Math 4/896: Seminar in Mathematics
Topic: Inverse Theory

Instructor: Thomas Shores
Department of Mathematics

Lecture 18, March 21, 2006
AvH 10

Outline

SVD Implementation

To solve the Tikhonov regularized problem, first recall:

- $\nabla\left(\|G\mathbf{m}-\mathbf{d}\|_2^2+\alpha^2\|\mathbf{m}\|_2^2\right)=\left(G^TG\mathbf{m}-G^T\mathbf{d}\right)+\alpha^2\mathbf{m}$
- Equate to zero and these are the normal equations for the system $\begin{bmatrix} G \\ \alpha I \end{bmatrix}\mathbf{m}=\begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix}$, or $\left(G^TG+\alpha^2 I\right)\mathbf{m}=G^T\mathbf{d}$
- To solve, calculate $\left(G^TG+\alpha^2 I\right)^{-1}G^T=$

$$V\begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\alpha^2} & & & & \\ & \ddots & & & \\ & & \frac{\sigma_p}{\sigma_p^2+\alpha^2} & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix}U^T$$

SVD Implementation

To solve the Tikhonov regularized problem, first recall:

- $\nabla \left(\|G\mathbf{m} - \mathbf{d}\|_2^2 + \alpha^2 \|\mathbf{m}\|_2^2 \right) = \left(G^T G\mathbf{m} - G^T\mathbf{d} \right) + \alpha^2 \mathbf{m}$
- Equate to zero and these are the normal equations for the system $\begin{bmatrix} G \\ \alpha I \end{bmatrix} \mathbf{m} = \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix}$, or $\left(G^T G + \alpha^2 I \right) \mathbf{m} = G^T \mathbf{d}$
- To solve, calculate $\left(G^T G + \alpha^2 I \right)^{-1} G^T =$

$$V \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\alpha^2} & & & & \\ & \ddots & & & \\ & & \frac{\sigma_p}{\sigma_p^2+\alpha^2} & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix} U^T$$

SVD Implementation

To solve the Tikhonov regularized problem, first recall:

- $\nabla \left(\|G\mathbf{m} - \mathbf{d}\|_2^2 + \alpha^2 \|\mathbf{m}\|_2^2 \right) = \left(G^T G\mathbf{m} - G^T\mathbf{d}\right) + \alpha^2\mathbf{m}$
- Equate to zero and these are the normal equations for the system $\begin{bmatrix} G \\ \alpha I \end{bmatrix} \mathbf{m} = \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix}$, or $\left(G^T G + \alpha^2 I\right)\mathbf{m} = G^T\mathbf{d}$
- To solve, calculate $\left(G^T G + \alpha^2 I\right)^{-1} G^T =$

$$V \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\alpha^2} & & & & \\ & \ddots & & & \\ & & \frac{\sigma_p}{\sigma_p^2+\alpha^2} & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix} U^T$$

SVD Implementation

To solve the Tikhonov regularized problem, first recall:

- $\nabla\left(\|G\mathbf{m}-\mathbf{d}\|_2^2+\alpha^2\|\mathbf{m}\|_2^2\right)=\left(G^TG\mathbf{m}-G^T\mathbf{d}\right)+\alpha^2\mathbf{m}$
- Equate to zero and these are the normal equations for the system $\begin{bmatrix} G \\ \alpha I \end{bmatrix}\mathbf{m}=\begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix}$, or $\left(G^TG+\alpha^2I\right)\mathbf{m}=G^T\mathbf{d}$
- To solve, calculate $\left(G^TG+\alpha^2I\right)^{-1}G^T=$

$$V\begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\alpha^2} & & & & \\ & \ddots & & & \\ & & \frac{\sigma_p}{\sigma_p^2+\alpha^2} & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix}U^T$$

