ROBUST PARAMETER ESTIMATION BASED ON THE K-DIVERGENCE

Yair Sorek and Koby Todros

School of ECE, Ben-Gurion University of the Negev

BACKGROUND – MAXIMUM LIKELIHOOD ESTIMATOR (MLE)

- Minimizes an empirical estimate of the Kullback-Leibler divergence (KLD).
- Asymptotically efficient at the true model.
- Highly sensitive to outliers when the hypothesized score function is unbounded.

ROBUST MLE ALTERNATIVES

- Minimum Helinger distance estimator (MHDE [1]), minimum α, β and γ-divergence estimators (MαDE [2], MβDE [3] and MγDE [4]).
- Mitigate the effect of distant outliers via density power weights.
- MHDE requires consistency of Parzen's non-parametric density estimator.
- MβDE, MγDE and MHDE often necessitate multivariate numerical integration under flexible distributional models.
- Consistency may be lost under the power density transform in the MαDE.

PROPOSED APPROACH

- We propose a new divergence, called $\mathcal{K}$-divergence.
- Applies non-parametric weighting to the hypothesized log-likelihood function.
- Avoids the use of density powers $\Rightarrow$ **Simple implementation.**
- Unlike the MHDE, the resulting estimator **do not require consistency of Parzen's non-parametric density estimator.**

THE $\mathcal{K}$-DIVERGENCE

- The $\mathcal{K}$-divergence between probability distributions G and F with pdf's $g(\cdot)$ and $f(\cdot)$, respectively:

$$\mathcal{K}_h[G||F] \triangleq E\left[\psi_G(\mathbf{x},h)\log\frac{g(x)}{f(x)};G\right] + \log E[\psi_G(\mathbf{x},h);F]$$

where

$$\psi_G(\mathbf{x},h) \triangleq \frac{(K_h * g)(\mathbf{r})}{E[(K_h * g)(\mathbf{x});G]},\quad (K_h * g)(\mathbf{r}) \triangleq \int_{\mathbb{R}^p} K_h(\mathbf{r}-\mathbf{s})g(\mathbf{s})d\lambda(\mathbf{r})$$

and $K_h(\mathbf{r}) \triangleq h^{-p}K(h^{-1}\mathbf{r})$, $K(\mathbf{r})$ is a strictly positive, bounded, integrable and continuous kernel function such that $\int_{\mathbb{R}^p} K(\mathbf{r})d\lambda(\mathbf{r}) = 1$ and $h \in \mathbb{R}_{++}$ is a bandwidth parameter.

Theorem 1 (Non-negativity)

$\mathcal{K}_h[G||F] \geq 0$, when equality holds if and only if $G = F$.

THE MINIMUM $\mathcal{K}$-DIVERGENCE ESTIMATOR (M$\mathcal{K}$DE)

- Consider a parametric family of probability distributions $\{F_{\boldsymbol{\theta}}\}$ and sequence of i.i.d samples $\{\mathbf{x}_n\}_{n=1}^N$ from G, the M$\mathcal{K}$DE

$$\hat{\boldsymbol{\theta}}_h \triangleq \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} J_h(\boldsymbol{\theta})$$

where

$$J_h(\boldsymbol{\theta}) \triangleq \sum_{n=1}^N w(\mathbf{x}_n,h)\log f(\mathbf{x}_n;\boldsymbol{\theta}) - \int_{\mathbb{R}^p}\hat{g}(\mathbf{r};h)f(\mathbf{r};\boldsymbol{\theta})d\lambda(\mathbf{r}),$$

$\hat{g}(\mathbf{r};h) \triangleq N^{-1}\sum_{n=1}^N K_h(\mathbf{r}-\mathbf{x}_n)$ - Parzen's kernel density estimator,

$w(\mathbf{r};h) \triangleq \tilde{g}(\mathbf{r};h)/\sum_{n=1}^N \tilde{g}(\mathbf{x}_n;h)$ and $\tilde{g}(\mathbf{r};h) \triangleq \hat{g}(\mathbf{r};h) - N^{-1}K_h(\mathbf{0})$.

