ENGSCI 721

INVERSE PROBLEMS

Oliver Maclaren
oliver.maclaren@auckland.ac.nz

MODULE OVERVIEW

Inverse Problems (*Oliver Maclaren*) [~8 lectures/2 tutorials]

1. *Basic concepts* [4 lectures]
   Forward vs inverse problems. Well-posed vs ill-posed problems. Algebra and calculus of inverse problems (matrix calculus, generalised inverses etc). Regularisation and trade-offs.
2. *More regularisation* [3 lectures]
   Higher-order Tikhonov regularisation, truncated singular value decompositions, iterative regularisation.

MODULE OVERVIEW

3. *Preview of the statistical view of inverse problems*
   [1 lectures]

   Bayesians, Frequentists and all that. Basic frequentist analysis.

LECTURE 8: INTRODUCTION TO STATISTICAL APPROACHES TO INVERSE PROBLEMS

Topics:

- From approximate models to probabilistic models
- Expectation as a generalised inverse; analog/plug-in methods
- Likelihood and maximum likelihood estimation
- Bayesians, Frequentists and all that
- Basic ideas of frequentist evaluation of methods

Engsci 721 Lecture 8.

Intro. to statistical approach to inverse problems.

Warning: fraught topic!

- Statistics lies at the intersection of
  - applied science methodology
  - experimental design
  - philosophy of science
  - mathematics
  - etc!
- statisticians (& non-statisticians) constantly argue about foundations!

↳ rule of thumb: assume everyone (incl. me..., but esp. the dogmatic advocates ---) are wrong about some parts.

↳ different approaches have diff. pros & cons & domains of application

How to navigate?

Wittgenstein's ladder:

→ use some 'ladder' concepts to help 'climb' towards understanding

→ be aware that you likely need to throw them away to reach a 'higher level' of understanding!

(True of all learning, but especially for statistics)

Starting point: constant model

"$y \approx \theta$": $\theta \rightarrow \boxed{F} \rightarrow y$

Above is theoretical 'template' for model runs. (realisations)

Consider 'running' model m times, get:

$$\{(\theta, y)_1, (\theta, y)_2, \ldots (\theta, y)_m\}$$

If θ is constant (fixed each time)

but only have $y \approx \theta$, get

$\{(\theta, y_1), (\theta, y_2), \ldots (\theta, y_m)\}$, where $y_i \neq y_j$ for $i \neq j$ (in gen.)

ie

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{bmatrix} \approx \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \theta = \vec{1}_m \theta$$

↑ using explicit over bar/arrow here to distinguish from scalar y, but will drop

↑ vector of m ones

From the perspective we've taken so far, to det. θ from y, we want to 'solve' the tall system:

'$y = 1_m \theta$' (vector eqn of obs., have dropped over-bars)

ie "$\begin{bmatrix} y \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix} [\theta]$"

for θ given y (vector of obs).

Note: Overdetermined system (m eqns, one unknown)

$A = \begin{bmatrix} 1 \end{bmatrix}$ is a tall matrix with LI cols (single col.)

$\Rightarrow$ no proper inverse, but has generalised inverse

Also: $1_m^T 1_m = \overbrace{[1\ 1\ 1 \ldots]}^{m} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \Big\}^{m}$

$= m$

& $1_m^T y = [1\ 1 \ldots 1] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix} = \Sigma y_i$

→

So:

$$A^+ = (A^T A)^{-1} A^T = (1_m^T 1_m)^{-1} 1_m^T$$

$$= (m)^{-1} 1_m^T$$

$$= \frac{1}{m} 1_m^T$$

& our (pseudo-inverse) soln is

$$\theta^+ = A^+ y = \frac{1}{m} 1_m^T y$$

$$= \frac{1}{m} \sum_i y_i$$

$$= \text{mean}(y)$$

$$= \mathbb{E}_m(y)$$

(see Ru's material for more on $\mathbb{E}$)

sample based expected value.

