Image Restoration from a Machine Learning Perspective

September 2012

NIST

Dianne P. O'Leary

Image Restoration
Using Machine Learning

Dianne P. O'Leary

Applied and Computational Mathematics Division, NIST

Computer Science Dept. and
Institute for Advanced Computer Studies
University of Maryland
oleary@cs.umd.edu

http://www.cs.umd.edu/users/oleary

Joint work with Julianne Chung (Virginia Tech),
Matthias Chung (Virginia Tech)

Support from NIST and NSF.

The Problem

- Focus on numerical solution of ill-posed problems.
- In particular, we try to reconstruct a clear image from a blurred one.
- Focus on methods that take advantage of the singular value decomposition (SVD) of a matrix (spectral methods).

Goal of our work:

To achieve better solutions than previously obtained from the SVD.

Ingredients:

- Exploiting training data.
- Using Bayesian estimation.
- Designing optimal filters.

Note: I'll focus in this talk on methods that take advantage of having the full SVD available, but our methods can exploit the savings of using iterative methods as well.

The Problem

We have m observations b_i resulting from convolution of a blurring function with a true image.

We model this as a linear system

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}_{\text{true}} + \boldsymbol{\delta},$$

where $\boldsymbol{b} \in \boldsymbol{R}^m$ is the vector of observed data, $\boldsymbol{x}_{\text{true}} \in \boldsymbol{R}^n$ is an unknown vector containing values of $x(t_j)$, matrix $\boldsymbol{A} \in \boldsymbol{R}^{m \times n}, m \geq n$, is known, and $\boldsymbol{\delta} \in \boldsymbol{R}^m$ represents noise in the data.

Goal: compute an approximation of $\boldsymbol{x}_{\text{true}}$, given $\boldsymbol{b}$ and $\boldsymbol{A}$.

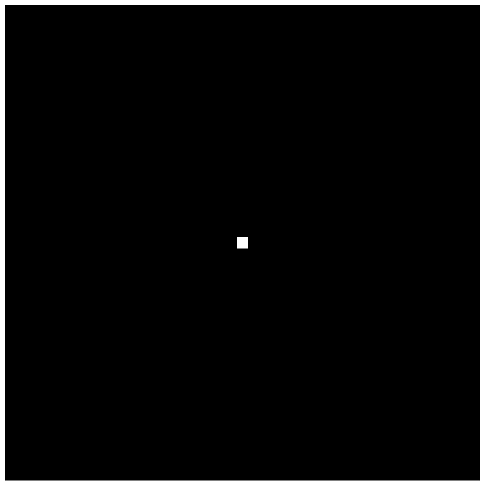
In other words: We need to learn the mapping between blurred images and true ones.

Problem characteristics

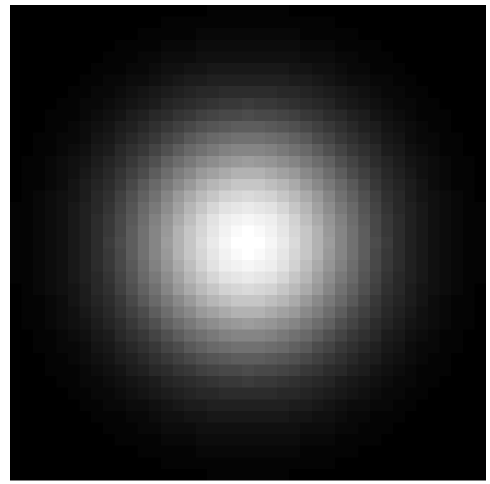
This is a discretization of an ill-posed inverse problem, meaning that small perturbations in the data may result in large errors in the solution.

Example

Suppose we have taken a picture but our lens gives us some Gaussian blur:



a single bright pixel



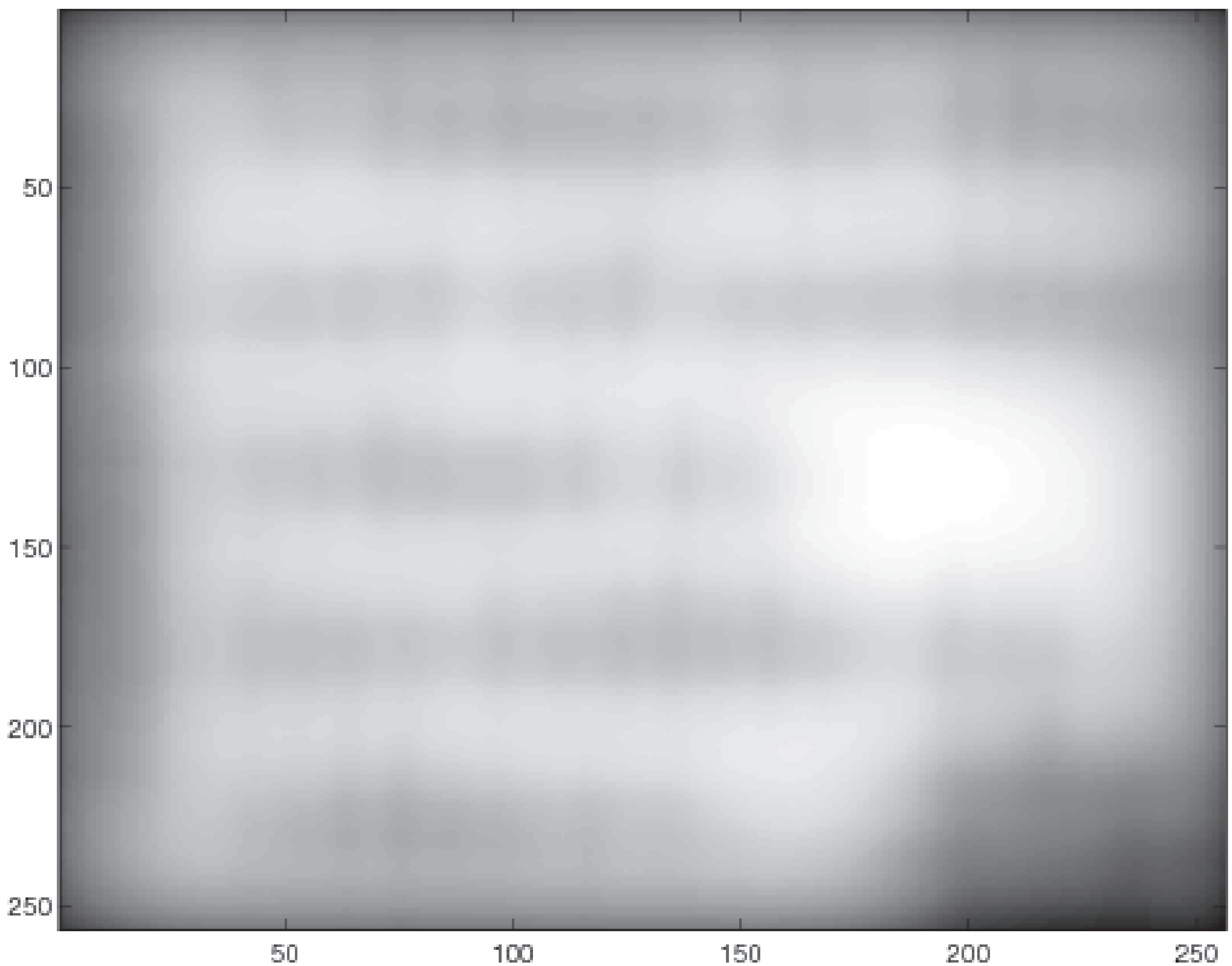
the blurred pixel

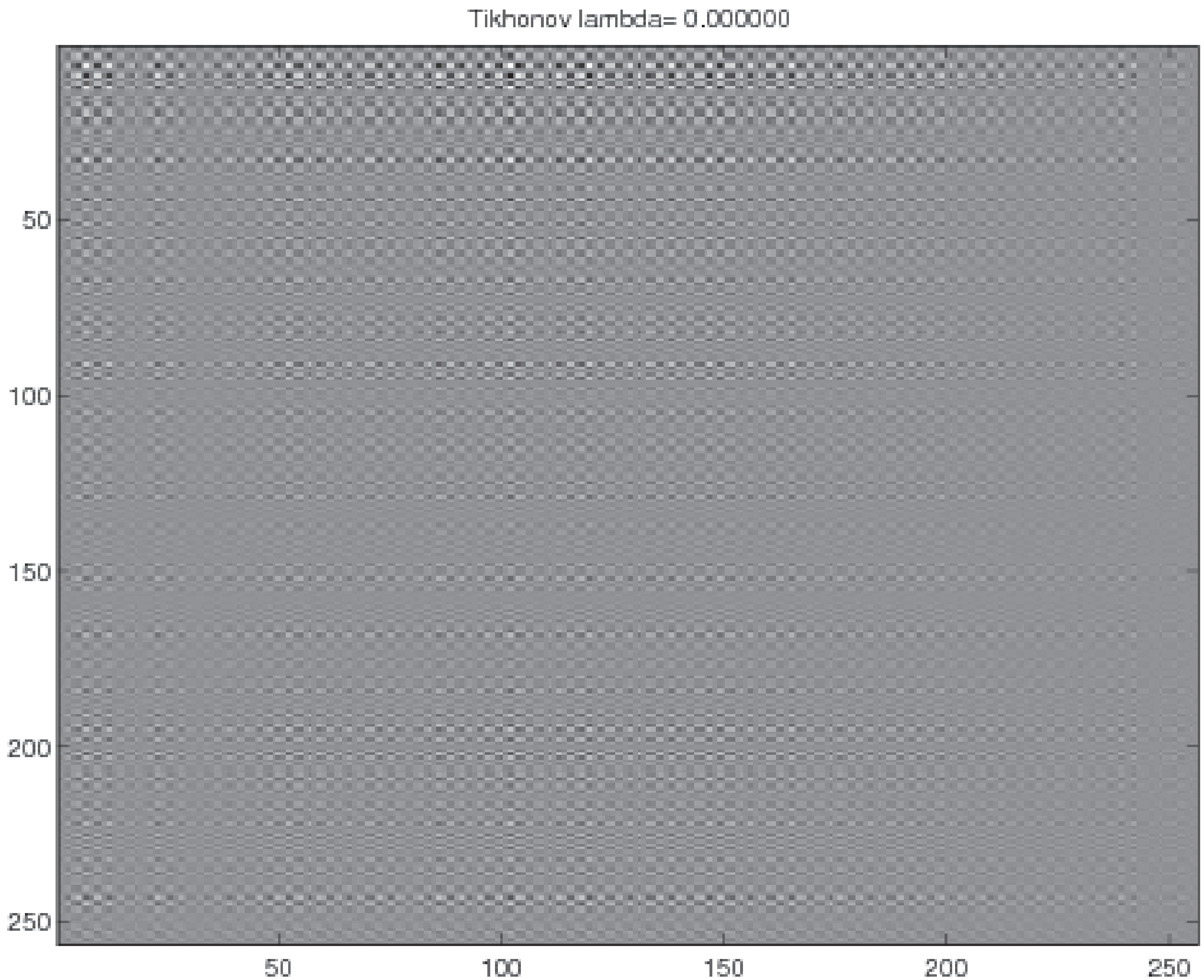
We construct the matrix $\boldsymbol{A}$ from the blurred image.

Our problem becomes

$$\min_{\boldsymbol{x}} \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \ .$$

Can we deblur this image?




Tikhonov lambda= 0.000000
50
100
150
200
250
50
100
150
200
250

Remedy

We regularize our problem by using extra information we have about the solution.

For example,

- We may have a bound on $\|\boldsymbol{x}\|_1$ or $\|\boldsymbol{x}\|_2$.
- We may know that $\mathbf{0} \le \boldsymbol{x}$, and we may have upper bounds, too.

Example, continued

Suppose we replace our problem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ by

$$\min_{\boldsymbol{x}} \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

subject to

$$\|\boldsymbol{x}\|_2 \leq \beta.$$

This formulation is called Tikhonov regularization.

Using a Lagrange multiplier λ, this problem becomes

$$\max_{\lambda} \min_{\boldsymbol{x}} \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda(\|\boldsymbol{x}\|_2 - \beta).$$

Write the solution to this problem using a spectral decomposition, the SVD of $\boldsymbol{A}$:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T,$$

where

- $\boldsymbol{\Sigma} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}} \\ 0 \end{bmatrix}$ is diagonal with entries equal to the singular values

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

- The singular vectors $\boldsymbol{u}_i$ $(i = 1, \ldots, m)$ and $\boldsymbol{v}_i$ $(i = 1, \ldots, n)$ are columns of the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively.
- The singular vectors are orthonormal, so $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_m$ and $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_n$.

