
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 $x$. Signal $b = Ax + n$

$A$ is discretization of point spread function (PSF), the blur.

$$b(t) = \int K(t - s)x(s)ds$$

$A$ discretizes $K$ PSF
Example 1-D Signal: **Restore Signal**

**The Unregularized Solution: Illustrates Sensitivity**

![Graph showing the unregularized solution for a 1-D signal, illustrating sensitivity.](image)
Two dimensional Image Deblurring

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 \( R(x) = \|D(x - x_0)\|^2 \): Tikhonov Regularized

### Generalized Tikhonov regularization: Given matrix \( D \) that is suitable.

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

- Assume \( \mathcal{N}(A) \cap \mathcal{N}(D) = \emptyset \)
- \( x_0 \) is a reference solution, often \( x_0 = 0 \).
- Solution
  
  \[
  \hat{x}(\lambda) = \arg\min J(x) = \arg\min \{\|Ax - b\|_W^2 + \lambda^2 \|D(x - x_0)\|^2\}. \tag{2}
  \]

### 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 $\lambda$, $D = I$. 

![Graph showing the solution for increasing lambda, with $D = I$.]
Solution for Increasing $\lambda$, $D = I$. 
Solution for Increasing $\lambda$, $D = I$. 

![Graph showing the solution for increasing $\lambda$, where $D = I$.]
Solution for Increasing $\lambda$, $D = I$. 
Solution for Increasing $\lambda, D = I.$
Solution for Increasing $\lambda$, $D = I$. 

![Graph showing solution for increasing $\lambda$ with $D = I$.](image)
Solution for Increasing $\lambda$, $D = I$. 

![Graph showing solution for increasing $\lambda$, with $D = I$.](image-url)
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: \( C_b = \sigma_b^2 I \).
- Find \( \lambda \) such that the regularized residual satisfies

\[
\sigma_b^2 = \frac{1}{m} \| b - Ax(\lambda) \|_2^2.
\]  

(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 
  \[ \frac{\|b - Ax(\lambda)\|^2_{W_b}}{[\text{trace}(I_m - A(\lambda))]^2}, \]
  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
  \[
  \| b - Ax(\lambda) \|^2_{W_b} + 2 \text{trace}(A(\lambda)) - m
  \]
- 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 \( Ax \approx b \).
- 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 $b \sim N(Ax, \sigma_b^2 I)$, (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 $b \sim N(Ax, C_b)$, and $W_b = C_b^{-1}$ then

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

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

Assume

- \( W_b \) 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: \( (b - Ax) = e \sim N(0, C_b) \), \( (x - x_0) = f \sim N(0, C_D) \)
  - \( x_0 \) 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 $R(W_D)$

\[
\hat{x} = x_0 + (A^T W_b A + D^T W_x D)^{-1} A^T W_b \mathbf{r},
\]
\[
= x_0 + R(W_D) W_b^{1/2} \mathbf{r}, \quad \mathbf{r} = b - Ax_0
\]
\[
= x_0 + y(W_D).
\]
\[
R(W_D) = (A^T W_b A + D^T W_x D)^{-1} A^T W_b^{1/2}
\]

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

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

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

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

\[
0 \leq \nu_1 \leq \cdots \leq \nu_p \leq 1, \quad 1 \geq \mu_1 \geq \cdots \geq \mu_p > 0, \quad \nu_i^2 + \mu_i^2 = 1, \quad i = 1, \ldots, p. \tag{12}
\]

**The Functional with the GSVD**

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

then \( J = \tilde{r}^T(I_m - A(W_D))\tilde{r} = \|\tilde{Q}U^T\tilde{r}\|_2^2 \),
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

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

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

\[
0 \leq \nu_1 \leq \cdots \leq \nu_p \leq 1, \quad 1 \geq \mu_1 \geq \cdots \geq \mu_p > 0, \tag{12}
\]

\[
\nu_i^2 + \mu_i^2 = 1, \quad i = 1, \ldots p.
\]

