Choosing the Regularization Parameter

At our disposal: several regularization methods, based on filtering of the SVD components.

Often fairly straightforward to "eyeball" a good TSVD truncation parameter from the Picard plot.

Need: a reliable and automated technique for choosing the regularization parameter, such as k (for TSVD) or λ (for Tikhonov).

Specifically: an efficient, robust, and reliable method for computing the regularization parameter from the given data, which does not require the computation of the SVD or any human inspection of a plot.

1. Perspectives on regularization
2. The discrepancy principle
3. Generalized cross validation (GCV)
4. The L-curve criterion
5. The NCP method

Once Again: Tikhonov Regularization

Focus on Tikhonov regularization; ideas carry over to many other methods.

Recall that the Tikhonov solution x_λ solves the problem

$$\min_x \left\{ \|A\,x - b\|_2^2 + \lambda^2 \|x\|_2^2 \right\},$$

and that it is formally given by

$$x_\lambda = (A^T A + \lambda^2 I)^{-1} A^T b = A_\lambda^{\#} b,$$

where $A_\lambda^{\#} = (A^T A + \lambda^2 I)^{-1} A^T$ is a "regularized inverse."

Our noise model

$$b = b^{\text{exact}} + e$$

where $b^{\text{exact}} = A\,x^{\text{exact}}$ and e is the error.

Classical and Pragmatic Parameter-Choice

Assume we are given the problem $A\,x = b$ with

$$b = b^{\text{exact}} + e \qquad \text{and} \qquad b^{\text{exact}} = A\,x^{\text{exact}} \ ,$$

and that we have a strategy for choosing the regularization parameter λ as a function of the “noise level” $\|e\|_2$.

Then *classical* parameter-choice analysis is concerned with the convergence rates of

$$x_\lambda \to x^{\text{exact}} \qquad \text{as} \qquad \|e\|_2 \to 0 \quad \text{and} \quad \lambda \to 0 \ .$$

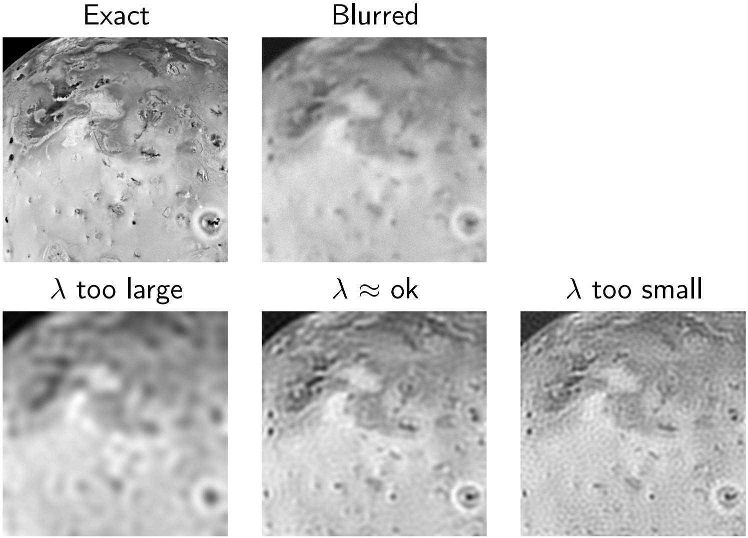
This is an important and natural requirement to algorithms for choosing λ.

Our focus here is on the typical situation in practice:

- The norm $\|e\|_2$ is not known, and
- the errors are fixed (not practical to repeat the measurements).

The *pragmatic* approach to choosing the regularization parameter is based on the forward/prediction error, or the backward error.

An Example (Image of Io, a Moon of Saturn)



Exact

Blurred

λ too large

$\lambda \approx$ ok

λ too small

Perspectives on Regularization

Problem formulation: balance the fit (residual) and the size of solution.

$$x_\lambda = \arg\min \left\{ \|A\,x - b\|_2^2 + \lambda^2 \|L\,x\|_2^2 \right\}$$

Cannot be used for choosing λ.