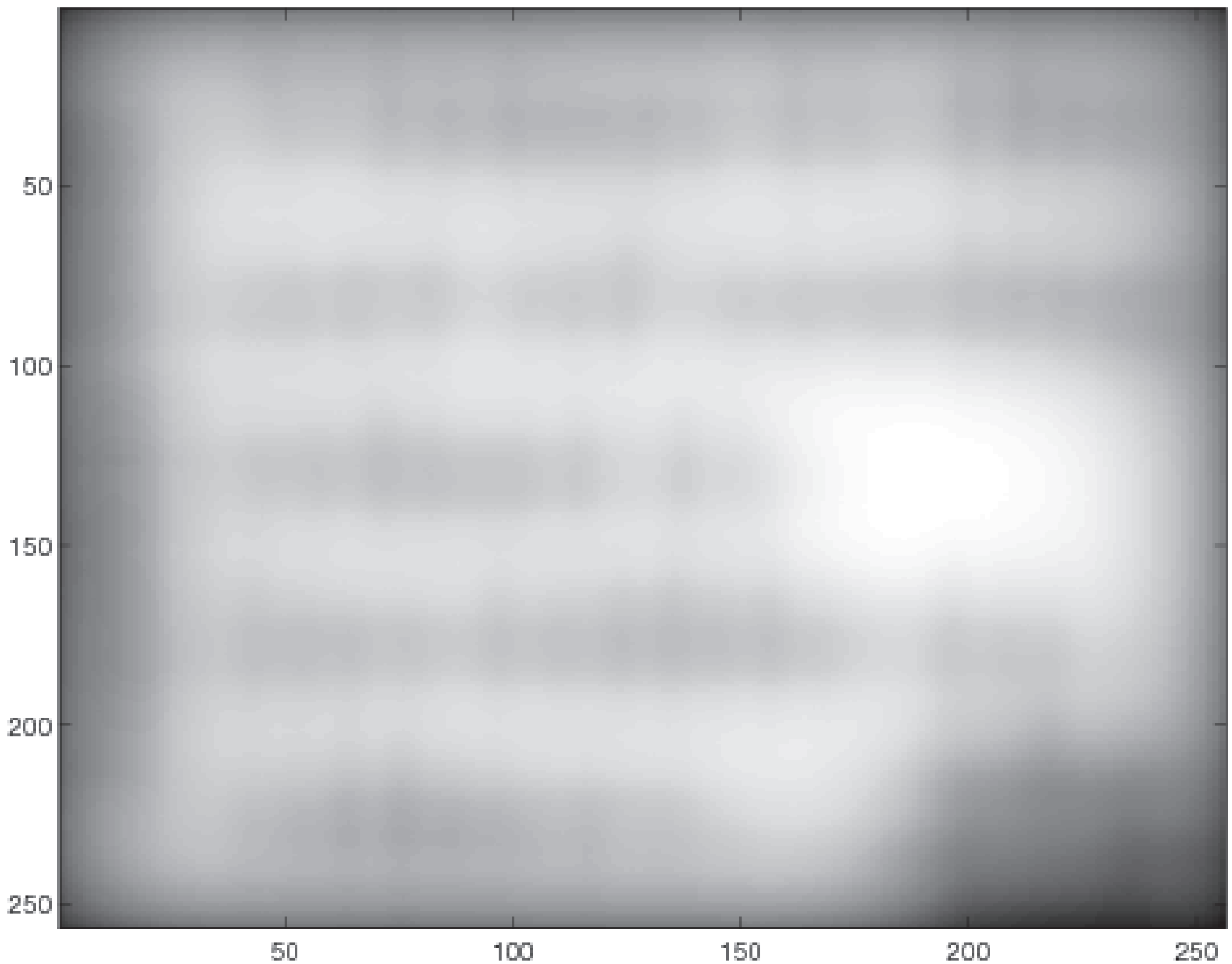
The solution becomes

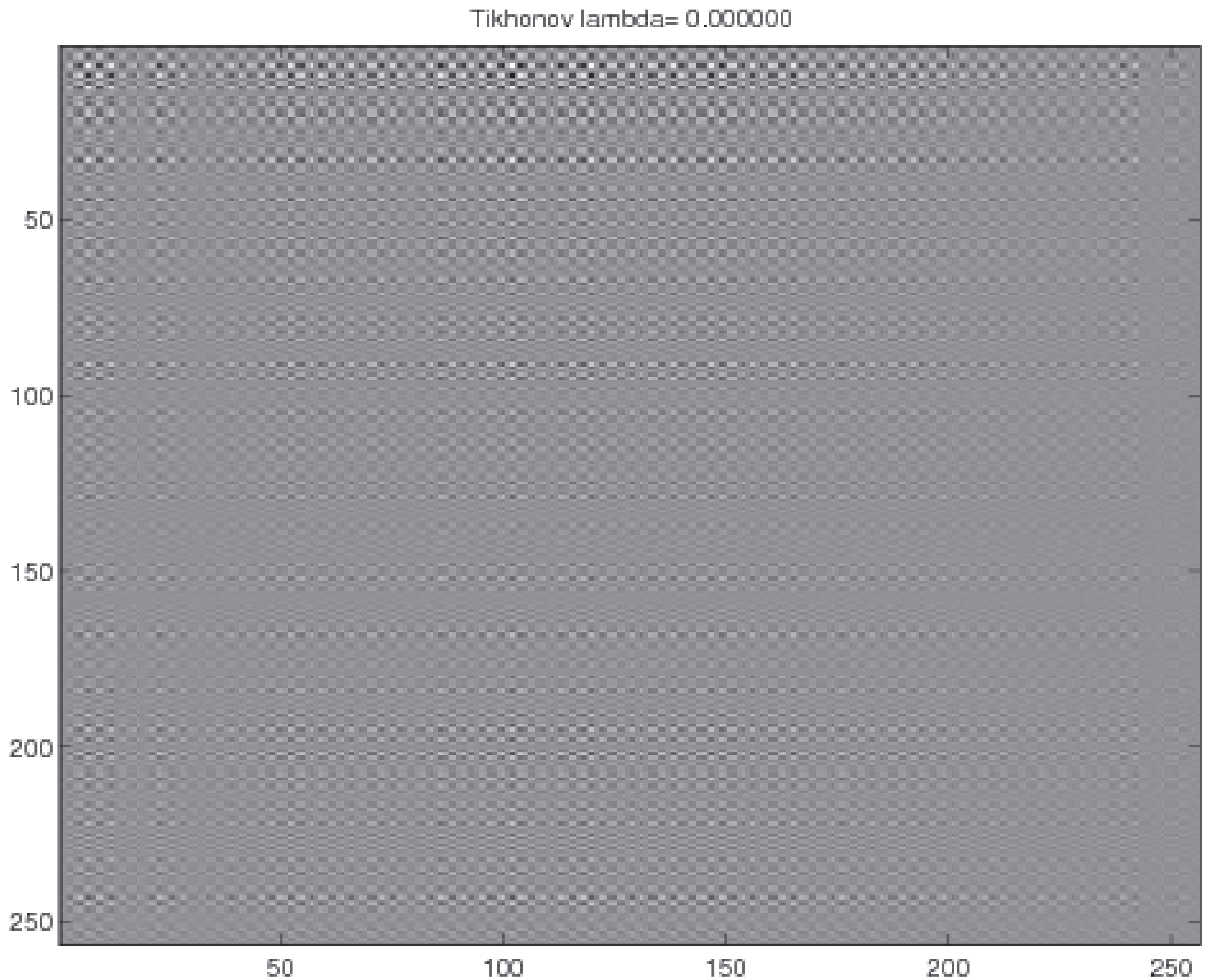
$$\boldsymbol{x} = \boldsymbol{V}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^T\boldsymbol{c},$$

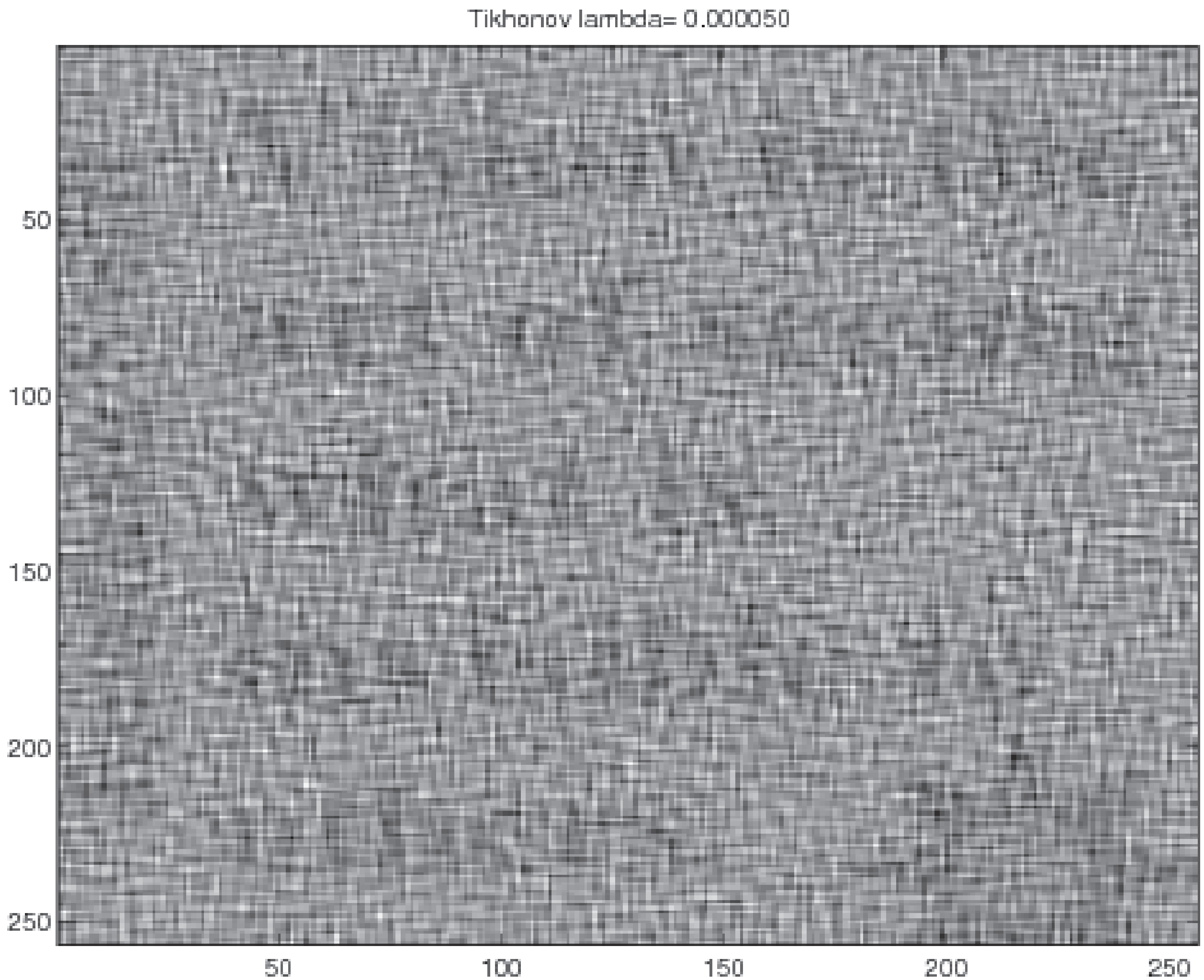
where $\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{b}$.

Unfortunately, we don't know λ, so a bit of trial-and-error is necessary.

