
Mathematical Institute

# Paradoxes in data-driven robust valuation

Samuel N. Cohen
*Mathematical Institute*
*University of Oxford*

*Research Supported by:*
The Oxford–Man Institute for Quantitative Finance
The Oxford–Nie Financial Big Data Laboratory

Oxford Mathematics

# An Ellsberg-type problem

If I wanted to gamble, I'd buy the casino — J.P. Getty, Snr



- Suppose we are going to bet on the toss of a (possibly unfair) coin
- Heads you get £1, Tails you get nothing.
- I supply the coin, but you pay to play.

# An Ellsberg-type problem

If I wanted to gamble, I'd buy the casino — J.P. Getty, Snr



- Suppose we are going to bet on the toss of a (possibly unfair) coin
- Heads you get £1, Tails you get nothing.
- I supply the coin, but you pay to play.
- Given I am untrustworthy, we will throw the coin $N$ times before we play.
- What is the maximum you are willing to pay? (or equivalently, what is the value of the game?)

For a classical frequentist, the solution is straightforward:

- There is only one (reasonable) way to estimate $p = P(H)$ from observations $\{X_i\}_{i=1}^N$
- $\hat{p} = \bar{X}_N$, where $\bar{X}_N = \sum_{i=1}^N X_i/N$, this has sampling variance $p(1-p)/N$.
- For a loss function $\phi$, wlog assume $\phi(0) = 0$, then you then calculate

$$\hat{E}[\phi(X)] = \hat{p}\phi(1).$$

- This is the estimated expected loss, and allows us to calculate prices.

For a classical frequentist, the solution is straightforward:

- There is only one (reasonable) way to estimate $p = P(H)$ from observations $\{X_i\}_{i=1}^N$
- $\hat{p} = \bar{X}_N$, where $\bar{X}_N = \sum_{i=1}^N X_i/N$, this has sampling variance $p(1-p)/N$.
- For a loss function $\phi$, wlog assume $\phi(0) = 0$, then you then calculate
$$\hat{E}[\phi(X)] = \hat{p}\phi(1).$$
- This is the estimated expected loss, and allows us to calculate prices.
- Strange, the variance doesn't appear...

We now have a problem: Suppose we could choose between two coins:

- The first we throw $N_1 = 3$ times, and observe 2 heads
- The second we throw $N_2 = 3000$ times, and observe 2000 heads.
- Which coin do you prefer?

We now have a problem: Suppose we could choose between two coins:

- The first we throw $N_1 = 3$ times, and observe 2 heads
- The second we throw $N_2 = 3000$ times, and observe 2000 heads.
- Which coin do you prefer?
- Everyone prefers the second. Most people still prefer the second if we only observe 1999 heads.
- This is inconsistent with our estimated expected loss rule!'

This is not a new observation:

*For in deciding on a course of action, it seems plausible to suppose that we ought to take account of the weight as well as the probability of different expectations.*
—J.M. Keynes, A Treatise on Probability, 1921

*The theoretical difference between the probability connected with an estimate and that involved in such phenomena as are dealt with by insurance is, however, of the greatest importance, and is clearly discernible in nearly any instance of the exercise of judgement.*
— F. Knight, Risk, Uncertainty and Profit, 1921

# Bayesian solution

Justice weighs out learning to those who suffer — Aeschylus

Mathematical Institute

The problem could be that $p$ is not part of our probabilistic framework, so let's be Bayesian...

- Let $\mathcal{F}_N$ denote the $\sigma$-algebra generated by the observations.
- Then we calculate

$$E[\phi(X)|\mathcal{F}_N] = E\Big[E[\phi(X)|p, \mathcal{F}_N]\Big|\mathcal{F}_N\Big]$$
$$= E[p\phi(1)|\mathcal{F}_N] = E[p|\mathcal{F}_N]\phi(1).$$

# Bayesian solution

Justice weighs out learning to those who suffer — Aeschylus

Mathematical Institute

The problem could be that $p$ is not part of our probabilistic framework, so let's be Bayesian...

- Let $\mathcal{F}_N$ denote the $\sigma$-algebra generated by the observations.
- Then we calculate

$$E[\phi(X)|\mathcal{F}_N] = E\Big[E[\phi(X)|p,\mathcal{F}_N]\Big|\mathcal{F}_N\Big]$$
$$= E[p\phi(1)|\mathcal{F}_N] = E[p|\mathcal{F}_N]\phi(1).$$

- Strange, the posterior variance still isn't there
- True for any rule based only on the posterior law of $X$
- We still don't value accurate knowledge of $p$.

# Bayesian solution

Justice weighs out learning to those who suffer — Aeschylus

Mathematical Institute

What's going on here?

