence suggests that (reactivation of) the Hedgehog (Hh) signaling pathway may support cancer development and progression in clear cell RCC (CCRCC) (Dormoy et al. 2009; D'Amato et al. 2014), the most common subtype of RCC. The Hh-signaling pathway is crucial in the sense that it "orchestrates tissue patterning" in embryonic development, making it "critical to normal kidney development, as it regulates the proliferation and differentiation of mesenchymal cells in the metanephric kidney" (D'Amato et al. 2014). Later in life Hh-signaling is largely silenced and constitutive reactivation may elicit and support tumor growth and vascularization (Dormoy et al. 2009; D'Amato et al. 2014). Our goal here is to explore if Hh-signaling might also be reactivated in ChRCC. The exploration will make use of network modeling (see Sect. 4.3) in which the network is taken as a representation of a biochemical pathway. This exercise hinges upon a well-conditioned precision matrix.

We attained data on RCC from the The Cancer Genome Atlas Research Network (2013) as queried through the Cancer Genomics Data Server (Cerami et al. 2012; Gao et al. 2013) using the `cgdsr` R-package (Jacobsen 2015). All ChRCC samples were retrieved for which messenger ribonucleic acid (mRNA) data is available, giving a total of $n = 15$ samples. The data stem from the IlluminaHiSeq_RNASeqV2 RNA sequencing platform and consist of normalized relative gene expressions. That is, individual gene expressions are given as mRNA z-scores relative to a reference population that consists of all tumors that are diploid for the gene in question. All

features were retained that map to the Hh-signaling pathway according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000), giving a total of $p = 56$ gene features. Regularization of the desired precision matrix is needed as $p > n$. Even though features from genomic data are often measured on or mapped to (approximately) the same scale, regularization on the standardized scale is often appropriate as the variability of the features may differ substantially when $p > n$; a point also made by Warton (2008). Note that we may use the correlation matrix $\mathbf{R} = (\hat{\mathbf{\Sigma}} \circ \mathbf{I}_p)^{-\frac{1}{2}} \hat{\mathbf{\Sigma}} (\hat{\mathbf{\Sigma}} \circ \mathbf{I}_p)^{-\frac{1}{2}}$ instead of $\hat{\mathbf{\Sigma}}$ in Eqs. (1) to (3) without loss of generality.

4.2 Penalty parameter evaluation and selection

The precision estimator of choice is the inverse of (3). The target matrix is chosen as $\mathbf{T} = \varphi \mathbf{I}_p$, with φ set to the reciprocal of the average eigenvalue of $\mathbf{R}$: 1. First, the approximate leave-one-out cross-validation (aLOOCV) procedure (Lian 2011; Vujačić et al. 2015) is used (on the negative log-likelihood) in finding an optimal value for λ_a under the given target and data settings. This procedure searches for the optimal value λ_a^* in the domain $\lambda_a \in [1 \times 10^{-5}, 20]$. A relatively fine-grained grid of 10,000 log-equidistant steps along this domain points to 1×10^{-5} as being the optimal value for the penalty (in the chosen domain). This value seems low given the p/n ratio of the data. This calls for usage-type (ii) of the condition number plot (Sect. 2.2), where one uses it to determine if an optimal penalty as proposed by some procedure indeed leads to a well-conditioned estimate. The condition number is plotted over the same penalty-domain considered by the aLOOCV procedure. The left-hand panel of Fig. 3 depicts this condition number plot. The dashed (green) vertical line represents the penalty-value that was chosen as optimal by the aLOOCV procedure. Clearly, the precision estimate at $\lambda_a = 1 \times 10^{-5}$ is not well-conditioned in the sense of the Heuristic Definition. This exemplifies that the (essentially large-sample) approximation to the LOOCV score may not be suitable for non-sparse situations and/or for situations in which the p/n ratio grows more extreme (the negative log-likelihood term then tends to completely dominate the bias term). At this point one could use the condition number plot in accordance with usage-type (i), in which one seeks a reasonable minimal penalty-value. This reasonable minimal value (in accordance with the Heuristic Definition) can be found at approximately $\exp(-6)$, at which $\mathcal{C}_2[\hat{\mathbf{\Omega}}^a(\exp(-6))] = \mathcal{C}_2[\hat{\mathbf{\Sigma}}^a(\exp(-6))] \approx 247.66$.

