COARSE-TO-FINE MANIFOLD LEARNING

Rui Castro, Rebecca Willett and Robert Nowak

Department of Electrical and Computer Engineering, Rice University, Houston, TX
Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI

ABSTRACT

In this paper we consider a sequential, coarse-to-fine estimation of a piecewise constant function with smooth boundaries. Accurate detection and localization of the boundary (a manifold) is the key aspect of this problem. In general, algorithms capable of achieving optimal performance require exhaustive searches over large dictionaries that grow exponentially with the dimension of the observation domain. The computational burden of the search hinders the use of such techniques in practice, and motivates our work. We consider a sequential, coarse-to-fine approach that involves first examining the data on a coarse grid, and then refining the analysis and approximation in regions of interest. Our estimators involve an almost linear-time (in two dimensions) sequential search over the dictionary, and converge at the same near-optimal rate as estimators based on exhaustive searches. Specifically, for two dimensions, our algorithm requires $O(n^{7/6})$ operations for an n-pixel image, much less than the traditional wedgelet approaches, which require $O(n^{11/6})$ operations.

1. INTRODUCTION

The dimensionality of signals is often lower than the ambient observation space. For example, a pure sinusoidal process occupies a one-dimensional linear subspace. Linear subspace models and subspace identification techniques have played a major role in modern signal processing. However, in many cases the signal may occupy a nonlinear lower-dimensional manifold of the observation space. A simple example of this phenomenon occurs in the analysis of images. For example, consider a binary image composed of a "white" region and a "black" region separated by a smooth boundary. This "image" is simply a one-dimensional curve (the boundary) embedded in the two-dimensional image space. Estimating or coding the image involves identifying or *learning* the manifold corresponding to the boundary. Several investigators have proposed new basis functions or dictionaries for describing $(d-1)$-dimensional manifolds embedded in d-dimensional spaces (for d=2,3), including wedgelet/beamlet dictionaries and curvelet frames [1, 2, 3]. While promising, the dictionaries are overcomplete and can be quite computationally demanding to implement. This computational hurdle motivates our work here. We consider sequential, coarse-to-fine manifold learning strategies. The basic idea is to first examine the data on a coarse grid, and then refine the analysis and approximation in regions predicted near the boundary. By carefully examining the bias and variance trade-offs in each stage, we show that manifolds can be optimally recovered through a sequential process in almost linear-time (in two dimensions), yielding significant computational savings. The work presented here has conceptual similarities with the work by Blanchard and Geman, which also exploits coarse-to-fine decision making [4].

Supported by the National Science Foundation, grants CCR-0310889 and ANI-0099148, and the Office of Naval Research, grant N00014-00-1-0390

2. PROBLEM FORMULATION

Consider a function f defined on the d-dimensional hypercube $[0,1]^d \subseteq \mathbb{R}^d$ (assume $d \geq 2$). The function consists of constant regions separated by $(d-1)$-dimensional boundaries. We assume these boundaries are Hölder-2 smooth (for example, twice continuously differentiable curves). In two dimensions this corresponds to the Horizon class of images described in [1], as shown in Figure 1(a). We do not observe the function f directly, but only samples which have been corrupted by noise. Consider a partition of the unit hypercube into n sub-hypercubes of sidelength $n^{-1/d}$ (assume without loss of generality that n is a power of d). Denote each hypercube by $V(i)$, $i \in \{1,\ldots,n\}$. In Figure 1(a) this procedure is shown for $d = 2$. Each one of these "small" sub-hypercubes corresponds to a voxel, and these determine the finest resolution we consider. This initial partition can be generated by a recursive dyadic partition (RDP). First divide the domain into 2^d sub-hypercubes of equal size. Repeat this process again on each sub-hypercube. Proceeding in this fashion $1/d \log_2 n$ times yields the initial partition. This gives rise to a complete RDP of resolution $1/n$ (*i.e.*, the original domain is divided into n cells). The RDP process can be represented with a rooted tree structure: the root node corresponds to the entire domain (*i.e.*, the unit hypercube), their children nodes correspond to the 2^d sub-hypercubes, and so on.

