Model selection under SCAD and MCP based on approximate message passing

Institute of Statistical Mathematics Ayaka Sakata
Tokyo Institute of Technology Tomoyuki Obuchi

We study model selection for linear regression problems penalized by nonconvex penalties; SCAD and MCP. The problem is formulated as

$$\min_{x} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + J(\boldsymbol{x}; \lambda, a),$$

where $\boldsymbol{y} \in \mathbb{R}^M$ and $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ denote response variables and predictor matrix, respectively. The regression coefficient to be estimated, $\boldsymbol{x} \in \mathbb{R}^N$, is penalized by $J(\boldsymbol{x}; \lambda, a)$, where λ and a are regularization parameters that control the form of SCAD and MCP. The density of non-zero components in the minimizer depends on the regularization parameters, and hence the determination of these parameters corresponds to the model selection. Hereafter, we denote the minimizer of the problem as $\hat{\boldsymbol{x}}(\boldsymbol{y}; \lambda, a)$.

The quantities to determine the regularization parameters considered here are in-sample error ϵ_{in} and extra-sample error ϵ_{ext} defined by

$$\epsilon_{in}(\boldsymbol{y}; \lambda, a) \equiv \frac{1}{M} E_{\boldsymbol{z}} \left[\sum_{\mu=1}^{M} \left(z_\mu - (\boldsymbol{A}\hat{\boldsymbol{x}}(\boldsymbol{y}; \lambda, a))_\mu \right)^2 \right],$$

$$\epsilon_{ext}(\boldsymbol{y}; \lambda, a) \equiv E_{w, \tilde{\boldsymbol{A}}} \left[\left(w - \tilde{\boldsymbol{A}}\hat{\boldsymbol{x}}(\boldsymbol{y}; \lambda, a) \right)^2 \right],$$

where the statistical property of $\boldsymbol{z}$ in the in-sample error is equivalent to that of response variable $\boldsymbol{y}$. The generative process of variable w is also equivalent to that of the response variables, but the predictor vector $\tilde{\boldsymbol{A}}$, which is not contained in the training of $\hat{\boldsymbol{x}}(\boldsymbol{y}; \lambda, a)$, is assigned to w.

The regularization parameters that achieve the minimum of the in-sample or extra-sample error is appropriate for the description of the response variables. However, the exact computation of these quantities is impossible because we do not know the true generative process of the response variables. We propose numerical methods to construct the estimators of these quantities using approximate message passing.