Journal de la Société Française de Statistique
Vol. 160 No. 3 (2019)

Discussion of "Minimal penalties and the slope heuristics: a survey" by Sylvain Arlot

Titre: Discussion sur "Pénalités minimales et heuristique de pente" par Sylvain Arlot

David Donoho[1] and Matan Gavish[2]

The Scree plot diagnostic for the number of factors (Catell, 1966) – and analogous procedures for selection of number of principal components, singular values, etc. – is phenomenally popular across science and engineering, with 10,000+ Google Scholar citations. This, despite being more than 50 years old – so that it predates the Internet and Google by decades; despite considerable vagueness in the specification of what one actually does in using the plot; and despite widespread ignorance among users of why this might or might not be a good idea.

Cattell's original observation was purely heuristic, based on a very preliminary and seemingly casual empirical observation. Cattell noticed that, when plotting eigenvalues in decreasing order versus eigenrank (1 = largest), one often observes, after some initial idiosyncrasies, a roughly straight line in the plot – which he called the "Scree", in analogy with tailings at the base of mountain landslides. He suggestively identified the Scree with a null model - eigenvalues caused by pure noise rather than signal. As it has come down to us across the decades, our task in looking at such a plot is to identify an 'elbow' separating the early idiosyncratic piece from the final Scree piece.

Modern random matrix theory has reached a state where we can now use it to understand mechanisms leading to Cattell's observations and thereby evaluate the whole Scree plot phenomenon. Indeed, the Scree plot of n eigenvalues or singular values, with the i-th largest eigenvalue plotted against the horizontal value i/n, is none other than a (sense-reversed) quantile (inverse CDF) function of the empirical eigenvalue or singular value distribution (Figure 1).

Using random matrix theory, we can today develop some generative noise models in which there are unambiguously no factors whatsoever, and where the quantile function is indeed well approximated by a straight line. We can also develop generative pure noise models where there is seemingly an elbow (Figure 2) – despite there being no factors at all!

In general, the linear slope heuristic is invalid for many different kinds of noise distributions. Even for the white noise model, when the aspect ratio m/n (m being the number of columns and n being the number of rows) is far from 1, the quantile function of noise-only eigenvalues is far from linear.

To shed further light, Gavish and Donoho (2014) plotted simple histograms of the empirical eigenvalues (Figure 3). Such histograms reveal the shape of the empirical eigenvalue density

[1] Department of Statistics, Stanford University
[2] School of Computer Science and Engineering, The Hebrew University of Jerusalem

 ISSN: 2102-6238

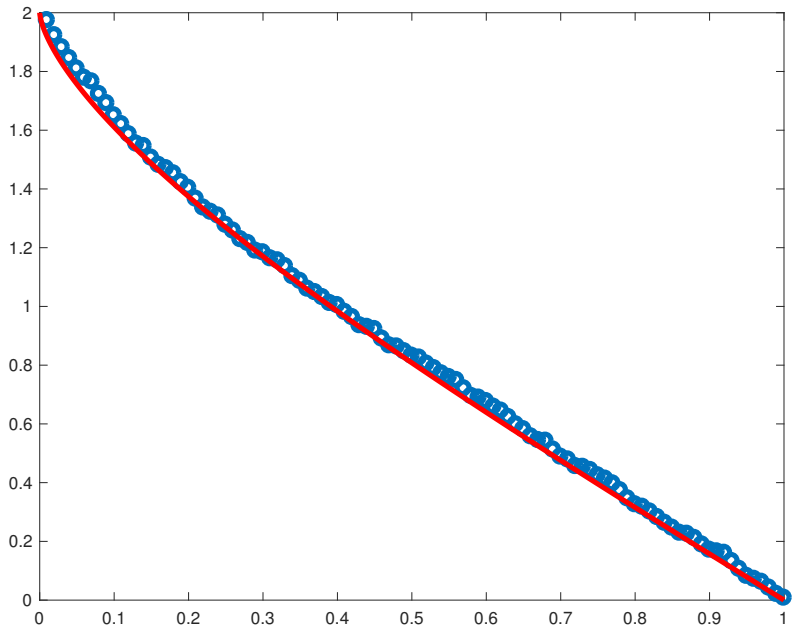

Figure 1: Horizontal axis: i/n. Vertical axis: size of i-th singular value. Blue circles: Scree plot of singular values of 100-by-100 white noise matrix. Red line: quantile function of the Quarter-Circle law, the limiting distribution of singular values for white-noise matrices, (Bais and Silverstein, 2010).