→ ie when estimating a constant from noisy data, the generalised inverse solves the over-determined tall system by using:

$$\theta = \text{sample mean}(y)$$

$$= \mathbb{E}_m(y).$$

→ The data resol. operator $R_D = AA^+ = 1_m \mathbb{E}_m$ projects y to the constant vector with same mean as y: can only 'resolve' average.

This approach is quite general: don't need explicit assumptions on the 'distribution of errors'

→ pro & con

→ many (not all!) statisticians prefer to explicitly model the errors probabilistically

---

Probabilistic 'error':

→ Capture all sources of 'noise'
- → measurement error
- → missing factors etc.

For example, additive error model:

$$Y = \theta + e, \quad \mathbb{E}[e] = 0$$

- Y: random variable
- θ: still const.
- e: random variable
- $\mathbb{E}[e] = 0$: mean of error is zero

$$\text{obs} = \text{signal} + \text{noise}$$

Realisations:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \theta + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix}$$

- y: observe
- θ: want
- e: don't observe, but know something about, $\mathbb{E}[e] = 0$

- What does this mean?
- How to solve in the statistical sense?

Meaning?

→ From approx. models to probability models:

Old story

$\theta \rightarrow [F] \rightarrow y$ } 'template' for approximate model (ie $y_{obs} \approx y$)

New story

$\theta \rightarrow [F] \rightarrow Y = \theta + e$ } 'template' for realisations from additive error probabilistic model.

(e: random vars)

More generally

$\theta \rightarrow [F] \rightarrow Y \sim \mathbb{P}_Y(Y)$ } 'template' for realisations from arbitr. probabilistic model.

($\mathbb{P}_Y$: probability model for Y.)

- Forward mapping maps from parameters, theories etc to random variables / prob. dist. for these random variables.
- We observe a finite number of realisations of these random vars / from these distr.

How to solve in statistical approach?

Many approaches! Depends on 'philosophy'.

Approach 1.

1A. Define 'ideal' (infinite data) target } population vs. sample

1B. 'Plug-in' sample analog }

Consider: $Y = \theta + e, \ \mathbb{E}[e] = 0$ } ideal 'template'/'recipe' for generating as many realisations as you want

(1A). $\mathbb{E}[Y] = \mathbb{E}[\theta + e]$

$= \mathbb{E}[\theta] + \mathbb{E}[e]$ (linearity)

$= \theta + 0$

θ constant by assumption

$\Rightarrow \theta = \mathbb{E}[Y]$ } note expected value under 'infinite' observations

(1B) Sample version of expected value:

$\theta = \mathbb{E}[Y] \approx \mathbb{E}_m[Y] = \frac{1}{m} \sum_i y_i$

$\Rightarrow \theta_{est} = \theta^+ =$ sample mean

→ same as before!

Morals: → There is a close correspondence between generalised inverses, projections etc & expectations of random vars.

→ If can't solve in usual sense (eg over-determined), we can solve in 'average sense'

↳ match expected value.

⇒ We can 'plug in' sample versions of our 'ideal' quantities to get estimates

↳ see eg 'An introduction to the bootstrap' by Efron & Tibshirani

'Analog estimation methods in econometrics' by Manski

→ Same sort of ideas carry over to non-trivial models, conditional expectations etc.

→ The 'functional analysis' approach to probability and statistics

- 'Probability via expectation' by Whittle
- Hilbert space methods in probability & statistics by Small & McLeish

[Uses prob. ideas but only 'minimally']

Estimate?

We used a sample-based estimate of θ:

$$\theta_{est} = \frac{1}{m} \sum_i y_i \approx \mathbb{E}[Y] = \theta$$

Is this a good estimate? What other ways are there of estimating things?

First: More detailed probability modelling.

Instead of just

$$Y = \theta + e, \quad \mathbb{E}[e] = 0$$

Introduce a more detailed model of e:

$$\boxed{Y = \theta + e}, \quad \boxed{e \sim N(0, \sigma^2)}$$

where $e \sim N(0, \sigma^2)$

means 'normally distributed'

with zero mean & variance σ^2

$$\mathbb{E}[e] = 0 \qquad \mathrm{Var}(e) = \sigma^2$$

→ see Ru's part for (much) more!

