Newton's method for obtaining the Regularization Parameter for Regularized Least Squares

Rosemary Renaut, Jodi Mead

Arizona State and Boise State

International Linear Algebra Society, Cancun

Outline

Least Squares for $A\mathbf{x} = \mathbf{b}$, (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

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

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

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

  - $\tilde{\mathbf{e}}$ 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:

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

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

- **Standard:** $W_{\mathbf{x}} = \lambda^2 I$, λ unknown penalty parameter.
- Statistically, $W_{\mathbf{x}}$ is inverse covariance matrix for the model $\mathbf{x}$ i.e. $\lambda = 1/\sigma_{\mathbf{x}}$, $\sigma^2_{\mathbf{x}}$ 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

What is the *correct* regularization parameter λ?
More generally, what is the correct $W_{\mathbf{x}}$?

Weighted Regularized Least Squares for numerically ill-posed systems

Formulation:

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

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

- **Standard:** $W_{\mathbf{x}} = \lambda^2 I$, λ unknown penalty parameter.
- Statistically, $W_{\mathbf{x}}$ is inverse covariance matrix for the model $\mathbf{x}$ i.e. $\lambda = 1/\sigma_{\mathbf{x}}$, $\sigma^2_{\mathbf{x}}$ 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

What is the *correct* regularization parameter λ?
More generally, what is the correct $W_{\mathbf{x}}$?

Formulation: Regularization with Solution Mapping

Generalized Tikhonov regularization, operator D acts on **x**.

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

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

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

GOAL

- Use statistics of the solution to find λ.
- Can we estimate λ efficiently when $W_\mathbf{b}$ is known?

Formulation: Regularization with Solution Mapping

Generalized Tikhonov regularization, operator D acts on **x**.

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

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

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

GOAL

- Use statistics of the solution to find λ.
- Can we estimate λ efficiently when $W_\mathbf{b}$ is known?

Other approaches

1. L-curve - trades off residual and the regularization.
2. Generalized Cross-Validation (GCV)- (Statistical) Minimizes

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

   which estimates predictive risk.
3. Unbiased Predictive Risk Estimator (UPRE) - Statistical - Minimize expected value of predictive risk:

$$\|\mathbf{b} - A\mathbf{x}(\lambda)\|^2_{W_{\mathbf{b}}} + 2\,\mathrm{trace}(A(\lambda)) - m$$

Problems

- Require generally multiple solutions
- Expensive to calculate
- Not always sufficient - might not converge to minimum, multiple minima

Other approaches

1. L-curve - trades off residual and the regularization.
2. Generalized Cross-Validation (GCV)- (Statistical) Minimizes

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

   which estimates predictive risk.
3. Unbiased Predictive Risk Estimator (UPRE) - Statistical - Minimize expected value of predictive risk:

$$\|\mathbf{b} - A\mathbf{x}(\lambda)\|^2_{W_\mathbf{b}} + 2\,\mathrm{trace}(A(\lambda)) - m$$

Problems

- Require generally multiple solutions
- Expensive to calculate
- Not always sufficient - might not converge to minimum, multiple minima

More Recent Developments

1. Residual Periodogram (O'Leary and Rust)
2. The core problem reduction (Page, Strakos, Hnetynkova)
3. Projection Methods (Hansen, Kilmer, Espanol)
4. HyBR Nagy - Copper Mountain 2008

A Statistical Technique: Mead 2007

- Use the degrees of freedom of the functional
- Exploits the χ^2 distribution of the functional-designed to find W_D not just λ

.

1. Residual Periodogram (O'Leary and Rust)
2. The core problem reduction (Page, Strakos, Hnetynkova)
3. Projection Methods (Hansen, Kilmer, Espanol)
4. HyBR Nagy - Copper Mountain 2008

A Statistical Technique: Mead 2007

- Use the degrees of freedom of the functional
- Exploits the χ^2 distribution of the functional-designed to find $W_{\mathbf{D}}$ not just λ

.

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 χ^2 distribution with $m-r$ degrees of freedom.