and show that noise distributions often give rise to a compactly supported eigenvalue density, known as the Bulk. Under many random matrix models for the noise, the Bulk has a defined, hard edge at an analytically predicted point which we call the *bulk edge*. In real data, one may observe that the bulk is well-described by such a model, but that a few individual outliers stick out beyond the bulk edge, and which are not predicted under the noise model. We call these the *signal* eigenvalues; their corresponding eigenvectors can be shown to carry useful information about out-of-sample data.

Such outlier eigenvalues correspond to the initial idiosyncratic piece at the left end of the scree plot; although the scree itself will not generally have a linear piece. Indeed it could only be linear if the bulk distribution were uniformly distributed – i.e. had a flat histogram across it support. In the useful models we know, the bulk is *never* uniformly distributed.

In the paper under discussion, Sylvain Arlot performs a useful service in connecting minimal penalty methods with Scree plots. The minimal penalty method corresponds to thresholding eigenvalues at the location of the bulk edge. And, indeed, there are formal proposals, such as "Universal Singular Value Thresholding" (Chatterjee, 2013) which amount to (a slight perturbation of) the minimal penalty method.

In the case $m = n$, the minimal penalty method would suggest the threshold $2 \cdot \sqrt{n}\sigma$, while

 ISSN: 2102-6238

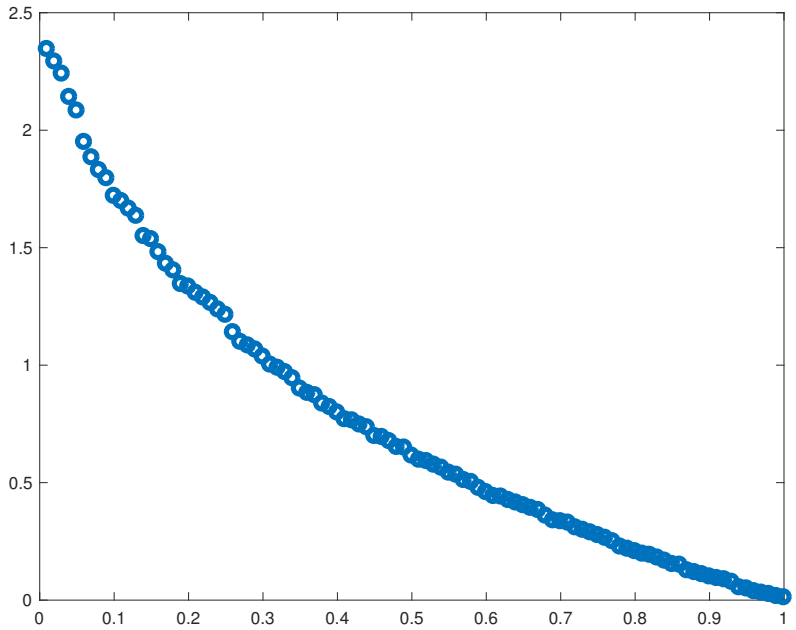

Figure 2: Horizontal axis: i/n. Vertical axis: size of i-th singular value. Blue circles: Scree plot of singular values of 100-by-100 matrix where each row has mean 0 and covariance $\sim \chi_{10}$, (Bais and Silverstein, 2010).

Chatterjee's proposal would suggest something slightly larger, say $2.01 \cdot \sqrt{n}\sigma$. (These values are particular to white noise and Frobenius loss function).

How do such proposals perform? Using results deriving from Random Matrix Theory (for example, Florent Benaych-Georges and Raj Rao-Nadakuditi), one can quantify the performance of the minimal penalty method and similar schemes in recovering factors under various noise models. Gavish and Donoho (2014) demonstrated that, under a spiked model appropriate for the case of a low-rank signal matrix, using white noise/Frobenius norm assumptions, thresholding at the bulk edge itself is noticeably sub-optimal. One gets MSE improvements of up to 60% by taking the threshold at about $2.31\ldots \cdot \sqrt{n}\sigma$. Moreover $2.31\ldots = 4/\sqrt{3}$ is the unique admissible choice. Since $2.31 > 2$ the minimal penalty/Scree prescription is seriously outperformed by some larger choice.

