Approximate Integral Methods for Fast Model Diagnostics

Dave Campbell
Haoxuan Zhou
Shaun McDonald

NSERC CRSNG
CANSSI INCASS
Canadian Statistical Sciences Institute
Institut canadien des sciences statistiques
SFU

@iamdavecampbell
www.stat.sfu.ca/~dac5

Outline

- Abstract Motivating Example: State Space Models
- Marginalizing Over Nuisance Parameters:

  Fast and Approximate or Slow and Precise

- Using Probabilistic Numerics to Decide if We Need UQ; how & where

State Space Models

- Observation Equation:

$$Y_t = f(X_t, \theta) + \epsilon_t$$

- (Unobservable) State Equation:

$$X_t = g(X_{t-1}, \phi) + \tau_t$$

- Full likelihood $L(Y, X \mid \theta, \phi, \ldots)$ is intractable.

- State Space Models

$$Y_t = f(X_t, \theta) + \epsilon_t \qquad X_t = g(X_{t-1}, \phi) + \tau_t$$

$$[Y_t \mid X_t, \theta, \sigma_\epsilon^2] \qquad [X_t \mid X_{t-1}, \phi, \sigma_\tau^2]$$

- Goal is to use:

$$\Theta = [\theta, \phi, \sigma_\epsilon^2, \sigma_\tau^2]$$

$$[Y \mid \Theta] = \int [Y \mid X, \Theta][X \mid \Theta] dX$$

$$[Y \mid \Theta] = \int [Y, X \mid \Theta] dX$$

Numerical strategy

$$[Y \mid \Theta] = \int [Y, X \mid \Theta] dX$$

- Choose one:

1. Realistically finite computational time

2. Accurate integral

MCMC strategy

$$[Y \mid \Theta] = \int [Y, X \mid \Theta] dX$$

- Perform Monte Carlo Integration
- Rely on vanishing uncertainty as samples $N \to \infty$
- Fancy algorithms exist but high dimensions tend to be slow & hard to assess convergence

$$[Y \mid \Theta] \approx \frac{1}{N} \sum_{i=1}^{N} [Y, X^{(i)} \mid \Theta]$$

Laplace Approximation

- We want to marginalize over the nuisance latent X

$$[Y \mid \Theta] = \int [Y, X \mid \Theta] dX$$

- High dimensional integrals are tricky, so approximate!

- At the MLE

$$\frac{d}{dX}[Y, X \mid \Theta]\Big|_{X=\hat{X}} = 0$$

Laplace Approximation

For likelihood:

$$[Y, X \mid \Theta] = exp\left\{log([Y, \hat{X} \mid \Theta])\right\}$$

The 2nd order Taylor approximation of the log(likelihood) is:

$$log([Y, X \mid \Theta]) \approx log([Y, \hat{X} \mid \Theta]) + \frac{1}{2}\frac{d^2}{dX^2} log\left([Y, \hat{X} \mid \Theta]\right)(X - \hat{X})^2$$

So

$$[Y, X \mid \Theta] \approx exp\left[log([Y, \hat{X} \mid \Theta]) + \frac{1}{2}\frac{d^2}{dX^2} log\left([Y, \hat{X} \mid \Theta]\right)(X - \hat{X})^2\right]$$

Laplace Approximation

- Recall the goal:

$$[Y \mid \Theta] = \int [Y, X \mid \Theta] dX$$

- So integrate:

$$\int [Y, X \mid \Theta) dX$$

$$\approx \int \left\{ exp\left[log([Y, \hat{X} \mid \Theta]) + \frac{1}{2} \frac{d^2}{dX^2} log\left([Y, \hat{X} \mid \Theta] \right) (X - \hat{X})^2 \right] \right\} dX$$

$$= [Y, \hat{X} \mid \Theta] \int exp\left[\frac{-1}{2} \left(-\frac{d^2}{dX^2} log\left([Y, \hat{X} \mid \Theta] \right) \right) (X - \hat{X})^2 \right] dX$$

