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.

Text book: "Discrete Inverse Problems: Insight and Algorithms"

Read sections 5.1, 5.2, 5.3, 5.4, 5.5, 5.6.

Once Again: Tikhonov Regularization

From now on, we consider a rectangular matrix A of dimensions $m \times n$.

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

Recall that the Tikhonov solution x_λ solves the problem

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

and that it is formally given by

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

where

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

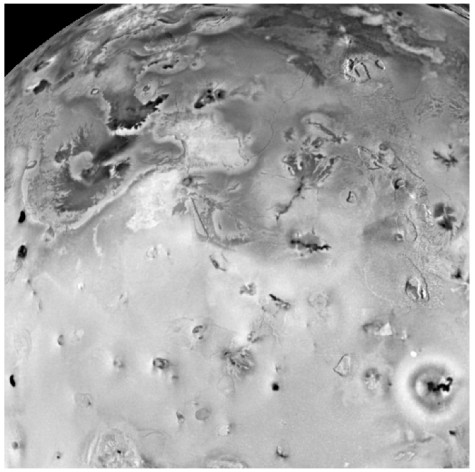
Our noise model

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

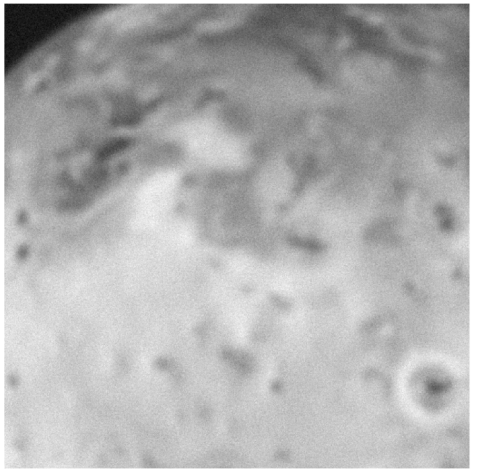
where $b^{\text{exact}} = A x^{\text{exact}}$ and e is the error.

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

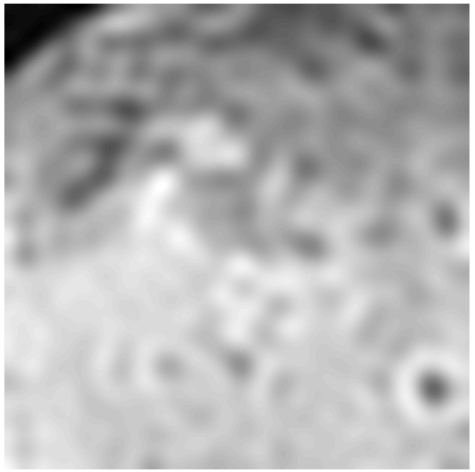
Exact



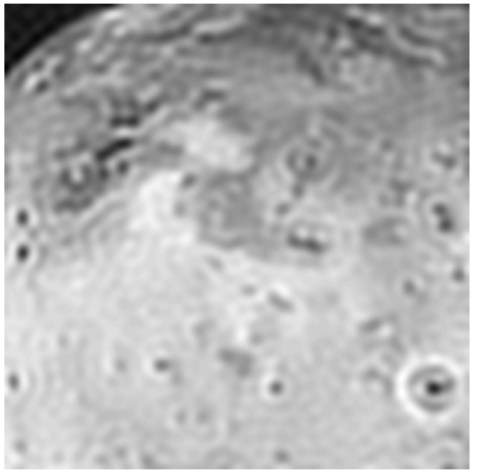
Blurred



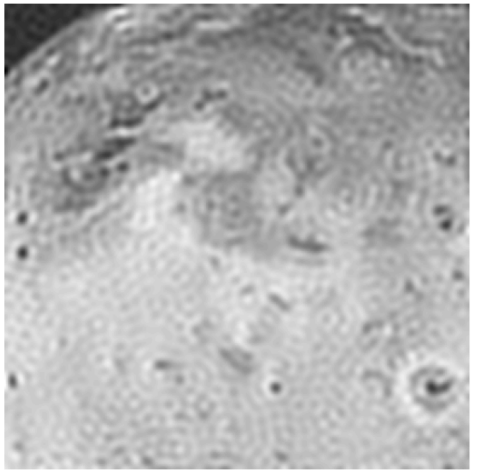
λ too large



$\lambda \approx$ ok



λ too small



Perspectives on Regularization

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

$$x_\lambda = \arg\min \left\{ \|Ax - b\|_2^2 + \lambda^2 \|x\|_2^2 \right\} \ .$$

Cannot be used for choosing λ.

Forward error: balance regularization errors and perturbation errors.

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

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*

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

which introduces a bias in the solution.

The second error term is the *perturbation error:*

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

which is caused by the errors in the data.

