Statistical Aspects of Inverse Problems

2019 RMMC Summer School
Inverse Problems in Imaging

Jodi Mead
Department of Mathematics

BOISE STATE UNIVERSITY

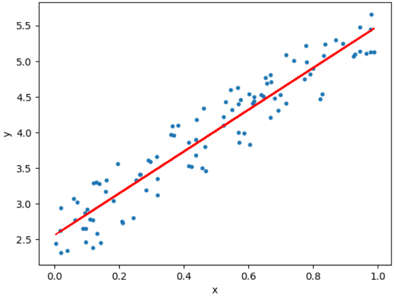
Inverse Problems

$$\mathbf{A}\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$$

- $\mathbf{A} \in R^{m \times n}$ - mathematical model
- $\mathbf{b} \in R^{m}$ - observed data
- $\mathbf{x} \in R^{n}$- unknown model parameters
- $\boldsymbol{\epsilon} \in R^{m}$ - additive noise or random error

Linear Regression

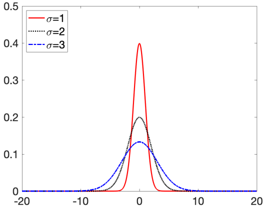
$$y = b + mx + \epsilon$$



$$\begin{array}{ccccccc} \mathbf{A} & \mathbf{x} & = & \mathbf{b} & + & \boldsymbol{\epsilon} \\ \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_m \end{bmatrix} & \begin{bmatrix} b \\ m \end{bmatrix} & = & \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{bmatrix} & + & \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_m \end{bmatrix} \end{array}$$

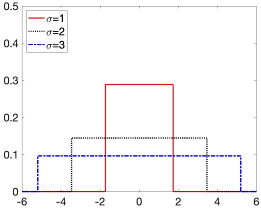
Noise Models

Normal: $f(\epsilon_i) = \frac{1}{\sqrt{\pi}\sigma} e^{-\frac{1}{2}\epsilon_i^2/\sigma^2}$

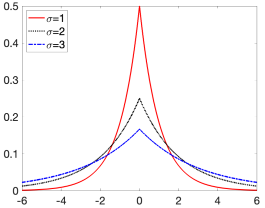


Uniform:

$$f(\epsilon_i) = \begin{cases} \frac{1}{\sigma 2\sqrt{3}} & |\epsilon_i| < \sigma\sqrt{3} \\ 0 & \text{otherwise} \end{cases}$$



Laplace: $f(\epsilon_i) = \frac{1}{\sqrt{2}\sigma} e^{-\sqrt{2}|\epsilon_i|/\sigma}$



PDF and Likelihood

Probability Density Function (PDF): $f(\boldsymbol{\epsilon}) = f(\mathbf{b}|\mathbf{x})$, $\mathbf{x}$ are fixed and $\mathbf{b}$ vary.

Likelihood function $L(\mathbf{x}) \equiv L(\mathbf{x}|\mathbf{b}) = f(\mathbf{b}|\mathbf{x})$, $\mathbf{b}$ are fixed $\mathbf{x}$ vary.

Common Noise Assumptions

Independent and identically distributed (i.i.d)

$$f(xs|\mathbf{x}) = f_1(\mathbf{b}_1|\mathbf{x}) f_2(\mathbf{b}_2|\mathbf{x}) \cdots f_m(\mathbf{b}_m|\mathbf{x})$$

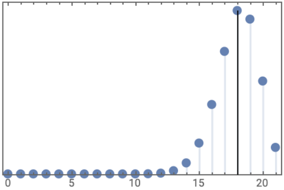
White noise

uncorrelated: $\rho(\epsilon_i, \epsilon_j) = \frac{\text{cov}(\epsilon_i, \epsilon_j)}{var(\epsilon_i) var(\epsilon_j)} = 0$

covariance: $\mathbf{C} = diag(\sigma_1^2, \ldots, \sigma_m^2)$

Frequentist

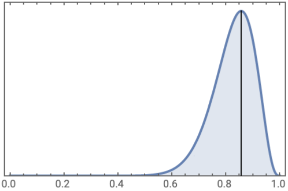
Likelihood function is a statistic that summarizes a single sample of data from a population. There is a “true" value of $\mathbf{x}$.



Number of successes

Bayesian

Likelihood function is information about the parameters provided by the data. The parameters $\mathbf{x}$ are random.



