Introduction

- **Inverse problem**: to recover an unknown object from indirect noisy observations

  **Example**: Deblurring of a signal $f$

  Forward problem: $\mu(x) = \int_0^1 A(x - t)f(t)\, dt$

  Noisy data: $y_i = \mu(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$

  Inverse problem: recover $f(t)$ for $t \in [0, 1]$
General Model

- \( A = (A_i), \ A_i : \mathcal{H} \to \mathbb{R}, \ f \in \mathcal{H} \)
- \( y_i = A_i[f] + \varepsilon_i, \ i = 1, \ldots, n \)
- The recovery of \( f \) from \( A[f] \) is ill-posed
- Errors \( \varepsilon_i \) modeled as random
- Estimate \( \hat{f} \) is an \( \mathcal{H} \)-valued random variable

Basic Question (UQ)
How good is the estimate \( \hat{f} \)?

- What is the distribution of \( \hat{f} \)?
  Summarize characteristics of the distribution of \( \hat{f} \)
  E.g., how ‘likely’ is it that \( \hat{f} \) will be ‘far’ from \( f \)?
Review: E, Var & Cov

(1) Expected value of random variable $X$ with pdf $f(x)$:

$$E(X) = \mu_X = \int x f(x) \, dx$$

(2) Variance of $X$: $\text{Var}(X) = E\left((X - \mu_X)^2\right)$

(3) Covariance of random variables $X$ and $Y$:

$$\text{Cov}(X, Y) = E\left((X - \mu_X)(Y - \mu_Y)\right)$$

(4) Expected value of random vector $X = (X_i)$:

$$E(X) = \mu_X = (E(X_i))$$
(5) Covariance matrix of $X$:

$$\text{Var}(X) = \mathbb{E}( (X - \mu_x)(X - \mu_x)^t ) = (\text{Cov}(X_i, X_j))$$

(6) For a fixed $n \times m$ matrix $A$

$$\mathbb{E}(AX) = A \mathbb{E}(X), \quad \text{Var}(AX) = A \text{Var}(X) A^t$$

Example: $X = (X_1, \ldots, X_n)$ correlated, $\text{Var}(X) = \sigma^2 \Sigma$, then

$$\text{Var} \left( \frac{X_1 + \cdots + X_n}{n} \right) = \text{Var}(1^t X / n) = \frac{\sigma^2}{n^2} 1^t \Sigma 1$$
Review: least-squares

\[ y = A\beta + \varepsilon \]

- \( A \) is \( n \times m \) with \( n > m \) and \( A^t A \) has stable inverse
- \( E(\varepsilon) = 0 \), \( \text{Var}(\varepsilon) = \sigma^2 I \)
- Least-squares estimate of \( \beta \):

\[
\hat{\beta} = \arg \min_b \| y - Ab \|^2 = (A^t A)^{-1} A^t y
\]

- \( \hat{\beta} \) is an unbiased estimator of \( \beta \):

\[
E_\beta(\hat{\beta}) = \beta \quad \forall \beta
\]
Covariance matrix of $\hat{\beta}$:

$$\text{Var}_\beta(\hat{\beta}) = \sigma^2 (A^t A)^{-1}$$

If $\varepsilon$ is Gaussian $N(0, \sigma^2 I)$, then $\hat{\beta}$ is the MLE and

$$\hat{\beta} \sim N(\beta, \sigma^2 (A^t A)^{-1})$$

Unbiased estimator of $\sigma^2$:

$$\hat{\sigma}^2 = \frac{\|y - A\hat{\beta}\|^2}{(n - m)}$$

$$= \frac{\|y - \hat{y}\|^2}{(n - \text{dof}(H))},$$

where $\hat{y} = H\hat{\beta}$, $H = A(A^t A)^{-1}A^t = \text{‘hat-matrix’}$
Confidence regions

1 − \( \alpha \) confidence interval for \( \beta_i \):

\[
\mathcal{I}_i(\alpha) : \ \hat{\beta}_i \pm t_{\alpha/2, n-m} \hat{\sigma} \sqrt{( (A^tA)^{-1} )_{i,i}}
\]

C.I.’s are pre-data

Pointwise vs simultaneous 1 − \( \alpha \) coverage:

- pointwise: \( P[ \beta_i \in \mathcal{I}_i(\alpha) ] = 1 - \alpha \ \forall i \)
- simultaneous: \( P[ \beta_i \in \mathcal{I}_i(\alpha) \ \forall i ] = 1 - \alpha \)

