A Non-Bayesian Interpretation of Bayesian Statistics
(and why estimation is an ill-posed problem)

Jonathan H. Manton

Department of Information Engineering
Research School of Information Sciences and Engineering
The Australian National University

Probability Conference, RSSS, ANU
2–4 March 2006

Outline

Opening Claims

- Estimating probabilities, estimating parameter values and confirming hypotheses are all ill-posed problems.

Opening Claims

- Estimating probabilities, estimating parameter values and confirming hypotheses are all ill-posed problems.
- This is analogous to a maths question which is not fully specified; depending on what extra assumptions people make (often unknowingly), different answers arise.

Opening Claims

- Estimating probabilities, estimating parameter values and confirming hypotheses are all ill-posed problems.
- This is analogous to a maths question which is not fully specified; depending on what extra assumptions people make (often unknowingly), different answers arise.
- We should not be asking these questions directly.

Opening Claims

- Estimating probabilities, estimating parameter values and confirming hypotheses are all ill-posed problems.
- This is analogous to a maths question which is not fully specified; depending on what extra assumptions people make (often unknowingly), different answers arise.
- We should not be asking these questions directly.
- Rather, every time, we should first work out the underlying reason why we want to estimate a parameter value, then work out how to solve this underlying problem directly.

Opening Claims

- Estimating probabilities, estimating parameter values and confirming hypotheses are all ill-posed problems.
- This is analogous to a maths question which is not fully specified; depending on what extra assumptions people make (often unknowingly), different answers arise.
- We should not be asking these questions directly.
- Rather, every time, we should first work out the underlying reason why we want to estimate a parameter value, then work out how to solve this underlying problem directly.
- By and large, the difficulty is not in the maths, but in how people map real world problems into mathematics, and how they interpret the results.

Parameter Estimation

- Assume we have a family of probability density functions $p(x;\theta)$ indexed by a parameter θ.

Parameter Estimation

- Assume we have a family of probability density functions $p(x;\theta)$ indexed by a parameter θ. (We stress the distinction with conditional probability $p(x|\theta)$.)

Parameter Estimation

- Assume we have a family of probability density functions $p(x;\theta)$ indexed by a parameter θ. (We stress the distinction with conditional probability $p(x|\theta)$.)
- Given an observation x we are asked to estimate θ.

Parameter Estimation

- Assume we have a family of probability density functions $p(x;\theta)$ indexed by a parameter θ. (We stress the distinction with conditional probability $p(x|\theta)$.)
- Given an observation x we are asked to estimate θ.
- This is an ill-posed problem! Yet people ask it all the time.

Parameter Estimation

- Assume we have a family of probability density functions $p(x;\theta)$ indexed by a parameter θ. (We stress the distinction with conditional probability $p(x|\theta)$.)
- Given an observation x we are asked to estimate θ.
- This is an ill-posed problem! Yet people ask it all the time. This does not make it right!

Parameter Estimation

- Assume we have a family of probability density functions $p(x;\theta)$ indexed by a parameter θ. (We stress the distinction with conditional probability $p(x|\theta)$.)
- Given an observation x we are asked to estimate θ.
- This is an ill-posed problem! Yet people ask it all the time. This does not make it right!
- Simple diagrams of pdf's show the difficulty.

Parameter Estimation

- Assume we have a family of probability density functions $p(x;\theta)$ indexed by a parameter θ. (We stress the distinction with conditional probability $p(x|\theta)$.)
- Given an observation x we are asked to estimate θ.
- This is an ill-posed problem! Yet people ask it all the time. This does not make it right!
- Simple diagrams of pdf's show the difficulty. Yet people have developed sophisticated estimation techniques.

Parameter Estimation

- Assume we have a family of probability density functions $p(x;\theta)$ indexed by a parameter θ. (We stress the distinction with conditional probability $p(x|\theta)$.)
- Given an observation x we are asked to estimate θ.
- This is an ill-posed problem! Yet people ask it all the time. This does not make it right!
- Simple diagrams of pdf's show the difficulty. Yet people have developed sophisticated estimation techniques.
- This false sense of security can come about in cases when there is a relatively large number of observations, in which case almost any "sensible estimator" will "work". (Consider very concentrated pdf's.)

James-Stein Estimator (there is no "best" estimator)

- Three-dimensional independent Gaussian r.v. $x \sim N(\mu, I)$.
- Only sensible estimator appears to be $\widehat{\mu} = x$.

