Maximum Entropy on the Mean and the Cramér Rate Function in Statistical Estimation

Yakov Vaisbourd, Tim Hoheisel, Rustum Choksi, Carola-Bibiane Schönlieb, Ariel Goodwin

McGill University *ariel.goodwin@mail.mcgill.ca*

March 31st, 2022

Motivation: Linear Inverse Problems

Canonical example:

- $C \in \mathbb{R}^{d \times d}$ (forward operator)
- $b \in \mathbb{R}^{d}$ corrupted data with noisy measurements

How can we accurately estimate $x \in \mathbb{R}^{d}$ such that $Cx \sim b$? Problem is ill-posed in general.

Example: Image deblurring with noise

- $C \in \mathbb{R}^{d \times d}$ is convolution matrix
- b is blurred and noisy image

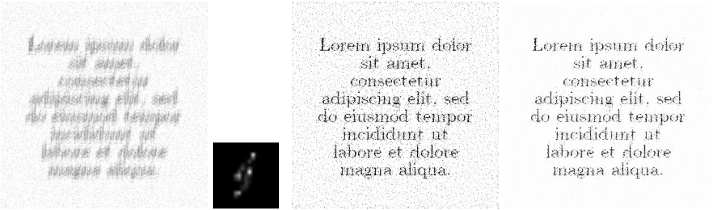


Figure: Rioux et al. (2021)

First Approach

We can solve the ill-posed problem naively via

$$\min\left\{\frac{1}{2}\|Cx-b\|^2 \mid x \in \mathbb{R}^n\right\}$$

- Add a regularizer to penalize violation of prior information, making the problem well-posed
- How can we incorporate knowledge about the problem to choose a good regularizer?

$$\min\left\{\frac{1}{2}\|Cx-b\|^2 + \kappa_R(x) \mid x \in \mathbb{R}^d\right\}$$

- Idea: Treat b as the expected value of some underlying distribution, and choose the distribution in a meaningful way

Roadmap

- Introduce distances induced by convex functions
- Grab some tools from probability and statistics
- Define Maximum Entropy on the Mean (MEM)
- Use our tools to show MEM is the convex conjugate of Cramér's function
- Obtain functions to regularize the problem, and get algorithms/proxs to solve it efficiently

The function class Γ_0

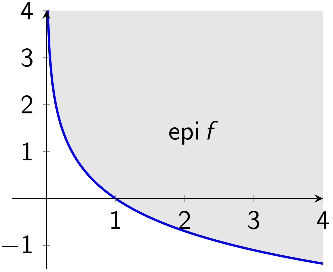
$\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is called:

- **closed** if $\operatorname{epi} \psi := \{(x, \alpha) \in \mathbb{R}^d \times \mathbb{R} : \psi(x) \leq \alpha\}$ is a closed set in $\mathbb{R}^d \times \mathbb{R}$
- **proper** if $\operatorname{dom} \psi := \psi^{-1}(\mathbb{R}) \neq \emptyset$
- **convex** if $\operatorname{epi} \psi \subseteq \mathbb{R}^d \times \mathbb{R}$ is a convex set

Example:

The function $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$

$$f(x) = \begin{cases} -\log x, & x > 0, \\ +\infty, & x \leq 0, \end{cases}$$



Set $\Gamma_0 := \{\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\} \mid \psi \text{ is convex, proper, and closed}\}$

Legendre Type

For $\psi \in \Gamma_0$ we define its **convex conjugate** $\psi^* \in \Gamma_0$ to be

$$\psi^*(y) = \sup\left\{\langle y, x\rangle - \psi(x) \mid x \in \mathbb{R}^d\right\}$$

Definition (Legendre Type)

A function $\psi \in \Gamma_0$ is **essentially smooth** if it satisfies the following conditions:

1. $\operatorname{int}(\operatorname{dom}\psi) \neq \emptyset$
2. ψ is differentiable on $\operatorname{int}(\operatorname{dom}\psi)$
3. $\|\nabla\psi(x^k)\| \to \infty$ for any $\left\{x^k\right\} \subseteq \operatorname{int}(\operatorname{dom}\psi)$ such that $x^k \to \bar{x} \in \partial(\operatorname{dom}\psi)$

If moreover ψ is strictly convex on $\operatorname{int}(\operatorname{dom}\psi)$ then ψ is of **Legendre type**.

