

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

Rosemary Renaut



ARIZONA STATE UNIVERSITY

DEPARTMENT OF MATHEMATICS AND STATISTICS

Prague 2008

## Introduction and Motivation

Some Standard (or NOT) Methods for Regularization Parameter Estimation

## Statistical Results for Least Squares

## Implications of Statistical Results for Regularized Least Squares

## Newton algorithm

## Algorithm with LSQR

## Results

## Conclusions and Future Work

## Other Detailed Results

Results for Validating LSQR Implementation with GSVD

- ▶ 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.**

### Include some additional information about the solution and/or the data

- ▶ Usually error in  $\mathbf{b}$ ,  $\mathbf{e}$  is an  $m$ -vector of **random measurement errors** with mean 0 and **positive definite covariance** matrix  $\mathbf{C}_b = \mathbf{E}(\mathbf{e}\mathbf{e}^T)$ .
- ▶ Suppose  $\mathbf{C}_b$  is known. (Calculate if given multiple  $\mathbf{b}$ )
  - ▶ For **uncorrelated** measurements  $\mathbf{C}_b$  is **diagonal** matrix of **standard deviations** of the errors. (Colored noise)
- ▶ Perhaps the fit to data can be calculated in a **weighted** norm.
- ▶ Let  $\mathbf{W}_b = \mathbf{C}_b^{-1}$  and  $\mathbf{L}_b\mathbf{L}_b^T = \mathbf{W}_b$  be the Choleski factorization of  $\mathbf{W}_b$  and weight the equation:  $\mathbf{L}_b\mathbf{A}\mathbf{x} = \mathbf{L}_b\mathbf{b} + \tilde{\mathbf{e}}$ , ie

$$\mathbf{x}_{WLS} = \arg \min_{\mathbf{x}} \{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_{\mathbf{W}_b}^2 \}$$

- ▶ Then, theoretically,  $\tilde{\mathbf{e}}$  are uncorrelated. (White noise).
- ▶ But system ill-conditioning is usually **deteriorated** if noise is far from white (see Hansen).

## Alternative: Introduce a Mechanism for Regularization

### Weighted Fidelity with Regularization

- Regularize

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

where  $R(\mathbf{x})$  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?
- ▶ Depends on  $\lambda$  - 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.

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

- ▶ Assume  $\mathcal{N}(A) \cap \mathcal{N}(D) = \emptyset$
- ▶ Weighting matrix  $W_{\mathbf{b}}$  is inverse covariance matrix for data  $\mathbf{b}$ .
- ▶  $\mathbf{x}_0$  is a reference solution, often  $\mathbf{x}_0 = 0$ .
- ▶ Solution

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

### Question

Given  $D$ , how do we find  $\lambda$ ?

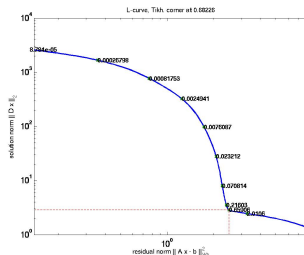
# Some standard approaches I: L-curve - *Find the corner*

- ▶ Let  $\mathbf{r}(\lambda) = (A(\lambda) - A)\mathbf{b}$ :  
Influence Matrix  
 $A(\lambda) = A(A^T W_b A + \lambda^2 D^T D)^{-1} A^T$
- ▶ Plot

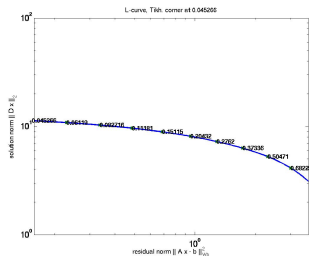
$$\log(\|D\mathbf{x}\|), \log(\|\mathbf{r}(\lambda)\|)$$

Trade off contributions.