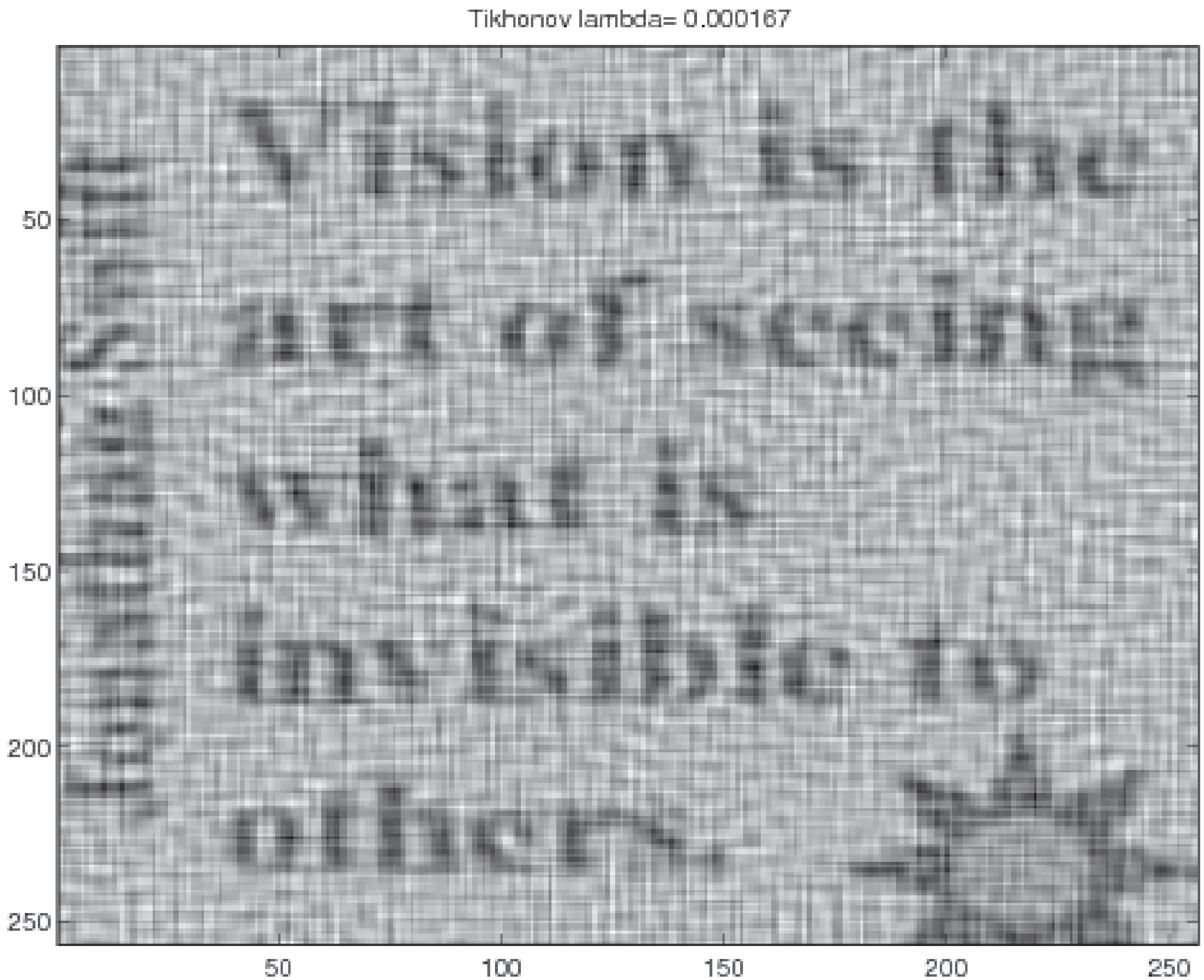
Can we deblur this image? (Revisited)



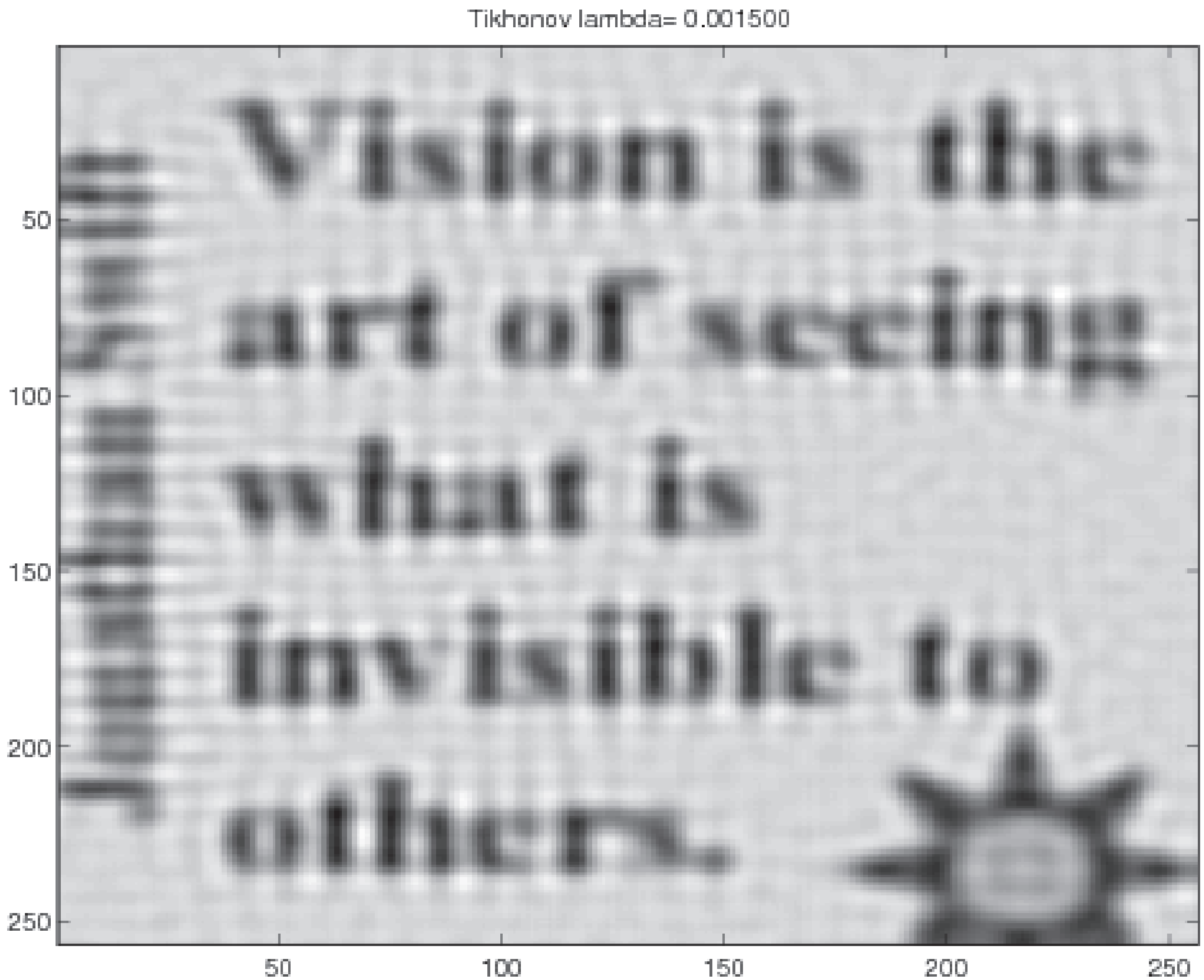

Tikhonov lambda= 0.000000
50
100
150
200
250
50
100
150
200
250

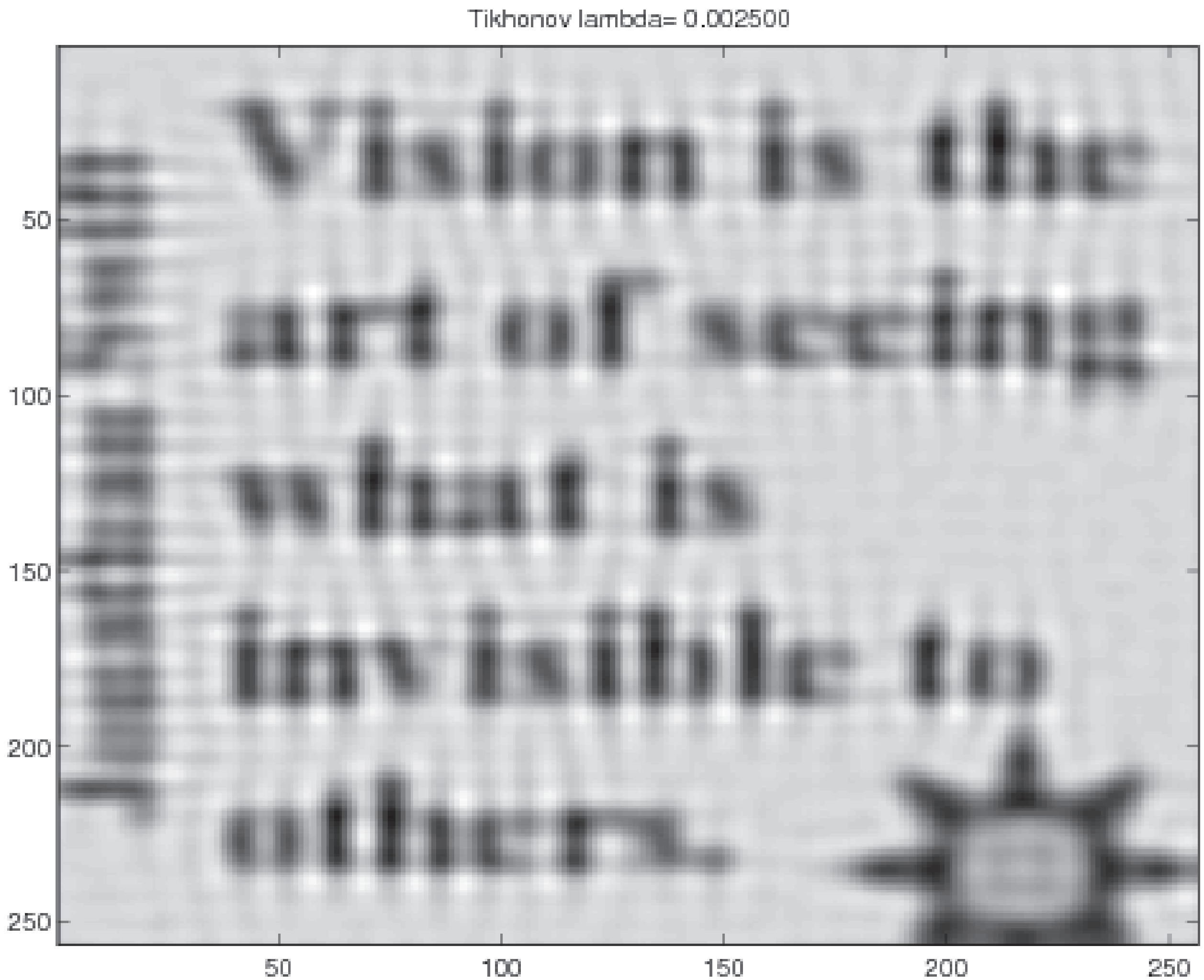


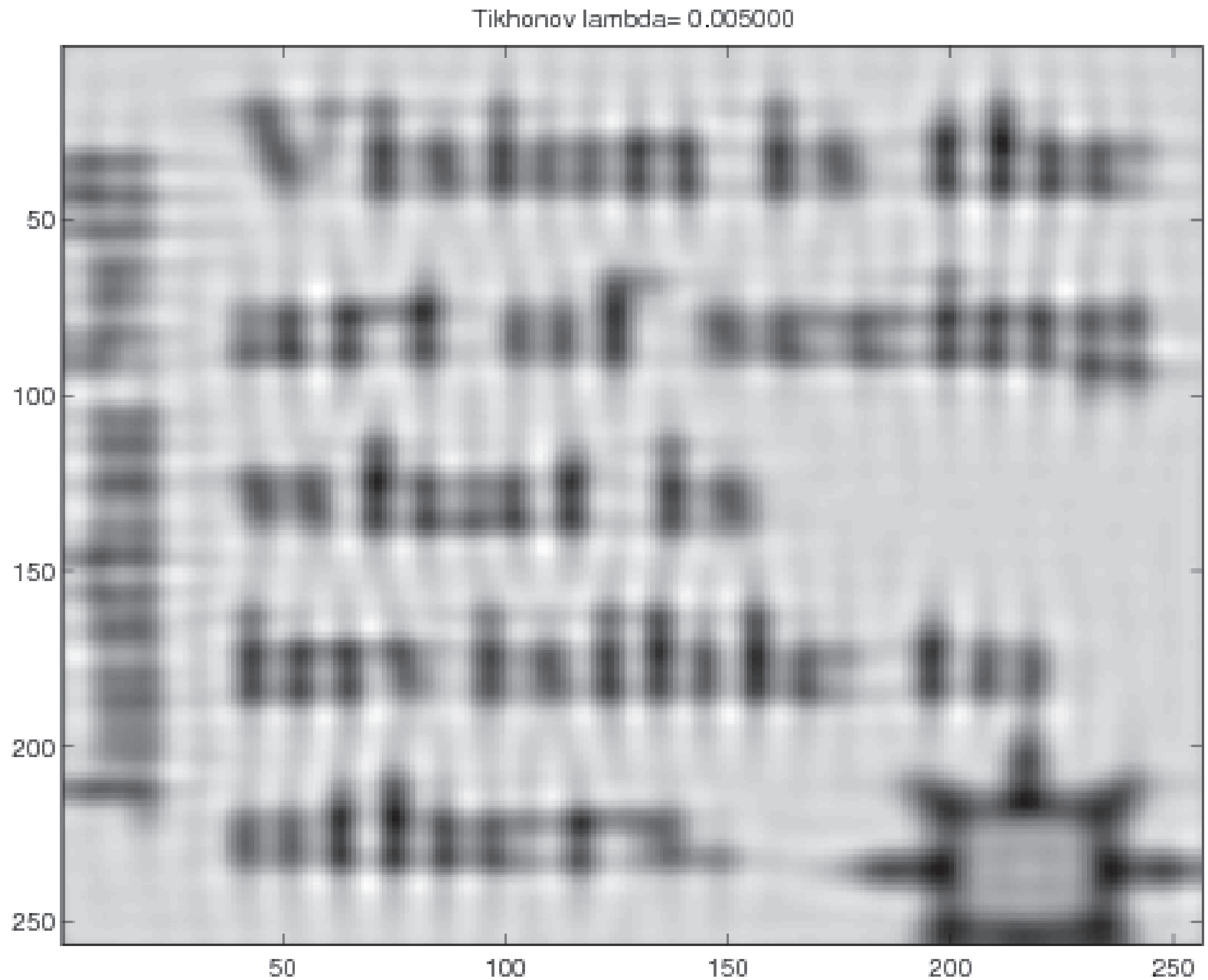
Tikhonov lambda= 0.000050
50
100
150
200
250
50
100
150
200
250

