Whaddya know? Bayesian and Frequentist approaches to inverse problems

Philip B. Stark

Department of Statistics
University of California, Berkeley

Inverse Problems: Practical Applications and Advanced Analysis
Schlumberger WesternGeco
Houston, TX
12–15 November 2012

Abstract

The difference between Bayesians and Frequentists is what they presume to know about how the data were generated. Frequentists generally presume to know that the data come from a *statistical model*. A statistical model says what the probability distribution of the data would be, as a function of one or more unknown *parameters*. Frequentists might also presume to know that the parameters satisfy some *constraints*.

Bayesians claim that and more: that the actual values of the parameters are selected at random from a set of possible values of those parameters, and that they know the probability distribution used to make that random selection.

Unsurprisingly, the difference in assumptions leads to different conclusions. It also leads to different interpretations and definitions of what a “good” estimator is. Frequentist criteria generally quantify error over repetitions of the experiment, keeping the parameters constant. Bayesian criteria generally quantify error over repetitions of selecting the parameters, keeping the observations constant.

Acknowledgments:

Much cribbed from joint work with Luis Tenorio.
See:

Stark, P.B. and L. Tenorio, 2010. A Primer of Frequentist and Bayesian Inference in Inverse Problems. In *Large Scale Inverse Problems and Quantification of Uncertainty*, Biegler, L., G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders and K. Willcox, eds. John Wiley and Sons, NY.

What's a Bayesian?

- Frequentists and Bayesians agree as matter of math that

$$P(A|B) = P(B|A)P(A)/P(B).$$

- The problem arises when A is an hypothesis.
  Then $P(A)$ is a *prior*.
- Bayesians have no trouble with priors, e.g.,
  "the probability that the hypothesis is true is p."
  ("the probability that there's oil in the formation is 10%.")
- To frequentists, doesn't make sense:
  Hypothesis is true or not—its truth is not random.
  Ignorance cannot always be expressed as a probability, and hence not as a prior.
- Frequentist focus: evidence.
  Bayesian focus: degree of belief given evidence.

What's $P(A)$? Several standard interpretations.

- Equally likely outcomes
- Frequency theory
- Subjective (Bayesian) theory
- Model-based

Arguments for Bayesianism

- Descriptive (mountains of evidence to the contrary)
- Normative (maybe, if you had a prior)
- Appeals to betting
  (Dutch book. If you had to bet, would you be Bayesian?)
- To capture constraints (frequentists can, too)
- “It doesn’t matter, if you have enough data”
  (depends: not true for improper priors or inverse problems)
- “Makes prejudices explicit” (does it?)
- Always gives an answer–with muy macho computations
- The error bars are smaller than for Frequentist methods

How does “probability” enter an inverse problem?

- Physics: quantum, thermo
- Deliberate randomization (experiments)
- Measurement error: real or an idealization
- Convenient approximate description (stochastic model): random media, earthquakes
- Priors

“Probability” means quite different things in these cases.

A partial taxonomy of problems

- Prediction or inference?
- “Philosophical” or practical?
- Falsifiable/testable on relevant timescale?
- Repeatable?

Earth’s core: inference, philosophical, not falsifiable, not repeatable.

Oil exploration: prediction, practical, falsifiable, repeatable:
If method makes more money than anything else in the toolkit, use it.
(But check whether it actually does.)

When are Bayesian Methods Justified?

- You really have a prior. Still:
  - Little reason anyone but you should be persuaded.
  - I've never met anyone with a prior, only people who use Bayesian computations.
  - Priors generally chosen for convenience or habit, not because anyone deeply believes they are true.
  - Mountains of evidence that nobody is Bayesian in daily life: biology
- Prediction, practical, falsifiable on a reasonable time scale, e.g., targeted marketing, predicting server load, aiming anti-aircraft guns.
  - Can test approach against data, repeatedly.
  - Still, error bars still don't mean much.
  - I feel the same about tea leaves and chicken bones.
- Checking theoretical properties of frequentist methods.
  - If an estimate is not Bayes for *some* prior (and the loss in question), it is *inadmissible*; might want a different method.

Bayesian uncertainties not to be trusted in high-stakes decisions, if less than frequentist.

Frequentist uncertainties based on invented model not to be trusted, either.

Does God play Dice with the Universe?

- Einstein: “no.”
- Bayesians: “yes.”
- PBS: “what’s the potential downside to acting as if she does if she doesn’t?”

