

The Chi-squared Distribution of the Regularized Least Squares Functional for Regularization Parameter Estimation

Rosemary Renaut

Collaborators: Jodi Mead and Iveta Hnetynkova



ARIZONA STATE UNIVERSITY

DEPARTMENT OF MATHEMATICS AND STATISTICS

Denmark 2008

Outline

- 1 Introduction and Motivation
 - Some Standard (or NOT) Statistical Methods for Regularization Parameter Estimation
- 2 Statistical Results for Least Squares
- 3 Implications of Statistical Results for Regularized Least Squares
- 4 Newton algorithm
- 5 Algorithm with LSQR
- 6 Results
- 7 Conclusions and Future Work

Least Squares for $A\mathbf{x} = \mathbf{b}$: A Quick Review

- Consider discrete systems: $A \in \mathcal{R}^{m \times n}$, $\mathbf{b} \in \mathcal{R}^m$, $\mathbf{x} \in \mathcal{R}^n$

$$A\mathbf{x} = \mathbf{b} + \mathbf{e},$$

- **Classical Approach** Linear Least Squares

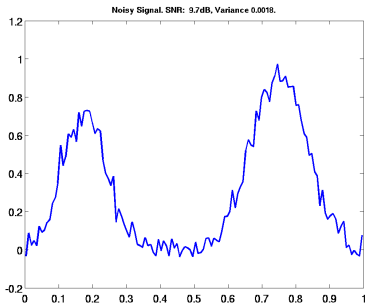
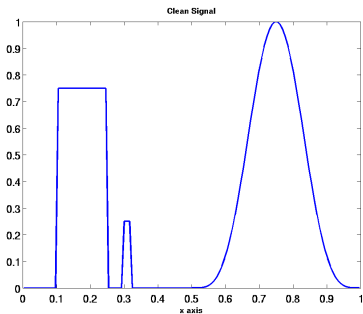
$$\mathbf{x}_{LS} = \arg \min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2$$

- **Difficulty** \mathbf{x}_{LS} is sensitive to changes in the right hand side \mathbf{b} when A is ill-conditioned.

System is numerically ill-posed.

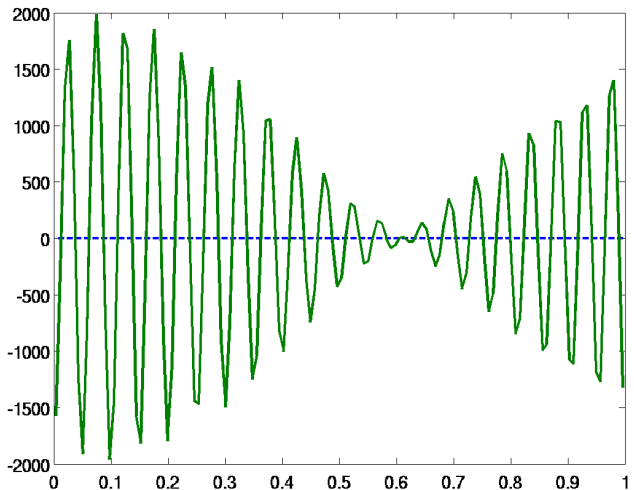
Example 1-D Signal: **Restore Signal**

1-D Original and Blurred Noisy Signal



Example 1-D Signal: Restore Signal

The Unregularized Solution: Illustrates Sensitivity



Alternative: Introduce a Mechanism for Regularization

Weighted Fidelity with Regularization

- Regularize

$$\mathbf{x}_{LS} = \arg \min_{\mathbf{x}} \{ \|\mathbf{b} - A\mathbf{x}\|_{W_b}^2 + \lambda^2 R(\mathbf{x}) \},$$

- Weighting matrix W_b is **inverse covariance matrix** for data \mathbf{b} .
- $R(\mathbf{x})$ is a regularization term
- λ is a regularization parameter which is unknown.
- Notice that the solution is $\mathbf{x}_{LS}(\lambda)$, dependent on λ . It also depends on choice of R .

Requirements

- Depends on R - what to chose?

A Specific Choice $R(\mathbf{x}) = \|D(\mathbf{x} - \mathbf{x}_0)\|^2$: Tikhonov Regularized

Generalized Tikhonov regularization: Given matrix D that is suitable.

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

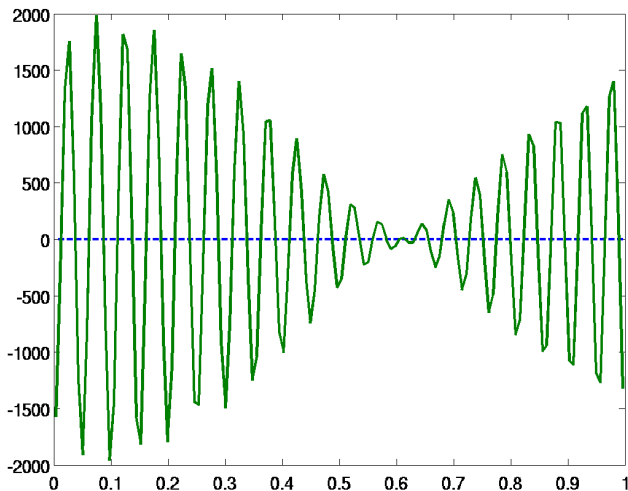
- Assume $\mathcal{N}(A) \cap \mathcal{N}(D) = \emptyset$
- \mathbf{x}_0 is a reference solution, often $\mathbf{x}_0 = 0$.
- Solution