SVD Implementation

From the previous equation we obtain that the Moore-Penrose inverse and solution to the regularized problem are given by

$$G_\alpha^\dagger = \sum_{j=1}^{p} \frac{\sigma_j}{\sigma_j^2 + \alpha^2} \mathbf{V}_j \mathbf{U}_j^T$$

$$\mathbf{m}_\alpha = G^\dagger \mathbf{d} = \sum_{j=1}^{p} \frac{\sigma_j^2}{\sigma_j^2 + \alpha^2} \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{\sigma_j} \mathbf{V}_j$$

which specializes to the generalized inverse solution we have seen in the case that G is full column rank and $\alpha = 0$. (Remember $\mathbf{d} = U\mathbf{h}$ so that $\mathbf{h} = U^T \mathbf{d}$.)

The Filter Idea

About Filtering:

The idea is simply to “filter” the singular values of our problem so that (hopefully) only “good” ones are used.

- We replace the σ_i by $f(\sigma_i)$. The function f is called a **filter**.
- $f(\sigma) = 1$ simply uses the original singular values.
- $f(\sigma) = \dfrac{\sigma^2}{\sigma^2 + \alpha^2}$ is the Tikhonov filter we have just developed.
- $f(\sigma) = \max\{\operatorname{sgn}(\sigma - \epsilon), 0\}$ is the TSVD filter with singular values smaller than ϵ truncated to zero.

The Filter Idea

About Filtering:

The idea is simply to "filter" the singular values of our problem so that (hopefully) only "good" ones are used.

- We replace the σ_i by $f(\sigma_i)$. The function f is called a **filter**.
- $f(\sigma) = 1$ simply uses the original singular values.
- $f(\sigma) = \dfrac{\sigma^2}{\sigma^2 + \alpha^2}$ is the Tikhonov filter we have just developed.
- $f(\sigma) = \max\{\text{sgn}(\sigma - \epsilon), 0\}$ is the TSVD filter with singular values smaller than ϵ truncated to zero.

The Filter Idea

About Filtering:

The idea is simply to "filter" the singular values of our problem so that (hopefully) only "good" ones are used.

- We replace the σ_i by $f(\sigma_i)$. The function f is called a **filter**.
- $f(\sigma) = 1$ simply uses the original singular values.
- $f(\sigma) = \dfrac{\sigma^2}{\sigma^2 + \alpha^2}$ is the Tikhonov filter we have just developed.
- $f(\sigma) = \max\{\operatorname{sgn}(\sigma - \epsilon), 0\}$ is the TSVD filter with singular values smaller than ϵ truncated to zero.

The Filter Idea

About Filtering:

The idea is simply to "filter" the singular values of our problem so that (hopefully) only "good" ones are used.

- We replace the σ_i by $f(\sigma_i)$. The function f is called a **filter**.
- $f(\sigma) = 1$ simply uses the original singular values.
- $f(\sigma) = \dfrac{\sigma^2}{\sigma^2 + \alpha^2}$ is the Tikhonov filter we have just developed.
- $f(\sigma) = \max\{\operatorname{sgn}(\sigma - \epsilon), 0\}$ is the TSVD filter with singular values smaller than ϵ truncated to zero.

The Filter Idea

About Filtering:

The idea is simply to "filter" the singular values of our problem so that (hopefully) only "good" ones are used.

- We replace the σ_i by $f(\sigma_i)$. The function f is called a **filter**.
- $f(\sigma) = 1$ simply uses the original singular values.
- $f(\sigma) = \dfrac{\sigma^2}{\sigma^2 + \alpha^2}$ is the Tikhonov filter we have just developed.
- $f(\sigma) = \max\{\operatorname{sgn}(\sigma - \epsilon), 0\}$ is the TSVD filter with singular values smaller than ϵ truncated to zero.

The L-curve

L-curves are one tool for choosing the regularization paramter α:

- Make a plot of the curve $\left(\|\mathbf{m}_\alpha\|_2, \|G\mathbf{m}_\alpha - \mathbf{d}\|_2\right)$
- Typically, this curve looks to be asymptotic to the axes.
- Choose the value of α closest to the corner.
- Caution: L-curves are NOT guaranteed to work as a regularization strategy.
- An alternative: (Morozov's discrepancy principle) Choose α so that the misfit $\|G\mathbf{m}_\alpha - \mathbf{d}\|_2$ is the same size as the data noise $\|\delta\mathbf{d}\|_2$.