→

Normal/Gaussian distribution

→ Ru's favourite topic

→ just mention briefly for now...

The probability density for a vector y representing an m-dimensional IID sample of size m from a normal (Gaussian) distribution with constant mean θ & constant variance σ^2 is:

$$\boxed{p_Y(Y = y; \theta) = \frac{1}{(2\pi\sigma^2)^{1/2}} \cdot \exp\left\{ -\frac{1}{2} \left\| \frac{y - 1\theta}{\sigma} \right\|^2 \right\}}$$

('given θ', when θ isn't rec. a random var. some, esp. Bayesians just use $P(Y|\theta)$ instead of $P(Y;\theta)$).

(assuming σ known here)

(θ scalar, 1θ vector)

Likelihood

Given an observation/realisation y of Y,

the **likelihood function** is defined as

$$\mathcal{L}(\theta; y) \propto P(Y = y; \theta)\, dy \qquad (P(Y=y)dy \approx P(Y=y))$$

i.e.

- is a function of the parameter
- for each parameter value is proportional to the probability of the observed data under the corresponding model.
- key is relative values eg

$$\frac{\mathcal{L}(\theta_1; y)}{\mathcal{L}(\theta_2; y)} = \frac{P(Y = y; \theta_1)}{P(Y = y; \theta_2)}$$

→ constants & 'dy' cancel.

Intuitively, the **likelihood** measures how well (relatively) the model corresponding to a given parameter value 'fits' the data.

→ if $P(Y = y; \theta)$ is (relatively) high for a given θ then the model gives 'high' relative probability to the data & hence provides a good 'fit'

→ note: often write $P(Y; \theta) = P(Y|\theta)$ but still: (cf. $P(\theta|Y)$)

$$P(Y = y; \theta) \neq P(\theta|Y)$$

→ likelihood is not a probability of the parameter (is prob. of data)

→ it measures how 'compatible' a model & data are

→ doesn't require θ to be random

→ isn't additive

etc.

'Maximum likelihood' estimation determines a point estimate by solving:

$$\theta_{ML} = \arg\max_{\theta} \mathcal{L}(\theta; y)$$

} "parameter that makes data most probable"

Note: $\arg\max_{\theta} F(\theta) = \arg\max_{\theta} \ln F(\theta)$ ← natural log.

under general conditions. (ln monotonic)

$$\arg\max_{\theta} A \cdot F(\theta) + B = \arg\max_{\theta} F$$

for A, B indep. of θ, $A > 0$

&

$$\arg\max_{\theta} F(\theta) = \arg\min_{\theta} -F(\theta)$$

Now:

Define $\ell(\theta; y) = \ln \mathcal{L}(\theta; y)$

= "log-likelihood function"

$$\Rightarrow \begin{aligned} \theta_{ML} &= \arg\max \mathcal{L}(\theta; y) \\ &= \arg\max \ell(\theta; y) \\ &= \arg\min -\ell(\theta; \theta) \end{aligned}$$

For normal distribution as before:

$$\mathcal{L}(\theta; y) = C \cdot \exp\left\{ -\tfrac{1}{2} \left\| \tfrac{y - 1\theta}{\sigma} \right\|^2 \right\}$$

for some constant C.

$$\Rightarrow \ln \mathcal{L}(\theta; y) = \ln C - \tfrac{1}{2} \left\| \tfrac{y - 1\theta}{\sigma} \right\|^2$$

$$\Rightarrow \arg\max_{\theta} \ln \mathcal{L}(\theta; y) = \arg\min \left\{ -\ln \mathcal{L}(\theta; y) \right\}$$

$$= \arg\min_{\theta} \tfrac{1}{2\sigma^2} \| y - 1\theta \|^2$$

$$= \arg\min_{\theta} \| y - 1\theta \|^2$$

Take away: max likelihood estimate under additive normal errors* = least squares estimate

(= estimate from generalised inverse = sample plug-in)

[* for fixed σ.]

See 'lectorial on maximum likelihood estimation' on Canvas (from a short course I presented at)

for more on likelihood-based estimation

Regularisation?