- ▶ **Expensive** - requires range of  $\lambda$ .
- ▶ GSVD makes calculations *efficient*.
- ▶ **Not statistically based**



**Find corner**



**No corner**

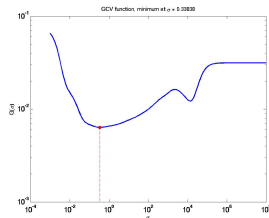
## Some standard approaches II: Generalized Cross-Validation (GCV)

- ▶ Minimizes GCV function

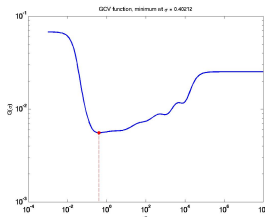
$$\frac{\|\mathbf{b} - A\mathbf{x}(\lambda)\|_{W_b}^2}{[\text{trace}(I_m - A(\lambda))]^2},$$

which estimates predictive risk.

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



**Multiple minima**



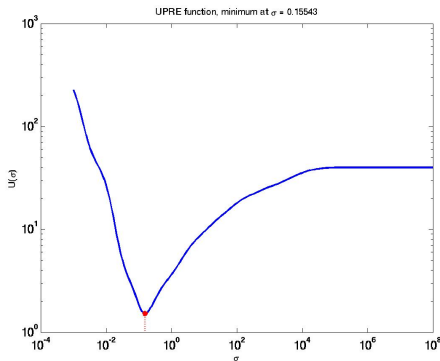
**Sometimes flat**

## Some standard approaches III: 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  $\lambda$ .
- ▶ GSVD makes calculations *efficient*.
- ▶ **Statistically based**
- ▶ **Minimum needed**



- ▶ Iterate LSQR to find solution and stop when noise starts to dominate (Hnetynkova and others)
- ▶ Solve the reduced system.
- ▶ Hybrid method - solve reduced system with additional regularization.
  - ▶ Cost of regularization of reduced system is minimal
  - ▶ Any regularization method may be used ( Nagy)
- ▶ Talk to the local experts!

## A More general formulation: Maximum A Posteriori Method

### Formulation:

$$\hat{\mathbf{x}} = \operatorname{argmin} J(\mathbf{x}) = \operatorname{argmin} \{ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_{W_b}^2 + \|(\mathbf{x} - \mathbf{x}_0)\|_{W_D}^2 \}. \quad (3)$$

Notice the regularization term includes a new weighting matrix, but no mapping  $D$ . (It is hidden in  $W_D$ )

- ▶ **Standard:**  $W_D = \lambda^2 D^T D$ ,  $\lambda$  unknown penalty parameter.
- ▶ **Ideally, statistically,**  $W_D$  is **inverse covariance matrix** for the mapped model  $D\mathbf{x}$  i.e.  $\lambda = 1/\sigma_x$ ,  $\sigma_x^2$  the common variance in  $D\mathbf{x}$ .
- ▶ Assumes the resulting estimates for  $D\mathbf{x}$  **uncorrelated**.
- ▶  $\hat{\mathbf{x}}$  is the standard **maximum a posteriori (MAP)** estimate of the solution, when all *a priori* information is provided.

**Can this provide additional information?**

## 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

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

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

### Corollary (Weighted Least Squares)

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

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

## Extension: Statistics of the Regularized Least Squares Problem

Two New Results to Help Find the Regularization parameter:

**Theorem:**  $\chi^2$  distribution of the regularized functional

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

Assume

- ▶  $W_b$  and  $W_x$  are symmetric positive definite.
- ▶ Problem is uniquely solvable  $\mathcal{N}(A) \cap \mathcal{N}(D) \neq 0$ .
- ▶ Moore-Penrose generalized inverse of  $W_D$  is  $C_D$
- ▶ Statistics:  $(\mathbf{b} - \mathbf{Ax}) = \mathbf{e} \sim N(0, C_b)$ ,  $(\mathbf{x} - \mathbf{x}_0) = \mathbf{f} \sim N(0, C_D)$ ,
  - ▶  $\mathbf{x}_0$  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  $\chi^2$  distribution of the regularized functional**

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

Assume all assumptions as before, but  $\mathbf{x}_1 \neq \mathbf{x}_0$  is the mean vector of the model parameters.

