Regularization Parameter Estimation for Least Squares: Using the χ^2-curve

Rosemary Renaut, Jodi Mead
Supported by NSF

Arizona State and Boise State

Harrachov, August 2007

Outline

- Ill-posed system: $A \in \mathcal{R}^{m\times n}$, $\mathbf{b} \in \mathcal{R}^m$, $\mathbf{x} \in \mathcal{R}^n$
- Generalized Tikhonov regularization with operator D on $\mathbf{x}$.

$$\hat{\mathbf{x}} = \operatorname{argmin} J(\mathbf{x}) = \operatorname{argmin}\{\|A\mathbf{x}-\mathbf{b}\|^2_{W_{\mathbf{b}}} + \|D(\mathbf{x}-\mathbf{x}_0)\|^2_{W_{\mathbf{x}}}\}. \quad (1)$$

Assume $\mathcal{N}(A) \cap \mathcal{N}(D) = \emptyset$

- Statistically $W_{\mathbf{b}}$ is inverse covariance matrix for data $\mathbf{b}$.
- Standard: $W_{\mathbf{x}} = \lambda^2 I$, λ unknown penalty parameter

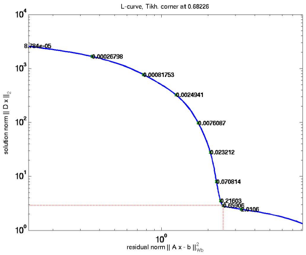
Focus: How to find λ?

Standard Methods I: L-curve - *Find the corner*

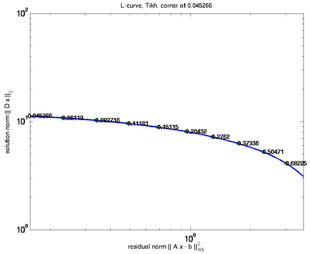
1. Let $\mathbf{r}(\lambda) = (A(\lambda) - A)\mathbf{b}$, where Influence Matrix $A(\lambda) = A(A^T W_{\mathbf{b}} A + \lambda^2 D^T D)^{-1} A^T$ Plot

$$\log(\|D\mathbf{x}\|), \log(\|\mathbf{r}(\lambda)\|)$$

2. Trade off contributions.
3. Expensive - requires range of λ.
4. GSVD makes calculations *efficient*.
5. No statistical information.



Find corner



No corner

Standard Methods II: Generalized Cross-Validation (GCV)

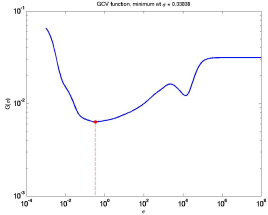
1. Minimizes GCV function

$$\frac{\|\mathbf{b} - A\mathbf{x}(\lambda)\|^2_{W_\mathbf{b}}}{[\mathrm{trace}(I_m - A(W_\mathbf{x}))]^2},$$

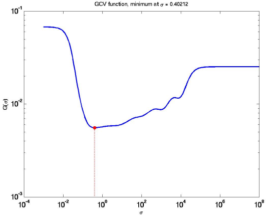
$$W_\mathbf{x} = \lambda^{-2} I_n$$

   which estimates predictive risk.

2. Expensive - requires range of λ.
3. GSVD makes calculations *efficient*.
4. Uses statistical information.



Multiple minima

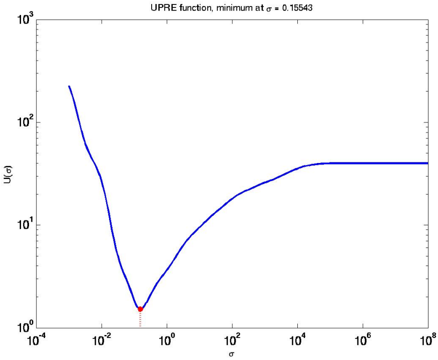


Sometimes flat

1. Minimize expected value of predictive risk: Minimize UPRE function

$$\|\mathbf{b} - A\mathbf{x}(\lambda)\|^2_{W_\mathbf{b}} + 2\,\mathrm{trace}(A(W_\mathbf{x})) - m$$

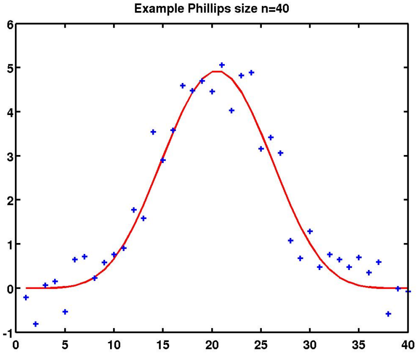
2. Expensive - requires range of λ.
3. GSVD makes calculations *efficient*.
4. Uses statistical information.
5. **Minimum needed**