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

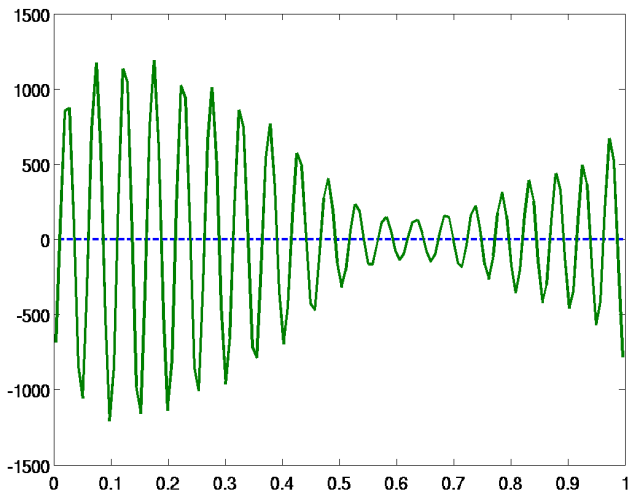
Question

Given D , how do we find λ ?
Choice of λ impacts the solution

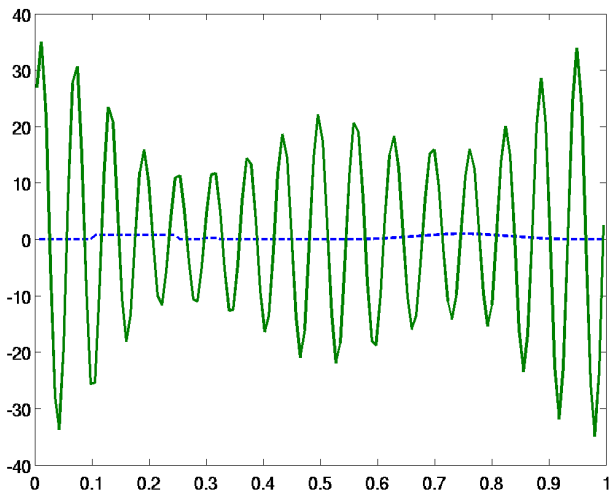
Solution for Increasing λ , $D = I$.



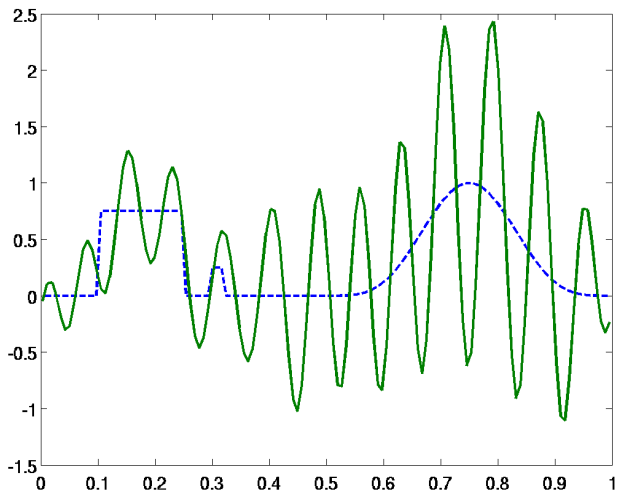
Solution for Increasing λ , $D = I$.



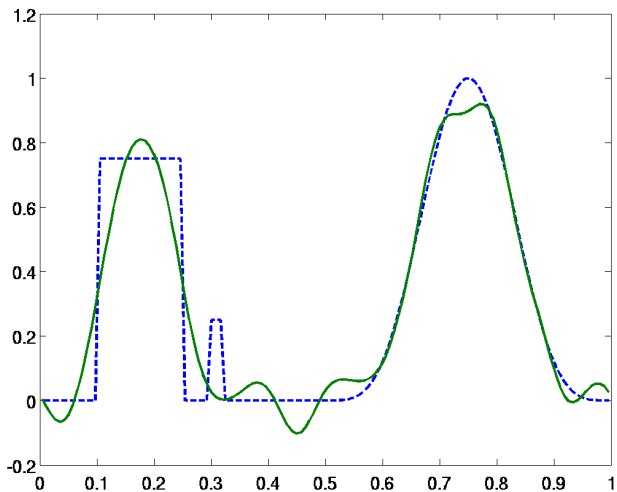
Solution for Increasing λ , $D = I$.



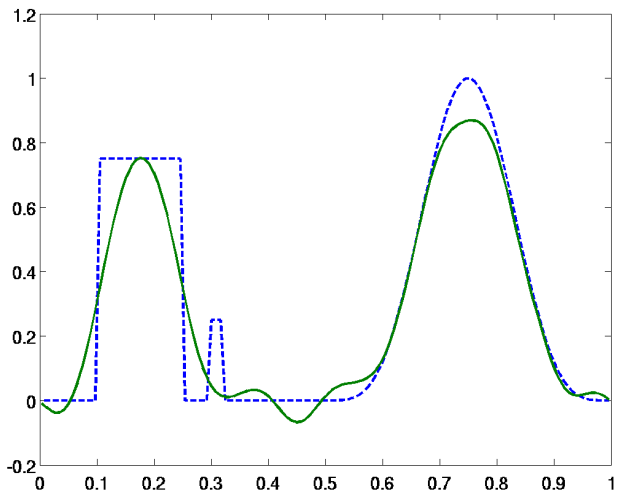
Solution for Increasing λ , $D = I$.