For each voxel we associate the value $\theta(i)$,

$$\theta(i) = \frac{1}{|V(i)|} \int_{V(i)} f \, ,$$

the average of f over the voxel $V(i)$, where $i \in \{1,\ldots,n\}$ and $|V(i)|$ denotes the volume of $V(i)$. We do not observe the value of each voxel directly, but instead only a noisy corrupted version. Our measurements, $x(i)$, are samples of the field $\theta(i)$ corrupted by additive white Gaussian noise with variance σ^2, that is, $x(i) \sim \mathcal{N}(\theta(i), \sigma^2)$, where we assume that all measurements are statistically independent. Given these measurements we would like to estimate the voxel values $\theta(i)$.

Let $\Theta = \{\theta(i)\}_i$ and $\mathbf{x} = \{x(i)\}_i$. Let $\widehat{\theta}_{\mathbf{x}}(i)$ be our estimate for the value of voxel i (in the following we will drop the dependence on $\mathbf{x}$ for the ease of notation). Define $\widehat{\Theta} = \{\widehat{\theta}(i)\}_i$. The measure of performance we consider is the Mean Square Error (MSE), defined as

$$\mathrm{MSE}(\widehat{\Theta}, \Theta) \equiv \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\theta(i) - \widehat{\theta}(i)\right)^2\right]. \qquad (1)$$

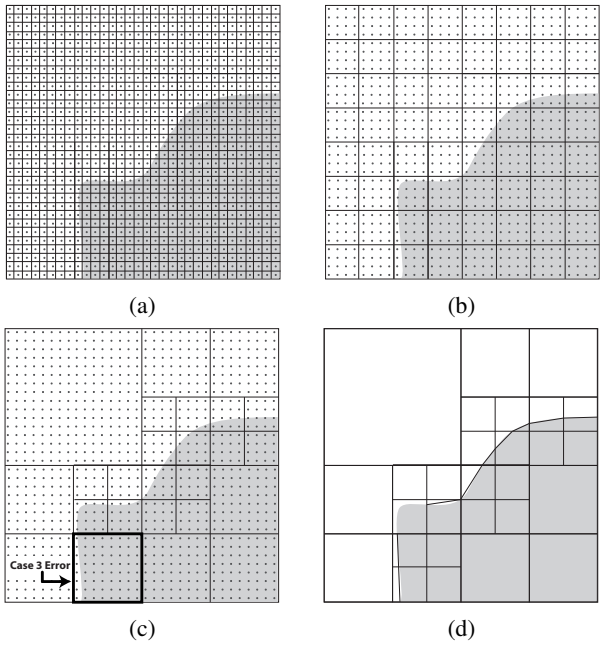


Fig. 1. Example boundary estimation problem. (a) Initial RDP used in traditional piecewise linear methods. (b) Initial coarse-resolution RDP used in the preview stage. (c) Partition generated in a preview stage. Note the example of a Case 3 error. (d) Final partition generated during the refinement stage.

3. MAXIMUM PENALIZED LIKELIHOOD ESTIMATORS

For any reasonable estimator $\widehat{\Theta}$ it is known that as the number of voxels n increases the MSE (1) decreases. For piecewise constant functions on $[0,1]^d$ separated by $(d-1)$-dimensional Hölder-2 boundaries, the MSE decays no faster than $O(n^{-2/(d+1)})$, the minimax lowerbound [5].

It turns out that it is possible to nearly achieve the optimal rates above using maximum penalized likelihood techniques. To see this, first note that the complete RDP can be pruned back to produce an RDP with non-uniform resolution. Let $\mathcal{P}_n$ be the class of all possible prunings. For each $P \in \mathcal{P}_n$, there is an associated tree structure (generally of non-uniform depth corresponding to the non-uniform resolution of most RDPs). The leafs of each tree represent dyadic (side-length equal to a power of 2) hypercube regions of the associated partition. Consider a certain RDP P, and define the estimator of the field on each leaf of the partition to be the least-squares fit of a wedgelet to the measurements in the corresponding hypercube. In d dimensions, a wedgelet fit consists of a $(d-1)$-dimensional hyperplane separating a hypercube into two regions and a constant fit to the data in each region. Let $\widehat{\Theta}_P$ denote a model of the field (based on the least-squares model fits on each square of P). The empirical measure of performance is the sum-of-squared errors between $\widehat{\Theta}_P$ and the data:

$$R\left(\widehat{\Theta}_P, \mathbf{x}\right) = \sum \left(\widehat{\theta}_P(i) - x(i)\right)^2.$$

