Statistically-Based hybrid LSQR Newton method for Finding the Regularization Parameter: Application in Image Deblurring and Signal Restoration

Rosemary Renaut
Joint work with Jodi Mead and Iveta Hnetynkova

April 27, 2009
Outline

1. Motivation
   - Quick Review

2. Statistical Results for Least Squares
   - Summary of LS Statistical Results
   - Implications of Statistical Results for Regularized Least Squares

3. Newton algorithm
   - Algorithm with LSQR (Paige and Saunders)

4. Results

5. Conclusions and Future Work
Signal/Image Restoration:

**Integral Model of Signal Degradation** \( b(t) = \int K(t, s)x(s)ds \)

- \( K(t, s) \) describes *blur* of the signal.
- Convolutional model: *invariant* \( K(t, s) = K(t - s) \) is Point Spread Function (PSF).
- Typically sampling includes noise \( e(t) \), model is
  \[
  b(t) = \int K(t - s)x(s)ds + e(t)
  \]

**Discrete model**: given discrete samples \( b \), find samples \( x \) of \( x \)

- Let \( A \) discretize \( K \), assume known, model is given by
  \[
  b = Ax + e.
  \]
- Naïvely *invert* the system to find \( x \)!
Example 1-D Original and Blurred Noisy Signal

Original signal $x$.

Blurred and noisy signal $b$, Gaussian PSF.
The Solution: Regularization is needed

Naïve Solution

A Regularized Solution
Least Squares for $Ax = b$: A Quick Review

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

$$Ax = b + e,$$

- **Classical Approach** Linear Least Squares

$$x_{LS} = \arg \min_x ||Ax - b||_2^2$$

- **Difficulty** $x_{LS}$ is sensitive to changes in the right hand side $b$ when $A$ is ill-conditioned.

For convolutional models system is ill-posed.
Introduce Regularization to Pick a Solution

**Weighted Fidelity with Regularization**

- Regularize

\[ x_{RLS}(\lambda) = \arg \min_x \{ \| b - Ax \|^2_{W_b} + \lambda^2 R(x) \}, \]

- Weighting matrix \( W_b \)
- \( R(x) \) is a regularization term
- \( \lambda \) is a regularization parameter which is unknown.
- Solution \( x_{RLS}(\lambda) \)
  - depends on \( \lambda \).
  - depends on regularization operator \( R \)
  - depends on the weighting matrix \( W_b \)
The Weighting Matrix: Some Assumptions

Given multiple measurements of data \( b \):

- Usually error in \( b \), \( e \) is an \( m \)-vector of random measurement errors with mean 0 and positive definite covariance matrix \( C_b = \text{E}(ee^T) \).
- For uncorrelated measurements \( C_b \) is diagonal matrix of standard deviations of the errors. (Colored noise)
- For white noise \( C_b = \sigma^2 I \).
- Weighting by \( W_b = C_b^{-1} \) in data fit term, theoretically, \( \tilde{e} \) are uncorrelated.
- Difficulty if \( W_b \) increases ill-conditioning of \( A \)!
Generalized Tikhonov Regularization With Weighting

Use $R(x) = \|D(x - x_0)\|^2$

\[
\hat{x} = \arg\min J(x) = \arg\min \{\|Ax - b\|_W^2 + \lambda^2 \|D(x - x_0)\|^2\}. \tag{1}
\]

- $D$ is a suitable operator, often derivative approximation.
- Assume $\mathcal{N}(A) \cap \mathcal{N}(D) = \emptyset$
- $x_0$ is a reference solution, often $x_0 = 0$.

**Question**

Given $D, W_b$ how do we find $\lambda$?
Example: solution for Increasing $\lambda$, $D = I$. 
Example: solution for Increasing $\lambda, D = I$. 
Example: solution for Increasing $\lambda$, $D = I$. 
Example: solution for Increasing $\lambda$, $D = I$. 
Choice of $\lambda$ crucial

- Different algorithms yield different solutions.
- Examples:
  - Discrepancy Principle
  - Generalized Cross Validation (GCV)
  - L-Curve
  - Unbiased Predictive Risk (UPRE)
  - Residual Periodogram and related approaches (O’Leary et al)
- General Difficulties
  - Expensive (GCV, L, UPRE)
  - Not necessarily unique solution (GCV)
  - Oversmoothing (Discrepancy)
  - No kink in the L-curve
  - Require some analysis of resulting data.