One could worry that the precision estimate retains too much noise under the heuristic minimal penalty-value. To this end, a proper LOOCV procedure is implemented that makes use of the root-finding Brent algorithm (Brent 1971). The expectation is that the proper data-driven LOOCV procedure will find an optimal penalty-value in the domain of λ_a for which the estimate is well-conditioned. The penalty-space of search is thus constrained to the region of well-conditionedness for additional speed, exemplifying usage-type (iii) of the condition number plot. Hence, the LOOCV procedure is told to search for the optimal value λ_a^* in the domain $\lambda_a \in [\exp(-6), 20]$. The optimal penalty-value is indeed found to the right of the heuristic minimal value at 5.2. At this value, indicated by the solid (red) vertical line in Fig. 3,

$\mathcal{C}_2[\hat{\boldsymbol{\Omega}}^a(5.2)] = \mathcal{C}_2[\hat{\boldsymbol{\Sigma}}^a(5.2)] \approx 8.76$. The precision estimate at this penalty-value is used in further analysis.

4.3 Further analysis

Biochemical networks are often reconstructed from data through graphical models. Graphical modeling refers to a class of probabilistic models that uses graphs to express conditional (in)dependence relations (i.e., Markov properties) between random variables. Let $\mathcal{V}$ denote a finite set of vertices that correspond to a collection of random variables with probability distribution $\mathcal{P}$, i.e., $\{Y_1, \ldots, Y_p\} \sim \mathcal{P}$. Let $\mathcal{E}$ denote a set of edges, where edges are understood to consist of pairs of distinct vertices such that Y_j is connected to $Y_{j'}$, i.e., $Y_j - Y_{j'} \in \mathcal{E}$. We then consider graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ under the basic assumption $\{Y_1, \ldots, Y_p\} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. In this Gaussian case, conditional independence between a pair of variables corresponds to zero entries in the precision matrix. Indeed, the following relations can be shown to hold for all pairs $\{Y_j, Y_{j'}\} \in \mathcal{V}$ with $j \neq j'$ (see, e.g., Whittaker 1990):

$$(\hat{\boldsymbol{\Omega}})_{jj'} = 0 \iff Y_j \perp\!\!\!\perp Y_{j'} | \mathcal{V} \backslash \{Y_j, Y_{j'}\} \iff Y_j \not- Y_{j'},$$

where $\not-$ indicates the absence of an edge. Hence, the graphical model can be selected by determining the support of the precision matrix. For support determination we resort to a local false discovery rate procedure proposed by Schäfer and Strimmer (2005), retaining only those edges whose empirical posterior probability of being present equals or exceeds .80.

Note that the coupling of a ridge estimate of the precision matrix with post-hoc edge selection differs from the dominant graphical lasso approach to graphical modeling (Friedman et al. 2008) which induces automatic sparsity. It is well-known that ℓ_1-based estimation (and thus support recovery) is consistent only under the assumption that the true graphical model is (very) sparse. When the number of truly non-null elements exceeds the sample size the ℓ_1-penalty is unable to retrieve the sparsity pattern (van Wieringen and Peeters 2016). This is undesirable as there is accumulating evidence that many networks, such as biochemical pathways involved in disease aetiology, are dense (Boyle et al. 2017). In such a situation the coupling of a non-sparsity-inducing penalty with a post-hoc selection step such as the local false discovery rate can outperform the (graphical) lasso (van Wieringen and Peeters 2016; Bilgrau et al. 2015). These considerations underlie our chosen approach.

The right-hand panel of Fig. 3 represents the retrieved Markov network on the basis of $\hat{\boldsymbol{\Omega}}^a(5.2)$. The vertex-labels are curated gene names of the Hh-signaling pathway genes. The graph seems to retrieve salient features of the Hh-signaling pathway. The Hh-signaling pathway involves a cascade from the members of the Hh-family (IHH, SHH, and DHH) via the SMO gene to ZIC2 and members of the Wnt-signaling pathway. The largest connected component is indicative of this cascade, giving tentative evidence of reactivation of the Hh-signaling pathway in (at least) rudimentary form in ChRCC.