Let

$$c = \|\mathbf{c}\|_2^2 = \|\tilde{\mathbf{Q}}\mathbf{U}^T \mathbf{W}_b^{1/2} \mathbf{A}(\mathbf{x}_1 - \mathbf{x}_0)\|_2^2$$

Then

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

### 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 2007)
- ▶ Mead algorithm finds  $W_D$  but is expensive
- ▶ Our proposal - 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  $\mathbf{x}_1$ .
- ▶ But  $\mathbf{x}_1$  and  $\mathbf{x}_0$  are not known.
- ▶ For repeated data measurements  $C_{\mathbf{b}}$  can be calculated. Also  $\mathbf{b}_1$  can be found, the mean of  $\mathbf{b}$ .
- ▶ But  $E(\mathbf{b}) = A E(\mathbf{x})$  implies  $\mathbf{b}_1 = A \mathbf{x}_1$ . Hence

$$c = \|\mathbf{c}\|_2^2 = \|\tilde{Q}U^T W_{\mathbf{b}}^{1/2}(\mathbf{b}_1 - A\mathbf{x}_0)\|_2^2$$

▶

$$E(J_D) = E(\|\tilde{Q}U^T W_{\mathbf{b}}^{1/2}(\mathbf{b} - A\mathbf{x}_0)\|_2^2) = m+p-n + \|\tilde{Q}U^T W_{\mathbf{b}}^{1/2}(\mathbf{b}_1 - A\mathbf{x}_0)\|_2^2$$

**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

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

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

## GOAL

Find  $W_D$  to make (6) 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

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

- ▶ Hence we have a basic Newton Iteration

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

- ▶ Add a line search

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

### GSVD

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

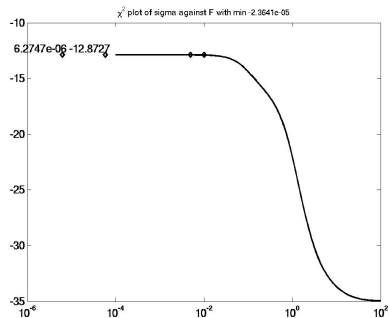
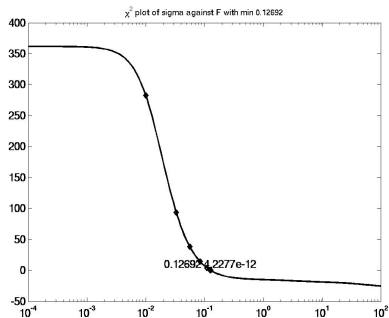
$$F(\sigma_{\mathbf{x}}) = \sum_{i=1}^p \left( \frac{1}{\gamma_i^2 \sigma_{\mathbf{x}}^2 + 1} \right) s_i^2 + \sum_{i=n+1}^m s_i^2 - \tilde{m} = 0$$

- ▶ 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  $\sigma$
- ▶ **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.



### 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.

### Algorithm

#### Initialization

- ▶ Convert generalized Tikhonov problem to standard form.
- ▶ Use LSQR algorithm to find the bidiagonal matrix spanning appropriate space of solution
- ▶ Obtain a solution of the bidiagonal problem for given  $\sigma$ .
- ▶ Reuse bidiagonalization in update of  $\sigma$  for Newton.
- ▶ Each  $\sigma$  calculation of algorithm reuses saved information from the Lancos bidiagonalization. The system is augmented if needed.

- ▶ Algorithm concurrently regularizes and solves the system.
- ▶ In contrast, standard hybrid LSQR solves projected system with regularization.
- ▶ 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  $\mathbf{C}_b = \sigma_b^2 I$  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  $\mathbf{b}_1$  as the mean of measured  $\mathbf{b}$ , when implemented with centrality parameter,  $\mathbf{x}_0 = 0$ .

### Tolerance on Convergence

- ▶ The convergence tolerance depends on the noise structure.
- ▶ Use  $\text{TOL} = \sqrt{2\tilde{m}}z_{\alpha/2}$ .
- ▶ No noise structure use  $\alpha = .001$ , generates large TOL
- ▶ Good noise information use  $\alpha = .95$ , generates small TOL

## 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  $\mathbf{b}_1$  as the mean of measured  $\mathbf{b}$ , when implemented with centrality parameter,  $\mathbf{x}_0 = 0$ .

