Functional Generalized Empirical Likelihood Estimation for Conditional Moment Restrictions

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Heiner Kremer[1], Jia-Jie Zhu[2], Krikamol Muandet[1], Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany;
[2]Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany;

Motivation

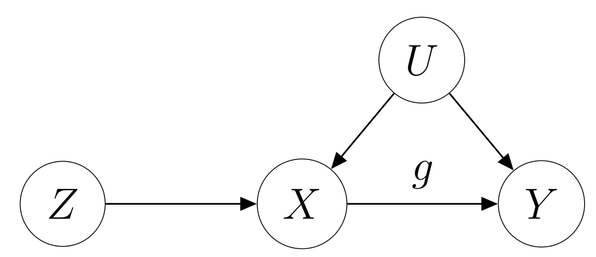
Conditional moment restrictions (CMR) identify a parameter θ_0 via:

$$E[\psi(X;\theta_0)\,|\,Z]=0 \quad P_Z\text{-a.s.,} \tag{1}$$

with $\psi:\mathcal{X}\times\Theta\to\mathbb{R}^n$ being an integrable function.

Examples:

- Instrumental variable regression [1]
- Off-policy evaluation in RL [2]
- Double/Debiased ML [4]



Equivalent unconditional moment restrictions:

$$E[\psi(X;\theta_0)^\top h(Z)]=0 \quad \forall h\in\mathcal{H} \tag{2}$$

⇒ Requires methods which can handle continua of moment restrictions

Method of Moments

Moment restrictions identify a parameter $\theta_0\in\Theta$ uniquely via:

$$E[\psi(X;\theta_0)]=0,$$

where $\psi:\mathcal{X}\times\Theta\to\mathbb{R}^m$.

Empirical counterpart:

$$E_{\hat{P}_n}[\psi(X;\theta)]=0, \quad \theta\in\Theta\subseteq\mathbb{R}^p, \tag{3}$$

where $\hat{P}_n=\sum_{i=1}^n\frac{1}{n}\delta_{x_i}$ is the empirical distribution.

In the over-identified case ($m\gg p$) it is generally impossible to fulfill all moment restrictions exactly → Constraints (3) need to be relaxed

Generalized Method of Moments (GMM)

The GMM relaxes the constraint (3) into a minimization of a quadratic form,

$$\theta^{\text{OWGMM}}=\operatorname*{argmin}_{\theta\in\Theta}\; E_{\hat{P}_n}[\psi(X;\theta)]^\top\left(\widehat{\Omega}_{\tilde\theta}\right)^{-1}E_{\hat{P}_n}[\psi(X;\theta)]. \tag{4}$$

- 2-step procedure:
  1. Compute initial parameter estimate $\tilde\theta$ to compute $\widehat{\Omega}_{\tilde\theta}=E_{\hat{P}_n}[\psi(X;\tilde\theta)\psi(X;\tilde\theta)^\top]$
  2. Optimize (4) using $\widehat{\Omega}_{\tilde\theta}$
- Multiple generalizations to continuum moment restrictions / CMR [1, 3, 6]

Generalized Empirical Likelihood (GEL)

The GEL relaxes the restrictions (3) by requiring $E_P[\psi(X;\theta)]=0$ to be fulfilled exactly but allowing the distribution P to deviate from the empirical distribution $\hat{P}_n$.

The GEL estimator for θ minimizes the *profile divergence*,

$$R(\theta)=\min_{P\ll\hat{P}_n} D_f(P||\hat{P}_n) \quad \text{s.t.} \quad E_P[\psi(X;\theta)]=0, \quad E_P[1]=1.$$

$$\theta^{\text{GEL}}=\operatorname*{argmin}_{\theta\in\Theta} R(\theta)$$

where $D_f(P||Q)=\int f\left(\frac{dP}{dQ}\right)dQ$ is the f-divergence between distributions P and Q.

- Asymptotically equivalent to GMM (contains GMM as special case)
- Improved small sample properties especially in the case $m\gg p$ [7]

Functional GEL

For a CMR of the form (1), a profile divergence can be defined as

$$R(\theta):=\min_{P\in\mathcal{P}} D_f(P\,\|\,\hat{P}_n) \quad \text{s.t.} \quad E_P[\psi(X;\theta)\,|\,Z]=0, \quad P_Z\text{-a.s.,}$$

with $\mathcal{P}:=\{P\ll\hat{P}_n : E_P[1]=1\}$.

