

# Statistical Properties of the Regularized Least Squares Functional and a hybrid LSQR Newton method for Finding the Regularization Parameter: Application in Image Deblurring and Signal Restoration

Rosemary Renaut

Collaborators: Jodi Mead and Iveta Hnetynkova



ARIZONA STATE UNIVERSITY

DEPARTMENT OF MATHEMATICS AND STATISTICS

Boise 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

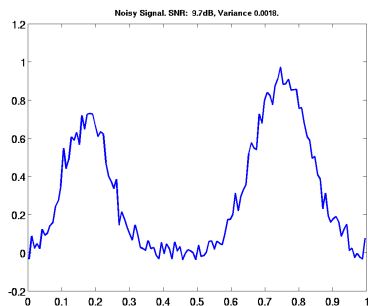
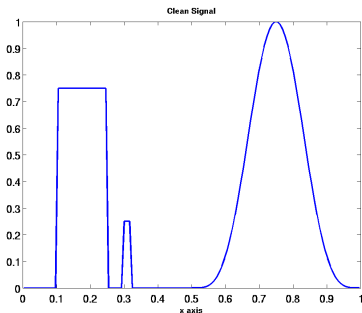
$$\mathbf{x}_{LS} = \arg \min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2$$

- **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



Original signal  $\mathbf{x}$ .

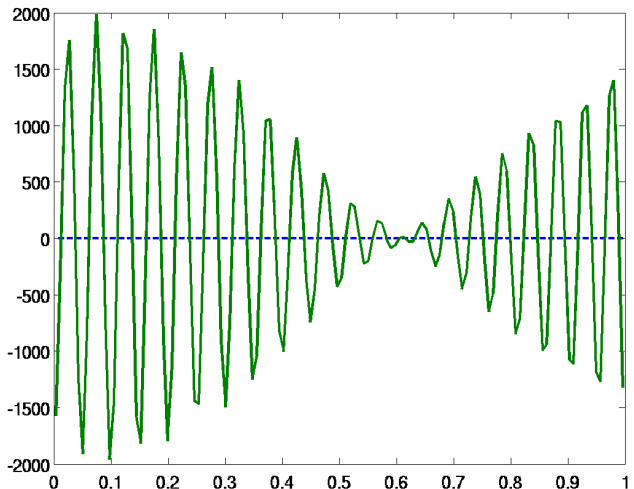
Signal  $\mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{n}$

$\mathbf{A}$  is discretization of point spread function (PSF), the blur.

$$b(t) = \int K(t-s)x(s)ds \quad \mathbf{A} \text{ discretizes } K \text{ PSF}$$

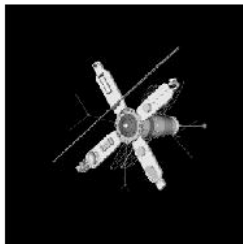
## Example 1-D Signal: Restore Signal

### The Unregularized Solution: Illustrates Sensitivity

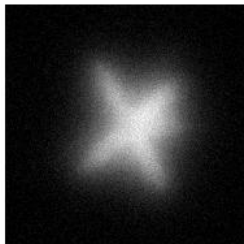


## Two dimensional Image Deblurring

True solution,  $x$



Noisy and Blurred Data



Deblurred



**Goal: Discuss how to do inversion to obtain deblurred image**

## Introduce Regularization to Pick a Solution

### 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
- $\lambda$  is a regularization parameter which is unknown.
- Notice that the solution is  $\mathbf{x}_{LS}(\lambda)$ , dependent on  $\lambda$ . 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.**

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

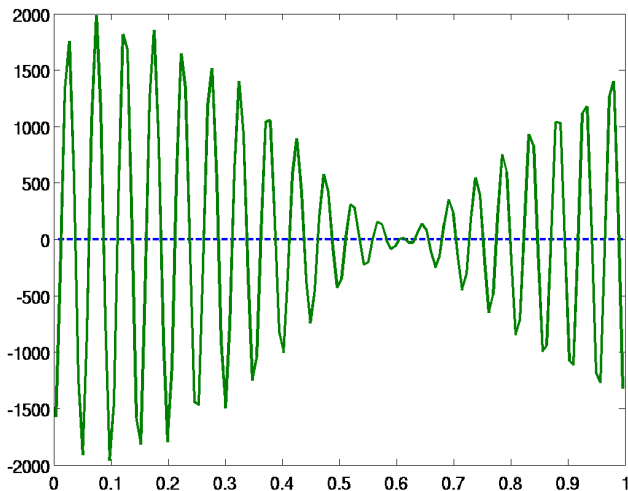
- Assume  $\mathcal{N}(A) \cap \mathcal{N}(D) = \emptyset$
- $\mathbf{x}_0$  is a reference solution, often  $\mathbf{x}_0 = 0$ .
- Solution