Easy to find 1 − \( \alpha \) joint confidence region \( R_\alpha \) for \( \beta \) which gives simultaneous coverage and if

\[
R^f_\alpha = \{ f(b) : b \in R_\alpha \} \ \Rightarrow \ P[f(\beta) \in R^f_\alpha] \geq 1 - \alpha
\]
Residuals

\[ r = y - A\hat{\beta} = y - \hat{y} \]

- \( E(r) = 0, \quad \text{Var}(r) = \sigma^2(I - H) \)
- If \( \varepsilon \sim N(0, \sigma^2 I) \), then: \( r \sim N(0, \sigma^2(I - H)) \)

\( \Rightarrow \) do not behave like original errors \( \varepsilon \)

- Corrected residuals \( \hat{r}_i = r_i / \sqrt{1 - H_{ii}} \)
- Corrected residuals may be used for model validation
Review: Bayes theorem

- Conditional probability measure $P(\cdot | B)$:
  
  $P(A | B) = \frac{P(A \cap B)}{P(B)}$ if $P(B) \neq 0$

- Bayes theorem: for $P(A) \neq 0$
  
  $P(A | B) = \frac{P(B | A)P(A)}{P(B)}$

- For $X$ and $\Theta$ with pdf’s
  
  $f_{\Theta|X}(\theta | x) \propto f_{X|\Theta}(x | \theta)f_\Theta(\theta)$

  More generally:

  $\frac{d\mu_{\Theta|X}}{d\mu_\Theta} \propto \frac{d\mu_{X|\Theta}}{d\nu}$
Basic tools: conditional probability measures and Radon-Nikodym derivatives

Conditional probabilities can be defined via Kolmogorov’s definition of conditional expectations or using conditional distributions (e.g., via disintegration of measures)
Review: Gaussian Bayesian linear model

\[ y \mid \beta \sim N(A\beta, \sigma^2 I), \quad \beta \sim N(\mu, \tau^2 \Sigma) \]

\[ \mu, \sigma, \tau, \Sigma \quad \text{known} \]

- Goal: update prior distribution of \( \beta \) in light of data \( y \).
  Compute posterior distribution of \( \beta \) given \( y \) (use Bayes theorem):

\[ \beta \mid y \sim N(\mu_y, \Sigma_y) \]

\[ \Sigma_y = \sigma^2 \left( A^t A + \left( \sigma^2 / \tau^2 \right) \Sigma^{-1} \right)^{-1} \]

\[ \mu_y = \left( A^t A + \left( \sigma^2 / \tau^2 \right) \Sigma^{-1} \right)^{-1} \left( A^t y + \left( \sigma^2 / \tau^2 \right) \Sigma^{-1} \mu \right) \]

= weighted average of \( \hat{\mu}_{1s} \) and \( \mu \)
Summarizing posterior:

\[ \mu_y = E(\mu \mid y) = \text{mean and mode of posterior (MAP)} \]
\[ \Sigma_y = \text{Var}(\mu \mid y) = \text{covariance matrix of posterior} \]

1 - \( \alpha \) credible interval for \( \beta_i \):

\[ (\mu_y)_i \pm z_{\alpha/2} \sqrt{(\Sigma_y)_{i,i}} \]

Can also construct 1 - \( \alpha \) joint credible region for \( \beta \)
Some properties

- $\mu_y$ minimizes $E\|\delta(y) - \beta\|^2$ among all $\delta(y)$

- $\mu_y \to \hat{\beta}_{1s}$ as $\sigma^2/\tau^2 \to 0$, $\mu_y \to \mu$ as $\sigma^2/\tau^2 \to \infty$

- $\mu_y$ can be used as a frequentist estimate of $\beta$:
  - $\mu_y$ is biased
  - its variance is NOT obtained by sampling from the posterior

- **Warning**: a $1 - \alpha$ credible regions MAY NOT have $1 - \alpha$ frequentist coverage

- If $\mu, \sigma, \tau$ or $\Sigma$ unknown $\Rightarrow$ use hierarchical or empirical Bayesian and MCMC methods
Parametric reductions for IP

- Transform infinite to finite dimensional problem: i.e., \( \mathcal{A}[f] \approx Aa \)

