

# Improving the Efficiency of the Chi-squared Method for Regularization Parameter Estimation

Rosemary Renaut, Iveta Hnetynkova, Jodi Mead

Arizona State, Charles University and Boise State

SIAM Annual Meeting 2008

- 1 Introduction
- 2 Statistical Results for Least Squares
- 3 Implications of Statistical Results for Regularized Least Squares
- 4 Newton algorithm
- 5 Results
- 6 Conclusions and Future Work

## Least Squares for $Ax = b$ , (Weighted)

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

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

- $\mathbf{e}$  is the  $m$ -vector of random measurement errors with mean 0 and **positive definite covariance** matrix

$$C_b = \mathbf{E}(\mathbf{e}\mathbf{e}^T).$$

- Assume that  $C_b$  is known. (Calculate if given multiple  $\mathbf{b}$ )
- For **uncorrelated** measurements  $C_b$  is **diagonal** matrix of **standard deviations** of the errors. (Colored noise)
- For **correlated** measurements, let  $W_b = C_b^{-1}$  and  $L_b L_b^T = W_b$  be the Choleski factorization of  $W_b$  and weight the equation:

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

- $\tilde{\mathbf{e}}$  are uncorrelated. (White noise).
- $\tilde{\mathbf{e}} \sim N(0, I)$ , normally distributed mean 0 and variance  $I$ .

# Weighted Regularized Least Squares for numerically ill-posed systems

## Formulation:

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

$\mathbf{x}_0$  is a reference solution, often  $\mathbf{x}_0 = 0$ .

- **Standard:**  $W_x = \lambda^2 I$ ,  $\lambda$  unknown penalty parameter.
- **Statistically,**  $W_x$  is **inverse covariance matrix** for the model  $\mathbf{x}$  i.e.  $\lambda = 1/\sigma_x$ ,  $\sigma_x^2$  the common variance in  $\mathbf{x}$ .
- Assumes the resulting estimates for  $\mathbf{x}$  **uncorrelated**.
- $\hat{\mathbf{x}}$  is the standard **maximum a posteriori (MAP)** estimate of the solution

## Question: The Problem

How do we find an *appropriate* regularization parameter  $\lambda$ ?  
More generally, what is the correct  $W_x$ ?

# Weighted Regularized Least Squares for numerically ill-posed systems

## Formulation:

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

$\mathbf{x}_0$  is a reference solution, often  $\mathbf{x}_0 = 0$ .

- **Standard:**  $W_x = \lambda^2 I$ ,  $\lambda$  unknown penalty parameter.
- **Statistically,**  $W_x$  is **inverse covariance matrix** for the model  $\mathbf{x}$  i.e.  $\lambda = 1/\sigma_x$ ,  $\sigma_x^2$  the common variance in  $\mathbf{x}$ .
- Assumes the resulting estimates for  $\mathbf{x}$  **uncorrelated**.
- $\hat{\mathbf{x}}$  is the standard **maximum a posteriori (MAP)** estimate of the solution

## Question: The Problem

How do we find an *appropriate* regularization parameter  $\lambda$ ?  
More generally, what is the correct  $W_x$ ?

# The General Case : Generalized Tikhonov Regularization

## Formulation: Regularization with Solution Mapping

Generalized Tikhonov regularization, operator  $D$  acts on  $\mathbf{x}$ .

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

- Assume **invertibility**  $\mathcal{N}(\mathbf{A}) \cap \mathcal{N}(D) = \emptyset$
- Then solutions depend on  $W_D = \lambda^2 D^T D$  :

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

## GOAL

- Can we estimate  $\lambda$  efficiently when  $W_b$  is known?
- Use statistics of the solution to find  $\lambda$ .

# The General Case : Generalized Tikhonov Regularization

## Formulation: Regularization with Solution Mapping

Generalized Tikhonov regularization, operator  $D$  acts on  $\mathbf{x}$ .

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

- Assume **invertibility**  $\mathcal{N}(A) \cap \mathcal{N}(D) = \emptyset$
- Then solutions depend on  $W_D = \lambda^2 D^T D$  :

$$\hat{\mathbf{x}}(\lambda) = \operatorname{argmin} J_D(\mathbf{x}) = \operatorname{argmin} \{ \|\mathbf{Ax} - \mathbf{b}\|_{W_b}^2 + \lambda^2 \|D(\mathbf{x} - \mathbf{x}_0)\|^2 \}. \quad (3)$$

## GOAL

- Can we estimate  $\lambda$  efficiently when  $W_b$  is known?
- Use statistics of the solution to find  $\lambda$ .

