Minimum Hellinger Distance in Model Selection and Estimation

Matt Hutmacher

Pharmacometrics – Pfizer, Inc.

14JUN2006

Joint work:

Anand Vidyashankar (Cornell University)

Debu Mukherjee (Statsystem Inc.)

Introduction

- Need for better model selection techniques
  - Subject matter can not always provide model form
  - Empirical models often used in exposure-response
  - Nonlinear mixed effects is flexible with respect to model forms
  - Marginal variance depends upon model form
  - Robust selection
- Population modelers (pharmacometricians) are well-suited for developing and applying new data-analytic techniques to enhance decision making

Objective

- Introduce Hellinger Distance as a principled methodology for selection between nonhierarchical models
- Introduce the concept of minimizing the Hellinger Distance as an efficient yet robust estimator – an alternative to the MLE (or ELS)

What is Hellinger Distance

- Definition

$$HD^2 = \int \left(f_y^{1/2} - g_y^{1/2} \right)^2 dy = 2 - 2\int f_y^{1/2} g_y^{1/2} dy$$

  - An absolute measure between two densities
  - $HD^2 = 0$ when $f \equiv g$ ranging from 0-2 inclusive

- HD for two univariate normal densities

$$HD^2 = 2 - 2\frac{\sqrt{2}\left(\sigma_1^2 \sigma_2^2\right)^{1/4}}{\left(\sigma_1^2 + \sigma_2^2\right)^{1/2}} \exp\left[-\frac{\left(\mu_1 - \mu_2\right)^2}{4\left(\sigma_1^2 + \sigma_2^2\right)} \right]$$

  - Extendable to multivariate normal for population model densities

HD for Model Selection

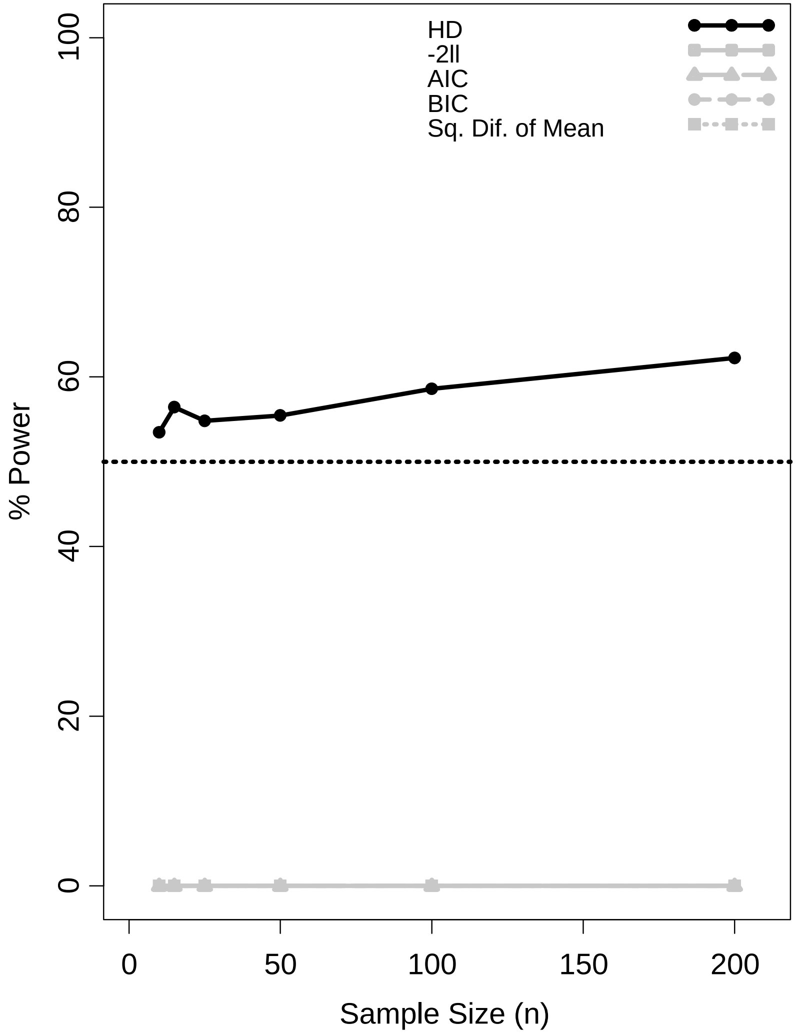
- Likelihood based methods (such as AIC, BIC) are often applied to nonhierarchical model selection
  - Debatable as to whether these methods are appropriate
  - Robust to outliers and data contamination?
- Ultimately interested in selecting a model that is “closest” to the underlying model
- HD for model selection
  - Targets the “closest” parametric model to an assumption-poor nonparametric model form
  - “Closeness” defined with respect to the first two moments (mean, variance)