Can compare performance of frequentist and Bayes estimators, both with and without a prior.

Can evaluate Bayes estimator from frequentist perspective and vice versa.

Comparing minimax risk to Bayes risk measures how much information the prior adds.

Terminology

- State of the world, θ: mathematical representation of physical property, e.g., seismic velocity.
- Θ: set of possible states of the world, incorporating any prior constraints.
  Know that $\theta \in \Theta$.
- Parameter $\lambda = \lambda[\theta]$ is a property of θ.
- Data Y take values in the sample space $\mathcal{Y}$.
- Statistical model gives distribution of Y for any $\theta \in \Theta$.
  If $\theta = \eta$, then $Y \sim \mathbf{P}_\eta$.
- Density of $\mathbb{P}_\eta$ (with respect to dominating measure μ) at y:

$$p_\eta(y) \equiv d\mathbf{P}_\eta / d\mu|_y$$

- Likelihood of η given $Y = y$ is $p_\eta(y)$ as a function of η, y fixed.
- Estimators: quantities that can be computed from the data Y without knowing the state of the world.

Priors and Posteriors

- Prior π is a probability distribution on Θ
- Assume $p_\eta(y)$ is jointly measurable wrt η, y.
  Then π and $\mathbf{P}_\eta$ determine the joint distribution of θ and Y.
  Marginal distribution of Y has density

$$m(y) = \int_\Theta p_\eta(y)\, \pi(d\eta).$$

- Posterior distribution of θ given $Y = y$:

$$\pi(d\eta | Y = y) = \frac{p_\eta(y)\ \pi(d\eta)}{m(y)}.$$

- marginal posterior distribution of $\lambda[\theta]$ given $Y = y$,
  $\pi_\lambda(d\ell | Y = y)$ defined by

$$\int_\Lambda \pi_\lambda(d\ell | Y = y) \equiv \int_{\eta:\lambda[\eta]\in\Lambda} \pi(d\eta | Y = y)$$

Bounded Normal Mean

- $\theta \in \Theta \equiv [-\tau, \tau]$
- $Y = \theta + Z$, $Z \sim N(0, 1)$
- density of $\mathbf{P}_\theta$ at y is

$$\phi_\theta(y) \equiv \frac{1}{\sqrt{2\pi}} e^{-(y-\theta)^2/2}$$

- likelihood of η given $Y = y$ is $\phi_\eta(y)$ as a function of η, y fixed.

Priors for the Bounded Normal Mean

- π needs to assign probability 1 to the interval $[-\tau, \tau]$.
- Every choice of π has more information than the constraint $\theta \in \Theta$, because it specifies chance that θ is in each subset of Θ.
- One choice: "uninformative" uniform distribution on $[-\tau, \tau]$, with density