Corollary (Weighted Least Squares)

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

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

ARIZONA STATE UNIVERSITY

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 χ^2 distribution with $m-r$ degrees of freedom.

Corollary (Weighted Least Squares)

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

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

Theorem: χ^2 distribution of the regularized functional

Assume

- Invertibility of the system $\mathcal{N}(A) \cap \mathcal{N}(D) \neq 0$,
- $(\mathbf{b} - A\mathbf{x}) \sim N(0, C_{\mathbf{b}})$, $D(\mathbf{x} - \mathbf{x}_0) \sim N(0, C_{\mathbf{D}})$

Then $J_D \sim \chi^2(m + p - n)$: ie is a random variable which follows a χ^2 distribution with $m + p - n$ degrees of freedom.

Proof: Sketch

Observe

$$J_D = \min_{\mathbf{x}} \| \begin{pmatrix} A \\ D \end{pmatrix} \mathbf{x} - \begin{pmatrix} \mathbf{b} \\ D\mathbf{x}_0 \end{pmatrix} \|_W^2 = \min_{\mathbf{x}} \| G\mathbf{x} - \mathbf{d} \|_W^2 \quad (4)$$

Definitions of system matrix $G \in \mathcal{R}^{(m+p)\times n}$, block diagonal weighting matrix W and right hand side vector $\mathbf{d} \in \mathcal{R}^{m+p}$ are immediate from the equation.

Result follows adapting from Rao:

Theorem: χ^2 distribution of the regularized functional

Assume

- Invertibility of the system $\mathcal{N}(A) \cap \mathcal{N}(D) \neq 0$,
- $(\mathbf{b} - A\mathbf{x}) \sim N(0, C_{\mathbf{b}})$, $D(\mathbf{x} - \mathbf{x}_0) \sim N(0, C_{\mathbf{D}})$

Then $J_D \sim \chi^2(m+p-n)$: ie is a random variable which follows a χ^2 distribution with $m+p-n$ degrees of freedom.

Proof: Sketch

Observe

$$J_D = \min_{\mathbf{x}} \| \begin{pmatrix} A \\ D \end{pmatrix} \mathbf{x} - \begin{pmatrix} \mathbf{b} \\ D\mathbf{x}_0 \end{pmatrix} \|_W^2 = \min_{\mathbf{x}} \|G\mathbf{x} - \mathbf{d}\|_W^2 \qquad (4)$$

Definitions of system matrix $G \in \mathcal{R}^{(m+p)\times n}$, block diagonal weighting matrix W and right hand side vector $\mathbf{d} \in \mathcal{R}^{m+p}$ are immediate from the equation.

Result follows adapting from Rao:

DESIGNING THE ALGORITHM: I

- If $C_\mathbf{b}$ and $C_\mathbf{D}$ are good estimates of the covariance matrices

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

should be **small.**

- Regularized solution given in terms of **resolution** matrix $R(W_\mathbf{D})$

$$\begin{aligned}
\hat{\mathbf{x}} &= \mathbf{x}_0 + (A^T W_\mathbf{b} A + D^T W_\mathbf{D} D)^{-1} A^T W_\mathbf{b} \mathbf{r}, & (5)\\
&= \mathbf{x}_0 + R(W_\mathbf{D}) {W_\mathbf{b}}^{1/2} \mathbf{r}, \quad \mathbf{r} = \mathbf{b} - A\mathbf{x}_0 & \\
&= \mathbf{x}_0 + \mathbf{y}(W_\mathbf{D}). & (6)\\
R(W_\mathbf{D}) &= (A^T W_\mathbf{b} A + D^T W_\mathbf{D} D)^{-1} A^T {W_\mathbf{b}}^{1/2} & (7)
\end{aligned}$$