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

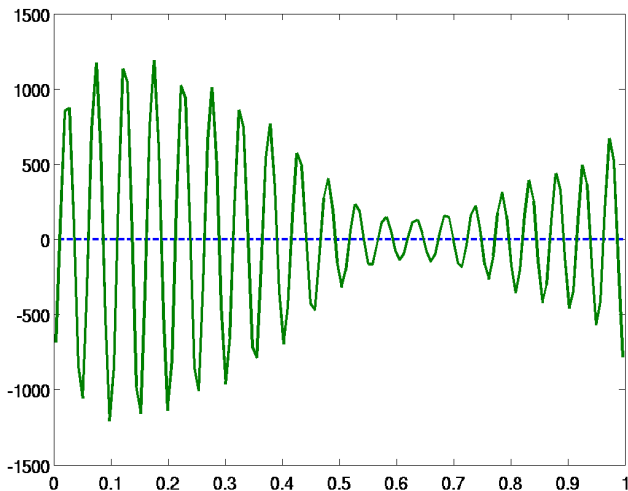
### Question

**Given  $D$ , how do we find  $\lambda$ ?**  
**Choice of  $\lambda$  impacts the solution**

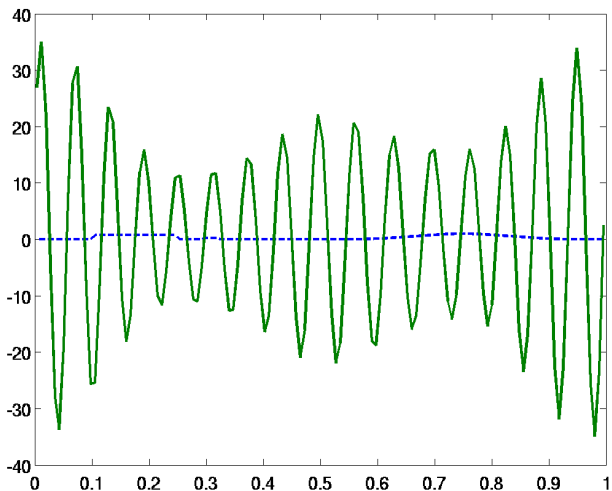
## Solution for Increasing $\lambda$ , $D = I$ .



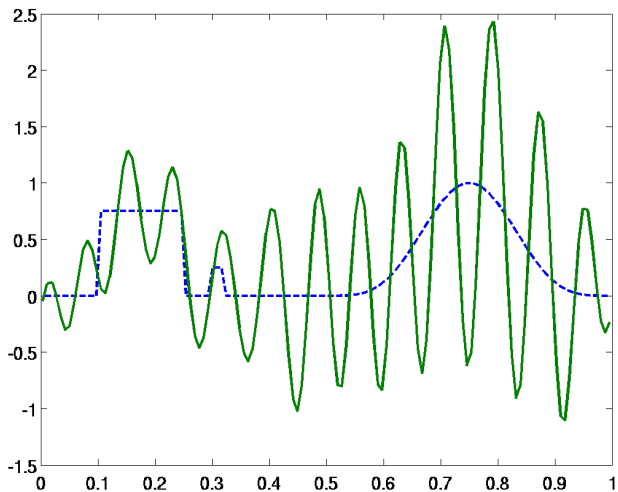
## Solution for Increasing $\lambda$ , $D = I$ .



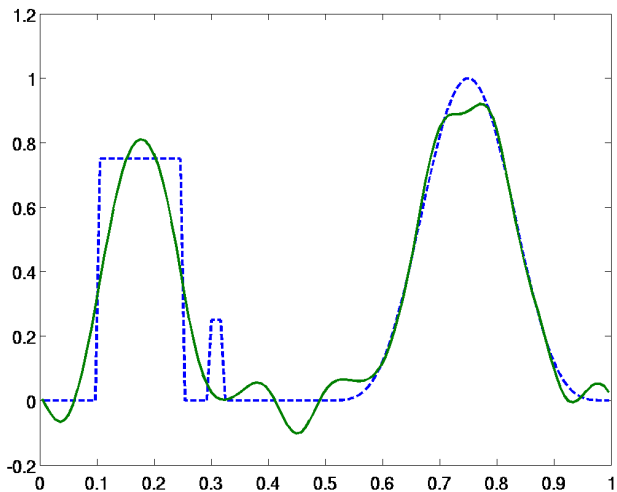
## Solution for Increasing $\lambda$ , $D = I$ .