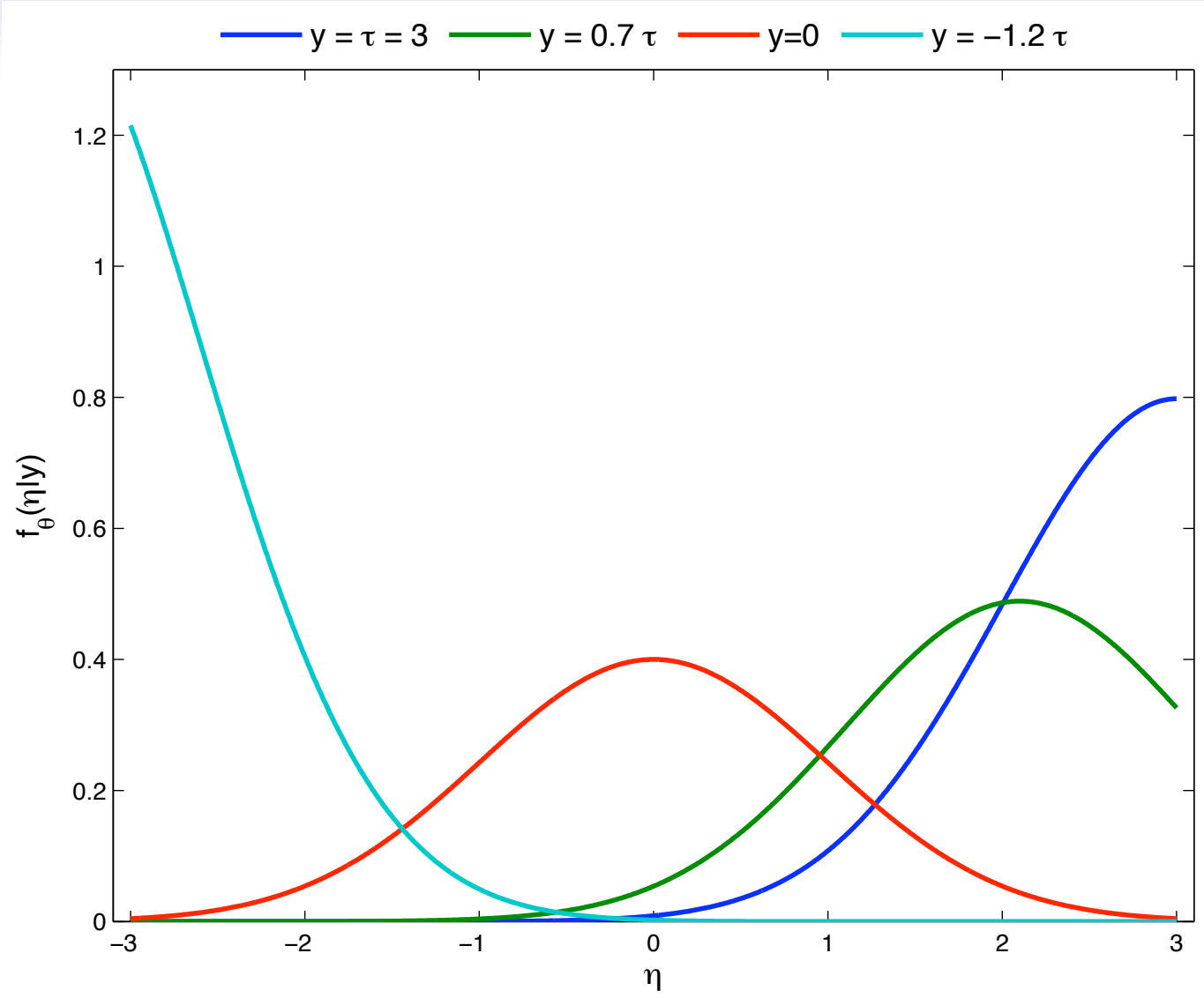
$$U_\tau(\eta) \equiv \frac{1}{2\tau} 1_{[-\tau,\tau]}(\eta).$$

- Density of predictive distribution of Y is

$$m(y) = \frac{1}{2\tau} \int_{-\tau}^{\tau} \phi_\eta(y) d\eta = \frac{1}{2\tau} (\Phi(y+\tau) - \Phi(y-\tau)).$$

- Posterior distribution of θ given $Y = y$ is

$$\pi(d\eta | Y = y) = \frac{\phi_\eta(y) \frac{1}{2\tau} 1_{\eta \in [-\tau,\tau]}(\eta)}{m(y)} = \frac{\phi_\eta(y) 1_{\eta \in [-\tau,\tau]}(\eta)}{\Phi(y+\tau) - \Phi(y-\tau)}.$$



Posterior densities of θ for a bounded normal mean with a uniform prior on the interval $[-3, 3]$ for four values of y.

Comparing estimators

- Risk function: way to compare estimators: expected "cost" of using a particular estimator when the world is in a given state.
- Choice of risk function should be informed by scientific goal, but often chosen for convenience.
- Generally, no estimator has the smallest risk for all $\theta \in \Theta$.
- Bayesian and frequentist methods trade off risks for different θ differently.

Decision Theory

- Two-player game: Nature versus analyst.
- Frequentist and Bayesian games differ.
- According to both, Nature and the analyst know Θ, $\mathbf{P}_\eta$ for all $\eta \in \Theta$, λ, and the payoff rule $L(\ell, \lambda[\eta])$
- Nature selects θ from Θ.
- Analyst selects estimator $\widehat{\lambda}$.
- Analyst does not know θ and Nature does not know $\widehat{\lambda}$.
- Data Y generated using the value of θ that Nature selected, plugged into $\widehat{\lambda}$, and $L(\widehat{\lambda}(Y), \lambda[\theta])$ is found.
- With θ fixed, a new value of Y is generated, and $L(\widehat{\lambda}(Y), \lambda[\theta])$ is calculated again; repeat.
- Analyst pays average value of $L(\widehat{\lambda}(Y), \lambda[\theta])$, *risk of* $\widehat{\lambda}$ *at* θ, denoted $\rho_\theta(\widehat{\lambda}, \lambda[\theta])$.