## Background: Statistics of the Least Squares Problem

### Theorem (Rao73: First Fundamental Theorem)

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

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

$J$  follows a  $\chi^2$  distribution with  $m - r$  degrees of freedom.

### Corollary (Weighted Least Squares)

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

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

## Background: Statistics of the Least Squares Problem

### Theorem (Rao73: First Fundamental Theorem)

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

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

$J$  follows a  $\chi^2$  distribution with  $m - r$  degrees of freedom.

### Corollary (Weighted Least Squares)

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

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

## 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:  $(\mathbf{b} - \mathbf{Ax}) = \mathbf{e} \sim N(0, C_b)$ ,  $(\mathbf{x} - \mathbf{x}_0) = \mathbf{f} \sim N(0, C_D)$ ,
  - $\mathbf{e}$  and  $\mathbf{f}$  are i.i.d. random variables.

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

$J_D$  is a random variable which follows a  $\chi^2$  distribution with  $m + p - n$  degrees of freedom.

## Algebraic Simplifications

- 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: 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

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

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

### The Functional with the GSVD

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

## Key Aspects of the Proof II: 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

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

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

### The Functional with the GSVD

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

# Key Aspects of the Proof III: 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

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

$$\mathbf{P} = \text{diag}(\mu_1^2, \dots, \mu_p^2, \mathbf{I}_{n-p}, \mathbf{I}_{m-n})$$

## The Functional is a rv

- Let  $\mathbf{k} = \mathbf{P}^{-1}\mathbf{U}^T\tilde{\mathbf{r}}$ , then  $\mathbf{k} \sim N(0, \mathbf{P}^{-1}\mathbf{U}^T(\mathbf{UP}^2\mathbf{U}^T)\mathbf{UP}^{-1}) \sim N(0, \mathbf{I}_m)$
- Now  $J = \tilde{\mathbf{r}}\mathbf{U}\mathbf{Q}\mathbf{U}^T\tilde{\mathbf{r}} = \mathbf{k}^T\mathbf{P}\mathbf{Q}\mathbf{P}\mathbf{k}$ . Thus

$$J_D = \sum_{i=1}^p k_i^2 + \sum_{i=n+1}^m k_i^2 \sim \chi^2(m+p-n).$$

# Key Aspects of the Proof III: 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

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

$$\mathbf{P} = \text{diag}(\mu_1^2, \dots, \mu_p^2, \mathbf{I}_{n-p}, \mathbf{I}_{m-n})$$

## The Functional is a rv

- Let  $\mathbf{k} = \mathbf{P}^{-1}\mathbf{U}^T\tilde{\mathbf{r}}$ , then  $\mathbf{k} \sim N(0, \mathbf{P}^{-1}\mathbf{U}^T(\mathbf{UP}^2\mathbf{U}^T)\mathbf{UP}^{-1}) \sim N(0, \mathbf{I}_m)$
- Now  $J = \tilde{\mathbf{r}}\mathbf{U}\mathbf{Q}\mathbf{U}^T\tilde{\mathbf{r}} = \mathbf{k}^T\mathbf{P}\mathbf{Q}\mathbf{P}\mathbf{k}$ . Thus

$$J_D = \sum_{i=1}^p k_i^2 + \sum_{i=n+1}^m k_i^2 \sim \chi^2(m+p-n).$$

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

### DESIGNING THE ALGORITHM: I

- If  $C_b$  and  $C_x$  are good estimates of the covariance matrices

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

should be **small**.

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

$$\tilde{m} - \sqrt{2\tilde{m}}z_{\alpha/2} < \mathbf{r}^T W_b^{1/2} (I_m - A(W_D)) W_b^{1/2} \mathbf{r} < \tilde{m} + \sqrt{2\tilde{m}}z_{\alpha/2}. \quad (13)$$

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

### GOAL

Find  $W_x$  to make (13) tight: Single Variable case find  $\lambda$

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

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

### DESIGNING THE ALGORITHM: I

- If  $C_b$  and  $C_x$  are good estimates of the covariance matrices

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

should be **small**.

