ROBUST ESTIMATION AND REGRESSION WITH MMD

PIERRE ALQUIER

Classification AMS 2020: 62F10, 62F35, 62J02, 68T05, 46E22.

Keywords: universal estimation; robust statistics; kernel methods; minimum distance estimation.

Popular estimation methods in statistics, such as the maximum likelihood estimator (MLE), or the method of moments, require strong assumptions on the statistical model and the data-generating process to converge. When these conditions are not met, these estimators can become very unstable. We are interested in universal estimators, that would converge without assumptions on the model. For example, [15] proved that so-called minimum distance estimators converge under far more general assumptions than the MLE. More recently, the ρ-estimators defined in [4] are not only convergent, but minimax-optimal, in a very large class of models. However, there are still a few limitating assumptions in [15, 4], and the computation of these estimators might be difficult in practice.

This talk will summarize a recent line of work on a variant of minimum distance estimators based on the so-called maximum mean discrepancy (MMD). In the case of i.i.d. data, these estimators converge without *any* assumption on the model nor on the data generating process. Moreover, relatively efficient algorithms are available to compute these estimators.

Let $X_1, \dots, X_n$ be $\mathcal{X}$-valued random variables i.i.d. from a probability distribution P^0. Let $\mathcal{M} = (P_\theta, \theta \in \Theta)$ be a statistial model. Note that we don't assume $P^0 \in \mathcal{M}$. Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) equipped with a scalar product $\langle \cdot, \cdot \rangle$, its associated norm $\| \cdot \|$ and a kernel k: there is a map $\varphi : \mathcal{X} \to \mathcal{H}$ with $k(x,y) = \langle \varphi(x), \varphi(y) \rangle$. The kernel mean embedding (KME) is defined for any probability P on $\mathcal{X}$ by $\mu(P) = \mathbb{E}_{X \sim P}[\varphi(X)]$. We refer the reader to [11] for more details on this construction. In particular: if k is *bounded*, then the KME is indeed well-defined for any P; if k is *characteristic* (see [11] for a definition), μ is one-to-one. Thus, if k is both bounded and characteristic, $D_k(P, P') = \|\mu(P) - \mu(P')\|$ defines a metric on probabilities over $\mathcal{X}$. Finally, [11] provides examples of kernels that are indeed bounded and characteristic: for example, when $\mathcal{X} = \mathbb{R}^d$, the Gaussian kernel $k(x,y) = \exp(-\|x-y\|^2/\gamma)$.

Definition 0.1. *We define the MMD-minimum distance estimator, or MMD-MDE, as*

$$\hat{\theta} = \arg\min_{\theta \in \Theta} D_k(P_\theta, \hat{P}_n)$$

where $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ *is the empirical distribution of the sample* $X_1, \dots, X_n$.

In the talk, I will prove the following result:

Theorem 0.2 (Theorem 3.1 in [6]). *Assume $k \leq 1$, then*

$$\mathbb{E}\big[D_k(P_{\hat{\theta}}, P^0)\big] \leq \inf_{\theta\in\Theta} D_k(P_\theta, P^0) + \frac{2}{\sqrt{n}}.$$

The fact that there are no assumptions on P^0 nor on $\mathcal{M} = (P_\theta, \theta \in \Theta)$ in the theorem is not due to lack of space: this is actually a nice feature of $\hat{\theta}$! Variants of Theorem 0.2, including a bound that holds with large probability rather than in expectation, can be found in [5, 6].

Various topics on $\hat{\theta}$ will be covered in this talk:

- the MMD distance can be written in terms of expectations, and thus, the minimization in Definition 0.1 can be done with a stochastic gradient method. Details are discussed in [5, 7].
- Theorem 1 leads to convergence: if the model is well-specified, $P^0 \in \mathcal{M}$, then $\mathbb{E}[D_k(P_{\hat{\theta}}, P^0)] \leq 2/\sqrt{n} \to 0$ when $n \to \infty$. It also leads to robustness as defined by Huber: if $P^0 = (1-\varepsilon)P_{\theta^0} + \varepsilon Q$ for an arbitrary contamination Q and a small ε, then $\mathbb{E}[D_k(P_{\hat{\theta}}, P^0)] \leq 4\varepsilon + 2/\sqrt{n}$. This is proven in [7], where we also prove more difficult robustness results under adversarial contamination of the data.
- the asymptotic normality of $\hat{\theta}$ is studied in [5] (this requires assumptions).
- an extention of Theorem 0.2 to non-i.i.d. observations (time series) is provided in [7], under a new mixing condition. We also prove that this mixing condition is less restrictive than the standard β-mixing condition.
- the term $2/\sqrt{n}$ in Theorem 0.2 cannot be improved *in general*. However, under assumptions on the variance of P^0, it can actually be improved: variance-aware versions of Theorem 0.2 are proven in [14].
- we can define Bayesian-flavored estimators by using the MMD to define a pseudo-likelihood. We studied such estimators in [7] and proved their convergence using tools from the PAC-Bayes theory [1]. Other Bayesian-inspired variants of $\hat{\theta}$ are based on sampling [8] and Approximate Bayesian Computation or ABC [10].
- this estimation strategy was successfully implemented in a wide range of applications: generative artificial intelligence [9], quantisation and clustering [13], estimation of copulas [2], estimation of parameters in stochastic PDEs [5] etc.