Laplace Approximation

$$\int [Y, X \mid \Theta) dX$$

$$\approx [Y, \hat{X} \mid \Theta] \int exp\left[\frac{-1}{2}\left(-\frac{d^2}{dX^2} log\left([Y, \hat{X} \mid \Theta]\right)\right)(X - \hat{X})^2\right] dX$$

- But:

$$exp\left[\frac{-1}{2}\left(-\frac{d^2}{dX^2} log\left([Y, \hat{X} \mid \Theta]\right)\right)(X - \hat{X})^2\right]$$

is the kernel of a Gaussian with mean $\hat{X}$ and variance:

$$\left(-\frac{d^2}{dX^2} log\left([Y, \hat{X} \mid \Theta]\right)\right)^{-1}$$

Laplace Approximation workflow:

- At each numerical optimization iteration:
  - Propose an updated Θ
  - Find the conditional MLE of X in [Y,X| Θ]
  - Use the Laplace approximation for $\int [Y, X \mid \Theta] dX$
  - Evaluate [Y|Θ]
  - Continue until [Y|Θ] is maximized

The Good News and the Bad News

- Good news: Having an approximation for

$$[Y \mid \Theta] = \int [Y, X \mid \Theta] dX$$

  makes parameter estimation tractable, and fast.

- Bad news:

Implies Gaussian Shape.

There are no fast diagnostic tools and poor approximations can give misleading results for estimating Θ.

Goal of this work

- Laplace and other short cuts: fast, occasionally exact, doesn't have a diagnostic This work

- Monte Carlo Integration: slow, asymptotically exact

Goal of this work

- Laplace and other short cuts: fast, occasionally exact, doesn't have a diagnostic This work

- Probabilistic integration as a diagnostic

- Monte Carlo Integration: slow, asymptotically exact

Target of this work: Probabilistic Numerical Marginalization

- Slower than Laplace Approximation, but faster than other Monte Carlo (or other) integration.
- Has a diagnostic statistical test.
- Possible outcomes:

❋Laplace is bad, use a method with more precision.

❋Laplace is good enough. No need for bigger tools.

Probabilistic Solution for DE models

- Goal: Obtain a sample from the distribution of the estimated integral to see if the Laplace approximation is good enough

$$[Y \mid \Theta] = \int [Y, X \mid \Theta] dX$$

- Method: Use a distribution on a function space and a sequential updating scheme to estimate the integral

- Tool: Bayesian Integrator for Frequentist Estimation

Terminology
(ignoring Θ)

- full function $g(Y, X)$
- function of interest $f(Y) = \int g(Y, X) dX$
- For a given (Y,X), g(Y,X) can be evaluated exactly.

Probabilistic integrator part

$$f(Y) = \int g(Y, X) dX$$

- Set up a GP model:

$$\begin{bmatrix} \mu_g(x) \\ \mu_f(x) \end{bmatrix} \sim \mathcal{GP}\left(\begin{bmatrix} m_g(x) \\ m_f(x) \end{bmatrix}, \begin{bmatrix} C_{gg} & C_{gf} \\ C_{fg} & C_{ff} \end{bmatrix} \right)$$

- Choose covariance kernel for g
- Use the integrated covariance kernel for f

Probabilistic integrator part

$$f(Y) = \int g(Y, X) dX$$

- Set up a GP model:

$$\begin{bmatrix} \mu_g(x) \\ \mu_f(x) \end{bmatrix} \sim \mathcal{GP}\left(\begin{bmatrix} m_g(x) \\ m_f(x) \end{bmatrix}, \begin{bmatrix} C_{gg} & C_{gf} \\ C_{fg} & C_{ff} \end{bmatrix} \right)$$

- Use the Laplace approximation to define the prior means.