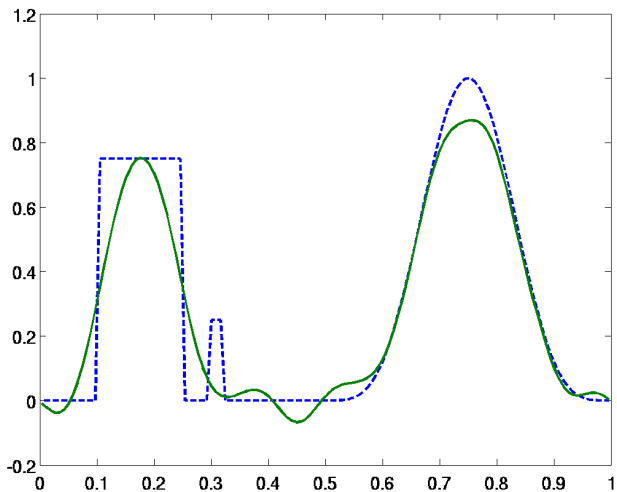
## Solution for Increasing $\lambda$ , $D = I$ .



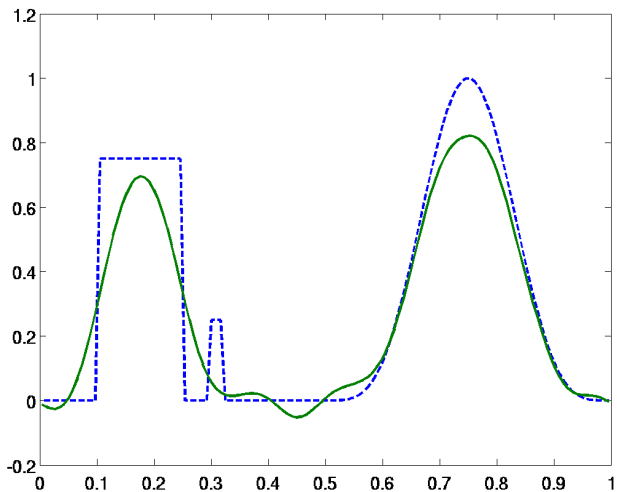
## Solution for Increasing $\lambda$ , $D = I$ .



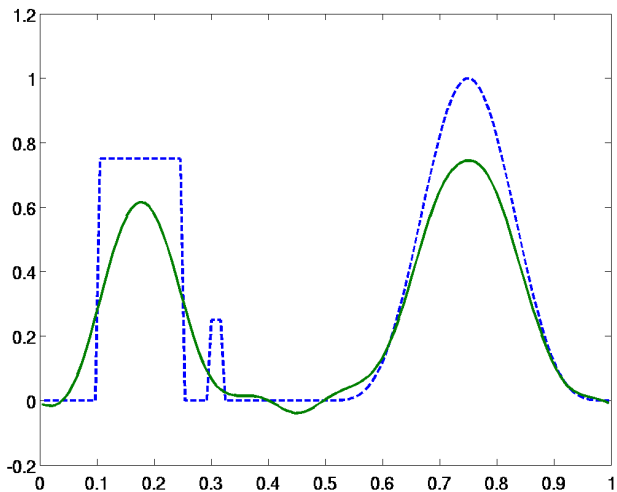
## Solution for Increasing $\lambda$ , $D = I$ .



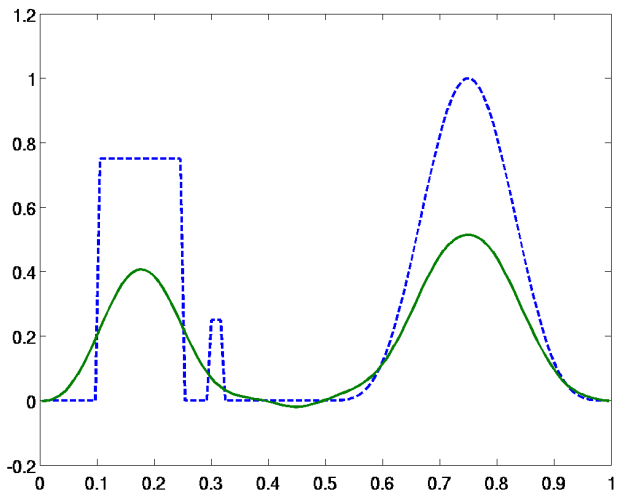
## Solution for Increasing $\lambda$ , $D = I$ .



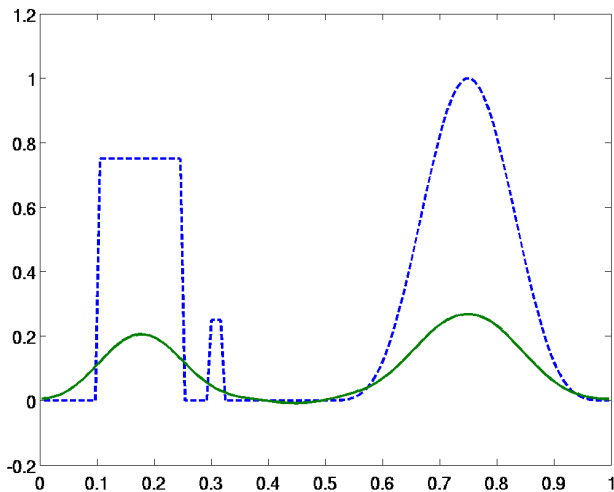
## Solution for Increasing $\lambda$ , $D = I$ .



