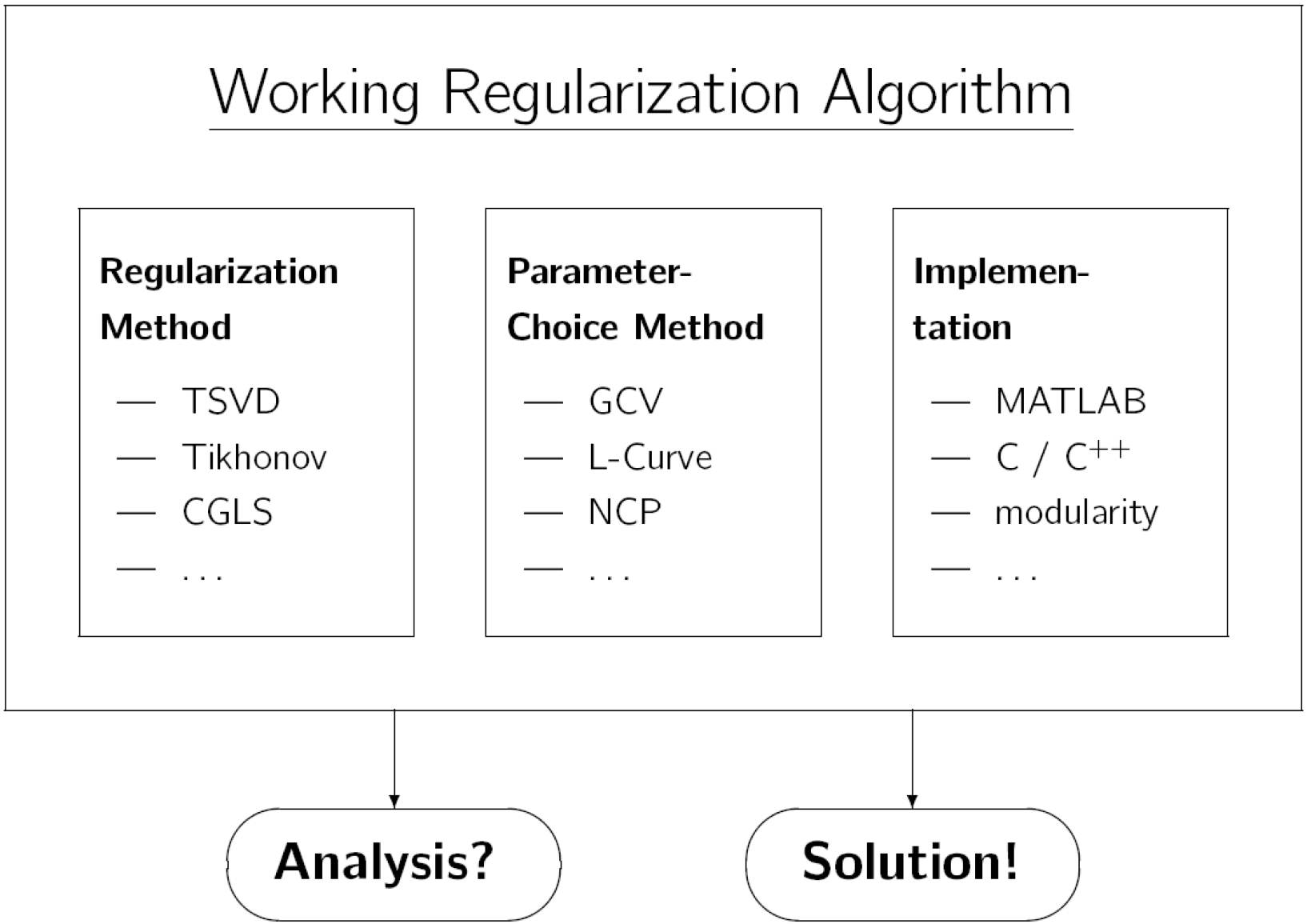
Chapter 7: Regularization Methods at Work



Checklist for designing, implementing, and testing a WRA.

1. Focus on a single application, or a specific/narrow class of applications; no WRA is guaranteed to work for a broad class of problems.
2. When implementing the WRA, focus on modularity and clarity of the computer code; it is guaranteed that you need to go back and modify/expand the software at some point in time.
3. Make sure you understand the performance of the implementation, including computing times, storage requirements, etc.
4. When testing the WRA, make sure to generate test problems that reflect as many aspects as possible of real, measured data.
5. When testing, also make sure to model the noise as realistically as possible, and use realistic noise levels.
6. Be aware of the concept of inverse crime:
    1. As a "proof-of-concept" first use tests that commit inverse crime; if the WRA does not work under such circumstances it can never work.
    2. Next, in order to check the robustness to model errors, test the WRA without committing inverse crime.
7. Carefully evaluate the regularized solutions; consider which characteristics are important, and use the appropriate measure of the

Barcode Reading

The blurring model (approximate, good for illustration):

$$\int_0^1 \exp\left(-\frac{(s-t)^2}{\varsigma^2}\right) f(t)\,dt = g(s), \qquad 0 \le s \le 1,$$

where the s-axis represents the one-dimensional photo conductor inside the scanner.

Special instance of the Fredholm integral equation with the kernel given by

$$K(s,t) = \exp\left(-\frac{(s-t)^2}{\varsigma^2}\right), \qquad s,t \in [0,1].$$

The parameter ς controls the amount of blurring. The larger the ς the narrower the Gaussian peak and thus the less blurring.