An Illustrative Example: **phillips** Fredholm integral equation (Hansen)

1. Add noise to **b**
2. Standard deviation $\sigma_{b_i} = .01|b_i| + .1 b_{\max}$
3. Covariance matrix $C_{\mathbf{b}} = \sigma_{\mathbf{b}}^2 I_m = W_{\mathbf{b}}{}^{-1}$
4. $\sigma_{\mathbf{b}}^2$ average of $\sigma_{b_i}^2$
5. — is the original **b** and $*$ noisy data.

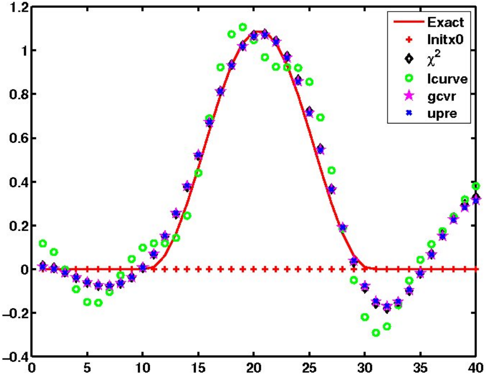
Example Error 10%



An Illustrative Example: **phillips** Fredholm integral equation (Hansen)

1. Add noise to **b**
2. Standard deviation $\sigma_{b_i} = .01|b_i| + .1 b_{\max}$
3. Covariance matrix $C_{\mathbf{b}} = \sigma_{\mathbf{b}}^2 I_m = W_{\mathbf{b}}{}^{-1}$
4. $\sigma_{\mathbf{b}}^2$ average of $\sigma_{b_i}^2$
5. $-$ is the original **b** and $*$ noisy data.
6. Each method gives different solution: o is L-curve
7. $+$ is reference

Comparison with new method



Theorem (Rao:73, Tarantola, Mead (2007))

$$J(\mathbf{x}) = (\mathbf{b} - A\mathbf{x})^T {C_\mathbf{b}}^{-1}(\mathbf{b} - A\mathbf{x}) + (\mathbf{x} - \mathbf{x}_0)^T {C_\mathbf{x}}^{-1}(\mathbf{x} - \mathbf{x}_0),$$

- $\mathbf{x}$ *and* $\mathbf{b}$ *are stochastic (need not be normal)*
- $\mathbf{r} = \mathbf{b} - A\mathbf{x}_0$ *are iid.*
- *Matrices* $C_\mathbf{b} = {W_\mathbf{b}}^{-1}$ *and* $C_\mathbf{x} = {W_\mathbf{x}}^{-1}$ *are SPD -*
- *Then for large m,*
  - *minimium value of J is a random variable*
  - *it follows a* χ^2 *distribution with* m *degrees of freedom.*

Implication: Find $W_{\mathbf{x}}$ such that J is χ^2 r.v.

- Theorem implies

$$m - \sqrt{2}z_{\alpha/2} < J(\hat{\mathbf{x}}) < m + \sqrt{2}z_{\alpha/2}$$

for confidence interval $(1-\alpha)$, $\hat{\mathbf{x}}$ the solution.

- Equivalently, when $D = I$,

$$m - \sqrt{2}z_{\alpha/2} < \mathbf{r}^T(AC_{\mathbf{x}}A^T + C_{\mathbf{b}})^{-1}\mathbf{r} < m + \sqrt{2}z_{\alpha/2}.$$

- Having found $W_{\mathbf{x}}$ posterior inverse covariance matrix is

$$\tilde{W}_{\mathbf{x}} = A^T W_{\mathbf{b}} A + W_{\mathbf{x}}$$

Note that $W_{\mathbf{x}}$ is completely general

Algorithm (Mead 07)

Given confidence interval parameter α*, initial residual* $\mathbf{r} = \mathbf{b} - A\mathbf{x}_0$ *and estimate of the data covariance* $C_{\mathbf{b}}$*, find* $L_{\mathbf{x}}$ *which solves the nonlinear optimization.*

Minimize	$\|L_{\mathbf{x}} {L_{\mathbf{x}}}^T\|_F^2$
Subject to	$m - \sqrt{2} z_{\alpha/2} < \mathbf{r}^T (A L_{\mathbf{x}} {L_{\mathbf{x}}}^T A^T + C_{\mathbf{b}})^{-1} \mathbf{r} < m + \sqrt{2} z_{\alpha/2}$
	$A L_{\mathbf{x}} {L_{\mathbf{x}}}^T A^T + C_{\mathbf{b}}$ well-conditioned.