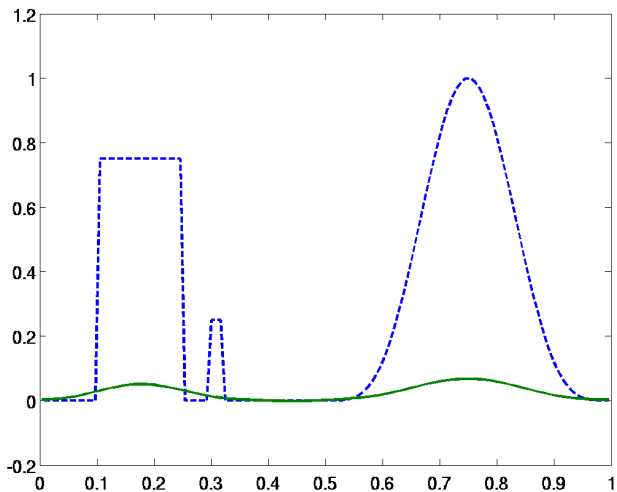
## Solution for Increasing $\lambda$ , $D = I$ .



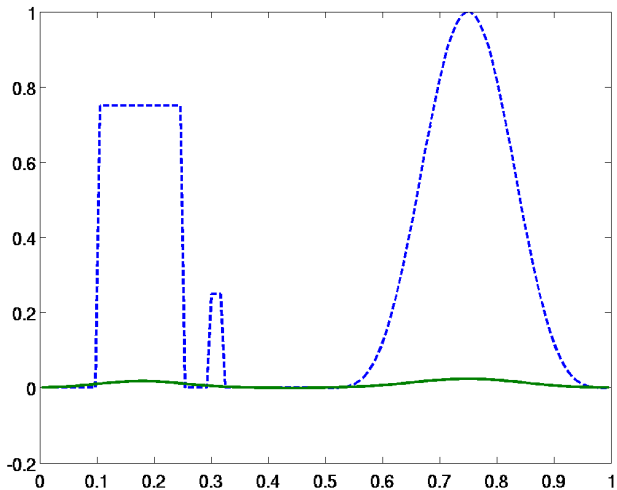
## Solution for Increasing $\lambda$ , $D = I$ .



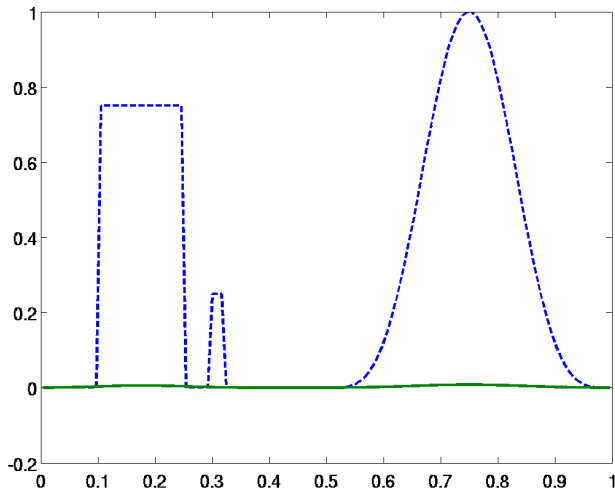
## Solution for Increasing $\lambda$ , $D = I$ .



## Solution for Increasing $\lambda$ , $D = I$ .



## Solution for Increasing $\lambda$ , $D = I$ .



## Choice of $\lambda$ 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  $\lambda$  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.

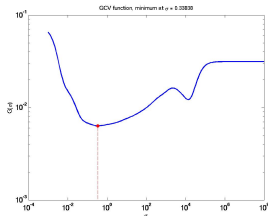
## Generalized Cross-Validation (GCV)

- 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

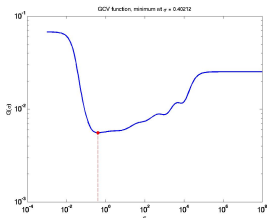
$$\frac{\|\mathbf{b} - A\mathbf{x}(\lambda)\|_{W_b}^2}{[\text{trace}(I_m - A(\lambda))]^2},$$

which estimates predictive risk.