Convolution and Deconvolution

A convolution kernel depends only on the difference $s - t$.

The measured signal $g(s)$ is computed as the *convolution*

$$\int_0^1 h(s-t)\, f(t)\, dt = g(s), \qquad 0 \le s \le 1$$

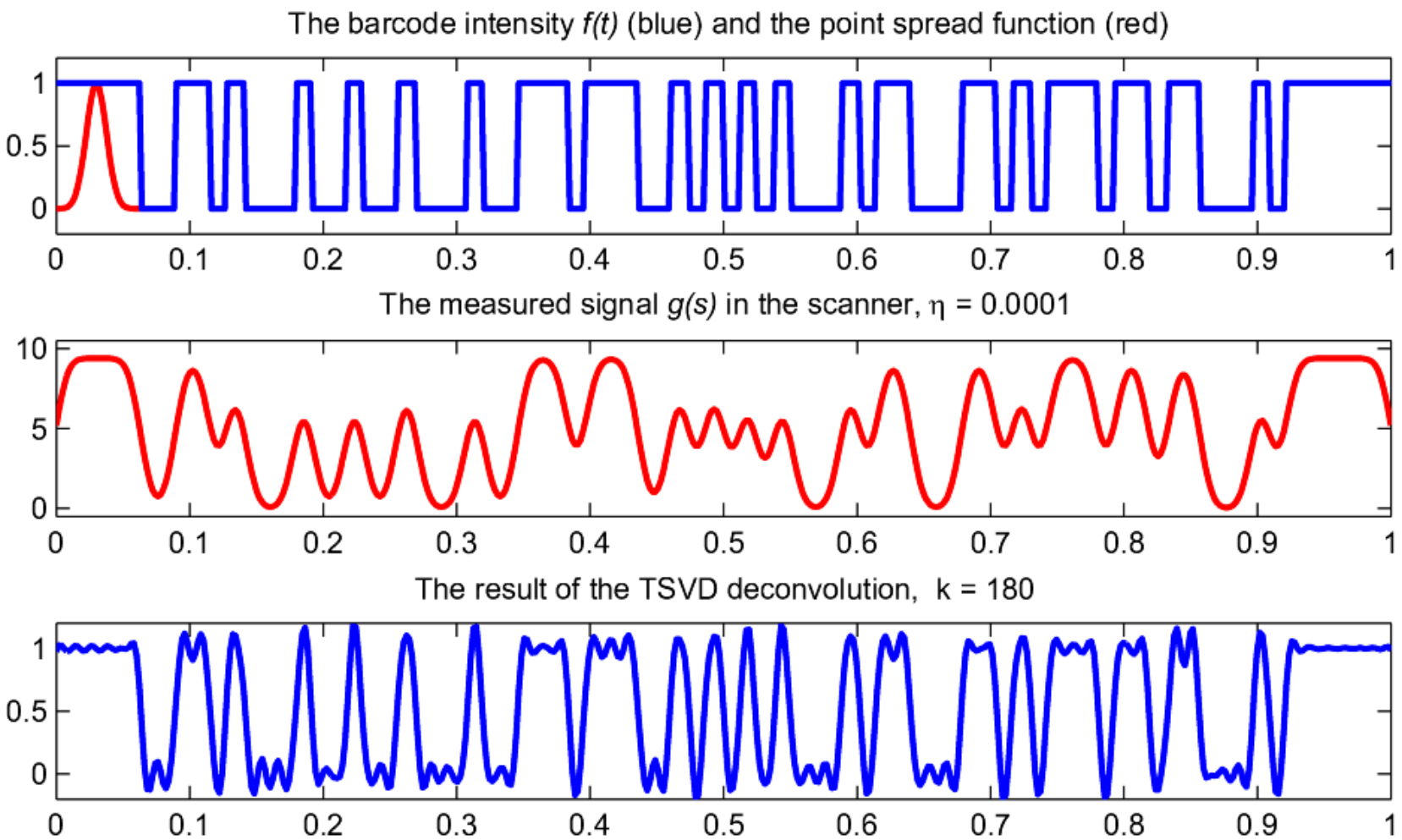
between the exact signal $f(t)$ and the point spread function $h(t)$, in this case given by

$$h(t) = \exp(-t^2/\varsigma^2), \qquad t \in \mathbb{R}.$$

Point spread functions typically peak at $t = 0$ and decay away from this peak such that we can think of $h(t)$ as located near the peak.

For obvious reasons, the process of computing the function $f(t)$, given the measured signal $g(s)$ and the point spread function $h(t)$, is called *deconvolution*.

Barcode Deconvolution



We used TSVD to solve the discretized barcode problem.

Discrete Deconvolution

Advantageous to discretize deconvolution problems by means of the midpoint quadrature rule with n quadrature and collocation points:

$$s_i = (i - \tfrac{1}{2})/n, \qquad t_j = (j - \tfrac{1}{2})/n, \qquad i, j = 1, \ldots, n.$$

Then the elements of the matrix A are:

$$a_{ij} = K(s_i, t_j) = \exp\left(-\frac{(i-j)^2}{\varsigma^2 n^2} \right), \qquad i, j = 1, \ldots, n.$$

If we define $h_{i-j} = a_{ij}$, then we can write the system in the form

$$\sum_{j=1}^{n} h_{i-j}\, x_j = b_i, \qquad i = 1, \ldots, n,$$

where h_{i-j} are allowed to have zero and negative indices. We refer to this problem as *discrete convolution*.