→ Make up more observations, stack & do max likelihood/least squares ('mixed estimation' in freq. econometrics, 'prior likelihood' in 'pure likelihood' approach)

→ prior distribution & 'MAP' estimation (Bayesian approach → see Rui's part)

(see L3 comments)

So.... many approaches give same point estimate

→ is this a 'good' estimate?

→ what about 'interval' estimates & uncertainty?

Ru will cover Bayes... So...

I will sketch some frequentist ideas.

Note: Beware of Bayesian descriptions of frequentist inference & vice-versa ... often dogmatic, often wrong!

Also applies to Bayesian descriptions of Bayes & freq. description of freq.!

↳ with that in mind....

What is a 'good estimate' of θ?

Bayesians

- want a 'posterior probability' $P(\theta|y)$ that represents their personal 'belief'/'state of information' etc about the true value of θ given data y & prior beliefs $P(\theta)$. Point estimate might be max, mean, median of $P(\theta|y)$
- to make it more 'objective', need to agree on the prior
- Can think of in terms of (personal) decision theory & maximising subjective expected utility.

Frequentists

- want more 'objective' approach: prob. represents 'variability in the world' rather than personal uncertainty
- focus on evaluating methods of drawing conclusions by evaluating performance under 'repeated use' given noisy observations/realisations of random processes
- a 'good' estimate is an output of a reliable method ... performs well in the 'long run' ... but did it perform well in this case?
- can think of in terms of a two player game between you & nature

Frequentist Statistics : trust the process

Provides methods of evaluating the performance of estimation methods, under repeated use in a 'noisy' environment (noisy observations/realisations of random variables), regardless of where the methods come from... Eg can do:

> 'frequentist performance of Bayesian methods'!

Methods?

> An estimator is a method (function) that provides estimates of an estimand (parameter/target) given observations

→ eg the generalised inverse is an estimator:

$$\theta^+ = A^+ y$$

(θ^+: estimate of θ; A^+: estimator; y: data (realisation of random var))

[where:

θ ⇄ y via A ← forward model (for mean)

θ: parameter; y: data (or sample)

A^+ ← estimator]

Frequentist statistics as a two player game.

- nature chooses fixed θ & generates random data $Y \sim \mathbb{P}(Y;\theta)$
- you come up with a 'guessing method' T (estimator) that gives an estimate of θ given any realisation of Y (ie you see data, not θ)

$$\text{estimate for } \theta = T(Y)$$

(T: method; Y: data realisation)

→ The estimate is random because Y (data) is random, even if θ fixed (but unknown)

Here:

- Probability applies to
  - 'actual variation' in nature (aleatoric)
  - the estimates that the guessing method 'spits out' (since estimator is a function of the data, $T(Y)$)
- Does not apply to 'fixed but unknown' things (eg θ)
  - Not rep. of personal uncertainty (epistemics)

Types of estimate: Many!

Point estimate: 'single best guess'
$T(Y) = \theta_{est} \in$ parameter space
} eg max likelihood estimate

Interval/set estimate: a set (eg interval) of parameter values indicating 'good', 'plausible' etc range of parameter values/estimates, ie $T(Y) = S_{\theta_{est}} \subseteq$ param space

↳ eg confidence interval.

Function or distribution* estimate: a distribution or function over parameter values, indicating 'good', 'plausible' etc ranges of parameter values/estimates.
→ likelihood function
→ posterior distribution
→ confidence distribution
etc.

(* more common for Bayes)

Performance:

Point estimator

- Want to guarantee that for all θ (since nature's choice!) we have good expected performance in repeated use, eg want:

$$\mathbb{E}\left[d(T(Y), \theta)\right] \text{ for } Y \sim \mathbb{P}(Y;\theta)$$

to be small.

Example: want $\mathbb{E}\left[\|T(Y) - \theta\|^2\right]$ small for all θ & $Y \sim \mathbb{P}(Y;\theta)$

(low mean squared error)

Interval estimator

- Want to guarantee that for all of nature's choices of θ, our (random) interval guesses will often (eg 95% of the time we get data) contain ('trap') the true but unknown value.