Interrogate the function full function to integrate the target function

$$f(Y) = \int g(Y, X) dX$$

- Obtain some interrogation points χ
- Point mass likelihood
- Condition on them to get a posterior

$$\begin{bmatrix} \mu_g(x) \\ \mu_f(x) \end{bmatrix} \middle| \chi \Bigg] \sim \mathcal{GP}\left(\begin{bmatrix} m_g(x) \\ m_f(x) \end{bmatrix} \middle| \chi \Bigg], \begin{bmatrix} C_{gg} & C_{gf} \\ C_{fg} & C_{ff} \end{bmatrix} \middle| \chi \Bigg] \right)$$

Probabilistic Diagnostic tool

$$f(Y) = \int g(Y, X) dX$$

- Ho: Laplace approximation is adequate
- Ha: The underlying g is “not gaussian enough”

- Obtain a large sample from $\boldsymbol{\mu_f(x)}$ and see if the Laplace approximation lies therein.

$$\begin{bmatrix} \mu_g(x) \\ \mu_f(x) \end{bmatrix} \Big| \chi \sim \mathcal{GP}\left(\begin{bmatrix} m_g(x) \\ m_f(x) \end{bmatrix} \Big| \chi, \begin{bmatrix} C_{gg} & C_{gf} \\ C_{fg} & C_{ff} \end{bmatrix} \Big| \chi \right)$$

Considerations

$$f(Y) = \int g(Y, X) dX$$

- Interrogation points must be “well chosen”
- Curse of dimensionality

Choosing Interrogation Points

$$f(Y) = \int g(Y, X) dX$$

- Criteria: Choose points $\{\hat{X} - \epsilon\sigma, \hat{X}, \hat{X} + \epsilon\sigma\}$
- σ is the standard deviation of the (rotated) marginal covariance directions (from Laplace)
- $\hat{X}$ is the MLE (from Laplace)
- ε is chosen to maximize power

Optimal ε

- Maximize KL divergence between $[f \mid \chi]$ and the result you'd get if Ho was true