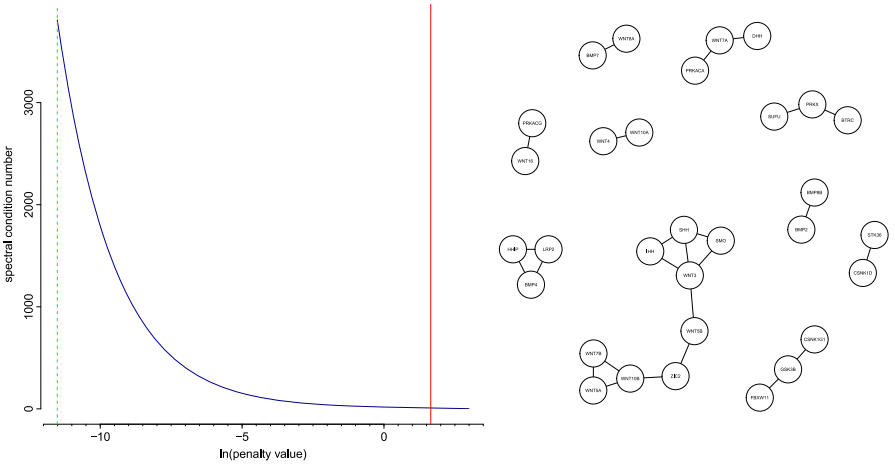

Fig. 3 Left-hand panel: Condition number plot of the Hedgehog (Hh) signaling pathway variables on the chromophobe kidney cancer data of The Cancer Genome Atlas Research Network (2013). The dashed vertical line indicates the value of the penalty parameter that was chosen as optimal by the aLOOCV procedure (1×10^{-5}). The solid vertical line indicates the value of the penalty that was chosen as optimal by the root-finding LOOCV procedure (5.2). The value indicated by the aLOOCV procedure does not lie in a region where the estimate can be deemed well-conditioned. Right-hand panel: The retrieved Markov network using the optimal penalty-value as indicated by the root-finding LOOCV procedure. The vertex-labels are Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC) curated gene names of the Hh-signaling pathway genes. Certain salient features of the Hh-signaling pathway are retrieved, indicating that this pathway may be reactivated in ChRCC

5 Software

The R-package `rags2ridges` (Peeters et al. 2019) implements the ridge estimators of Sect. 2.1 and the condition number plot through the function `CNplot`. This function outputs visualizations such as Figs. 1 and 2. The condition number is efficiently determined by the calculation of (solely) the largest and smallest eigenvalues using an implicitly restarted Lanczos method (IRLM; Colvetti et al. 1994) available through the `RSpectra` package (Qiu and Mei 2019). For most practical purposes this calculation is fast enough, especially in rotation equivariant situations for which only a single IRLM run is required to obtain the complete solution path of the condition number. Additional computational speed is achieved by the implementation of core operations in C++ via `Rcpp` and `RcppArmadillo` (Eddelbuettel and François 2011; Eddelbuettel 2013). For example, producing the basic condition number plot for the estimator (3) on the basis of data with dimensions $p = 1000$ and $n = 200$, using a scalar target matrix and 1000 steps along the penalty-domain, will take approximately .35 s (see Section 3 of the Supplementary Material for a benchmark exercise). The additional computational cost of the interpretational aids is linear: producing the panel of plots (including interpretational aids) for the situation just given takes approximately .38 s. When λ is very small and $p \gg n$ the calculation of the condition number may suffer from rounding problems (much like the imaginary linear system $\mathbf{\Sigma x} = \mathbf{b}$), but remains adequate in its indication of ill-conditioning.