DESIGNING THE ALGORITHM: II

- Functional is given in terms of **influence matrix** $A(W_{\mathbf{D}})$

$$J_D(\hat{\mathbf{x}}) = \mathbf{r}^T {W_{\mathbf{b}}}^{1/2}(I_m - A(W_{\mathbf{D}})){W_{\mathbf{b}}}^{1/2}\mathbf{r}, \quad (8)$$

$$A(W_{\mathbf{D}}) = {W_{\mathbf{b}}}^{1/2} A R(W_{\mathbf{D}}) \quad (9)$$

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

$$\tilde{m} - \sqrt{2\tilde{m}} z_{\alpha/2} < \mathbf{r}^T {W_{\mathbf{b}}}^{1/2}(I_m - A(W_{\mathbf{D}})){W_{\mathbf{b}}}^{1/2}\mathbf{r} < \tilde{m} + \sqrt{2\tilde{m}} z_{\alpha/2}. \quad (10)$$

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

GOAL

Find $W_{\mathbf{D}}$ to make (10) tight: Single Variable case find λ

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

DESIGNING THE ALGORITHM: II

- Functional is given in terms of **influence matrix** $A(W_{\mathbf{D}})$

$$J_D(\hat{\mathbf{x}}) = \mathbf{r}^T {W_{\mathbf{b}}}^{1/2}(I_m - A(W_{\mathbf{D}})){W_{\mathbf{b}}}^{1/2}\mathbf{r}, \quad (8)$$

$$A(W_{\mathbf{D}}) = {W_{\mathbf{b}}}^{1/2} A R(W_{\mathbf{D}}) \quad (9)$$

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

$$\tilde{m} - \sqrt{2\tilde{m}} z_{\alpha/2} < \mathbf{r}^T {W_{\mathbf{b}}}^{1/2}(I_m - A(W_{\mathbf{D}})){W_{\mathbf{b}}}^{1/2}\mathbf{r} < \tilde{m} + \sqrt{2\tilde{m}} z_{\alpha/2}. \quad (10)$$

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

GOAL

Find $W_{\mathbf{D}}$ to make (10) tight: Single Variable case find λ

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

A Newton-line search Algorithm to find λ.

Newton Step

- We use $\sigma = 1/\lambda$, and $\mathbf{y}(\sigma^{(k)})$ is the current solution,

$$\mathbf{x}(\sigma^{(k)}) = \mathbf{y}(\sigma^{(k)}) + \mathbf{x}_0$$

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

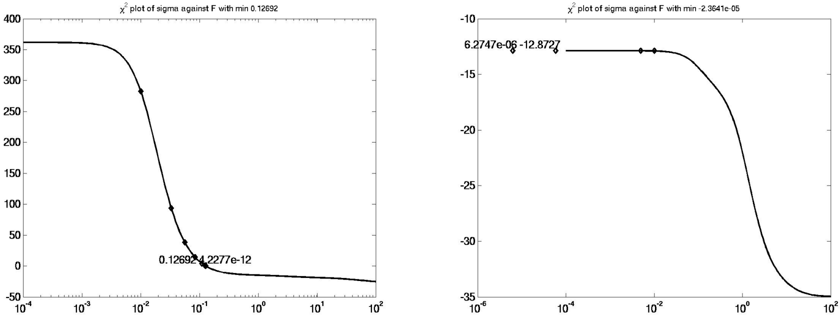
- Then to solve $F(\sigma) = J_D(\sigma) - \tilde{m} = 0$:
- Newton Iteration

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

Arizona State University

Convergence

- F is **monotonic decreasing**
- Solution either exists and is **unique** for positive σ
- **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

Algorithm

- **Step 1:** Bracket the root by logarithmic search on σ to handle the asymptotes: yields $\mathrm{sigmamax}$ and $\mathrm{sigmamin}$
- **Step 2:** Calculate step, with steepness controlled by tolD

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

- **Step 3:** Introduce line search $\alpha^{(k)}$ in Newton

$$\mathrm{sigmanew} = \sigma^{(k)}(1 + \alpha^{(k)}\mathrm{step})$$

$\alpha^{(k)}$ chosen such that $\mathrm{sigmanew}$ within bracket.