The L-curve

L-curves are one tool for choosing the regularization paramter α:

- Make a plot of the curve $(\|\mathbf{m}_\alpha\|_2, \|G\mathbf{m}_\alpha - \mathbf{d}\|_2)$
- Typically, this curve looks to be asymptotic to the axes.
- Choose the value of α closest to the corner.
- Caution: L-curves are NOT guaranteed to work as a regularization strategy.
- An alternative: (Morozov's discrepancy principle) Choose α so that the misfit $\|G\mathbf{m}_\alpha - \mathbf{d}\|_2$ is the same size as the data noise $\|\delta\mathbf{d}\|_2$.

The L-curve

L-curves are one tool for choosing the regularization paramter α:

- Make a plot of the curve $(\|\mathbf{m}_\alpha\|_2, \|G\mathbf{m}_\alpha - \mathbf{d}\|_2)$
- Typically, this curve looks to be asymptotic to the axes.
- Choose the value of α closest to the corner.
- Caution: L-curves are NOT guaranteed to work as a regularization strategy.
- An alternative: (Morozov's discrepancy principle) Choose α so that the misfit $\|G\mathbf{m}_\alpha - \mathbf{d}\|_2$ is the same size as the data noise $\|\delta\mathbf{d}\|_2$.

The L-curve

L-curves are one tool for choosing the regularization paramter α:

- Make a plot of the curve $\left(\|\mathbf{m}_\alpha\|_2, \|G\mathbf{m}_\alpha - \mathbf{d}\|_2\right)$
- Typically, this curve looks to be asymptotic to the axes.
- Choose the value of α closest to the corner.
- Caution: L-curves are NOT guaranteed to work as a regularization strategy.
- An alternative: (Morozov's discrepancy principle) Choose α so that the misfit $\|G\mathbf{m}_\alpha - \mathbf{d}\|_2$ is the same size as the data noise $\|\delta\mathbf{d}\|_2$.

The L-curve

L-curves are one tool for choosing the regularization paramter α:

- Make a plot of the curve $\left(\|\mathbf{m}_\alpha\|_2, \|G\mathbf{m}_\alpha - \mathbf{d}\|_2\right)$
- Typically, this curve looks to be asymptotic to the axes.
- Choose the value of α closest to the corner.
- Caution: L-curves are NOT guaranteed to work as a regularization strategy.
- An alternative: (Morozov's discrepancy principle) Choose α so that the misfit $\|G\mathbf{m}_\alpha - \mathbf{d}\|_2$ is the same size as the data noise $\|\delta\mathbf{d}\|_2$.

The L-curve

L-curves are one tool for choosing the regularization paramter α:

- Make a plot of the curve $(\|\mathbf{m}_\alpha\|_2, \|G\mathbf{m}_\alpha - \mathbf{d}\|_2)$
- Typically, this curve looks to be asymptotic to the axes.
- Choose the value of α closest to the corner.
- Caution: L-curves are NOT guaranteed to work as a regularization strategy.
- An alternative: (Morozov's discrepancy principle) Choose α so that the misfit $\|G\mathbf{m}_\alpha - \mathbf{d}\|_2$ is the same size as the data noise $\|\delta\mathbf{d}\|_2$.

Historical Notes

Tikhonov's original interest was in operator equations

$$d(s) = \int_a^b k(s,t)\, m(t)\, dt$$

or $d = Km$ where K is a compact (**bounded = continuous**) linear operator from one Hilbert space H_1 into another H_2. In this situation:

- Such an operator $K : H_1 \to H_2$ has an **adjoint operator** $K^* : H_2 \to H_1$ (analogous to transpose of matrix operator.)
- Least squares solutions to $\min \|Km - d\|$ are just solutions to the **normal** equation $K^*Km = K^*d$ (and exist.)
- There is a Moore-Penrose inverse operator $K^\dagger$ such that $m = K^\dagger d$ is the least squares solution of least 2-norm. But this operator is generally **unbounded** (not continuous.)