- Assumptions

\[
y = \mathcal{A}[f] + \varepsilon = Aa + \delta + \varepsilon
\]

Function representation: \( f(x) = \phi(x)^t a \)

Penalized least-squares estimate:

\[
\hat{a} = \arg \min_b \| y - Ab \|^2 + \lambda^2 b^t Sb
\]

\[
= (A^t A + \lambda^2 S)^{-1} A^t y
\]

\[
\hat{f}(x) = \phi(x)^t \hat{a}
\]
Example: \( y_i = \int_0^1 A(x_i - t)f(t) \, dt + \varepsilon_i \)

Quadrature approximation:

\[
\int_0^1 A(x_i - t)f(t) \, dt \approx (1/m) \sum_{j=1}^m A(x_i - t_j)f(t_j)
\]

Discretized system:

\[ y = Af + \delta + \varepsilon, \quad A_{ij} = A(x_i - t_j), \quad f_i = f(x_i) \]

Regularized estimate:

\[
\hat{f} = \arg \min_{g \in \mathbb{R}^m} \| y - Ag \|^2 + \lambda^2 \| Dg \|^2
\]

\[
= (A^t A + \lambda^2 D^t D)^{-1} A^t y \equiv Ly
\]

\[
\hat{f}(x_i) = f_i = e_i^t Ly
\]
Example: $\mathcal{A}_i$ continuous on Hilbert space $\mathcal{H}$. Then:

$\exists$ orthonormal $\phi_k \in \mathcal{H}$, orthonormal $v_k \in \mathbb{R}^n$ and non-decreasing $\lambda_k \geq 0$ such that $\mathcal{A}[\phi_k] = \lambda_k v_k$,

$f = f_0 + f_1$, with $f_0 \in \text{Null}(\mathcal{A})$, $f_1 \in \text{Null}(\mathcal{A})^\perp$

$f_0$ not constrained by data, $f_1 = \sum_{i=1}^n a_k \phi_k$

Discretized system:

$$y = Va + \varepsilon, \quad V = (\lambda_1 v_1 \cdots \lambda_n v_n)$$

Regularized estimate:

$$\hat{a} = \arg \min_{b \in \mathbb{R}^n} \|y - Vb\|^2 + \lambda^2 \|b\|^2 \equiv Ly$$

$$\hat{f}(x) = \phi(x)^t Ly, \quad \phi(x) = (\phi_i(x))$$
Example: $A_i$ continuous on RKHS $H = W_m([0, 1])$. Then: $H = \mathcal{N}_{m-1} \oplus H_m$ and $\exists \phi_j \in H$ and $\kappa_j \in H$ such that

$$\|y - A[f]\|^2 + \lambda^2 \int_0^1 (f_1^{(m)})^2 = \sum_i (y_i - \langle \kappa_i, f \rangle)^2 + \lambda^2 \|P_{H_m} f\|^2$$

$$f = \sum_j a_j \phi_j + \eta, \quad \eta \in \text{span}\{\phi_k\}^\perp$$

Discretized system:

$$y = Xa + \delta + \varepsilon$$

Regularized estimate:

$$\hat{a} = \arg \min_{b \in \mathbb{R}^n} \|y - Xb\|^2 + \lambda^2 b^t Pb \equiv Ly$$

$$\hat{f}(x) = \phi(x)^t Ly, \quad \phi(x) = (\phi_i(x))$$
To summarize:

\[ y = Aa + \delta + \varepsilon \]

\[ f(x) = \phi(x)^t a \]

\[ \hat{a} = \arg \min_{b \in \mathbb{R}^n} \| y - Xb \|^2 + \lambda^2 b^t P b \]

\[ = \left( A^t A + \lambda^2 P \right)^{-1} A^t y \]

\[ \hat{f}(x) = \phi(x)^t \hat{a} \]
Bias, variance and MSE of $\hat{f}$

Is $\hat{f}(x)$ close to $f(x)$ on average?

$$\text{Bias}(\hat{f}(x)) = \mathbb{E}(\hat{f}(x)) - f(x) = \phi(x)^t B_\lambda a + \phi(x)^t G_\lambda A^t \delta$$

Pointwise sampling variability:

$$\text{Var}(\hat{f}(x)) = \sigma^2 \| AG_\lambda \phi(x) \|^2$$

Pointwise MSE:

$$\text{MSE}(\hat{f}(x)) = \text{Bias}(\hat{f}(x))^2 + \text{Var}(\hat{f}(x))$$

Integrated MSE:

$$\text{IMSE}(\hat{f}(x)) = \mathbb{E} \int |\hat{f}(x) - f(x)|^2 \, dx = \int \text{MSE}(\hat{f}(x)) \, dx$$
Review: are unbiased estimators good?

