Penalized high-dimensional empirical likelihood

Chengyong Tang and Chenlei Leng

Recent research on statistical inference gets to a focus on the challenge of high data dimensionality. This is in-line with the growth of the areas resulting abundant data, for instance the high frequency trading in finance, genetics and image processing. When data dimensionality is high, variable selection through regularization has proven to be effective through properly adjusting the bias variance trade-off so that the performance improvement can be achieved.

The main objective of the research is to explore the dimensional effect and regularization on the empirical likelihood approach. Empirical likelihood is a nonparametric statistical instrument and has successful implementations in various areas. It has advantages of being robust and flexible in incorporating multiple source of information.

We propose in our work (Biometrika, volume 97, Issue 4, Page 905-920) a penalized empirical likelihood approach for parameter estimation and variable selection for problems with diverging numbers of parameters. Our results are demonstrated for estimating the mean vector in multivariate analysis and regression coefficients in linear models. By using an appropriate penalty function, we show that penalized empirical likelihood has the oracle property. That is, with probability tending to one, penalized empirical likelihood identifies the true model and estimates the nonzero coefficients as efficiently as if the sparsity of the true model was known in advance.

The following table for multivariate mean vector estimation demonstrates the performance of the proposed approach in variable selection and parameter estimation.

Simulation results for penalized empirical likelihood in estimating the mean vector, in terms of the root mean square errors and model selection. The root mean square errors are multiplied by a factor of 10^3

				Root mean square errors			Zero coefficients	
n	p	ρ		μ_1	μ_2	μ_3	True	False
50	10		$\bar{X}$	197	201	197	–	–
		0·3	$\hat{\mu}$	190	218	243	6·04	0·43
			$\hat{\mu}_{ST}$	281	279	225	4·35	0·34
			$\hat{\mu}_{HT}$	205	324	284	6·06	0·88
			$\hat{\mu}_{QL}$	255	261	225	3·27	0·30
		0·7	$\hat{\mu}$	137	149	175	5·86	0·19
			$\hat{\mu}_{ST}$	296	287	238	4·47	0·33
			$\hat{\mu}_{HT}$	208	308	284	5·73	0·82
			$\hat{\mu}_{QL}$	180	181	175	3·02	0·10
100	20		$\bar{X}$	143	141	142	–	–
		0·3	$\hat{\mu}$	133	147	185	15·62	0·19
			$\hat{\mu}_{ST}$	222	225	202	12·42	0·22
			$\hat{\mu}_{HT}$	142	195	257	15·59	0·68
			$\hat{\mu}_{QL}$	187	189	183	10·98	0·12
		0·7	$\hat{\mu}$	89	95	117	15·67	0·05
			$\hat{\mu}_{ST}$	227	227	206	11·79	0·19
			$\hat{\mu}_{HT}$	143	182	245	14·06	0·56
			$\hat{\mu}_{QL}$	129	131	128	10·23	0·01

$\bar{X}$, the sample mean; $\hat{\mu}$, the penalized empirical likelihood estimate; $\hat{\mu}_{ST}$, the soft-threshold estimator; $\hat{\mu}_{HT}$, the hard-threshold estimator; $\hat{\mu}_{QL}$, the estimator using quadratic loss.

Leng Chenlei

Tang Chengyong