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

$$\tilde{m} - \sqrt{2\tilde{m}}z_{\alpha/2} < \mathbf{r}^T W_b^{1/2} (I_m - A(W_D)) W_b^{1/2} \mathbf{r} < \tilde{m} + \sqrt{2\tilde{m}}z_{\alpha/2}. \quad (13)$$

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

### GOAL

Find  $W_x$  to make (13) tight: Single Variable case find  $\lambda$   
 $J_D(\hat{\mathbf{x}}(\lambda)) \approx \tilde{m}$

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

### 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

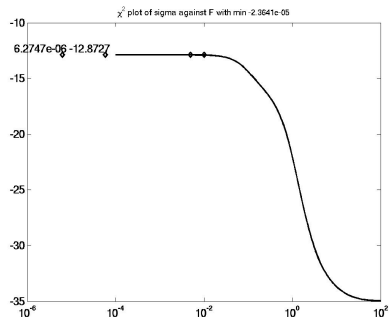
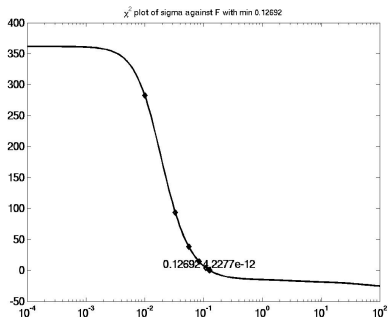
$$\frac{\partial}{\partial \sigma} J(\sigma) = -\frac{2}{\sigma^3} \|D\mathbf{y}(\sigma)\|^2 < 0$$

- Hence we have a basic Newton Iteration

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

## Convergence

- $F$  is **monotonic decreasing**
- 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

## Initialization

- Convert generalized Tikhonov problem to standard form.
- Lancos-hybrid projects to smaller problem with bidiagonal matrix.
- Each  $\sigma$  calculation of algorithm reuses saved information from the Lancos bidiagonalization. The system is augmented if 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 toID. Let  $\mathbf{t} = D\mathbf{y}/\sigma^{(k)}$  then

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

- **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

## Initialization

- Convert generalized Tikhonov problem to standard form.
- Lancos-hybrid projects to smaller problem with bidiagonal matrix.
- Each  $\sigma$  calculation of algorithm reuses saved information from the Lancos bidiagonalization. The system is augmented if 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 toID. Let  $\mathbf{t} = D\mathbf{y}/\sigma^{(k)}$  then

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

- **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.

## Covariance of Error: Statistics of Measurement Errors

- Information on the covariance structure of errors in  $\mathbf{b}$  needed.
- Use  $\mathbf{C}_b = \sigma_b^2 \mathbf{I}$  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  $\mathbf{C}_b = \mathbf{I}$ .

## 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

## Covariance of Error: Statistics of Measurement Errors

- Information on the covariance structure of errors in  $\mathbf{b}$  needed.
- Use  $\mathbf{C}_b = \sigma_b^2 \mathbf{I}$  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  $\mathbf{C}_b = \mathbf{I}$ .

## 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

## 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.
  - 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?

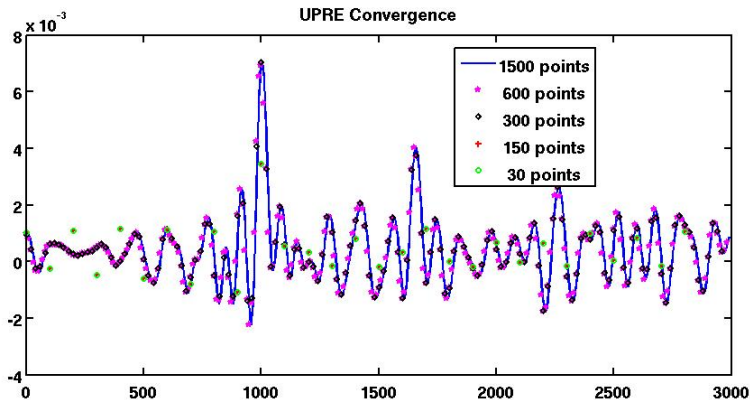
## The Data Set and Goal

- 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

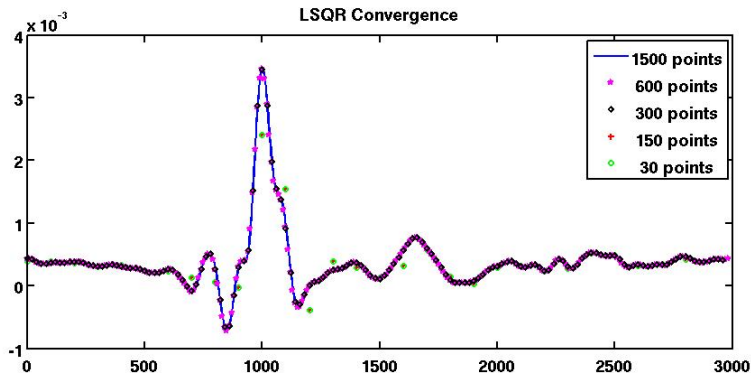
- No exact solution.
  - Downsample the signal and restore for different resolutions
- |            |       |       |        |        |         |
|------------|-------|-------|--------|--------|---------|
| Resolution | 2 : 1 | 5 : 1 | 10 : 1 | 20 : 1 | 100 : 1 |
| Points     | 1500  | 600   | 300  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