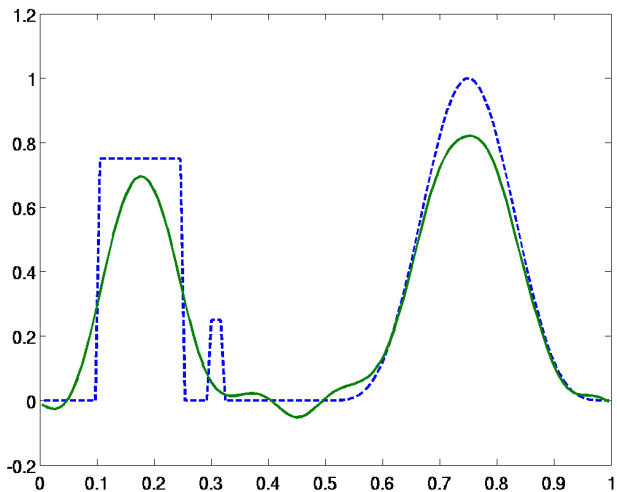
Solution for Increasing λ , $D = I$.



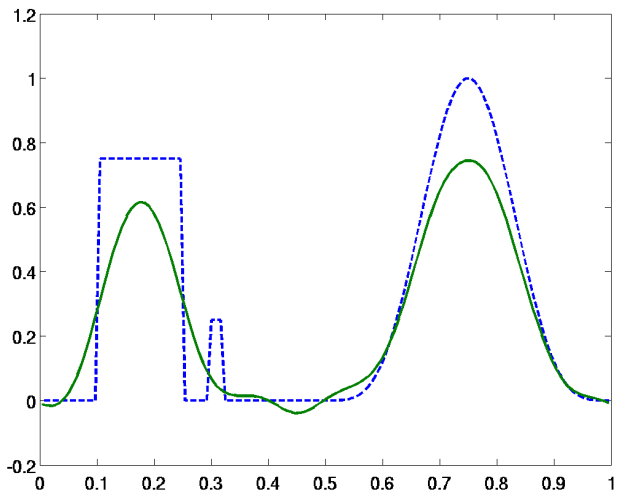
Solution for Increasing λ , $D = I$.



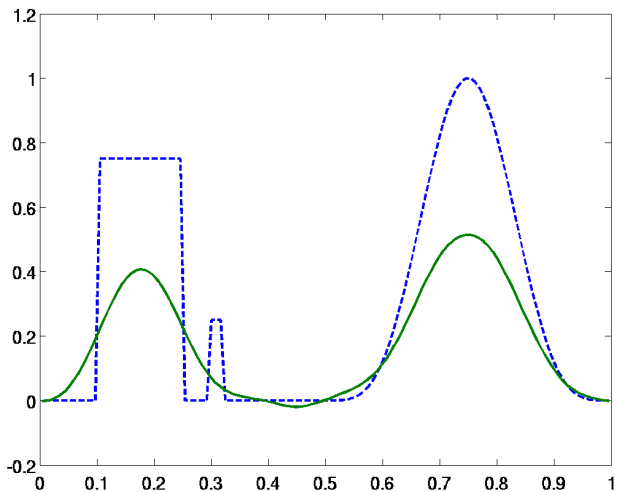
Solution for Increasing λ , $D = I$.



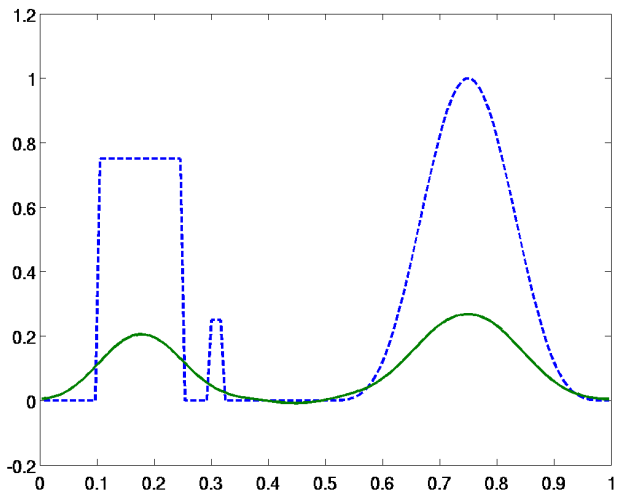
Solution for Increasing λ , $D = I$.



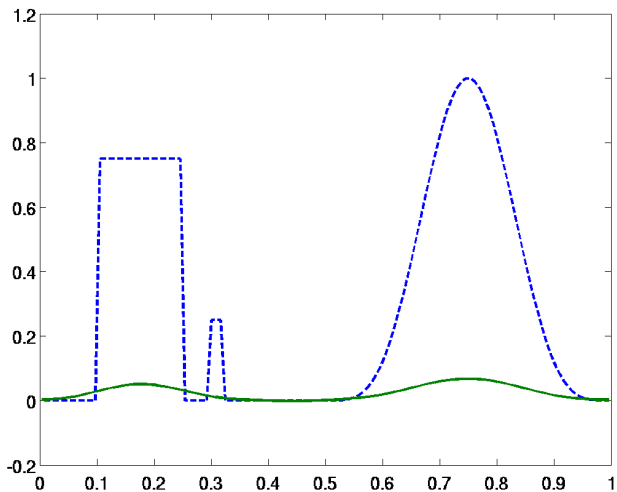
Solution for Increasing λ , $D = I$.