Historical Notes

Tikhonov's original interest was in operator equations

$$d(s) = \int_a^b k(s,t)\, m(t)\, dt$$

or $d = Km$ where K is a compact (**bounded = continuous**) linear operator from one Hilbert space H_1 into another H_2. In this situation:

- Such an operator $K : H_1 \rightarrow H_2$ has an **adjoint operator** $K^* : H_2 \rightarrow H_1$ (analogous to transpose of matrix operator.)
- Least squares solutions to min $\|Km - d\|$ are just solutions to the **normal** equation $K^*Km = K^*d$ (and exist.)
- There is a Moore-Penrose inverse operator $K^\dagger$ such that $m = K^\dagger d$ is the least squares solution of least 2-norm. But this operator is generally **unbounded** (not continuous.)

Historical Notes

Tikhonov's original interest was in operator equations

$$d(s) = \int_a^b k(s,t)\, m(t)\, dt$$

or $d = Km$ where K is a compact (**bounded = continuous**) linear operator from one Hilbert space H_1 into another H_2. In this situation:

- Such an operator $K : H_1 \to H_2$ has an **adjoint operator** $K^* : H_2 \to H_1$ (analogous to transpose of matrix operator.)
- Least squares solutions to min $\|Km - d\|$ are just solutions to the **normal** equation $K^*Km = K^*d$ (and exist.)
- There is a Moore-Penrose inverse operator $K^\dagger$ such that $m = K^\dagger d$ is the least squares solution of least 2-norm. But this operator is generally **unbounded** (not continuous.)

Tikhonov's original interest was in operator equations

$$d(s) = \int_a^b k(s,t)\, m(t)\, dt$$

or $d = Km$ where K is a compact (**bounded = continuous**) linear operator from one Hilbert space H_1 into another H_2. In this situation:

- Such an operator $K : H_1 \to H_2$ has an **adjoint operator** $K^* : H_2 \to H_1$ (analogous to transpose of matrix operator.)
- Least squares solutions to min $\|Km - d\|$ are just solutions to the **normal** equation $K^*Km = K^*d$ (and exist.)
- There is a Moore-Penrose inverse operator $K^\dagger$ such that $m = K^\dagger d$ is the least squares solution of least 2-norm. But this operator is generally **unbounded** (not continuous.)

More on Tikhonov's operator equation:

- The operator $(K^*K+\alpha I)$ is bounded with bounded inverse and the **regularized problem** $(K^*K+\alpha I)\,m=K^*d$ has a unique solution m_α.
- Given that $\delta=\|\delta d\|$ is the noise level and that the problem actually solved is $(K^*K+\alpha I)\,m=K^*d^\delta$ with $d^\delta=d+\delta d$ yielding m_α^δ Tikhonov defines a **regular algorithm** to be a choice $\alpha=\alpha(\delta)$ such that

$$\alpha(\delta)\to 0 \text{ and } m^\delta_{\alpha(\delta)}\to K^\dagger d \text{ as } \delta\to 0.$$

- Morozov's discrepancy principle is a regular algorithm.

Finish Section 5.2 by exploring the Example 5.1 file, which constructs the L-curve of the Shaw problem using tools from the Regularization Toolbox.

More on Tikhonov's operator equation:

- The operator $(K^*K + \alpha I)$ is bounded with bounded inverse and the **regularized problem** $(K^*K + \alpha I)\,m = K^*d$ has a unique solution m_α.
- Given that $\delta = \|\delta d\|$ is the noise level and that the problem actually solved is $(K^*K + \alpha I)\,m = K^*d^\delta$ with $d^\delta = d + \delta d$ yielding m_α^δTikhonov defines a **regular algorithm** to be a choice $\alpha = \alpha(\delta)$ such that

$$\alpha(\delta) \to 0 \text{ and } m^\delta_{\alpha(\delta)} \to K^\dagger d \text{ as } \delta \to 0.$$

- Morozov's discrepancy principle is a regular algorithm.