Eg want, for all θ:

$C(Y)$ is 95% confidence interval procedure →

$$\left\{ \mathbb{P}(C(Y) \ni \theta ; \theta) \geq 0.95, \text{ for } Y \sim \mathbb{P}(Y;\theta) \right.$$

- $C(Y)$: random interval
- θ: fixed
- 0.95: coverage level
- $Y \sim \mathbb{P}(Y;\theta)$: (confidences)

Confidence intervals?

A confidence interval procedure is a method of guessing intervals given data realisations:

eg $[L(Y), U(Y)]$ } random* interval guess

($L(Y)$: lower limit; $U(Y)$: upper limit; Y: random var)

(*depends on random data)

The method has a given coverage probability (eg 95% prob) which is the prob. of containing the true (usually fixed but unknown) value θ, ie $\mathbb{P}([L(Y), U(Y)] \ni \theta ; \theta)$, under repeated samples/realisations of Y.

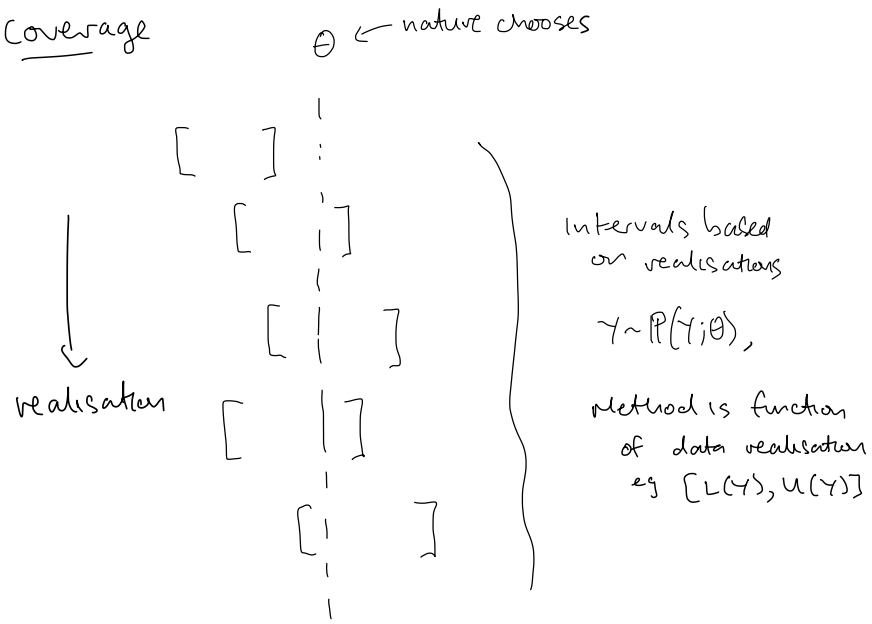
A given interval realisation either contains the true value or doesn't. Not 95% prob.

↳ Compare $\mathbb{P}(\tilde{Y} = 3)$ (random var) to $\mathbb{P}(4 = 3)$ (realisation) } doesn't make sense

eg $\mathbb{P}_Y([L(Y), U(Y)] \ni \theta ; \theta)$ ✓

$\mathbb{P}_Y([3, 5] \ni \theta ; \theta)$ × (either 0 or 1.)

Coverage



Above 4/5 intervals trap the true value (empirical coverage = 0.8)

Coverage checks

- given a method, you can play the role of [nature] to check performance.
  - → choose a θ, generate data from $P(Y; \theta)$, see how often interval traps true value.
  - → get coverage for each choice of θ
  - → usually want uniform coverage, eg 95% coverage for all θ
    - ↳ can do for selection or prove mathematically for all.

Other topics & further reading:

Intervals from point estimates, intervals from tests

Linearisation & uncertainty propagation

Estimation vs testing? p-values & NHST?

Bayesian vs Frequentist ways?

Machine learning vs statistics?

Objective vs subjective Bayes?

Confidence intervals/sets for ill-posed inverse problems?

Simulation-based inference?

Causal inference?

Randomisation, experimental design, computer experiment design?

→ Will (maybe) put some on Canvas.
Can always ask me directly for suggestions!