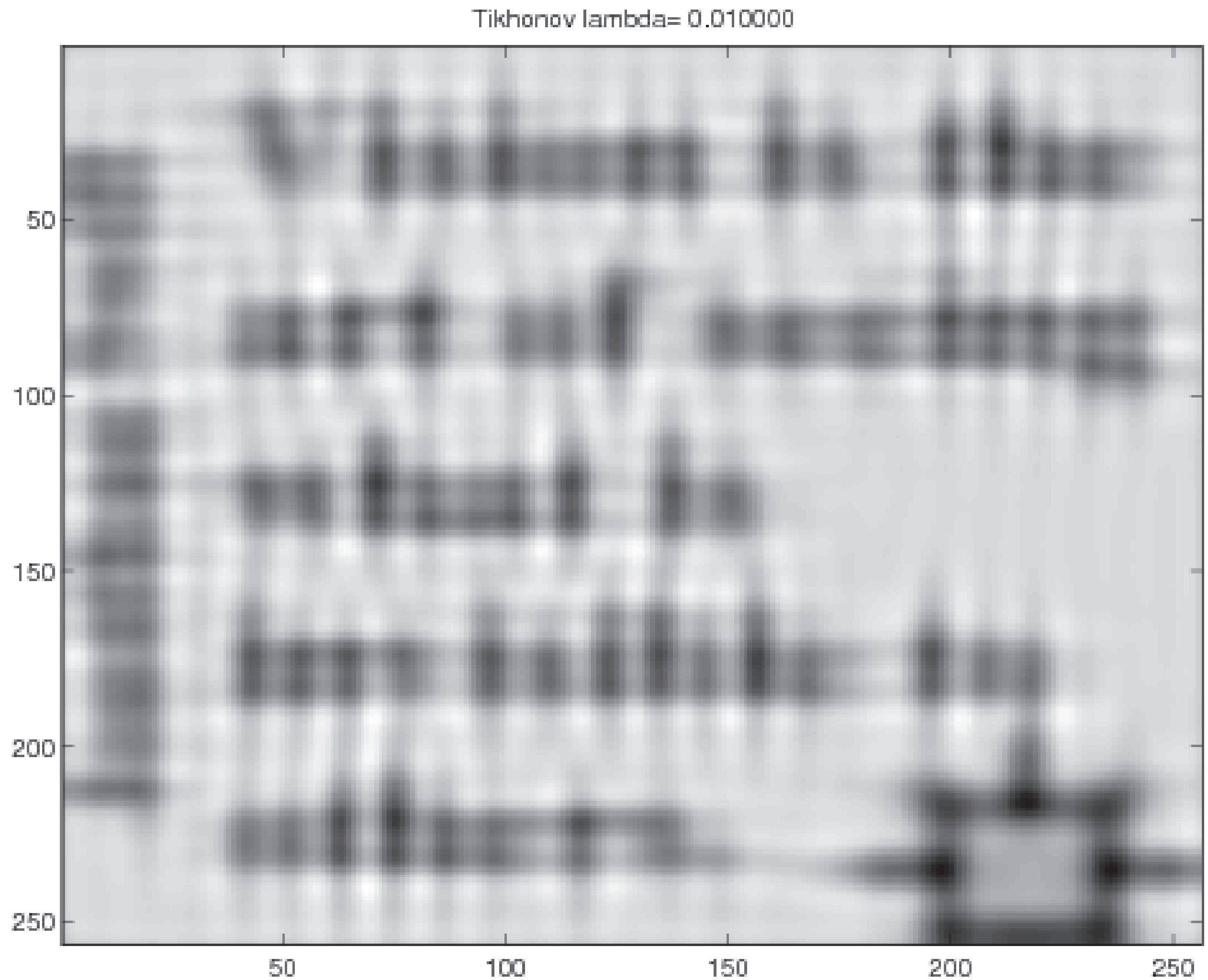

Tikhonov lambda= 0.000167
Vision is the
art of seeing
what is
invisible to
others.
50
100
150
200
250
50
100
150
200
250

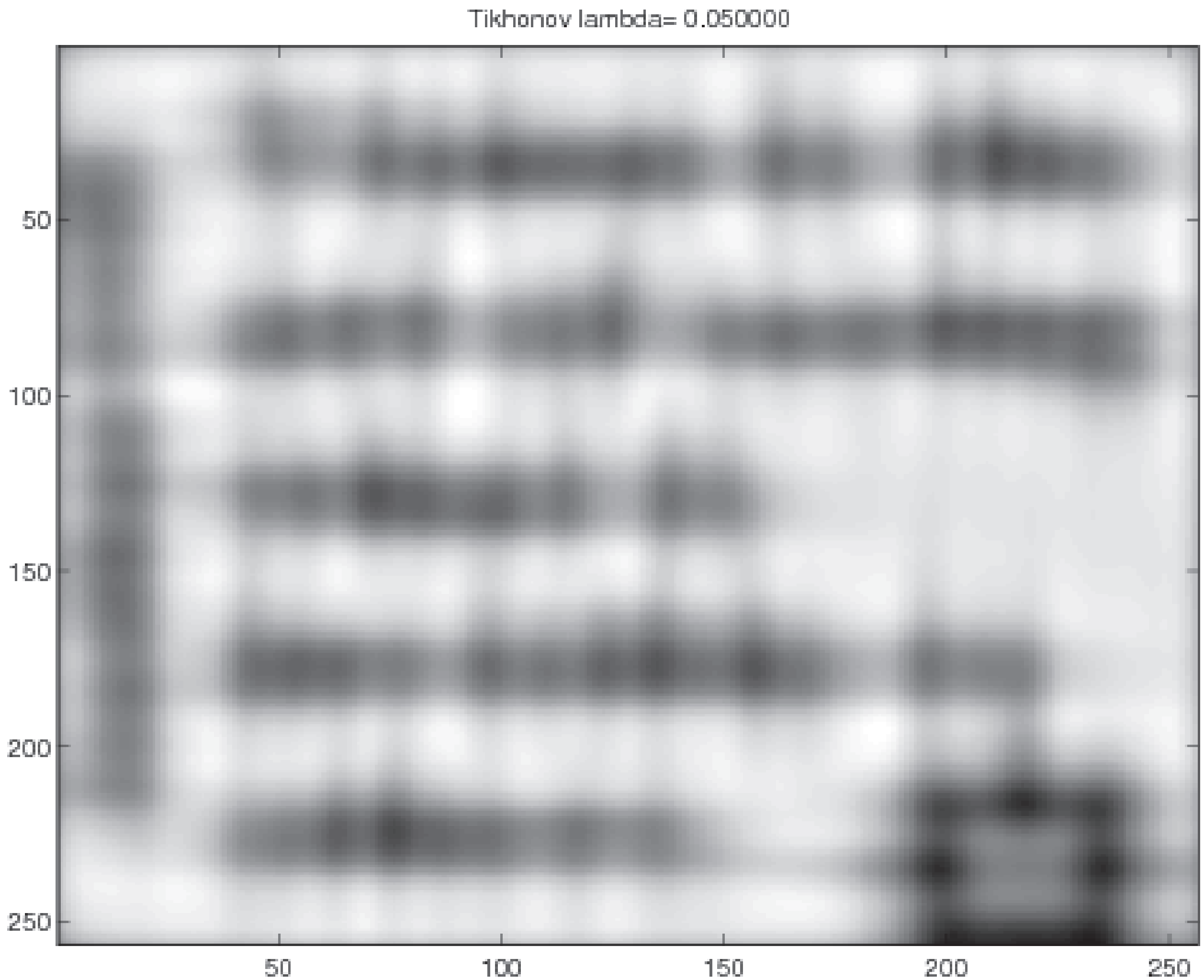

Tikhonov lambda= 0.001000
Vision is the
art of seeing
what is
invisible to
others.
50
100
150
200
250
50
100
150
200
250


Tikhonov lambda= 0.001500
Vision is the
art of seeing
what is
invisible to
others.
50
100
150
200
250
50
100
150
200
250


Tikhonov lambda= 0.002500
Vision is the
art of seeing
what is
invisible to
others.
50
100
150
200
250
50
100
150
200
250


Tikhonov lambda= 0.005000
50
100
150
200
250
50
100
150
200
250


Tikhonov lambda= 0.010000
50
100
150
200
250
50
100
150
200
250


Tikhonov lambda= 0.050000
50
100
150
200
250
50
100
150
200
250

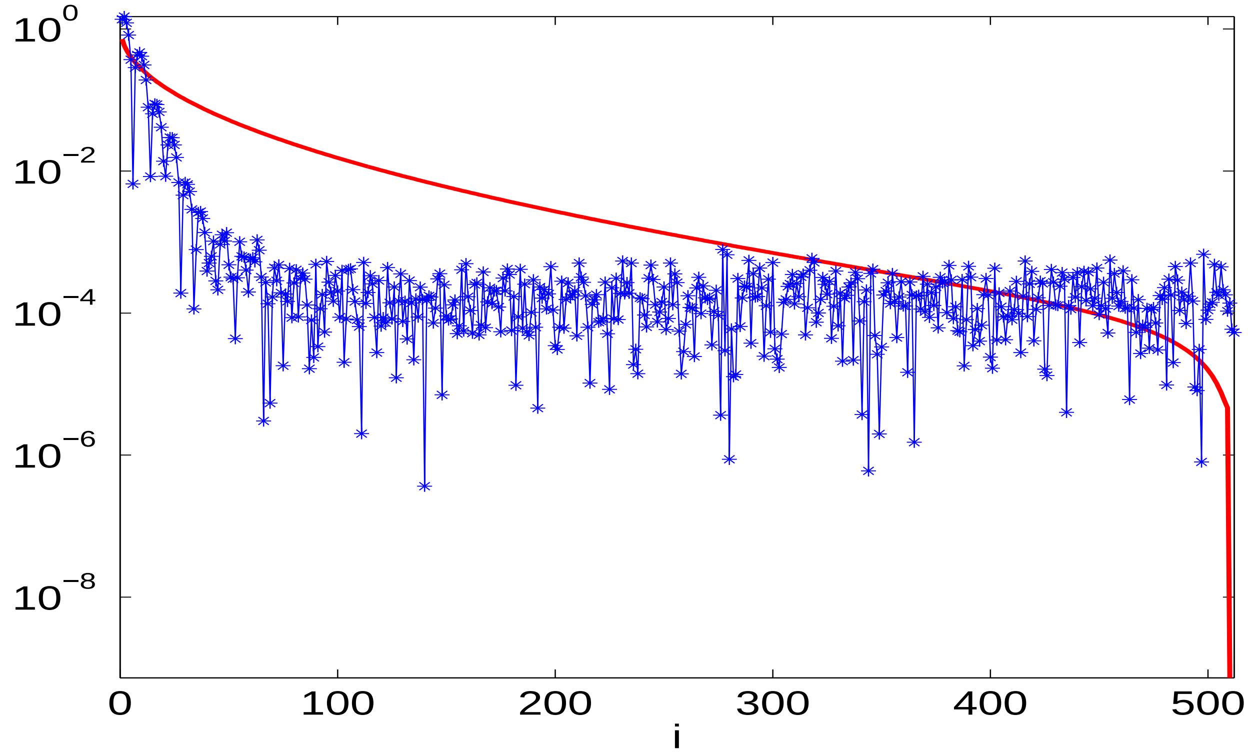
What makes spectral methods work?