Bregman Divergence

Theorem (Rockafellar)

If $\psi \in \Gamma_0$ is of Legendre type then

1. *The convex conjugate ψ^* is of Legendre type*
2. $\nabla\psi$ *is a bijection from* $\operatorname{int}(\operatorname{dom}\psi)$ *to* $\operatorname{int}(\operatorname{dom}\psi^*)$ *with inverse* $(\nabla\psi)^{-1} = \nabla\psi^*$

Legendre type functions induce **Bregman divergences**:

$$D_\psi(y, x) := \psi(y) - \psi(x) - \langle \nabla\psi(x), y - x \rangle$$

for $x \in \operatorname{int}(\operatorname{dom}\psi), y \in \operatorname{dom}\psi$

- $D_\psi(y, x) \geq 0$ with equality iff $y = x$
- Not symmetric in general

Probability Theory

Let ρ, μ σ-finite measures on measurable $\Omega \subseteq \mathbb{R}^d$. Some definitions:

- $\Omega_\rho = \text{supp}(\rho)$ (support of ρ)
- $\Omega_\rho^{cc} = \text{cl}(\text{conv}\,\Omega_\rho)$ (convex support of ρ)
- $\mu \ll \rho$ (μ is **absolutely continuous** w.r.t. ρ) if $\rho(A) = 0$ implies $\mu(A) = 0$
- (Radon-Nikodym) If $\mu \ll \rho$ then $\exists!\ h = \dfrac{\mathrm{d}\mu}{\mathrm{d}\rho}$ called the **Radon-Nikodym derivative** satisfying
$$\mu(A) = \int_A \frac{\mathrm{d}\mu}{\mathrm{d}\rho} d\rho$$

Probability Theory

We consider two cases:

1. $\Omega = \mathbb{R}^d$, underlying measure ν is Lebesgue
2. $\Omega \subseteq \mathbb{R}^d$ is countable, underlying measure ν is counting measure

Define $\mathcal{P}(\Omega) := \{P \text{ probability measure on } \Omega \mid P \ll \nu\}$.

Each such P has Radon-Nikodym Derivative $f_P := \frac{\mathrm{d}P}{\mathrm{d}\nu}$, expected value E_P, and **moment-generating function** M_P:

$$E_P := \int_\Omega y dP(y) \in \mathbb{R}^d$$

$$M_P(\theta) := \int_\Omega \exp(\langle y, \theta \rangle) dP(y)$$

Exponential Families

Let P be σ-finite, $P \ll \nu$. The **natural parameter space** for P is defined by

$$\Theta_P := \left\{ \theta \in \mathbb{R}^d \mid M_P(\theta) = \int_\Omega \exp(\langle y, \theta \rangle) dP(y) < \infty \right\}$$

Definition (Log-Normalizer)

The function $\psi_P \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ by

$$\psi_P(\theta) = \begin{cases} \log M_P(\theta), & \theta \in \Theta_P \\ +\infty, & \theta \notin \Theta_P \end{cases}$$

is called the **log-normalizer**.

Definition (Exponential Family)

The **standard exponential family generated by** P is

$$\mathcal{F}_P := \{ f_{P_\theta}(y) := \exp(\langle y, \theta \rangle - \psi_P(\theta)) \mid \theta \in \Theta_P \}$$

Exponential Family Properties

WLOG we assume $\operatorname{int}\Theta_P \neq \emptyset, \operatorname{int}\Omega_P^{cc} \neq \emptyset$ (an exponential family satisfying this is called **minimal**)

Theorem (Regularity of ψ_P, Brown 1986)

Let $\mathcal{F}_P$ be a minimal exponential family. Then:

1. *The log-normalizer ψ_P is strictly convex on the convex set Θ_P*
2. $\psi_P \in C^\infty(\operatorname{int}\Theta_P)$, $\nabla\psi_P(\theta) = E_{P_\theta}$

If ψ_P is essentially smooth we say $\mathcal{F}_P$ is **steep**.
Conclusion: If $\mathcal{F}_P$ is minimal and steep then ψ_P is of Legendre type.

Corollary (Mean Value Parametrization)

The natural parameter θ can be expressed as

$$\theta = \nabla\psi_P^*(\mu)$$

where $\mu = E_{P_\theta} = \nabla\psi_P(\theta)$.