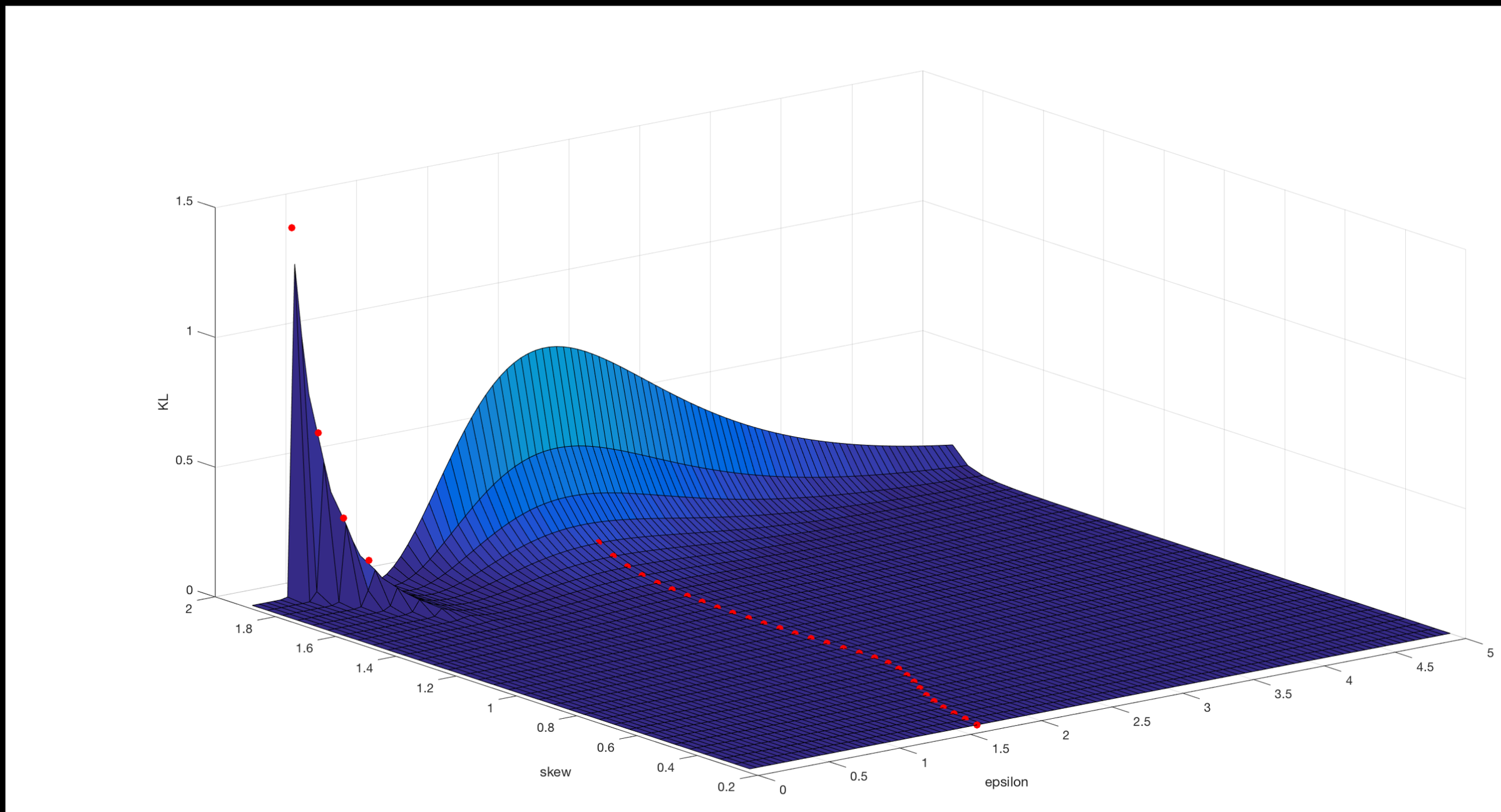
$$\mathrm{KL} = \mathrm{bias}^2/(2 * \mathrm{variance})$$

- Bias increases as ε grows, but eventually all targets converge
- .Variance increases as ε grows