Structured Matrices: Toeplitz

When discrete convolution equations are written in the matrix form $A\,x = b$, then we note that the matrix elements $a_{ij} = h_{i-j}$ are always a function of the difference $i - j$ only, i.e.,

$$a_{ij} = a_{i+\ell,j+\ell} \quad \text{for all relevant integers } i,\ j,\ \text{and } \ell.$$

Such a matrix A, which has constant values along all its diagonals, is called a *Toeplitz matrix*. Example:

$$A = \begin{pmatrix} 5 & 4 & 3 & 2 & 1 & 0 \\ 4 & 5 & 4 & 3 & 2 & 1 \\ 3 & 4 & 5 & 4 & 3 & 2 \\ 2 & 3 & 4 & 5 & 4 & 3 \\ 1 & 2 & 3 & 4 & 5 & 4 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{pmatrix}.$$

Barcode matrix: `toeplitz(-((0:n-1)/(sigma*n)).^2)`.

Structured Matrices: Circulant

In some problems, the function h is periodic with period 1,

$$h(t) = h(t+p), \qquad t \in \mathbb{R}, \quad p = \mathbb{N}.$$

In the discrete case this corresponds to the condition that

$$h_{i-j} = h_{(i-j) \bmod n}, \qquad i, j \in \mathbb{N}.$$

The matrix elements satisfy $a_{ij} = a_{(i+\ell) \bmod n,(j+\ell) \bmod n}$.

A Toeplitz matrix, where the elements also "wrap around" the borders, is called a *circulant matrix*. Example:

$$A = \begin{pmatrix} 5 & 0 & 1 & 2 & 3 & 4 \\ 4 & 5 & 0 & 1 & 2 & 3 \\ 3 & 4 & 5 & 0 & 1 & 2 \\ 2 & 3 & 4 & 5 & 0 & 1 \\ 1 & 2 & 3 & 4 & 5 & 0 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

Computations with Structured Matrices

Discrete deconvolution problems are thus special instances of discrete inverse problems with Toeplitz or circulant matrices.

They can be solved very efficiently by means of methods that take into account the structure of the matrix, such as FFT-based og DCT-based methods of complexity $O(n \log n)$.

- P. C. Hansen, *Deconvolution and regularization with Toeplitz matrices*, Numerical Algorithms, 29 (2002), pp. 323–378.
- P. C. Hansen, J. G. Nagy, and D. P. O'Leary, *Deblurring Images: Matrices, Spectra, and Filtering*, SIAM, Philadelphia, 2006 (130 pages). Korean translation, Jin Publishing Co., 2007.
- P. A. Jansson (Ed.), *Deconvolution of Images and Spectra*, 2. Edition, Academic Press, San Diego, 1997.

Data/Model Mismatch

The integral equation model

$$\int_0^1 K(s,t)\, f(t)\, dt = g(s), \qquad 0 \le s \le 1$$

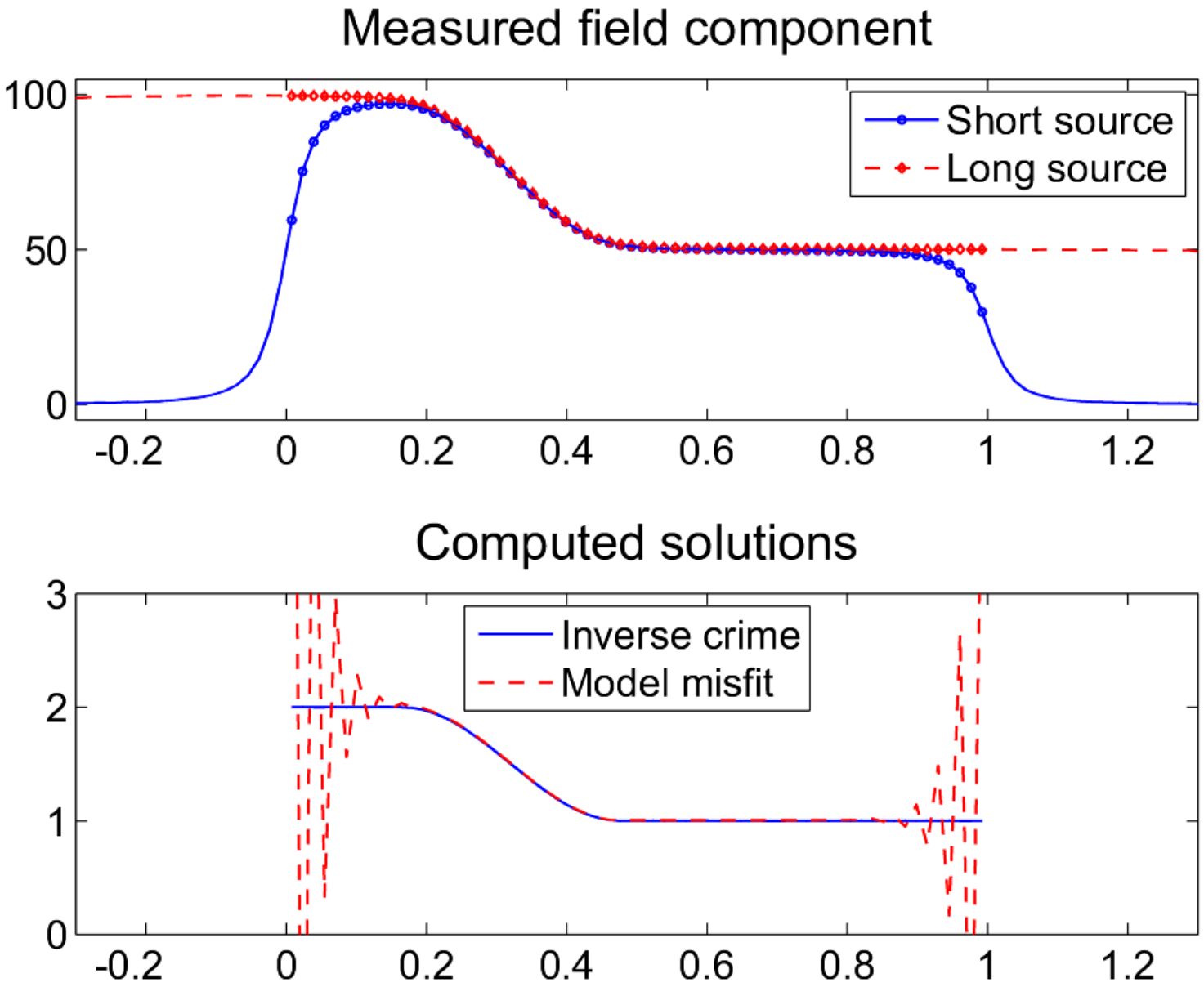
assumes that the data $g(s)$ is solely due to contributions from $t(t)$ in the interval $[0, 1]$.