- They don’t have to exist
- If they exist they may be very difficult to find
- A biased estimator may still be better: variance variance trade off
  i.e., don’t get too attached to them
What about the bias of $\hat{f}$?

- Upper bounds are sometimes possible
- Geometric information can be obtained from bounds
- Optimization approach: find max/min of bias s.t. $f \in S$
- Average bias: choose a prior for $f$ and compute mean bias over prior

Example: $f$ in RKHS $\mathcal{H}$, $f(x) = \langle \rho_x, f \rangle$

$$\text{Bias}(\hat{f}(x)) = \langle A_x - \rho_x, f \rangle$$

$$|\text{Bias}(\hat{f}(x))| \leq \|A_x - \rho_x\| \|f\|$$
Confidence intervals

- If $\varepsilon$ Gaussian and $|\text{Bias}(\hat{f}(x))| \leq B(x)$, then
  \[\hat{f}(x) \pm z_{\alpha/2} (\sigma \|AG_\lambda \phi(x)\| + B(x))\]
is a $1 - \alpha$ C.I. for $f(x)$

- Simultaneous $1 - \alpha$ C.I.’s for $E(\hat{f}(x)), x \in S$: find $\beta$ such that
  \[P\left[\sup_{x \in S} |Z^t V(x)| \geq \beta\right] \leq \alpha\]

  $Z_i$ iid $N(0, 1), V(x) = KG_\lambda \phi(x)/\|KG_\lambda \phi(x)\|$ and use results from maxima of Gaussian processes theory.
Residuals

\[ r = y - A\hat{a} = y - \hat{y} \]

- Moments:

\[
\begin{align*}
E(r) &= -A \text{Bias}(\hat{a}) + \delta \\
\text{Var}(r) &= \sigma^2(1 - H_\lambda)^2
\end{align*}
\]

**Note:** Bias(\(A\hat{a}\)) = A Bias(\(\hat{a}\)) may be small even if Bias(\(\hat{a}\)) is significant

- Corrected residuals: \(\hat{r}_i = r_i / (1 - (H_\lambda)_{ii})\)
Estimating $\sigma$ and $\lambda$

- $\lambda$ can be chosen using same data $y$ (with GCV, $L$-curve, discrepancy principle, etc.) or independent training data
  This selection introduces an additional error
- If $\delta \approx 0$, $\sigma^2$ can be estimated as in least-squares:
  \[
  \hat{\sigma}^2 = \|y - \hat{y}\|^2/(n - \text{dof}(H_\lambda))
  \]
  otherwise one may consider $y = \mu + \varepsilon$, $\mu = A[f]$ and use methods from nonparametric regression
Resampling methods

**Idea:** If \( \hat{f} \) and \( \hat{\sigma} \) are ‘good’ estimates, then one should be able to generate synthetic data consistent with actual observations

- If \( \hat{f} \) is a good estimate \( \Rightarrow \) make synthetic data

\[
y_1^* = A[\hat{f}] + \varepsilon_i^*, \ldots, y_b^* = A[\hat{f}] + \varepsilon_b^*
\]

with \( \varepsilon_i^* \) from same distribution as \( \varepsilon \). Use same estimation procedure to get \( \hat{f}_i^* \) from \( y_i^* \):

\[
y_1^* \rightarrow \hat{f}_1^*, \ldots, y_b^* \rightarrow \hat{f}_b^*
\]

Approximate distribution of \( \hat{f} \) with that of \( \hat{f}_1^*, \ldots, \hat{f}_b^* \)
Generating $\varepsilon^*$ similar to $\varepsilon$

- **Parametric resampling:** $\varepsilon^* \sim N(0, \hat{\sigma}^2 I)$

- **Nonparametric resampling:** sample $\varepsilon^*$ with replacement from corrected residuals

**Problems:**
- Possibly badly biased and computational intensive
- Residuals may need to be corrected also for correlation structure
- A bad $\hat{f}$ may lead to misleading results