- Expensive** - requires range of  $\lambda$ .
- GSVD makes calculations *efficient*.
- Requires minimum



**Multiple minima**



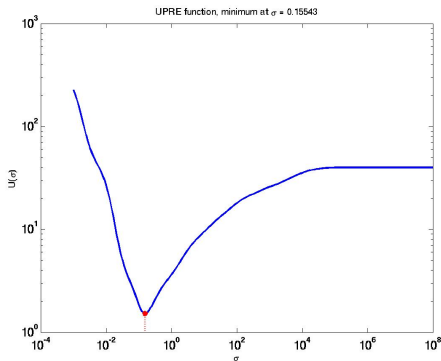
**Sometimes flat**

## Unbiased Predictive Risk Estimation (UPRE)

- Minimize expected value of predictive risk: Minimize UPRE function

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

- Expensive** - requires range of  $\lambda$ .
- 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  $\chi^2$  distribution with  $m - r$  degrees of freedom:

***Basically the Discrepancy Principle***

### Corollary (Weighted Least Squares)

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

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

## Extension: Statistics of the Regularized Least Squares Problem

Two New Results to Help Find the Regularization parameter:

**Theorem:  $\chi^2$  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

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

## Key Aspects of the Proof I: The Functional $J$

### Algebraic Simplifications: Rewrite functional as quadratic form

- Regularized solution given in terms of **resolution** matrix  $R(W_D)$

$$\hat{\mathbf{x}} = \mathbf{x}_0 + (A^T W_b A + D^T W_x D)^{-1} A^T W_b \mathbf{r}, \quad (5)$$

$$= \mathbf{x}_0 + R(W_D) W_b^{1/2} \mathbf{r}, \quad \mathbf{r} = \mathbf{b} - A \mathbf{x}_0$$

$$= \mathbf{x}_0 + \mathbf{y}(W_D). \quad (6)$$

$$R(W_D) = (A^T W_b A + D^T W_x D)^{-1} A^T W_b^{1/2} \quad (7)$$

- Functional is given in terms of **influence matrix**  $A(W_D)$

$$A(W_D) = W_b^{1/2} A R(W_D) \quad (8)$$

$$J_D(\hat{\mathbf{x}}) = \mathbf{r}^T W_b^{1/2} (I_m - A(W_D)) W_b^{1/2} \mathbf{r}, \quad \text{let } \tilde{\mathbf{r}} = W_b^{1/2} \mathbf{r} \quad (9)$$

$$= \tilde{\mathbf{r}}^T (I_m - A(W_D)) \tilde{\mathbf{r}}. \quad (10)$$

## Key Aspects of the Proof II : Properties of a Quadratic Form

### $\chi^2$ distribution of Quadratic Forms $\mathbf{x}^T P \mathbf{x}$ for normal variables (Fisher-Cochran Theorem)

- Components  $x_i$  are independent normal variables  $x_i \sim N(0, 1)$ ,  $i = 1 : n$ .
- A necessary and sufficient condition that  $\mathbf{x}^T P \mathbf{x}$  has a **central  $\chi^2$  distribution** is that  $P$  is **idempotent**,  $P^2 = P$ . In which case the degrees of freedom of  $\chi^2$  is  $\text{rank}(P) = \text{trace}(P) = n$ .
- When the means of  $x_i$  are  $\mu_i \neq 0$ ,  $\mathbf{x}^T P \mathbf{x}$  has a **non-central  $\chi^2$  distribution**, with **non-centrality parameter**  $c = \mu^T P \mu$
- A  $\chi^2$  random variable with  $n$  degrees of freedom and centrality parameter  $c$  has **mean**  $n + c$  and **variance**  $2(n + 2c)$ .

## Key Aspects of the Proof III: Requires the GSVD

### Lemma

Assume invertibility and  $m \geq n \geq p$ . There exist unitary matrices  $U \in \mathcal{R}^{m \times m}$ ,  $V \in \mathcal{R}^{p \times p}$ , and a nonsingular matrix  $X \in \mathcal{R}^{n \times n}$  such that

$$A = U \begin{bmatrix} \Upsilon \\ \mathbf{0}_{(m-n) \times n} \end{bmatrix} X^T \quad D = V[M, \mathbf{0}_{p \times (n-p)}]X^T, \quad (11)$$

$$\begin{aligned} \Upsilon &= \text{diag}(v_1, \dots, v_p, 1, \dots, 1) \in \mathcal{R}^{n \times n}, \quad M = \text{diag}(\mu_1, \dots, \mu_p) \in \mathcal{R}^{p \times p}, \\ 0 &\leq v_1 \leq \dots \leq v_p \leq 1, \quad 1 \geq \mu_1 \geq \dots \geq \mu_p > 0, \\ &v_i^2 + \mu_i^2 = 1, \quad i = 1, \dots, p. \end{aligned} \quad (12)$$

### The Functional with the GSVD

$$\begin{aligned} \text{Let } \tilde{Q} &= \text{diag}(\mu_1, \dots, \mu_p, \mathbf{0}_{n-p}, I_{m-n}) \\ \text{then } J &= \tilde{\mathbf{r}}^T (I_m - A(W_D)) \tilde{\mathbf{r}} = \|\tilde{Q}U^T \tilde{\mathbf{r}}\|_2^2, \end{aligned}$$