Probability of success

Maximum Likelihood Estimation (MLE)

i.i.d. $\rightarrow f(\boldsymbol{\epsilon}) = \prod_{i=1}^{m} f(\epsilon_i)$

Normal: $\epsilon_i \sim N(0, \sigma_i)$

$$f(\boldsymbol{\epsilon}) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}(\mathbf{Ax}-\mathbf{b})_i^2/\sigma_i^2} = \frac{1}{(\sqrt{2\pi})^m \prod_{i=1}^{m} \sigma_i} \sum_{i=1}^{m} e^{-\frac{1}{2}(\mathbf{Ax}-\mathbf{b})_i^2/\sigma_i^2}$$

$$\max_{\mathbf{x}} \; f(\boldsymbol{\epsilon}) = \min_{\mathbf{x}} \; \sum_{i=1}^{m} (\mathbf{Ax} - \mathbf{b})_i^2/\sigma_i^2$$

Laplace: $\epsilon_i \sim \mathcal{L}(0, \sigma_i)$

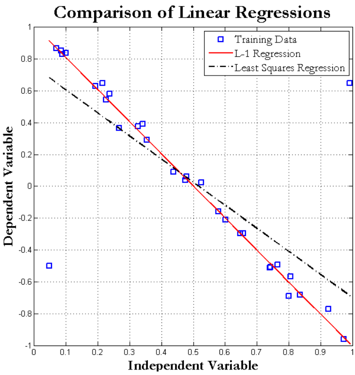
$$f(\boldsymbol{\epsilon}) = \prod_{i=1}^{m} \frac{1}{\sqrt{2}\sigma_i} e^{-\sqrt{2}|(\mathbf{Ax}-\mathbf{b})_i|/\sigma} = \frac{1}{(\sqrt{2})^m \prod_{i=1}^{m} \sigma_i} \sum_{i=1}^{m} e^{-\sqrt{2}|(\mathbf{Ax}-\mathbf{b})_i|/\sigma}$$

$$\max_{\mathbf{x}} \; f(\boldsymbol{\epsilon}) = \min_{\mathbf{x}} \; \sum_{i=1}^{m} |(\mathbf{Ax} - \mathbf{b})_i|/\sigma_i$$

MLE Estimates

Least squares:

$$\mathbf{x}_{\mathrm{L2}} = \arg\min_{\mathbf{x}} \ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$



L1:

$$\mathbf{x}_{\mathrm{L1}} = \arg\min_{\mathbf{x}} \ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1$$

χ^2 Goodness of Fit

How "close" are data $\mathbf{b}$ to those which would be expected under the fitted model $\mathbf{A}\mathbf{x}_{\mathsf{L2}}$ or $\mathbf{A}\mathbf{x}_{\mathsf{L1}}$?

$$\begin{aligned}
\chi^2 &= \quad (\mathbf{A}\mathbf{x} - \mathbf{b})^T \mathbf{C}^{-1} (\mathbf{A}\mathbf{x} - \mathbf{b}) && \sim \chi^2(m) \\
\chi^2_{\mathsf{obs}} &= (\mathbf{A}\mathbf{x}_{\mathsf{L2}} - \mathbf{b})^T \mathbf{C}^{-1} (\mathbf{A}\mathbf{x}_{\mathsf{L2}} - \mathbf{b}) && \sim \chi^2(m-n) \\
\chi^2_{\mathsf{obs}} &= (\mathbf{A}\mathbf{x}_{\mathsf{L1}} - \mathbf{b})^T \mathbf{C}^{-1} (\mathbf{A}\mathbf{x}_{\mathsf{L1}} - \mathbf{b}) && \sim \chi^2(m-n)
\end{aligned}$$

Null Hypothesis: $\boldsymbol{\epsilon} = \mathbf{A}\mathbf{x} - \mathbf{b} \sim N(0, \mathbf{C})$ (or $\sim \mathcal{L}(0, \mathbf{C})$)

Fail to reject if χ^2_{obs} exceeds desired level of significance, e.g.

$$(\mathbf{A}\mathbf{x}_{\mathsf{L2}} - \mathbf{b}_{\mathsf{obs}})^T \mathbf{C}^{-1} (\mathbf{A}\mathbf{x}_{\mathsf{L2}} - \mathbf{b}_{\mathsf{obs}}) \approx m - n$$

Posterior uncertainty estimates

If $\chi^2_{\text{obs}} \approx m - n$ then $\mathbf{x}_{\text{L2}}$ or $\mathbf{x}_{\text{L1}}$ is the MLE and

$$\text{cov}(\mathbf{x}_{\text{L2}}) = (\mathbf{A}_w^T \mathbf{A}_w)^{-1} \mathbf{A}_w^T \text{cov}(\mathbf{b}_w) \mathbf{A}_w (\mathbf{A}_w^T \mathbf{A}_w)^{-1}$$

with $\mathbf{A}_w = \mathbf{C}^{-1/2}\mathbf{A}$, $\mathbf{b}_w = \mathbf{C}^{-1/2}\mathbf{b}$.