**Training sets:** Let $\{f_1, \ldots, f_k\}$ be a training set (e.g., historical data)
- For each $f_j$ use resampling to estimate $\text{MSE}(\hat{f}_j)$
- Study variability of $\text{MSE}(\hat{f}_j)/\|f_j\|$
Bayesian inversion

- Hierarchical model (simple Gaussian case):

\[ y \mid a, \theta \sim N(Aa + \delta, \sigma^2 I), \quad a \mid \theta \sim F_a(\cdot \mid \theta), \quad \theta \sim F_\theta \]

\[ \theta = (\delta, \sigma, \tau) \]

- It all reduces to sampling from posterior \( F(a \mid y) \)

⇒ use MCMC methods

Possible problems:

- Selection of priors
- Convergence of MCMC
- Computational cost of MCMC in high dimensions
- Interpretation of results
Warning

Parameters can be tuned so that

\[ \hat{f} = \mathbb{E}(f \mid y) \]

but this does not mean that the uncertainty of \( \hat{f} \) is obtained from the posterior of \( f \) given \( y \).
Warning

Bayesian or frequentist inversions can be used for uncertainty quantification but:

- Uncertainties in each have different interpretations that should be understood

- Both make assumptions (sometimes stringent) that should be revealed

- The validity of the assumptions should be explored

⇒ exploratory analysis and model validation
Exploratory data analysis (by example)

Example: $\ell^2$ and $\ell^1$ estimates

data \quad \Rightarrow \quad y = Af + \varepsilon = AW\beta + \varepsilon

\[
\hat{f}_{\ell^2} = \arg\min_g \| y - Ag \|_2^2 + \lambda^2 \| Dg \|_2^2
\]

$\lambda$: from GCV

\[
\hat{f}_{\ell^1} = W\hat{\beta}, \quad \hat{\beta} = \arg\min_b \| y - AWb \|_2^2 + \lambda \| b \|_1
\]

$\lambda$: from discrepancy principle

$E(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = \sigma^2 I$

$\hat{\sigma} = \text{smoothing spline estimate}$
Example: $y$, $Af$ & $f$
Tools: robust simulation summaries

- Sample mean and variance of $x_1, \ldots, x_m$:
  \[
  \bar{X} = \frac{x_1 + \cdots + x_m}{m}
  \]
  \[
  S^2 = \frac{1}{m} \sum_{i}(x_i - \bar{X})^2
  \]

- Recursive formulas:
  \[
  \bar{X}_{n+1} = \bar{X}_n + \frac{1}{n+1} (x_{n+1} - \bar{X}_n)
  \]
  \[
  T_{n+1} = T_n + \frac{n}{n+1} (x_{n+1} - \bar{X}_n)^2, \quad S^2_n = T_n/n
  \]

- Not robust
Robust measures

- Median and median absolute deviation from median (MAD)

\[
\tilde{X} = \text{median}\{x_1, \ldots, x_m\}
\]

\[
\text{MAD} = \text{median}\left\{|x_1 - \tilde{X}_m|, \ldots, |x_m - \tilde{X}_m|\right\}
\]

- Approximate recursive formulas with good asymptotics
Stability of $\hat{f}_{\ell^2}$
Stability of $\hat{f}_{\ell_1}$
Tools: trace estimators

\[ \text{Trace}(\mathbf{H}) = \sum_i H_{ii} = \sum_i \mathbf{e}_i^t \mathbf{H} \mathbf{e}_i \]

- Expensive for large \( \mathbf{H} \)

- Randomized estimator: \( \mathbf{H} \) symmetric \( n \times n \), \( \mathbf{U}_1, \ldots, \mathbf{U}_m \) with \( \mathbf{U}_{ij} \) iid \( \text{Unif}\{-1, 1\} \). Define

\[ \hat{T}_m(\mathbf{H}) = \frac{1}{m} \sum_{i=1}^m \mathbf{U}_i^t \mathbf{H} \mathbf{U}_i \]

- \( \mathbb{E}(\hat{T}_m(\mathbf{H})) = \text{Trace}(\mathbf{H}) \)

- \( \text{Var}(\hat{T}_m(\mathbf{H})) = \frac{2}{m} \left[ \text{Trace}(\mathbf{H}^2) - \| \text{diag}(\mathbf{H}) \|^2 \right] \)
Some bounds