Forward error: balance regularization errors and perturbation errors.

$$\begin{aligned} x^{\text{exact}} - x_\lambda &= x^{\text{exact}} - A_\lambda^{\#}(b^{\text{exact}} + e) \\ &= \left(I - A_\lambda^{\#} A\right) x^{\text{exact}} - A_\lambda^{\#} e \; . \end{aligned}$$

Backward/prediction error: balance contributions from the exact data and the perturbation.

$$\begin{aligned} b^{\text{exact}} - A\,x_\lambda &= b^{\text{exact}} - A\,A_\lambda^{\#}(b^{\text{exact}} + e) \\ &= \left(I - A\,A_\lambda^{\#}\right) b^{\text{exact}} - A\,A_\lambda^{\#} e \; . \end{aligned}$$

More About the Forward Error

The forward error in the SVD basis:

$$\begin{aligned} x^{\text{exact}} - x_\lambda &= x^{\text{exact}} - V\,\Phi^{[\lambda]}\,\Sigma^{-1}\,U^T b \\ &= x^{\text{exact}} - V\,\Phi^{[\lambda]}\,\Sigma^{-1}\,U^T A\,x^{\text{exact}} - V\,\Phi^{[\lambda]}\,\Sigma^{-1}\,U^T e \\ &= V\left(I - \Phi^{[\lambda]}\right) V^T x^{\text{exact}} - V\,\Phi^{[\lambda]}\,\Sigma^{-1}\,U^T e. \end{aligned}$$

The first term is the *regularization error:*

$$\Delta x_{\text{bias}} = V\left(I - \Phi^{[\lambda]}\right) V^T x^{\text{exact}} = \sum_{i=1}^{n} \left(1 - \varphi_i^{[\lambda]}\right) \left(v_i^T x^{\text{exact}}\right) v_i,$$

and we recognize this as (minus) the bias term.

The second error term is the *perturbation error:*

$$\Delta x_{\text{pert}} = V\,\Phi^{[\lambda]}\,\Sigma^{-1}\,U^T e.$$

Regularization and Perturbation Errors – TSVD

For TSVD solutions, the regularization and perturbation errors take the form

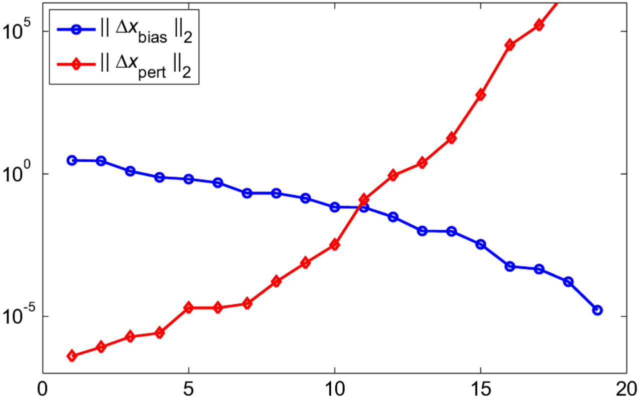
$$\Delta x_{\text{bias}} = \sum_{i=k+1}^{n} (v_i^T x^{\text{exact}})\, v_i, \qquad \Delta x_{\text{pert}} = \sum_{i=1}^{k} \frac{u_i^T e}{\sigma_i}\, v_i.$$

We use the truncation parameter k to prevent the perturbation error from blowing up (due to the division by the small singular values), at the cost of introducing bias in the regularized solution.

A "good" choice of the truncation parameter k should balance these two components of the forward error (see next slide).

The behavior of $\|x_k\|_2$ and $\|A\, x_k - b\|_2$ is closely related to these errors – see the analysis in §5.1.

The Regularization and Perturbation Errors



The norm of the regularization and perturbation error for TSVD as a function of the truncation parameter k. The two different errors approximately balance each other for $k = 11$.