Regularization and Perturbation Errors – TSVD

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

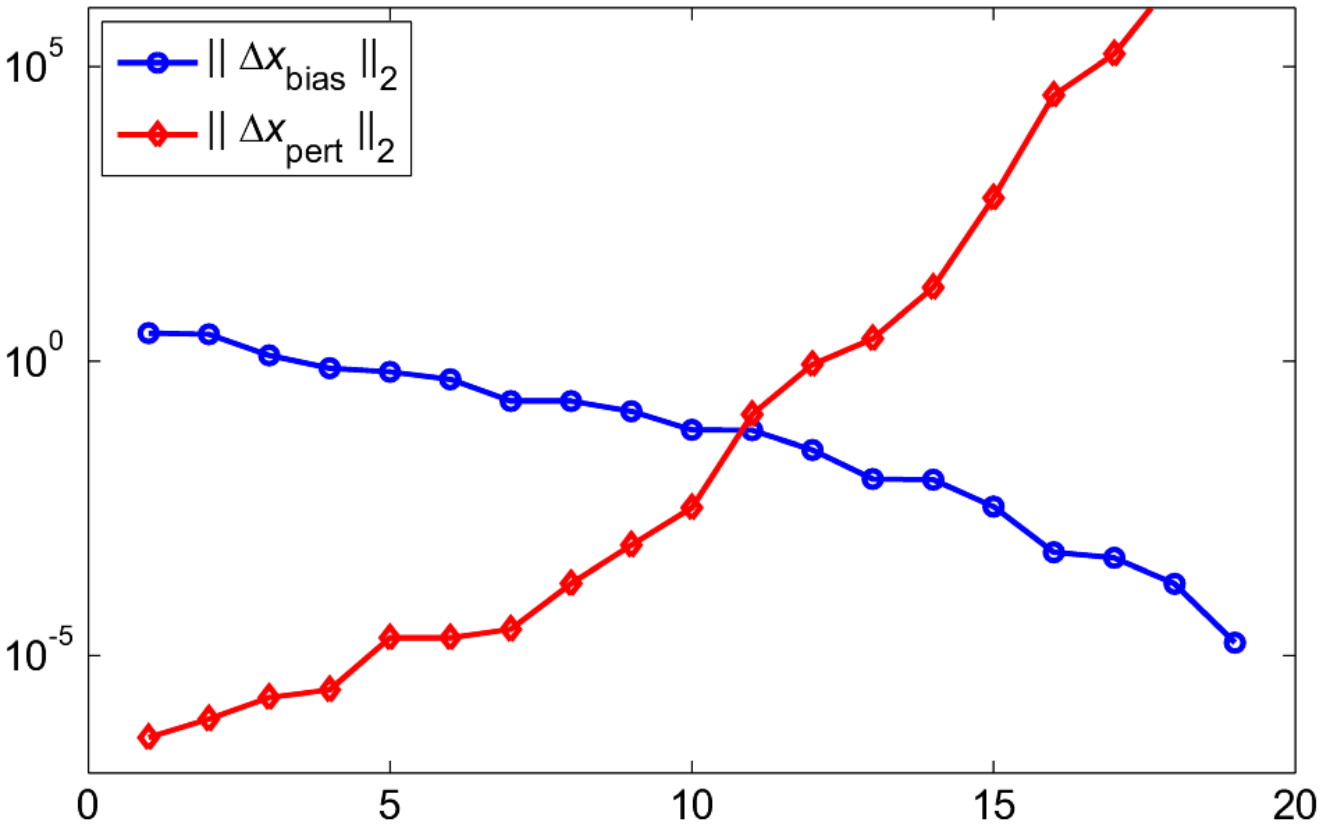
$$\Delta x_{\text{bias}} = \sum_{i=k+1}^{n} (v_i^T x^{\text{exact}})\, v_i, \qquad \Delta x_{\text{pert}} = \sum_{i=1}^{k} \frac{u_i^T e}{\sigma_i}\, v_i.$$

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 Discrepancy Principle

The *discrepancy principle* (DP) seeks to find a regularized solution such that the residual is of the same size as the errors, by solving

$$\|A\,x_\lambda - b\|_2^2 = \tau\,\|e\|_2^2\ ,$$

where τ is some parameter $\tau = O(1)$.

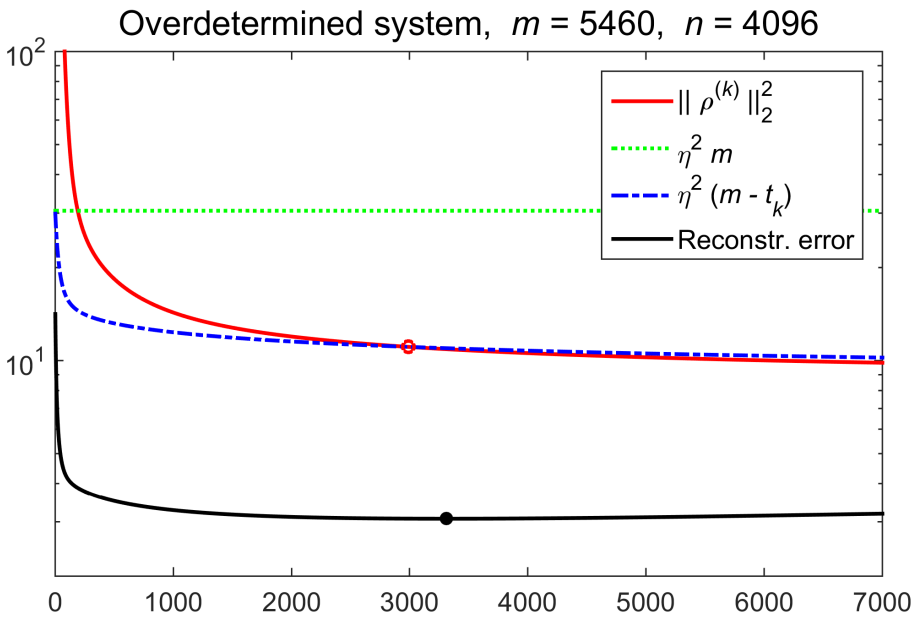
A statistician's point of view. Write $x_\lambda = A_\lambda^\# b$ and assume that $\mathrm{Cov}(b) = \eta^2 I$; choose the λ that solves