- **Relative variance:**

\[
V_r(\hat{T}_m) = \frac{\text{Var}(\hat{T}_m(H))}{\text{Trace}(H)^2} \leq \frac{2}{mn} \frac{S_\sigma^2}{\bar{\sigma}^2},
\]

\[
\bar{\sigma} = \frac{1}{n} \sum_i \sigma_i, \quad S_\sigma^2 = \frac{1}{n} \sum_i (\sigma_i - \bar{\sigma})^2
\]

- **Concentration:** For any \( t > 0 \),

\[
P \left( \hat{T}_m(H) \geq \text{Trace}(H)(1 + t) \right) \leq e^{-mt^2/4V_r(\hat{T}_1)}
\]

\[
P \left( \hat{T}_m(H) \leq \text{Trace}(H)(1 - t) \right) \leq e^{-mt^2/4V_r(\hat{T}_1)}
\]
**Example:** ‘Hat matrix’ $H$ in least-squares; projects onto $k$-dimensional space

$$V_r(\hat{T}_m) \leq \frac{2}{mk} \left( 1 - \frac{k}{n} \right)$$

$$P \left( \hat{T}_m(H) \geq \text{Trace}(H) \left( 1 + t \right) \right) \leq e^{-mkt^2/8}.$$  

**Example:** Approximating a determinant of $A > 0$:

$$\log(\text{Det}(A)) = \text{Trace}(\log(A)) \approx \frac{1}{m} \sum_i U_i^t \log(A) U_i$$

$log(A)U$ requires only matrix-vector products with $A$
CI for bias of $\hat{Af}_\ell^2$

- Good estimate of $f \Rightarrow$ good estimate of $Af$

- $y$: direct observation of $Af$; less sensitive to $\hat{\lambda}$

- $1 - \alpha$ CIs for bias of $\hat{Af}$ ($\hat{y} = Hy$):

$$\frac{\hat{y}_i - y_i}{1 - H_{ii}} \pm z_{\alpha/2} \hat{\sigma} \sqrt{1 + \frac{(H^2)_{ii} - (H_{ii})^2}{(1 - H_{ii})^2}}$$

$$\frac{\hat{y}_i - y_i}{\text{Trace}(I - H)/n} \pm z_{\alpha/2} \hat{\sigma} \frac{1}{\sqrt{\text{Trace}(I - H)/n}}$$
95% CIs for bias
Coverage of 95% CIs for bias
Bias bounds for $\hat{f}_{\ell^2}$

For fixed $\lambda$:

$$\text{Var}(\hat{f}_{\ell^2}) = \sigma^2 G(\lambda)^{-1} A^t A G(\lambda)^{-1}$$

$$\text{Bias}(\hat{f}_{\ell^2}) = \mathbb{E}(\hat{f}_{\ell^2}) - f = B(\lambda),$$

$$G(\lambda) = A^t A + \lambda^2 D^t D, \quad B(\lambda) = -\lambda^2 G(\lambda)^{-1} D^t Df.$$ 

Median bias: $\text{Bias}_M(\hat{f}_{\ell^2}) = \text{median}(\hat{f}_{\ell^2}) - f$

For $\lambda$ estimated:

$$\text{Bias}_M(\hat{f}_{\ell^2}) \approx \text{median}[B(\hat{\lambda})]$$
Hölder’s inequality

\[ |B(\hat{\lambda})_i| \leq \hat{\lambda}^2 U_{p,i}(\hat{\lambda}) \|Df\|_q \]
\[ |B(\hat{\lambda})_i| \leq \hat{\lambda}^2 U_{p,i}^w(\hat{\lambda}) \|\beta\|_q \]

\[ U_{p,i}(\hat{\lambda}) = \|DG(\hat{\lambda})^{-1}e_i\|_p \text{ or } U_{p,i}^w(\hat{\lambda}) = \|WD^tDG(\hat{\lambda})^{-1}e_i\|_p \]

Plots of \( U_{p,i}(\hat{\lambda}) \) and \( U_{p,i}^w(\hat{\lambda}) \) vs \( x_i \)
$L^2$ bias and bounds

$L^2$ bias and wavelet bounds
Assessing fit

- Compare $y$ to $\hat{A}f + \hat{\varepsilon}$

- Parametric/nonparametric bootstrap samples $y^*$:
  
  (i) (parametric) $\varepsilon_i^* \sim N(0, \hat{\sigma}^2 I)$, or (nonparametric) from corrected residuals $r_c$
  