## Key Aspects of the Proof III: Requires the GSVD

### Lemma

Assume invertibility and  $m \geq n \geq p$ . There exist unitary matrices  $U \in \mathcal{R}^{m \times m}$ ,  $V \in \mathcal{R}^{p \times p}$ , and a nonsingular matrix  $X \in \mathcal{R}^{n \times n}$  such that

$$A = U \begin{bmatrix} \Upsilon & \\ & \mathbf{0}_{(m-n) \times n} \end{bmatrix} X^T \quad D = V[M, \mathbf{0}_{p \times (n-p)}]X^T, \quad (11)$$

$$\begin{aligned} \Upsilon &= \text{diag}(v_1, \dots, v_p, 1, \dots, 1) \in \mathcal{R}^{n \times n}, \quad M = \text{diag}(\mu_1, \dots, \mu_p) \in \mathcal{R}^{p \times p}, \\ 0 &\leq v_1 \leq \dots \leq v_p \leq 1, \quad 1 \geq \mu_1 \geq \dots \geq \mu_p > 0, \\ &v_i^2 + \mu_i^2 = 1, \quad i = 1, \dots, p. \end{aligned} \quad (12)$$

### The Functional with the GSVD

$$\begin{aligned} \text{Let } \tilde{Q} &= \text{diag}(\mu_1, \dots, \mu_p, \mathbf{0}_{n-p}, I_{m-n}) \\ \text{then } J &= \tilde{\mathbf{r}}^T (I_m - A(\mathbf{W}_D)) \tilde{\mathbf{r}} = \|\tilde{Q}U^T \tilde{\mathbf{r}}\|_2^2, \end{aligned}$$

## Key Aspects of the Proof IV: Statistical Distribution of the Weighted Residual

### Covariance Structure

- $\mathbf{e} = \mathbf{Ax} - \mathbf{b} \sim N(0, \mathbf{C}_b)$  hence we can show  $\mathbf{b} \sim N(\mathbf{Ax}_0, \mathbf{C}_b + \mathbf{AC}_D\mathbf{A}^T)$   
**Note that  $\mathbf{b}$  depends on  $\mathbf{x}$ .**
- $\mathbf{r} \sim N(0, \mathbf{C}_b + \mathbf{AC}_D\mathbf{A}^T)$ , and  $\tilde{\mathbf{r}} \sim N(0, \mathbf{I} + \tilde{\mathbf{A}}\mathbf{C}_D\tilde{\mathbf{A}}^T)$ ,  $\tilde{\mathbf{A}} = \mathbf{W}_b^{-1/2}\mathbf{A}$ .
- Use the GSVD

$$\mathbf{I} + \tilde{\mathbf{A}}\mathbf{C}_D\tilde{\mathbf{A}}^T = \mathbf{U}\mathbf{Q}^{-2}\mathbf{U}^T,$$

$$\mathbf{Q} = \text{diag}(\mu_1, \dots, \mu_p, \mathbf{I}_{n-p}, \mathbf{I}_{m-n})$$

### The Functional is a rv

- Let  $\mathbf{k} = \mathbf{QU}^T\tilde{\mathbf{r}}$ , then  $\mathbf{k} \sim N(0, \mathbf{QU}^T(\mathbf{U}\mathbf{Q}^{-2}\mathbf{U}^T)\mathbf{U}\mathbf{Q}) \sim N(0, \mathbf{I}_m)$
- But  $J = \|\tilde{\mathbf{Q}}\mathbf{U}^T\tilde{\mathbf{r}}\|^2 = \|\tilde{\mathbf{k}}\|^2$ , where  $\tilde{\mathbf{k}}$  is the vector  $\mathbf{k}$  excluding components  $p+1:n$ . Thus

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

## Key Aspects of the Proof IV: Statistical Distribution of the Weighted Residual

### Covariance Structure

- $\mathbf{e} = \mathbf{Ax} - \mathbf{b} \sim N(0, \mathbf{C}_b)$  hence we can show  $\mathbf{b} \sim N(\mathbf{Ax}_0, \mathbf{C}_b + A\mathbf{C}_D A^T)$   
**Note that  $\mathbf{b}$  depends on  $\mathbf{x}$ .**
- $\mathbf{r} \sim N(0, \mathbf{C}_b + A\mathbf{C}_D A^T)$ , and  $\tilde{\mathbf{r}} \sim N(0, I + \tilde{A}\mathbf{C}_D \tilde{A}^T)$ ,  $\tilde{A} = \mathbf{W}_b^{-1/2} A$ .
- Use the GSVD

$$I + \tilde{A}\mathbf{C}_D \tilde{A}^T = UQ^{-2}U^T,$$

$$Q = \text{diag}(\mu_1, \dots, \mu_p, I_{n-p}, I_{m-n})$$

### The Functional is a rv

- Let  $\mathbf{k} = QU^T \tilde{\mathbf{r}}$ , then  $\mathbf{k} \sim N(0, QU^T(UQ^{-2}U^T)UQ) \sim N(0, I_m)$
- But  $J = \|\tilde{Q}U^T \tilde{\mathbf{r}}\|^2 = \|\tilde{\mathbf{k}}\|^2$ , where  $\tilde{\mathbf{k}}$  is the vector  $\mathbf{k}$  excluding components  $p+1 : n$ . Thus