Assumptions

Covariance of Error: Statistics of Measurement Errors

- Information on the covariance structure of errors in **b** needed.
- Use $C_{\mathbf{b}} = \sigma_{\mathbf{b}}^2 I$ for common covariance, white noise.
- Use $C_{\mathbf{b}} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$ for colored uncorrelated noise.
- With no noise information $C_{\mathbf{b}} = I$.

Tolerance on Convergence

- The convergence tolerance depends on the noise structure.
- Use $\mathrm{TOL} = \sqrt{2\tilde{m}}z_{\alpha/2}$.
- 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

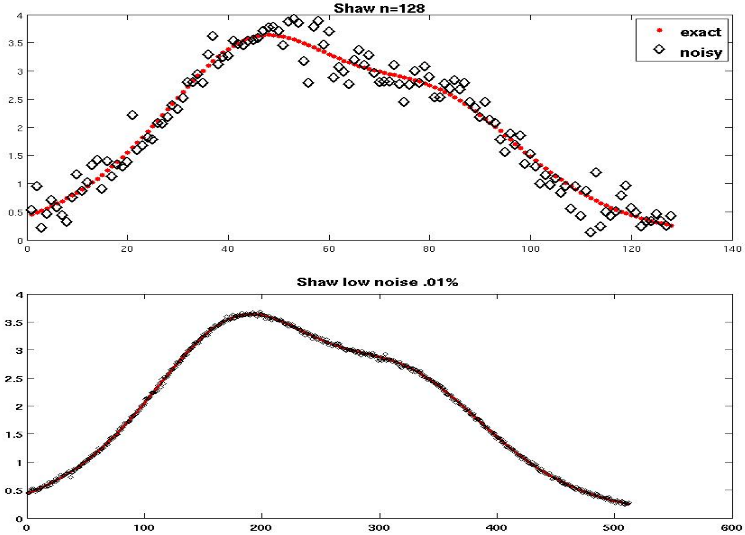
- Information on the covariance structure of errors in **b** needed.
- Use $C_{\mathbf{b}} = \sigma_{\mathbf{b}}^2 I$ for common covariance, white noise.
- Use $C_{\mathbf{b}} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$ for colored uncorrelated noise.
- With no noise information $C_{\mathbf{b}} = I$.

Tolerance on Convergence

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

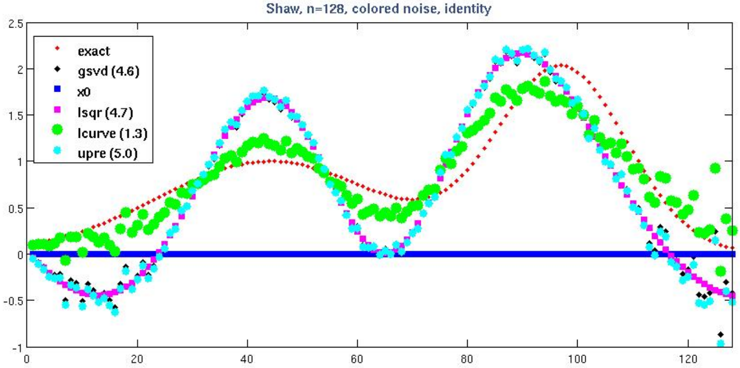
ARIZONA STATE UNIVERSITY

Error on the Measurements (10%) noise



An Example : Shaw Hansen's Toolbox (10%) noise

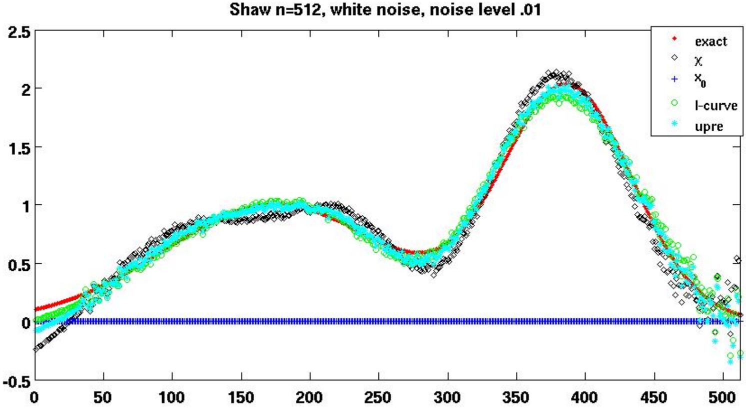
Comparative Results: $D = I$, $C_{\mathbf{b}} = \sigma_{\mathbf{b}}^2 I$