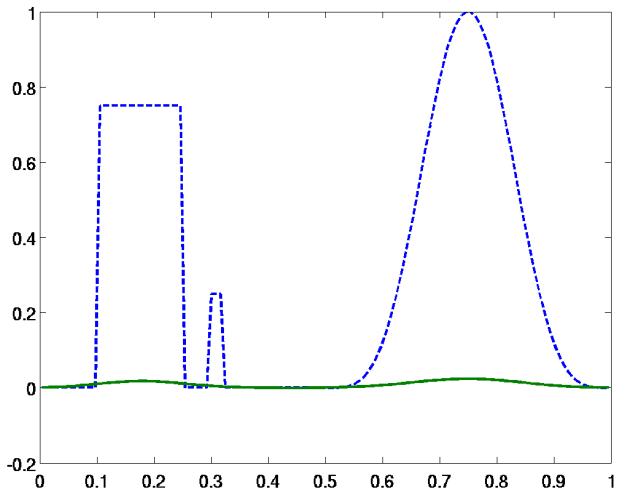
Solution for Increasing λ , $D = I$.



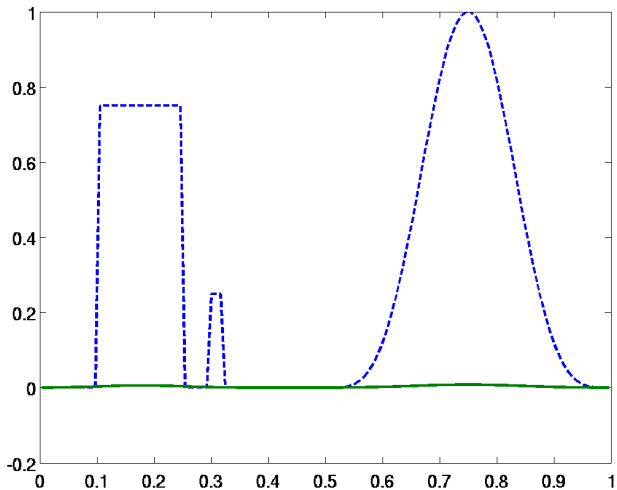
Solution for Increasing λ , $D = I$.



Solution for Increasing λ , $D = I$.



Solution for Increasing λ , $D = I$.



Choice of λ crucial

- Different algorithms yield different solutions.
- What is the **correct** choice?
- Use some **prior** information.
- But there is no one correct choice.

The Discrepancy Principle

- Suppose noise is **white**: $C_{\mathbf{b}} = \sigma_{\mathbf{b}}^2 I$.
- Find λ such that the regularized residual satisfies

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

- Can be implemented by a Newton root finding algorithm.
- But discrepancy principle typically oversmooths.

Generalized Cross-Validation (GCV)

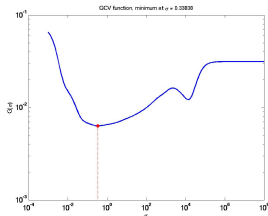
- Let

$$A(\lambda) = A(A^T W_b A + \lambda^2 D^T D)^{-1} A^T$$
- Can pick $W_b = I$.
- Minimize GCV function

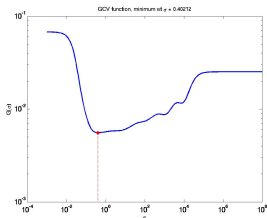
$$\frac{\|\mathbf{b} - A\mathbf{x}(\lambda)\|_{W_b}^2}{[\text{trace}(I_m - A(\lambda))]^2},$$

which estimates predictive risk.

- Expensive** - requires range of λ .
- GSVD makes calculations *efficient*.
- Requires minimum



Multiple minima



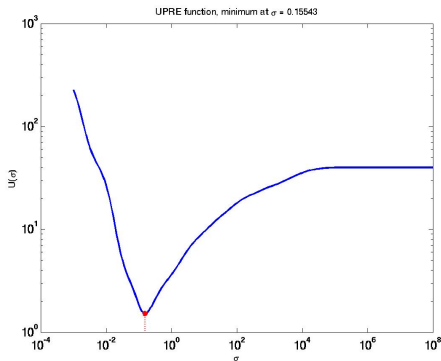
Sometimes flat

Unbiased Predictive Risk Estimation (UPRE)

- Minimize expected value of predictive risk: Minimize UPRE function

$$\| \mathbf{b} - A\mathbf{x}(\lambda) \|_{W_b}^2 + 2 \operatorname{trace}(A(\lambda)) - m$$

- Expensive** - requires range of λ .
- GSVD makes calculations *efficient*.
- Need estimate of **trace**
- Minimum needed**



Iterative Methods with Stopping Criteria

- Iterate to find approximate solution of $A\mathbf{x} \approx \mathbf{b}$.
- Introduce stopping criteria based on determining noise in the solution. e.g. Residual Periodogram (O'Leary and Rust).
- Hybrid LSQR iterate to reduce problem size.
- Stop when noise dominates Hnetynkova et al.
- Hybrid method - solve reduced system with additional regularization.
 - Cost of regularization of reduced system is minimal
 - **Advantage** any regularization may be used for subproblem. (Nagy etc)
 - How to find the appropriate regularization approach?
 - How to be sure when to stop the LSQR iteration?
 - How is statistics included in sub problem?