James-Stein Estimator (there is no “best” estimator)

- Three-dimensional independent Gaussian r.v. $x \sim N(\mu, I)$.
- Only sensible estimator appears to be $\widehat{\mu} = x$.
- Assume we judge goodness by the risk function $E[\|\widehat{\mu} - \mu\|^2; \mu]$.
- Is $\widehat{\mu} = x$ the best estimator possible?

James-Stein Estimator (there is no "best" estimator)

- Three-dimensional independent Gaussian r.v. $x \sim N(\mu, I)$.
- Only sensible estimator appears to be $\widehat{\mu} = x$.
- Assume we judge goodness by the risk function $E[\|\widehat{\mu} - \mu\|^2; \mu]$.
- Is $\widehat{\mu} = x$ the best estimator possible?
- The most striking post-war result in statistics is that the James-Stein estimator

$$\widehat{\mu} = \left(1 - \frac{1}{\|x\|^2}\right) x$$

dominates the estimator $\widehat{\mu} = x$. (For all μ, it is better.)

James-Stein Estimator (there is no “best” estimator)

- Three-dimensional independent Gaussian r.v. $x \sim N(\mu, I)$.
- Only sensible estimator appears to be $\widehat{\mu} = x$.
- Assume we judge goodness by the risk function $E[\|\widehat{\mu} - \mu\|^2; \mu]$.
- Is $\widehat{\mu} = x$ the best estimator possible?
- The most striking post-war result in statistics is that the James-Stein estimator

$$\widehat{\mu} = \left(1 - \frac{1}{\|x\|^2}\right) x$$

  dominates the estimator $\widehat{\mu} = x$. (For all μ, it is better.)
- Initially counter-intuitive: Using the price of tea in China to influence our estimate of the temperature in Canberra!

Which Estimator is Best?

- The James-Stein Estimator is to be preferred if we wish to minimise our total squared error. A multi-national insurance company does not care about its individual loses, only its total lose.

Which Estimator is Best?

- The James-Stein Estimator is to be preferred if we wish to minimise our total squared error. A multi-national insurance company does not care about its individual loses, only its total lose.
- If individual loses are important, which they usually are, then $\widehat{\mu} = x$ should be used.

Which Estimator is Best?

- The James-Stein Estimator is to be preferred if we wish to minimise our total squared error. A multi-national insurance company does not care about its individual loses, only its total lose.
- If individual loses are important, which they usually are, then $\widehat{\mu} = x$ should be used.
- The best estimator is application dependent.

Which Estimator is Best?

- The James-Stein Estimator is to be preferred if we wish to minimise our total squared error. A multi-national insurance company does not care about its individual loses, only its total lose.
- If individual loses are important, which they usually are, then $\widehat{\mu} = x$ should be used.
- The best estimator is application dependent.
- In other words, parameter estimation is an ill-posed problem. Until we know the application, rarely can we claim one estimator is better than another.

There are No Estimation Problems...

- Claim: In this world, there are no (important / useful / naturally occurring) estimation problems.

There are No Estimation Problems...

- Claim: In this world, there are no (important / useful / naturally occurring) estimation problems.
- If we estimate the height of a tree for the sake of it, maybe we feel warm and fuzzy inside, but we would feel this way regardless of whether we are right or wrong.

There are No Estimation Problems...

- Claim: In this world, there are no (important / useful / naturally occurring) estimation problems.
- If we estimate the height of a tree for the sake of it, maybe we feel warm and fuzzy inside, but we would feel this way regardless of whether we are right or wrong.
- If we don't estimate something for the sake of it, then we will do something with the result. Invariably we will make a decision (about what to do).

There are No Estimation Problems...

- Claim: In this world, there are no (important / useful / naturally occurring) estimation problems.
- If we estimate the height of a tree for the sake of it, maybe we feel warm and fuzzy inside, but we would feel this way regardless of whether we are right or wrong.
- If we don't estimate something for the sake of it, then we will do something with the result. Invariably we will make a decision (about what to do).
- There are no estimation problems, only decision problems.
  - We are used to estimating temperature to work out what to wear tomorrow, but we are really only interested in what to wear tomorrow.

There are No Estimation Problems...

- Claim: In this world, there are no (important / useful / naturally occurring) estimation problems.
- If we estimate the height of a tree for the sake of it, maybe we feel warm and fuzzy inside, but we would feel this way regardless of whether we are right or wrong.
- If we don't estimate something for the sake of it, then we will do something with the result. Invariably we will make a decision (about what to do).
- There are no estimation problems, only decision problems.
    - We are used to estimating temperature to work out what to wear tomorrow, but we are really only interested in what to wear tomorrow.
    - We are used to debating the existence of God, but really we are interested in deciding how we should behave (go to church, worry about the ten commandments,...).