Note: $\mathbf{x}_{\text{L2}} = (\mathbf{A}_w^T \mathbf{A}_w)^{-1} \mathbf{A}_w^T \mathbf{b}_w$

while there's not closed form expression for $\text{cov}(\mathbf{x}_{\text{L1}})$ so ...

$$\text{cov}(\mathbf{x}_{\text{L1}}) \approx \ldots$$

via Monte Carlo error propatation[1]

[1] Parameter Estimation and Inverse Problems, Aster et all, 2018

Confidence vs Credible Intervals

Frequentist

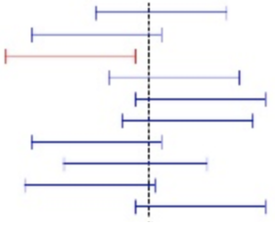
95% confidence interval (and p-values):

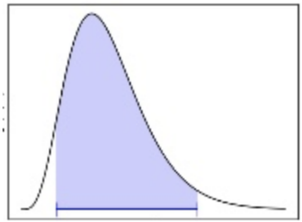
Collect data 100 times and 95 of the confidence intervals would contain the true parameters.

Bayesian

95% credible interval:

95% chance true parameters are in the interval.





Confidence Intervals and Regions

$$\mathrm{cov}(\mathbf{x}_{\mathrm{L2}}) \ (\text{or } \mathrm{cov}(\mathbf{x}_{\mathrm{L1}})) \ = \hat{\mathbf{C}} \ = \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_1\hat{\sigma}_2 & \dots & \\ \hat{\sigma}_1\hat{\sigma}_2 & \hat{\sigma}_2^2 & \hat{\sigma}_2\hat{\sigma}_3 & \dots \\ \vdots & \vdots & \ddots & \vdots \\ & \dots & \hat{\sigma}_{n-1}\hat{\sigma}_n & \hat{\sigma}_n^2 \end{bmatrix}$$

Confidence interval: $(\mathbf{x}_{\mathrm{L2}})_i \pm \hat{\sigma}_i z_{\alpha/2}$, i.e.

$$(\mathbf{x}_{\mathrm{true}})_i \in [(\mathbf{x}_{\mathrm{L2}})_i - \hat{\sigma}_i z_{\alpha/2}, (\mathbf{x}_{\mathrm{L2}})_i + \hat{\sigma}_i z_{\alpha/2}]$$

Confidence ellipsoid:

No correlation: $\frac{((\mathbf{x}_{\mathrm{true}})_i - (\mathbf{x}_{\mathrm{L}})_i)^2}{\hat{\sigma}_i^2} + \frac{((\mathbf{x}_{\mathrm{true}})_j - (\mathbf{x}_{\mathrm{L}})_j)^2}{\hat{\sigma}_j^2} \leq \Delta^2$

where Δ represents $1-\alpha$ confidence region for χ_1^2.

Correlation: $(\mathbf{x}_{\mathrm{true}} - \mathbf{x}_{\mathrm{L}})^T \hat{\mathbf{C}} (\mathbf{x}_{\mathrm{true}} - \mathbf{x}_{\mathrm{L}}) \leq \Delta^2$

Rank deficiency and ill-conditioning

$(\mathbf{A}^T\mathbf{A})^{-1}$ typically not possible. Approaches that have been talked about so far

- Truncated SVD
- Stopping iterations
- Pre-conditioning
- **Regularization or prior information**

Tikhonov Regularization

$$\mathbf{x}_{\mathbf{L}_p} = \arg\min_{\mathbf{x}} \left\{ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha^2 ||\mathbf{L}_p(\mathbf{x} - \mathbf{x}_0)||_2^2 \right\}$$

$\mathbf{x}_0$ - initial estimate of $\mathbf{x}$

$\mathbf{L}_p$ - typically represents the first $(p = 1)$ or second derivative $(p = 2)$

α - regularization parameter

This gives estimates

$$\mathbf{x}_{\mathbf{L}_p} = \mathbf{x}_0 + (\mathbf{A}^T\mathbf{A} + \alpha^2\mathbf{L}_p^T\mathbf{L}_p)^{-1}\mathbf{A}^T\mathbf{b}$$

Choice of regularization parameter

Methods: L-curve, Generalized Cross Validation (GCV) and Morozov's Discrepancy Principle, UPRE, χ^2 method[2].