Include statistics directly

Background: Statistics of the Least Squares Problem

Theorem (Rao73: First Fundamental Theorem)

Let r be the rank of A and for $\mathbf{b} \sim N(A\mathbf{x}, \sigma_{\mathbf{b}}^2 I)$, (errors in measurements are normally distributed with mean 0 and covariance $\sigma_{\mathbf{b}}^2 I$), then

$$J = \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \sim \sigma_{\mathbf{b}}^2 \chi^2(m - r).$$

*J follows a χ^2 distribution with $m - r$ degrees of freedom:
Basically the Discrepancy Principle*

Corollary (Weighted Least Squares)

For $\mathbf{b} \sim N(A\mathbf{x}, C_{\mathbf{b}})$, and $W_{\mathbf{b}} = C_{\mathbf{b}}^{-1}$ then

$$J = \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_{W_{\mathbf{b}}}^2 \sim \chi^2(m - r).$$

Extension: Statistics of the Regularized Least Squares Problem

Two New Results to Help Find the Regularization parameter:

Theorem: χ^2 distribution of the regularized functional

$$\hat{\mathbf{x}} = \operatorname{argmin} J_D(\mathbf{x}) = \operatorname{argmin} \{ \|\mathbf{Ax} - \mathbf{b}\|_{\mathbf{W}_b}^2 + \|(\mathbf{x} - \mathbf{x}_0)\|_{\mathbf{W}_D}^2 \}, \quad \mathbf{W}_D = D^T \mathbf{W}_x D. \quad (4)$$

Assume

- \mathbf{W}_b and \mathbf{W}_x are symmetric positive definite.
- Problem is uniquely solvable $\mathcal{N}(A) \cap \mathcal{N}(D) \neq 0$.
- Moore-Penrose generalized inverse of \mathbf{W}_D is \mathbf{C}_D
- Statistics: $(\mathbf{b} - \mathbf{Ax}) = \mathbf{e} \sim N(0, \mathbf{C}_b)$, $(\mathbf{x} - \mathbf{x}_0) = \mathbf{f} \sim N(0, \mathbf{C}_D)$,
 - \mathbf{x}_0 is the mean vector of the model parameters.

Then

$$J_D \sim \chi^2(m + p - n)$$

Corollary: a-priori information not mean value, e.g. $\mathbf{x}_0 = 0$

Corollary: non-central χ^2 distribution of the regularized functional

$$\hat{\mathbf{x}} = \operatorname{argmin} J_D(\mathbf{x}) = \operatorname{argmin} \{ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_{\mathbf{W}_b}^2 + \|(\mathbf{x} - \mathbf{x}_0)\|_{\mathbf{W}_D}^2 \}, \quad \mathbf{W}_D = D^T \mathbf{W}_x D. \quad (5)$$

Assume all assumptions as before, but $\bar{\mathbf{x}} \neq \mathbf{x}_0$ is the mean vector of the model parameters.

Let

$$c = \|\mathbf{c}\|_2^2 = \|\tilde{Q}U^T \mathbf{W}_b^{1/2} \mathbf{A}(\bar{\mathbf{x}} - \mathbf{x}_0)\|_2^2$$

Then

$$J_D \sim \chi^2(m + p - n, c)$$

Implications of the Result

Statistical Distribution of the Functional

- Mean and Variance are prescribed

$$E(J_D) = m + p - n + c \quad E(J_D J_D^T) = 2(m + p - n) + 4c$$

- Can we use this?
- **YES**
- Try to find W_D so that $E(J) = m - n + p + 2c$
- Mead presented expensive nonlinear algorithm when $c = 0$.
- But it does find W_D .
- Our proposal - find λ only.

What do we need to apply the Theory?

Requirements

- **Covariance** information \mathbf{C}_b on data parameters \mathbf{b} (or on model parameters \mathbf{x} !)
- **A priori** information either \mathbf{x}_0 is the mean, or mean value $\bar{\mathbf{x}}$.
- But $\bar{\mathbf{x}}$ and \mathbf{x}_0 are not known.
- For repeated data measurements \mathbf{C}_b can be calculated. Also $\bar{\mathbf{b}}$ can be found, the mean of \mathbf{b} .
- But $E(\mathbf{b}) = AE(\mathbf{x})$ implies $\bar{\mathbf{b}} = A\bar{\mathbf{x}}$. Hence

$$c = \|\mathbf{c}\|_2^2 = \|\tilde{Q}U^T \mathbf{W}_b^{1/2}(\bar{\mathbf{b}} - A\mathbf{x}_0)\|_2^2$$

-

$$E(J_D) = E(\|\tilde{Q}U^T \mathbf{W}_b^{1/2}(\mathbf{b} - A\mathbf{x}_0)\|_2^2) = m+p-n + \|\tilde{Q}U^T \mathbf{W}_b^{1/2}(\bar{\mathbf{b}} - A\mathbf{x}_0)\|_2^2$$

Then we can use $E(J)$ to find λ

Assume \mathbf{x}_0 is the mean (experimentalists do know something about the model parameters)

DESIGNING THE ALGORITHM: I

- Recall: if \mathbf{C}_b and \mathbf{C}_x are good estimates of covariance

$$|J_D(\hat{\mathbf{x}}) - (m + p - n)|$$

should be **small**.