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

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

**Corollary: non-central  $\chi^2$  distribution of the regularized functional**

$$\hat{\mathbf{x}} = \operatorname{argmin} J_D(\mathbf{x}) = \operatorname{argmin} \{ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_{\mathbf{W}_b}^2 + \|(\mathbf{x} - \mathbf{x}_0)\|_{\mathbf{W}_D}^2 \}, \quad \mathbf{W}_D = D^T \mathbf{W}_x D. \quad (13)$$

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  $E(J) = m - n + p + c$
- Mead presented nonlinear algorithm when  $c = 0$ .
- But it does find  $W_D$ .
- Here- find  $\lambda$  only.

## What do we need to apply the Theory?

### Requirements

- **Covariance** information  $C_{\mathbf{b}}$  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  $C_{\mathbf{b}}$  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

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

- 

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

**Then we can use  $E(J)$  to find  $\lambda$**

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

$$\tilde{m} - \sqrt{2\tilde{m}}z_{\alpha/2} < J(\mathbf{x}(W_D)) < \tilde{m} + \sqrt{2\tilde{m}}z_{\alpha/2}. \quad (14)$$

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

## GOAL

Find  $W_D$  to make (14) tight: Single Variable case find  $\lambda$

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

## A Newton-line search Algorithm to find $\lambda$ . (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

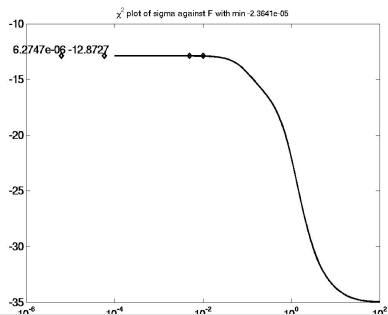
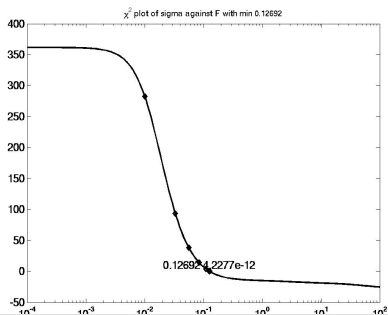
$$\sigma^{(k+1)} = \sigma^{(k)} \left( 1 + \frac{1}{2} \left( \frac{\sigma^{(k)}}{\|\mathbf{D}\mathbf{y}\|} \right)^2 (J_D(\sigma^{(k)}) - \tilde{m}) \right).$$

- Add a line search

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

## 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  $\sigma$
- **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  $\sigma$  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.( if  $L$  is not invertible you just need to know how to find  $Ax$  and  $A^T x$ , and the null space of  $L$ )
- Use LSQR algorithm to find the bidiagonal matrix for the projected problem.
- Obtain a solution of the bidiagonal problem for given initial  $\sigma$ .

#### Subsequent Steps

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

## 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  $\sigma$ .
- The regularization introduced by LSQR projection may be useful for preventing problems with GSVD expansion.
- Makes algorithm viable for large scale problems.

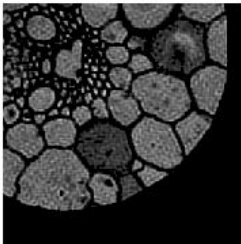
## Recall: Implementation Assumptions

### Covariance of Error: Statistics of Measurement Errors

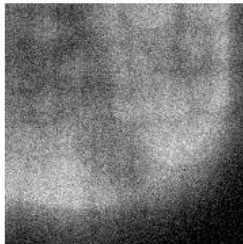
- Information on the covariance structure of errors in  $\mathbf{b}$  needed.
- Use  $\mathbf{C}_{\mathbf{b}} = \sigma_{\mathbf{b}}^2 I$  for common covariance, **white noise**.
- Use  $\mathbf{C}_{\mathbf{b}} = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_m^2)$  for **colored uncorrelated noise**.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