If Truth is Gamma

- Scale parameter fixed, manipulating shape α

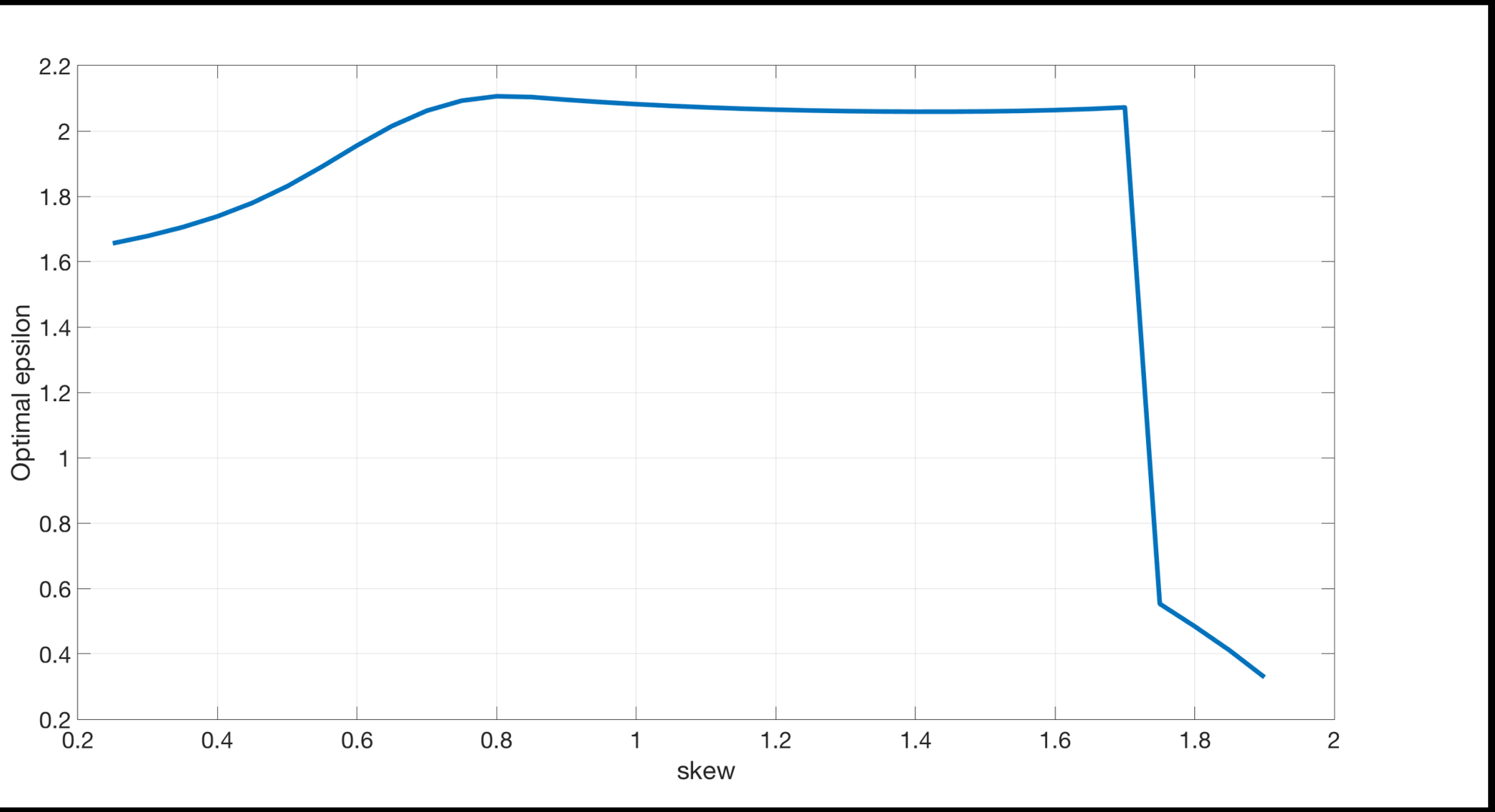
$$\text{Skew} = 2/\sqrt{\alpha} \qquad \text{Kurtosis} = 6/\alpha$$