- Thus, let $\tilde{m} = m + p - n$ then we want

$$\tilde{m} - \sqrt{2\tilde{m}}z_{\alpha/2} < J(\mathbf{x}(W_D)) < \tilde{m} + \sqrt{2\tilde{m}}z_{\alpha/2}. \quad (6)$$

- $z_{\alpha/2}$ is the relevant z -value for a χ^2 -distribution with \tilde{m} degrees

GOAL

Find W_D to make (6) tight: Single Variable case find λ

$$J_D(\hat{\mathbf{x}}(\lambda)) \approx \tilde{m}$$

A Newton-line search Algorithm to find λ . (Basic algebra)

Newton to Solve $F(\sigma) = J_D(\sigma) - \tilde{m} = 0$

- We use $\sigma = 1/\lambda$, and $\mathbf{y}(\sigma^{(k)})$ is the current solution for which

$$\mathbf{x}(\sigma^{(k)}) = \mathbf{y}(\sigma^{(k)}) + \mathbf{x}_0$$

Then

$$\frac{\partial}{\partial \sigma} J(\sigma) = -\frac{2}{\sigma^3} \|\mathbf{D}\mathbf{y}(\sigma)\|^2 < 0$$

- Hence we have a basic Newton Iteration

$$\sigma^{(k+1)} = \sigma^{(k)} \left(1 + \frac{1}{2} \left(\frac{\sigma^{(k)}}{\|\mathbf{D}\mathbf{y}\|} \right)^2 (J_D(\sigma^{(k)}) - \tilde{m}) \right).$$

- Add a line search

$$\sigma^{(k+1)} = \sigma^{(k)} \left(1 + \frac{\alpha^{(k)}}{2} \left(\frac{\sigma^{(k)}}{\|\mathbf{D}\mathbf{y}\|} \right)^2 (J_D(\sigma^{(k)}) - \tilde{m}) \right).$$

Algorithm Using the GSVD

GSVD

- Use GSVD of $[W_{\mathbf{b}}^{1/2}A, D]$
- For γ_i the generalized singular values, and $\mathbf{s} = U^T W_{\mathbf{b}}^{1/2} \mathbf{r}$
- $\tilde{m} = m - n + p$
- $\tilde{s}_i = s_i / (\gamma_i^2 \sigma_{\mathbf{x}}^2 + 1)$, $i = 1, \dots, p$, $t_i = \tilde{s}_i \gamma_i$.
- Find root of

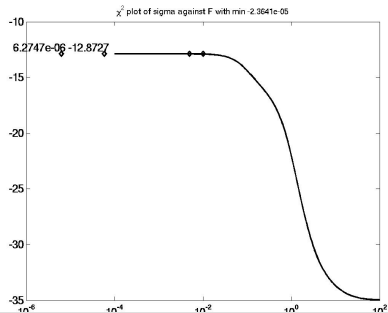
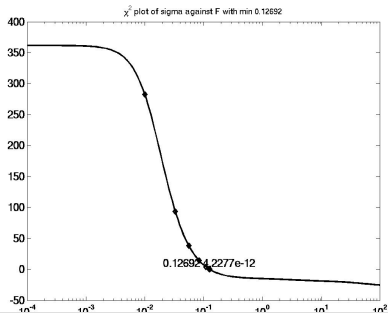
$$F(\sigma_{\mathbf{x}}) = \sum_{i=1}^p \left(\frac{1}{\gamma_i^2 \sigma_{\mathbf{x}}^2 + 1} \right) s_i^2 + \sum_{i=n+1}^m s_i^2 - \tilde{m} = 0$$

- Equivalently: solve $F = 0$, where

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

Discussion on Convergence

- F is **monotonic decreasing** ($F'(\sigma_{\mathbf{x}}) = -2\sigma_{\mathbf{x}}\|\mathbf{t}\|_2^2$)
- Solution either exists and is **unique** for positive σ
- **Or no solution exists** $F(0) < 0$.
 - implies incorrect statistics of the model
- Theoretically, $\lim_{\sigma \rightarrow \infty} F > 0$ possible.
 - Equivalent to $\lambda = 0$. No regularization needed.



Practical Details of Algorithm

Find the parameter

- **Step 1:** Bracket the root by logarithmic search on σ to handle the asymptotes: yields **sigmamax** and **sigmamin**
- **Step 2:** Calculate step, with steepness controlled by tolD. Let $\mathbf{t} = D\mathbf{y}/\sigma^{(k)}$, where \mathbf{y} is the current update, given from the GSVD, then

$$\text{step} = \frac{1}{2} \left(\frac{1}{\max \{ \|\mathbf{t}\|, \text{tolD} \}} \right)^2 (J_D(\sigma^{(k)}) - \tilde{m})$$

- **Step 3:** Introduce line search $\alpha^{(k)}$ in Newton

$$\text{sigmanew} = \sigma^{(k)} (1 + \alpha^{(k)} \text{step})$$

$\alpha^{(k)}$ chosen such that sigmanew within bracket.