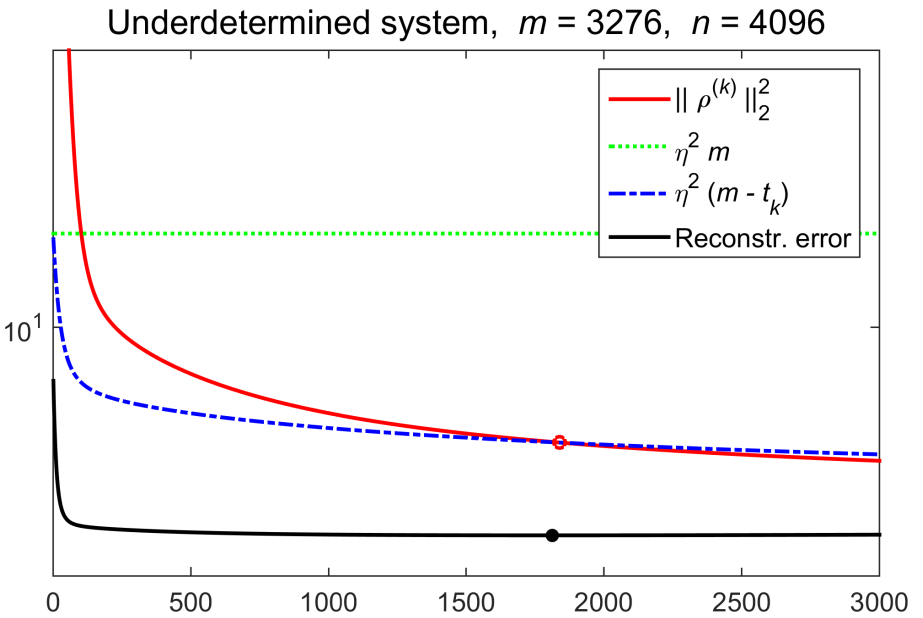
$$\|A\,x_\lambda - b\|_2^2 = \|e\|_2^2 - \eta^2\,\mathrm{trace}(A\,A_\lambda^\#)\ .$$

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

If e is white noise with variance η^2 then $\mathcal{E}(\|e\|_2^2) = n\,\eta^2$, which we can use in the DP. In the alternative approach we can use $\eta^2\big(m - \mathrm{trace}(A\,A_\lambda^\#)\big)$.