Kullback-Leibler Divergence

The **Kullback-Leibler (KL) divergence** [Kullback, Leibler (1951)] between σ-finite P and $Q \in \mathcal{P}(\Omega)$ is defined by

$$D_{\mathrm{KL}}(Q||P) := \begin{cases} \int_\Omega \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right) \mathrm{d}Q, & Q \ll P \\ +\infty, & \text{otherwise} \end{cases}$$

Properties:

- If $P \in \mathcal{P}(\Omega)$ then $D_{\mathrm{KL}}(Q||P) \geq 0$ with equality iff $Q = P$
- Not symmetric in general
- Convex in (P, Q)

Principle of Minimum Discriminative Information: Given new information, a new distribution should be chosen that is hard to discriminate from the prior distribution in the sense of KL

The MEM Approach

- **Maximum Entropy on the Mean:** The state best describing a system is the mean of a distribution maximizing some measure of entropy [Jaynes, 1957]

Definition (MEM Function)

The Maximum Entropy on the Mean (MEM) Function $\kappa_P : \mathbb{R}^d \to (-\infty, \infty]$ is defined by [Rietsch, 1977]:

$$\kappa_P(y) := \inf \{ D_{\mathrm{KL}}(Q||P) \mid E_Q = y, Q \ll P \}$$

- Information-driven approach: Measure compliance of y with P via $\kappa_P(y)$
- Applications: crystallography [Navaza (1985)], seismic tomography [Fermín et al. (2006)], medical imaging [Amblard et al. (2004), Deslauriers-Gauthier et al. (2017), Cai et al. (2022)], image processing [Rioux et al. (2020)]

Alternative Representation

$$\kappa_P(y) := \inf\left\{D_{\mathrm{KL}}(Q||P) \mid E_Q = y, Q \ll P\right\}$$

- This representation of κ_P is computationally challenging. How can we use it?
- Under some conditions, $\kappa_P = \psi_P^*$, and ψ_P^* is called the **Cramér rate function** (c.f. large deviations theory)

$$\inf\left\{D_{\mathrm{KL}}(Q||P) \mid E_Q = y, Q \ll P\right\} = \sup\left\{\langle y, \theta\rangle - \log\int_\Omega e^{\langle y,\theta\rangle} dP(y) \mid \theta \in \mathbb{R}^d\right\}$$

Domain of Cramér Rate Function vs. MEM Function

Theorem (Domain of ψ_P^*, Barndorff-Nielsen)

Suppose $P \in \mathcal{P}(\Omega)$ generates a minimal and steep exponential family. Then:

$$\operatorname{int} \Omega_P^{cc} \subseteq \operatorname{dom} \psi_P^* \subseteq \Omega_P^{cc}$$

Moreover, the following hold:

1. *If Ω_P is finite then $\operatorname{dom} \psi_P^* = \Omega_P^{cc}$*
2. *If Ω_P is countable then $\operatorname{dom} \psi_P^* \supseteq \operatorname{conv} \Omega_P$*
3. *If Ω_P is uncountable then $\operatorname{dom} \psi_P^* = \operatorname{int} \Omega_P^{cc}$*

Theorem (Domain of κ_P)

Suppose P satisfies the same assumptions above. Then:

- *If Ω_P is countable then $\operatorname{dom} \kappa_P = \operatorname{conv} \Omega_P$*
- *If Ω_P is uncountable then $\operatorname{dom} \kappa_P = \operatorname{int} \Omega_P^{cc}$*

Key Inequality

Lemma (Cramér vs. MEM Inequality)

Suppose $P \in \mathcal{P}(\Omega)$ generates a minimal and steep exponential family. Then:

$$\psi_P^*(y) \leq \kappa_P(y) \leq \psi_P^*(y) + D_{KL}(Q||P_\theta) - D_{\psi_P^*}(y, \nabla\psi_P(\theta))$$

for any $y \in \operatorname{dom}\kappa_P$, $Q \ll P$ with $E_Q = y$, and $\theta \in \operatorname{int}\Theta_P$. Recall P_θ is defined by density $f_{P_\theta} = \exp(\langle\cdot,\theta\rangle - \psi_P(\theta)) \in \mathcal{F}_P$.

Equivalence of Cramér and MEM

Theorem (Equality Conditions)

Suppose $P \in \mathcal{P}(\Omega)$ generates a minimal and steep exponential family. Moreover, suppose one of the following holds:

- Ω_P *is uncountable*
- Ω_P *is countable and* $\operatorname{conv} \Omega_P$ *is closed*

Then $\kappa_P = \psi_P^*$. In particular, κ_P is closed, proper, and convex.