Bringing Back "Estimators"

- The optimal thing to do is to decide the best way of going from the observations to the decision directly.

Bringing Back “Estimators”

- The optimal thing to do is to decide the best way of going from the observations to the decision directly.
- Often this is impractical; imaging determining what to wear tomorrow from today’s global weather observations!

Bringing Back “Estimators”

- The optimal thing to do is to decide the best way of going from the observations to the decision directly.
- Often this is impractical; imaging determining what to wear tomorrow from today’s global weather observations!
- Hence it makes sense to compress the observations into an “approximate sufficient statistic” and then make (sub-optimal) decisions based on this more manageable quantity. (Can make multiple decisions based on a single approximate sufficient statistic.)

Bringing Back "Estimators"

- The optimal thing to do is to decide the best way of going from the observations to the decision directly.
- Often this is impractical; imaging determining what to wear tomorrow from today's global weather observations!
- Hence it makes sense to compress the observations into an "approximate sufficient statistic" and then make (sub-optimal) decisions based on this more manageable quantity. (Can make multiple decisions based on a single approximate sufficient statistic.)
- Moreover, sometimes the simple rule of "estimating" θ then acting as if the estimate were the true value works well enough. Doing so is a conscious choice though!

(True) Bayesian Statistics

- Assume both x and θ are random variables; joint pdf.

(True) Bayesian Statistics

- Assume both x and θ are random variables; joint pdf.
- Aside: In (advanced) probability theory, one works with expectation and conditional expectation in favour of conditional probability. Indeed, conditional probability is "not nice" but conditional expectation is.

(True) Bayesian Statistics

- Assume both x and θ are random variables; joint pdf.
- Aside: In (advanced) probability theory, one works with expectation and conditional expectation in favour of conditional probability. Indeed, conditional probability is "not nice" but conditional expectation is.
- Bayes rule (deducible from Kolmogorov's axioms): $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ if $p(x) > 0$.
- Given prior $p(\theta)$ and an observation x, can compute posterior $p(\theta|x)$.

(True) Bayesian Statistics

- Assume both x and θ are random variables; joint pdf.
- Aside: In (advanced) probability theory, one works with expectation and conditional expectation in favour of conditional probability. Indeed, conditional probability is “not nice” but conditional expectation is.
- Bayes rule (deducible from Kolmogorov’s axioms): $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ if $p(x) > 0$.
- Given prior $p(\theta)$ and an observation x, can compute posterior $p(\theta|x)$.
- Bayesians assert $p(\theta|x)$ contains all the knowledge there is to know about θ.

(True) Bayesian Statistics

- Assume both x and θ are random variables; joint pdf.
- Aside: In (advanced) probability theory, one works with expectation and conditional expectation in favour of conditional probability. Indeed, conditional probability is "not nice" but conditional expectation is.
- Bayes rule (deducible from Kolmogorov's axioms): $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ if $p(x) > 0$.
- Given prior $p(\theta)$ and an observation x, can compute posterior $p(\theta|x)$.
- Bayesians assert $p(\theta|x)$ contains all the knowledge there is to know about θ.
- e.g., Can compute MAP estimate $\widehat{\theta} = \arg\max_\theta p(\theta|x)$.

(True) Bayesian Statistics

- Assume both x and θ are random variables; joint pdf.
- Aside: In (advanced) probability theory, one works with expectation and conditional expectation in favour of conditional probability. Indeed, conditional probability is "not nice" but conditional expectation is.
- Bayes rule (deducible from Kolmogorov's axioms): $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ if $p(x) > 0$.
- Given prior $p(\theta)$ and an observation x, can compute posterior $p(\theta|x)$.
- Bayesians assert $p(\theta|x)$ contains all the knowledge there is to know about θ.
- e.g., Can compute MAP estimate $\widehat{\theta} = \arg\max_\theta p(\theta|x)$.
- Everything here is fine (except for thinking in terms of estimates - why should MAP be "best"?).

What are the Problems?

- In a nutshell, problems arise when people use Bayesian statistics when θ is not truly a random variable.

What are the Problems?

- In a nutshell, problems arise when people use Bayesian statistics when θ is not truly a random variable.
- There are many examples:
  - Witness sees yellow taxi. Tests show witness is 99% reliable. But in the town, there are 9999 white taxis and only 1 yellow taxi. So, Bayesian concludes Pr(yellow taxi) is 0.01; the taxi must have been white.

What are the Problems?