Illustration of the Discrepancy Principle



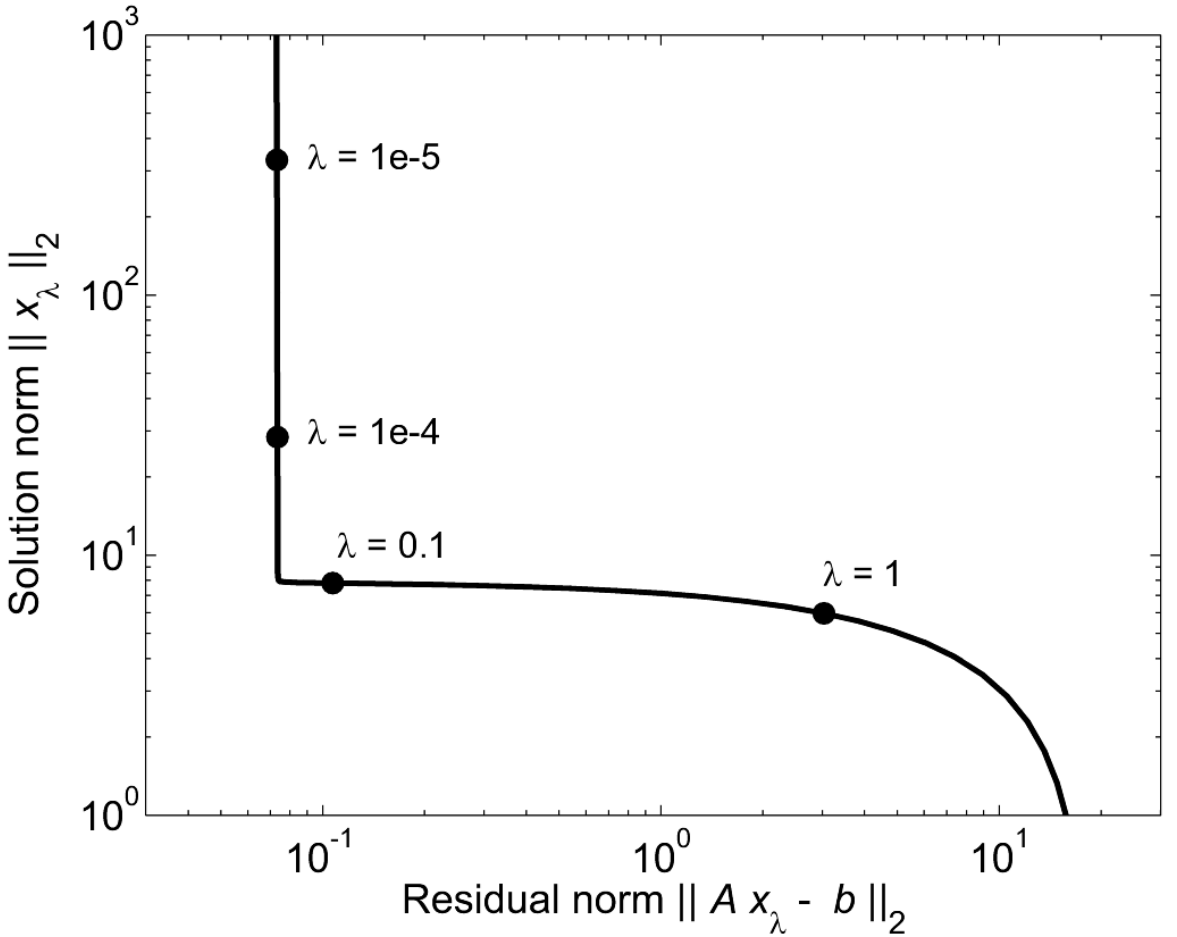


Parallel-beam CT example: 64×64 image; 91 detector pixels; projection angles $3°, 6°, 9°, \ldots, 180°$ (left) and $8°, 16°, 24°, \ldots, 180°$ (right).

Figures show the TSVD reconstruction error $\|\bar{x} - x_k\|_2$ and residual norm $\|b - A x_k\|_2$ versus k, together with threshold $\eta^2 m$ and the function $\eta^2 (m - t_k)$ where t_k = trace term. Plain vanilla DP is not doing well.

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. It is very useful for monitoring the influence of λ.



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 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.

- Intuitively, we expect that 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

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

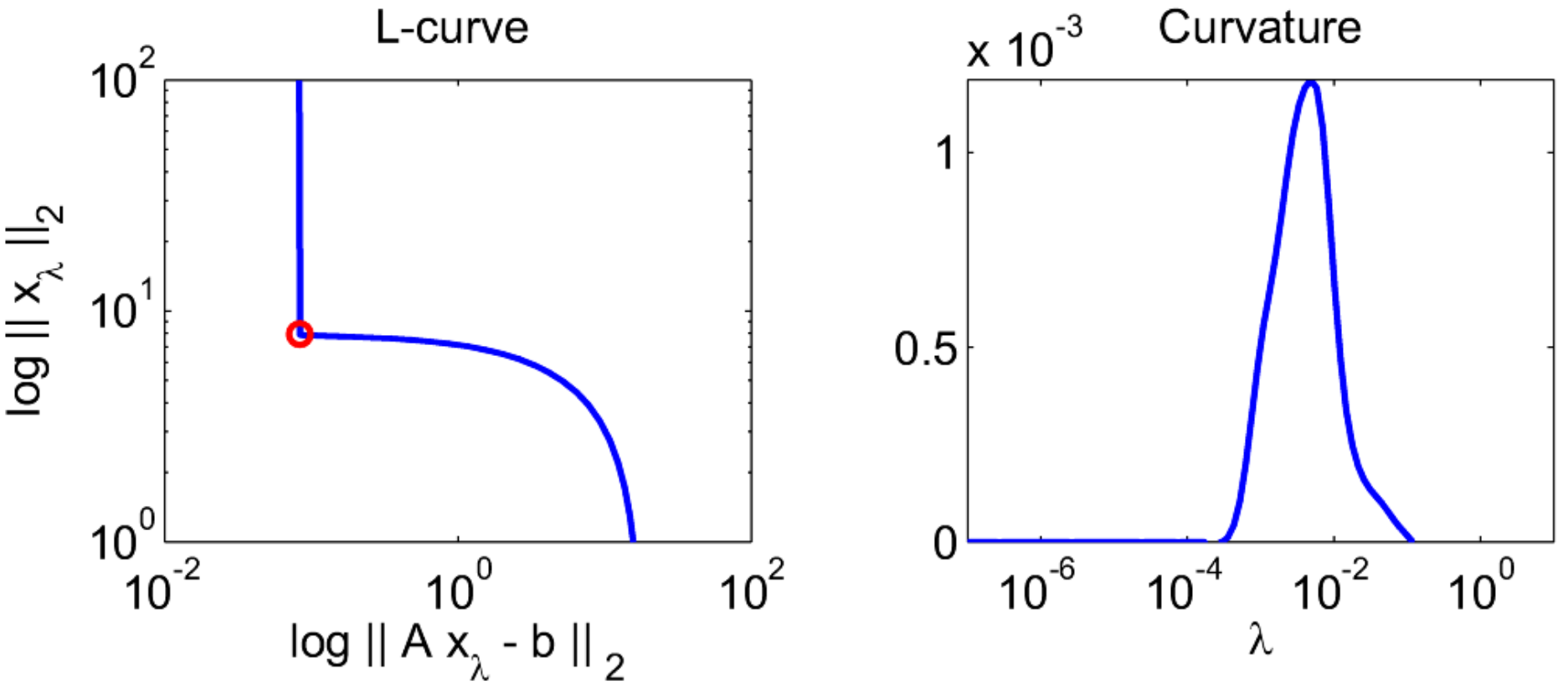
Then the curvature is given by

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

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 $\lambda_L = 4.86 \cdot 10^{-3}$.