Finish Section 5.2 by exploring the Example 5.1 file, which constructs the L-curve of the Shaw problem using tools from the Regularization Toolbox.

More on Tikhonov's operator equation:

- The operator $(K^*K + \alpha I)$ is bounded with bounded inverse and the **regularized problem** $(K^*K + \alpha I)\, m = K^*d$ has a unique solution m_α.
- Given that $\delta = \|\delta d\|$ is the noise level and that the problem actually solved is $(K^*K + \alpha I)\, m = K^*d^\delta$ with $d^\delta = d + \delta d$ yielding m_α^δ Tikhonov defines a **regular algorithm** to be a choice $\alpha = \alpha(\delta)$ such that

$$\alpha(\delta) \to 0 \text{ and } m_{\alpha(\delta)}^\delta \to K^\dagger d \text{ as } \delta \to 0.$$

- Morozov's discrepancy principle is a regular algorithm.

Finish Section 5.2 by exploring the Example 5.1 file, which constructs the L-curve of the Shaw problem using tools from the Regularization Toolbox.

More on Tikhonov's operator equation:

- The operator $(K^*K + \alpha I)$ is bounded with bounded inverse and the **regularized problem** $(K^*K + \alpha I)\, m = K^*d$ has a unique solution m_α.
- Given that $\delta = \|\delta d\|$ is the noise level and that the problem actually solved is $(K^*K + \alpha I)\, m = K^*d^\delta$ with $d^\delta = d + \delta d$ yielding m_α^δTikhonov defines a **regular algorithm** to be a choice $\alpha = \alpha(\delta)$ such that

$$\alpha(\delta) \to 0 \text{ and } m_{\alpha(\delta)}^\delta \to K^\dagger d \text{ as } \delta \to 0.$$

- Morozov's discrepancy principle is a regular algorithm.

Finish Section 5.2 by exploring the Example 5.1 file, which constructs the L-curve of the Shaw problem using tools from the Regularization Toolbox.

Outline

Resolution Matrix

Definition:

Resolution matrix for a regularized problem starts with this observation:

- Let $G^\natural \equiv \left(G^T G + \alpha^2 I\right)^{-1} G^T$.
- Then $\mathbf{m}_\alpha = G^\natural \mathbf{d} = \sum_{j=1}^{p} f_j \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{\sigma_j} \mathbf{V}_j = VFS^\dagger U^T \mathbf{d}$.
- Model resolution matrix: $R_{\mathbf{m},\alpha} = G^\natural G = VFV^T$
- Data resolution matrix: $R_{\mathbf{d},\alpha} = GG^\natural = UFU^T$

The Example 5.1 file constructs the model resolution matrix of the Shaw problem and shows poor resolution in this case.

Resolution Matrix

Definition:

Resolution matrix for a regularized problem starts with this observation:

- Let $G^\natural \equiv \left(G^T G + \alpha^2 I\right)^{-1} G^T$.
- Then $\mathbf{m}_\alpha = G^\natural \mathbf{d} = \sum_{j=1}^{p} f_j \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{\sigma_j} \mathbf{V}_j = V F S^\dagger U^T \mathbf{d}$.
- Model resolution matrix: $R_{\mathbf{m},\alpha} = G^\natural G = V F V^T$
- Data resolution matrix: $R_{\mathbf{d},\alpha} = G G^\natural = U F U^T$

The Example 5.1 file constructs the model resolution matrix of the Shaw problem and shows poor resolution in this case.