HD for MS (Implementation)

- Compare models to a nonparametric assessment of model form
  - Estimate μ nonparametrically (g_n)
    - Loess, kernel smoother, mean
    - $g_n \rightarrow$ true μ under mild conditions
  - Estimate σ_g^2 using residuals
  - Compute $HD(f_1,g)$ and $HD(f_2,g)$
  - Select the f with the smallest HD

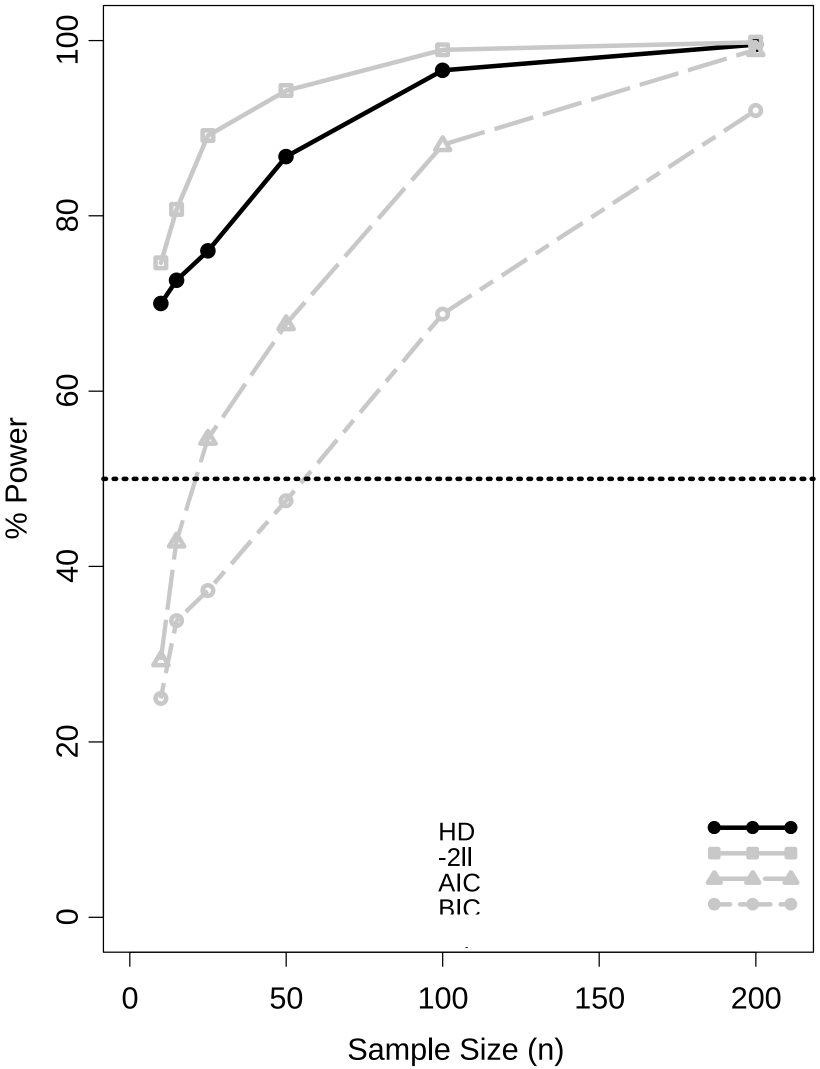
HD for Model Selection (Examples)