The TSVD Residual

Let k_η denote the index that marks the transition between decaying and flat coefficients $|u_i^T b|$.

Due to the discrete Picard condition, the coefficients $|u_i^T b|/\sigma_i$ will also decay, on the average, for all $i < k_\eta$.

$$k < k_\eta : \; \|A\,x_k - b\|_2^2 \approx \sum_{i=k+1}^{k_\eta} (u_i^T b)^2 + (n - k_\eta)\eta^2 \approx \sum_{i=k+1}^{k_\eta} (u_i^T b^{\text{exact}})^2$$

$$k > k_\eta : \; \|A\,x_k - b\|_2^2 \approx (n-k)\,\eta^2 .$$

For $k < k_\eta$ the residual norm decreases steadily with k.

For $k > k_\eta$ it decreases much more slowly.

The transition between the two types of behavior occurs at $k = k_\eta$ when the regularization and perturbation errors are balanced.

The Discrepancy Principle

Recall that $\mathcal{E}(\|e\|_2) \approx n^{1/2}\eta$.

We should ideally choose k such that $\|A\,x_k - b\|_2 \approx (n-k)^{1/2}\,\eta$.

The *discrepancy principle* (DP) seeks to combine this:

Assume we have an upper bound δ_e for the noise level, then solve

$$\|A\,x_\lambda - b\|_2 = \tau\,\delta_e\ , \quad \text{where} \qquad \|e\|_2 \le \delta_e$$

and τ is some parameter $\tau = O(1)$. See next slide.

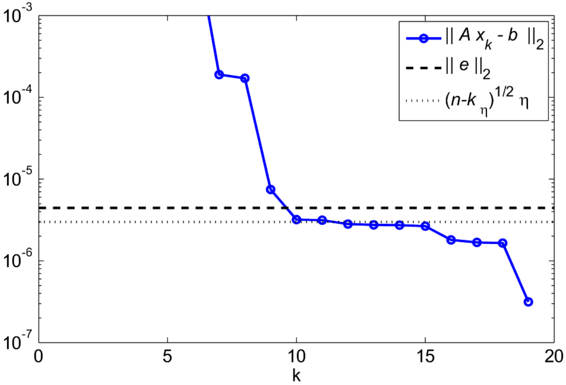
A statistician's point of view. Write $x_\lambda = A_\lambda^{\#}\,b$ and assume that $\mathrm{Cov}(b) = \eta^2 I$; choose the λ that solves

$$\|A\,x_\lambda - b\|_2 = \left(\|e\|_2^2 - \eta^2\,\mathrm{trace}(A\,A_\lambda^{\#})\right)^{1/2} .$$

Note that the right-hand side now depends on λ.

Both versions of the DP are very sensitive to the estimate δ_e.

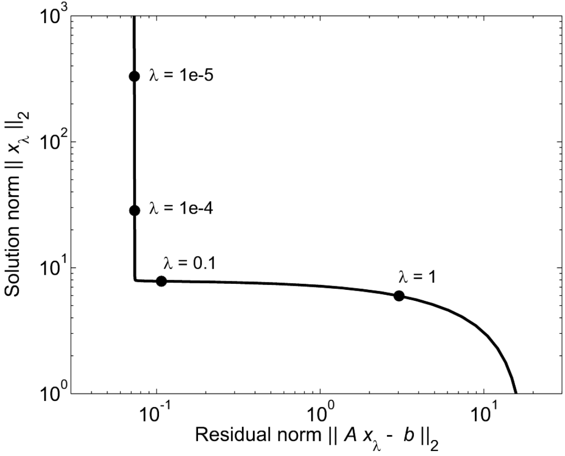
Illustration of the Discrepancy Principle



The choice $\|A\,x_k - b\|_2 \approx (n-k_\eta)^{1/2}\eta$ leads to a too large value of the truncation parameter k, while the more conservative choice $\|A\,x_k - b\|_2 \approx \|e\|_2$ leads to a better value of k.