### Tolerance on Convergence

- ▶ The convergence tolerance depends on the noise structure.
- ▶ Use  $\text{TOL} = \sqrt{2\tilde{m}}z_{\alpha/2}$ .
- ▶ No noise structure use  $\alpha = .001$ , generates large TOL
- ▶ Good noise information use  $\alpha = .95$ , generates small TOL

## An example of the method: 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

- ▶ No exact solution.
  - ▶ Downsample the signal and restore for different resolutions
- |            |       |       |        |        |         |
|------------|-------|-------|--------|--------|---------|
| Resolution | 2 : 1 | 5 : 1 | 10 : 1 | 20 : 1 | 100 : 1 |
| Points     | 1500  | 600   | 300    | 150    | 30      |
- ▶ Do results converge? Compare with UPRE and L-Curve.

## An example of the method: Seismic Signal Restoration

### The Data Set and Goal

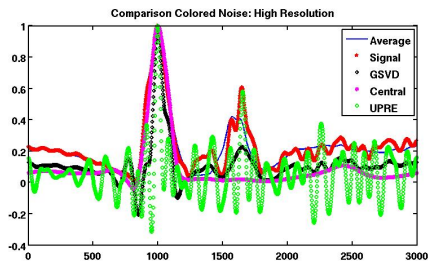
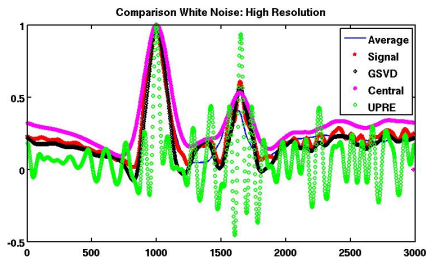
- ▶ Real data set of 48 signals of length 3000.
- ▶ The point spread function is derived from the signals.
- ▶ Calculate the signal variance pointwise over all 48 signals.
- ▶ Goal: restore the signal  $\mathbf{x}$  from  $A\mathbf{x} = \mathbf{b}$ , where  $A$  is psf matrix and  $\mathbf{b}$  is given blurred signal.

### Method of Comparison- no exact solution known

- ▶ No exact solution.
- ▶ Downsample the signal and restore for different resolutions  

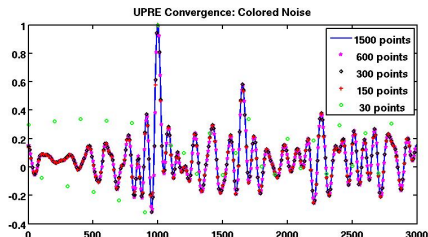
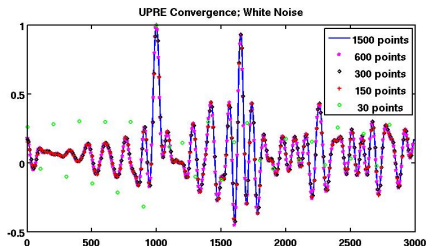
Resolution	2 : 1	5 : 1	10 : 1	20 : 1	100 : 1
Points	1500	600	300	150	30
- ▶ Do results converge? Compare with UPRE and L-Curve.

## 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

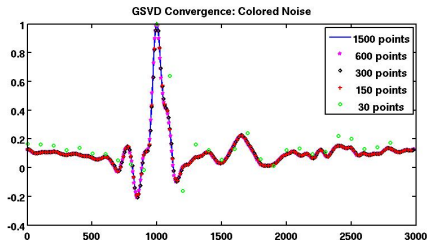
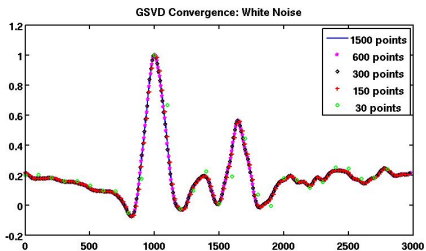
# THE UPRE SOLUTION: White Noise and Colored Noise $x_0 = 0$



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

# THE GSVD SOLUTION: White Noise (left) and Colored Noise (right)

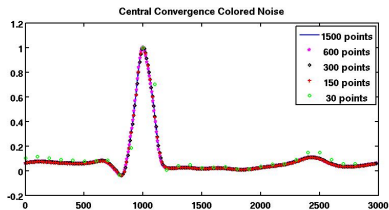
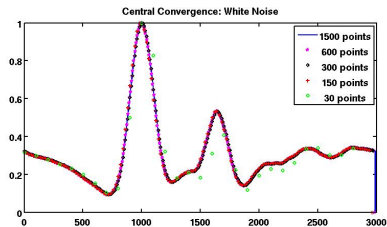
$$x_0 = 0$$



