1. Discretization of the Integral Equation
2. Galerkin approach for discretization leads to relationship of SVD and SVE
3. Discrete Picard Condition
4. Truncated SVD and filter factors
5. Basic Tikhonov Solution
6. Parameter determines the filtering
Review the Tikhonov Regularization
L-Curve for Parameter Estimation
Definition of Resolution/Influence Matrix
Review of Statistical Results
Unbiased Predictive Risk
Generalized Cross Validation
Rosemary Renaut
November 9, 2011
Least Squares Solution

Consider general overdetermined discrete problem

\[ A\mathbf{x} = \mathbf{b}, \quad A \in \mathbb{R}^{m \times n}, \quad \mathbf{b} \in \mathbb{R}^m, \quad \mathbf{x} \in \mathbb{R}^n, \quad m \geq n. \]

Fit to data functional of the least squares problem is \( \|A\mathbf{x} - \mathbf{b}\|^2 \)

Define \( \mathbf{x}_{LS} = \arg \min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|^2 \) and \( \hat{\mathbf{x}} \) by \( A\hat{\mathbf{x}} = \hat{\mathbf{b}} \), i.e. \( \hat{\mathbf{b}} \in \text{Range}(A) \)

We know that \( \mathbf{x}_{LS} \) is noise contaminated if \( A \) is ill-conditioned.

Example: noise in \( \mathbf{b} \) is \( \eta \sim \mathcal{N}(0, 10^{-7}) \) (normally distributed, mean 0 and variance \( 10^{-7} \))
Tikhonov Regularization

**Tikhonov Regularization:** add a penalty term with regularization parameter $\lambda > 0$

$$x(\lambda) = \arg\min_x \{ \|Ax - b\|^2 + \lambda^2 \|x\|^2 \}$$

Regularized solution trades of $\|x(\lambda)\|^2$ against $\|b - Ax(\lambda)\|^2$.
For small $\lambda$ the fit to data term is more closely enforced and the solution may be noisy.
For larger $\lambda$ the regularization term is enforced and so the solution is smoothed.
1-D Original and Noisy Signal

![Clean Signal](image1)

![Noisy Signal, SNR: 9.7dB, Variance 0.0018](image2)
Solution for Different Choices of $\lambda$

Solutions $x(\lambda)$

The Signal to Noise Ratio of the solution

$$10 \log_{10} \frac{\| \hat{x} \|}{\| \hat{x} - x(\lambda) \|},$$

true solution $\hat{x}$
Solution for Different Choices of $\lambda$

Solutions $\mathbf{x}(\lambda)$

The Signal to Noise Ratio of the solution

$$10 \log_{10} \frac{\|\hat{\mathbf{x}}\|}{\|\hat{\mathbf{x}} - \mathbf{x}(\lambda)\|},$$

true solution $\hat{\mathbf{x}}$
Solution for Different Choices of $\lambda$

The Signal to Noise Ratio of the solution

$$10 \log_{10} \frac{||\hat{x}||}{||\hat{x} - x(\lambda)||},$$
true solution $\hat{x}$
Solution for Different Choices of $\lambda$

Solutions $\mathbf{x}(\lambda)$

The Signal to Noise Ratio of the solution

$$10 \log_{10} \frac{\|\hat{\mathbf{x}}\|}{\|\hat{\mathbf{x}} - \mathbf{x}(\lambda)\|},$$

true solution $\hat{\mathbf{x}}$
Solution for Different Choices of $\lambda$

Solutions $\mathbf{x}(\lambda)$

The Signal to Noise Ratio of the solution

$$10 \log_{10} \frac{\|\hat{\mathbf{x}}\|}{\|\hat{\mathbf{x}} - \mathbf{x}(\lambda)\|},$$

true solution $\hat{\mathbf{x}}$
Solution for Different Choices of $\lambda$

Solutions $x(\lambda)$

The Signal to Noise Ratio of the solution

$10 \log_{10} \frac{\|\hat{x}\|}{\|\hat{x} - x(\lambda)\|}$,

true solution $\hat{x}$
Solution for Different Choices of $\lambda$

Solutions $\mathbf{x}(\lambda)$

The Signal to Noise Ratio of the solution

$$10 \log_{10} \frac{\| \hat{\mathbf{x}} \|}{\| \hat{\mathbf{x}} - \mathbf{x}(\lambda) \|}, \quad \text{true solution } \hat{\mathbf{x}}$$
Solutions $x(\lambda)$