But what if the observed $g(t)$ actually has contributions from sources $f(t)$ outside the interval [0.1]?

Example next page: $g(s)$ is produced by a source $f(t)$ that is nonzero for $t \in [-0.5, 1.5]$.

If we solve with the wrong assumption that $f(t)$ is zero outside $[0, 1]$, then we get severe artifacts at the boundaries.

Illustration



Inverse Crime

This refers to a perfect agreement between the model and the data in the discretized inverse problem. It arises when the *same model* is used to generate the test data and to compute the reconstruction.

No WRA is useful if it cannot solve such as inverse problem.

But real data are obviously not generated by a computer model, but come from a physical (or some other) phenomenon.

The term "inverse crime" is used to signal the danger of testing a working regularization algorithm using only data generated with the same model as that used in the inversion.

A rigorous testing of a working regularization algorithm must reveal how sensitive the algorithm is to model/data mismatch.

Boundary Conditions

A convenient mechanism for avoiding the severe effects of model/data mismatch.

Must be considered whenever the measured data come from a function $g(s)$ that can be influenced by values of the source function $f(t)$ outside the integration interval.

Not needed if $f(t)$ is zero outside the integration interval.

Here we consider *reflexive boundary conditions:*

the underlying assumption is that the behavior of the function $f(t)$ outside $[0, 1]$ is a "reflection" of its behavior inside the interval.

More BC

Define the extended function f_{BC} in the interval $[-1, 2]$ as follows:

$$f_{\mathrm{BC}}(t) = \begin{cases} f(-t), & -1 < t < 0 \\ f(t), & 0 \le t \le 1 \\ f(2-t), & 1 < t < 2 \end{cases}$$