The L-Curve for Tikhonov Regularization

Recall that the L-curve is a log-log-plot of the solution norm versus the residual norm, with λ as the parameter.



Parameter-Choice and the L-Curve

Recall that the L-curve basically consists of two parts.

- A “flat” part where the regularization errors dominates.
- A “steep” part where the perturbation error dominates.

The optimal regularization parameter (in the pragmatic sense) must lie somewhere near the L-curve’s corner.

The component b^{exact} dominates when λ is large:

$$\|x_\lambda\|_2 \approx \|x^{\text{exact}}\|_2 \text{ (constant)}$$

$$\|b - A\,x_\lambda\|_2 \text{ increases with } \lambda.$$

The error e dominates when λ is small:

$$\|x_\lambda\|_2 \text{ increases with } \lambda^{-1}$$

$$\|b - A\,x_\lambda\|_2 \approx \|e\|_2 \text{ (constant.)}$$

The L-Curve Criterion

The flat and the steep parts of the L-curve represent solutions that are dominated by regularization errors and perturbation errors.

- The balance between these two errors must occur near the L-curve's corner.
- The two parts – and the corner – are emphasized in log-log scale.
- Log-log scale is insensitive to scalings of A and b.

An *operational* definition of the corner is required.

Write the L-curve as

$$(\log \|A\, x_\lambda - b\|_2\, ,\ \log \|x_\lambda\|_2)$$

and seek the point with maximum curvature.

The Curvature of the L-Curve

We want to derive an analytical expression for the L-curve's curvature ζ in log-log scale. Define

$$\xi = \|x_\lambda\|_2^2 \, , \qquad \rho = \|A\, x_\lambda - b\|_2^2$$

and

$$\hat{\xi} = \log \xi \, , \qquad \hat{\rho} = \log \rho \, .$$

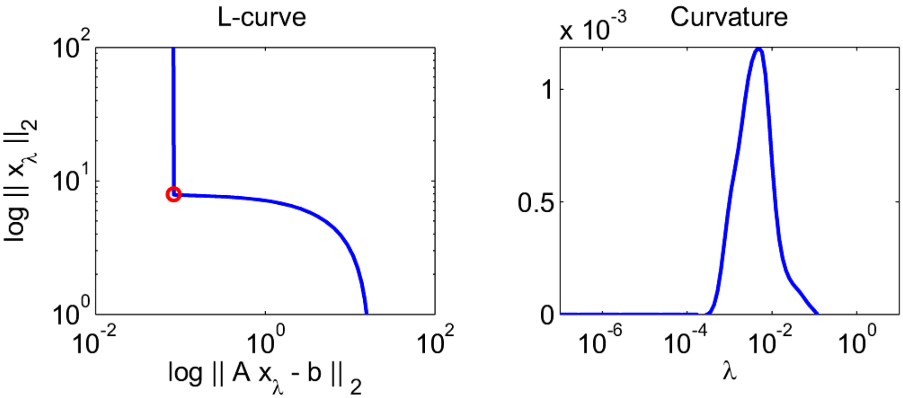
Then the curvature is given by

$$\hat{c}_\lambda = 2 \, \frac{\hat{\rho}' \hat{\xi}'' - \hat{\rho}'' \hat{\xi}'}{((\hat{\rho}')^2 + (\hat{\xi}')^2)^{3/2}} \, ,$$

where a prime denotes differentiation with respect to λ.

This can be used to define the "corner" of the L-curve as the point with maximum curvature.

Illustration



An L-curve and the corresponding curvature $\hat{c}_\lambda$ as a function of λ. The corner, which corresponds to the point with maximum curvature, is marked by the red circle; it occurs for $\lambda_\mathrm{L} = 4.86 \cdot 10^{-3}$.