Bayesian Rules v Frequentist Rules

- Bayesian version: Nature selects θ at random according to the prior distribution π, and the analyst knows π.
- Frequentist version: analyst does not know how Nature will select θ from Θ.
- Essential difference between the frequentist and Bayesian viewpoints: Bayesians claim to know more about how Nature generates the data.
- Cautious frequentist might select $\widehat{\lambda}$ to minimize her worst-case risk, on the assumption that Nature might play intelligently to win as much as possible. Minimax estimator.
- Bayesian would select estimator that minimizes the average risk on the assumption that Nature selects θ at random from π. Bayes estimator.

Duality between Bayes Risk and Minimax Risk

- Bayes risk depends on Θ, $\{\mathbf{P}_\eta : \eta \in \Theta\}$, λ, L, and π.
- Consider allowing π to vary over a rich set of possible priors.
- Prior for which Bayes risk is largest is *least favorable*.
- Typically not "uninformative" prior.
- Under some technical conditions, the Bayes risk for the least favorable prior is equal to the minimax risk.
- If Bayes risk is much smaller than the minimax risk, prior added information not present in the constraint itself.

What's it all about, α?

- MSE
- Posterior MSE
- Confidence levels
- Credible levels

Mean Squared Error

- Suppose $\lambda[\theta]$ takes values in a Hilbert space.
- Estimate $\lambda[\theta]$ by $\widehat{\lambda}(Y)$.
- MSE of $\widehat{\lambda}$ when $\theta = \eta$ is

$$\mathrm{MSE}(\widehat{\lambda}(Y), \eta) \equiv \mathbb{E}_\eta \|\widehat{\lambda}(Y) - \lambda[\eta]\|^2.$$

- Depends on η: Expectation is wrt $\mathbb{P}_\eta$.
- True θ is unknown, so can't select $\widehat{\lambda}$ to minimize MSE.
- Might choose $\widehat{\lambda}$ to make the largest MSE for $\eta \in \Theta$ as small as possible: *minimax MSE estimator*.

Posterior Mean Squared Error

- $\mathrm{PMSE}(\widehat{\lambda}(y), \pi) \equiv \mathbf{E}_\pi \|\widehat{\lambda}(y) - \lambda[\eta]\|^2$.
- Depends on π and observed value y.
- Expectation is wrt posterior distribution of θ given $Y = y$.
- π is known, so can select (for each y) the estimator w/ smallest PMSE.
- Bayes estimator for PMSE is posterior marginal mean:

$$\widehat{\lambda}_\pi(y) \equiv \int \ell \pi_\lambda(d\ell | Y = y).$$

MSE v PMSE

- Both involve expectations of the squared norm of the difference between the estimate and the true value of the parameter.
- MSE: expectation wrt distribution of Y, holding $\theta = \eta$ fixed.
- PMSE: expectation wrt posterior distribution of θ, holding $Y = y$ fixed.

Confidence Sets and Credible Regions

- Pick $\alpha \in (0, 1)$.
- A random set $\mathcal{I}(Y)$ of possible values of λ is $1 - \alpha$ confidence set for $\lambda[\theta]$ if

$$\mathbb{P}_\eta\{\mathcal{I}(Y) \ni \lambda[\eta]\} \geq 1 - \alpha, \quad \forall \eta \in \Theta.$$

- Probability is wrt distribution of Y holding η fixed.
- "Coverage probability" is the (smallest) chance that $\mathcal{I}(Y) \ni \lambda[\eta]$ for $\eta \in \Theta$, with $Y \sim \mathbb{P}_\eta(y)$.

Posterior Credible Region

- Pick $\alpha \in (0, 1)$.
- Set $\mathcal{I}(Y)$ of possible values of λ is $1 - \alpha$ posterior credible region for $\lambda[\theta]$ if