and consider data $g_{\mathrm{BC}}(s)$ generated by the extended model

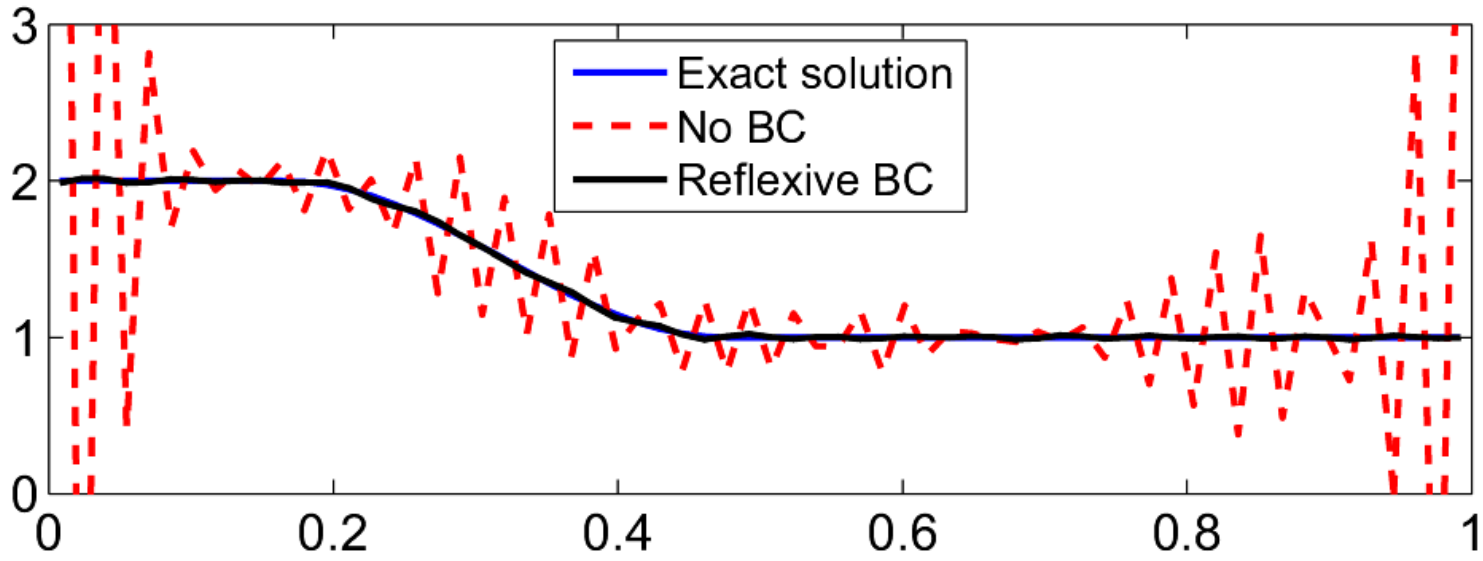
$$\begin{aligned} g_{\mathrm{BC}}(s) &= \int_{-1}^{2} K(s,t)\, f_{\mathrm{BC}}(t)\, dt \\ &= \int_{-1}^{0} K(s,t)\, f_{\mathrm{BC}}(t)\, dt + \int_{0}^{1} K(s,t)\, f_{\mathrm{BC}}(t)\, dt + \int_{1}^{2} K(s,t)\, f_{\mathrm{BC}}(t)\, dt \\ &= \int_{0}^{1} K(s,-t)\, f(t)\, dt + \int_{0}^{1} K(s,t)\, f(t)\, dt + \int_{0}^{1} K(s,2-t)\, f(t)\, dt \\ &= \int_{0}^{1} \big(K(s,-t) + K(s,t) + K(s,2-t)\big)\, f(t)\, dt. \end{aligned}$$

Reflexive BC correspond to working with a modified integral equation where $K(s,t)$ is replaced by the extended kernel

$$K_{\mathrm{BC}}(s,t) = K(s,-t) + K(s,t) + K(s,2-t).$$

Use of Reflexive BC

The use of reflexive BC improves the solution:



When we discretize the modified integral equation, then we obtain a modified coefficient matrix A_{BC} that is the sum of the original matrix A plus correction terms coming from discretization of the terms $K(s,-t)$ and $K(s,2-t)$ in $K_{\mathrm{BC}}(s,t)$.

More Matrix Structure

Incorporation of the reflexive BC in the barcode problem means that the must add correction terms A^{l} and A^{r} for the left and right boundary conditions, respectively.

The elements of these two matrices are, for $i, j = 1, \ldots, n$,

$$a^{\mathsf{l}}_{ij} = K(s_i, -t_j) = \exp\left(-\frac{(s_i + t_j)^2}{\varsigma^2}\right) = \exp\left(-\frac{(i + j - 1)^2}{\varsigma^2\, n^2}\right)$$

and

$$a^{\mathsf{r}}_{ij} = K(s_i, 2 - t_j) = \exp\left(-\frac{(s_i + t_j - 2)^2}{\varsigma^2}\right) = \exp\left(-\frac{(i + j - 2n - 1)^2}{\varsigma^2\, n^2}\right)$$

(again we ignored the factor n^{-1}). Note that a_{ij} is a function of $i + j$, such that $a_{ij} = a_{i+\ell, j-\ell}$ for all relevant integers i, j, and ℓ.

A matrix with this structure, where the elements are constant along all the anti-diagonals, is called a *Hankel matrix*.

Hankel Structure

The elements of the two Hankel matrices A^{l} and A^{r} can be expressed in terms of the elements of the Toeplitz matrix A:

$$a^{\mathrm{l}}_{i1} = \exp\left(-\frac{((i+1)-1)^2}{\varsigma^2\, n^2}\right) = a_{i+1,1}, \qquad i = 1, 2, \ldots, n-1,$$

$$a^{\mathrm{r}}_{nj} = \exp\left(-\frac{(1-(2+n-j))^2}{\varsigma^2\, n^2}\right) = a_{1,2+n-j}, \qquad i = 2, 3, \ldots, n.$$

The remaining elements are undetermined.

But the point spread function $h(t)$ can be considered as zero away from the peak, and thus we can safely define these remaining elements of A^{l} and A^{r} to be zero.

- No overlap between the nonzero elements of A^{l} and A^{r}.
- The matrix $A^{\mathrm{l}} + A^{\mathrm{r}}$ is also Hankel:

```
hankel([A(2:n,1);0],[0,A(1,n:-1:2)]).
```

Symmetric Toeplitz Plus Hankel Structure

The matrix $A_{\mathrm{BC}} = A + A^{\mathrm{l}} + A^{\mathrm{r}}$ is a symmetric Toeplitz plus Hankel matrix, with the Hankel part derived from the Toeplitz part:

$$A = \begin{pmatrix} 3 & 2 & 1 & 0 & 0 \\ 2 & 3 & 2 & 1 & 0 \\ 1 & 2 & 3 & 2 & 1 \\ 0 & 1 & 2 & 3 & 2 \\ 0 & 0 & 1 & 2 & 3 \end{pmatrix} \Rightarrow A^{\mathrm{l}} + A^{\mathrm{r}} = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix}$$

$$A_{\mathrm{BC}} = \begin{pmatrix} 5 & 3 & 1 & 0 & 0 \\ 3 & 3 & 2 & 1 & 0 \\ 1 & 2 & 3 & 2 & 1 \\ 0 & 1 & 2 & 3 & 3 \\ 0 & 0 & 1 & 3 & 5 \end{pmatrix}.$$

We refer to a matrix with this particular structure as a *symmetric Toeplitz-plus-Hankel matrix*, or STH matrix.