When spectral computation is deemed too costly in terms of floating-point operations, or when one wants more speed in determining the condition number, the `CNplot` function offers the option to cheaply approximate the ℓ_1-condition number, which amounts to

$$\mathcal{C}_1[\hat{\boldsymbol{\Sigma}}(\lambda)] = \| \hat{\boldsymbol{\Sigma}}(\lambda) \|_1 \| \hat{\boldsymbol{\Omega}}(\lambda) \|_1 = \left\{ \max_{j'} \sum_{j} |[\hat{\boldsymbol{\Sigma}}(\lambda)]_{jj'}| \right\} \left\{ \max_{j'} \sum_{j} |[\hat{\boldsymbol{\Omega}}(\lambda)]_{jj'}| \right\}.$$

The ℓ_1-condition number is computationally less complex than the calculation of $\mathcal{C}_2[\hat{\boldsymbol{\Sigma}}(\lambda)]$ in non-rotation equivariant settings. The machinery of ridge-type regularization is, however, less directly connected to this ℓ_1-condition number (in comparison to the ℓ_2-condition number). The approximation of $\mathcal{C}_1[\hat{\boldsymbol{\Sigma}}(\lambda)]$ uses LAPACK routines (Anderson et al. 1999) and avoids overflow. This approximation is accessed through the `rcond` function from R (R Development Core Team 2011). The package `rags2ridges` is freely available from the Comprehensive R Archive Network (http://cran.r-project.org/) (R Development Core Team 2011).

6 Discussion

The condition number plot is a heuristic tool and heuristics should be handled with care. Below, some cases are presented that serve as notes of caution. They exemplify that the proposed heuristic accompanying the condition number plot should not be applied (as any statistical technique) without proper inspection of the data.

6.1 Artificial ill-conditioning

A first concern is that, when the variables are measured on different scales, artificial ill-conditioning may ensue (see, e.g., Gentle 2007). In case one worries if the condition number is an adequate indication of error propagation when using variables on their original scale one can ensure that the columns (or rows) of the input matrix are on the same scale. This is easily achieved by scaling the input covariance matrix to be the correlation matrix. Another issue is that it is not guaranteed that the condition number plot will give an unequivocal point of relative stabilization for every data problem (which hinges in part on the chosen domain of the penalty parameter). Such situations can be dealt with by extending the domain of the penalty parameter or by determining the value of λ that corresponds to the loss of $\lfloor \log_{10} \mathcal{C}[\hat{\boldsymbol{\Sigma}}(\lambda)] \rfloor$ digits (in the imaginary linear system $\boldsymbol{\Sigma}\mathbf{x} = \mathbf{b}$) one finds acceptable.

6.2 Naturally high condition numbers

Some covariance matrices may have high condition numbers as their 'natural state'. Consider the following covariance matrix: $\boldsymbol{\Sigma}_{\text{equi}} = (1 - \varrho)\mathbf{I}_p + \varrho \mathbf{J}_p$, with $-1/(p - 1) < \varrho < 1$ and where $\mathbf{J}_p$ denotes the $(p \times p)$-dimensional all-ones matrix. The

variates are thus equicorrelated with unit variance. The eigenvalues of this covariance matrix are $p\varrho + (1 - \varrho)$ and (with multiplicity $p - 1$) $1 - \varrho$. Consequently, its condition number equals $1 + p\varrho/(1-\varrho)$. The condition number of $\mathbf{\Sigma}_{\text{equi}}$ thus becomes high when the number of variates grows large and/or the (marginal) correlation ϱ approaches one (or $-1/(p-1)$). The large ratio between the largest and smallest eigenvalues of $\mathbf{\Sigma}_{\text{equi}}$ in such situations mimics a high-dimensional setting in which any non-zero eigenvalue of the sample covariance estimate is infinitely larger than the smallest (zero) eigenvalues. However, irrespective of the number of samples, any sample covariance estimate of an $\mathbf{\Sigma}_{\text{equi}}$ with large p and ϱ close to unity (or $-1/(p-1)$) exhibits such a large ratio. Would one estimate the $\mathbf{\Sigma}_{\text{equi}}$ in penalized fashion (even for reasonable sample sizes) and choose the penalty parameter from the condition number plot as recommended, then one would select a penalty parameter that yields a 'well-conditioned' estimate. Effectively, this amounts to limiting the difference between the penalized eigenvalues, which need not give a condition number representative of $\mathbf{\Sigma}_{\text{equi}}$. Thus, the recommendation to select the penalty parameter from the well-conditioned domain of the condition number plot may in some (perhaps exotic) cases lead to a choice that crushes too much relevant signal (shrinking the largest eigenvalue too much). For high-dimensional settings this may be unavoidable, but for larger sample sizes this is undesirable.