A More Practical Formula

The first derivatives of $\hat{\xi}$ and $\hat{\rho}$ satisfy

$$\hat{\xi}' = \xi'/\xi \ , \qquad \hat{\rho}' = \rho'/\rho, \qquad \rho' = -\lambda^2 \xi' \ .$$

The second derivatives satisfy

$$\hat{\xi}'' = \frac{\xi'' \xi - (\xi')^2}{\xi^2}, \qquad \hat{\rho}'' = \frac{\rho'' \rho - (\rho')^2}{\rho^2},$$

as they are interrelated by

$$\rho'' = \frac{d}{d\lambda} \left(-\lambda^2 \xi' \right) = -2\, \lambda\, \xi' - \lambda^2 \xi'' \ .$$

When all this is inserted into the equation for $\hat{c}_\lambda$, we get

$$\hat{c}_\lambda = 2\, \frac{\xi\, \rho}{\xi'}\, \frac{\lambda^2 \xi' \rho + 2\, \lambda\, \xi\, \rho + \lambda^4 \xi\, \xi'}{(\lambda^2 \xi^2 + \rho^2)^{3/2}} \ .$$

Efficient Computation of the Curvature

The quantities ξ and ρ readily available.

Straightforward to show that

$$\xi' = \frac{4}{\lambda} x_\lambda^T z_\lambda$$

where z_λ is given by

$$z_\lambda = \left(A^T A + \lambda^2 I\right)^{-1} A^T (A\, x_\lambda - b) \ ,$$

i.e., z_λ is the solution to the problem

$$\min \left\| \begin{pmatrix} A \\ \lambda\, I \end{pmatrix} z - \begin{pmatrix} A\, x_\lambda - b \\ 0 \end{pmatrix} \right\|_2 \ .$$

This can be used to compute z_λ efficiently, when we already have a factorization of the coefficient matrix.

Discrete L-Curves

The L-curve may be discrete – corresponding to a discrete regularization parameter k. May have local, fine-grained "corners" (that do not appear with a continuous parameter).

Two-step approach (older versions of Reg. Tools):

1. Perform a local smoothing of the L-curve points.
2. Use the smoothed points as control points for a cubic spline curve, compute its "corner," and return the original point closest to this corner.

Another two-step approach (current version of Reg. Tools):

1. Prune the discrete L-curve for small local corners.
2. Use the remaining points to determine the largest angle between neighbor points.

The Prediction Error

A different kind of goal: find the value of λ or k such that $A\,x_\lambda$ or $A\,x_k$ predicts the *exact* data $b^{\mathrm{exact}} = A\,x^{\mathrm{exact}}$ as well as possible.

We split the analysis in two cases, depending on k:

$$\begin{aligned} k < k_\eta : &\qquad \|A\,x_k - b^{\mathrm{exact}}\|_2^2 \approx k\,\eta^2 + \sum_{i=k+1}^{k_\eta} (u_i^T b^{\mathrm{exact}})^2 \\ k > k_\eta : &\qquad \|A\,x_k - b^{\mathrm{exact}}\|_2^2 \approx k\,\eta^2 . \end{aligned}$$

For $k < k_\eta$ the norm of the prediction error decreases with k.

For $k > k_\eta$ the norm increases with k.

The minimum arises near the transition, i.e., for $k \approx k_\eta$. Hence it makes good sense to search for the regularization parameter that minimizes the prediction error. But b^{exact} is unknown ...

(Ordinary) Cross-Validation

Leave-one-out approach:
skip ith element b_i and predict this element.

$$\begin{aligned} A^{(i)} &= A([1\colon i-1, i+1\colon m], :\,) \\ b^{(i)} &= b([1\colon i-1, i+1\colon m]) \\ x_\lambda^{(i)} &= \big(A^{(i)}\big)_\lambda^{\#} b^{(i)} \quad \text{(Tikh. sol. to reduced problem)} \\ b_i^{\text{predict}} &= A(i, :\,)\, x_\lambda^{(i)} \quad \text{(prediction of ``missing'' element.)} \end{aligned}$$

The optimal λ minimizes the quantity

$$\mathcal{C}(\lambda) = \sum_{i=1}^{m} \big(b_i - b_i^{\text{predict}}\big)^2 \ .$$

But λ is hard to compute, and depends on the ordering of the data.