- Expected loss can't see our uncertainty in parameters
- Bayesian methods feel like they incorporate this but....

# Bayesian solution

Justice weighs out learning to those who suffer — Aeschylus

Mathematical Institute

What's going on here?

- Expected loss can't see our uncertainty in parameters
- Bayesian methods feel like they incorporate this but....
- The outcome $X$ is Bernoulli, which is a one parameter distribution family.
- Bayesian methods only replace distributions with mixtures, but these are still just Bernoulli.
- The setting is too simple for Bayesian methods to (even appear to) work.

# Bayesian solution

Justice weighs out learning to those who suffer — Aeschylus

UNIVERSITY OF OXFORD
Mathematical Institute

What's going on here?

- Expected loss can't see our uncertainty in parameters
- Bayesian methods feel like they incorporate this but....
- The outcome $X$ is Bernoulli, which is a one parameter distribution family.
- Bayesian methods only replace distributions with mixtures, but these are still just Bernoulli.
- The setting is too simple for Bayesian methods to (even appear to) work.

If Bayesian methods can't handle a single coin toss, why would we expect them to properly handle something complicated?

# Multiple prior models

If I take refuge in ambiguity, I assure you that it's quite conscious. — Kingman Brewster, Jr.



One approach to these problems, in economics, is 'multiple prior models'/'Bayes Minimax estimation'

- Key modern names: Gilboa–Schmeidler / Fagin / Jaffray / Dempster–Schafer
- The only user-input into the Bayesian approach is the prior
- Perhaps we should be 'robust' to varying that input.
- Given a family of priors, we look at the maximum expectation in the family.
- Let's consider this with our simple problem.

# Multiple prior models

If I take refuge in ambiguity, I assure you that it's quite conscious. — Kingman Brewster, Jr.



- Consider a Beta prior for $p$
- Let's parameterize our Betas $B(\alpha, \beta)$ by the mean $\mu = \frac{\alpha}{\alpha+\beta}$ and 'precision'/'equivalent sample size' $n = \alpha + \beta$.
- A $B(\mu, n)$ has mean $\mu$, variance $\mu(1-\mu)/(n+1)$.

# Multiple prior models

If I take refuge in ambiguity, I assure you that it's quite conscious. — Kingman Brewster, Jr.



- Consider a Beta prior for $p$
- Let's parameterize our Betas $B(\alpha, \beta)$ by the mean $\mu = \frac{\alpha}{\alpha+\beta}$ and 'precision'/'equivalent sample size' $n = \alpha + \beta$.
- A $B(\mu, n)$ has mean $\mu$, variance $\mu(1-\mu)/(n+1)$.
- It's an undergrad exercise to show that with prior $B(\mu_0, n_0)$ the posterior is

$$B\left(\frac{n_0}{N+n_0}\mu_0 + \frac{N}{N+n_0}\bar{X}_N, n_0 + N\right)$$

- As usual, the credible interval for $p$ (and hence $E[\phi(X)]$) has width $O(N^{-1/2})$.

# Multiple prior models

If I take refuge in ambiguity, I assure you that it's quite conscious. — Kingman Brewster, Jr.



- Let's vary the prior mean within an interval.
- We fix $n_0$, but consider priors with $\mu_0 \in [\mu_*, \mu^*]$
- The corresponding family of posteriors have precisions $N + n_0$ and means

$$\mu \in \frac{N}{n_0 + N}\bar{X}_N + \frac{n_0}{N + n_0}\Big[\mu_*, \mu^*\Big].$$

- This is an interval of width $O(N^{-1})$.
- The range of posteriors collapses much quicker than our statistical uncertainty.

# Multiple prior models

If I take refuge in ambiguity, I assure you that it's quite conscious. — Kingman Brewster, Jr.

Mathematical Institute

- Let's vary the prior mean within an interval.
- We fix $n_0$, but consider priors with $\mu_0 \in [\mu_*, \mu^*]$
- The corresponding family of posteriors have precisions $N + n_0$ and means

$$\mu \in \frac{N}{n_0 + N}\bar{X}_N + \frac{n_0}{N + n_0}\left[\mu_*, \mu^*\right].$$

- This is an interval of width $O(N^{-1})$.
- The range of posteriors collapses much quicker than our statistical uncertainty.

Simply using multiple priors does not represent our uncertainty.

# Multiple prior models

If I take refuge in ambiguity, I assure you that it's quite conscious. — Kingman Brewster, Jr.

