Uncertainty Quantification for Inverse Problems

November 7, 2011
Outline

UQ and inverse problems

Review: least-squares

Review: Gaussian Bayesian linear model

Parametric reductions for IP

Bias, variance and MSE for Tikhonov

Resampling methods

Bayesian inversion

Exploratory data analysis (by example)
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

- \( A = (A_i), \ A_i : \mathcal{H} \rightarrow \mathbb{R}, \ f \in \mathcal{H} \)
- \( y_i = A_i[f] + \varepsilon_i, \ i = 1, \ldots, n \)
- 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: least-squares

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

- \( A \) is \( n \times m \) with \( n > m \) and \( A^tA \) has stable inverse
- \( E(\varepsilon) = 0, \ Var(\varepsilon) = \sigma^2 I \)
- Least-squares estimate of \( \beta \): 
  \[ \hat{\beta} = \text{arg min}_b \| y - Ab \|^2 = (A^tA)^{-1} A^t y \]
- \( \hat{\beta} \) 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 $\sigma^2$:

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

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

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

- $1 - \alpha$ confidence interval for $\beta_i$:

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

- C.I.’s are pre-data

- Pointwise vs simultaneous $1 - \alpha$ coverage:

  pointwise: $P[\beta_i \in I_i(\alpha)] = 1 - \alpha \quad \forall i$

  simultaneous: $P[\beta_i \in I_i(\alpha) \forall i] = 1 - \alpha$

- Easy to find $1 - \alpha$ joint confidence region $R_\alpha$ for $\beta$ which gives simultaneous coverage
Residuals

\[ r = y - \hat{y} = (I - H)y \]

- \( E(r) = 0 \), \( \text{Var}(r) = \sigma^2(I - H) \)
- If \( \varepsilon \sim N(0, \sigma^2 I) \), then: \( r \sim N(0, \sigma^2(I - H)) \)
  \[ \Rightarrow \text{do not behave like original errors } \varepsilon \]
- Corrected residuals \( \hat{r}_i = r_i / \sqrt{1 - H_{ii}} \)
- Corrected residuals may be used for model validation
- Other types of residuals (e.g., recursive)
Review: Gaussian Bayesian linear model

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

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

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

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

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

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

\[ = \text{weighted average of } \hat{\mu}_{ls} \text{ and } \mu \]
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 $\mathbb{E}\|\delta(y) - \beta\|^2$ among all $\delta(y)$

- $\mu_y \rightarrow \hat{\beta}_{ls}$ as $\sigma^2/\tau^2 \rightarrow 0$, $\mu_y \rightarrow \mu$ as $\sigma^2/\tau^2 \rightarrow \infty$

- $\mu_y$ can be used as a **frequentist estimate** of $\beta$ but:
  - $\mu_y$ 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., $\mathcal{A}[f] \approx \mathbf{Aa}$

- Assumptions

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

**Function representation:**

$$f(x) = \phi(x)^t a + \eta(x), \quad \mathcal{A}[\eta] = 0$$

**Penalized least-squares estimate:**

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

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

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

Quadrature approximation:

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

Discretized system:

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

Regularized estimate:

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

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

\[
\hat{f}(x_i) = f_i = e_i^t Ly
\]
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: \( \mathcal{A}_i \) continuous on RKHS \( \mathcal{H} = W_m([0, 1]) \).

Then: \( \mathcal{H} = N_{m-1} \oplus \mathcal{H}_m \) and \( \exists \phi_j \in \mathcal{H} \) and \( \kappa_j \in \mathcal{H} \) such that

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

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

Discretized system:

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

Regularized estimate:

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

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

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

\[ f(x) = \phi(x)^t a + \eta(x), \quad A[\eta] = 0 \]

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

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

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

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

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

- 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:

  \[
  \text{IMSE}(\hat{f}(x)) = E \int |\hat{f}(x) - f(x)|^2 \, dx = \int \text{MSE}(\hat{f}(x)) \, dx
  \]
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

$$\hat{f}(x) \pm (z_{\alpha/2} \sigma \| A G_\lambda \phi(x) \| + B(x))$$

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:

\[ E(r) = -A \text{Bias}(\hat{a}) + \delta \]
\[ \text{Var}(r) = \sigma^2 (I - H_\lambda)^2 \]

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

  $$\hat{\sigma}^2 = \|y - \hat{y}\|^2 / (n - \text{dof}(\hat{y}))$$

  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

  $$y_1^* = A[\hat{f}] + \epsilon_i^*, \ldots, y_b^* = A[\hat{f}] + \epsilon_b^*$$

  with $\epsilon_i^*$ from same distribution as $\epsilon$. 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 $\text{MSE}(\hat{f}_j)/\|f_j\|$
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, improper priors
- Convergence of MCMC
- Computational cost of MCMC in high dimensions
- Interpretation of results
Warning

Parameters can be tuned so that

$$\text{Tikhonov estimate } = \hat{f} = E(f \mid y)$$

but this does not mean that the uncertainty of $\hat{f}$ is obtained from the posterior of $f$ given $y$
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: $l^2$ and $l^1$ estimates

data = \[ y = Af + \epsilon = AW\beta + \epsilon \]

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

$\lambda$: from GCV

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

$\lambda$: from discrepancy principle

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

$\hat{\sigma} = \text{smoothing spline estimate}$
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:
  \[
  \bar{X}_{n+1} = \bar{X}_n + \frac{1}{n+1} (x_{n+1} - \bar{X}_n)
  \]
  \[
  T_{n+1} = T_n + \frac{n}{n+1} (x_{n+1} - \bar{X}_n)^2, \quad S_n^2 = T_n/n
  \]

- 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 $\hat{f}_{\ell^1}$
Cl for bias of $\hat{A}f_{\ell^2}^*$

- 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 $\hat{A}f$ ($\hat{y} = Hy$):

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

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

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

Plots of \( U_{p,i}(\hat{\lambda}) \) and \( U_{p,i}(\hat{\lambda}) \) vs \( x_i \)
Assessing fit

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

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

- 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

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

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

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

- Hierarchical model:

\[
\begin{align*}
    y | f, \sigma, \gamma & \sim N(Af, \sigma^2 I) \\
    f | \gamma & \propto \exp \left( -f^t D^t D f / 2\gamma^2 \right) \\
    (\sigma, \gamma) & \sim \pi \\
    \text{with } \quad E(f | y) = \left( A^t A + \frac{\sigma^2}{\gamma^2} D^t D \right)^{-1} A^t y
\end{align*}
\]

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

- Sample \((f^*, \sigma^*)\) from posterior of \((f, \sigma) | y \) \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
- 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
Readings

Basic Bayesian methods:

Bayesian methods for inverse problems:

Discussion of frequentist vs Bayesian statistics

Tutorial on frequentist and Bayesian methods for inverse problems

Statistics and inverse problems
- Marzouk, Y & Tenorio, L, 2012, *Uncertainty Quantification for Inverse Problems*
- Special section in *Inverse Problems*: Volume 24, Number 3, June 2008

Mathematical statistics