If Truth is Gamma

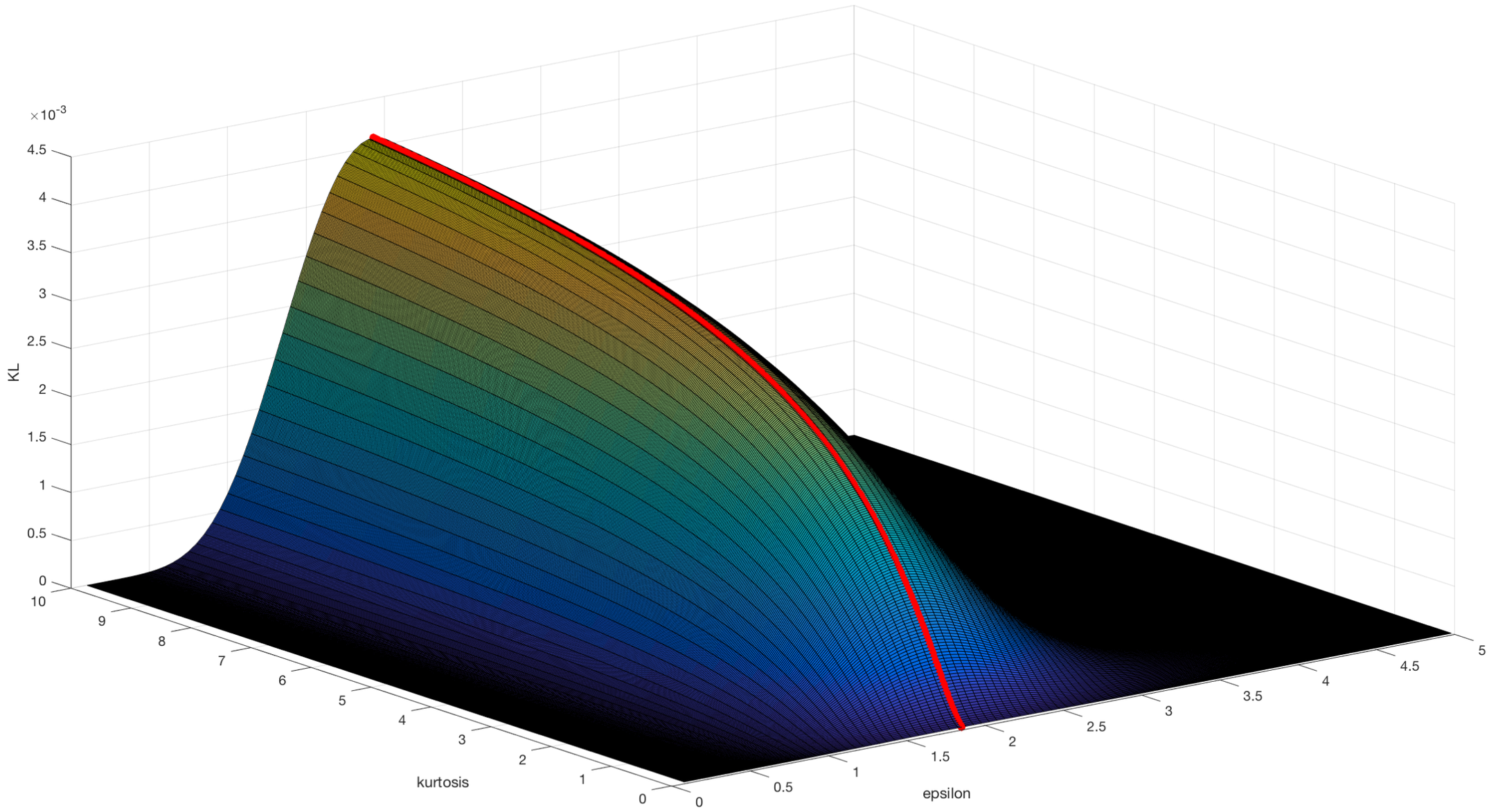
- Scale parameter fixed, manipulating shape α