6.3 Contamination

Real data is often contaminated with outliers. To illustrate the potential effect of outliers on the usage of the condition number plot, consider data $\mathbf{y}_i$ drawn from a contaminated distribution, typically modeled by a mixture distribution: $\mathbf{y}_i \sim (1-\phi)\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}) + \phi\mathcal{N}_p(\mathbf{0}, c\mathbf{I}_p)$ for $i = 1, \ldots, n$, some positive constant $c > 0$, and mixing proportion $\phi \in [0, 1]$. Then, the expectation of the sample covariance matrix $\mathbb{E}(\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}) = (1-\phi)\mathbf{\Sigma} + c\phi\mathbf{I}_p$. Its eigenvalues are: $d[(1-\phi)\mathbf{\Sigma} + c\phi\mathbf{I}_p]_j = (1-\phi)d(\mathbf{\Sigma})_j + c\phi$, for $j = 1, \ldots, p$. In high-dimensional settings with few samples the presence of any outlier corresponds to mixing proportions clearly deviating from zero. In combination with any substantial c the contribution of the outlier(s) to the eigenvalues may be such that the contaminated sample covariance matrix is represented as better conditioned (vis-à-vis its uncontaminated counterpart). It is the outlier(s) that will determine the domain of well-conditionedness in such a situation. Then, when choosing the penalty parameter in accordance with the Heuristic Definition, undershrinkage may ensue. In situations in which the results are influenced by outliers one has several options at disposal. One could simply trim the data as a preprocessing step before obtaining $\hat{\mathbf{\Sigma}}$. Another option would be to use techniques for identifying (and subsequently removing) multivariate outliers such as those based on the robust Mahalanobis distance (Mahalanobis 1936). One may also opt to use a robust estimator for $\hat{\mathbf{\Sigma}}$, such as the well-known Minimum Covariance Determinant estimator (Rousseeuw 1984), in producing the condition number plot.

7 Conclusion

We have proposed a simple visual display that may be of aid in determining the value of the penalty parameter in ridge-type estimation of the covariance or precision matrix when the number of variables is large relative to the sample size. The visualization we propose plots the spectral condition number against the domain of the penalty parameter. As the value of the penalty parameter increases, the covariance (or precision) matrix will move away from (near) singularity. In some lower-end of the domain this will mean that small increments in the value of the penalty parameter will lead to large decreases of the condition number. At some point, the condition number can be expected to stabilize, in the sense that small increments in the value of the penalty have (relatively) little effect on the condition number. The point of relative stabilization may be deemed to indicate a reasonable (minimal) value for the penalty parameter. Hence, in analogy to usage of the scree plot in factor analysis (Cattell 1966), initial interest will lie with the assessment of the 'elbow' of the plot.

Usage of the condition number plot was exemplified in situations concerned with the direct estimation of covariance or precision matrices. The plot may also be of interest in situations in which (scaled versions of) these matrices are conducive to further computational procedures. For example, it may support the ridge approach to the regression problem $\mathbf{x} = \mathbf{Y}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. We would then assess the conditioning of $\mathbf{Y}^{\mathrm{T}}\mathbf{Y} + \lambda\mathbf{I}_p$ for use in the ridge-solution to the regression coefficients: $\hat{\boldsymbol{\beta}} = (\mathbf{Y}^{\mathrm{T}}\mathbf{Y} + \lambda\mathbf{I}_p)^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{x}$.

We explicitly state that we view the proposed condition number plot as an *heuristic tool*. We emphasize 'tool', as it gives easy and fast access to penalty-value assessment and determination without proposing an optimal (in some sense) value. Also, in the tradition of exploratory data analysis (Tukey 1977), usage of the condition number plot requires good judgment. As any heuristic method, it is not without flaws.