For discretizations of ill-posed problems:

- The singular values $\sigma_i > 0$ have a clusterpoint at 0 as $m, n \to \infty$.
- There is no noticeable gap in the singular values, and therefore the matrix $\boldsymbol{A}$ should be considered to be full-rank.
- The small singular values correspond to oscillatory singular vectors.

We need two further features:

- The discretization is fine enough that to satisfy the *discrete Picard condition*: the sequence $\{|\boldsymbol{u}_i^T \boldsymbol{b}_{\text{true}}|\}$ decreases to 0 faster than $\{\sigma_i\}$.
- The noise components δ_j, $j = 1, \ldots, m$, are uncorrelated, zero mean, and have identical but unknown variance.



Picard plot: The singular values, represented with a red solid line, exhibit gradual decay to 0. The coefficients $|\boldsymbol{u}_i^T \boldsymbol{b}|$ are represented by blue stars.

The Plan

- The Problem
- Spectral Filtering
- Learning the Filter: Data to the Rescue
- Judging Goodness
- Results
- Conclusions

Spectral Filtering

We wrote our Tikhonov solution as

$$\boldsymbol{x} = \boldsymbol{V}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma}^T \boldsymbol{c}\,,$$

where $\boldsymbol{c} = \boldsymbol{U}^T \boldsymbol{b}$.

We can express this as

$$\boldsymbol{x} = \boldsymbol{V}\boldsymbol{\Phi}(\lambda)\boldsymbol{\Sigma}^\dagger \boldsymbol{c}\,,$$

where the diagonal matrix $\boldsymbol{\Phi}$ is

$$\boldsymbol{\Phi}(\lambda) = (\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}.$$

For Tikhonov, λ is a single parameter.

- Can we do better by using more parameters, resulting in a filter matrix $\boldsymbol{\Phi}(\boldsymbol{\alpha})$?
- If so, how can we choose $\boldsymbol{\alpha}$? We will learn it!

The Plan

- The Problem
- Spectral Filtering
- Learning the Filter: Data to the Rescue
- Judging Goodness
- Results
- Conclusions

Learning the Filter: Data to the Rescue

What do we need?

Informally, we need:

- Knowledge of $\boldsymbol{A}$.
- A universe of possible true images.
- A blurred image corresponding to one of these true images, chosen at random.
- Knowledge of some characteristics of the noise.
- Some training data.

More formally, we need:

- Knowledge of $\boldsymbol{A}$.
  We assume we know it exactly.
- A universe of possible true images.
  We assume that the true images that resulted in the ones presented to us are chosen from a known probability distribution $\mathcal{P}_{\boldsymbol{\xi}}$ on images in $\boldsymbol{\Xi} \subset \boldsymbol{R}^n$ that has finite second moments.
- A blurred image corresponding to one of these true images, chosen at random, according to $\mathcal{P}_{\boldsymbol{\xi}}$.
- Knowledge of some characteristics of the noise:
  mean zero, finite second moments, known probability distribution $\mathcal{P}_{\boldsymbol{\delta}}$ on noise vectors in $\boldsymbol{\Delta} \subset \boldsymbol{R}^n$.
- Some training data:
  pairs consisting of a true image and its resulting blurred image.

Where does the training data come from?

When an expensive imaging device is powered on, there is often a calibration procedure.

For example, for an MRI machine, we might use a phantom, made of material with density similar to that of the brain, and insert a small sphere with density similar to that of a tumor.

Taking images of the phantom at different positions in the field of view, or at different well-controlled rotations, gives us pairs of truth and measured values.

The Plan

- The Problem
- Spectral Filtering
- Learning the Filter: Data to the Rescue
- Judging Goodness
- Results
- Conclusions

How do we judge goodness of parameters?

We want to minimize the error in our reconstruction!

We settle for minimizing the expected error in our reconstruction:

Error vector

$$\boldsymbol{e}(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\delta}) = \boldsymbol{x}_{\text{filter}}(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\delta}) - \boldsymbol{\xi} \,,$$

Measure error as

$$\mathrm{ERR}(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\delta}) = \frac{1}{n} \sum_{i=1}^{n} \rho(e_i(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\delta})) \,,$$

where, for example,

$$\rho(z) = \frac{1}{p} |z|^p \,,$$

for $p \geq 1$, related to the p-norm of the error vector.

Choice of ρ

We use 1-norm, 2-norm, p-norm ($p = 4$, as an approximation to the ∞-norm).

We also use the Huber function to reduce the effects of outliers.

$$\rho(z) = \begin{cases} |z| - \frac{\beta}{2}, & \text{if } |z| \geq \beta, \\ \\ \frac{1}{2\beta}\, z^2, & \text{if } |z| < \beta, \end{cases}$$

Bayes risk minimization

An optimal filter would minimize the expected value of the error:

$$\check{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) = \boldsymbol{E}_{\boldsymbol{\delta},\boldsymbol{\xi}}\{\mathrm{ERR}(\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\delta})\},$$

Given our training data, we approximate this problem by minimizing the empirical Bayes risk

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} f_N(\alpha),$$

where

$$f_N(\boldsymbol{\alpha}) = \frac{1}{nN}\sum_{k=1}^{N}\sum_{i=1}^{n}\rho(e_i^{(k)}(\boldsymbol{\alpha})),$$

where the samples $\boldsymbol{\xi}^{(k)}$, and noise realizations, $\boldsymbol{\delta}^{(k)}$, for $k = 1, ..N$, constitute a *training set*.

Convergence theorems: Shapiro 2009.
Statistical learning theory: Vapnik 1998.

Standard choices for the parameters $\boldsymbol{\alpha}$

Two standard choices:

- Truncated SVD:

$$\phi_i^{\text{tsvd}}(\alpha) = \begin{cases} 1, & \text{if } i \leq \alpha, \\ 0, & \text{else,} \end{cases}$$

  with $\alpha \in \mathcal{A}^{\text{tsvd}} = \{1, \ldots, n\}$.

- Tikhonov filtering:,

$$\phi_i^{\text{tik}}(\alpha) = \frac{\sigma_i^2}{\sigma_i^2 + \alpha}.$$

  for $\alpha \in \mathcal{A}^{\text{tik}} = \boldsymbol{R}_+$.

Advantage: 1-parameter optimization problems are easy.

Disadvantage: The filters are quite limited by their form.

Most general choice of parameters

We let

$$\phi_i^{\mathrm{err}}(\boldsymbol{\alpha}) = \alpha_i, \;\; i = 1, \ldots, n$$

Advantage: The filters are now quite general.

Disadvantage: n-parameter optimization problems are hard and the resulting filter can be very oscillatory.

A compromise: smoothing filters

Take an n-parameter optimal filter and apply a smoothing operator to it:

$$\phi^{\text{smooth}} = \boldsymbol{K}\hat{\phi}^{\text{err}},$$

where $\boldsymbol{K}$ denotes a smoothing matrix (e.g., a Gaussian).

Advantage: The filter is now smoother.

Disadvantage: It is no longer optimal.

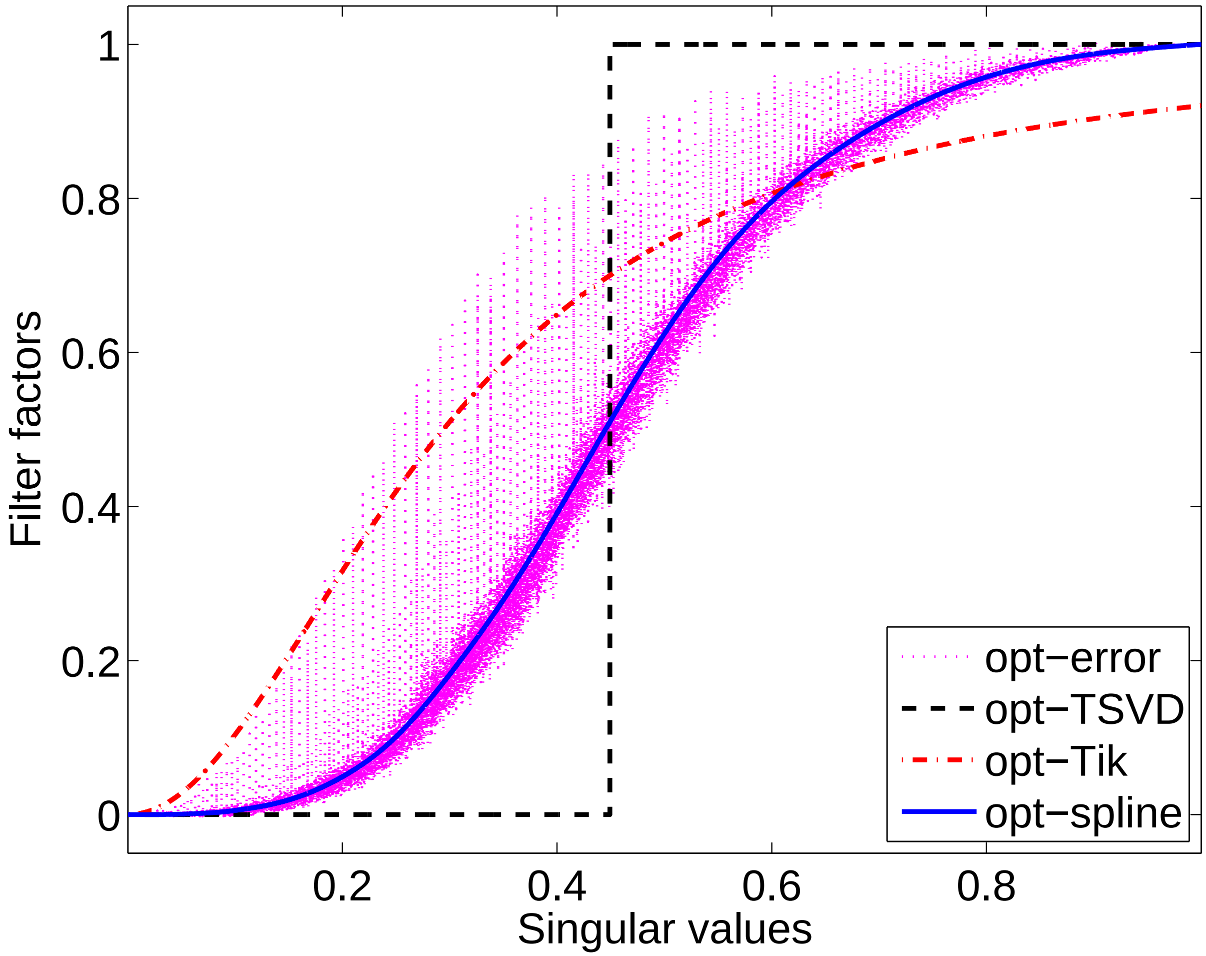
A second compromise: spline filters

Constrain the filter function $\phi(\boldsymbol{\alpha})$ to be a cubic spline with m (given) knots. (We used knots equally spaced on a log scale.)