Resolution Matrix

Definition:

Resolution matrix for a regularized problem starts with this observation:

- Let $G^\natural \equiv \left(G^T G + \alpha^2 I\right)^{-1} G^T$.
- Then $\mathbf{m}_\alpha = G^\natural \mathbf{d} = \sum_{j=1}^{p} f_j \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{\sigma_j} \mathbf{V}_j = VFS^\dagger U^T \mathbf{d}$.
- Model resolution matrix: $R_{\mathbf{m},\alpha} = G^\natural G = VFV^T$
- Data resolution matrix: $R_{\mathbf{d},\alpha} = GG^\natural = UFU^T$

The Example 5.1 file constructs the model resolution matrix of the Shaw problem and shows poor resolution in this case.

Resolution Matrix

Definition:

Resolution matrix for a regularized problem starts with this observation:

- Let $G^\natural \equiv \left(G^T G + \alpha^2 I\right)^{-1} G^T$.
- Then $\mathbf{m}_\alpha = G^\natural \mathbf{d} = \sum_{j=1}^{p} f_j \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{\sigma_j} \mathbf{V}_j = VFS^\dagger U^T \mathbf{d}$.
- Model resolution matrix: $R_{\mathbf{m},\alpha} = G^\natural G = VFV^T$
- Data resolution matrix: $R_{\mathbf{d},\alpha} = GG^\natural = UFU^T$

The Example 5.1 file constructs the model resolution matrix of the Shaw problem and shows poor resolution in this case.

Resolution Matrix

Definition:

Resolution matrix for a regularized problem starts with this observation:

- Let $G^{\natural} \equiv \left(G^T G + \alpha^2 I\right)^{-1} G^T$.
- Then $\mathbf{m}_\alpha = G^{\natural}\mathbf{d} = \sum_{j=1}^{p} f_j \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{\sigma_j}\mathbf{V}_j = VFS^{\dagger}U^T\mathbf{d}$.
- Model resolution matrix: $R_{\mathbf{m},\alpha} = G^{\natural}G = VFV^T$
- Data resolution matrix: $R_{\mathbf{d},\alpha} = GG^{\natural} = UFU^T$

The Example 5.1 file constructs the model resolution matrix of the Shaw problem and shows poor resolution in this case.

Outline

Higher Order Regularization

Basic Idea

We can think of the regularization term $\alpha^2 \|\mathbf{m}\|_2^2$ as favoring minimizing the 0-th order derivative of a function $m(x)$ under the hood. Alternatives:

- Minimize a matrix approximation to $m'(x)$. This is a first order method.
- Minimize a matrix approximation to $m''(x)$. This is a second order method.
- These lead to new minimization problems: to minimize

$$\|G\mathbf{m} - \mathbf{d}\|_2^2 + \alpha^2 \|L\mathbf{m}\|_2^2 .$$

- How do we resolve this problem as we did with $L = I$?

Higher Order Regularization

Basic Idea

We can think of the regularization term $\alpha^2 \|\mathbf{m}\|_2^2$ as favoring minimizing the 0-th order derivative of a function $m(x)$ under the hood. Alternatives:

- Minimize a matrix approximation to $m'(x)$. This is a first order method.
- Minimize a matrix approximation to $m''(x)$. This is a second order method.
- These lead to new minimization problems: to minimize

$$\|G\mathbf{m} - \mathbf{d}\|_2^2 + \alpha^2 \|L\mathbf{m}\|_2^2.$$

- How do we resolve this problem as we did with $L = I$?

Higher Order Regularization

Basic Idea

We can think of the regularization term $\alpha^2 \|\mathbf{m}\|_2^2$ as favoring minimizing the 0-th order derivative of a function $m(x)$ under the hood. Alternatives:

- Minimize a matrix approximation to $m'(x)$. This is a first order method.
- Minimize a matrix approximation to $m''(x)$. This is a second order method.
- These lead to new minimization problems: to minimize

$$\|G\mathbf{m} - \mathbf{d}\|_2^2 + \alpha^2 \|L\mathbf{m}\|_2^2 .$$

- How do we resolve this problem as we did with $L = I$?