Expensive

1. Let $W_{\mathbf{x}} = \sigma_{\mathbf{x}}^{-2} I$, where regularization parameter $\lambda = 1/\sigma_{\mathbf{x}}$.
2. Use SVD to implement $U_{\mathbf{b}} \Sigma_{\mathbf{b}} V_{\mathbf{b}}^T = {W_{\mathbf{b}}}^{1/2} A$, svs $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_p$ and define $\mathbf{s} = U_{\mathbf{b}} {W_{\mathbf{b}}}^{1/2} \mathbf{r}$:
3. Find $\sigma_{\mathbf{x}}$ such that

$$m - \sqrt{2} z_{\alpha/2} < \mathbf{s}^T \mathrm{diag}\left(\frac{1}{\sigma_i^2 \sigma_{\mathbf{x}}^2 + 1}\right) \mathbf{s} < m + \sqrt{2} z_{\alpha/2}.$$

4. Equivalently, find $\sigma_{\mathbf{x}}^2$ such that

$$F(\sigma_{\mathbf{x}}) = \mathbf{s}^T \mathrm{diag}\left(\frac{1}{1 + \sigma_{\mathbf{x}}^2 \sigma_i^2}\right) \mathbf{s} - m = 0.$$

Scalar Root Finding: Newton's Method

Define

$$\hat{\mathbf{x}}_{\mathrm{GTik}} = \mathrm{argmin} J_D(\mathbf{x}) = \mathrm{argmin}\{\|A\mathbf{x}-\mathbf{b}\|^2_{W_{\mathbf{b}}} + \|D(\mathbf{x}-\mathbf{x}_0)\|^2_{W_{\mathbf{x}}}\}, \quad (2)$$

Theorem

For large m, the minimium value of J_D is a random variable which follows a χ^2 distribution with $m-n+p$ degrees of freedom.

Proof.

Use the Generalized Singular Value Decomposition for $[{W_{\mathbf{b}}}^{1/2}A, {W_{\mathbf{x}}}^{1/2}D]$ □

Find $W_{\mathbf{x}}$ such that J_D is χ^2 with $m-n+p$ d.o.f.

Newton Root Finding $W_{\mathbf{x}} = \sigma_{\mathbf{x}}^{-2} I_p$

- GSVD of $[W_{\mathbf{b}}{}^{1/2} A, D]$

$$A = U \begin{bmatrix} \Upsilon \\ 0_{m-n \times n} \end{bmatrix} X^T \quad D = V[M, 0_{p \times n-p}] X^T,$$

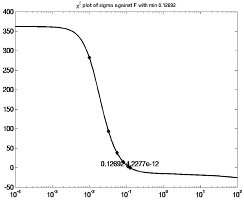
- γ_i are the generalized singular values
- $\tilde{m} = m - n + p - \sum_{i=1}^{p} s_i^2 \delta_{\gamma_i 0} - \sum_{i=n+1}^{m} s_i^2$,
- $\tilde{s}_i = s_i / (\gamma_i^2 \sigma_{\mathbf{x}}^2 + 1)$, $i = 1, \dots, p$
- $t_i = \tilde{s}_i \gamma_i$.

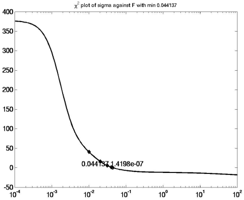
Solve $F = 0$**, where**

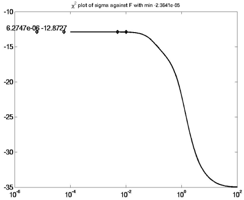
$$F(\sigma_{\mathbf{x}}) = \mathbf{s}^T \tilde{\mathbf{s}} - \tilde{m} \quad \text{and } F'(\sigma_{\mathbf{x}}) = -2\sigma_{\mathbf{x}} \|\mathbf{t}\|_2^2.$$

Observations: Example F

- Initialization GCV, UPRE, L-curve, χ^2 all use GSVD (or SVD).
- Algorithm is cheap as compared to GCV, UPRE, L-curve.
- F is **monotonic decreasing**, even
- Solution either exists and is **unique** for positive σ
- **Or no solution exists $F(0) < 0$.**