Let $\mathcal{H}$ be a sufficiently large Hilbert space of functions such that

$$E[\psi(X;\theta_0)\,|\,Z]=0 \quad P_Z\text{-a.s.} \quad\Longleftrightarrow\quad E[\psi(X;\theta_0)^\top h(Z)]=0 \quad \forall h\in\mathcal{H}. \tag{5}$$

Define the *moment functional*, a statistical functional $H(X,Z;\theta)\in\mathcal{H}^*$, as

$$\begin{aligned} H(X,Z;\theta):\ \ &\mathcal{H}\to\mathbb{R} \\ &h\mapsto H(X,Z;\theta)(h)=\psi(X;\theta)^\top h(Z). \end{aligned}$$

Then, the computation of the profile likelihood can be written as a *functionally constrained* optimization problem

$$R(\theta)=\inf_{P\in\mathcal{P}} D_f(P||\hat{P}_n) \quad \text{s.t.} \quad \|E_P[H(X,Z;\theta_0)]\|_{\mathcal{H}^*}=0.$$

Relax the problem to restore strong duality:

$$R_\lambda(\theta):=\inf_{P\in\mathcal{P}} D_f(P||\hat{P}_n) \quad \text{s.t.} \quad \|E_P[H(X,Z;\theta)]\|_{\mathcal{H}^*}\le\lambda.$$

Motivate FGEL estimator from the exact dual formulation:

$$R_\lambda(\theta)=\sup_{\substack{h\in\mathcal{H}\\ \mu\in\mathbb{R}}}\mu-\frac{1}{n}\sum_{i=1}^n f^*(\mu+H(x_i,z_i;\theta)(h))-\lambda\|h\|_{\mathcal{H}},$$

where $f^*(v)=\sup_{p\in\mathbb{R}^n}\langle v,p\rangle-f(p)$.

FGEL estimation

Let $V\subseteq\mathbb{R}$ be an open interval containing zero and $\phi:V\to\mathbb{R}$ be a twice differentiable concave function with first and second derivatives $\phi_1(0)\neq 0$ and $\phi_2(0)<0$. Then we define the empirical FGEL objective $G:\Theta\times\widehat{\mathcal{H}}_\theta\to\mathbb{R}$ as

$$G_{\lambda_n}(\theta,h):=\frac{1}{n}\sum_{i=1}^n\phi\left(H(x_i,z_i;\theta)(h)\right)-\frac{\lambda_n}{2}\|h\|^2_{\mathcal{H}},$$

where $H(x_i,z_i;\theta)(h)=\psi(x_i;\theta)^\top h(z_i)$ and $\widehat{\mathcal{H}}_\theta:=\{h\in\mathcal{H}:\ \psi(x_i;\theta)^\top h(z_i)\in\operatorname{dom}(\phi),\ 1\le i\le n\}$. The FGEL estimate $\hat\theta$ of θ_0 is then given by

$$\hat\theta=\operatorname*{argmin}_{\theta\in\Theta}\ \sup_{h\in\widehat{\mathcal{H}}_\theta} G_{\lambda_n}(\theta,h).$$

- Allows leveraging arbitrary ML models as instrument functions h
- Divergence functions beyond the Cressie-Reed family, in particular $\neq\chi^2$ ($\hat{=}$ GMM)
- Can benefit from recent progress in saddle point optimization (e.g. [5])

Asymptotic properties

Let $\lambda_n=O_p(n^{-\xi})$, then under several technical assumptions we have as $n\to\infty$:

- Consistency:

$$\hat\theta\xrightarrow{p}\theta_0 \quad\text{and}\quad \|E[H(X,Z;\hat\theta)]\|_{\mathcal{H}^*}=O_p(n^{-1/2+\xi})$$

- Asymptotic normality:

$$\sqrt{n}(\hat\theta-\theta_0)\xrightarrow{d}N(0,\Sigma_\theta), \quad \Sigma_\theta=((\nabla_\theta H^*)\Omega^{-1}(\nabla_{\theta^\top}H))^{-1},$$

where $\widehat{\Omega}_{\lambda_n}:=E_{\hat{P}_n}[H(X,Z,\theta_0)H(X,Z,\theta_0)^*]+\lambda_n I\otimes I\xrightarrow{p}\Omega$

Choice of Divergence and Instrument Function

Choice of Divergence

	$f(p)$	$\phi(v)$	$\operatorname{dom}(\phi)$
χ^2	$\frac{1}{2}(p-1)^2$	$-\frac{1}{2}(1+v)^2$	$\mathbb{R}$
Burg	$-\log(p)$	$-\log(1-v)$	$\left(-\infty,1-\frac{1}{n}\right]$
KL	$p\log(p)$	$-e^v$	$\mathbb{R}$

- Contains continuous updating version of VMM [1] as special case ($f=\chi^2$)
- Continuum generalizations of the original EL (Burg) and exponential tilting estimators (KL)