Higher Order Regularization

Basic Idea

We can think of the regularization term $\alpha^2 \|\mathbf{m}\|_2^2$ as favoring minimizing the 0-th order derivative of a function $m(x)$ under the hood. Alternatives:

- Minimize a matrix approximation to $m'(x)$. This is a first order method.
- Minimize a matrix approximation to $m''(x)$. This is a second order method.
- These lead to new minimization problems: to minimize

$$\|G\mathbf{m} - \mathbf{d}\|_2^2 + \alpha^2 \|L\mathbf{m}\|_2^2 .$$

- How do we resolve this problem as we did with $L = I$?

Higher Order Regularization

Basic Idea

We can think of the regularization term $\alpha^2 \|\mathbf{m}\|_2^2$ as favoring minimizing the 0-th order derivative of a function $m(x)$ under the hood. Alternatives:

- Minimize a matrix approximation to $m'(x)$. This is a first order method.
- Minimize a matrix approximation to $m''(x)$. This is a second order method.
- These lead to new minimization problems: to minimize

$$\|G\mathbf{m} - \mathbf{d}\|_2^2 + \alpha^2 \|L\mathbf{m}\|_2^2 .$$

- How do we resolve this problem as we did with $L = I$?

Example Matrices

We will explore approximations to first and second derivatives at the board.

Key Idea: Generalized SVD (GSVD)

Theorem

Let G be an $m \times n$ matrix and L a $p \times n$ matrix. Then there exist $m \times m$ orthogonal U, $p \times p$ orthogonal V and $n \times n$ nonsingular matrix X with $m \geq n \geq \min\{p, n\} = q$ such that

$$
\begin{aligned}
U^T G X &= \operatorname{diag}\{c_1, c_2, \ldots, c_n\} \\
V^T L X &= \operatorname{diag}\{s_1, s_2, \ldots, s_q\} \\
C^T C + S^T S &= 1 \\
0 \leq c_1 \leq c_2 \cdots \leq c_n \leq 1 & \\
1 \geq s_1 \geq s_2 \cdots \geq s_n \geq 0 &
\end{aligned}
$$

The numbers $\gamma_i = c_i / s_i$, $i = 1, \ldots, q$ are called the ***generalized singular values*** *of G and L and $0 \leq \gamma_1 \leq \gamma_2 \cdots \leq \gamma_q$.*

Notes: If $\operatorname{rank}(L) = q$, then the singular values are finite.

Application to Higher Order Regularization

The minimization problem is shown, just as we did earlier, to be equivalent to the problem

$$\left(G^TG + \alpha^2 L^T L\right) \mathbf{m} = G^T \mathbf{d}$$

which has solution

$$\mathbf{m}_{\alpha,L} = \left(G^TG + \alpha^2 L^T L\right) G^T \mathbf{d} \equiv G^\natural \mathbf{d}.$$

With some work:

$$\mathbf{m}_{\alpha,L} = \sum_{j=1}^{p} \frac{\gamma_j^2}{\gamma_j^2 + \alpha^2} \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{c_j} \mathbf{X}_j + \sum_{j=p+1}^{n} \left(\mathbf{U}_j^T \mathbf{d}\right) \mathbf{X}_j$$

Outline

TGSVD:

We have seen this idea before. Simply apply it to formula above, remembering that the generalized singular values are reverse ordered.

- Formula becomes

$$\mathbf{m}_{\alpha,L} = \sum_{j=k}^{p} \frac{\gamma_j^2}{\gamma_j^2 + \alpha^2} \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{c_j} \mathbf{X}_j + \sum_{j=p+1}^{n} \left(\mathbf{U}_j^T \mathbf{d}\right) \mathbf{X}_j$$

- Key question: where to start k.

Example 5.6 gives a nice illustration of the ideas. We'll use the CD script to explore it.

GCV

Basic Idea:

Comes from statistical “leave-one-out” cross validation.