STH Matrices and the DCT

Let $W_n = (w_1, \ldots, w_n)$ denote the $n \times n$ "DCT matrix" such that $\text{dct}(x) = W_n^T x$ is the discrete cosine transform (DCT) of x.

Any STH matrix has the eigenvalue decomposition

$$A_{\text{BC}} = W_n \, D \, W_n^T, \qquad D = \text{diag}(d_1, \ldots, d_n),$$

where the eigenvalues d_i are given by

$$d_i = [\text{dct}(A_{\text{BC}}(:,1))]_i \,/\, [\text{dct}(e_1)]_i, \qquad i = 1, \ldots, n,$$

in which $e_1 = (1, 0, \ldots, 0)^T$ and $[\,\cdot\,]_i$ denotes the ith component. The alternative relation

$$d_j = a_{11} + 2 \sum_{i=2}^{n} a_{i,1} \cos((j-1)(k-1)\pi/n),$$

requires $O(n^2)$ flops. Nice but too expensive.

SVD and DCT

The good news is that the SVD of A_{BC} is very closely related to its eigenvalue decomposition, since we have

$$A_{\mathrm{BC}} = (W_n \, \Omega) \, \mathrm{diag}(|d_1|, \ldots, |d_n|) \; W_n^T,$$

in which the diagonal matrix $\Omega = \mathrm{diag}(\mathrm{sign}(d_1), \ldots, \mathrm{sign}(d_n))$ holds the signs of the eigenvalues, such that

$$W_n \, \Omega = (\, \mathrm{sign}(d_1) \, w_1 \, , \, \mathrm{sign}(d_2) \, w_2 \, , \, \ldots \,).$$

Almost the SVD of A_{BC}, except that the singular values in $\mathrm{diag}(|d_1|, \ldots, |d_n|)$ not guaranteed to appear in decreasing order.

Since large singular values are typically associated with low- frequency singular vectors, the eigenvalues d_i tend to appear in order of decreasing magnitude.

Efficient Implementation

Now it follows immediately that the Tikhonov solution to a discrete inverse problem with an STH matrix takes the form

$$x_\lambda = W_n \, \Phi^{[\lambda]} \, D^{-1} W_n^T b$$

and similarly for other regularized solutions, with the Tikhonov filter matrix $\Phi^{[\lambda]}$ replaced by the appropriate filter matrix.

Since the multiplications with W_n and W_n^T represent the DCT and its inverse, it follows immediately that we can compute the Tikhonov regularized solution in $O(n \log n)$ complexity:

```
d = dct(Abc(:,1)) ./ dct(eye(n,1));
q = d ./ (d.^2 + lambda^2);
x_reg = idct( q .* dct(b) );
```

No need to store the full A and A_{BC} matrices, since the first column of each matrix suffice for the computations we need to carry out.
For TSVD, use `q = [1./d(1:k);zeros(n-k,1)]`.

PIXE Depth Profiling

PIXE = Particle-Induced X-ray Emission spectroscopy.

Reconstruction of hidden layers of paint in an image via solving an inverse problem that involves X-ray measurements.

At a specific position, an X-ray source sends a beam at an angle s into the material, and the energy penetrates to a depth of $d\cos(s)$, where d depends on the energy in the beam.

The beam that penetrates the material is partially reflected according to the material parameter $f(t)$, where t denotes the depth into the material, with $0 \leq t \leq d$.

A detector located perpendicular to the surface records the reflected signal $g(s)$, which depends on the incidence angle s and the material parameter $f(t)$.

The PIXE Model

The pixe model:

$$g(s) = \int_0^{d\cos(s)} \exp(-\mu\, t)\, f(t)\, dt, \qquad 0 \le s \le \tfrac{1}{2}\pi.$$

Reconstruction of $f(t)$ is an inverse problem given by the first-kind Fredholm integral equation with kernel

$$K(s,t) = \begin{cases} \exp(-\mu\, t)\ , & 0 \le t \le d\ \cos(s) \\ 0\ , & d\ \cos(s) \le t \le d \end{cases}$$

with s and t in the intervals $s \in [0, \frac{1}{2}\pi]$ and $t \in [0, d]$.

We use $d = 1$, $\mu = 8$, and an artificial solution which models three layers of paint at depths $t = 0.25$, 0.5, and 0.75:

$$f(t) = \exp\big(-30(t-0.25)^2\big) + 0.4\exp\big(-30(t-0.5)^2\big) + 0.5\exp\big(-50(t-0.75)^2\big)$$

Discretization Issues

Discretization by the midpoint quadrature rule with equidistant abscissas t_j, and with equidistant sampling points s_i for $g(s)$:

$$a_{ij} = \begin{cases} n^{-1} \exp(-\mu\, t_j), & t_j \leq \cos(s_i) \\ 0 & \text{else.} \end{cases}$$

This matrix has a large cluster of singular values at or below the machine precision. Unfortunate since the integral operator has only the trivial null space.