The Signal to Noise Ratio of the solution

$$10 \log_{10} \frac{\| \hat{x} \|}{\| \hat{x} - x(\lambda) \|}, \quad \text{true solution } \hat{x}$$
Solution for Different Choices of $\lambda$

Solutions $\mathbf{x}(\lambda)$

The Signal to Noise Ratio of the solution

$$10 \log_{10} \frac{\|\hat{\mathbf{x}}\|}{\|\hat{\mathbf{x}} - \mathbf{x}(\lambda)\|},$$

true solution $\hat{\mathbf{x}}$
Solutions $x(\lambda)$

The Signal to Noise Ratio of the solution

$$10 \log_{10} \frac{\|\hat{x}\|}{\|\hat{x} - x(\lambda)\|},$$

true solution $\hat{x}$
Solution for Different Choices of $\lambda$

Solutions $\mathbf{x}(\lambda)$

The Signal to Noise Ratio of the solution

$$10 \log_{10} \frac{\|\hat{\mathbf{x}}\|}{\|\hat{\mathbf{x}} - \mathbf{x}(\lambda)\|},$$

true solution $\hat{\mathbf{x}}$
Solution for Different Choices of $\lambda$

Solutions $\mathbf{x}(\lambda)$

The Signal to Noise Ratio of the solution

$$10 \log_{10} \frac{\|\hat{\mathbf{x}}\|}{\|\hat{\mathbf{x}} - \mathbf{x}(\lambda)\|},$$
true solution $\hat{\mathbf{x}}$
Solution for Different Choices of $\lambda$

Solutions $x(\lambda)$

The Signal to Noise Ratio of the solution

$$10 \log_{10} \frac{\| \hat{x} \|}{\| \hat{x} - x(\lambda) \|},$$

true solution $\hat{x}$
Investigating the Regularized solution

\[ x(\lambda) = \arg \min_x \{ \| Ax - b \|_2^2 + \lambda^2 \| x \|_2^2 \} \]

has equivalent formulation

\[ x(\lambda) = \arg \min_x \left\| \begin{pmatrix} A \\ \lambda I_n \end{pmatrix} x - \begin{pmatrix} b \\ 0_n \end{pmatrix} \right\|_2^2 \]

In practice we use iterative methods for the second formulation. For theoretical discussion we solve the normal equations:

Regularization matrix

\[ x(\lambda) = R(\lambda)b \text{ where } R(\lambda) = \left( A^T A + \lambda^2 I \right)^{-1} A^T \]

Matrix \( R \) is also sometimes referred to as the generalized inverse. Denoted by \( A^\# \) in Vogel.
Bias and Variance in the Solution: exact values $\hat{x}$ and $\hat{b}$ not available

Model error (not computable)

$$e(\lambda) = x(\lambda) - \hat{x} = R(\lambda)(b) - \hat{x}$$

$$= R(\lambda)(\hat{b} + \eta) - \hat{x} = R(\lambda)A\hat{x} - \hat{x} + R(\lambda)\eta$$

$$= \underbrace{(R(\lambda)A - I_n)\hat{x}} + \underbrace{R(\lambda)\eta}$$

$$= \text{bias} \quad \text{variance}$$

$$= (V\Gamma V^T - I_n)\hat{x} + V\Gamma\Sigma^\dagger U^T\eta$$

**Bias** the loss of information introduced by regularization

Express $\hat{x} = VV^T\hat{x}$ and rewrite $(V\Gamma V^T - I_n)\hat{x}$ then we can filter so that the bias tends to zero:

$$\Gamma_{\text{TVD}} = \text{diag}(I_k, 0_{n-k}) \quad \left(\sum_{i=k+1}^n (v_i^T\hat{x})v_i\right) \to 0 \text{ as } k \to n$$

$$\Gamma_{\text{TIK}} = \text{diag}(\gamma_i) \quad \sum_{i=1}^n (1 - \gamma_i)(v_i^T\hat{x})v_i = (\sum_{i=1}^n \frac{\lambda^2}{\sigma_i^2 + \lambda^2}(v_i^T\hat{x})v_i) \to 0 \text{ as } \lambda \to 0$$

**Variance** is the amplification of the noise in the error - the perturbation error also tends to zero by appropriate choice of $\lambda$
Estimating the model error

First express the variance as a sum

\[ V \Gamma \Sigma^\dagger U^T \eta = \sum_{i=1}^{n} (u_i^T \eta) \frac{\gamma_i}{\sigma_i} v_i = \sum_{i=1}^{n} (u_i^T \eta) \frac{\sigma_i^2}{(\lambda^2 + \sigma_i^2)\sigma_i} v_i \]

Suppose in TSVD \( k \) is chosen by a threshold:

\[ \gamma_k = 1 \text{ for } \sigma_k^2 > \lambda^2, \text{ and } \gamma_i = 0, i > k. \]

Consider \( i \leq k \) then \( \gamma_i = 1 \leq \sigma_i / \lambda \). Thus \( \gamma_i \leq \sigma_i / \lambda \) for all \( i \).

