Complexity-Regularized Multiresolution Density Estimation

Rebecca M. Willett
Rice University
Houston, TX, USA
email: willett@rice.edu

Robert D. Nowak[1]
University of Wisconsin
Madison, WI, USA
email: nowak@ece.wisc.edu

Abstract — **The density estimation method proposed in this paper employs piecewise polynomial fits on adaptive dyadic partitions. The proposed estimator enjoys the minimax adaptivity associated with wavelet-based density estimators as well as the following additional advantages: estimates are guaranteed to be non-negative, theoretical bounds provide an indication of performance even for small sample sizes, and the method can be extended to free-degree piecewise polynomial estimation, which allows the data to adaptively determine the smoothness of the underlying basis functions.**

I. Multiscale penalized likelihood estimation

We estimate a density, $f : [0,1] \to [0,\infty)$, from n observations using a minimum description length/coding theoretic approach to regularization and information theoretic techniques based on the Li-Barron bound and its extension [1, 2]. We accomplish this by sub-dividing $[0,1]$ into $N \sim n$ intervals, each of length $1/N$, and record the number of samples falling into each interval. This results in a vector of count measurements $\boldsymbol{x} = \{x_i\}_{i=0}^{N-1}$, where x_i is the number of samples in the i-th interval Q_i. The $\boldsymbol{x}$ are multinomially distributed according to the probabilty mass function (pmf) $\boldsymbol{f} = \{f_i\}_{i=0}^{N-1}$, where $f_i = \int_{Q_i} f$. Thus, the likelihood of observing $\boldsymbol{x}$, given $\boldsymbol{f}$, is multinomial and denoted by $p(\boldsymbol{x}|\boldsymbol{f})$.

The proposed method calculates density estimates by determining the adaptive dyadic partition P of the range of observations (assumed to be $[0,1]$) and using maximum likelihood estimation to fit a polynomial to each interval in the partition. Thus, a density $f_n(P)$ is associated with every possible partition. The space of possible partitions, $\mathcal{P}$, is a nested hierarchy defined through a recursive dyadic partition (RDP) of $[0,1]$, and the optimal partition is selected by optimally pruning the complete RDP of the data range. This gives our estimators the capability of spatially varying the resolution to automatically increase the smoothing in very regular regions of the density and to preserve detailed structure in less regular regions.

The RDP is pruned to optimize a penalized likelihood criterion, wherein the penalization is based on the complexity of the underlying partition:

$$\widehat{P}_n \equiv \arg\min_{P \in \mathcal{P}} \left[-\log p(\boldsymbol{x} \mid f(P)) + \text{pen}(P,r)\right], \quad (1)$$

where r is the polynomial order and pen(P,r) is the prefix codelength required to uniquely describe an order-r piecewise polynomial on P. The codelengths are proportional to the size of the partition associated with each model, and thus penalization leads to estimates that favor smaller partitions. We take $\widehat{f}_n \equiv f(\widehat{P}_n)$ as our density estimator. The degree r can also be adaptively selected using a similar procedure.

[1]R. Nowak was partially supported by the National Science Foundation, grants CCR-0310889 and ANI-0099148, and the Office of Naval Research, grant N00014-00-1-0390

II. Bounds on estimation error

The L_1 error of the estimator described in the previous section is bounded in the following theorem:

Theorem 1 *Let f be a density in the Besov space $B^\alpha_\tau(L_\tau([0,1]))$ where $0 < \alpha \leq r$, $1/\tau = r + 1/2$ and $L_2([0,1])$ is the approximation space. Further assume $0 < C_\ell \leq f(\cdot) \leq C_u$. Let $\widehat{f}_n$ be the multiscale piecewise polynomial estimator satisfying (1) using the penalty pen$(P,r) = (2|P|-1)\log_e 2 + \frac{|P|r}{2}\log_e n$. Then for some constant $C > 0$*

$$\mathbb{E}\left[\|\widehat{f}_n - f\|_{L_1}\right] \leq C \left(\frac{\log 2n}{n}\right)^{\frac{\alpha}{2\alpha+1}}$$

At the heart of the proof of this theorem is a key information-theoretic inequality derived by Li and Barron [1] and extended by Kolaczyk and Nowak [2]. This upper bound is within a logarithmic factor of the lower bound on the minimax risk, demonstrating the near-optimality of the proposed method. Note that the result of Theorem 1 holds for free-degree estimators, too; however, free-degree estimation allows the data to adaptively determine the smoothness of the underlying basis function instead of forcing the user to select a polynomial order or wavelet smoothness.

While there is no closed-form solution for the MLE of the polynomial coefficients with respect to the multinomial likelihood, the multinomial likelihood function is a convex function and the set of possible polynomial coefficient vectors is a convex set, which ensures that the polynomial coefficients, and hence the entire density estimate, can be computed quickly, as summarized in Theorem 2 below. For complete proofs and detailed discussion, see [3].

Theorem 2 *A fixed-degree piecewise polynomial PLE can be computed in $O(N)$ calls to a convex minimization routine and $O(N)$ comparisons of the resulting (penalized) likelihood values, where N is the number of intervals in the initial complete RDP. A free-degree piecewise polynomial PLE can be computed in $O(N \log N)$ calls to a convex minimization routine and $O(N)$ comparisons of the resulting (penalized) likelihood values.*

References

[1] Q. Li and A. Barron. *Advances in Neural Information Processing Systems 12*, chapter Mixture Density Estimation. MIT Press, 2000.

[2] E. Kolaczyk and R. Nowak. Multiscale likelihood analysis and complexity penalized estimation. To appear in Annals of Stat., 2003. Available at http://www.ece.wisc.edu/ nowak/pubs.html.

[3] R. Willett and R. Nowak. Multiscale density estimation. Technical report, Rice University, 2003. Submitted to *IEEE Transactions on Information Theory*; available at http://www.ece.rice.edu/∼willett/papers/WillettIT2003.pdf.