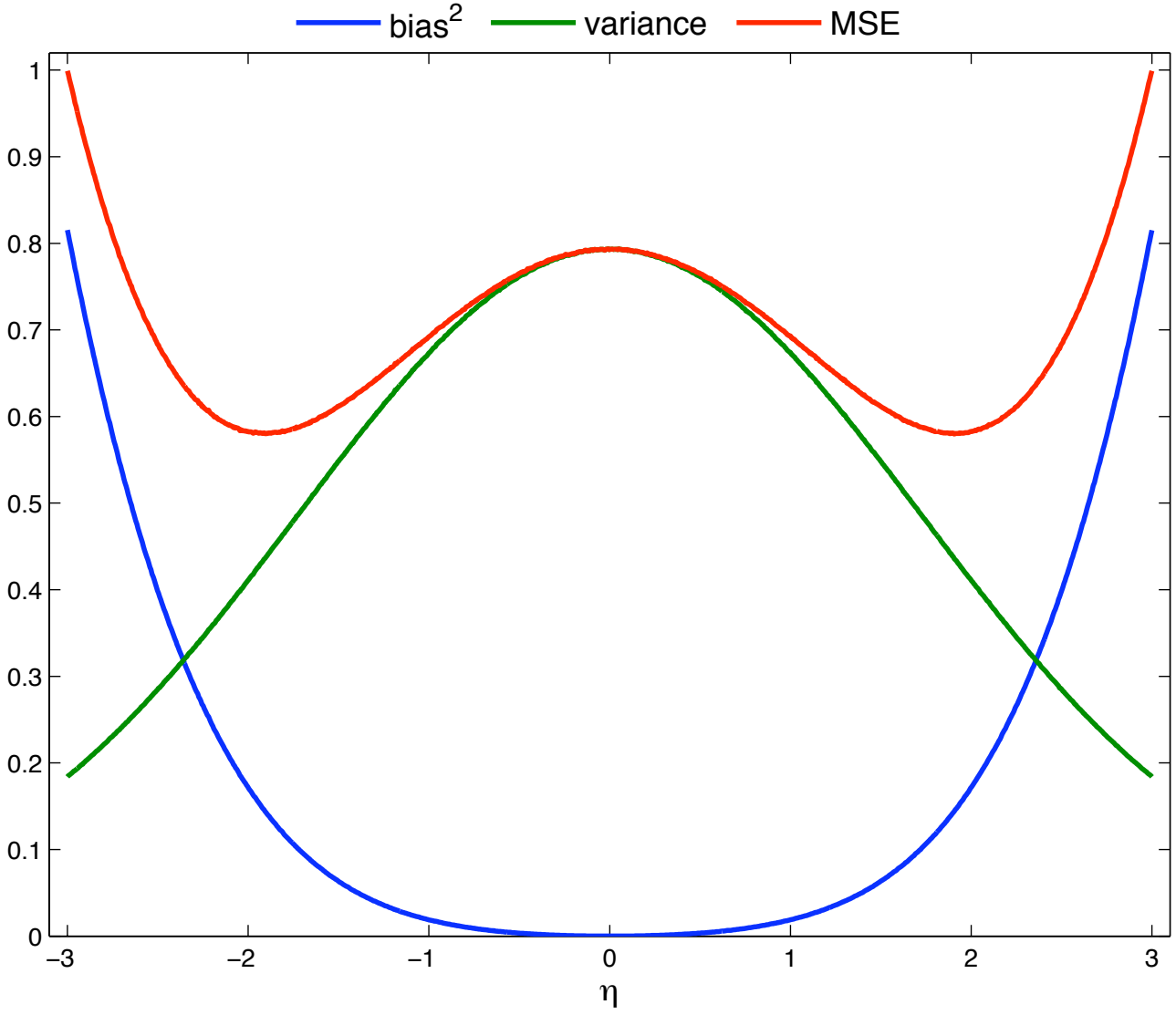
$$\mathbf{P}_{\pi(d\theta|Y=y)}(\lambda[\theta] \in \mathcal{I}(y)) \equiv \int_{\mathcal{I}(y)} \pi_\lambda(d\ell|Y = y) \geq 1 - \alpha.$$

- Probability is wrt posterior distribution of θ, holding $Y = y$.
- Posterior probability that $\mathcal{I}(y) \ni \lambda[\theta]$ given $Y = y$.