  (ii) $y_i^* = \hat{A}f + \varepsilon_i^*$

- Compare statistics of $y$ and $\{y_1^*, \ldots, y_B^*\}$ (blue, red)

- Compare statistics of $r_c$ and $N(0, \hat{\sigma}^2 I)$ (black)

- Compare parametric vs nonparametric results (red)

- Example: Simulations for $\hat{A}f_{\ell^2}$ with:
  
  $y^{(1)}$ good
  $y^{(2)}$ with skewed noise
  $y^{(3)}$ biased
Example

\[ T_1(y) = \min(y) \]
\[ T_2(y) = \max(y) \]
\[ T_k(y) = \frac{1}{n} \sum_{i=1}^{n} (y - \text{mean}(y))^k / \text{std}(y)^k \quad (k = 3, 4) \]
\[ T_5(y) = \text{sample median}(y) \]
\[ T_6(y) = \text{MAD}(y) \]
\[ T_7(y) = \text{1st sample quartile}(y) \]
\[ T_8(y) = \text{3rd sample quartile}(y) \]
\[ T_9(y) = \# \text{runs above/below median}(y) \]

\[ P_j = P \left( |T_j(y^*)| \geq |T_j^o| \right), \quad T_j^o = T_j(y) \]
Assessing fit: Bayesian case

- **Hierarchical model:**

  
  \[
  y \mid f, \sigma, \gamma \sim N(Af, \sigma^2 I)
  \]

  \[
  f \mid \gamma \propto \exp \left( -f^t D^t D f / 2\gamma^2 \right)
  \]

  \[
  (\sigma, \gamma) \sim \pi
  \]

  with \( E(f \mid y) = \left( A^t A + \frac{\sigma^2}{\gamma^2} D^t D \right)^{-1} A^t y \)

- **Posterior mean** = \( \hat{f}_{\ell^2} \) for \( \gamma = \sigma/\lambda \)

- Sample \( (f^*, \sigma^*) \) from posterior of \( (f, \sigma) \mid y \Rightarrow \) and simulated data \( y^* = Af^* + \sigma^* \varepsilon^* \) with \( \varepsilon^* \sim N(0, I) \)

- Explore consistency of \( y \) with simulated \( y^* \)
Simulating $f$

- Modify simulations from $y_i^* = \hat{A}f + \varepsilon_i^*$ to
  \[ y_i^* = Af_i + \varepsilon_i^*, \quad f_i \text{ ‘similar’ to } f \]

- Frequentist ‘simulation’ of $f$?
  One option: $f_i$ from historical data (training set)

- Check relative MSE for the different $f_i$
Bayesian or frequentist?

- Results have different interpretations. Different meanings of probability: $P \neq P$
- Frequentists can use Bayesian methods to derive procedures; Bayesians can use frequentist methods to evaluate procedures
- Consistent results for small problems and lots of data
- Possibly very different results for complex large-scale problems
- Theorems do not address relation of theory to reality
Personal view

- Neither is a substitute for detective work, common sense and honesty

- Some say that Bayesian is the only way, some others that frequentist is the best way, while those with vision see
Application: experimental design

Classic framework

\[ y_i = f(x_i)^t \theta + \varepsilon_i \Rightarrow y = F\theta + \varepsilon, \]

- Experimental configurations \( x_i \in \mathcal{X} \subset \mathbb{R}^p \)
- \( \theta \in \mathbb{R}^m \) unknown parameters
- \( \varepsilon_i \) iid zero-mean, variance \( \sigma^2 \)

Experimental design question: How many times should each configuration \( x_i \) be used to get ‘best’ estimate of \( \theta \)?

i.e., variance \( \sigma_i^2 \) for each selected \( x_i \) is controlled by averaging independent replications
More general approach: Control the $\sigma_i^2$ to get ‘best’ estimate of $\theta$ (i.e., not necessarily by averaging)

Let $\sigma_{t,i}$ be ‘optimal’ target variance for $x_i$. Set constraint:

$$\sum_i \frac{\sigma_i^2}{\sigma_{t,i}^2} = N$$

for fixed $N \in \mathbb{R}$.