Notwithstanding these concerns, the condition number plot gives access to a fast and generic (i.e., non-target and non-ridge-type specific) procedure for regularization parameter determination that is of use when analytic solutions are not available and when other procedures fail. In addition, the condition number plot may aid more formal procedures, in terms of assessing if a well-conditioned estimate is indeed obtained, and in terms of proposing a reasonable minimal value for the regularization parameter for usage in a search grid.

Acknowledgements The Authors would like to thank two anonymous reviewers whose constructive comments have led to an improvement in presentation.

References

Anderson E, Bai Z, Bischof C, Blackford S, Demmel J, Dongarra J, Du Croz J, Greenbaum A, Hammarling S, McKenney A, Sorensen D (1999) LAPACK users' guide, 3rd edn. Society for Industrial and Applied Mathematics, Philadelphia

Bien J, Tibshirani R (2011) Sparse estimation of a covariance matrix. Biometrika 98:807–820
Bilgrau AE, Peeters CFW, Eriksen PS, Boegsted M, van Wieringen WN (2015) Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes. Technical report. arXiv:1509.07982 [stat.ME]
Boyle EA, Li YI, Pritchard JK (2017) An expanded view of complex traits: from polygenic to omnigenic. Cell 169:1177–1186
Brent RP (1971) An algorithm with guaranteed convergence for finding a zero of a function. Comput J 14:422–425
Cattell RB (1966) The scree test for the number of factors. Multivar Behav Res 1:245–276
Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2:401–404
Cheney W, Kincaid D (2008) Numerical computing and mathematics, 6th edn. Thomson Brooks, Belmont
Chi EC, Lange K (2014) Stable estimation of a covariance matrix guided by nuclear norm penalties. Comput Stat Data Anal 80:117–128
Colvetti D, Reichel L, Sorensen DC (1994) An implicitely restarted Lanczos method for large symmetric eigenvalue problems. Electron Trans Numer Anal 2:1–21
D'Amato C, Rosa R, Marciano R, D'Amato V, Formisano L, Nappi L, Raimondo L, Di Mauro C, Servetto A, Fulciniti F, Cipolletta A, Bianco C, Ciardiello F, Veneziani BM, De Placido S, Bianco R (2014) Inhibition of Hedgehog signalling by NVP-LDE225 (Erismodegib) interferes with growth and invasion of human renal cell carcinoma cells. Br J Cancer 111:1168–1179
Daniels MJ, Kass RE (2001) Shrinkage estimators for covariance matrices. Biometrics 57:1173–1184
Demmel JW (1987) On condition numbers and the distance to the nearest ill-posed problem. Numer Math 51:251–289
Devlin SJ, Gnanadesikan R, Kettenring JR (1975) Robust estimation and outlier detection with correlation coefficients. Biometrika 62:531–545
Dormoy V, Danilin S, Lindner V, Thomas L, Rothhut S, Coquard C, Helwig JJ, Jacqmin D, Lang H, Massfelder T (2009) The sonic hedgehog signaling pathway is reactivated in human renal cell carcinoma and plays orchestral role in tumor growth. Mol Cancer 8:123
Eddelbuettel D (2013) Seamless R and C++ integration with `Rcpp`. Springer, New York
Eddelbuettel D, François R (2011) `Rcpp`: seamless R and C++ integration. J Stat Softw 40(8):1–18
Fisher TJ, Sun X (2011) Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. Comput Stat Data Anal 55:1909–1918
Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9:432–441
Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 6:pl1
Gentle JE (2007) Matrix algebra: theory, computations, and applications in statistics. Springer, New York
Ha MJ, Sun W (2014) Partial correlation matrix estimation using ridge penalty followed by thresholding and re-estimation. Biometrics 70:765–773
Haff LR (1980) Empirical Bayes estimation of the multivariate normal covariance matrix. Ann Stat 8:586–597
Haff LR (1991) The variational form of certain Bayes estimators. Ann Stat 19:1163–1190
Higham DJ (1995) Condition numbers and their condition numbers. Linear Algebra Appl 214:193–213
Hoerl AE, Kennard R (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67
IEEE Computer Society (2008) IEEE standard for floating-point arithmetic. IEEE Std 754–2008, pp 1–70
Jacobsen A (2015) cgdsr: R-based API for accessing the MSKCC Cancer Genomics Data Server (CGDS). R package version 1.2.5. http://CRAN.R-project.org/package=cgdsr. Accessed 13 Apr 2019
Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucl Acids Res 28(1):27–30
Ledoit O, Wolf M (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J Empir Finance 10:603–621
Ledoit O, Wolf M (2004a) Honey, I shrunk the sample covariance matrix. J Portf Manag 30:110–119
Ledoit O, Wolf M (2004b) A well-conditioned estimator for large-dimensional covariance matrices. J Multivar Anal 88:365–411