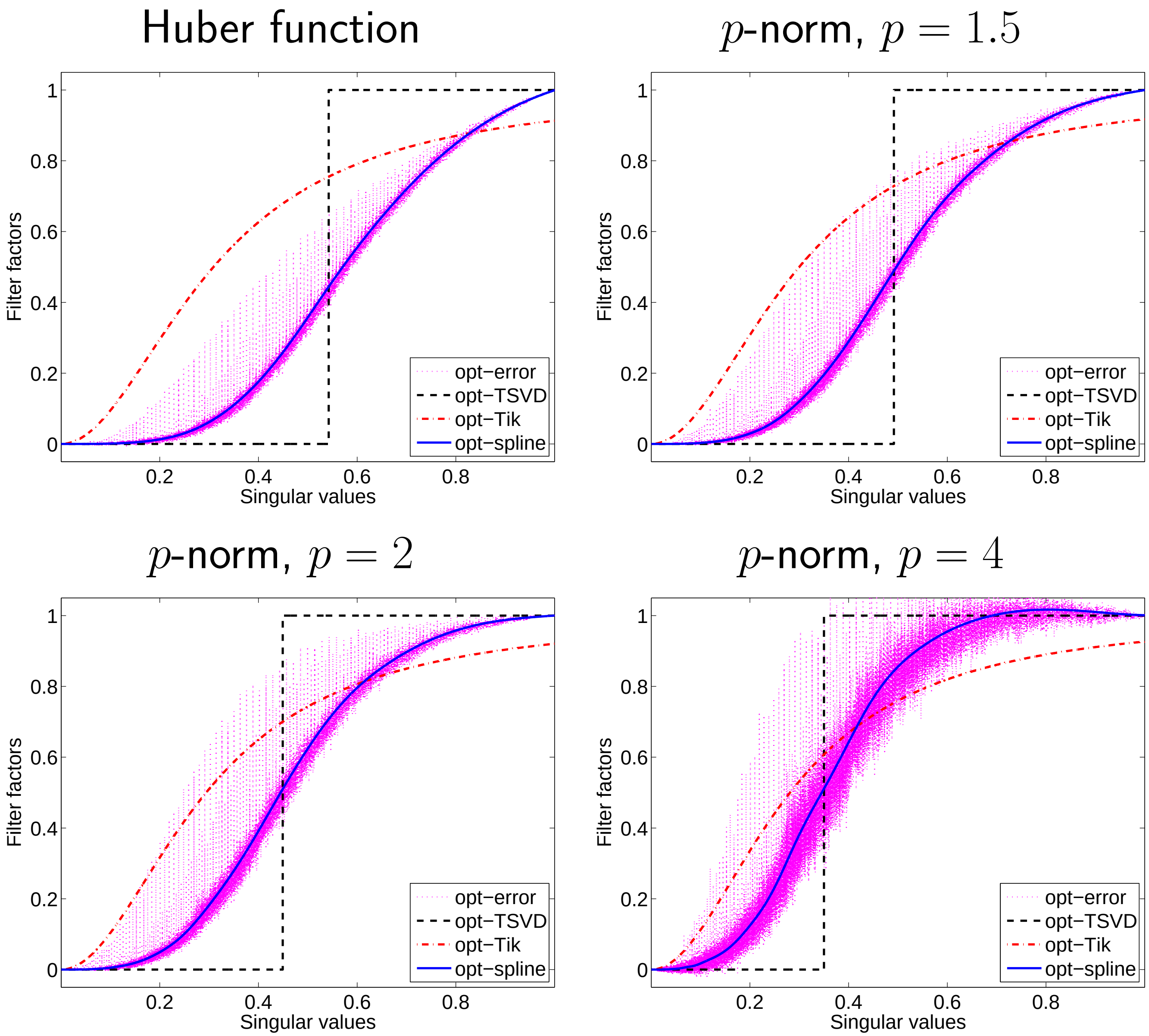
Advantage: This simplifies the optimization problem to have approx. m variables and prevents wild oscillations or abrupt changes.

Disadvantage: Knots and boundary conditions need to be specified or chosen by optimization.

Typical optimal filters



(Smooth filter (not shown) follows trend of optimal-error filter.)



Huber function
p-norm, $p = 1.5$
p-norm, $p = 2$
p-norm, $p = 4$
Filter factors
Singular values
opt−error
opt−TSVD
opt−Tik
opt−spline

Computational considerations

- Computational Issue 1: The Jacobian matrix contains the very ill-conditioned matrix $\mathbf{\Sigma}^{-1}$.

  Solution: We use a change of variables $\tilde{\boldsymbol{\phi}}^{\mathrm{err}} = \mathbf{\Sigma}^{-1}\boldsymbol{\phi}^{\mathrm{err}}$.

- Computational Issue 2: Choosing a minimization algorithm.

  Solution:

  - Golden section search for Tikhonov; discrete version for TSVD.
  - Linear programming interior-point method (IPM) for the 1-norm or ∞-norm.
  - A Newton variant for the p-norm with $2 \leq p < \infty$.
  - A gradient-based method or Newton for the Huber function.

- Computational Issue 3: The problem may be very large, with a large number of parameters or a large training set.

  Solution:

  - Iterative methods (e.g., conjugate gradient) can be used in the Newton variants (without forming derivative matrices) and in the IPM.
  - An object-oriented implementation makes this easy.
  - If a preconditioner is needed to accelerate convergence, a natural choice arises from using a subset of the training data.

The Plan

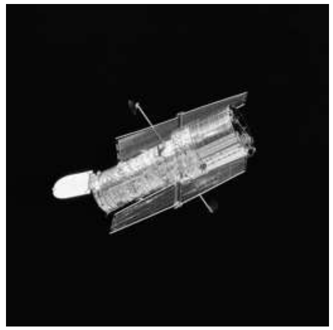
- The Problem
- Spectral Filtering
- Learning the Filter: Data to the Rescue
- Judging Goodness
- Results
- Conclusions

Example 1

Test problem:

- Training: 800 images, 256×256 generated from 8 satellite images, each with 100 rigid transformations (rotation, translation, magnification).
- Blur: symmetric Gaussian point spread function.
- Blurred images: Blur and add Gaussian random noise, with standard deviation uniformly sampled from [0.1, 0.15].
- Validation: 800 different satellite images with 100 rigid transformations, blurred with noise added.

3 Training and 3 validation images





Cost of training: $p = 2$

Macbook Pro with OS-X 10.6 and 8GB memory, running MATLAB 7.10.0 (64-bit).

optimal TSVD filter	1 parameter	606 sec.
optimal Tikhonov filter	1 parameter	1787 sec.
optimal spline filter	50 parameters	265 sec.
optimal error filter	256^2 parameters	237 sec.

Performance measures

- Error (ERR):

$$\mathrm{ERR}(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\delta}) = \frac{1}{n} \sum_{i=1}^{n} \rho(e_i(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\delta})).$$

- Relative Error (REL):

$$\frac{\mathrm{ERR}}{\frac{1}{n}\sum_{i=1}^{n} \rho(\xi_i)}.$$

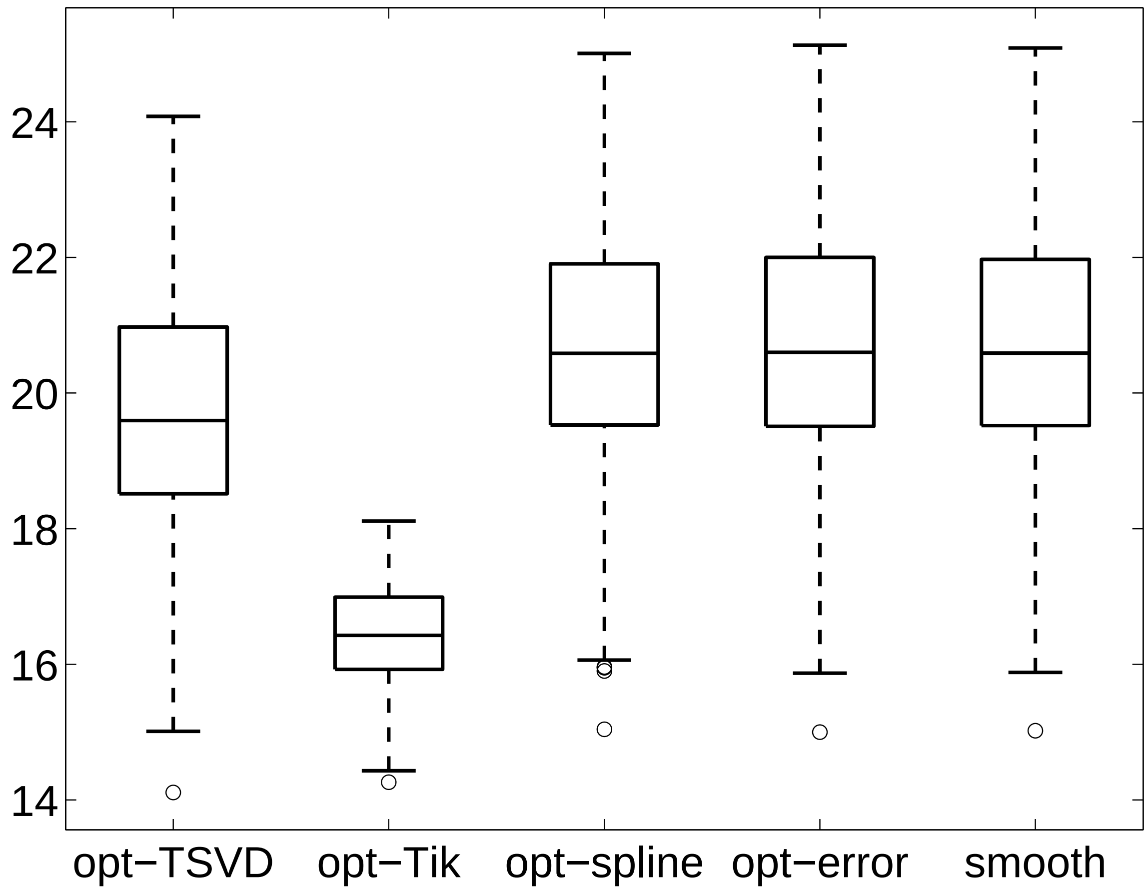
- Signal-to-noise ratio (SNR_ρ) with respect to ρ:

$$10 \cdot \log_{10}\left(\frac{1}{\mathsf{REL}}\right).$$

- Standard signal-to-noise ratio (SNR):

$$10 \cdot \log_{10}\left(\frac{\|\boldsymbol{\xi}\|_2^2}{\|\boldsymbol{x}_{\text{filter}} - \boldsymbol{\xi}\|_2^2}\right).$$

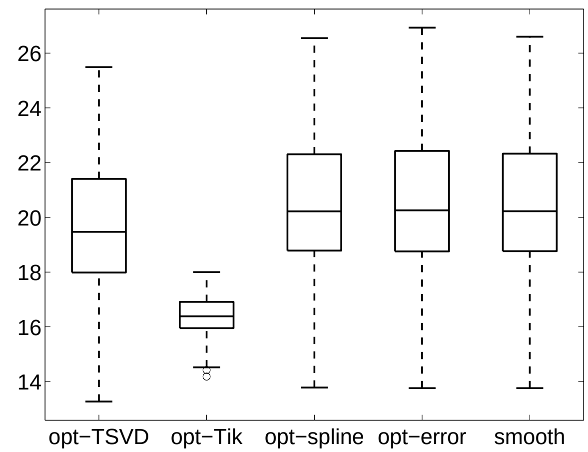
Results: SNR for all validation images



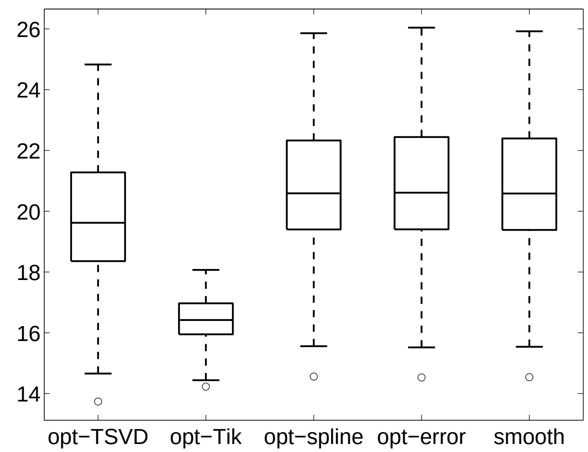
2-norm

SNR_p is similar.