## Illustrating the Deblurring Result: Problem Size 65536

True



Blurred



Chi



**Computational Cost is Minimal: Projected Problem Size is 15,  $\lambda = .58$**

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

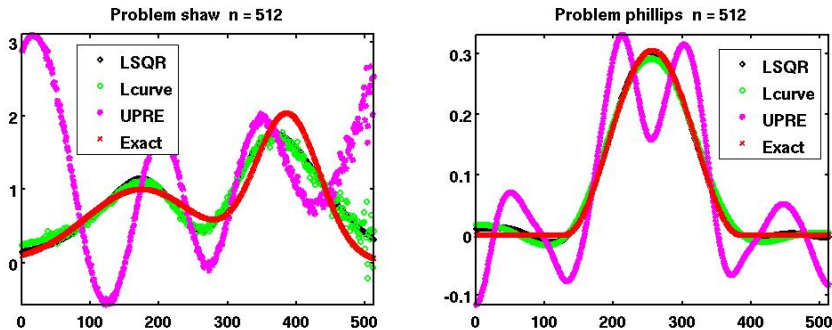
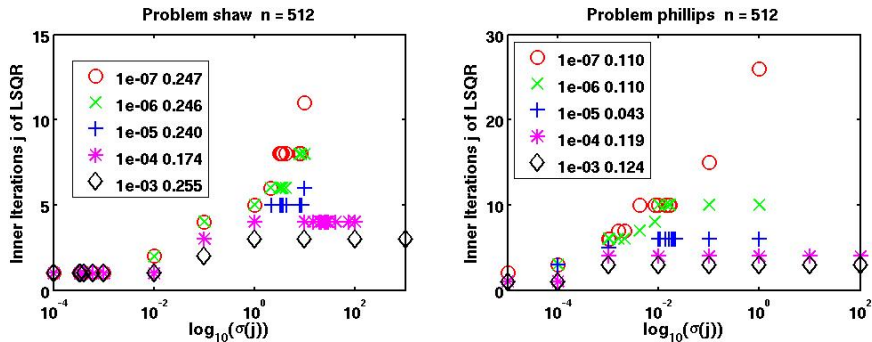


Figure: Comparison for noise level 10%. On left  $D = I$  and on right  $D$  is first derivative

- Notice L-curve and  $\chi^2$ -LSQR perform well.
- UPRE does not perform well.

## Sensitivity of LSQR to Parameter Choices



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

- Subproblem size increases with increasing  $\sigma$ .
- Subproblem size decreases with decreasing tolerance.

## 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.

## Future Work

### Other Results and Future Work

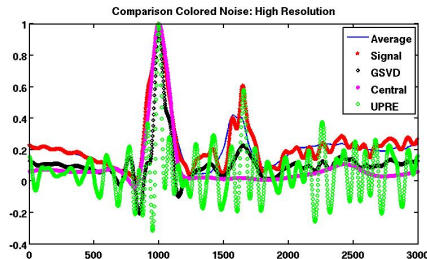
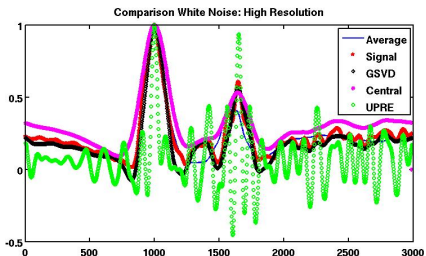
- Preconditioning
- Software Package!
- Diagonal Weighting Schemes
- Edge preserving regularization - Total Variation

## Real Data: Seismic Signal Restoration

### The Data Set and Goal

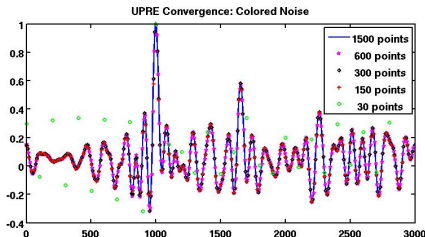
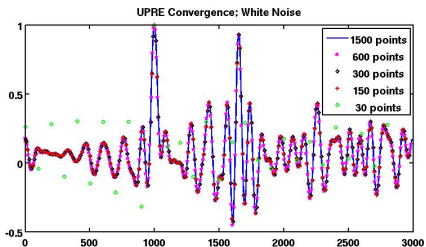
- Real data set of 48 signals of length 3000.
- The point spread function is derived from the signals.
- Calculate the signal variance pointwise over all 48 signals.
- Goal: restore the signal  $\mathbf{x}$  from  $A\mathbf{x} = \mathbf{b}$ , where  $A$  is psf matrix and  $\mathbf{b}$  is given blurred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

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



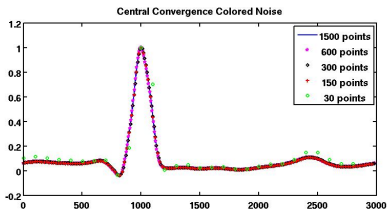
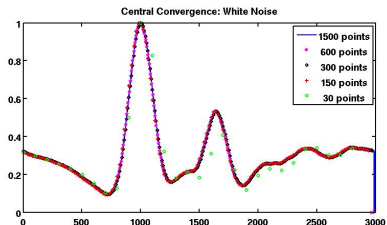
Greater contrast with  $\chi^2$ . UPRE is insufficiently regularized.  
 L-curve severely undersmooths (not shown). Parameters not consistent across resolutions

# 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.