Mathematical Institute

- This scaling is completely typical.
- We usually have estimation error of $O(N^{-1/2})$.
- For exchangeable data, the posterior log-density is

$$\log(f_{\text{post.}}(\theta)) = N\Big(\text{'average' log-likelihood}\Big) + \log f_{\text{prior}}(\theta)$$

  so the impact of the prior is $O(N^{-1})$ relative to observations.
- In the uncertain-prior approach, we are not *learning* our model, but simply conditioning using Bayes' rule.
- The problem is not the posterior distribution of the parameters, but the assumption that our expectations depend only on the posterior law of the future outcomes.

# Alternative approaches

If I take refuge in ambiguity, I assure you that it's quite conscious. — Kingman Brewster, Jr.



- One approach which allows for learning our model is given by confidence intervals
- Assuming a central limit theorem or likelihood approach, these are special cases of the 'DR-expectation'
- For iid observations, under appropriate continuity assumptions, these expectations have the property that

$$\mathcal{E}(\xi) \approx \hat{E}[\xi] + k\left(\mathrm{Var}(\hat{E}[\xi])\right)^{k'/(2k'-1)} + o(N^{-k'/(2k'-1)})$$

  for chosen parameters $k, k'$.
- Taking $k' \to \infty$, we have confidence intervals.

# Including learning increases uncertainty

It is the worst of madness to learn what has to be unlearnt. — Erasmus



## Paradox

*If we ignore trying to learn our model from data, and simply place an uncertain prior over our parameters and use Bayesian methods to condition on observations, we underestimate our uncertainty.*

- We now leave this problem, and look at confidence intervals specifically, in a simple estimation context.
- Our setting is simple, but surprisingly generic.
- Suppose we have iid observations

$$X_n \sim N(\theta, I_d)$$

  where $\theta, X_n \in \mathbb{R}^d$, and $I_d$ is the $d \times d$ identity matrix.
- We are interested in giving a value to $Y$, using observations $\{X_n\}_{n=1}^N$ where we know $E[Y] = \theta^1$.
- For the sake of robustness, we will use the upper/lower bounds of a confidence interval for $Y$

Approach 1:

- As $X_n^i$ and $X_n^j$ are jointly Gaussian and uncorrelated, they are independent.
- To estimate $\theta^1 = E[Y]$, we ignore observations $X^j$ for $j \neq 1$.
- We have the classical estimate of $\theta^1$

$$\bar{X}_N^1 = N^{-1} \sum X_n^1,$$

and the confidence interval

$$\bar{X}_N^1 \pm \frac{z_\alpha}{\sqrt{N}}$$

where $z_\alpha = F_{N(0,1)}^{-1}(1-\alpha) = \sqrt{F_{\chi_1^2}^{-1}(1-\alpha)}$

- This approach has an issue – it is very specific to estimating $E[Y] = \theta^1$.
- We have no way to compare $Y$ with $Y'$, where (for example) $E[Y'] = \sum c_k \theta^k$.
- In order to make comparisons, we need to work with all the components of $\theta$ simultaneously.

Approach 2:

- We can estimate $\theta$ by $\bar{X}_N = N^{-1} \sum X_n \in \mathbb{R}^d$
- The usual confidence region is then the ellipse

$$\Theta = \left\{ \theta : N\|\theta - \bar{X}_N\|^2 \leq F^{-1}_{\chi^2_d}(1-\alpha) \right\}$$

- A basic asymptotic gives, for large $d$

$$F^{-1}_{\chi^2_d}(1-\alpha) \approx d + \sqrt{2d} z_\alpha + O(1).$$

- The upper/lower bounds on $E[Y]$ are then

$$\sup_{\theta \in \Theta} / \inf_{\theta \in \Theta} E_\theta[Y] = \bar{X}^1_N \pm \sqrt{\frac{d}{N}} + O(d/N)$$

# Irrelevant information is not irrelevant

What is important is to spread confusion, not eliminate it – S. Dali



## Paradox

*Dimensions of the parameter space which are known to be irrelevant to our outcome have an impact on our estimation when using a confidence region.*

UNIVERSITY OF OXFORD
Mathematical Institute

- For large samples, the DR-expectation asymptotics rely on

$$\mathrm{Var}(\hat{E}[Y]) \approx (\partial_\theta E_\theta[Y])^\top \mathrm{Var}(\hat{\theta})(\partial_\theta E_\theta[Y])$$

  in particular, only the first order sensitivity of $E_\theta[Y]$ is important.
- We define a *confident projection region* to be a random set $\tilde{\Theta} \subset \mathbb{R}^d$ such that the linear projection $\pi^\top \tilde{\Theta}$ is a confidence interval for the parameter $\pi^\top \tilde{\theta}$, for all $\pi \in \mathbb{R}^d$.
- In our setting

$$\tilde{\Theta} = \left\{ \theta : N\|\theta - \bar{X}_N\|^2 \leq F^{-1}_{\chi^2_1}(1-\alpha) = z_\alpha \right\}$$

- Importantly, $\tilde{\Theta}$ does not grow with $d$.