- In a nutshell, problems arise when people use Bayesian statistics when θ is not truly a random variable.
- There are many examples:
  - Witness sees yellow taxi. Tests show witness is 99% reliable. But in the town, there are 9999 white taxis and only 1 yellow taxi. So, Bayesian concludes Pr(yellow taxi) is 0.01; the taxi must have been white.
  - Test for disease is 99% reliable. Therefore, no point testing anyone for a rare disease!

What are the Problems?

- In a nutshell, problems arise when people use Bayesian statistics when θ is not truly a random variable.
- There are many examples:
  - Witness sees yellow taxi. Tests show witness is 99% reliable. But in the town, there are 9999 white taxis and only 1 yellow taxi. So, Bayesian concludes Pr(yellow taxi) is 0.01; the taxi must have been white.
  - Test for disease is 99% reliable. Therefore, no point testing anyone for a rare disease!
- Aside: If our jury system is only interested in locking up the criminals *on average*, then the Bayesian approach is fine! Same if we are only interested in curing people *on average*.

What are the Problems?

- In a nutshell, problems arise when people use Bayesian statistics when θ is not truly a random variable.
- There are many examples:
  - Witness sees yellow taxi. Tests show witness is 99% reliable. But in the town, there are 9999 white taxis and only 1 yellow taxi. So, Bayesian concludes Pr(yellow taxi) is 0.01; the taxi must have been white.
  - Test for disease is 99% reliable. Therefore, no point testing anyone for a rare disease!
- Aside: If our jury system is only interested in locking up the criminals *on average*, then the Bayesian approach is fine! Same if we are only interested in curing people *on average*.
- In general, the problem is not with Bayes' rule, it is with the incorrect mapping of the problem to mathematics, and the incorrect interpretation of the results.

Further Problems

- Students may approach a problem by deciding whether to take a Bayesian approach, and if so, working out what prior to assign to θ. (Both these steps I disagree with.)

Further Problems

- Students may approach a problem by deciding whether to take a Bayesian approach, and if so, working out what prior to assign to θ. (Both these steps I disagree with.)
- Example: Design a system to detect if an aeroplane is about to crash into a mountain.

Further Problems

- Students may approach a problem by deciding whether to take a Bayesian approach, and if so, working out what prior to assign to θ. (Both these steps I disagree with.)
- Example: Design a system to detect if an aeroplane is about to crash into a mountain.
- First we estimate the height of the plane, then we ask if this is less than or equal to the height of the mountain.

Further Problems

- Students may approach a problem by deciding whether to take a Bayesian approach, and if so, working out what prior to assign to θ. (Both these steps I disagree with.)
- Example: Design a system to detect if an aeroplane is about to crash into a mountain.
- First we estimate the height of the plane, then we ask if this is less than or equal to the height of the mountain.
- To estimate the height, we first need a prior. Aeroplanes typically fly at 30,000 feet, so choose a normal distribution about this height for the prior.

Further Problems

- Students may approach a problem by deciding whether to take a Bayesian approach, and if so, working out what prior to assign to θ. (Both these steps I disagree with.)
- Example: Design a system to detect if an aeroplane is about to crash into a mountain.
- First we estimate the height of the plane, then we ask if this is less than or equal to the height of the mountain.
- To estimate the height, we first need a prior. Aeroplanes typically fly at 30,000 feet, so choose a normal distribution about this height for the prior.
- The resulting MAP is certainly a "good estimator" for some problems, but not for ours. A plane flying at a height just below that of the mountain will appear to be ok since the prior will distort the estimate upwards.

Babies and Bathwater

- What is good about Bayesian statistics?

Babies and Bathwater

- What is good about Bayesian statistics?
- First, let's forget about any meaning attributed to the formula, and think in terms of a mathematical rule going from the observation x to a function $p(\theta; x)$.

Babies and Bathwater

- What is good about Bayesian statistics?
- First, let's forget about any meaning attributed to the formula, and think in terms of a mathematical rule going from the observation x to a function $p(\theta; x)$.
- Historically, Bayesian statistics have had a resurgence twice.
  - As a way of finding admissible estimators. The MAP estimator is always admissible.
  - Computationally, the recursive form of Bayes (update) rule is suited for efficient implementations of estimation problems over time (e.g., tracking the location of an aeroplane).

Babies and Bathwater

- What is good about Bayesian statistics?
- First, let's forget about any meaning attributed to the formula, and think in terms of a mathematical rule going from the observation x to a function $p(\theta; x)$.
- Historically, Bayesian statistics have had a resurgence twice.
  - As a way of finding admissible estimators. The MAP estimator is always admissible.
  - Computationally, the recursive form of Bayes (update) rule is suited for efficient implementations of estimation problems over time (e.g., tracking the location of an aeroplane).
- Therefore, we want to keep Bayesian estimation as an option, but interpret it differently, so it gets used correctly in practice.