In practical terms, even though an eigenvalue sticks out of the bulk – and even though the corresponding eigenvector carries some solid information – it should still properly be ignored, *unless it sticks out a certain specific distance*, specific to the noise model generating the bulk. The optimal penalty is noticeably larger than the minimal penalty.

This conclusion signals a general phenomenon, not an artifact of any particular probabilistic model or loss function. Taking into account the inaccuracy of the empirical eigenvectors of the data matrix in estimating the underlying population eigenvectors, caused by the noise, inevitably

 ISSN: 2102-6238

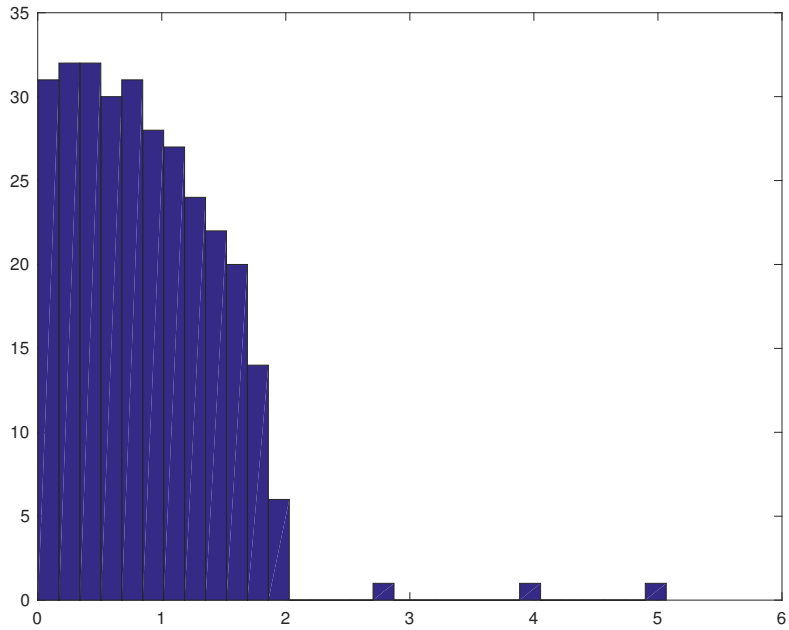

Figure 3: Histogram of singular values of 300-by-300 matrix $Y = X + Z$, where X is rank-3 and Z is a white noise matrix. Note the quarter-circle shape of the bulk (corresponding to quantile function in Figure 1) and three spikes emerging from the bulk.

leads to the conclusion that the optimal threshold for eigenvalues is necessarily noticeably larger than the bulk edge. The exact location of the optimal threshold depends on the noise distribution and the function at hand. In short, the minimal penalty is suboptimal, and some larger choice will be optimal, based on the loss function and noise model.

In summary: the once-mysterious observation, that noise-only models sometimes yield roughly straight Scree plots of eigenvalues and singular values, is made clear by modern random matrix theory. The Scree plot just shows an empirical quantile function. The straight-line approximation sometimes works and often fails. A histogram seems to be a more suitable way to understand this phenomenon. The Scree plot method, and similar methods visually looking for departure from the Scree, correspond to bulk-edge thresholding, where the bulk is apparent in the histogram representation. It was demonstrated generally that bulk-edge thresholding is suboptimal due to eigenvector rotation caused by the noise; a more careful analysis obtains the optimal threshold suitable to the particular noise distribution at hand. Modern random matrix theory thus helps explain traditional heuristics, evaluate them, and replace them with well-understood optimal procedures. The threshold $(1+\sqrt{m/n})\sqrt{n}\sigma$ is particular to white noise and a specific choice of loss function; it is suboptimal *even under* the assumptions which derived it; and it should certainly not be used universally.

 ISSN: 2102-6238

References

Gavish, M., and D. L. Donoho. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. IEEE Transactions on Information Theory 60.8: 5040-5053.

Chatterjee (2013). Matrix estimation by universal singular value thresholding. *The Annals of Statistics* 43.1: 177-214.

Bais, Z. and Silverstein, J. W.(2010). Bais, Z. and Silverstein, J. W. *Spectral Analysis of Large Dimensional Random Matrices (2nd Edition).* Springer New York

 ISSN: 2102-6238