- Leave out one data point and use model to predict it.
- Sum these up and choose regularization parameter α that minimizes the sum of the squares of the predictive errors

$$V_0(\alpha) = \frac{1}{m}\sum_{k=1}^{m}\left(\left(Gm_{\alpha,L}^{[k]}\right)_k - d_k\right)^2.$$

- One can show a good approximation is

$$V_0(\alpha) = \frac{m\left\|G\mathbf{m}_\alpha - \mathbf{d}\right\|_2}{\operatorname{Tr}\left(I - GG^\natural\right)^2}$$

GCV

Basic Idea:

Comes from statistical "leave-one-out" cross validation.

- Leave out one data point and use model to predict it.
- Sum these up and choose regularization parameter α that minimizes the sum of the squares of the predictive errors

$$V_0(\alpha) = \frac{1}{m}\sum_{k=1}^{m}\left(\left(Gm_{\alpha,L}^{[k]}\right)_k - d_k\right)^2.$$

- One can show a good approximation is

$$V_0(\alpha) = \frac{m\left\|G\mathbf{m}_\alpha - \mathbf{d}\right\|_2}{\mathrm{Tr}\left(I - GG^\natural\right)^2}$$

GCV

Basic Idea:

Comes from statistical "leave-one-out" cross validation.

- Leave out one data point and use model to predict it.
- Sum these up and choose regularization parameter α that minimizes the sum of the squares of the predictive errors

$$V_0(\alpha) = \frac{1}{m}\sum_{k=1}^{m}\left(\left(Gm_{\alpha,L}^{[k]}\right)_k - d_k\right)^2.$$

- One can show a good approximation is

$$V_0(\alpha) = \frac{m\left\|G\mathbf{m}_\alpha - \mathbf{d}\right\|_2}{\operatorname{Tr}\left(I - GG^\natural\right)^2}$$

GCV

Basic Idea:

Comes from statistical "leave-one-out" cross validation.

- Leave out one data point and use model to predict it.
- Sum these up and choose regularization parameter α that minimizes the sum of the squares of the predictive errors

$$V_0(\alpha) = \frac{1}{m}\sum_{k=1}^{m}\left(\left(Gm_{\alpha,L}^{[k]}\right)_k - d_k\right)^2.$$

- One can show a good approximation is

$$V_0(\alpha) = \frac{m\left\|G\mathbf{m}_\alpha - \mathbf{d}\right\|_2}{\mathrm{Tr}\left(I - GG^\natural\right)^2}$$

Outline

Error Bounds

Error Estimates:

They exist, even in the hard cases where there is error in both G and d.

- In the simpler case, G known exactly, they take the form

$$\frac{\|\mathbf{m}_\alpha - \widetilde{\mathbf{m}}_\alpha\|_2}{\|\mathbf{m}_\alpha\|_2} \leq \kappa_\alpha \frac{\left\|\mathbf{d} - \widetilde{\mathbf{d}}\right\|_2}{\|G\mathbf{m}_\alpha\|_2}$$

  where κ_α is inversely proportional to α.

-

Error Bounds

Error Estimates:

They exist, even in the hard cases where there is error in both G and d.

- In the simpler case, G known exactly, they take the form

$$\frac{\|\mathbf{m}_\alpha - \widetilde{\mathbf{m}}_\alpha\|_2}{\|\mathbf{m}_\alpha\|_2} \leq \kappa_\alpha \frac{\left\|\mathbf{d} - \widetilde{\mathbf{d}}\right\|_2}{\|G\mathbf{m}_\alpha\|_2}$$

where κ_α is inversely proportional to α.

Error Bounds

Error Estimates:

They exist, even in the hard cases where there is error in both G and d.

- In the simpler case, G known exactly, they take the form

$$\frac{\left\|\mathbf{m}_\alpha - \widetilde{\mathbf{m}}_\alpha\right\|_2}{\left\|\mathbf{m}_\alpha\right\|_2} \leq \kappa_\alpha \frac{\left\|\mathbf{d} - \widetilde{\mathbf{d}}\right\|_2}{\left\|G\mathbf{m}_\alpha\right\|_2}$$

where κ_α is inversely proportional to α.