Practical Details of Algorithm: Large Scale problems

Algorithm

Initialization

- Convert generalized Tikhonov problem to standard form.
- Use LSQR algorithm to find the bidiagonal matrix for the projected problem.
- Obtain a solution of the bidiagonal problem for given initial σ .

Subsequent Steps

- Increase dimension of space if needed with reuse of existing bidiagonalization.
- Each σ calculation of algorithm reuses saved information from the Lancos bidiagonalization.
- The system is augmented if needed. May also use smaller size system if appropriate.

Comparison with Standard LSQR hybrid Algorithm

- Algorithm concurrently regularizes and solves the system.
- Standard hybrid LSQR solves projected system then adds regularization.

Advantages

Costs

- Needs only cost of standard LSQR algorithm with some updates for solution solves for iterated σ .
- The regularization introduced by LSQR projection may be useful for preventing problems with GSVD expansion.
- Makes algorithm viable for large scale problems.

Recall: Implementation Assumptions

Covariance of Error: Statistics of Measurement Errors

- Information on the covariance structure of errors in \mathbf{b} needed.
- Use $\mathbf{C}_b = \sigma_b^2 I$ for common covariance, **white noise**.
- Use $\mathbf{C}_b = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_m^2)$ for **colored uncorrelated noise**.
- With no noise information $\mathbf{C}_b = I$.
- Use $\bar{\mathbf{b}}$ as the mean of measured \mathbf{b} , when implemented with centrality parameter, $\mathbf{x}_0 = 0$.

Illustrating the Results for Problem Size 512: Two Standard Test Problems

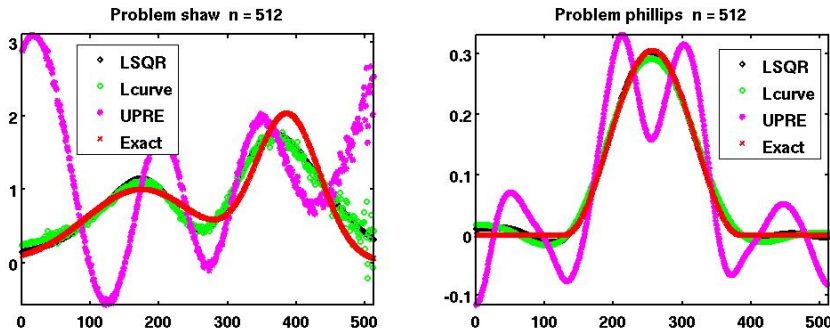


Figure: Comparison for noise level 10%. On left $D = I$ and on right D is first derivative

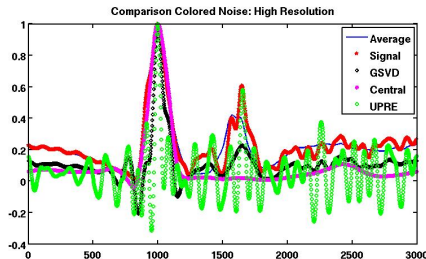
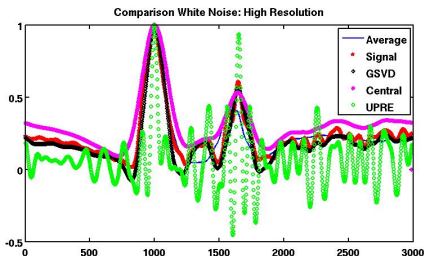
- Notice L-curve and χ^2 -LSQR perform well.
- UPRE does not perform well.

Real Data: Seismic Signal Restoration

The Data Set and Goal

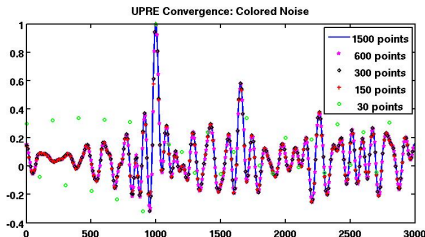
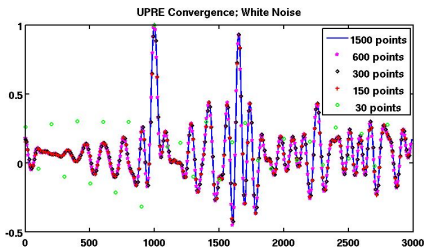
- Real data set of 48 signals of length 3000.
- The point spread function is derived from the signals.
- Calculate the signal variance pointwise over all 48 signals.
- Goal: restore the signal \mathbf{x} from $A\mathbf{x} = \mathbf{b}$, where A is psf matrix and \mathbf{b} is given blurred signal.
- Method of Comparison- no exact solution known: use convergence with respect to downsampling.

Comparison High Resolution White noise (left) and Colored Noise (right)



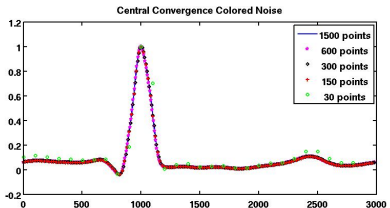
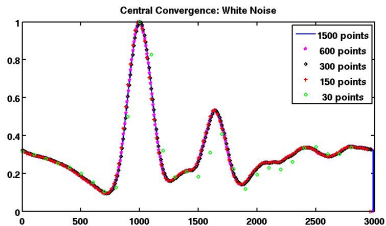
Greater contrast with χ^2 . UPRE is insufficiently regularized.
 L-curve severely undersmooths (not shown). Parameters not consistent across resolutions

THE UPRE SOLUTION: White Noise and Colored Noise $x_0 = 0$



Regularization Parameters are consistent: $\sigma = 0.01005$ all resolutions

THE LSQR Hybrid SOLUTION: White Noise (left) and Colored Noise (right) $x_0 = 0$



Regularization quite consistent resolution 2 to 100

$\sigma = 0.0000029, .0000029, .0000029, .0000057, .0000057$ (left)

$\sigma = 0.00007, .00007, .00007, .00007, .00012$ (right).

