Sparse Unbiased Estimating Equations via Likelihood Score Alignment

Learning informative estimating equations through adaptive penalization

Giulia Bertagnolli, Z. Huang and D. Ferrari

Free University of Bozen-Bolzano, Italy

2026-06-27

Motivation

- Why another sparse estimator?
- Can sparsity be imposed on the estimating equations instead of directly on the parameters?

unibz

Estimating Equations

Estimating Equations (EE) are everywhere:

- likelihood score
- GEE
- composite likelihood
- M-estimation and Weighted Likelihood

$$\sum_{i=1}^{n} S(\theta, Y^{(i)}) = 0$$

Some notation and conditions:

- $Y^{(1)}, \dots, Y^{(n)}$ i.i.d. with density $f(\cdot; \theta)$.
- $\theta \in \Theta \subseteq \mathbb{R}^p$
- unbiasedness $\mathbb{E}(S(\theta_0, Y)) = 0$, e.g. $S = U = \nabla_\theta \log f(y; \theta)$
- variability matrix: $\Sigma_S = \mathrm{Var}(S(\theta_0))$
- sensitivity matrix $H_S = -\mathbb{E}(\nabla S(\theta_0))$

unibz

Optimal estimating equations

- General theory (Godambe 1960; Heyde 1997)
- $p < n$: well-posed problem, asymptotically efficient estimator…
- $p > n$ or $p \gg n$: ill-conditioned estimating system

Sparsity: existence of an active set $A_0 := \{j : \theta_{0j} \neq 0\}$ with cardinality $p_0 \ll p$

Typically,

- regularise of the optimisation problem, e.g. imposing $\|\theta\|_q \leq t$
- yields corresponding penalised EE: $0 = S_n(\theta) - \nabla P_\lambda(\theta)$
- sparsity prior on the parameter space induces bias in the EE

unibz

Which estimating equations are actually useful? I

Can we automatically remove redundant estimating equations?

(Heyde 1997):

- given a set of (zero mean, square integrable) estimating functions $\mathcal{H}$, an *optimal estimating equation within this class* is the orthogonal projection of the score function U onto the chosen space $\mathcal{H}$;
- maximum correlation with the generally unknown score function.

unibz

Which estimating equations are actually useful? II

Our formulation:

- We consider the class of linear transformations of $S(\theta; Y) =: S(\theta)$ i.e.

$$WS(\theta) \text{ with } W \in \mathbb{R}^{p \times p}$$

- minimise the objective

$$\begin{aligned} &\frac{1}{2}\mathbb{E}\left\|U\left(\theta_0\right) - WS\left(\theta_0\right)\right\|^2 \\ \Leftrightarrow &\frac{1}{2}\operatorname{tr}\left(W\Sigma_S W^\top\right) - \operatorname{tr}\left(H_S W^\top\right), \end{aligned}$$

  subject to a sparsity condition on W.

Which estimating equations are actually useful? III

- Sparsity: only a small subset of columns $W_{\cdot k}$ are non-zero,
- we select only p_0 of the the original p estimating functions

$$\min_{W} \frac{1}{2}\mathbb{E}\left\|U\left(\theta_0\right) - WS\left(\theta_0\right)\right\|^2 + \lambda \sum_{k=1}^{p} \frac{\left\|W_{\cdot k}\right\|_2}{\left|\theta_{0k}\right|^2} \tag{1}$$

with $\lambda \geq 0$.

Which estimating equations are actually useful? IV

Each column of W corresponds to one estimating equation, not one parameter.

$$W = \begin{pmatrix} w_{11} & \dots & 0 & \dots \\ w_{21} & \dots & 0 & \dots \\ \vdots & & \vdots & \\ w_{p1} & \dots & 0 & \dots \end{pmatrix}$$

The optimisation problem is solved alternating between estimation of W and θ.

unibz

Theoretical Guarantees I

- Existence and uniqueness (sub-gradient equations)
- Oracle property (non-zero columns of W correspond to indices in the active set)
- Selection consistency
- Consistency and asymptotic normality on the empirical active set

$$P\left(\hat{A} = A_0\right) \to 1$$
$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \Rightarrow N\left(0, G^{-1}\right).$$

unibz

Example: Linear Regression I

Gaussian linear model: $Y^{(i)} = x^{(i)\top}\theta + \varepsilon^{(i)}$, $\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$,with $x^{(i)} \in \mathbb{R}^p$ and design matrix $X = (x^{(1)}, \ldots, x^{(n)})^\top \in \mathbb{R}^{n \times p}$. The score-based estimating function is

$$S^{(i)}(\theta) = \frac{1}{\sigma^2} x^{(i)} \{Y^{(i)} - x^{(i)\top}\theta\}$$

$$H_S(\theta) = -\mathbb{E}\left\{\frac{\partial S^{(i)}(\theta)}{\partial \theta^\top}\right\} = \frac{1}{n\sigma^2} X^\top X, \qquad \widehat{\Sigma}_S(\theta) = \mathrm{Var}\{S^{(i)}(\theta)\} = \frac{1}{n\sigma^4} X^\top r r^\top X,$$

with $r = Y - X\theta$.