- Example 1
  - – Emax model + ε
    - True model
    - 3 parameters
  - – Emax model × exp(ε)
    - False model
    - 3 parameters
  - – <25% CV
  - – 2000 simulations



HD for Model Selection (Examples)

- Example 2
  - – Emax + ε
    - True model
    - 3 parameters
  - – Linear + ε
    - False model
    - 2 parameters
  - – 10% outliers
    - 4-6 σ range
  - – 2000 simulations



Minimum HD Estimation (MHDE)

- Recall HD definition

$$HD^2(\theta) = 2 - 2\int f_\theta^{1/2} h_n^{1/2} dy, \quad \gamma(\theta) = \int f_\theta^{1/2} h_n^{1/2} dy$$

- Consider:
  - f_θ a model density of interest (eg, Normal)
  - h_n a nonparametric density estimator
- The parameters (θ) can be estimated
  - HD (or γ) is a well-defined objective function (bounded)
  - HD definition suggests minimizing HD^2 or maximizing γ for estimation of θ [Beran (1977)]

Minimum HD Estimation

- Integral evaluation

$$\gamma(\theta) = \int f_\theta^{1/2} h_n^{1/2} dy = \int \frac{f_\theta^{1/2}}{h_n^{1/2}} h_n dy$$

$$= E_y\left(\frac{f_\theta^{1/2}}{h_n^{1/2}}\right) \approx \tfrac{1}{M} \sum_j \left[\frac{f_\theta(y_j^*)}{h_n(y_j^*)}\right]^{1/2}, \quad y_j^* \sim h_n$$

- Integral by SLLN [Cheng & Vidyashankar (2003)]
- Simulating $y_j^* \sim h_n$ is the key

Minimum HD Estimation

- Sampling $y_j^* \sim h_n$
  - Let $y_i = m_\theta(x_i) + \varepsilon_i,\ 1 \le i \le n$
  - Estimate $m(x_i)$ nonparametrically

$$\hat{m}(x_i) = g_n(x_i);\quad \text{e.g.,}\ g_n(x) = \frac{\sum y_i K\left(\frac{x - x_i}{c_n}\right)}{\sum K\left(\frac{x - x_i}{c_n}\right)}$$

  - Calculate residuals $\tilde{\varepsilon}_i = y_i - g_n(x_i)$
  - Estimate the density of the residuals

$$h_n(\varepsilon) = \frac{1}{nc_n} \sum_i^n K\left(\frac{\varepsilon - \tilde{\varepsilon}_i}{c_n}\right) = \frac{1}{nc_n} \sum_i^n K_i$$

Minimum HD Estimation

– Recall $\gamma(\theta) = \int \frac{f_\theta^{1/2}}{h_n^{1/2}} h_n dy \approx \frac{1}{M} \sum_j \left[\frac{f_\theta(y_j^*)}{h_n(y_j^*)} \right]^{1/2}, \; y_j^* \sim h_n$

– For j-th term in the integral approximation

- Sample a K_i with probability p such as $1/n$ (i')
- Sample a random variable from this $K_{i'}$, i.e., $\varepsilon_j^* = \tilde{\varepsilon}_{i'} + c_n \delta_j$, where $\delta \sim K$ (e.g., $N[0,1]$)
- Then $y_j^* = g_n(x) + \varepsilon_j^*$ or $y_j^* \sim h_n$ where h_n is a smoothed empirical density

– Optimization of θ is now possible

- SAS PROC NLP (or PROC MODEL)

MHDE (Examples)