A new $\chi^2$ result extending the discrepancy principle
Background: Statistics of the Least Squares Problem

**Theorem (Rao73: First Fundamental Theorem)**

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

\[
J = \min_x \|Ax - b\|^2 \sim \sigma_b^2 \chi^2(m - r).
\]

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

**Corollary (Weighted Least Squares)**

For \( b \sim N(Ax, C_b) \), \( W_b = C_b^{-1} \) then

\[
J = \min_x \|Ax - b\|^2_{W_b} \sim \chi^2(m - r).
\]

Sanity check: matrix \( A \) is square and full rank \( m = r \) mean(\( J \)) = 0.
Extension: Statistics of the Regularized Least Squares Problem

Theorem: $\chi^2$ distribution of the regularized functional (Renaut/Mead 2008)

NOTE: Weighting Matrix on Regularization term.

\[ \hat{x} = \arg\min_J J_D(x) = \arg\min \left\{ \|Ax - b\|_{W_b}^2 + \|(x - x_0)\|_{W_D}^2 \right\}, \quad W_D = D^T W_x D. \tag{2} \]

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: Errors in the right hand side $e \sim N(0, C_b)$, and $x_0$ is known so that $(x - x_0) = f \sim N(0, C_D)$,
- $x_0$ is the mean vector of the model parameters.

Then

\[ J_D(\hat{x}(W_D)) \sim \chi^2(m + p - n) \]
**Significance of the $\chi^2$ result**

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

For sufficiently large $\tilde{m} = m + p - n$

\[ E(J(x(W_D))) = m + p - n \quad E(JJ^T) = 2(m + p - n) \]

Moreover

\[ \tilde{m} - \sqrt{2\tilde{m}}z_{\alpha/2} < J(\hat{x}(W_D)) < \tilde{m} + \sqrt{2\tilde{m}}z_{\alpha/2}. \quad (3) \]

$z_{\alpha/2}$ is the relevant $z$-value for a $\chi^2$-distribution with $\tilde{m} = m + p - n$ degrees
Key Aspects of the Proof I: The Functional $J$

**Algebraic Simplifications: Rewrite functional as quadratic form**

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

\[
\hat{x} = x_0 + (A^T W_b A + D^T W_x D)^{-1} A^T W_b r, \quad (4)
\]

\[
= x_0 + R(W_D) W_b^{1/2} r, \quad r = b - A x_0
\]

\[
= x_0 + y(W_D). \quad (5)
\]

\[
R(W_D) = (A^T W_b A + D^T W_x D)^{-1} A^T W_b^{1/2} \quad (6)
\]

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

\[
A(W_D) = W_b^{1/2} A R(W_D) \quad (7)
\]

\[
J_D(\hat{x}) = r^T W_b^{1/2} (I_m - A(W_D)) W_b^{1/2} r, \quad \text{let} \quad \tilde{r} = W_b^{1/2} r \quad (8)
\]

\[
= \tilde{r}^T (I_m - A(W_D)) \tilde{r}. \quad \text{A Quadratic Form} \quad (9)
\]
Key Aspects of the Proof II: Properties of a Quadratic Form

\( \chi^2 \) distribution of Quadratic Forms \( x^T P x \) 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 \( x^T P x \) 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 \( \mu_i \neq 0 \), \( x^T P x \) 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 \( U \in \mathbb{R}^{m \times m} \), \( V \in \mathbb{R}^{p \times p} \), and a nonsingular matrix \( X \in \mathbb{R}^{n \times n} \) such that

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

(10)

\( \Upsilon = \text{diag}(\nu_1, \ldots, \nu_p, 1, \ldots, 1) \in \mathbb{R}^{n \times n}, \quad M = \text{diag}(\mu_1, \ldots, \mu_p) \in \mathbb{R}^{p \times p}, \)

(11)

\( 0 \leq \nu_1 \leq \cdots \leq \nu_p \leq 1, \quad 1 \geq \mu_1 \geq \cdots \geq \mu_p > 0, \quad \nu_i^2 + \mu_i^2 = 1, \quad i = 1, \ldots, p. \)

The Functional with the GSVD

Let \( \tilde{Q} = \text{diag}(\mu_1, \ldots, \mu_p, 0_{n-p}, I_{m-n}) \)

then \( J = \tilde{r}^T(I_m - A(W_D))\tilde{r} = \|\tilde{Q}U^T\tilde{r}\|_2^2 = \|k\|_2^2 \)
Proof IV: Statistical Distribution of the Weighted Residual

Covariance Structure

Errors in $b$ are $\mathbf{e} \sim N(0, \mathbf{C}_b)$. Now $b$ depends on $x$, $b = Ax$ hence we can show $b \sim N(Ax_0, \mathbf{C}_b + A\mathbf{C}_dA^T)$ ($x_0$ is mean of $x$)

Residual $\mathbf{r} = b - Ax \sim N(0, \mathbf{C}_b + A\mathbf{C}_dA^T)$.