Relationship to Discrepancy Principle

- The discrepancy principle can be implemented by a Newton method.
- Finds $\sigma_{\mathbf{x}}$ such that the regularized residual satisfies

$$\sigma_{\mathbf{b}}^2 = \frac{1}{m}\|\mathbf{b} - A\mathbf{x}(\sigma)\|_2^2. \tag{3}$$

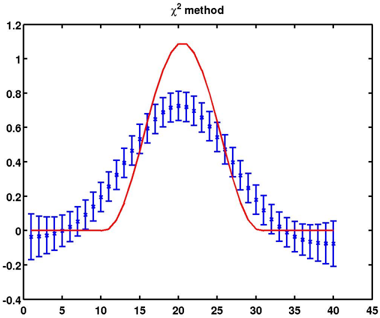
- Consistent with our notation

$$\sum_{i=1}^{p}\left(\frac{1}{\gamma_i^2\sigma^2+1}\right)^2 s_i^2 + \sum_{i=n+1}^{m} s_i^2 = m, \tag{4}$$

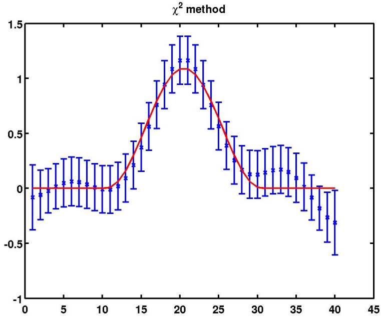
- Similar but note that the weight in the first sum is squared in this case.

Some Solutions: with no prior information $\mathbf{x}_0$

Illustrated are solutions and error bars



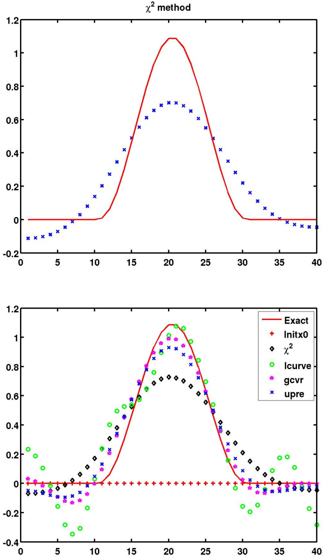
No Statistical Information
Solution is Smoothed



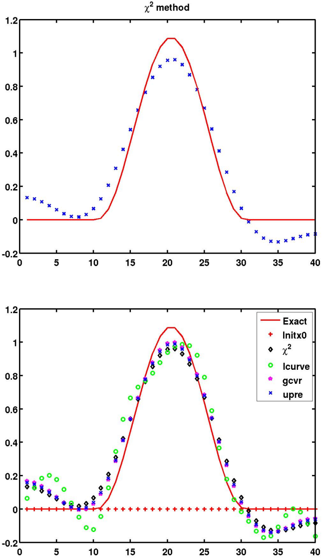
With statistical information
$C_{\mathbf{b}} = \mathrm{diag}(\sigma^2_{b_i})$

Some Generalized Tikhonov Solutions: First Order Derivative

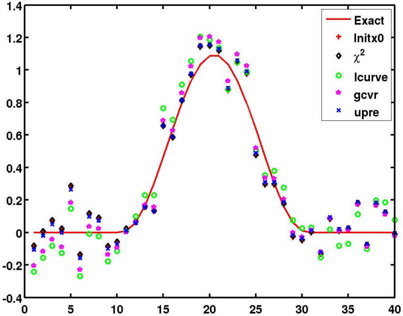
No Statistical Information



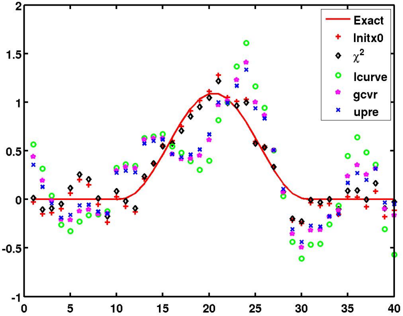
$C_{\mathbf{b}} = \mathrm{diag}(\sigma_{b_i}^2)$