Use the variable transformation $t = \sin(\tau)$:

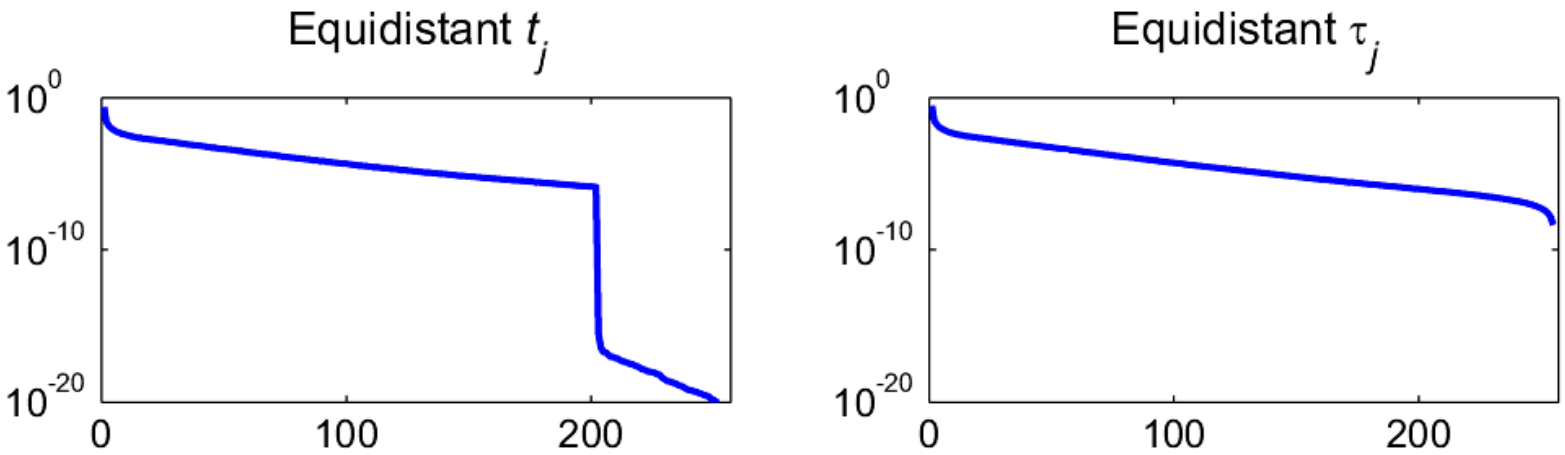
$$g(s) = \int_0^{\arcsin(d\cos(s))} \exp\bigl(-\mu\, \sin(\tau)\bigr)\, \cos(\tau)\, f(\tau)\, d\tau, \quad 0 \leq s \leq \tfrac{1}{2}\pi.$$

Be Careful

Using again the midpoint quadrature rule, this time with equidistant $\tau_j \in [0, \frac{1}{2}\pi]$, we obtain a new coefficient matrix:

$$a_{ij} = \begin{cases} (\pi/2n)\exp\big(-\mu \sin(\tau_j)\big)\cos(\tau_j), & \sin(\tau_j) \le s_i \\ 0 & \text{else.} \end{cases}$$

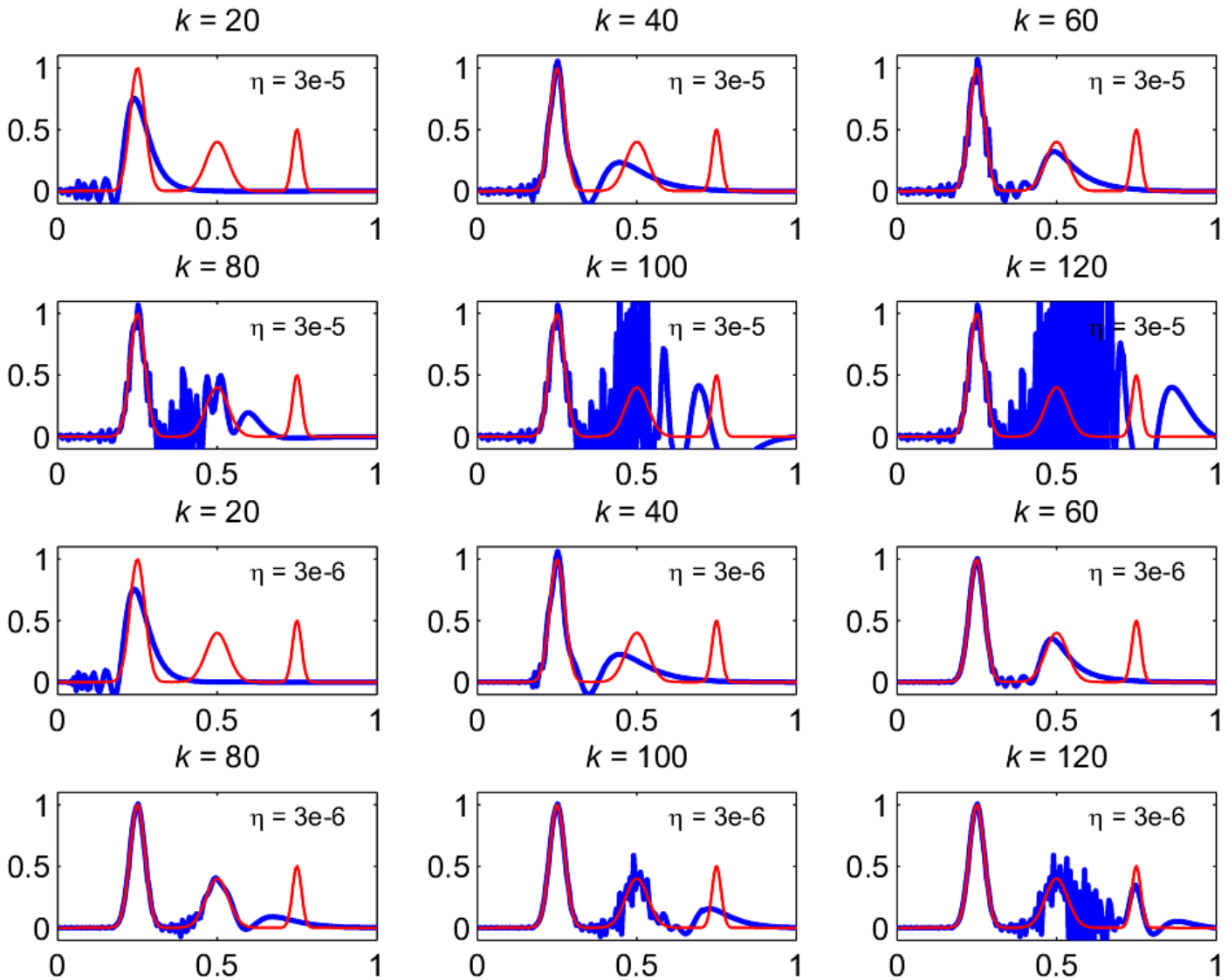
The singular values of this matrix show a much more satisfactory behavior free from rounding error artifacts.