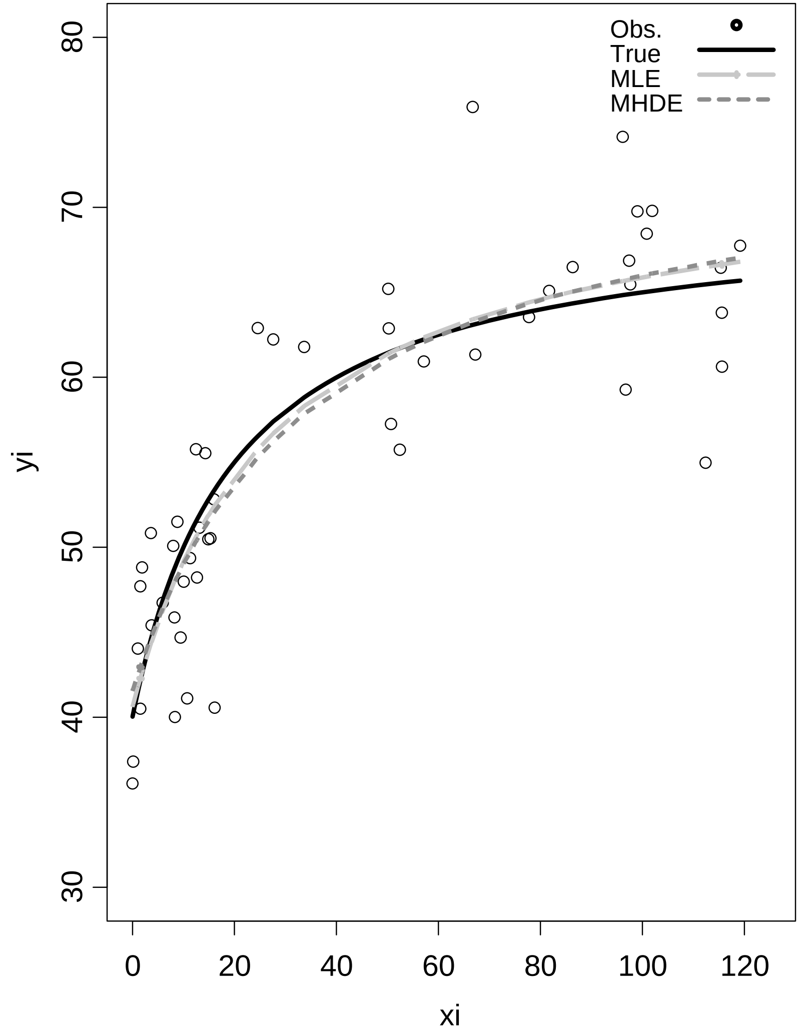
- Example 3
  - Model:

$$y_i = E_o + \frac{E_{\max} \cdot x_i}{EC_{50} + x_i} + \varepsilon_i$$

  - Distribution:

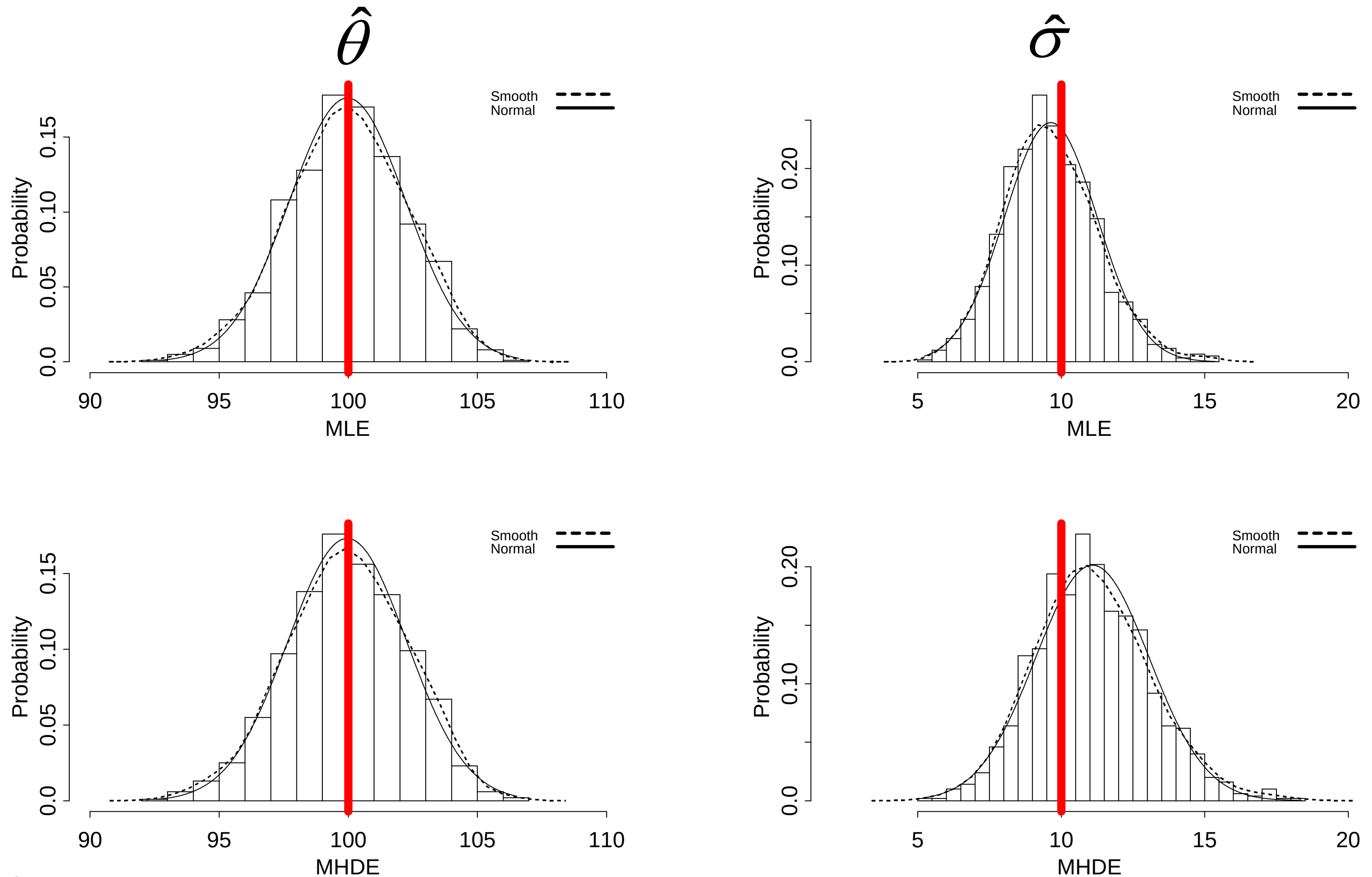
$$\varepsilon_i \sim N\left(0, \sigma^2 = 25\right)$$

  - N=50



Ex. 4 $y_i = \theta + \varepsilon_i \quad \theta = 100, \sigma = 10 \quad \varepsilon_i \sim N(0, \sigma^2)$