LeVeque RJ (2007) Finite difference methods for ordinary and partial differential equations: steady state and time dependent problems. Society for Industrial and Applied Mathematics (SIAM), Philadelphia
Lian H (2011) Shrinkage tuning parameter selection in precision matrices estimation. J Stat Plan Inference 141:2839–2848
Lin S, Perlman M (1985) A Monte Carlo comparison of four estimators of a covariance matrix. In: Krishnaiah PR (ed) Multivariate analysis, 6th edn. North Holland, Amsterdam, pp 411–429
Mahalanobis PC (1936) On the generalised distance in statistics. Proc Natl Inst Sci India 2:49–55
Peeters CFW, Bilgrau AE, van Wieringen WN (2019) rags2ridges: Ridge estimation of precision matrices from high-dimensional data. R package version 2.2.1. http://cran.r-project.org/package=rags2ridges. Accessed 13 Apr 2019
Pourahmadi M (2013) High-dimensional covariance estimation. Wiley, Hoboken
Qiu Y, Mei J (2019) RSpectra: solvers for large-scale eigenvalue and SVD problems. R package version 0.14-0. https://CRAN.R-project.org/package=RSpectra. Accessed 13 Apr 2019
R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. ISBN 3-900051-07-0. Accessed 13 Apr 2019
Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–880
Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4:art. 32
Shuch B, Amin A, Armstrong AJ, Eble JN, Ficarra V, Lopez-Beltran A, Martignoni G, Rini BI, Kutikov A (2015) Understanding pathologic variants of renal cell carcinoma: distilling therapeutic opportunities from biologic complexity. Eur Urol 67:85–97
Stec R, Grala B, Mączewski M, Bodnar L, Szczylik C (2009) Chromophobe renal cell cancer-review of the literature and potential methods of treating metastatic disease. J Exp Clin Cancer Res 28:134
Stein C (1975) Estimation of a covariance matrix. Rietz Lecture. 39th Annual Meeting IMS. Atlanta, Georgia
Stein C (1986) Lectures on the theory of estimation of many parameters. J Math Sci 34:1373–1403
Subramanya AR, Ellison DH (2014) Distal convoluted tubule. Clin J Am Soc Nephrol 9:2147–2163
The Cancer Genome Atlas Research Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499:43–49
Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Boston
Turing AM (1948) Rounding-off errors in matrix processes. Q J Mech Appl Math 1:287–308
van Wieringen WN, Peeters CFW (2016) Ridge estimation of inverse covariance matrices from high-dimensional data. Comput Stat Data Anal 103:284–303
Von Neumann J, Goldstine HH (1947) Numerical inverting of matrices of high order. Bull Am Math Soc 53:1021–1099
Vujačić I, Abbruzzo A, Wit EC (2015) A computationally fast alternative to cross-validation in penalized Gaussian graphical models. J Stat Comput Simul 85:3628–3640
Warton DI (2008) Penalized normal likelihood and ridge regularization of correlation and covariance matrices. J Am Stat Assoc 103:340–349
Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, Chichester
Won JH, Lim J, Kim SJ, Rajaratnam B (2013) Condition-number-regularized covariance estimation. J R Stat Soc Ser B 75:427–450
Yang R, Berger JO (1994) Estimation of a covariance matrix using the reference prior. Ann Stat 22:1195–1211
Yuan KH, Chan W (2008) Structural equation modeling with near singular covariance matrices. Comput Stat Data Anal 52:4842–4858