Define $w_i = \frac{\sigma_i^2}{N\sigma_{t,i}^2} \Rightarrow$

$$w_i \geq 0, \quad \sum_i w_i = 1 \quad \text{and} \quad \sigma_{t,i}^2 = \frac{\sigma_i^2}{N} w_i^{-1}$$

i{}th experimental configuration: $(x_i, w_i)$

Experiment $\xi$: $\{x_1, \ldots, x_n, w_1, \ldots, w_n\}$
Well-posed case: $F^tF = \sum_i f(x_i)f(x_i)^t$ well conditioned

$$\hat{\theta} = \arg \min \sum_i w_i(y_i - f(x_i)^t\theta)^2 = \left( F^tWF \right)^{-1} F^tWy$$

$W = \text{diag}(w_i)$

$\hat{\theta}$ unbiased with

$$\text{Var}(\hat{\theta}) = \sigma^2 \left( F^tWF \right)^{-1} = \sigma^2 M(\xi)^{-1},$$

$M(\xi) = \text{information matrix} = F^tWF = \sum_i w_i f(x_i)f(x_i)^t$

Write $D(\xi) = M(\xi)^{-1}$
Optimization problem:

\[
\min_w \psi(D(\xi)) \quad \text{s.t.} \quad w \geq 0, \quad \sum w_i = 1,
\]

\(D\)-design: \(\psi = \det\)

\(A\)-design: \(\psi = \text{Trace} \quad \Rightarrow \quad \text{minimization of } \text{MSE}(\hat{\theta})\)

Optimal \(w_i\) independent of \(\sigma\) and \(N\)
Ill-posed problems

- 1D magnetotelluric example:

\[
d_j = \int_0^1 \exp(-\alpha_j x) \cos(\gamma_j x) m(x) \, dx + \varepsilon_j \quad j = 1, \ldots, n
\]

- \(\gamma_j\): recording frequencies
- \(\alpha_j\): determined by frequencies and background conductivity
- \(m(x)\): conductivity

- Goal: Choose \(\gamma_j, \alpha_j\) to determine smooth \(m\) consistent with data
Ill-posed problems

- Ill-conditioning of $F$ requires regularization
- Regularized estimate

$$
\hat{\theta} = \arg \min \sum_i w_i [y_i - f(x_i)^t \theta]^2 + \lambda^2 \|D\theta\|^2
$$

$$
= \left( F^t W F + \lambda^2 D^t D \right)^{-1} F^t W y
$$

- MSE($\hat{\theta}$) depends on unknown $\theta$ ⇒ control ‘average’ MSE
- Use prior moment conditions:

$$
E(y \mid \theta, \gamma) = F\theta, \quad \text{Var}(y \mid \theta, \gamma) = \sigma^2 W^{-1}/N
$$

$$
E(\theta \mid \gamma) = \theta_0, \quad \text{Var}(\theta \mid \gamma) = \gamma^2 \Sigma
$$

$$
E(\gamma^2) = \mu_\gamma^2,
$$

with $\sigma$ and $\Sigma$ known and $\gamma = \sigma/\lambda$
Affine estimate minimizing Bayes risk: \( \delta^2 = E(1/\lambda^2) \)

\[
\hat{\theta} = \left( N\delta^2 F^tWF + \Sigma^{-1} \right)^{-1} \left( N\delta^2 F^tWy + \Sigma^{-1} \theta_0 \right)
\]

\[
E[ (\hat{\theta} - \theta)(\hat{\theta} - \theta)^t ] = \frac{\tau \sigma^2}{N} \left( \tau F^tWF + (1 - \tau)\Sigma^{-1} \right)^{-1}
\]

\( \tau \in (0, 1) \)

A-design optimization:

\[
\begin{align*}
\min_w & \quad \text{Trace} \left( \tau F^tWF + (1 - \tau)\Sigma^{-1} \right)^{-1} \\
\text{s.t.} & \quad w \geq 0, \sum w_i = 1.
\end{align*}
\]

Optimal \( w_i \) depend on \( \sigma \) and \( N \)
Readings

Basic Bayesian methods:

Bayesian methods for inverse problems:

Discussion of frequentist vs Bayesian statistics

Tutorial on frequentist and Bayesian methods for inverse problems
- Stark, PB & Tenorio, L, 2011, A primer of frequentist and Bayesian inference in inverse problems. In *Computational Methods for Large-Scale Inverse Problems and Quantification of Uncertainty*, pp.9–32, Wiley

Statistics and inverse problems
- Special section in *Inverse Problems*: Volume 24, Number 3, June 2008

Mathematical statistics