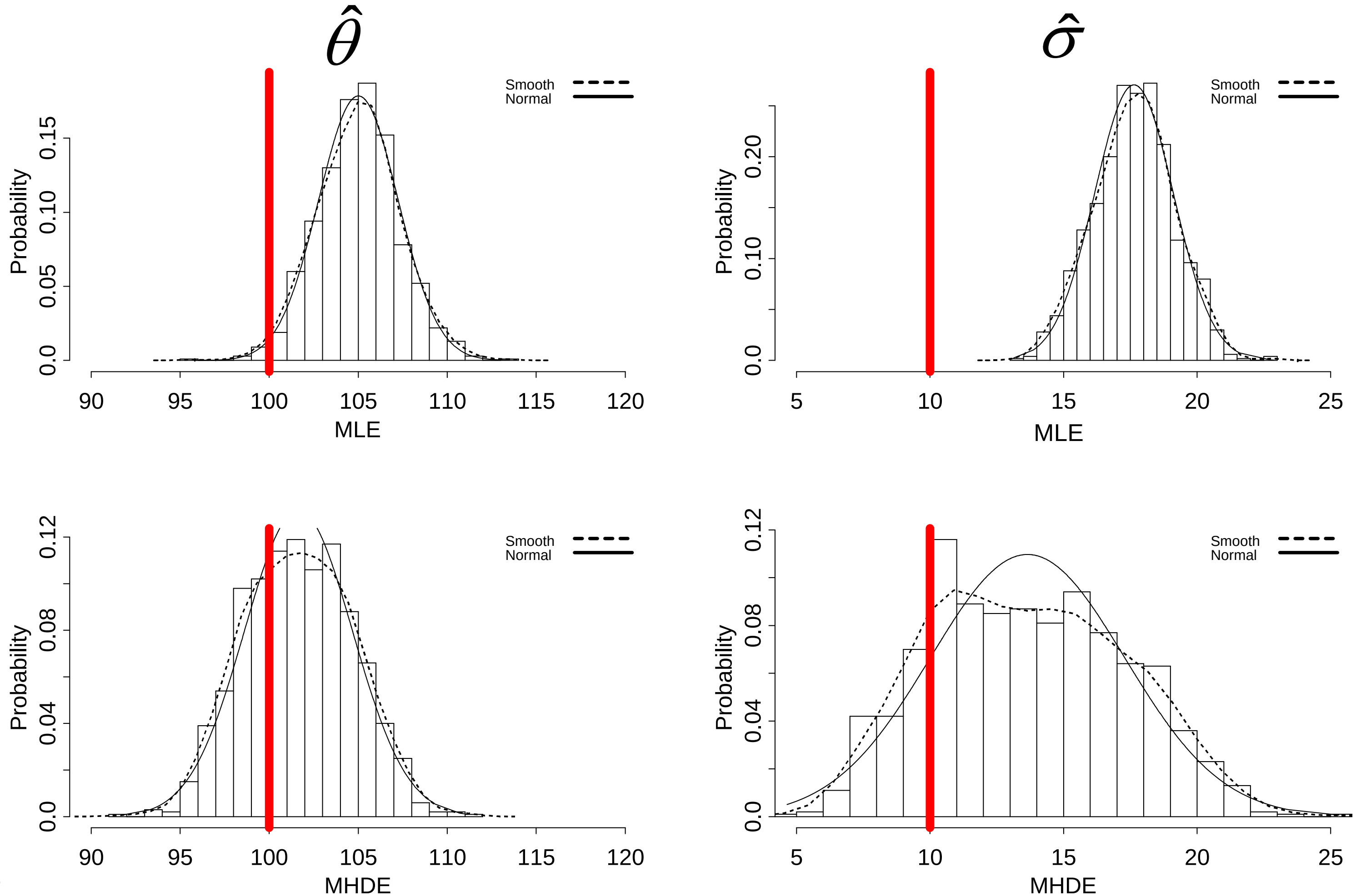
N=20 – 1000 Simulations



Ex. 5 $y_i = \theta + \sigma\varepsilon_i \;\; \theta = 100, \sigma = 10$

90%: $\varepsilon_i \sim N(0, \sigma^2)$ **10%:** $\varepsilon_i \sim 4 - 6\sigma$ outliers $(+)$

N=20 – 1000 Simulations



Remarks

- Properties of a good estimator [Beran]
  - Efficient when model is true
  - Not much loss when model is approximately true
- MLE (ELS)
  - Is efficient when the model is true
  - Can suffer instability under data contamination (inefficiencies and lack of robustness)
  - Uses a squared error with a penalty for increasing the variance

Remarks

- MHDE is efficient with increased robustness
  - Efficient when model is true
  - Increased efficiency relative to total nonparametric estimation
  - Nonparametric empirical density estimator reduces the influence of outliers (less loss when approximately true)
  - Novel methodology
    - Smoothed mixture of kernels to emulate empirical density of data
    - Generalizes MHDE to typical data-analytic problems
    - Well-suited to simulate data adequately