- α large $\rightarrow$ $\arg\min_{\mathbf{x}} \|\mathbf{L}_p(\mathbf{x}-\mathbf{x}_0)\|_2^2$

  $\mathbf{L}_p\mathbf{x} \approx \mathbf{0}$, i.e. $\mathbf{x}_{\mathbf{L}_p}$ is smooth

- α small $\rightarrow$ $(\mathbf{A}^T\mathbf{A} + \alpha^2\mathbf{L}_p^T\mathbf{L}_p)^{-1}$ DNE

  problem may stay ill-conditioned

[2] Mead et al, 2008, 2009, 2010, 2016

Choice of L_p

$$\mathbf{x}_{\mathbf{L}_p} = \arg\min_{\mathbf{x}} \left\{ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha^2 ||\mathbf{L}_p(\mathbf{x} - \mathbf{x}_0)||_2^2 \right\}$$

$\mathbf{L}_0 (= \mathbf{I})$ - requires good initial estimate $\mathbf{x}_0$, otherwise may not regularize the problem.

$\mathbf{L}_1$ - requires first derivative estimate $\mathbf{L}_1\mathbf{x}_0$, i.e. changes in $\mathbf{x}_0$, which is less information than $\mathbf{x}_0$.

$\mathbf{L}_2$ - requires $\mathbf{L}_2\mathbf{x}_0$, leaves more degrees of freedom than first derivative so that data has more opportunities to inform changes in parameter estimates.

Statistical view of Regularization (Bayesian?)

Assume parameters $\mathbf{x}$ are random variables.

- Tikhonov $\alpha^2||\mathbf{L}_p(\mathbf{x}-\mathbf{x}_0)||_2^2 \rightarrow$

  $$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}}|\mathbf{C}_x|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T\mathbf{C}_x^{-1}(\mathbf{x}-\mathbf{x}_0)}$$

  with $\mathbf{C}_x^{-1/2} = \alpha\mathbf{L}_p$

- Total variation $\lambda||\mathbf{L}_1\mathbf{x}||_1 \rightarrow$

  $$f(\mathbf{x}) = \frac{1}{2^{n/2}}|\mathbf{C}_x|^{-1/2}e^{-\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n|(\mathbf{L}_1)_{ij}(\mathbf{x}_{\mathsf{TV}})_j|_i}$$

  with $\mathbf{C}_x^{-1/2} = \lambda\mathbf{L}_1$

Hypothesis testing

Let

$$
\begin{aligned}
\mathbf{b} &= \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon} \\
\mathbf{x} &= \mathbf{x}_0 + \mathbf{f}
\end{aligned}
$$

with Null Hypothesis

$$
\begin{aligned}
\bar{\boldsymbol{\epsilon}} &= \mathbf{0} \\
\bar{\mathbf{f}} &= \mathbf{0} \\
\overline{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T} &= \mathbf{C}_b \\
\overline{\mathbf{f}\mathbf{f}^T} &= \mathbf{C}_x
\end{aligned}
$$

On what basis do we reject or fail to reject the Null Hypothesis?

How do we determine values which comprise the Null Hypotheses?

Discrepancy Principle as a χ^2 test

$$\chi^2_{\text{obs}} = (\mathbf{A}\mathbf{x}_{\mathbf{L}_p} - \mathbf{b})^T \mathbf{C}_b^{-1} (\mathbf{A}\mathbf{x}_{\mathbf{L}_p} - \mathbf{b}) \sim \chi^2(m)$$

or

$$\|\mathbf{A}_w \mathbf{x}_{\mathbf{L}_p} - \mathbf{b}_w\| \approx m$$

incorrect degrees of freedom!