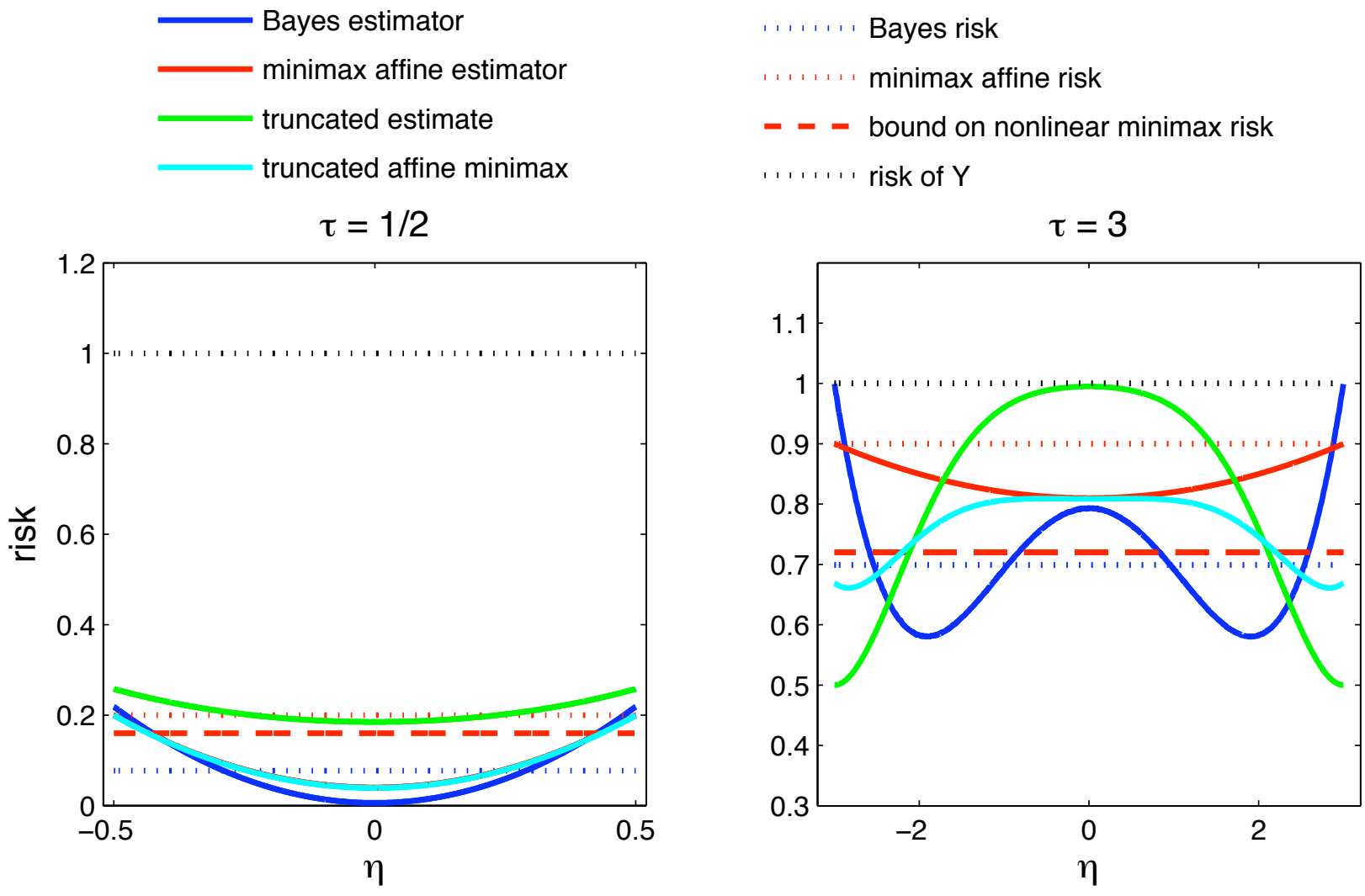
Frequentist performance of Bayes estimators for a BNM

- Maximum MSE of the Bayes estimator of θ for MSE risk
- Frequentist coverage probability and expected length of the 95% Bayes credible interval for θ.