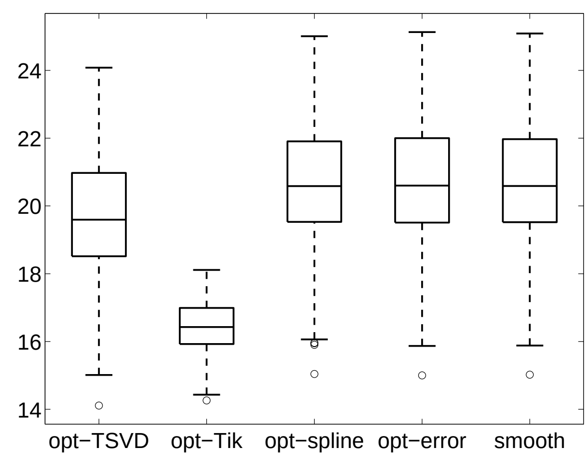
Results: SNR



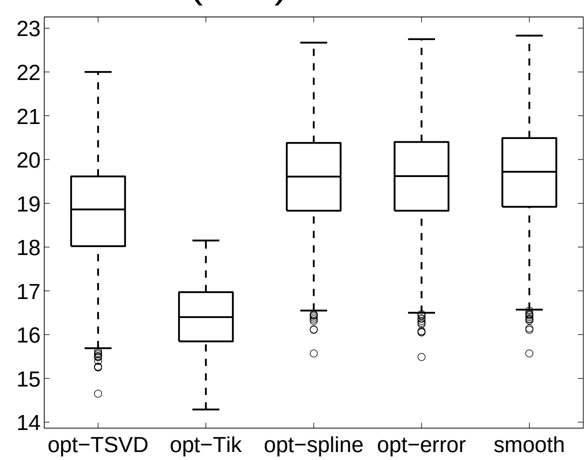
Huber



(1.5)-norm

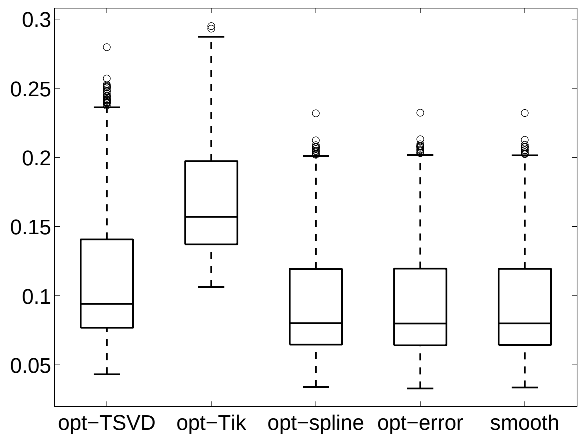


2-norm

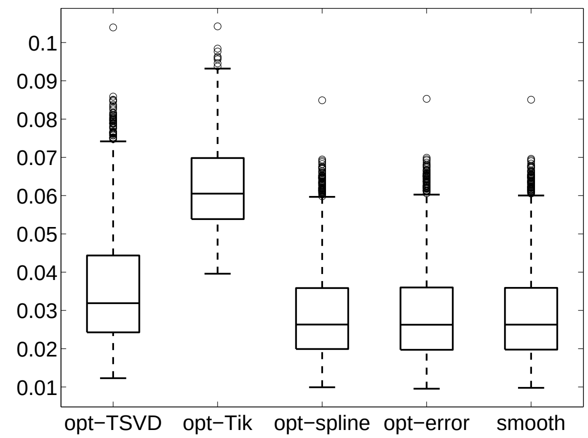


4-norm

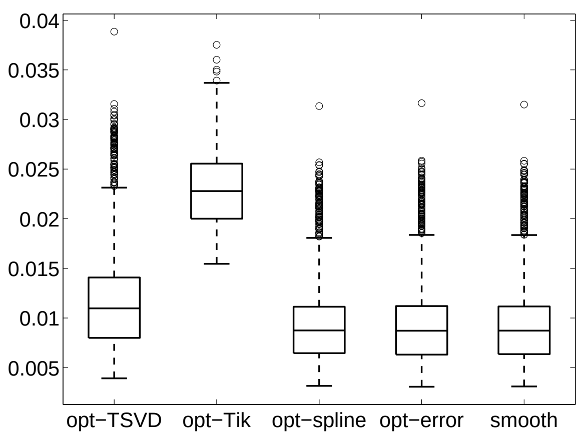
Results: REL for all validation images



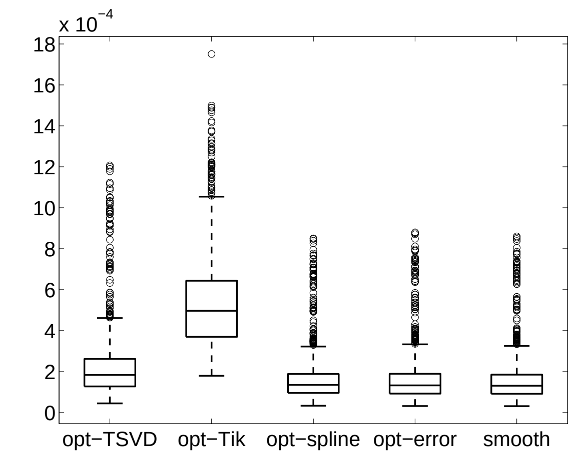
Huber



(1.5)-norm

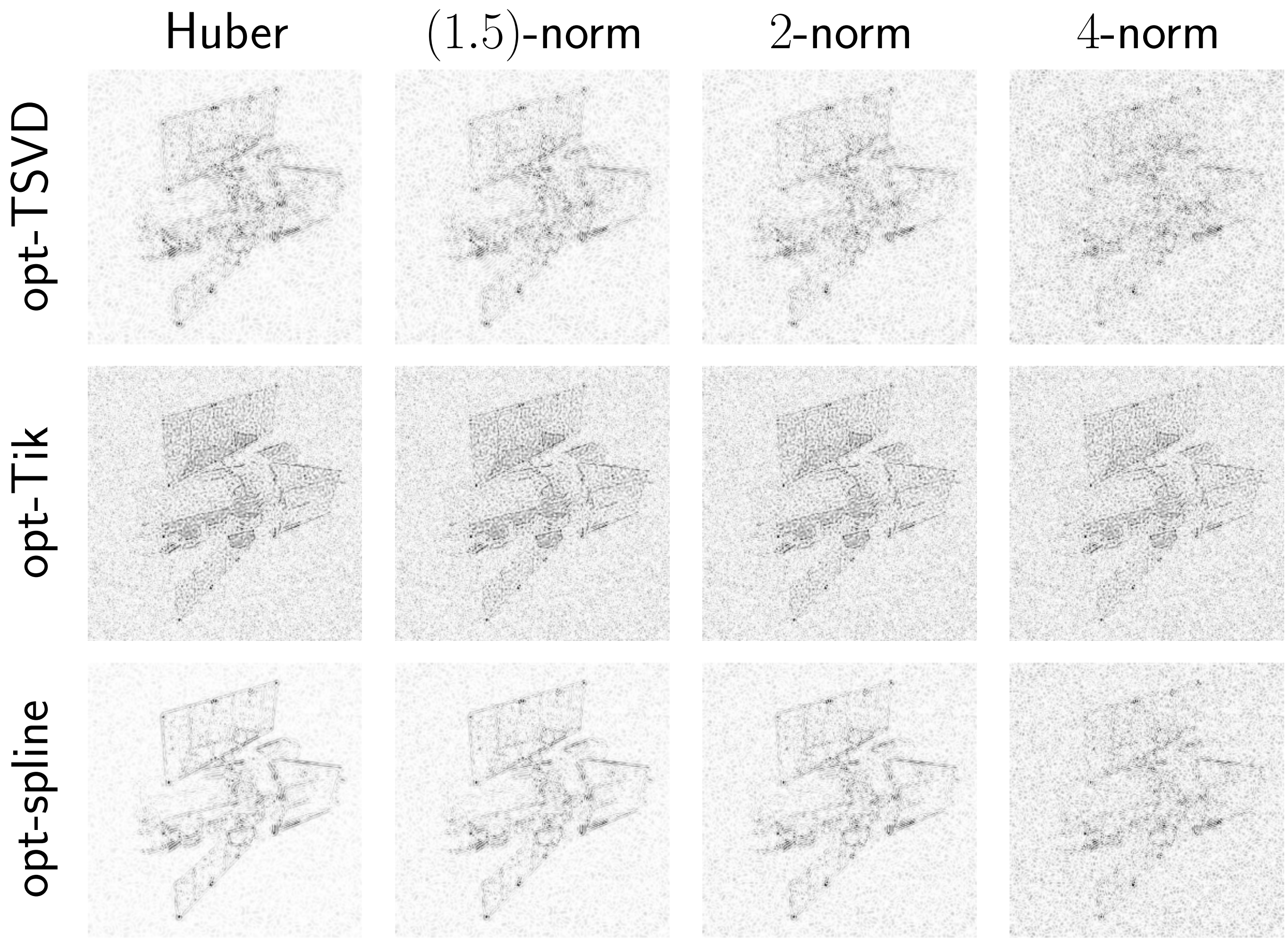


2-norm



4-norm

Absolute error images for one validation image



Opt-error and Smooth similar.

Confidence

The training images can be used to obtain uncertainty estimates:

- For each computed optimal filter, we reconstruct all of the training images and evaluate the average error per pixel,
- The expected error in each pixel is approximated by the sample mean

$$\mu_i = \frac{1}{N} \sum_{k=1}^{N} e_i^{(k)} , \qquad i = 1, \ldots, n ,$$

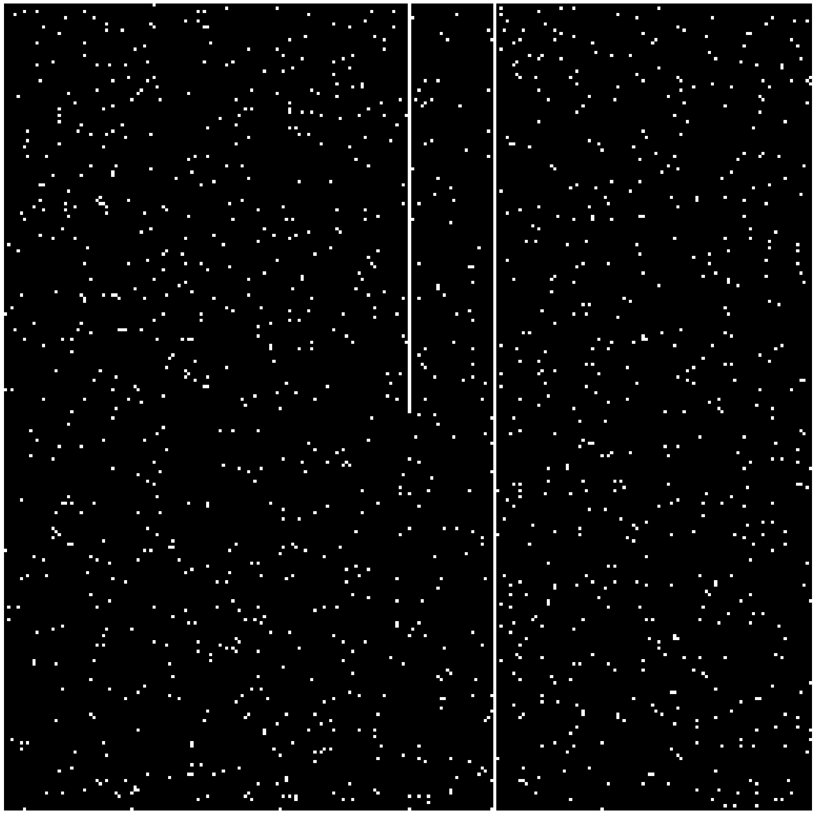
- This is very close to the average error we see in the training images, as it should be if our assumptions hold.

Average (standard deviation) of pixel error

opt-		Huber	(1.5)-norm	2-norm	4-norm
TSVD	T	-2.35e-4 (6e-3)	-2.04e-4 (5e-3)	-1.79e-4 (5e-3)	-1.29e-4 (4e-3)
	V	-2.54e-4 (6e-3)	-2.21e-4 (5e-3)	-1.95e-4 (5e-3)	-1.43e-4 (5e-3)
Tik	T	-1.94e-2 (8e-3)	-1.84e-2 (7e-3)	-1.77e-2 (7e-3)	-1.63e-2 (7e-3)
	V	-2.32e-2 (1e-2)	-2.20e-2 (1e-2)	-2.12e-2 (9e-3)	-1.96e-2 (9e-3)
spline	T	-5.50e-4 (6e-3)	-3.64e-4 (5e-3)	-2.18e-4 (4e-3)	1.01e-4 (4e-3)
	V	-6.08e-4 (6e-3)	-4.00e-4 (5e-3)	-2.34e-4 (5e-3)	1.33e-4 (4e-3)
error	T	-5.61e-4 (5e-3)	-3.30e-4 (4e-3)	-1.42e-4 (4e-3)	2.61e-4 (3e-3)
	V	-6.34e-4 (5e-3)	-3.71e-4 (4e-3)	-1.57e-4 (4e-3)	3.25e-4 (3e-3)

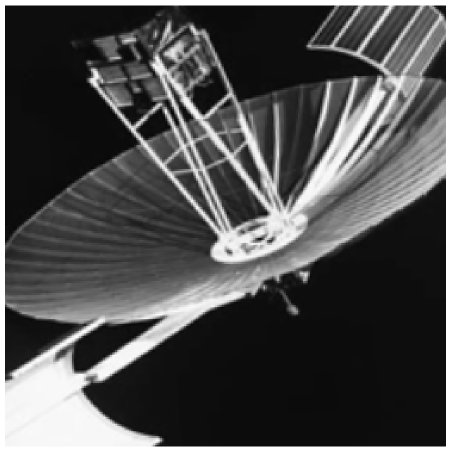
Example 2

Test problem: Suppose our camera is imperfect, having a substantial number of dead pixels:



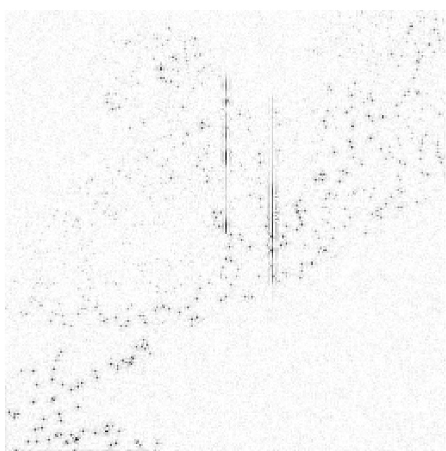
We use 40 training images to learn the filter function.