$$\text{Skew} = 2/\sqrt{\alpha} \qquad \text{Kurtosis} = 6/\alpha$$



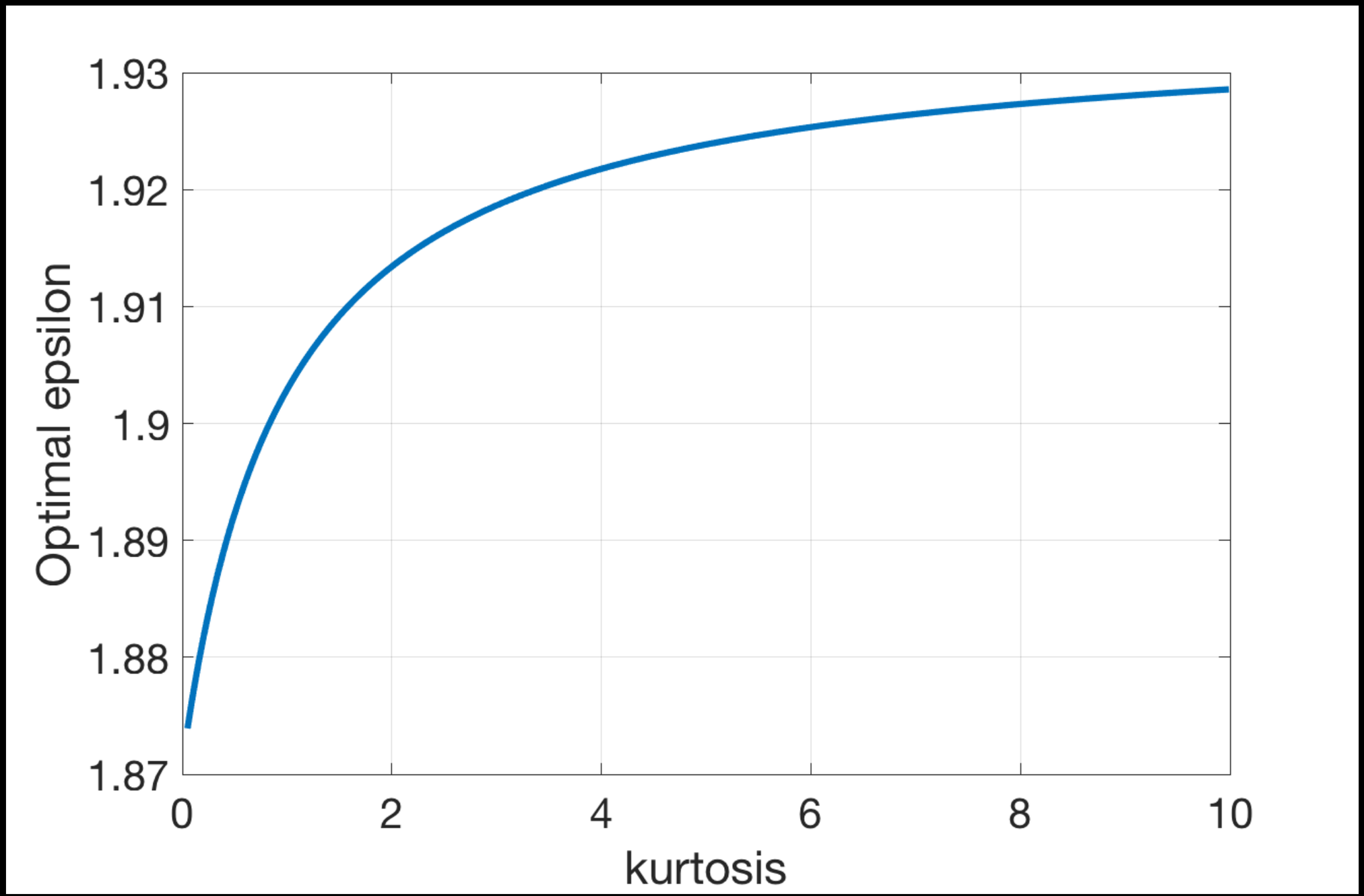
If Truth is T

$$\text{Kurtosis} = 6/(\nu - 4)$$

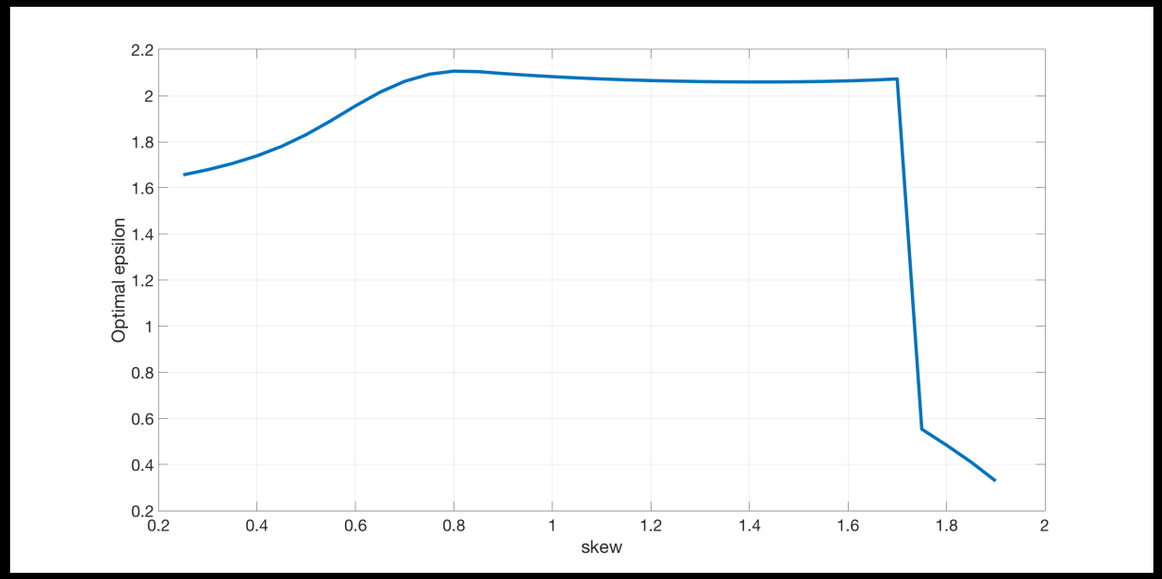


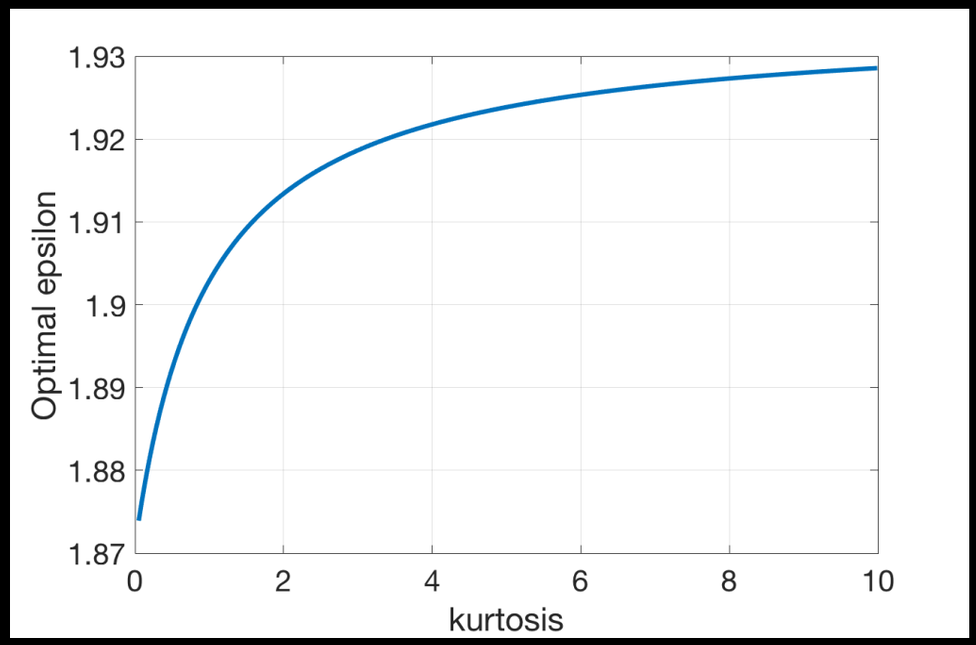
If Truth is T

$$\text{Kurtosis} = 6/(\nu - 4)$$



- A value of ε in (1.9,2.1) performs well under skew and kurtosis scenarios.





Dimensionality problem

$$f(Y) = \int g(Y, X) dX$$

- X is already **d** dimensional
- Laplace discretization = 1 point, 1 hessian

Numerical Integration:

- dense grid in **d** dimensions

Our approach uses:

- $3^{\mathbf{d}}$ -1 additional points
- gpus
- sparse covariance kernels

Ongoing

- Dimensionality through blockwise assessment
- Moving into sparse covariances to buy a few more
- Diagnosing a need for robust estimators