- Confident projections have the property
  - For each expectation, we have a convex expectation, with interpretation as a confidence interval.
  - The *uniform* confidence level of this expectation (in terms of parameter values) is low, even if the individual levels (in terms of scalar expectations) is high.
- Generally, this will lead to far less conservative decision making.

- The above problem is generic – it's the dimension of $\Theta$ which is important, not the observation dimension.
- Another approach is to use a nonparametric confidence region, for example using Wasserstein distance.
- Let's suppose we have iid observations $X_n$ valued in $[0,1]$, generating an empirical measure $\hat{\mu}$.
- We take a ball around $\hat{\mu}$ in $W_1$, and associate this with a confidence region.
- This generates robust estimates for $Y = f(X)$, whenever $f$ is $\mathrm{Lip}_1$.

- Using the bound on $W_1$ obtained by Kloeckner, $E[W_1(\hat{\mu}_n, \mu)] \leq \frac{1+\sqrt{2}}{2\sqrt{n}}$
- Approximating using Markov's inequality, we obtain the $1-\alpha$ confidence bound

$$N^{-1}\sum f(X_n) \pm \frac{E[W_1]}{\alpha} = N^{-1}\sum f(X_n) \pm \frac{1+\sqrt{2}}{2\alpha\sqrt{n}}$$

- A direct CI, for a specific $f$, using the CLT is

$$N^{-1}\sum f(X_n) \pm z_\alpha \frac{\mathrm{sd}(f(X_n))}{\sqrt{n}}$$

and as $f$ is $\mathrm{Lip}_1$, we know $\mathrm{sd}(f(X_n)) \leq 1/2$ (Popoviciu).

- As $\alpha \to 0$, the CLT bound is much better. (e.g., $\pm 2/\sqrt{n}$ vs $\pm 24/\sqrt{n}$ for a 95% interval)
- The Wasserstein bound is uniform over $f$ (within $\mathrm{Lip}_1$), while the direct CI is not.
- If we wish to compare multiple $f$'s we have to start using multivariate confidence regions again.

## Paradox

*The Wasserstein norm gives tighter intervals than the CLT, provided the dimension of the parameter space (or number of variables, in a nonparametric setting) is sufficiently high. For 95% intervals, the critical dimension is at most* 2221, *provided at least one of our outcomes under consideration is maximally spread out.*

This is based on $z_{0.05} = \sqrt{F^{-1}_{\chi^2_{2221}}(0.95)} > (1+\sqrt{2})/0.05$.

Numerical experiments suggest *far* tighter bounds are possible, but I am unaware of better *explicit* estimates.

Other issues are possible to observe, particularly when working through time

**Paradox**

*The sequence of confidence intervals generally shrinks, but almost every observation will cause it to include models that were previously excluded.*

**Paradox**

*With probability one, the current best estimate of a parameter will leave the future confidence interval. This typically happens at a logarithmic frequency in time.*

Both of these facts make 'pasting' approaches to time-consistency difficult. Some approaches use recursive constructions, but there is no measure fixed, so the CI is not a Markov process.

## Paradox

- *Statistics cares a lot about null sets – we find methods which converge (a.s. or in probability) to the correct parameter value except on null sets.*
  - *By considering all the parameters (and their associated distributions), we have a non-dominated family of measures (over infinite horizon, equivalent on finite horizon).*
  - *We make decisions based on excluding models with sufficiently small likelihood for our observations.*
- *Suppose we look at an uncertain volatility framework with learning...*

## Paradox

- *Statistics cares a lot about null sets – we find methods which converge (a.s. or in probability) to the correct parameter value except on null sets.*
  - *By considering all the parameters (and their associated distributions), we have a non-dominated family of measures (over infinite horizon, equivalent on finite horizon).*
  - *We make decisions based on excluding models with sufficiently small likelihood for our observations.*
- *Suppose we look at an uncertain volatility framework with learning...*
  - *Then we want to keep a family of non-dominated measures, but have no classical likelihood in continuous time.*
  - *Inference becomes difficult conceptually*

- Further issues arise when looking at decision making, through time: to obtain time-consistency, one often needs to include an uncertainty premium to the level of the uncertainty premium that one will have in the future (and so on to higher degrees).
- Moving away from iid data (e.g. to filtering approaches) provides further challenges – what does one assume is fixed, and what is learnt from observations? What is filtered/conditioned/Bayesian and what is robust?
- New methods, based on but different from statistical ideas, will need to be developed to understand robust decision making based on data.