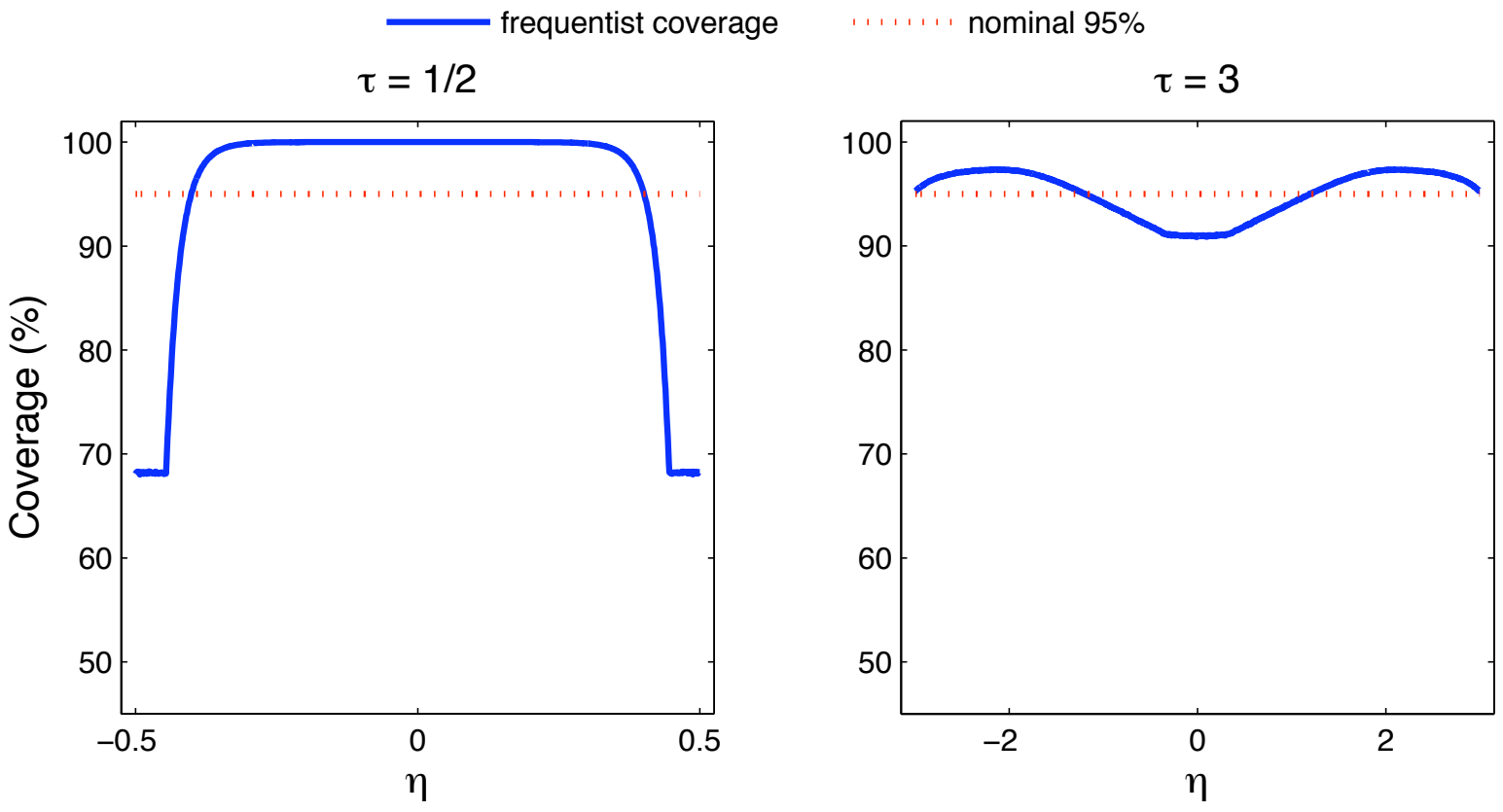
Bias^2, variance, MSE of Bayes estimator of BNM, $\pi \sim U[-3, 3]$