Regularization Parameters are consistent:

$\sigma = 0.00058$  (left),  $\sigma = 0.00069$  (right) all resolutions

# THE NONCENTRAL GSVD SOLUTION: White Noise (left) and Colored Noise (right) $x_0 = 0$



Regularization quite consistent resolution 2 to 100

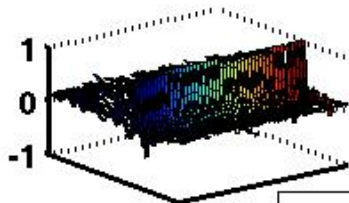
$\sigma = 0.0000029, .0000029, .0000029, .0000057, .0000057$  (left)

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

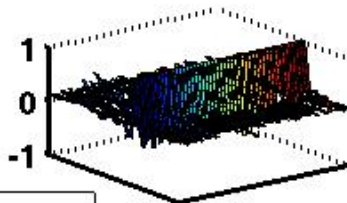
Notice that colored noise eliminates second arrival of signal but excellent contrast to identify primary arrival.

# More Signals! UPRE and L-curve exhibit under regularization

## Central

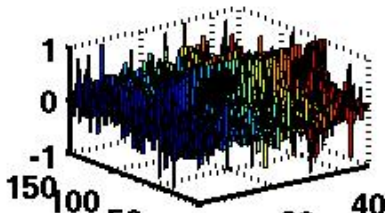


## Non-Central

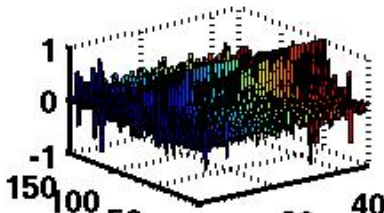


Dataset 1  
Colored Noise

## L-Curve



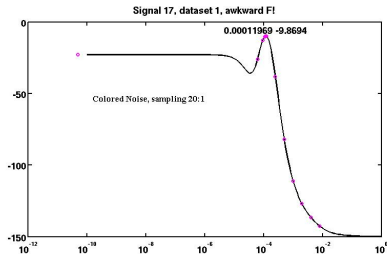
## UPRE



### 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  $\mathbf{x}_0$  is the mean of  $\mathbf{x}$  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  $\mathbf{x}_0 = 0$ .)
- ▶  $\sigma$  can be bounded by result of central case.
- ▶ Range of  $\sigma$  given by range of  $\gamma_i$ .
- ▶ May oversmooth the solution if good range of  $\sigma$  not found.

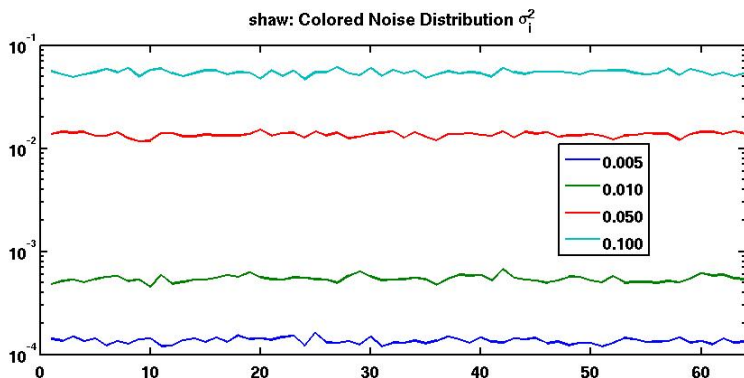
### Other Results and Future Work

- ▶ Degrees of freedom reduced when using the GSVD.
- ▶ How to apply Picard condition for GSVD to handle problems with robustness due to conditioning of  $C_b$
- ▶ Image deblurring. (Implementation to use minimal storage)
- ▶ Diagonal Weighting Schemes
- ▶ Edge preserving regularization
- ▶ Constraint implementation ( with Mead submitted).

