Multivariate, Heteroscedastic Empirical Bayes via Nonparametric Maximum Likelihood

Sen Bodhisattva

Columbia University

Abstract:

Multivariate, heteroscedastic errors complicate statistical inference in many large-scale denoising problems. Empirical Bayes is attractive in such settings, but standard parametric approaches rest on assumptions about the form of the prior distribution which can be hard to justify and which introduce unnecessary tuning parameters. In this talk we extend the nonparametric maximum likelihood estimator (NPMLE) for Gaussian location mixture densities to allow for multivariate, heteroscedastic errors. NPMLEs estimate an arbitrary prior by solving an infinite-dimensional, convex optimization problem; we show that this convex optimization problem can be tractably approximated by a finite-dimensional version. We introduce a dual mixture density whose modes contain the atoms of every NPMLE, and we leverage the dual both to show non-uniqueness in multivariate settings as well as to construct explicit bounds on the support of the NPMLE. The empirical Bayes posterior means based on an NPMLE have low regret, meaning they closely target the oracle posterior means one would compute with the true prior in hand. We prove an oracle inequality implying that the empirical Bayes estimator performs at nearly the optimal level (up to logarithmic factors) for denoising without prior knowledge. We provide finite-sample bounds on the average Hellinger accuracy of an NPMLE for estimating the marginal densities of the observations. We also demonstrate the adaptive and nearly-optimal properties of NPMLEs for deconvolution. We apply the method to two astronomy datasets, constructing a fully data-driven color-magnitude diagram of 1.4 million stars in the Milky Way and investigating the distribution of chemical abundance ratios for 27 thousand stars in the red clump.