THE MINIMUM $\mathcal{K}$-DIVERGENCE ESTIMATOR (M$\mathcal{K}$DE)

Remark

- $\hat{\boldsymbol{\theta}}_h \to$ MLE as $h \to \infty$.
- The integral term comprising $J_h(\boldsymbol{\theta})$ has analytical solution whenever the convolution $(K_h * f)(\mathbf{r};\boldsymbol{\theta})$ can be computed, e.g., when the assumed distribution is GMM and $K_h(\boldsymbol{r})$ is Gaussian.

ASYMPTOTIC PERFORMANCE ANALYSIS

Theorem 2 (Consistency)

Under some regularity conditions

$$\hat{\boldsymbol{\theta}}_h \xrightarrow[N\to\infty]{p} \boldsymbol{\theta}_h^*$$

where $\boldsymbol{\theta}_h^* \triangleq \arg\min_{\boldsymbol{\theta}} \mathcal{K}_h[G||F_{\boldsymbol{\theta}}]$

Conclusion (Consistency at the true model)

When $G = F_{\boldsymbol{\theta}_0}$ and $\{F_{\boldsymbol{\theta}}\}$ is identifiable:

$$\hat{\boldsymbol{\theta}}_h \xrightarrow[N\to\infty]{p} \boldsymbol{\theta}_0$$

Theorem 3 (Asymptotic normality and unbiasedness)

Under some regularity conditions

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h^*\right) \xrightarrow[N\to\infty]{d} N\left(\mathbf{0}, \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_h^*, h\right)\right)$$

Conclusion (Asymptotic efficiency at the true model)

When $G = F_{\boldsymbol{\theta}_0}$, the asymptotic MSE $\to$ Cramér-Rao lower bound as $h \to \infty$.

BANDWIDTH PARAMETER SELECTION

- A lower bound on the asymptotic weighted MSE:

$$E\left[\left\|\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0\right\|_{\mathbf{W}}^2;G\right] \overset{a}{\approx} \mathrm{tr}\left[\mathbf{R}\left(\boldsymbol{\theta}_h^*,h\right)\mathbf{W}\right] + \left\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_h^*\right\|_{\mathbf{W}}^2 \geq \mathrm{tr}\left[\mathbf{R}\left(\boldsymbol{\theta}_h^*,h\right)\mathbf{W}\right]$$

where $\mathbf{R}(\boldsymbol{\theta}_h^*,h) = N^{-1}\boldsymbol{\Sigma}(\boldsymbol{\theta}_h^*,h)$ and $\mathbf{W}$ is a positive-semidefinite weight matrix.

- The optimal bandwidth parameter minimizes an empirical estimate of the lower bound obtained from the data sample $\{\mathbf{x}_n\}_{n=1}^N$

$$h_{\text{opt}} = \arg\min_{h\in I} \mathrm{tr}\left[\hat{\mathbf{R}}\left(\hat{\boldsymbol{\theta}}_h,h\right)\mathbf{W}\right]$$

EXAMPLE-SETTINGS

- We consider Huber's ε-contamination model:

$$G = (1-\varepsilon)F_{\boldsymbol{\theta}_0} + \varepsilon Q$$

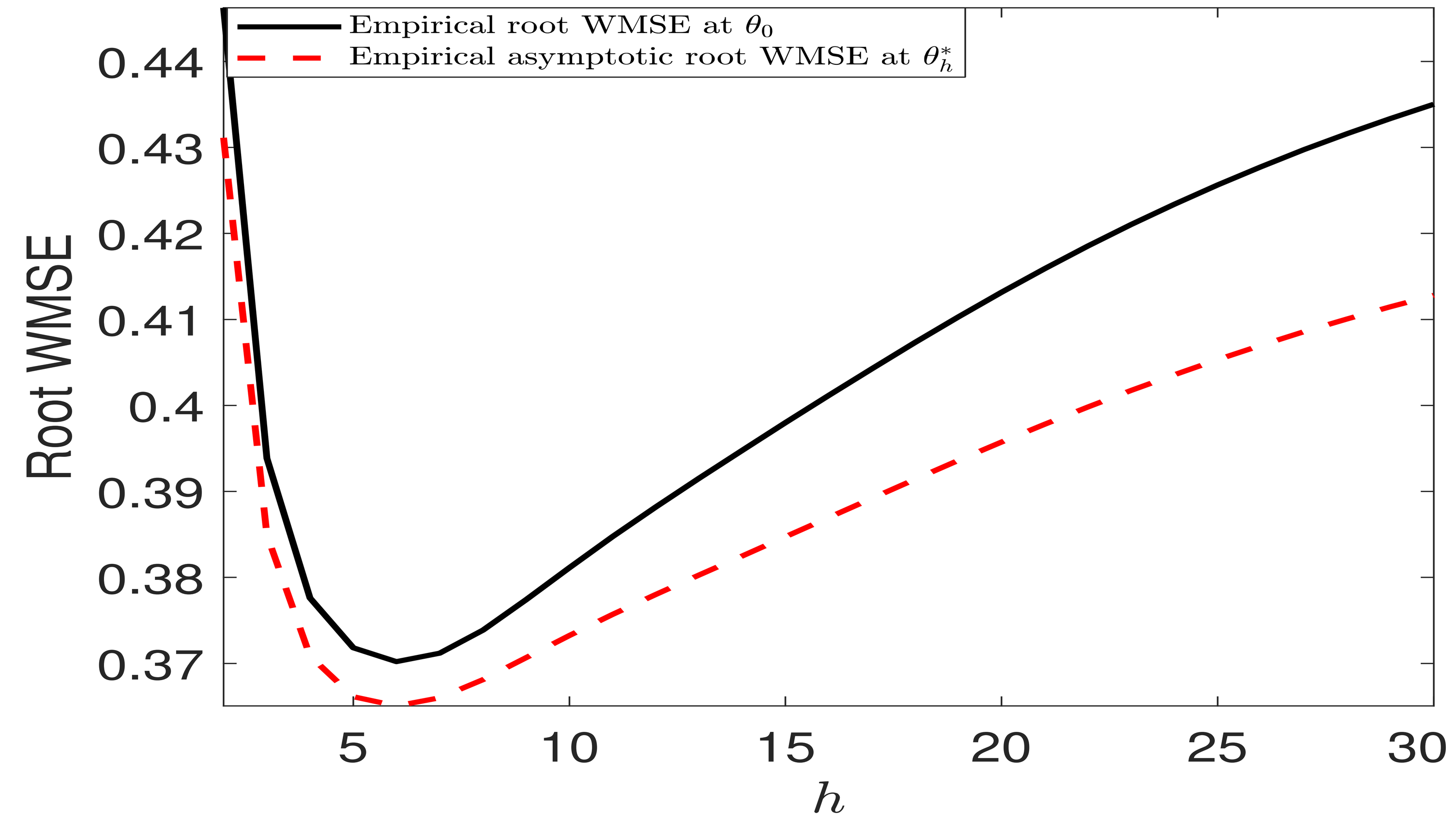
where Q is a contaminating distribution, selected here to be Gaussian.