### Details

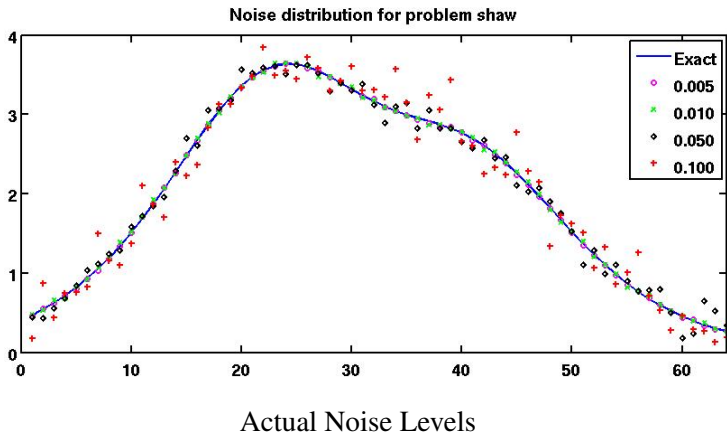
- ▶ Take example from Hansen's toolbox, eg shaw, phillips, heat, ilaplace.
- ▶ Generate 500 copies for each noise level, here .005, .01, .05, .1.
- ▶ Solve for 500 cases using GSVD and LSQR Newton.
- ▶ Pairwise t test on obtained  $\sigma$ : verify equivalence GSVD and LSQR.
- ▶ Compare results with statistical technique : UPRE
  - ▶ Errors - relative least squares, and max error. Calculate over all errors less than .5.
  - ▶ Regularization parameter (not given).

## Example of the Colored Noise distribution

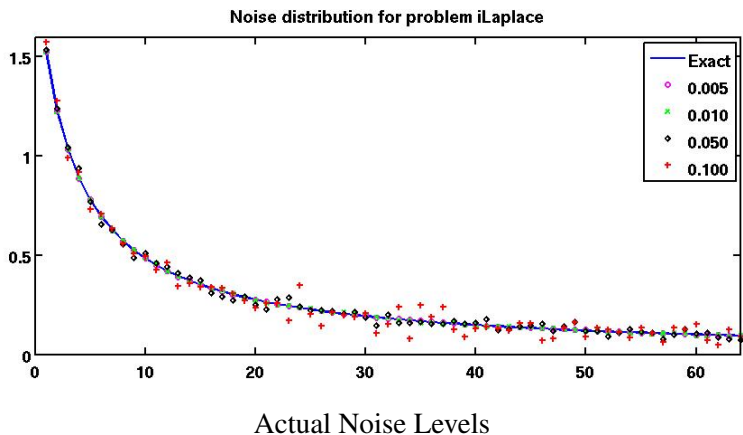


The pointwise variance for each noise level

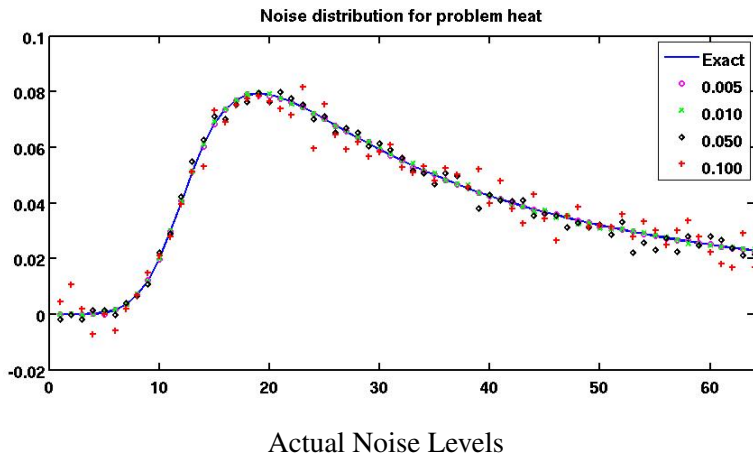
## Noise and Exact Right Hand Side Vector