The Prediction Error and (Ordinary) Cross-Validation

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 $b^{\text{exact}} = A\,x^{\text{exact}}$ as well as possible.

(Ordinary) cross validation is based on a leave-one-out approach: skip ith element b_i and predict this element.

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

The optimal λ minimizes the quantity

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

But λ is really 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

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

where

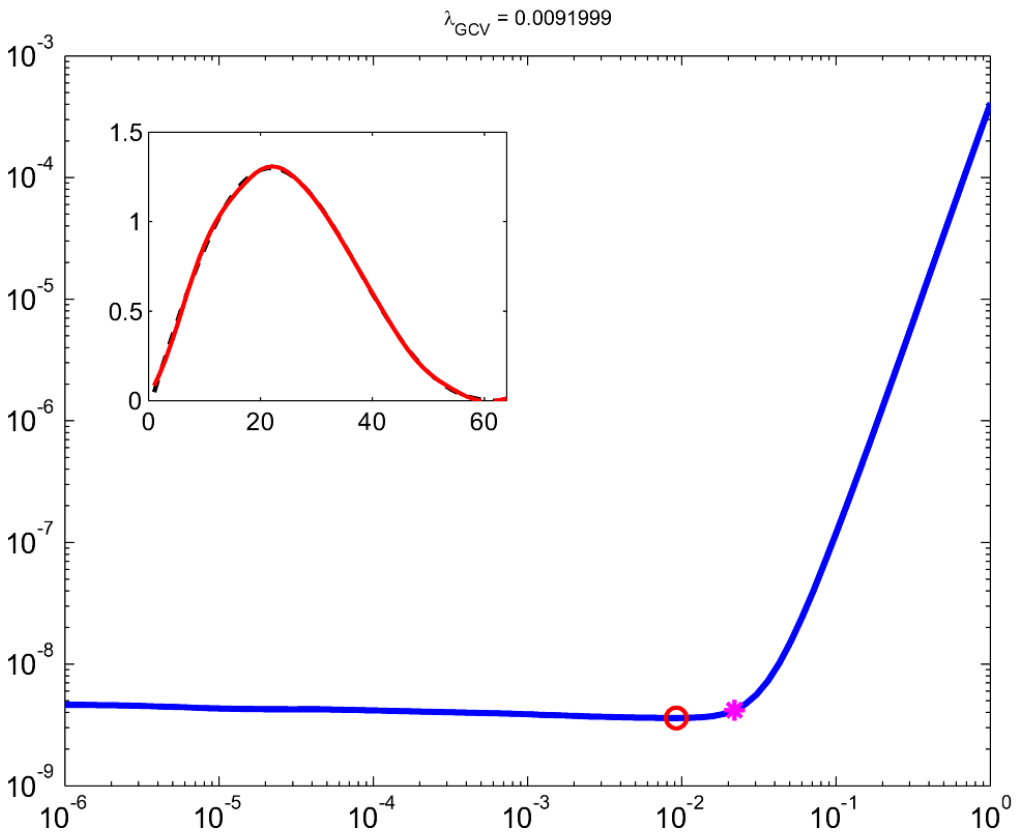
$$\text{trace}(I_m - A\,A_\lambda^{\#}) = m - \sum_{i=1}^{n} \varphi_i^{[\lambda]} \ .$$

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

For TSVD the trace term is particularly simple:

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

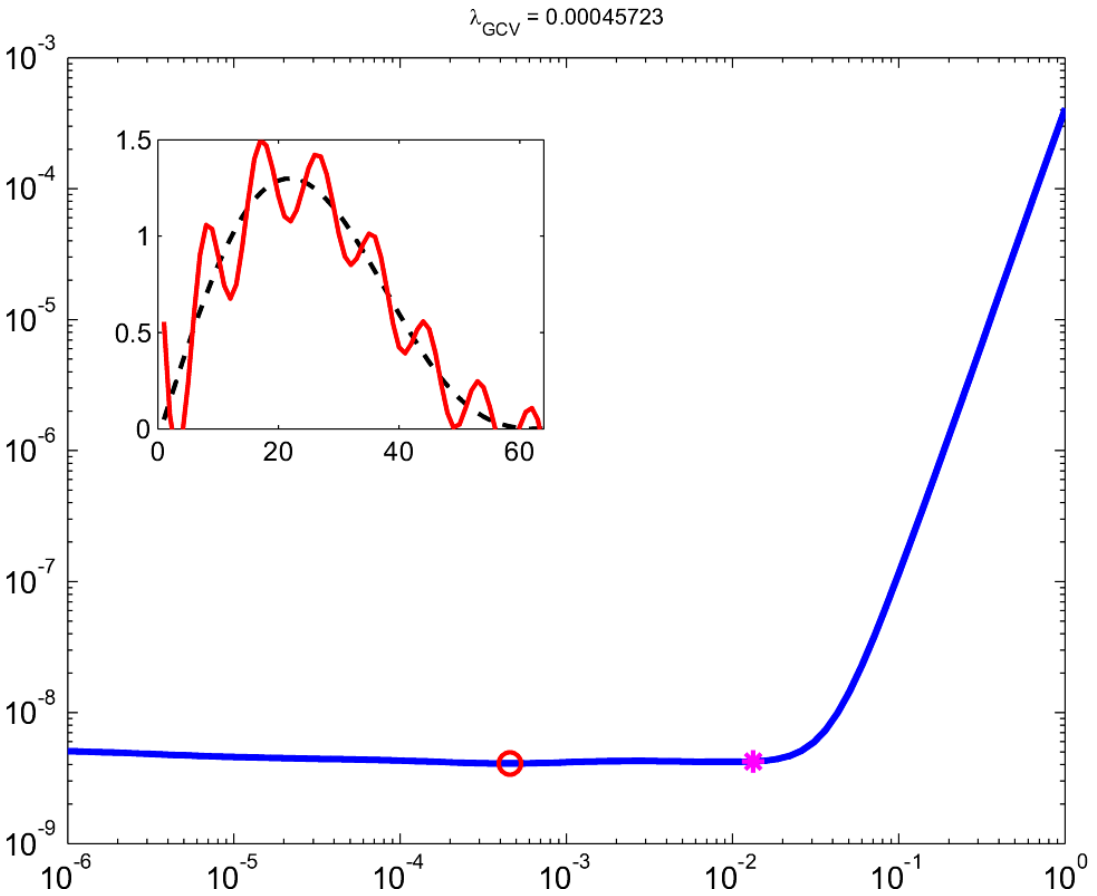
The GCV Function



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

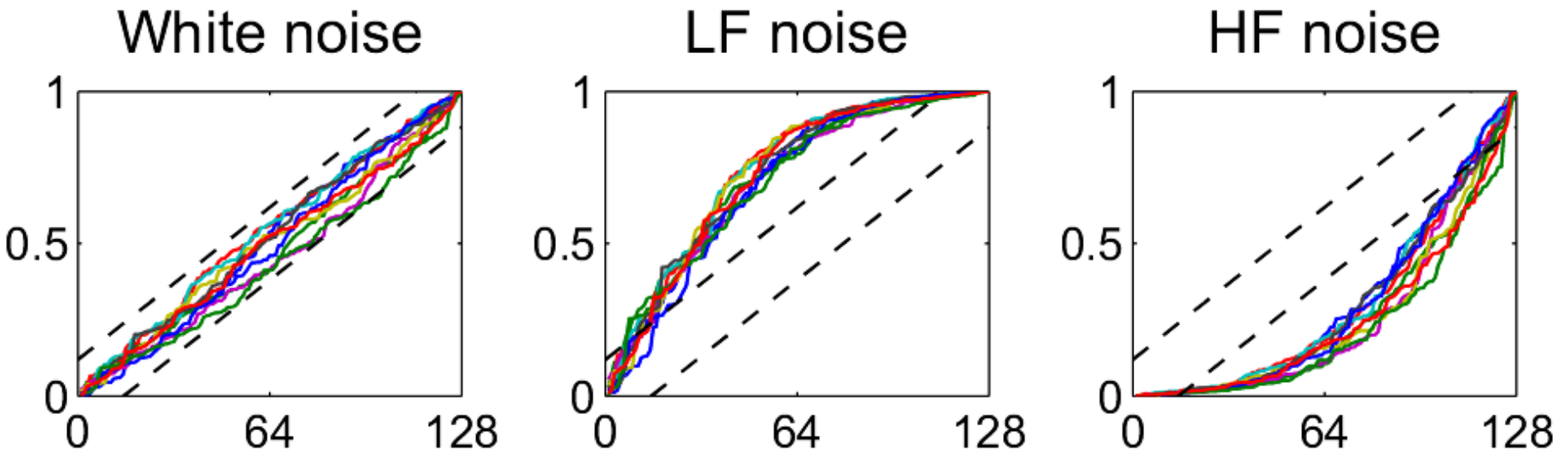
$$(p_\lambda)_k = |\text{dft}(A x_\lambda - b)_k|^2, \qquad k = 1, 2, \ldots, n/2 \ .$$

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.