Corollary (Properties of κ_P)

Suppose the assumptions of the previous theorem hold. Then:

1. $\kappa_P(y) \geq 0$ *with equality iff* $y = E_P$
2. κ_P *is of Legendre type*
3. κ_P *is coercive, i.e.,* $\lim_{\|y\| \to \infty} \kappa_P(y) = +\infty$

Independence and Separability

Suppose the reference distribution $P \in \mathcal{P}(\Omega)$ has a separable structure i.e.,

$$P(y) = P_1(y_1)P_2(y_2)\cdots P_d(y_d)$$

In other words, the coordinates are independent. Then:

$$M_P(\theta) = \prod_{i=1}^{d} M_{P_i}(\theta_i)$$

It follows that:

$$\begin{aligned}\psi_P^*(y) &= \sup\left\{\langle y, \theta\rangle - \log M_P(\theta) \mid \theta \in \mathbb{R}^d\right\} \\ &= \sum_{i=1}^{d} \sup\left\{y_i\theta_i - \log M_{P_i}(\theta_i) \mid \theta_i \in \mathbb{R}\right\}\end{aligned}$$

Examples

Reference Distribution (P)	Cramér Rate Function ($\psi_P^*(y)$)	dom ψ_P^*
Multivariate Normal $\mu \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d, \Sigma \succ 0$	$\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)$	$\mathbb{R}^d$
Poisson ($\lambda \in \mathbb{R}_{++}$)	$y \log(y/\lambda) - y + \lambda$	$\mathbb{R}_+$
Gamma ($\alpha, \beta \in \mathbb{R}_{++}$)	$\beta y - \alpha + \alpha \log\left(\frac{\alpha}{\beta y}\right)$	$\mathbb{R}_{++}$
Normal-inverse Gaussian $\alpha, \beta, \delta \in \mathbb{R} \colon \alpha \geq \lvert\beta\rvert$, $\delta > 0, \gamma := \sqrt{\alpha^2 - \beta^2}$	$\alpha\sqrt{\delta^2 + (y-\mu)^2} - \beta(y-\mu) - \delta\gamma$	$\mathbb{R}$
Multinomial ($p \in \Delta_d, n \in \mathbb{N}$)	$\sum_{i=1}^d y_i \log\left(\frac{y_i}{np_i}\right)$	$n\Delta_d \cap I(p)$[1]

[1] $I(p) := \{x \in \mathbb{R}^d \mid x_i = 0 \text{ if } p_i = 0\}$

The MEM Estimator

Maximum likelihood (ML) is a popular principle of statistical estimation

$$\theta_{ML} = \theta_{ML}(\hat{y}, F_\Theta, S) := \text{argmax}\left\{\log f_{P_\theta}(\hat{y}) \mid \theta \in S \cap \Theta\right\}$$

where:

- $S \subseteq \mathbb{R}^d$ are admissible parameters
- F_Θ parameterized family of distributions $P_\theta, \theta \in \Theta \subseteq \mathbb{R}^d$ with densities f_{P_θ}
- $\hat{y} \in \mathbb{R}^d$ is a sample of observed data

Definition/Theorem

The MEM estimator $y_{MEM} \in \mathbb{R}^d$ is defined by:

$$y_{MEM} = y_{MEM}(\hat{y}, F_\Theta, S^*) := \text{argmin}\left\{\psi^*_{P_{\hat{\theta}}}(y) \mid y \in S^*\right\}$$

where $\hat{\theta}$ is such that $\hat{y} = E_{P_{\hat{\theta}}}$. The existence and uniqueness of y_{MEM} is guaranteed under some mild technical assumptions.