Conclusion: be careful with the discretization!

TSVD Solutions

Selected TSVD solutions for two different noise levels:



Depth Resolution

As k increases we obtain information about deeper structures, until the noise starts to dominate the reconstruction.

With a lower noise level (where the noise starts to dominate for a larger k) the deep part of the reconstruction is more reliable.

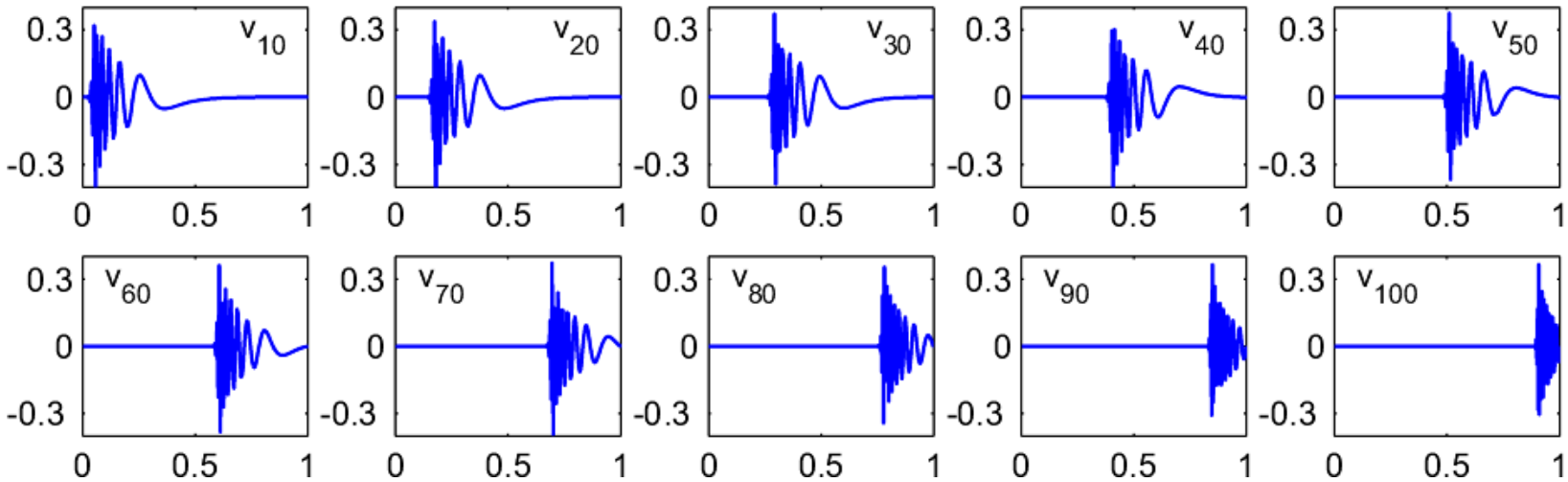
Only with a small noise level do we have a reliable reconstruction of the bottom layer of paint.

Depth resolution: to which depth can we, for a given problem and with a given noise level, reliably resolve details in the solution?

It is no surprise that the smaller the noise level, the deeper we can "look into" the material in this problem.

Depth Resolution Studies via Singular Vectors

Selected right singular vectors plotted versus the depth t:



Each singular vector carries information about the solution in a certain depth interval only, starting from the smallest depths.

When we include the first k SVD components in a regularized solution, then we can recover details in those depth ranges for which the singular vectors $v_1, \ldots, v_k$ carry information.