Now for the Tikhonov filter observe \((\sigma_i^2/(\sigma_i^2 + \lambda^2))/\sigma_i = (\sigma_i + \lambda^2/\sigma_i)^{-1}\)

- For \( \sigma_i^2 > \lambda^2, \sigma_i + \lambda^2/\sigma_i > \lambda \)
- For \( \sigma_i^2 \leq \lambda^2, 1/\sigma_i > 1/\lambda, \) so \( \lambda^2/\sigma_i > \lambda \) and \( \sigma_i + \lambda^2/\sigma_i > \lambda \).

Hence in each case \( \gamma_i/\sigma_i = (\sigma_i + \lambda^2/\sigma_i)^{-1} \leq 1/\lambda \) and \( \gamma_i \leq \sigma_i / \lambda \) for all \( i \)

Suppose that \( \| \eta \|_2 < \delta^2 \) then

\[ \| \sum_{i=1}^{n} (u_i^T \eta) \gamma_i/\sigma_i v_i \|^2 = \sum_{i=1}^{n} |(u_i^T \eta)|^2 (\gamma_i/\sigma_i)^2 < (\delta/\lambda)^2. \]

If \( \lambda = \delta^p \) where \( p < 1 \) then as \( \delta \to 0 \) variance error goes to zero. Notice that we use \( \| U^T \eta \|^2 = \| \eta \|^2 \) for \( U \) orthogonal.

If in addition \( p > 0 \) then the bias also goes to zero.

We conclude that \( \lambda \) can be chosen so that \( e \) goes to zero for both TSVD and Tikhonov.
Other Measures of the Solution

Note first that the model error cannot be computed. Predictive error (not computable) requires

\[ p(\lambda) = Ae(\lambda) = Ax(\lambda) - A\hat{x} = AR(\lambda)b - \hat{b} \]

Define Influence Matrix or Resolution Matrix \( A(\lambda) = AR(\lambda) \).

\[ A(\lambda) = AA^\# = A(A^TA + \lambda^2I)^{-1}A^T \]

Regularized Residual or Predictive Risk is computable

\[ p(\lambda) = Ax(\lambda) - A\hat{x} \approx (Ax(\lambda) - b) = (A(\lambda) - I_m)b := r(\lambda) \]

It can be used to estimate the error.
How do we find $\lambda$? - L-curve

Plot regularization term against the fidelity term for $\lambda$

$$\log(\|x(\lambda)\|), \log(\|Ax(\lambda) - b\|)$$

**Figure:** On the left a corner and on the right no corner. The L-curve is expensive for general matrices $A$, but is very general and straightforward. It does not consider any information on the noise structure.
Consider the predictive error - cannot be calculated

\[ p(\lambda) = A R b - \hat{b} = A(\lambda)(\hat{b} + \eta) - \hat{b} \]
\[ = (A(\lambda) - I_m)\hat{b} + A(\lambda)\eta \]
\[ = \text{deterministic } + \text{ stochastic} \]

Consider the predictive risk - can be calculated

\[ r(\lambda) = (A(\lambda) - I_m)b \]
\[ = (A(\lambda) - I_m)\hat{b} + (A(\lambda) - I_m)\eta \]
\[ = \text{deterministic } + \text{ stochastic} \]

Both expressions use the noise \( \eta \).

We need to some statistical results.
Necessary Statistical Results

Mean-Variance  Suppose random vector $x$ has mean $x_0$, covariance-variance matrix $\Sigma$.
  - we say $x \sim (x_0, \Sigma)$
  - Then $b \sim (Ax_0, A\Sigma A^T)$

Trace Operator is linear. $\text{trace}(A + B) = \text{trace}(A) + \text{trace}(B)$ and $\text{trace}(A^T) = \text{trace}(A)$. We note also the cyclic property $\text{trace}(ABC) = \text{trace}(CAB)$, provided that dimensions are consistent.