χ^2 tests for regularization parameter selection

$$\chi^2_{\text{reg}} = (\mathbf{A}\mathbf{x}_{\mathbf{L}_p} - \mathbf{b})^T \mathbf{C}_b^{-1} (\mathbf{A}\mathbf{x}_{\mathbf{L}_p} - \mathbf{b}) + (\mathbf{x}_{\mathbf{L}_p} - \mathbf{x}_0)^T \mathbf{C}_x^{-1} (\mathbf{x}_{\mathbf{L}_p} - \mathbf{x}_0) \sim \chi^2(m)$$

and with appropriate assumptions on $\mathbf{A}$ that are valid in most imaging applicaitons

$$\chi^2_{\text{tvreg}} = (\mathbf{A}\mathbf{x}_{\text{TV}} - \mathbf{b})^T \mathbf{C}_b^{-1} (\mathbf{A}\mathbf{x}_{\text{TV}} - \mathbf{b}) + \sum_{i=1}^{n} \sum_{j=1}^{n} |(\mathbf{L}_1)_{ij} (\mathbf{x}_{\text{TV}})_j|_i \sim \chi^2(m)$$

Bias

Least squares estimate $\mathbf{x}_{\mathsf{L2}} = (\mathbf{A}_w^T\mathbf{A})^{-1}\mathbf{A}_w^T\mathbf{b}$ has no bias, i.e.

$$\begin{aligned}\mathbb{E}[\mathbf{x}_{\mathsf{L2}}] - \mathbf{x}_{\mathsf{true}} &= (\mathbf{A}_w^T\mathbf{A})^{-1}\mathbf{A}_w^T\mathbb{E}[\mathbf{b}] - \mathbf{x}_{\mathsf{true}} \\ &= (\mathbf{A}_w^T\mathbf{A})^{-1}\mathbf{A}_w^T\mathbb{E}[\mathbf{A}_w\mathbf{x} + \boldsymbol{\epsilon}] - \mathbf{x}_{\mathsf{true}} \\ &= \mathbf{x}_{\mathsf{true}} - \mathbf{x}_{\mathsf{true}}\end{aligned}$$

Tikhonov estimate $\mathbf{x}_{\mathbf{L}_p} = (\mathbf{A}_w^T\mathbf{A}_w + \alpha^2\mathbf{L}_p^T\mathbf{L}_p)^{-1}\mathbf{A}_w^T(\mathbf{b} - \mathbf{A}_w\mathbf{x}_0) = \mathbf{A}^\dagger(\mathbf{b} - \mathbf{A}_w\mathbf{x}_0)$ is biased, i.e.

$$\begin{aligned}\mathbb{E}[\mathbf{x}_{\mathbf{L}_p}] &= \mathbf{A}^\dagger\mathbb{E}(\mathbf{b}) - \mathbf{A}^\dagger\mathbf{A}_w\mathbb{E}[\mathbf{x}_0] \\ &= \mathbf{A}^\dagger\mathbf{A}_w\mathbb{E}[\mathbf{x}] - \mathbf{A}^\dagger\mathbf{A}_w\mathbf{x}_0 \\ &= \mathbf{A}^\dagger\mathbf{A}_w(\mathbf{x}_{\mathsf{true}} - \mathbf{x}_0)\end{aligned}$$

Resolution

Model

$$\mathbf{x}_{\mathbf{L}_p} = \mathbf{A}^\dagger \mathbf{d}_{\text{true}} = \mathbf{A}^\dagger \mathbf{A}_w \mathbf{x}_{\text{true}}$$

$\mathbf{A}^\dagger \mathbf{A}_w = \mathbf{I}$ (least squares) $\rightarrow \mathbf{x}_{\mathbf{L}_p} = \mathbf{x}_{\text{true}}$. Model resolution: $\mathbf{R}_m = \mathbf{A}^\dagger \mathbf{A}_w$.

Data

$$\mathbf{A}_w \mathbf{x}_{\mathbf{L}_p} = \mathbf{d}_{\mathbf{L}_p} \quad \text{or} \quad \mathbf{A}_w \mathbf{A}^\dagger \mathbf{d} = \mathbf{d}_{\mathbf{L}_p}$$

$\mathbf{A}_w \mathbf{A}^\dagger = \mathbf{I}$ (least squares) $\rightarrow \mathbf{d}_{\mathbf{L}_p} = \mathbf{d}$. Data resolution: $\mathbf{R}_d = \mathbf{A}_w \mathbf{A}^\dagger$.