ADAPTIVE ESTIMATION OF REGRESSION PARAMETERS FOR THE GAUSSIAN SCALE MIXTURE MODEL

ROGER KOENKER

Abstract. A proposal of van der Vaart (1996) for an adaptive estimator of a location parameter from a family of normal scale mixtures is explored. Recent developments in convex optimization have dramatically improved the computational feasibility of the Kiefer and Wolfowitz (1956) nonparametric maximum likelihood estimator for general mixture models and yield an effective strategy for estimating the efficient score function for the location parameter in this setting. The approach is extended to regression and performance is evaluated with a small simulation experiment.

1. Introduction

The Princeton Robustness Study, Andrews, Bickel, Hampel, Huber, Rogers, and Tukey (1972), arguably the most influential simulation experiment ever conducted in statistics, compared performance of a 68 distinct location estimators focusing almost exclusively scale mixtures of Gaussian models. While such scale mixtures do not constitute an enormous class, see for example Efron and Olshen (1978), they are convenient for several reasons: their symmetry ensures a well-defined location estimand, their unimodality affirms Tukey's dictum that "all distributions are normal in the middle," and probably most significantly, conditional normality facilitates some nice Monte-Carlo tricks that lead to improvements in simulation efficiency.

A prototypical problem is the Tukey contaminated normal location model,

$$Y_i = \alpha + u_i \tag{1}$$

with iid u_i from the contaminated normal distribution, $F_{\epsilon,\sigma}(u) = (1-\epsilon)\Phi(u) + \epsilon\Phi(u/\sigma)$. We would like to estimate the center of symmetry, α, of the distribution of the Y_i's. Yet we do not know ϵ, nor the value of σ; how should we proceed? Of course we could adopt any one of the estimators proposed in the Princeton Study, or one of the multitude of more recent proposals. But we are feeling greedy, and would like to have an estimator that is also asymptotically fully efficient.

Version: January 5, 2016. This research was partially supported by NSF grant SES-11-53548. The paper will appear in J. Beran et al. (eds.), *Empirical Economic and Financial Research, a Festschrift for Siegfried Heiler.*

The Tukey model is a very special case of a more general Gamma mixture model in which we have (1), and the u_i^2's are iid with density,

$$g(v) = \int_0^\infty \gamma(v|\theta) dF(\theta)$$

where $\theta = \sigma^2$, and γ is the $\chi^2(1)$ density with free scale parameter θ,

$$\gamma(v|\theta) = \frac{1}{\Gamma(1/2)\sqrt{2\theta}} v^{-1/2} \exp(-v/(2\theta))$$

Our strategy will be to estimate this mixture model *nonparametrically* and employ it to construct an adaptive M-estimator for α. This strategy may be viewed as an example of the general proposal of van der Vaart (1996) for constructing efficient MLEs for semiparametric models.

2. Empirical Bayes and the Kiefer-Wolfowitz MLE

Given iid observations, $V_1, \cdots, V_n$, from the density,

$$g(v) = \int_0^\infty \gamma(v|\theta) dF(\theta)$$

we can estimate F and hence the density g by maximum likelihood. This was first suggested by Robbins (1951) and then much more explicitly by Kiefer and Wolfowitz (1956). It is an essential piece of the empirical Bayes approach developed by Robbins (1956) and many subsequent authors. The initial approach to computing the Kiefer-Wolfowitz estimator was provided by Laird (1978) employing the EM algorithm, however EM is excruciatingly slow. Fortunately, there is a better approach that exploits recent developments in convex optimization.

The Kiefer-Wolfowitz problem can be reformulated as a convex maximum likelihood problem and solved by standard interior point methods. To accomplish this we define a grid of values, $\{0 < v_1 < \cdots < v_m < \infty\}$, and let $\mathcal{F}$ denote the set of distributions with support contained in the interval, $[v_1, v_m]$. The problem,

$$\max_{f \in \mathcal{F}} \sum_{i=1}^n \log(\sum_{j=1}^m \gamma(V_i, v_j) f_j),$$

can be rewritten as,

$$\min\{-\sum_{i=1}^n \log(g_i) \mid Af = g,\ f \in \mathcal{S}\},$$

where $A = (\gamma(V_i, v_j))$ and $\mathcal{S} = \{s \in \mathbb{R}^m | 1^\top s = 1,\ s \geq 0\}$. So f_j denotes the estimated mixing density estimate $\hat{f}$ at the grid point v_j, and g_i denotes the estimated mixture density estimate, $\hat{g}$, evaluated at V_i.