Definition (Discrete White Noise Vector)
A random vector $\eta = (\eta_1, \eta_2, \ldots, \eta_n)$ is a discrete white noise vector provided that $E(\eta) = 0$ and $\text{cov}(\eta) = \sigma^2 I_n$. i.e.

$$E(\eta_i) = 0, \quad E(\eta_i \eta_j) = \sigma^2 \delta_{ij}$$

$\sigma^2$ is the variance of the white noise

Lemma (Trace Lemma)
Let $y$ be deterministic and $\eta$ a discrete white noise vector. Then

$$E(\|y + A\eta\|^2) = \|y\|^2 + \sigma^2 \text{trace}(A^T A)$$
Obtaining the Estimate

Use the Trace lemma and assume that the noise vector $\eta$ is a discrete white noise vector. Estimate mean predictive error from $E(\|p(\lambda)\|^2)$ and $E(\|r(\lambda)\|^2)$

$$E(\|r(\lambda)\|^2) = E(\|(A(\lambda) - I_m)\hat{b} + (A(\lambda) - I_m)\eta\|^2))$$
$$= \|(A(\lambda) - I_m)\hat{b}\|^2 + \sigma^2\text{trace}((A(\lambda) - I_m)^T(A(\lambda) - I_m)) \text{ and}$$
$$E(\|p(\lambda)\|^2) = E(\|(A(\lambda) - I_m)\hat{b} + A(\lambda)\eta\|^2))$$
$$= \|(A(\lambda) - I_m)\hat{b}\|^2 + \sigma^2\text{trace}(A(\lambda)^TA(\lambda))$$
$$= E(\|r(\lambda)\|^2) + \sigma^2\text{trace}(A(\lambda)^TA(\lambda)) - \sigma^2\text{trace}((A(\lambda) - I_m)^T(A(\lambda) - I_m))$$
$$= E(\|r(\lambda)\|^2) + \sigma^2(2\text{trace}(A(\lambda)) - m) \text{ by linearity of trace}$$
$$\approx \|r(\lambda)\|^2 + \sigma^2(2\text{trace}(A(\lambda)) - m) := U(\lambda)$$

Notice that $E(U(\lambda)) = E(\|p(\lambda)\|^2)$ so that $U$ is an unbiased estimator.

Thus seek $\lambda$ such that $U$ is minimum:

$$\lambda = \arg\min_{\lambda} U(\lambda).$$
An Example with UPRE

Figure: Notice the well-defined minimum. But requires the calculation of the trace
How do we use UPRE: First with the SVD rewrite $A(\lambda) - I_m$

\[ A(\lambda) - I_m = A(A^T A + \lambda^2 I_n)^{-1} A^T - I_m \]
\[ = U\Sigma V^T (V\Sigma^T U^T U\Sigma V^T + \lambda^2 VV^T)^{-1} V\Sigma^T U -UU^T \]
\[ = U(\Sigma(\Sigma^T \Sigma + \lambda^2 I_n)^{-1} \Sigma^T - I_m)U^T \quad \text{Hence with } \sigma_i = 0, \ i > n \]

\[ \text{trace}(A(\lambda)) = \sum_{i=1}^n \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2} \text{ and } (A(\lambda) - I_m)b = \sum_{i=1}^m (u_i^T b) \frac{-\lambda^2}{\sigma_i^2 + \lambda^2} u_i. \quad \text{Thus} \]

\[ \|r(\lambda)\|^2 = \sum_{i=1}^m |u_i^T b|^2 \left( \frac{\lambda^2}{\sigma_i^2 + \lambda^2} \right)^2 \quad \text{and} \]

\[ U(\lambda) = \sum_{i=1}^m |u_i^T b|^2 \left( \frac{\lambda^2}{\sigma_i^2 + \lambda^2} \right)^2 + \sigma^2(2\sum_{i=1}^n \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2} - m). \]

Entries $|u_i^T b|$ and $\sigma_i$ are calculated once. Hence given the SVD the cost is basically $O(n)$ for each $\lambda$. Basic idea is to bracket the minimum and then minimize within the bracket.

For the Fourier expansion a similar result can be obtained. (Suggest you try to derive this)

Otherwise the trace is estimated, say using randomization. (Reading material?)
Generalized Cross Validation: GCV

Based on omitting a data value and testing predictability of solution for this missing value: $\lambda$ is chosen to minimize

$$G(\lambda) = \frac{\|r(\lambda)\|_2^2}{(\text{trace}(I_m - A(\lambda)))^2}.$$ 

GCV provides another estimate of the predictive risk: (see e.g. Vogel)

There are the same difficulties of using the trace for large scale problems.

Often $G$ is relatively flat near the minimum, or has multiple minima and thus difficult to apply.