The Functional with the GSVD

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

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

### Covariance Structure

- **e = Ax − b ∼ N(0, C_b)** hence we can show b ∼ N(Ax_0, C_b + AC_DA^T)
  - Note that b depends on x.
- **r ∼ N(0, C_b + AC_DA^T)**, and \( \tilde{r} \sim N(0, I + \tilde{A}C_D\tilde{A}^T) \), \( \tilde{A} = W_b^{1/2}A \).
- Use the GSVD
  \[
  I + \tilde{A}C_D\tilde{A}^T = UQ^{-2}U^T,
  \]
  \[
  Q = \text{diag}(\mu_1, \ldots, \mu_p, I_{n-p}, I_{m-n})
  \]

### The Functional is a rv

- Let \( k = QU^T\tilde{r} \), then \( k \sim N(0, QU^T(UQ^{-2}U^T)UQ) \sim N(0, I_m) \)
- But \( J = \|\tilde{Q}U^T\tilde{r}\|^2 = \|\tilde{k}\|^2 \), where \( \tilde{k} \) is the vector 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**

- \( e = Ax - b \sim N(0, C_b) \) hence we can show \( b \sim N(Ax_0, C_b + AC_D A^T) \)
  
  Note that \( b \) depends on \( x \).

- \( r \sim N(0, C_b + AC_D A^T) \), and \( \tilde{r} \sim N(0, I + \tilde{A}C_D \tilde{A}^T) \), \( \tilde{A} = W_b^{1/2}A \).

- Use the GSVD

  \[
  I + \tilde{A}C_D \tilde{A}^T = UQ^{-2}U^T,
  
  Q = \text{diag}(\mu_1, \ldots, \mu_p, I_{n-p}, I_{m-n})
  \]

**The Functional is a rv**

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

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

  \[
  J_D \sim \chi^2(m + p - n)
  \]
Corollary: a-priori information not mean value, e.g. $x_0 = 0$

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

$$\hat{x} = \arg\min J_D(x) = \arg\min \{ \|Ax - b\|_W^2 + \| (x - x_0) \|_{W_D}^2 \}, \quad W_D = D^T W_x D. \quad (13)$$

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

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

**Statistical Distribution of the Functional**

- Mean and Variance are prescribed

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

- 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 $C_b$ on data parameters $b$ (or on model parameters $x$!)
- **A priori** information either $x_0$ is the mean, or mean value $\bar{x}$.
- But $\bar{x}$ and $x_0$ are not known.
- For repeated data measurements $C_b$ can be calculated. Also $\bar{b}$ can be found, the mean of $b$.
- But $E(b) = AE(x)$ implies $\bar{b} = A\bar{x}$. Hence

$$c = \|c\|_2^2 = \|\tilde{Q}U^TW_b^{1/2}(\bar{b} - Ax_0)\|_2^2$$

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.

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

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

- $z_{\alpha/2}$ 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 $\lambda$

$$J_D(\hat{x}(\lambda)) \approx \tilde{m}$$
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 $y(\sigma^{(k)})$ 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
  \[ \sigma^{(k+1)} = \sigma^{(k)} \left( 1 + \frac{1}{2} \left( \frac{\sigma^{(k)}}{||Dy||}\right)^2(J_D(\sigma^{(k)}) - \tilde{m}) \right) . \]

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

- $F$ is **monotonic decreasing** ($F'(\sigma_x) = -2\sigma_x \|t\|_2^2$)
- Solution either exists and is **unique** for positive $\sigma$
- **Or no solution exists** $F(0) < 0$.
  - implies incorrect statistics of the model
- Theoretically, $\lim_{\sigma \to \infty} F > 0$ 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 $\text{sigmamax}$ 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, given from the GSVD, 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.
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 $A^Tx$, 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 $C_{\mathbf{b}} = \sigma_{\mathbf{b}}^2 I$ for common covariance, **white noise**.
- Use $C_{\mathbf{b}} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$ for **colored uncorrelated noise**.
- With no noise information $C_{\mathbf{b}} = I$.
- Use $\bar{\mathbf{b}}$ as the mean of measured $\mathbf{b}$, when implemented with centrality parameter, $x_0 = 0$. 
Illustrating the Deblurring Result: Problem Size 65536

Computational Cost is Minimal: Projected Problem Size is $15, \lambda = 0.58$
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.
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.
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

$\sigma = 0.0000029, 0.0000029, 0.0000029, 0.0000057, 0.0000057$ (left)

$\sigma = 0.000007, 0.000007, 0.000007, 0.000007, 0.00012$ (right).

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