Notice that colored noise eliminates second arrival of signal but excellent contrast to identify primary arrival.

Sensitivity of LSQR to Parameter Choices

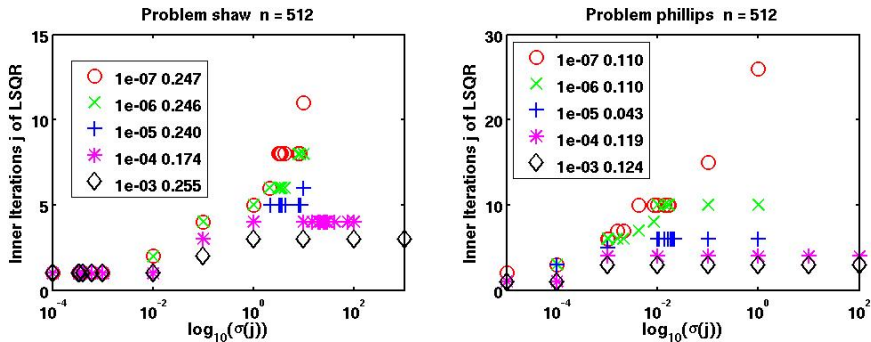


Figure: Illustrating the dependence of $j(\sigma)$ on σ for problem with noise level 10% for increasing inner tolerance in LSQR iteration.

- Subproblem size increases with increasing σ .
- Subproblem size decreases with decreasing tolerance.

Some Comparisons of LSQR and GSVD Implementations Problem size 64

shaw							
ϵ	Average Inner Steps	Newton Steps		P value		Relative Error	
		GSVD	LSQR	Iteration	σ		
$5e-02$	7.4	26.7	26.5	0.9	1.0	0.189	0.189
$1e-01$	6.9	26.5	25.8	0.8	1.0	0.215	0.215
ilaplace							
$5e-02$	4.7	25.6	30.8	0.9	1.0	0.161	0.162
$1e-01$	4.3	24.1	30.9	0.7	1.0	0.211	0.211
heat							
$5e-02$	14.8	26.1	30.3	1.0	1.0	0.304	0.304
$1e-01$	11.1	24.5	30.3	0.8	1.0	0.375	0.375
phillips							
$5e-02$	8.0	24.6	30.2	0.9	1.0	0.079	0.079
$1e-01$	7.0	25.4	30.5	0.9	1.0	0.112	0.112

Some Comparisons with Other Methods: Problem Size 128

shaw								
ϵ	Least Squares Error				Failure Count			
	GSVD	LSQR	LC	UPRE	GSVD	LSQR	LC	UPRE
$5e-02$	0.229	0.206	0.200	0.239	1	1	0	63
$1e-01$	0.276	0.236	0.241	0.259	5	4	0	71
ilaplace								
$5e-02$	0.200	0.204	0.166	0.195	22	18	0	71
$1e-01$	0.247	0.230	0.231	0.230	67	53	63	112
heat								
$5e-02$	0.327	0.324	<i>NaN</i>	0.305	46	40	250	21
$1e-01$	0.408	0.404	<i>NaN</i>	0.396	119	109	250	61
phillips								
$5e-02$	0.100	0.099	0.080	0.109	0	0	0	34
$1e-01$	0.137	0.136	0.123	0.139	2	2	0	28

Main Observations

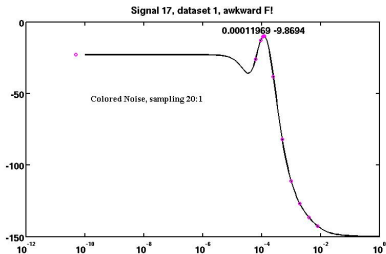
- GSVD and LSQR implementations consistent
- Method requires few iterations of root finding
- LSQR uses small bidiagonal system.
- LSQR is relatively robust to inner tolerance
- Increased σ implies reduced regularization by Tikhonov but increased regularization by LSQR.

Conclusions

Observations

- A new statistical method for estimating regularization parameter
 - Compares favorably with UPRE with respect to performance and compared to L-curve. (GCV is not competitive).
- Method can be used for large scale problems.
- Method is very efficient, Newton method is robust and fast.
- But a priori information is needed.

Difficulties when central parameter is required



What are the issues?

- Function need not be monotonic
- More problematic for NonCentral version with \mathbf{x}_0 not the mean. (ie $\mathbf{x}_0 = 0$.)
- σ can be bounded by result of central case.
- Range of σ given by range of γ_i .

Future Work

Other Results and Future Work

- Degrees of freedom reduced when using the GSVD.
- Image deblurring. (Implementation to use minimal storage)
- Diagonal Weighting Schemes
- Edge preserving regularization
- Constraint implementation (with Mead submitted).