Analogy between MEM and ML for Exponential Families

Theorem (Brown, 1981)

If $\mathcal{F}_P$ is an exponential family then the following characterizations hold (under technical assumptions on $P, \hat{\theta} := \nabla \psi_P^*(\hat{y}), S, S^*$):

1. *(Primal)* $y_{MEM} = \nabla \psi_P(\theta_{MEM})$ *where*

$$\theta_{MEM} \in \operatorname{argmin} \left\{ D_{KL}(P_\theta || P_{\hat{\theta}}) \mid \theta \in S \right\}$$

$$\theta_{ML} \in \operatorname{argmin} \left\{ D_{KL}(P_{\hat{\theta}} || P_\theta) \mid \theta \in S \right\}$$

2. *(Dual)*

$$y_{MEM} \in \operatorname{argmin} \left\{ D_{\psi_P^*}(y, \hat{y}) \mid y \in S^* \right\}$$

$$\theta_{ML} \in \operatorname{argmin} \left\{ D_{\psi_P}(\theta, \hat{\theta}) \mid \theta \in S \right\}$$

Linear Models

- Applications: bioinformatics, image processing, machine learning, ...
- $A \in \mathcal{C} \subseteq \mathbb{R}^{m\times d}$ (dictated by the problem)
- $F_\Theta = \{P_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^m\} \subseteq \mathcal{P}(\Omega)$

Reference distribution $P_{\hat\theta}$ is specified via $\hat y = E_{P_{\hat\theta}}$ where $\hat y$ is our observation vector. Thus the MEM estimator of the linear model is:

$$\min\left\{\psi^*_{P_{\hat\theta}}(Ax) \mid x \in X\right\}, \quad (A \in \mathcal{C}, \hat\theta \in \Theta \colon E_{P_{\hat\theta}} = \hat y)$$

Reference Family	Objective Function ($\psi^*_{P_{\hat\theta}} \circ A$)
Normal	$\frac{1}{2}\|Ax - \hat y\|^2$
Poisson	$\sum_{i=1}^m [\langle a_i, x\rangle \log(\langle a_i, x\rangle / \hat y_i) - \langle a_i, x\rangle + \hat y_i]$
Gamma ($\beta = 1$)	$\sum_{i=1}^m [\langle a_i, x\rangle - \hat y_i \log(\langle a_i, x\rangle) - (\hat y_i - \hat y_i \log \hat y_i)]$

Regularized Model

- Remark: If $X = \mathbb{R}^d, \mathrm{rge} A = \mathbb{R}^m$ with $m < d$ the MEM and ML estimators coincide due to ill-posedness of the model.
- Idea: Regularize to create well-posed problem

$$\min \left\{ \psi^*_{P_{\hat{\theta}}}(Ax) + \varphi(x) \mid x \in X \right\}, \quad (A \in \mathcal{C}, \hat{\theta} \in \Theta \colon E_{P_{\hat{\theta}}} = \hat{y})$$

- Here $\varphi \colon \mathbb{R}^d \to (-\infty, \infty]$ is closed, proper, convex.
- We can use Cramér's function to regularize
- Take $R \in \mathcal{P}(\Omega)$ as a prior distribution encoding info about the desired solution

$$\min \left\{ \psi^*_{P_{\hat{\theta}}}(Ax) + \psi^*_R(x) \mid x \in X \right\}$$

Examples - Applications

Barcode image deblurring:

$$\min \left\{ \frac{1}{2}\|Ax - \hat{y}\|_2^2 + \kappa_R(x) : x \in \mathbb{R}^d \right\}$$

- $\hat{y}$ - blurred and noisy image
- A - blurring matrix
- R - reference distribution (Bernoulli)

Fig. 11. Out of focus image of a QR code.

Fig. 12. Result of applying our method to a processed version of Fig. 11.

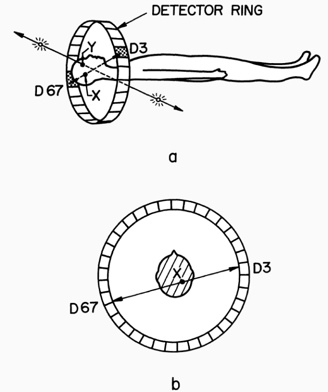
From Rioux et al. (2021)

Examples - Applications

Positron Emission Tomography:

$$\min \left\{ \sum_{i=1}^{m} [\langle a_i, x\rangle \log(\langle a_i, x\rangle / \hat{y}_i) - \langle a_i, x\rangle + \hat{y}_i] + \kappa_R(x) : x \in \mathbb{R}^d_+ \right\}$$

- $\hat{y}$ - measurements vector
- A - experiment model matrix
- R - reference distribution



From Vardi et al. (1985)

Algorithms

Our problems of interest belong to the additive composite model:

$$\min \left\{ f(x) + g(x) \mid x \in \mathbb{R}^d \right\}$$

for $f, g \in \Gamma_0$.