Results on a validation image

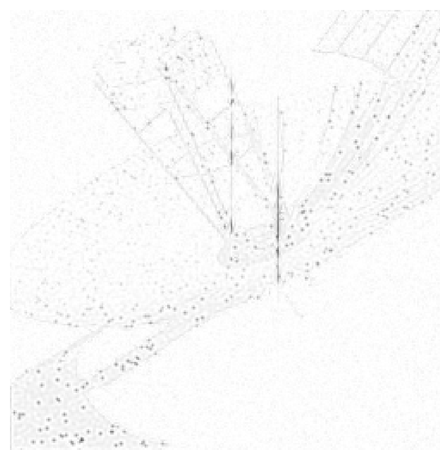


Validation image

2-norm error

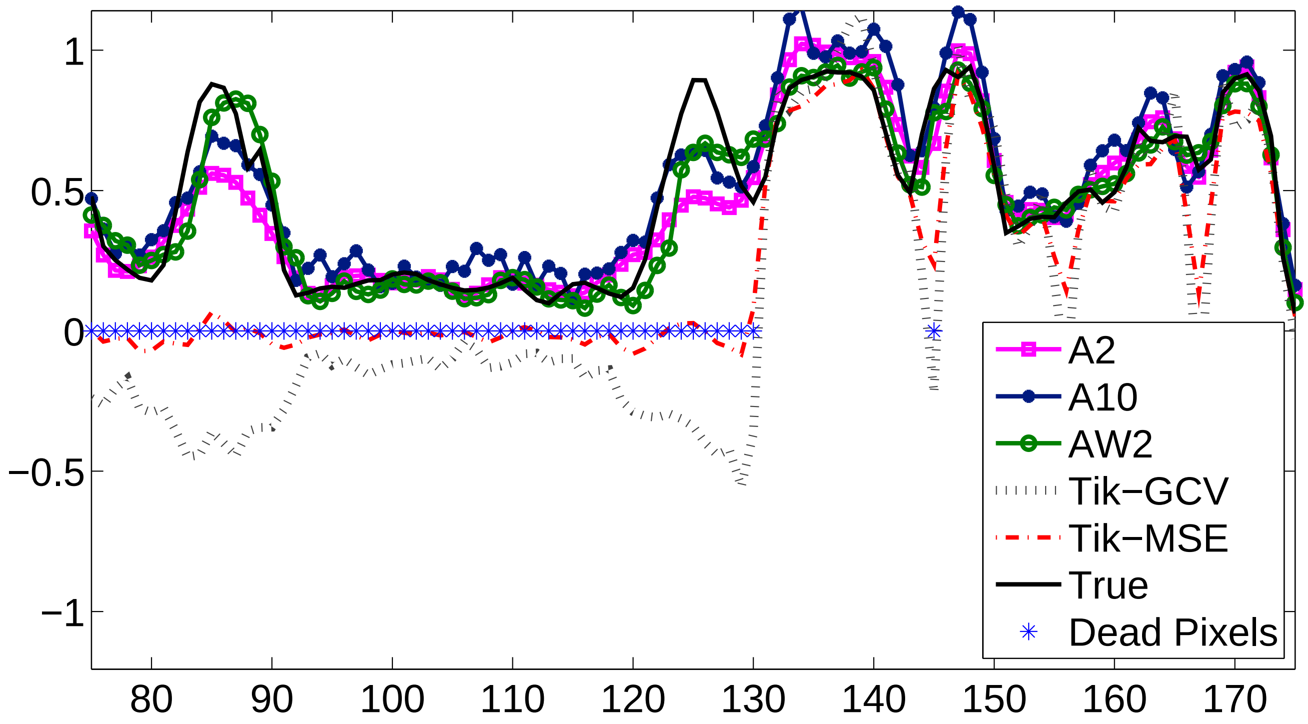


Tikhonov-GCV



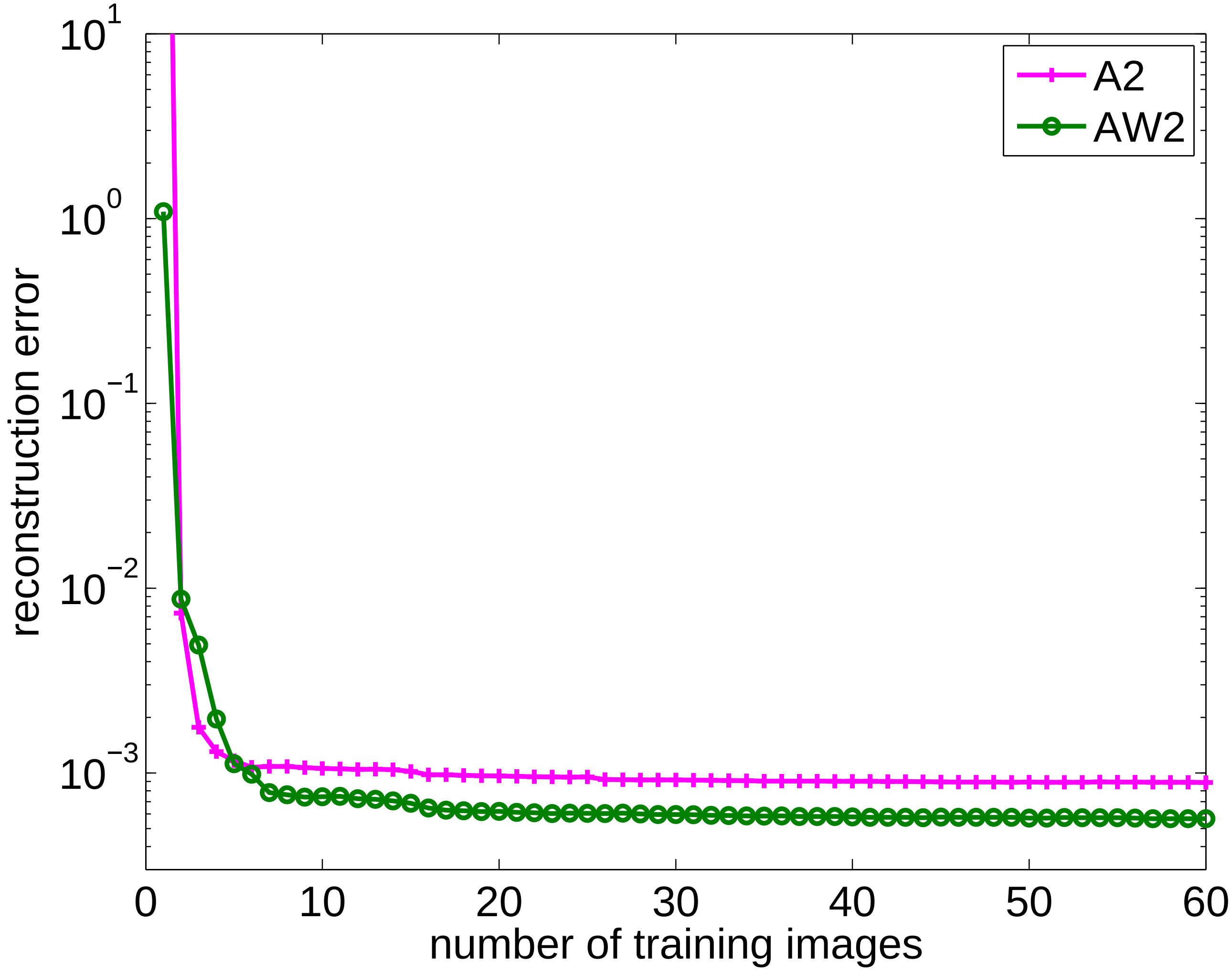
Tikhonov-MSE (not computable)

View down a single column of the image

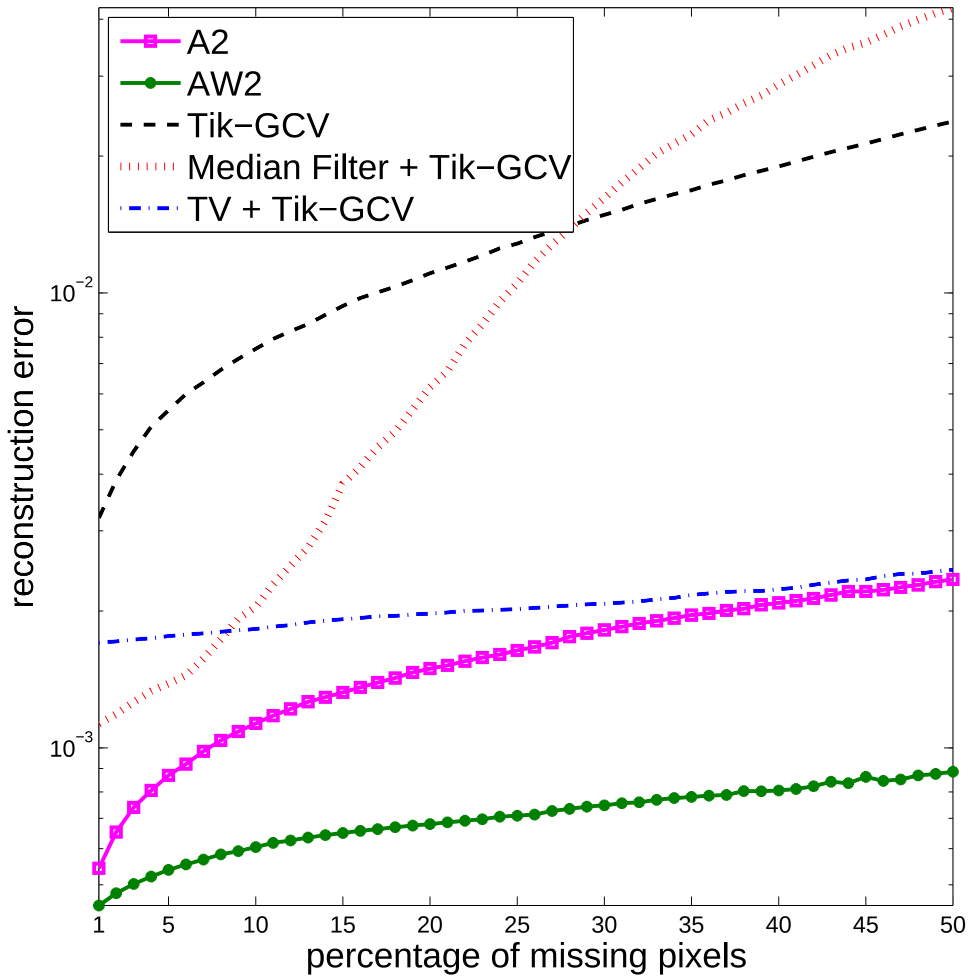


Results even better with noise post-processing.

Median reconstruction errors vs. number of training images



Results with unlearned missing pixels



The Plan

- The Problem
- Spectral Filtering
- Learning the Filter: Data to the Rescue
- Judging Goodness
- Example
- Conclusions

Conclusions

- Computing regularization parameters for ill-posed problems is generally difficult.
- We developed an optimal filtering approach for spectral regularization.
- Our formulation uses empirical Bayes risk minimization.
- A variety of error measures and filter representations are considered.
- Optimal filters are computed off-line.
- Reconstructions of test problems are very good.

Julianne Chung, Matthias Chung, and Dianne P. O'Leary
"Designing Optimal Spectral Filters for Inverse Problems"
SIAM Journal on Scientific Computing
33(6) (2011) pp. 3132-3152

Julianne M. Chung, Matthias Chung, and Dianne P. O'Leary
"Optimal Filters from Calibration Data for Image Deconvolution with Data Acquisition Errors,"
Journal of Mathematical Imaging and Vision,
44(3), pp. 366–374 (2012) DOI:10.1007/s10851-012-0332-4.

Julianne Chung   Matthias Chung