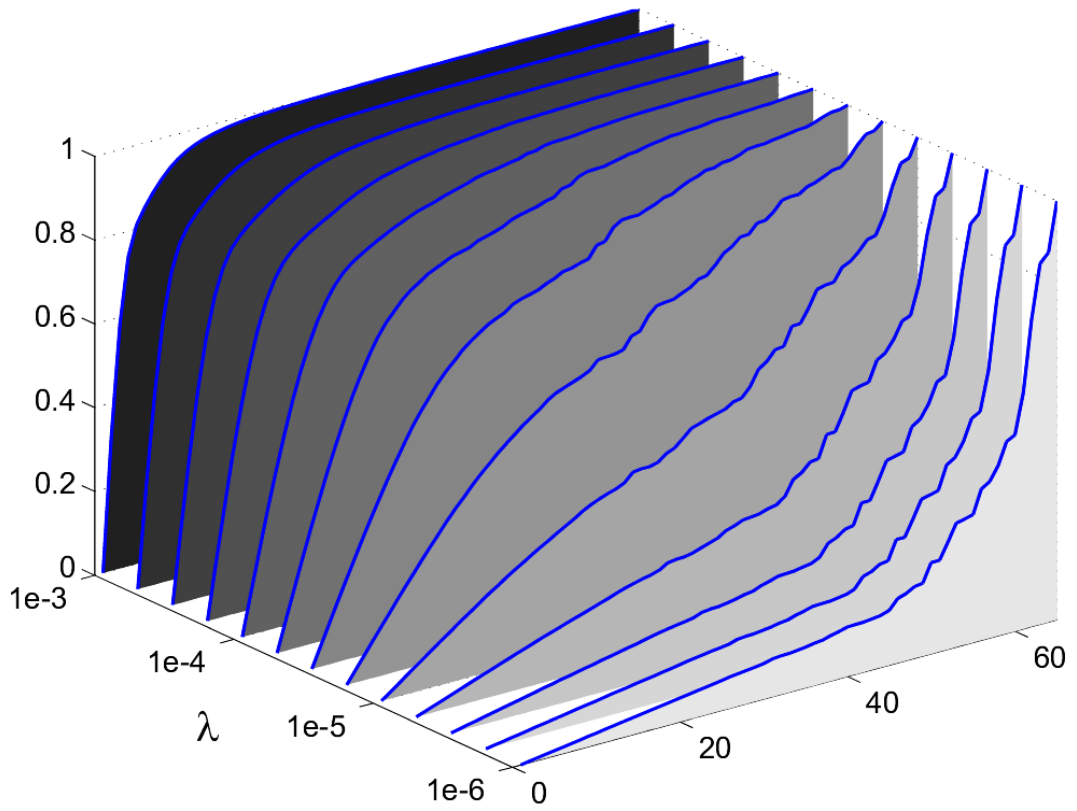
NCP Analysis



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 $\|e\|_2/\|b^{\text{exact}}\|_2 = 10^{-5}$.

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}:

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

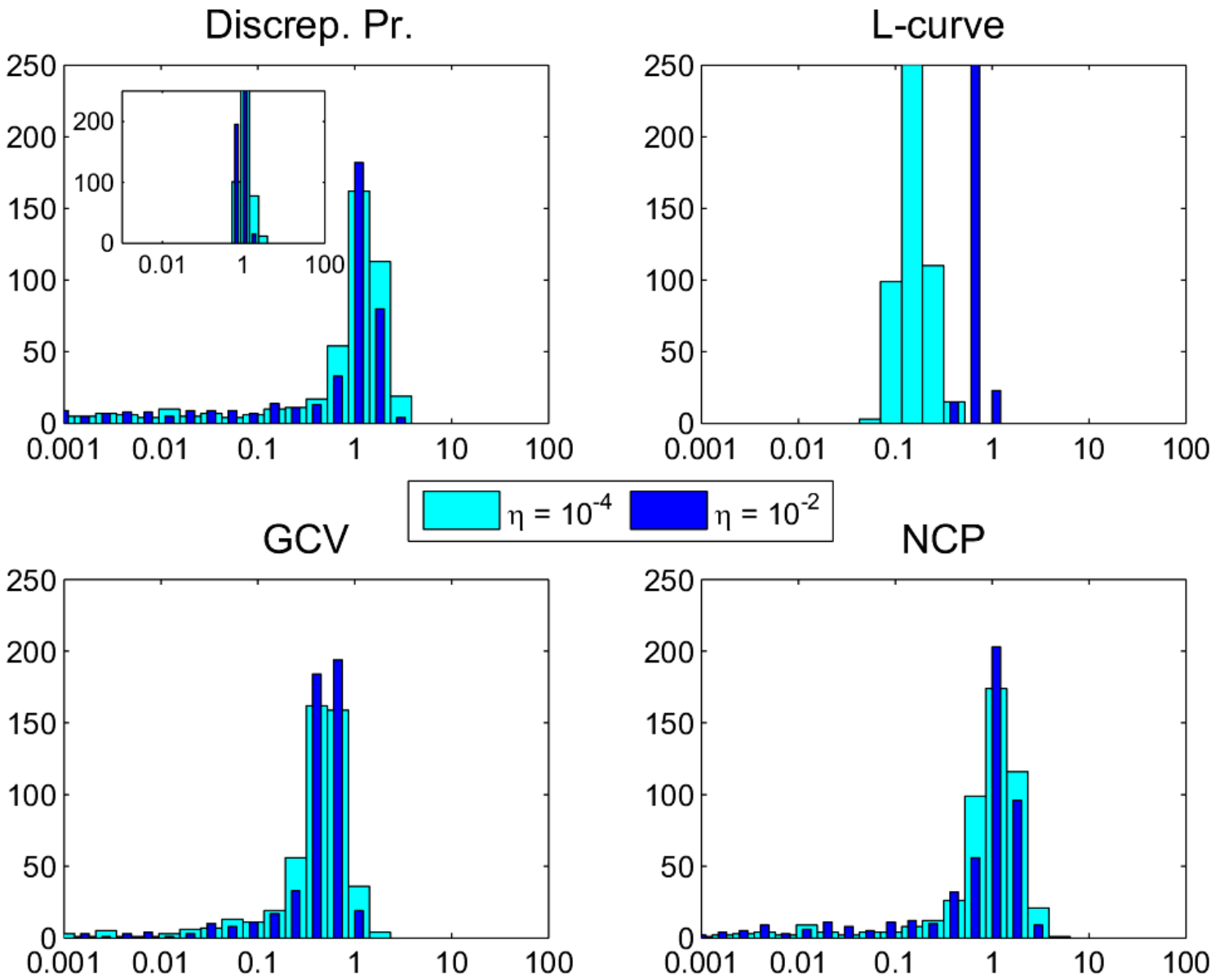
This allows us to compute the four ratios