# THE UPRE SOLUTION



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

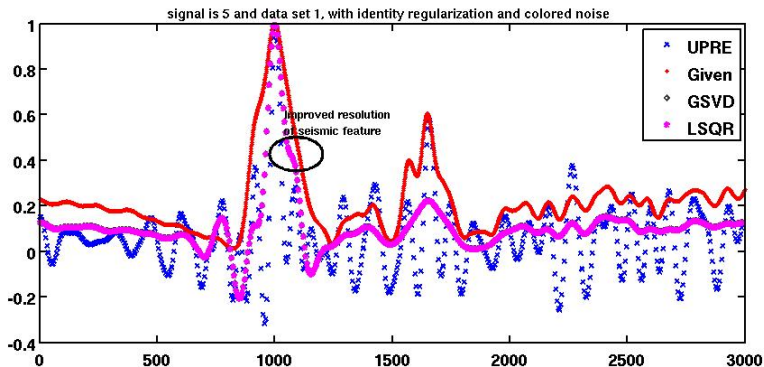
# THE LSQR SOLUTION



Regularization Parameters are consistent:

$\sigma = 0.00069, .00069, .00069, .00065, .00065$ , resolution from 2 to 100

# Comparison

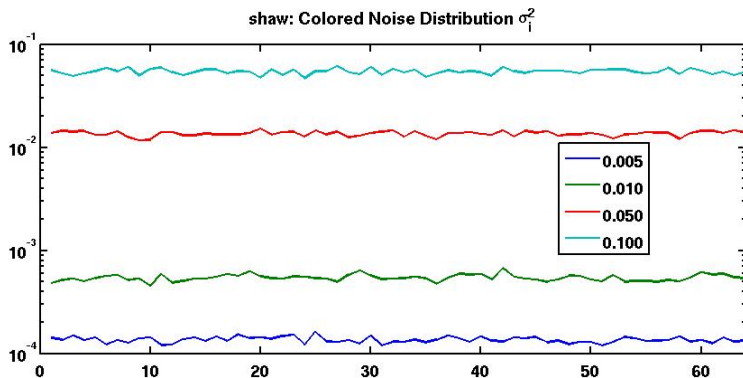


Greater contrast with  $\chi^2$  parameter estimation than with UPRE.  
Regularization Parameters of L-curve (not shown) are not consistent  
across resolutions, solution is poor

## Details

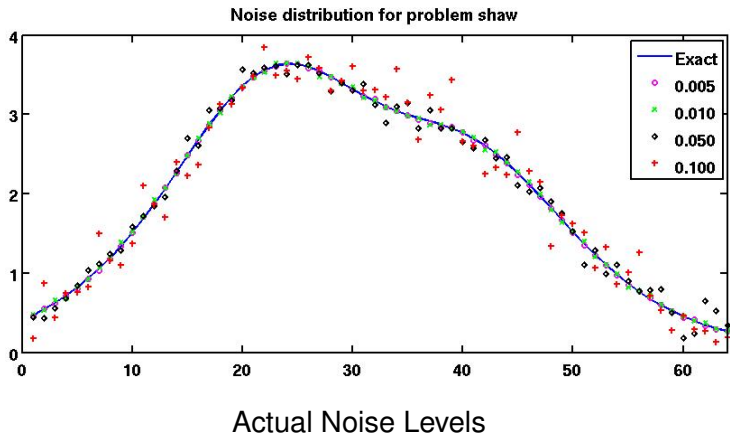
- Take example from Hansen's toolbox, eg `shaw`, `phillips`, `heat`, `ilaplace`.
- Generate 500 copies for each noise level, here `.005`, `.01`, `.05`, `.1`.
- Solve for 500 cases using GSVD and LSQR Newton.
- Pairwise t test on obtained  $\sigma$ : verify equivalence GSVD and LSQR.
- Compare results 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