A large part of the talk will be dedicated to parametric regression (linear, or not). In this case, the observations are pairs input-output: $X_i = (Z_i, Y_i)$. Theorem 0.2 guarantees that we can estimate the joint distribution of these pairs. But this is not relevant in regression, where the objective is rather the estimation of the conditional distribution of Y_i given Z_i.

The problem is that, while the theory of KME looks simple and elegant, a rigorous definition of conditional KME turns out to be far more difficult and cumbersome! We refer the reader to [12] for a recent account and a general approach to solve the problem.

In our recent paper [3], we proved that, in standard regression models: linear regression with Gaussian noise, logistic regression, Poisson regression, etc., the conditions for the existence of conditional KMEs are met. This can be used to define consistent and robust estimation of the regression parameters in the spirit of

Definition 0.1 above. These estimators turn out to perform extremely when compared to existing robust regression procedures.

REFERENCES

[1] Pierre Alquier. User-friendly Introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2), 174–203, 2024.

[2] Pierre Alquier, Badr-Eddine Chérief-Abdellatif, Alexis Derumigny and Jean-David Fermanian. Estimation of copulas via maximum mean discrepancy. *Journal of the American Statistical Association*, 118(543), 1997-2012, 2023.

[3] Pierre Alquier and Mathieu Gerber. Universal robust regression via maximum mean discrepancy. *Biometrika*, 111(1), 71–92, 2024.

[4] Yannick Baraud, Lucien Birgé and Mathieu Sart. A new method for estimation and model selection: ρ-estimation. *Inventiones mathematicae*, 207(2), 425–217, 2017.

[5] Francois-Xavier Briol, Alessandro Barp, Andrew B. Duncan, and Mark Girolami. *Statistical inference for generative models with maximum mean discrepancy*, arXiv preprint arXiv:1906.05944, 2019.

[6] Badr-Eddine Chérief-Abdellatif and Pierre Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. *Symposium on Advances in Approximate Bayesian Inference*, Proceedings in Machine Learning Research 118, 1–21, 2020.

[7] Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence. *Bernoulli*, 28(1), 181–213, 2022.

[8] Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas and Francois-Xavier Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. *International Conference on Artificial Intelligence and Statistics*, Proceedings in Machine Learning Research 151, 943–970, 2022.

[9] Gintare Karolina Dzuigaite, Daniel Roy and Zoubin Ghahramani. *Training generative neural networks via maximum mean discrepancy optimization*, arXiv preprint arXiv:1505.03906, 2015.

[10] Sirio Legramanti, Daniele Durante, and Pierre Alquier. *Concentration and robustness of discrepancy-based ABC via Rademacher complexity*, arXiv preprint arXiv:2206.06991, 2022.

[11] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur and Bernhard Schölkopf. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends® in Machine Learning*, 10(1–2), 1–141, 2017.

[12] Junhyung Park and Krikamol Muandet A measure-theoretic approach to kernel conditional mean embeddings. *Advances in neural information processing systems* 33, 21247–21259, 2020.

[13] Onur Teymur, Jackson Gorham, Marina Riabiz, and Chris Oates. Optimal quantisation of probability measures using maximum mean discrepancy. *International Conference on Artificial Intelligence and Statistics*, Proceedings in Machine Learning Research 130, 1027–1035, 2021.

[14] Geoffrey Wolfer and Pierre Alquier. *Variance-aware estimation of kernel mean embedding*, arXiv preprint arXiv:2210.06672, 2022.

[15] Yannis Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13(2), 768–774, 1985.

ESSEC BUSINESS SCHOOL, ASIA-PACIFIC CAMPUS, 5 NEPAL PARK, 139408 SINGAPORE
Email address: alquier@essec.edu