Weighting Function Statistics

- Assume we make a conscious decision that we want to compute an approximate sufficient statistic $\widehat{\theta}$ of θ.

Weighting Function Statistics

- Assume we make a conscious decision that we want to compute an approximate sufficient statistic $\widehat{\theta}$ of θ.
- We know that in general, no function $\widehat{\theta}(x)$ will be “good” at being close to θ for all true parameter values of θ.

Weighting Function Statistics

- Assume we make a conscious decision that we want to compute an approximate sufficient statistic $\widehat{\theta}$ of θ.
- We know that in general, no function $\widehat{\theta}(x)$ will be "good" at being close to θ for all true parameter values of θ.
- Therefore, we introduce a weighting function $w(\theta)$ which is peaked in regions where we want $\widehat{\theta}$ to be close to θ. (e.g. For aeroplane example, choose $w(\theta)$ to be peaked at the height of mountain.)

Weighting Function Statistics

- Assume we make a conscious decision that we want to compute an approximate sufficient statistic $\widehat{\theta}$ of θ.
- We know that in general, no function $\widehat{\theta}(x)$ will be "good" at being close to θ for all true parameter values of θ.
- Therefore, we introduce a weighting function $w(\theta)$ which is peaked in regions where we want $\widehat{\theta}$ to be close to θ. (e.g. For aeroplane example, choose $w(\theta)$ to be peaked at the height of mountain.)
- We compute the function $p(\theta; x) = \frac{p(x;\theta)w(\theta)}{\int p(x;\theta)w(\theta)\,d\theta}$.

Weighting Function Statistics

- Assume we make a conscious decision that we want to compute an approximate sufficient statistic $\widehat{\theta}$ of θ.
- We know that in general, no function $\widehat{\theta}(x)$ will be "good" at being close to θ for all true parameter values of θ.
- Therefore, we introduce a weighting function $w(\theta)$ which is peaked in regions where we want $\widehat{\theta}$ to be close to θ. (e.g. For aeroplane example, choose $w(\theta)$ to be peaked at the height of mountain.)
- We compute the function $p(\theta; x) = \frac{p(x;\theta)w(\theta)}{\int p(x;\theta)w(\theta)\,d\theta}$.
- We may decide to set $\widehat{\theta}(x) = \arg\max_\theta p(\theta; x)$.

Weighting Function Statistics

- Assume we make a conscious decision that we want to compute an approximate sufficient statistic $\widehat{\theta}$ of θ.
- We know that in general, no function $\widehat{\theta}(x)$ will be "good" at being close to θ for all true parameter values of θ.
- Therefore, we introduce a weighting function $w(\theta)$ which is peaked in regions where we want $\widehat{\theta}$ to be close to θ. (e.g. For aeroplane example, choose $w(\theta)$ to be peaked at the height of mountain.)
- We compute the function $p(\theta; x) = \frac{p(x;\theta)w(\theta)}{\int p(x;\theta)w(\theta)\,d\theta}$.
- We may decide to set $\widehat{\theta}(x) = \arg\max_\theta p(\theta; x)$.
- We do not claim $p(\theta; x)$ contains all the knowledge there is about θ. (Different $w(\theta)$ will bring out different knowledge about θ.)
- (There is likely a connection with imprecise probabilities.)

When faced with a parameter estimation problem...

- Determine the underlying decision problem.
- Decide if computing a low dimensional approximate sufficient statistic might be accurate enough (does it capture enough information to enable a good decision to be made?).
- Decide if the approximate sufficient statistic can be of the form $\widehat{\theta}$, where $\widehat{\theta}$ will be chosen to be "close to" θ on "average". (In other words, can good decisions be made by "estimating" θ and acting as if this estimate is correct.)
- Choose a weighting function $w(\theta)$ based on the underlying decision problem (what are critical values of θ where a "good estimate" is required?).
- See if the resulting decision rule makes sufficiently good decisions.

Summary

- Bayesian statistics is only applicable if θ is truly a random variable.
- The words we use to describe things with matter; they influence how we map real world problems into mathematics, which is where the difficulties lie.
- We introduce the concept of weighting function statistics, which looks identical to Bayesian statistics, but has a very different interpretation, hence hopefully it will be used differently (correctly!) in practice.
- We emphasise there is no estimation problem, only decision problems.
- What used to be called estimators should be called approximate sufficient statistics, and used accordingly, to simplify the decision problem at the expense of optimality.