No Statistical Information

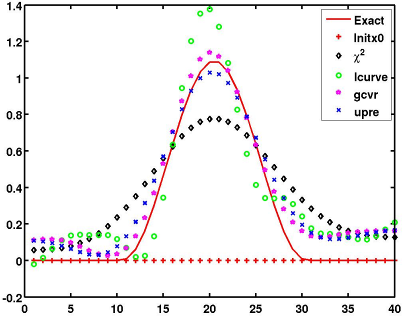


$C_{\mathbf{b}} = \mathrm{diag}(\sigma_{b_i}^2)$

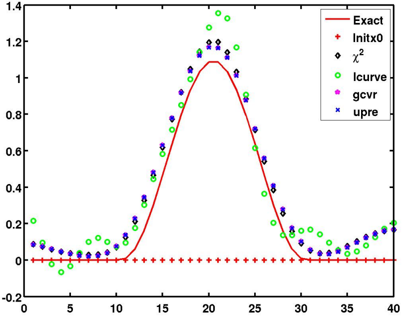


Some Generalized Tikhonov Solutions: $\mathbf{x}_0 = 0$: Exponential noise

No Statistical Information



$C_{\mathbf{b}} = \mathrm{diag}(\sigma_{b_i}^2)$



ARIZONA STATE UNIVERSITY

Newton's Method converges in 5 − 10 Iterations

l	cb	Iterations k	
		mean	std
0	1	$8.23e+00$	$6.64e-01$
0	2	$8.31e+00$	$9.80e-01$
0	3	$8.06e+00$	$1.06e+00$
1	1	$4.92e+00$	$5.10e-01$
1	2	$1.00e+01$	$1.16e+00$
1	3	$1.00e+01$	$1.19e+00$
2	1	$5.01e+00$	$8.90e-01$
2	2	$8.29e+00$	$1.48e+00$
2	3	$8.38e+00$	$1.50e+00$

Table: Convergence characteristics for problem phillips with $n = 40$ over 500 runs

Newton's Method converges in $5-10$ Iterations

l	cb	Iterations k	
		mean	std
0	1	$6.84e+00$	$1.28e+00$
0	2	$8.81e+00$	$1.36e+00$
0	3	$8.72e+00$	$1.46e+00$
1	1	$6.05e+00$	$1.30e+00$
1	2	$7.40e+00$	$7.68e-01$
1	3	$7.17e+00$	$8.12e-01$
2	1	$6.01e+00$	$1.40e+00$
2	2	$7.28e+00$	$8.22e-01$
2	3	$7.33e+00$	$8.66e-01$

Table: Convergence characteristics for problem blur with $n=36$ over 500 runs

		Error			
l	cb	χ^2	L	GCV	UPRE
		mean	mean	mean	mean
0	2	$4.37e-03$	$4.39e-03$	$4.21e-03$	$4.22e-03$
0	3	$4.32e-03$	$4.42e-03$	$4.21e-03$	$4.22e-03$
1	2	$4.35e-03$	$5.17e-03$	$4.30e-03$	$4.30e-03$
1	3	$4.39e-03$	$5.05e-03$	$4.38e-03$	$4.37e-03$
2	2	$4.50e-03$	$6.68e-03$	$4.39e-03$	$4.56e-03$
2	3	$4.37e-03$	$6.66e-03$	$4.43e-03$	$4.54e-03$

Table: Error characteristics for problem phillips with $n = 60$ over 500 runs with error contaminated $\mathbf{x}_0$. Relative errors larger than .009 removed.

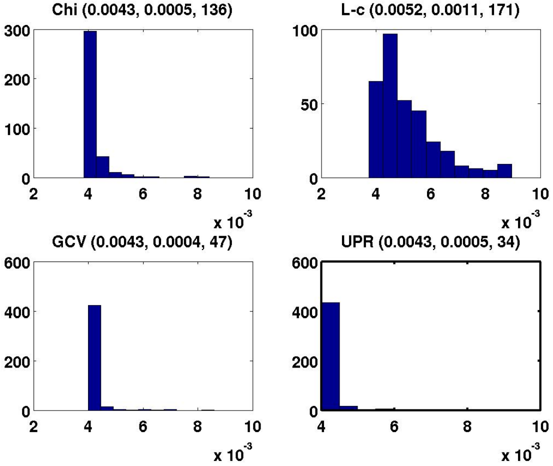
Results are comparable

		Risk			
l	cb	χ^2	L	GCV	UPRE
		mean	mean	mean	mean
0	2	$3.78e-02$	$5.22e-02$	$3.15e-02$	$2.92e-02$
0	3	$3.88e-02$	$5.10e-02$	$2.97e-02$	$2.90e-02$
1	2	$3.94e-02$	$5.71e-02$	$3.02e-02$	$2.74e-02$
1	3	$1.10e-01$	$5.90e-02$	$3.27e-02$	$2.79e-02$
2	2	$3.41e-02$	$6.00e-02$	$3.35e-02$	$3.79e-02$
2	3	$3.61e-02$	$5.98e-02$	$3.35e-02$	$3.82e-02$