Generalized Cross-Validation

Want a scheme for which λ is independent of any orthogonal transformation of b (incl. a permutation of the elements).

Minimize the GCV function

$$G(\lambda) = \frac{\| A\, x_\lambda - b \|_2^2}{\mathrm{trace}(I_m - A\, A_\lambda^{\#})^2}$$

where

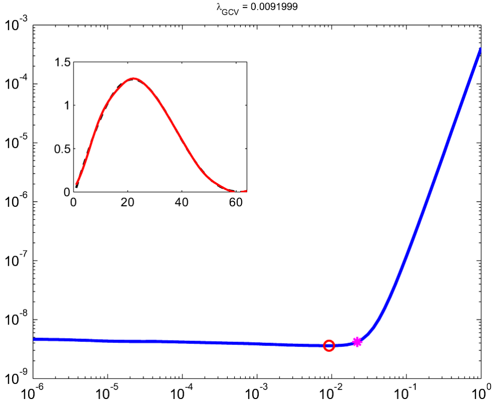
$$\mathrm{trace}(I_m - A\, A_\lambda^{\#}) = m - \sum_{i=1}^{n} \varphi_i^{[\lambda]} \ .$$

Easy to compute the trace term when the SVD is available.

For TSVD the trace term is particularly simple:

$$m - \sum_{i=1}^{n} \varphi_i^{[\lambda]} = m - k \ .$$

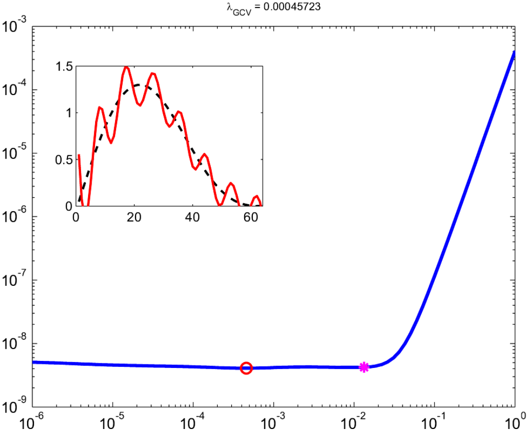
The GCV Function



The GCV function $G(\lambda)$ for Tikhonov regularization; the red circle shows the parameter λ_{GCV} as the minimum of the GCV function, while the cross indicates the location of the optimal parameter.

Occasional Failure

Occasional failure leading to a too small λ; more pronounced for correlated noise.



Extracting Signal in Noise

An observation about the residual vector.

- If λ is too large, not all information in b has not been extracted.
- If λ is too small, only noise is left in the residual.

Choose the λ for which the residual vector changes character from "signal" to "noise."

Our tool: the **normalized cumulative periodogram** (NCP).
Let $p_\lambda \in \mathbb{R}^{n/2}$ be the residual's power spectrum, with elements

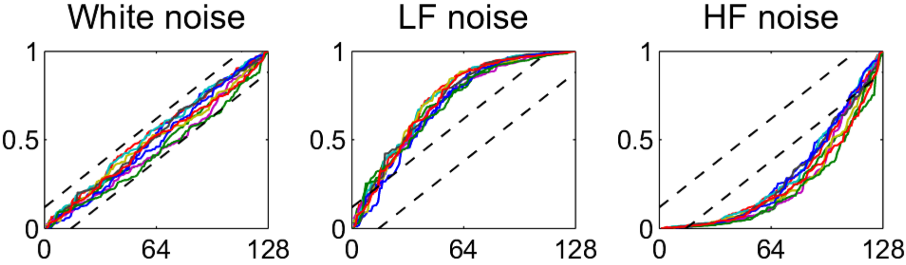
$$(p_\lambda)_k = |\mathrm{dft}(A\, x_\lambda - b)_k|^2, \qquad k = 1, 2, \ldots, n/2 \ .$$