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

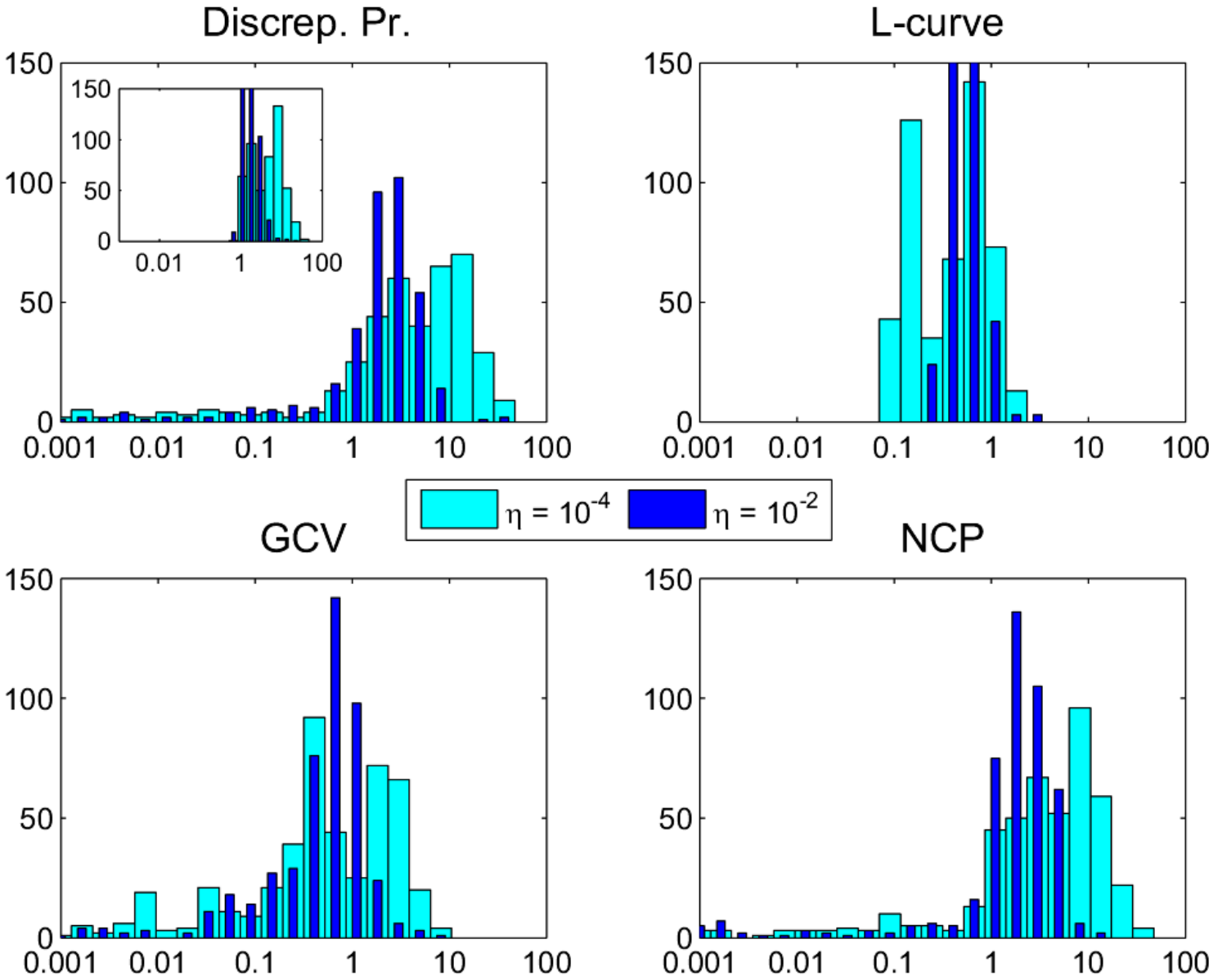
one for each parameter-choice method, and study their distributions via plots of their histograms (in log scale).

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.