Choice of Instrument Function Class

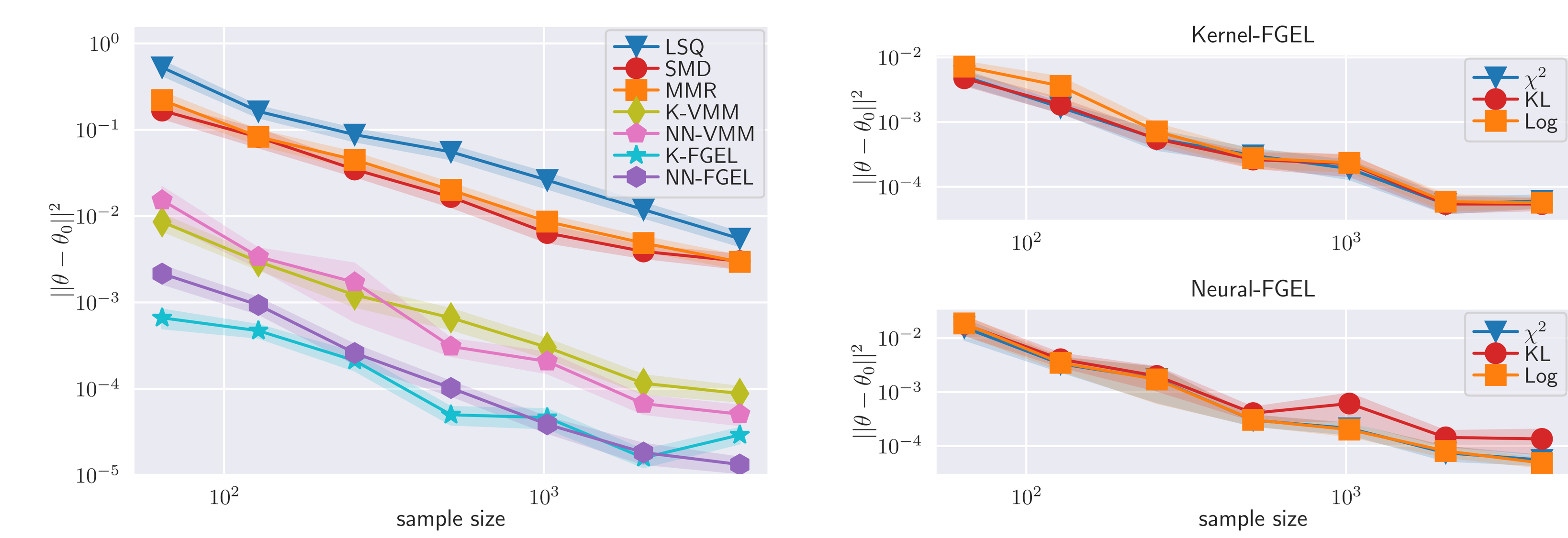
- Kernel-FGEL: $G_{\lambda_n}(\theta,\alpha)=\frac{1}{n}\sum_{i=1}^n\phi\left(\sum_{r=1}^m(\alpha_r^\top K_r)_i\psi_r(x_i;\theta)\right)-\frac{\lambda_n}{2}\sum_{r=1}^m\alpha_r^\top K_r\alpha_r$
  - Inner optimization over α convex → Solve with e.g. 2-layer LBFGS
  - Provably fulfills equivalence relation (5)
- Neural-FGEL: $G_{\lambda_n}(\theta,\omega):=\frac{1}{n}\sum_{i=1}^n\phi\left(\psi(x_i;\theta)^\top h_\omega(z_i)\right)-\frac{\lambda_n}{2n}\sum_{i=1}^n\|h_\omega(z_i)\|^2_{\mathbb{R}^m}$
  - Non-convex saddle point problem → Solve with optimistic Adam
  - Strong empirical performance and superior scaling due to mini-batch training

Experiments

Regression under heteroskedastic noise:

$$y=x^\top\theta+\varepsilon, \quad x\sim\text{Uniform}([-1.5,1.5]), \quad \varepsilon|x\sim\mathcal{N}(0,\sigma=5x^2)$$

Conditional moment restriction: $E[Y-X^\top\theta|X]=0 \quad P_X$-a.s.



Discussion

- Many problems in ML can naturally be expressed as risk minimizations
- (Conditional) moment restrictions appear in emerging areas such as causal inference and robust ML and require dedicated solution methods
- We extended the powerful GEL framework to CMR and proved its asymptotics

References

[1] A. Bennett and N. Kallus. The variational method of moments, 2020.
[2] A. Bennett, N. Kallus, L. Li, and A. Mousavi. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pages 1999–2007. PMLR, 2021.
[3] M. Carrasco and J.-P. Florens. Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, 16, 2000.
[4] V. Chernozhukov et al. Double/debiased machine learning for treatment and structural parameters, 2018.
[5] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism, 2018.
[6] G. Lewis and V. Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.
[7] W. K. Newey and R. J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72, 2004.