$\tilde{\mathbf{r}} = W_b^{1/2}\mathbf{r} \sim N(0, I + \tilde{A}\mathbf{C}_d\tilde{A}^T)$, $\tilde{A} = W_b^{1/2}A$.

Use the GSVD

$I + \tilde{A}\mathbf{C}_d\tilde{A}^T = UQ^{-2}U^T$, $Q = \text{diag}(\mu_1, \ldots, \mu_p, I_{n-p}, I_{m-n})$

Now $\mathbf{k} = QU^T\tilde{r}$ then $\mathbf{k} \sim N(0, QU^T(UQ^{-2}U^T)QU) \sim N(0, I_m)$

But $J = ||QU^T\tilde{r}||^2 = ||\tilde{\mathbf{k}}||^2$, where $\tilde{\mathbf{k}}$ is the vector $\mathbf{k}$ excluding components $p + 1 : n$. Thus

$J_D \sim \chi^2(m + p - n)$. 
When mean of the parameters is not known, or \( x_0 = 0 \) is not the mean

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

Recall

\[
\hat{x} = \text{argmin } J_D(x) = \text{argmin}\{ \|Ax - b\|_{W_b}^2 + \| (x - x_0) \|_{W_D}^2 \}, \quad W_D = D^T W_x D.
\]

Assume all assumptions as before, but \( \bar{x} \neq x_0 \) is the mean vector of the model parameters.
Let

\[
c = \|c\|_2^2 = \| \tilde{Q} U^T W_b^{1/2} A (\bar{x} - x_0) \|_2^2
\]

Then

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

The functional at optimum follows a non central \( \chi^2 \) distribution
A further result when $A$ is not of full column rank

The Rank Deficient Solution

Suppose $A$ is not full column rank. Then the filtered solution can be written in terms of the GSVD

$$x_{\text{FILT}}(\lambda) = \sum_{i=p+1-r}^{p} \frac{\gamma_i^2}{\nu_i(\gamma_i^2 + \lambda^2)} s_i \tilde{x}_i + \sum_{i=p+1}^{n} s_i \tilde{x}_i = \sum_{i=1}^{p} \frac{f_i}{\nu_i} s_i \tilde{x}_i + \sum_{i=p+1}^{n} s_i \tilde{x}_i.$$ 

Here $f_i = 0, i = 1 : p - r, f_i = \gamma_i^2 / (\gamma_i^2 + \lambda^2), i = p - r + 1 : p$. This yields

$$J(x_{\text{FILT}}(\lambda)) \sim \chi^2(m - n + r, c)$$

If rank reduction can be found, degrees of freedom are reduced.
The Cost Functional follows a $\chi^2$ Statistical Distribution

- Suppose degrees of freedom $\tilde{m}$ and centrality parameter $c$ then

$$E(J_D) = \tilde{m} + c \quad E(J_D J_D^T) = 2(\tilde{m}) + 4c$$

- Can we use this?
- **YES**: First Steps:
  - Try to find $W_D$ so that $E(J) = \tilde{m} + c$
  - Mead presented expensive **nonlinear** algorithm when $c = 0$ for general $W_D$.
  - First find $\lambda$ only.

Find $W_x = \lambda^2 I$
What do we need to apply the Theory?

Requirements

- **Covariance** $C_b$ on data parameters $b$ (or on model parameters $x$!)
- A priori information $x_0$, mean $\bar{x}$.
- But $\bar{x}$ (and hence $x_0$) are not known.
- If not known use repeated data measurements calculate $C_b$ and mean $\bar{b}$.
- Hence estimate the **centrality** parameter $E(b) = A E(x)$ implies $\bar{b} = A \bar{x}$.

Hence

$$c = \|c\|_2^2 = \|\tilde{Q}U^T W_b^{1/2} (\bar{b} - A x_0)\|_2^2$$

$$E(J_D) = E(\|\tilde{Q}U^T W_b^{1/2} (b - A x_0)\|_2^2) = m + p - n + \|c\|_2^2$$

- Given the GSVD estimate the degrees of freedom $\tilde{m}$.

Then we can use $E(J)$ to find $\lambda$
Assume $x_0$ is the mean (experimentalists do know something about the model parameters)

**DESIGNING THE ALGORITHM: I**

- Recall: if $C_b$ and $C_x$ are good estimates of covariance

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

should be small.