Observations

- Statistical Methods UPRE(cyan) and χ^2 (black and magenta) give similar solutions.
- L-curve (green) is very different

Comparative Results: $D = I$, $C_{\mathbf{b}} = \sigma_{\mathbf{b}}^2 I$



Observations

- Solutions are comparable for smaller noise level

Experiments

- Take example from Hansen's toolbox, eg `shaw`, `phillips`.
- Generate 100 copies for each noise level .001, .005, .01, .05, .1.
- Solve for 100 cases using GSVD and inexact Newton with increasing number of inner iterations.
- Do pairwise t test on the obtained σ to determine dependence on inexact solves
- Determine convergence speed: Number of Newton Iterations.

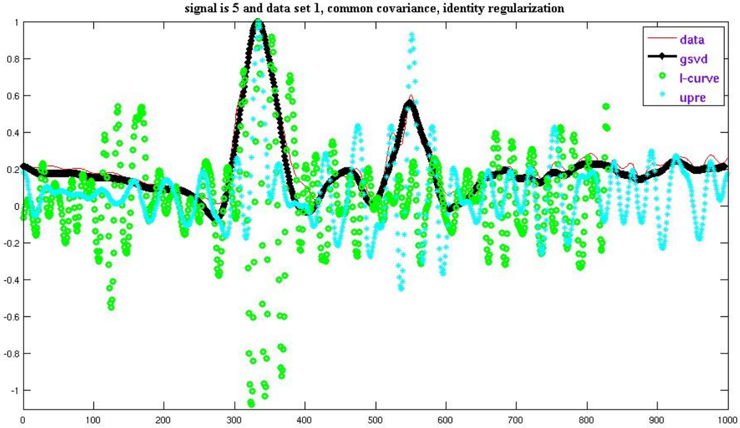
Major Observations

1. High correlation between σ for both Newton Methods (GSVD and LSQR approach) is obtained. $p-$ values at or near 1.
2. Algorithms converge with very few solution evaluations, typically fewer than 10.
3. With LSQR approach we may use inexact solves each Newton step.
4. With LSQR approach we will be able to reuse solves from each Newton step.

Example: Seismic Signal Restoration

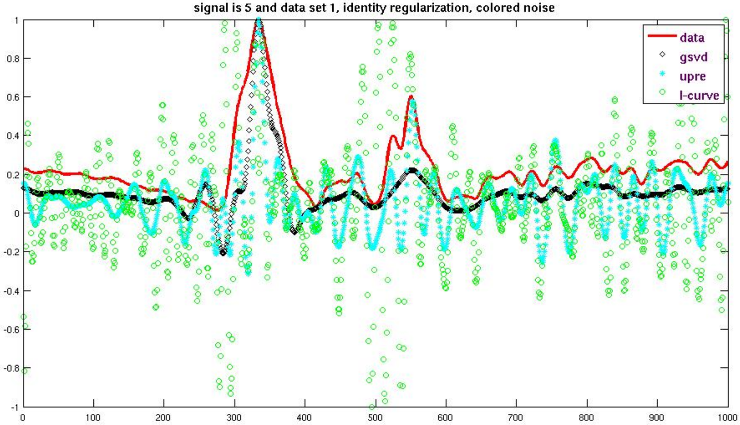
- Real data set of 48 signals of length 1000.
- The point spread function is derived from the signals
- Solve $Pf = g$, where P is psf matrix, g is signal and restore f.
- Calculate the signal variance pointwise over all 48 signals.