This is easily recognized as a convex optimization problem with an additively separable convex objective function subject to linear equality and inequality constraints, hence amenable to modern interior point methods of solution. For this purpose, we rely on the Mosek system of Andersen (2010) and its R interface, Friberg (2012). Implementations of all the procedures described here are available in the R package REBayes, Koenker (2012). For further details on computational aspects see Koenker and Mizera (2014).

Given a consistent initial estimate of α, for example as provided by the sample median, the Kiefer-Wolfowitz estimate of the mixing distribution can be used to construct an estimate of the optimal influence function, $\hat{\psi}$, that can be used in turn to produce an asymptotically efficient M-estimator of the location parameter. More explicitly, we define our estimator, $\hat{\alpha}_n$, as follows:

(1) Preliminary estimate: $\tilde{\alpha} = \text{median}(Y_1, \cdots, Y_n)$
(2) Mixture estimate: $\hat{f} = \text{argmax}_{f \in \mathcal{F}} \sum_{i=1}^n \log(\sum_{j=1}^m \gamma(Y_i - \tilde{\alpha}, v_j) f_j)$,
(3) Solve for $\hat{\alpha}$ such that $\hat{\psi}(Y_i - \alpha) = 0$, where $\hat{\psi}(u) = (\log \hat{g}(u))'$, and $\hat{g}(u) = \int \gamma(u, v) d\hat{F}(v)$.

Theorem 1. *(van der Vaart (1996)) For the Gaussian scale mixture model* (1) *with F supported on $[v_1, v_m]$, the estimator $\hat{\alpha}$ is asymptotically efficient, that is, $\sqrt{n}(\hat{\alpha}_n - \alpha) \rightsquigarrow \mathcal{N}(0, 1/\mathcal{I}(g))$, where $\mathcal{I}(g)$ is the Fisher information for location of the density, $g(u) = \int \gamma(u, v) dF(v)$.*

This result depends crucially on the orthogonality of the score function for the location parameter with that of the score of the (nuisance) mixing distribution and relies obviously on the symmetry inherent in the scale mixture model. In this way it is closely related to earlier literature on adaptation by Stein (1956), Stone (1975), Bickel (1982) and others. But it is also much more specialized since it covers a much smaller class of models. The restriction on the domain of $\mathcal{F}$ could presumably be relaxed by letting $v_1 \to 0$ and $v_m \to \infty$ (slowly) as $n \to \infty$. From the argument for the foregoing result in van der Vaart it is clear that the location model can be immediately extended to linear regression which will be considered in the next section.

3. Some Simulation Evidence

To explore the practical benefits of such an estimator we consider two simple simulation settings: the first corresponds to our prototypical Tukey model in which the scale mixture is composed of only two mass points, and the other is a smooth mixture in which scale is generated as $\sqrt{\chi_3^2/3}$, so the Y_i's are marginally Student t on three degrees of freedom. We will consider the simple bivariate linear regression model,

$$Y_i = \beta_0 + x_i \beta_1 + u_i$$

where the u_i's are iid from the scale mixture of Gaussian model described in the previous section. The x_i's are generated iidly from the standard Gaussian distribution, so intercept and slope estimators for the model have the same asymptotic variance. The usual median regression (least absolute error) estimator will be used as an initial

n	LAE	LSE	Adaptive
100	1.756	1.726	1.308
200	1.805	1.665	1.279
400	1.823	1.750	1.284
800	1.838	1.753	1.304
∞	1.803	1.800	1.256

Table 1. MSE scaled by sample size, n, for Tukey scale mixture of normals

estimator for our adaptive estimator and we will compare performance of both with the ubiquitous least squares estimator.

3.1. **Some Implementation Details.** Our implementation of the Kiefer-Wolfowitz estimator requires several decisions about the grid $v_1, \cdots, v_m$. For scale mixtures of the type considered here it is natural to adopt an equally spaced grid on a log scale. I have used $m = 300$ points with $v_1 = \log(\max\{0.1, \min\{r_1, \cdots, r_n\}\})$ and $v_m = \log(\max\{r_1, \cdots, r_n\})$. Bounding the support of the mixing distribution away from zero seems to be important, but a corresponding upper bound on the support has not proven to be necessary.