Example: Linear Regression II

- Blockwise proximal scoring algorithm with scoring step only on selected coordinates:

$$\theta^{(t+1)}_{A^{(t)}} = \theta^{(t)}_{A^{(t)}} + \left\{ X^\top_{A^{(t)}} X_{A^{(t)}} \right\}^{-1} X^\top_{A^{(t)}} r^{(t)}$$

and W-update

$$W^{(t+1)}_{\cdot,j} = \left(\frac{1}{n\sigma^4} x_j^\top r^{(t+1)} r^{(t+1)\top} x_j \right)^{-1} \left(1 - \frac{\lambda \gamma_j}{2\|R_j\|_2} \right)_+ R_j,$$

where the norm of the partial residual $\|R_j\|_2$ measures how strongly the score direction associated with predictor x_j is aligned with the current residual $r = Y - X\theta$ after removing the contribution explained by other predictors.

Numerical Experiments: LR I

- fixed sample size $n = 200$
- varying dimension $p \in 25, 50, 500, 1000$ and sparsity levels (90% and 99%)
- $\lambda = 5\sqrt{\frac{\log p}{n}}$

Table 1: Monte Carlo estimates for the support recovery performance metrics across simulation scenarios based on $B = 100$ Monte Carlo replications.

p	p_0	SN	SP	AC
25	1	1.00	1.00	1.00
50	1	1.00	1.00	1.00
500	5	1.00	1.00	1.00
1000	10	0.90	1.00	0.999

unibz

Numerical Experiments: LR II

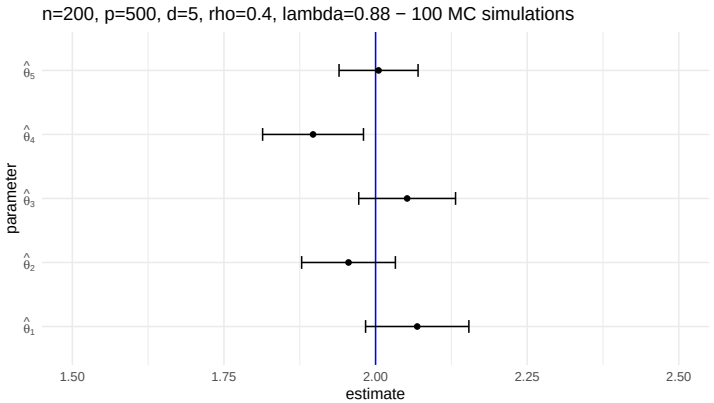

Figure 1: Monte Carlo Estimates (same setting). The blue vertical line represents the true value $\theta_k = 2$.

Numerical Experiments: LR III

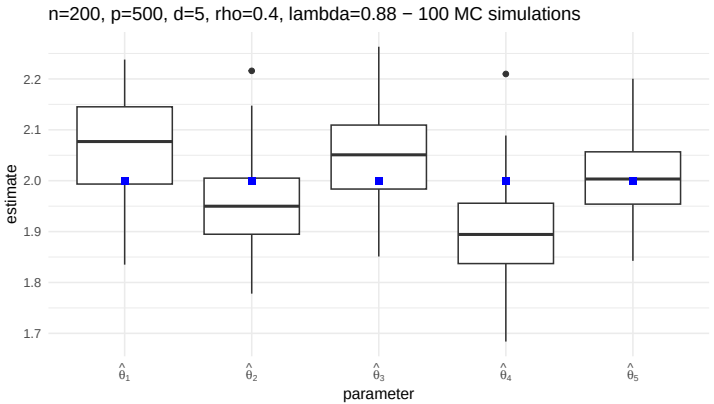

Figure 2: Monte Carlo Estimates (same setting). Boxplots. Blue dots represents the true value $\theta_k = 2$.

Numerical Experiments: LR IV

Main observations:

- Accurate support recovery
- Robust with respect to λ
- Works in high dimensions
- General framework

unibz

Beyond linear regression

- weighted scores (WLE), yielding an iteratively reweighted version of the proposed algorithm
- precision/covariance estimation
- multinomial model

unibz

Conclusions

1. Sparsity is imposed on estimating equations.
2. The method is model-agnostic.
3. Theory supports selection consistency and asymptotic inference.
4. The framework applies to many statistical models.

unibz

Thanks for your attention!

Giulia Bertagnolli

giulia.bertagnolli@unibz.it

gbertagnolli.github.io

unibz

References

Godambe, Vidyadhar P. 1960. “An Optimum Property of Regular Maximum Likelihood Estimation.” *The Annals of Mathematical Statistics* 31 (4): 1208–11.

Heyde, Christopher C. 1997. *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation*. Springer.

unibz