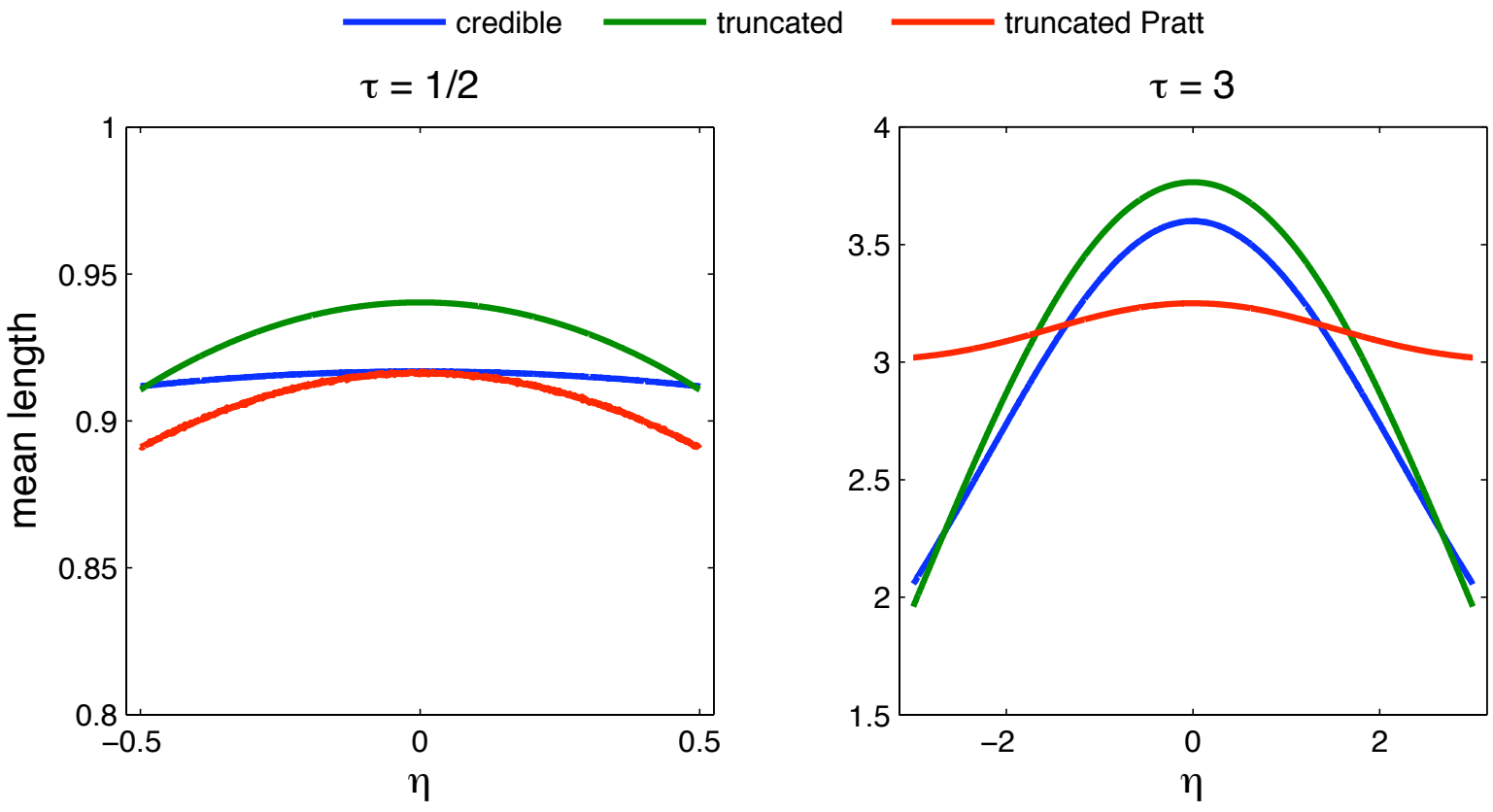
MSE Risk


Bayes estimator
minimax affine estimator
truncated estimate
truncated affine minimax
Bayes risk
minimax affine risk
bound on nonlinear minimax risk
risk of Y
τ = 1/2
τ = 3
risk
η
η
1.2
1
0.8
0.6
0.4
0.2
0
−0.5
0
0.5
1.1
0.9
0.7
0.5
0.3
−2
2

Frequentist coverage of 95% Credible Region



Expected length of the Bayesian Credible Region for BNM



Summary

- Bayesian methods require augmenting data and physical constraints with priors.
- Difference between frequentist and Bayesian viewpoints: Bayesians claim to know more about how the data are generated.
- Frequentists claim to know $\theta \in \Theta$, but not how θ is to be selected from Θ.
- Bayesians claim to know θ is selected at random from Θ according to prior distribution π, known to them.
- Both claim to know $\mathbf{P}_\eta$, the probability distribution of data if θ is η, for $\eta \in \Theta$.

Summary

- Bayesian: probability measures what the analyst thinks.
- Frequentist: probability has to do with empirical regularities.
- Model-based inference: what does probability mean, outside the model?
- Bayesian and frequentist measures of uncertainty differ.
- MSE and PMSE are expectations of the same thing, but wrt different distributions:
  MSE wrt distribution of the data, holding the parameter fixed
  PMSE wrt posterior distribution of the parameter, holding data fixed.
- Coverage probability and credible level are the chance that a set contains θ
  - Coverage probability is wrt distribution of data, holding parameter fixed and allowing set to vary randomly.
  - Credible level is wrt posterior distribution of parameter, holding data and set fixed and allowing θ to vary randomly.

Summary

- Priors add information not present in constraints.
- Worse in higher dimensions.
- With mild conditions, the largest Bayes risk as the prior is allowed to vary is the minimax risk as the parameter is allowed to vary
- If Bayes risk for a given prior is less than the minimax risk, beware: Do your beliefs match the analyst's?