Then the vector $c(r_\lambda) \in \mathbb{R}^{n/2-1}$ with elements

$$c(r_\lambda) = \frac{\|p_\lambda(2\colon k{+}1)\|_1}{\|p_\lambda(2\colon n/2)\|_1}, \qquad k = 1, \ldots, n/2 - 1$$

is the NCP for the residual vector.

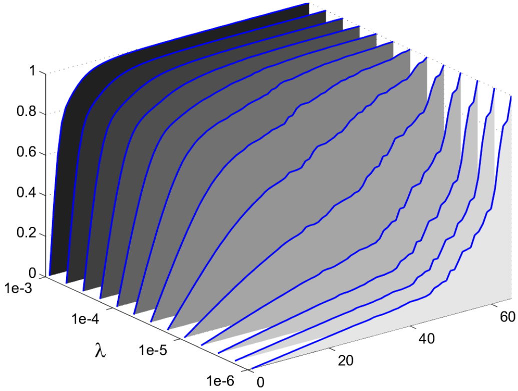
NCP Analysis



Left to right: 10 instances of white-noise residuals, 10 instances of residuals dominated by low-frequency components, and 10 instances of residuals dominated by high-frequency components.

The dashed lines show the Kolmogorov-Smirnoff limits $\pm 1.35\, q^{-1/2} \approx \pm 0.12$ for a 5% significance level, with $q = n/2 - 1$.

The Transition of the NCPs



Plots of NCPs for various regularization parameters λ, for the test problem `deriv2(128,2)` with rel. noise level $\|e\|_2/\|b^{\mathrm{exact}}\|_2 = 10^{-5}$.

Implementation of NCP Criterion

Two ways to implement a pragmatic NCP criterion.

- Adjust the regularization parameter until the NCP lies solely within the K-S limits.
- Choose the regularization parameter for which the NCP is closest to a straight line $c_{\text{white}} = (1/q, 2/q, \ldots, 1)^T$.

The latter is implemented in Regularization Tools.

Summary of Methods (Tikhonov)

Discrepancy principle (`discrep`):

$$\text{Choose } \lambda = \lambda_{\text{DP}} \text{ such that } \|A\, x_\lambda - b\|_2 = \nu_{\text{dp}} \|e\|_2 .$$

L-curve criterion (`l_curve`):

$$\text{Choose } \lambda = \lambda_{\text{L}} \text{ such that the curvature } \hat{c}_\lambda \text{ is maximum.}$$

GCV criterion (`gcv`):

$$\text{Choose } \lambda = \lambda_{\text{GCV}} \text{ as the minimizer of } G(\lambda) = \frac{\|A\, x_\lambda - b\|_2^2}{\left(m - \sum_{i=1}^{n} \varphi_i^{[\lambda]} \right)^2} .$$

NCP criterion (`ncp`):

$$\text{Choose } \lambda = \lambda_{\text{NCP}} \text{ as the minimizer of } d(\lambda) = \|c(r_\lambda) - c_{\text{white}}\|_2 .$$

Comparison of Methods

To evaluate the performance of the four methods, we need the optimal regularization parameter λ_{opt}:

$$\lambda_{\text{opt}} = \text{argmin}_{\lambda} \| x^{\text{exact}} - x_{\lambda} \|_2 .$$