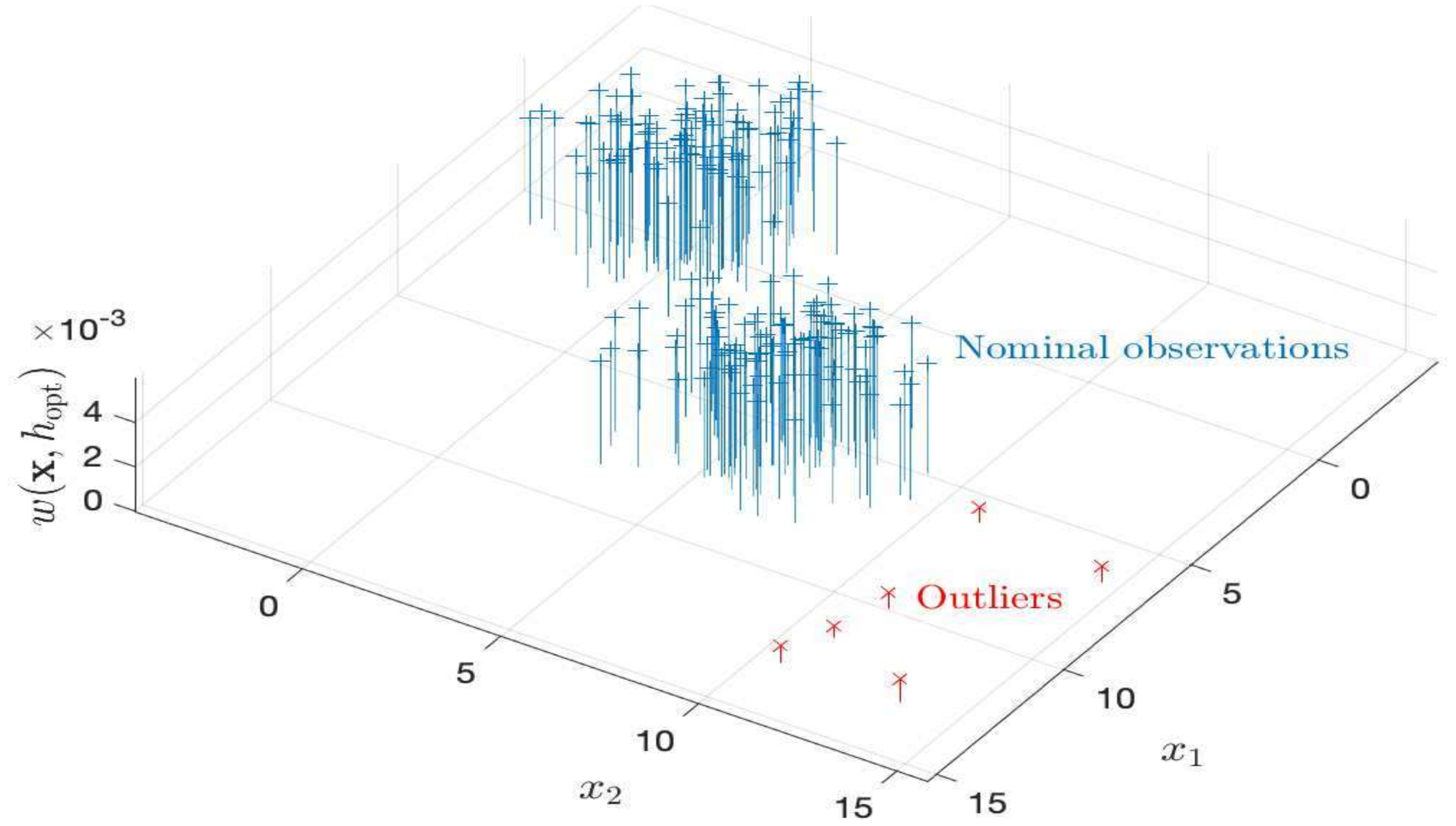
- The p.d.f of the nominal distribution $F_{\boldsymbol{\theta}}$:

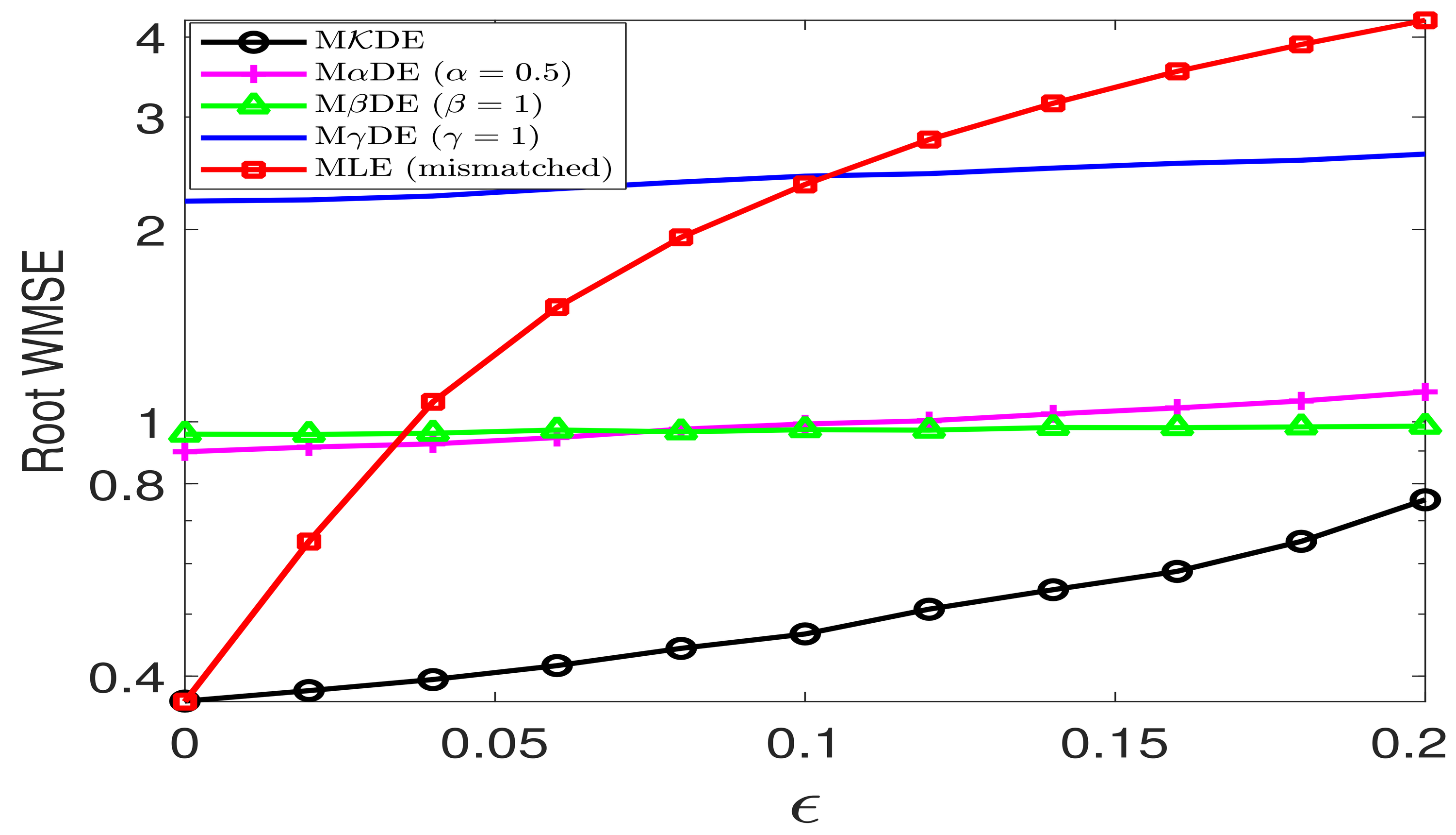
$$f(\mathbf{r};\boldsymbol{\theta}) \triangleq \frac{1}{2}\phi\left(\mathbf{r};\mathbf{0},\sigma^2\mathbf{I}_p\right) + \frac{1}{2}\phi\left(\mathbf{r};\boldsymbol{\eta},\sigma^2\mathbf{I}_p\right)$$

where $\boldsymbol{\theta} \triangleq [\boldsymbol{\eta}^T,\sigma^2]^T$, $\boldsymbol{\eta}$ is the parameter of interest and σ^2 is a nuisance parameter.

EXAMPLE-RESULTS







REFERENCES

1. R. Beran, “Minimum Helinger distance estimates for parametric models,” *The Annals of statistics*, vol. 5, no. 3, pp. 445-463, 1977.
2. A. Iqbal and A-K. Seghouane, “An α-divergence-based approach for robust dictionary learning,” *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5729-5739, 2019.
3. A. Basu, I. R. Harris, N. L. Hjort and M. C. Jones, “Robust and efficient estimation by minimising a density power divergence,” *Biometrika*, vol. 85, no. 3, pp. 549-559, 1998.
4. H. Fujisawa and S. Eguchi, “Robust parameter estimation with a small bias against heavy contamination,” *Journal of Multivariate Analysis*, vol. 99, no. 9, pp. 2053-2081, 2008.