An Example of Comparison with L-curve and UPRE Solutions: white noise



Regularization parameters: χ : 1716 l:5 UPRE:99

An Example of Comparison with L-curve and UPRE Solutions: colored noise



Observation: Greater contrast at major peak for χ method.

Regularization parameters: χ : 1499 l:4 UPRE:99

Conclusions and Future Work

Conclusions

- A new statistical method for estimating regularization parameter
- Method can be used for large scale problems, without GSVD
- Method is very efficient, Newton method is robust
- Method converges with very few solution evaluations
- For Multiple right hand side, calculate $C_{\mathbf{b}}$

Future Work

- Standard Form Implementation and image deblurring (with Hnetynkova in prep.)
- Diagonal Weighting Schemes
- Edge preserving regularization
- Constraints - See Mead Friday ILAS Cancun

Conclusions

- A new statistical method for estimating regularization parameter
- Method can be used for large scale problems, without GSVD
- Method is very efficient, Newton method is robust
- Method converges with very few solution evaluations
- For Multiple right hand side, calculate $C_{\mathbf{b}}$

Future Work

- Standard Form Implementation and image deblurring (with Hnetynkova in prep.)
- Diagonal Weighting Schemes
- Edge preserving regularization
- Constraints - See Mead Friday ILAS Cancun

THANK YOU!

GSVD

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

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

- γ_i are the generalized singular values
- $\tilde{m} = m - n + p - \sum_{i=1}^{p} s_i^2 \delta_{\gamma_i 0} - \sum_{i=n+1}^{m} s_i^2$,
- $\tilde{s}_i = s_i/(\gamma_i^2 \sigma_{\mathbf{x}}^2 + 1)$, $i = 1, \dots, p$
- $t_i = \tilde{s}_i \gamma_i$.
- Find root of $\sum_{i=1}^{p} (\frac{1}{\gamma_i^2 \sigma_{\mathbf{x}}^2 + 1}) s_i^2 + \sum_{i=n+1}^{m} s_i^2 = m$
- Solve $F = 0$, where

$$F(\sigma_{\mathbf{x}}) = \mathbf{s}^T \tilde{\mathbf{s}} - \tilde{m} \quad \text{and} \quad F'(\sigma_{\mathbf{x}}) = -2\sigma_{\mathbf{x}} \|\mathbf{t}\|_2^2.$$

- Implementatation details the same.

Idempotent Quadratic Form

- Rewrite functional
  $J_D = \min_{\mathbf{x}} \|\tilde{G}\mathbf{x} - \tilde{\mathbf{d}}\|^2, \quad \tilde{G} = W^{1/2}G, \tilde{\mathbf{d}} = W^{1/2}\mathbf{d},$
- $\tilde{\mathbf{d}}$ are i.i.d., $\mathbf{d} \sim N(\tilde{G}\hat{\mathbf{x}}, I_{m+p})$
- J_D is minimum when $\tilde{G}\hat{\mathbf{x}}$ is projection P of $\tilde{\mathbf{d}}$ on $\mathrm{ran}(\tilde{G})$, $\tilde{G}\hat{\mathbf{x}} = P\tilde{\mathbf{d}}$
- $J_D(\hat{\mathbf{x}})$ is perpendicular distance to range of $\tilde{G}$:

$$J_D(\hat{\mathbf{x}}) = \tilde{\mathbf{d}}^T(P - I)^T(P - I)\tilde{\mathbf{d}} = \tilde{\mathbf{d}}^T(P - I)\tilde{\mathbf{d}}.$$

- A quadratic form for $\tilde{\mathbf{d}}$, $(I - P)$ **symmetric idempotent**, hence **non-central** χ^2 distribution

$$J_D(\hat{\mathbf{x}}) \sim \chi^2(m + p - n, \omega),$$

- But $\omega = J_D(E(\tilde{\mathbf{d}})) = 0$ due to $P\tilde{G} = \tilde{G}$.
- Hence $J_D(\hat{\mathbf{x}})$ is non-central, $J_D \sim \chi^2(m + p - n)$