**GOAL**

Find $W_x$ to make (3) tight: Single Variable case find $\lambda$

$$J_D(\hat{x}(\lambda)) \approx \tilde{m}$$
A Newton-line search Algorithm to find $\lambda = 1/\sigma$. (Basic algebra)

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

- We use $\sigma = 1/\lambda$, and $y(\sigma^{(k)})$ is the current solution for which
  
  $$x(\sigma^{(k)}) = y(\sigma^{(k)}) + x_0$$

  Then
  
  $$\frac{\partial}{\partial \sigma} J(\sigma) = -\frac{2}{\sigma^3} \|Dy(\sigma)\|^2 < 0$$

- Hence we have a basic Newton Iteration
  
  $$\sigma^{(k+1)} = \sigma^{(k)} (1 + \frac{1}{2} (\frac{\sigma^{(k)}}{\|Dy\|})^2 (J_D(\sigma^{(k)}) - \tilde{m})).$$

- Add a line search
  
  $$\sigma^{(k+1)} = \sigma^{(k)} (1 + \frac{\alpha^{(k)}}{2} (\frac{\sigma^{(k)}}{\|Dy\|})^2 (J_D(\sigma^{(k)}) - \tilde{m})).$$
Discussion on Convergence

- \( F \) is **monotonic decreasing** \( (F'(\sigma_x) = -2\sigma_x\|Dy\|_2^2) \)
- Solution either exists and is **unique** for positive \( \sigma \)
- **Or no solution exists** \( F(0) < 0 \).
  - implies incorrect statistics of the model
- Theoretically, \( \lim_{\sigma \to \infty} F > 0 \) possible.
  - Equivalent to \( \lambda = 0 \). No regularization needed.

![Graph 1](image1.png)
![Graph 2](image2.png)
Practical Details of Algorithm

Find the parameter

- **Step 1**: Bracket the root by logarithmic search on $\sigma$ to handle the asymptotes: yields $\text{sigmamax}$ and $\text{sigmamin}$

- **Step 2**: Calculate step, with steepness controlled by $\text{tolD}$. Let $t = D y / \sigma^{(k)}$, where $y$ is the current update, then

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

- **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.
Prominent 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 can concurrently regularize and find $\lambda$
- Standard hybrid LSQR solves projected system then adds regularization.
- This is also possible: results of both approaches.

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.
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.
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 $x$ from $Ax = b$, where $A$ is PSF matrix and $b$ is given blurred signal.
- Method of Comparison- no exact solution known: use convergence with respect to downsampling.
Comparison High Resolution White noise

Greater contrast with $\chi^2$. UPRE is insufficiently regularized. L-curve severely undersmooths (not shown). Parameters not consistent across resolutions.
THE UPRE SOLUTION: $x_0 = 0$ White Noise

Regularization Parameters are consistent: $\sigma = 0.01005$ all resolutions
Regularization quite consistent resolution 2 to 100
\[ \sigma = 0.0000029, .0000029, .0000029, .0000057, .0000057 \]
Illustrating the Deblurring Result: Problem Size 65536

Example taken from RESTORE TOOLS Nagy et al 2007-8: 15% Noise

True  Blurred  Chi

Computational Cost is Minimal: Projected Problem Size is 15, $\lambda = 0.58$
Illustrating the progress of the Newton algorithm post LSQR

\[
\begin{align*}
\sigma &= 1 \quad \text{snr} = 3.4 \\
\sigma &= 2.3 \quad \text{snr} = 2.1 \\
\sigma &= 1.1 \quad \text{snr} = 3.4 \\
\sigma &= 10 \quad \text{snr} = -5.3 \\
\sigma &= 1.5 \quad \text{snr} = 3.2 \\
\sigma &= 1.1 \quad \text{snr} = 3.4 \\
\sigma &= 3.6 \quad \text{snr} = -0.15 \\
\sigma &= 1.2 \quad \text{snr} = 3.4 \\
\text{LSQR snr} &= -8.3
\end{align*}
\]
Illustrating the progress of the Newton algorithm with LSQR

\[ \sigma = 1 \text{ snr} = 3.4 \]

\[ \sigma = 10 \text{ snr} = -5.3 \]

\[ \sigma = 5.2 \text{ snr} = -2.2 \]

\[ \sigma = 1.1 \text{ snr} = 3.4 \]

\[ \sigma = 1.1 \text{ snr} = 3.4 \]

LSQR snr = -8.3
Problem Grain noise 15% added for increasing subproblem size

Signal to noise ratio $10 \log_{10}(1/e)$
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.
- **NOTE THAT THE PROBLEM SIZE OF HYBRID LSQR USED IS VERY SMALL**
- Requires estimate of column rank of $A$.
- Method is very efficient, Newton method is robust and fast.
- But *a priori* information is needed. This can be obtained directly from the data. e.g. Use local statistical information of image
Future Work

Other Results and Future Work

- Preconditioning
- How to apply Picard condition for GSVD to handle problems with robustness due to conditioning of $C_b$
- Software Package!
- Properties of the residual for finding rank?
- Diagonal Weighting Schemes
- Edge preserving regularization - Total Variation
- Bound Constraints (with Mead accepted).