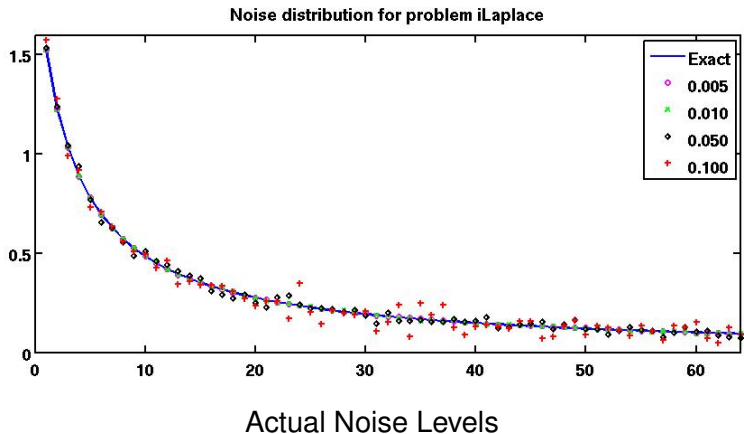


The pointwise variance for each noise level

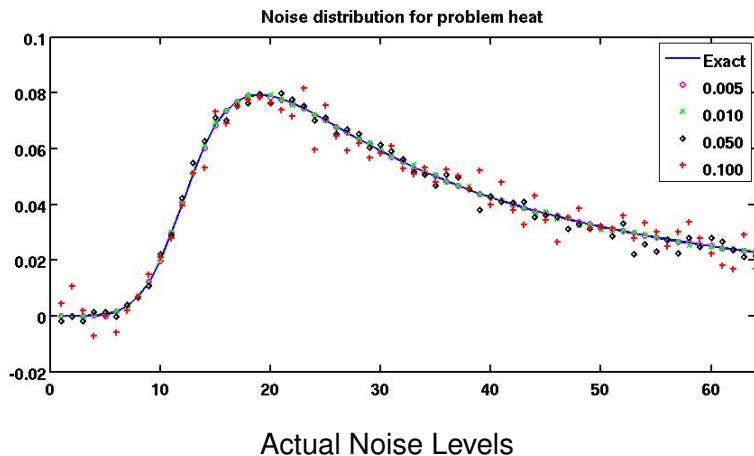
# Noise and Exact Right Hand Side Vector



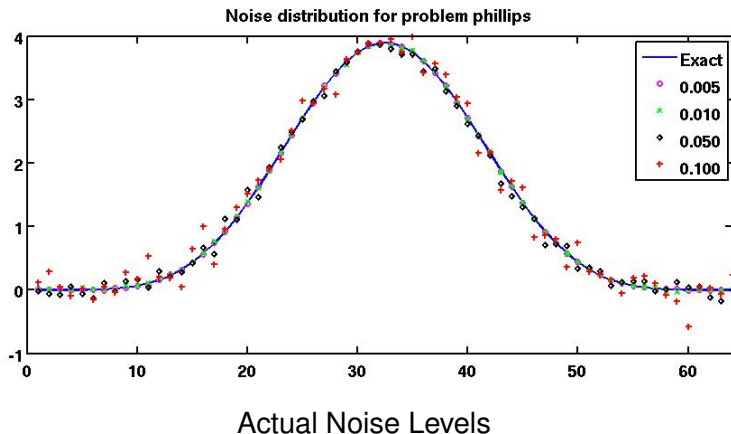
# Noise and Exact Right Hand Side Vector



# Noise and Exact Right Hand Side Vector



# Noise and Exact Right Hand Side Vector



# 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).

## Conclusions

- A new statistical method for estimating regularization parameter
  - Compares favorably with UPRE with respect to performance
- Method can be used for large scale problems, without GSVD
- Method is very efficient, Newton method is robust and fast.

## Future Work

- Investigate robustness with respect to conditioning of  $C_b$
- Image deblurring ( with Hnetynkova in prep.)
- Diagonal Weighting Schemes
- Edge preserving regularization

## Conclusions

- A new statistical method for estimating regularization parameter
  - Compares favorably with UPRE with respect to performance
- Method can be used for large scale problems, without GSVD
- Method is very efficient, Newton method is robust and fast.

## Future Work

- Investigate robustness with respect to conditioning of  $C_b$
- Image deblurring ( with Hnetynkova in prep.)
- Diagonal Weighting Schemes
- Edge preserving regularization

THANK YOU!