Given an estimate of the mixing distribution, $\hat{F}$, the score function for the efficient M-estimator is easily calculated to be,

$$\hat{\psi}(u) = (-\log \hat{g}(u))' = \frac{\int u\varphi(u/\sigma)/\sigma^3 d\hat{F}(\sigma)}{\int \varphi(u/\sigma)/\sigma d\hat{F}(\sigma)}.$$

We compute this estimate again on a relatively fine grid, and pass a spline representation of the score function to a slightly modified version of the robust regression function, `rlm()` of the R package MASS, Venables and Ripley (2002), where the final M-estimate is computed using iteratively reweighted least squares.

3.2. **Simulation Results.** For the Tukey scale mixture model (1) with $\epsilon = 0.1$ and $\sigma = 3$ mean and median regression have essentially the same asymptotic variance of about 1.80, while the efficient (MLE) estimator has asymptotic variance of about 1.25. In Table 1 we see that the simulation performance of the three estimators is in close accord with these theoretical predictions. We report the combined mean squared error for intercept and slope parameters scaled by the sample size so that each row of the table is comparable to the asymptotic variance reported in the last row.

It seems entirely plausible that the proposed procedure, based as it is on the Kiefer-Wolfowitz nonparametric estimate of the mixing distribution, would do better with discrete mixture models for scale like the Tukey model than for continuous mixtures like the Student t(3) model chosen as our second test case. Kiefer-Wolfowitz delivers a discrete mixing distribution usually with only a few mass points. Nevertheless,

n	LAE	LSE	Adaptive
100	1.893	2.880	1.684
200	1.845	2.873	1.579
400	1.807	2.915	1.540
800	1.765	2.946	1.524
∞	1.851	3.000	1.500

Table 2. MSE scaled by sample size, n, for Student t(3) mixture of normals

in Table 2 we see that the proposed adaptive estimator performs quite well for the Student t(3) case achieving close to full asymptotic efficiency for sample sizes 400 and 800.

4. Conclusions

Various extensions naturally suggest themselves. One could replace the Gaussian mixture model with an alternative; van der Vaart (1996) suggests the logistic as a possibility. As long as one maintains the symmetry of the base distribution adaptivity is still tenable, but symmetry, while an article of faith in much of the robustness literature, may be hard to justify. Of course, if we are only interested in slope parameters in the regression setting and are willing to maintain the iid error assumption, then symmetry can be relaxed as Bickel (1982) has noted.

The challenge of achieving full asymptotic efficiency while retaining some form of robustness has been a continuing theme of the literature. Various styles of ψ-function carpentry designed to attenuate the influence of outliers may improve performance in small to modest sample sizes. Nothing, so far, has been mentioned about the evil influence of outlying design observations; this too could be considered in further work.

References

Andersen, E. D. (2010): "The MOSEK Optimization Tools Manual, Version 6.0," Available from `http://www.mosek.com`.

Andrews, D., P. Bickel, F. Hampel, P. Huber, W. Rogers, and J. W. Tukey (1972): *Robust Estimates of Location: Survey and Advances.* Princeton University Press: Princeton.

Bickel, P. J. (1982): "On adaptive estimation," *The Annals of Statistics*, 10, 647–671.

Efron, B., and R. A. Olshen (1978): "How broad is the class of normal scale mixtures?," *The Annals of Statistics*, 5, 1159–1164.

Friberg, H. A. (2012): "Users Guide to the R-to-MOSEK Interface," Available from `http://cran.r-project.org`.

Kiefer, J., and J. Wolfowitz (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, 27, 887–906.

Koenker, R. (2012): "REBayes: An R package for empirical Bayes methods," Available from `http://cran.r-project.org`.

Koenker, R., and I. Mizera (2014): “Shape Constraints, Compound Decisions and Empirical Bayes Rules,” *Journal of the American Statistical Association*, 109, 674–685.

Laird, N. (1978): “Nonparametric Maximum Likelihood Estimation of a Mixing Distribution,” *Journal of the American Statistical Association*, 73, 805–811.

Robbins, H. (1951): “Asymptotically subminimax solutions of compound statistical decision problems,” in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, vol. I. University of California Press: Berkeley.

——— (1956): “An empirical Bayes approach to statistics,” in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. I. University of California Press: Berkeley.

Stein, C. (1956): “Efficient nonparametric testing and estimation,” in *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 187–195.

Stone, C. J. (1975): “Adaptive maximum likelihood estimators of a location parameter,” *The Annals of Statistics*, 3, 267–284.

van der Vaart, A. (1996): “Efficient maximum likelihood estimation in semiparametric mixture models,” *The Annals of Statistics*, 24, 862–878.

Venables, W. N., and B. D. Ripley (2002): *Modern Applied Statistics with S*. Springer, New York, fourth edn.