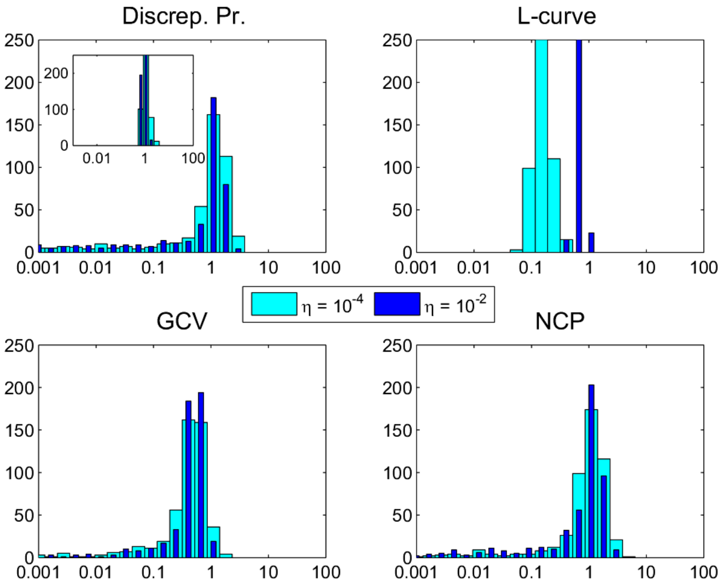
This allows us to compute the four ratios

$$R_{\text{DP}} = \frac{\lambda_{\text{DP}}}{\lambda_{\text{opt}}}, \qquad R_{\text{L}} = \frac{\lambda_{\text{L}}}{\lambda_{\text{opt}}}, \qquad R_{\text{GCV}} = \frac{\lambda_{\text{GCV}}}{\lambda_{\text{opt}}}, \qquad R_{\text{NCP}} = \frac{\lambda_{\text{NCP}}}{\lambda_{\text{opt}}},$$

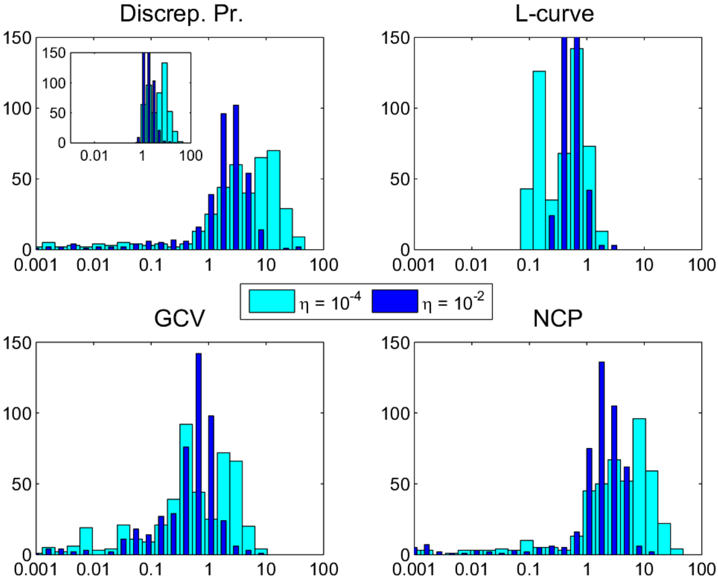
one for each parameter-choice method, and study their distributions via plots of their histograms (in log scale).

The closer these ratios are to one, the better, so a spiked histogram located at one is preferable.

First Example: gravity



Second Example: shaw



Summary

- The *discrepancy principle* is a simple method that seeks to reveal when the residual vector is noise-only. It relies on a good estimate of $\|e\|_2$ which may be difficult to obtain in practise.
- The *L-curve criterion* is based on an intuitive heuristic and seeks to balance the two error components via inspection (manually or automated) of the L-curve. This method fails when the solution is very smooth.
- The *GCV criterion* seeks to minimize the prediction error, and it is often a very robust method – with occasional failure, often leading to ridiculous under-smoothing that reveals itself.
- The *NCP criterion* is a statistically-based method for revealing when the residual vector is noise-only, based on the power spectrum. It can mistake LF noise for signal and thus lead to under-smoothing.