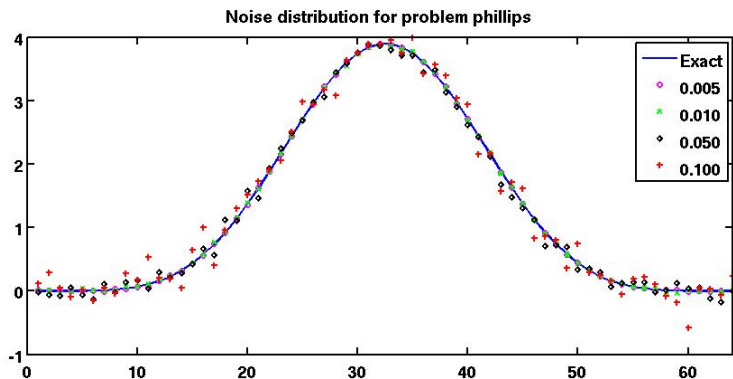
# Noise and Exact Right Hand Side Vector



# Noise and Exact Right Hand Side Vector



## Noise and Exact Right Hand Side Vector



Actual Noise Levels

## Problem size 64, Regularization First Order Derivative, Colored Noise

<b>shaw</b>					
	Bidiagonal	Newton Steps		P value	
$\epsilon$	Average	GSVD	LSQR	Iteration	$\sigma$
$5e - 02$	7.2	7.1	7.1	$1.000e + 00$	$1.000e + 00$
$1e - 01$	6.2	7.0	7.0	$9.166e - 01$	$8.489e - 01$
<b>ilaplace</b>					
$5e - 02$	8.0	6.5	6.5	$9.384e - 01$	$9.991e - 01$
$1e - 01$	7.4	6.5	6.5	$8.610e - 01$	$4.870e - 01$
<b>heat</b>					
$5e - 02$	29.4	5.8	5.8	$1.000e + 00$	$1.000e + 00$
$1e - 01$	22.4	6.0	6.0	$1.000e + 00$	$1.000e + 00$
<b>phillips</b>					
$5e - 02$	15.4	6.1	6.1	$1.000e + 00$	$1.000e + 00$
$1e - 01$	12.8	6.1	6.1	$1.000e + 00$	$1.000e + 00$

**Table:** Convergence characteristics and P-values comparing GSVD and LSQR iteration steps and  $\sigma$  values

## Problem size 64, Regularization First Order Derivative, Colored Noise

<b>shaw</b>						
	Least Squares Error			Max Error		
$\epsilon$	GSVD	LSQR	UPRE	GSVD	LSQR	UPRE
.05	.34(207)	.34(207)	.34(146)	.31(111)	.31(111)	.31(84)
.1	.39(105)	.39(105)	.35(29)	.35(43)	.35(43)	.35(28)
<b>ilaplace</b>						
.05	.23(361)	.23(361)	.16(398)	.12(495)	.12(495)	.07(425)
.1	.25(317)	.25(317)	.21(370)	.15(490)	.15(490)	.10(428)
<b>heat</b>						
.05	.27(500)	.27(500)	.28(479)	.19(500)	.19(500)	.20(489)
.1	.35(497)	.35(497)	.36(461)	.27(500)	.27(500)	.27(489)
<b>phillips</b>						
.05	.10(500)	.10(500)	.10(455)	.10(500)	.10(500)	.09(455)
.1	.12(500)	.12(500)	.11(431)	.12(500)	.12(500)	.11(431)

**Table:** Comparison with UPRE, Relative Least Squares and max error, in parentheses the number of accepted values, large values ( $> .5$ ) excluded from the average

### Major Observations

- GSVD-LSQR**
- ▶ High correlation between  $\sigma$  for both  $\chi^2$  (GSVD and LSQR) obtained.  $p$ - values at or near 1.
  - ▶ Algorithms converge with very few regularization calculations, average  $\approx 7$ .
  - ▶ The bidiagonalized system is on average much smaller than the system size.
  - ▶ Errors distributions equivalent.
  - ▶ Total cost of LSQR is the cost of bidiagonalization plus use of bidiagonalization to obtain a solution, eg column 2 plus column 4 solves.
- UPRE**
- ▶ Errors of UPRE are similar to  $\chi^2$  approaches.
  - ▶ UPRE fails more often ( more discarded solutions with high noise).

THANK YOU!