Table: Error characteristics for problem phillips with $n = 60$ over 500 runs

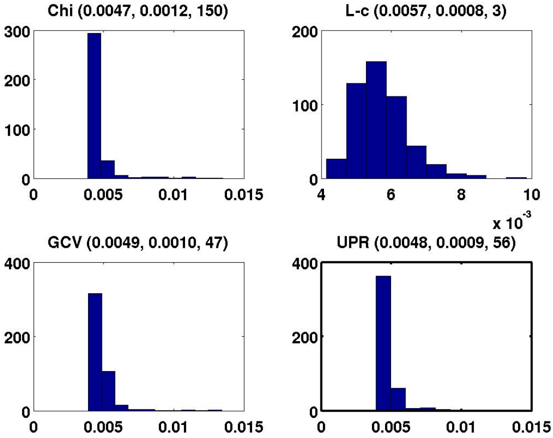
χ^2 method does not give best estimate of risk

Error Histogram



Normal noise on rhs, first order derivative, $C_{\mathbf{b}} = \sigma^2 I$

Error Histogram



Exponential noise on rhs, first order derivative, $C_{\mathbf{b}} = \sigma^2 I$

Conclusions

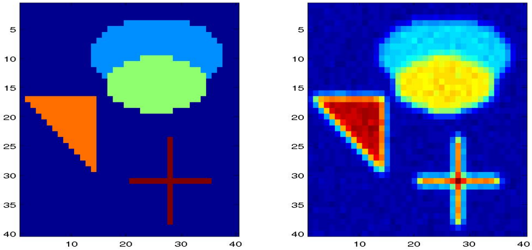
- χ^2 Newton algorithm is cost effective
- It performs as well (or better) than GCV and UPRE when statistical information is available.
- Should be method of choice when statistical information is provided
- Method can be adapted to find $W_{\mathbf{b}}$ if $W_{\mathbf{x}}$ is provided.

Future Work

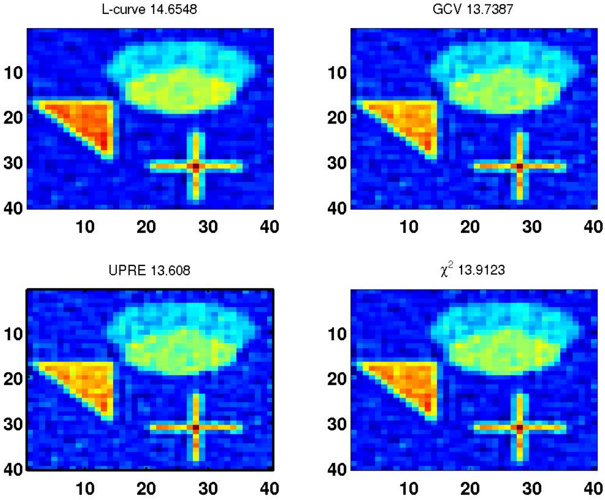
- Analyse for truncated expansions (TSVD and TGSVD) -reduce the degrees of freedom.
- Further theoretical analysis and simulations with other noise distributions.
- Can it be extended for nonlinear regularization terms? (TV?)
- Development of the nonlinear least squares for general diagonal $W_{\mathbf{x}}$.
- Efficient calculation of uncertainty information, covariance matrix.
- Nonlinear problems?

1. Bennett A, 2005 *Inverse Modeling of the Ocean and Atmosphere* (Cambridge University Press)
2. Hansen, P. C., 1994, Regularization Tools: A Matlab Package for Analysis and Solution of Discrete Ill-posed Problems, *Numerical Algorithms* **6** 1-35.
3. Mead J., 2007, A priori weighting for parameter estimation, *J. Inv. Ill-posed Problems*, to appear.
4. Rao, C. R., 1973, *Linear Statistical Inference and its applications*, Wiley, New York.
5. Tarantola A 2005 *Inverse Problem Theory and Methods for Model Parameter Estimation* (SIAM).
6. Vogel, C. R., 2002. *Computational Methods for Inverse Problems*, (SIAM), Frontiers in Applied Mathematics.

Solution on Left and Degraded on the Right



Solutions using $\mathbf{x}_0 = 0$, Generalized Tikhonov Second Derivative 5% noise



Solutions using $\mathbf{x}_0 = 0$, Generalized Tikhonov Second Derivative 10% noise