The Bregman proximal gradient algorithm is specified by a *kernel* function h that [Bauschke et al. (2017)]:

- is *smooth adaptable* with respect to f (generalized Lipschitz-convexity condition with constant $L > 0$)
- induces a well-defined and computationally tractable *Bregman proximal operator* with respect to g

Definition (Bregman Proximal Operator)

Let $g, h \colon \mathbb{R}^d \to (-\infty, +\infty]$ such that g is proper and closed, and h is Legendre type. Then for $\bar{x} \in \operatorname{int}(\operatorname{dom} h)$ we define the **Bregman proximal operator** to be

$$\operatorname{prox}_g^h(\bar{x}) := \operatorname{argmin} \left\{ g(x) + D_h(x, \bar{x}) \mid x \in \mathbb{R}^d \right\}$$

Bregman Proximal Gradient Algorithm

Algorithm 1: Bregman Proximal Gradient (BPG) Method

Input: Set $t \in (0, 1/L]$ and $x^0 \in \mathrm{int}(\mathrm{dom}\, h)$.
for $k = 0, 1, 2, \ldots$ **do**
  $x^{k+1} = \mathrm{prox}_{tg}^{h}(\nabla h^*(\nabla h(x^k) - t\nabla f(x^k)))$;
end

- $h = \frac{1}{2}\|\cdot\|_2^2$ - proximal gradient method
- $h = \frac{1}{2}\|\cdot\|_2^2,\ g = \delta_S$ - gradient projection method
- $h = \frac{1}{2}\|\cdot\|_2^2,\ g = 0$ - gradient descent method

Other variants and methods (acceleration, decomposition) rely on the same operators we derive in this work.

Bregman Proximal Gradient Algorithm

Algorithm 2: Bregman Proximal Gradient (BPG) Method

Input: Set $t \in (0, 1/L]$ and $x^0 \in \mathrm{int}(\mathrm{dom}\, h)$.
for $k = 0, 1, 2, \ldots$ **do**
  $x^{k+1} = \mathrm{prox}_{tg}^{h}(\nabla h^*(\nabla h(x^k) - t\nabla f(x^k)))$;
end

- Sublinear convergence rate (linear with more assumptions)
- Choice of the kernel h can simplify computations

Reference Family	Kernel (h_j)	Constant (L)
Normal	$(1/2)x_j^2$	$\|A\|_2$
Poisson	$x_j \log x_j$	$\|A\|_1$
Gamma ($\beta = 1$)	$-\log x_j$	$\|\hat{y}\|_1$

BPG for Linear Models - Bregman Proximal Operator

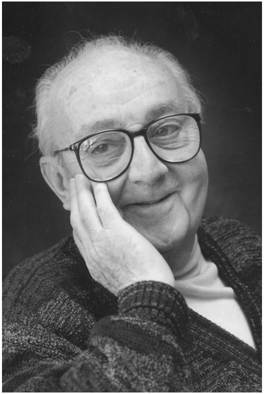
Example ($h = (1/2)\|\cdot\|_2^2$):

Reference Distribution	Proximal Operator
Gamma ($\alpha, \beta \in \mathbb{R}_{++}$)	$x^+ = \left(\bar{x} - t\beta + \sqrt{(\bar{x} - t\beta)^2 + 4t\alpha}\right)/2$
Laplace ($\mu \in \mathbb{R},\ b \in \mathbb{R}_{++}$)	$x^+ = \begin{cases} \mu, & \mu = \bar{x}, \\ \mu + b\rho, & \mu \neq \bar{x}, \end{cases}$ where $\rho \in \mathbb{R}: \quad \alpha_1\rho^3 + \alpha_2\rho^2 + \alpha_3\rho + \alpha_4 = 0,$ with $\alpha_1 = (b/t)^2 b^2,\ \alpha_2 = 2(b/t)^2 b(\mu - \bar{x}),$ $\alpha_3 = (b/t)^2(\mu - \bar{x})^2 - 2(b/t)b - 1,\ \alpha_4 = -2(b/t)(\mu - \bar{x}).$
Poisson ($\lambda \in \mathbb{R}_{++}$)	$x^+ \in \mathbb{R}_+: \ \log(x^+/\lambda) + (x^+ - \bar{x})/t = 0$

Summary

All models are wrong, but some are useful.

George E. P. Box



MEM provides a customizable information-driven framework for performing estimation, and is amenable to efficient first-order solution methods

Thank you!