For fixed partition P, the choice of $\widehat{\Theta}_P$ that minimizes $R(\widehat{\Theta}_P, \mathbf{x})$ is simply given by the least-squares fits on each square, as discussed above. Now define the complexity penalized estimator as

$$\widehat{\Theta} = \arg \min_{\widetilde{\Theta}_P : P \in \mathcal{P}_n} R\left(\widetilde{\Theta}_P, \mathbf{x}\right) + 2\sigma^2 c|P| \log n \qquad (2)$$

where $|P|$ denotes the number of hypercubes in the partition P and c is constant which will be defined later. This optimization can be solved using a bottom-up pruning algorithm [1, 6]. It has the further advantage that upper bounds on the estimation error can be established using several recent information-theoretic results, most notably the Li-Barron bound [7] and Nowak and Kolaczyk's generalization of this bound [8]. Specifically, if c is chosen so that the dictionary of estimators satisfies the Kraft inequality,

$$\sum_{\widetilde{\Theta}_P : P \in \mathcal{P}_n} e^{-c|P| \log n} \leq 1,$$

then

$$\text{MSE}(\widehat{\Theta}, \Theta) \leq \min_{\widetilde{\Theta}_P : P \in \mathcal{P}_n} \frac{2}{n} R\left(\widetilde{\Theta}, \Theta\right) + \frac{8\sigma^2 c|P| \log n}{n}. \qquad (3)$$

For the remainder of the paper assume that f has Hölder-2 $(d-1)$-dimensional smooth boundaries. For this class of functions it is known that the minimax MSE rate is bounded below by $O(n^{-2/(d+1)})$. It can be shown that solving the optimization in (2) yields a partition which balances the approximation error and estimation error terms in the bound on the MSE in (3), resulting in a final bound of $O\left((\log n/n)^{-2/(d+1)}\right)$.

The main challenge associated with such methods involves the optimization in (2). In general, the algorithms needed to achieve the performance rates described above must exhaustively examine all models in each of the candidate partitions $P \in \mathcal{P}_n$. For example, in the two-dimensional wedgelet case, the number of wedgelet models required depends on the δ-resolution of the wedgelet analysis, which determines the approximation accuracy [1]. Specifically, for a hypercube of sidelength ℓ the number of wedgelet models that must be evaluated is $O(\ell/\delta)^2$ (and each evaluation requires $O(n\ell^2)$ operations). Wedgelet analysis nearly achieves the minimax performance rate for the class of images under consideration when $\delta \sim n^{-2/3}$. Since wedgelet approximations must be calculated for each candidate partition $P \in \mathcal{P}_n$, a total of $O(n^{4/3})$ wedgelet fits must be computed [1], resulting in an overall computational complexity of $O(n^{7/3})$. This can be improved slightly by noting that the evaluation of the wedgelet models can be done in a incremental fashion, where each wedgelet model fit is evaluated using previously evaluated models; this means that for a given hypercube of sidelength ℓ, after the first wedgelet fit is calculated, each successive wedgelet fit can be calculated in $O(\sqrt{n}\ell)$ operations as opposed to $O(n\ell^2)$ operations. Taking this into account yields an overall computational complexity of $O(n^{11/6})$, which is prohibitive in practice.

4. COARSE-TO-FINE ESTIMATION

The heavy computational complexity of the above techniques motivates our work here. In the constant regions of the function f, the piecewise constant approximation estimates perform well. The task that limits the rates of convergence of the MSE is the estimation of f in the vicinity of the boundaries. The main idea is therefore to perform the estimation task in a sequential, coarse-to-fine fashion: In the first step (termed the *Preview* step), a coarse

estimation of the field is performed, using only piecewise constant approximations, as an attempt to identify the approximate location of the boundary. In a second step (termed the *Refinement* step) we perform a wedgelet boundary fit in areas that were identified as possible boundary regions. If the Preview step is effective then we will perform wedgelet fits (which are computationally demanding) only in the regions were they are needed, instead of applying and testing them throughout the entire domain.

In the remainder of this section, we show that this method (a) nearly achieves the minimax performance rate of $O(n^{-2/(d+1)})$ and (b) requires significantly fewer computational resources than wedgelet methods based on exhaustive dictionary searches. In the following we are going to omit the logarithmic factors to make the presentation lighter.

4.1. Error Analysis

In the first step of our approximation we start with an uniform RDP with n^γ voxels, as shown in Figure 1(b). We generate an estimator by pruning this RDP, as shown in Figure 1(c), and decorating each leaf with a constant model. Let $V_c(i)$ denote the voxels corresponding to the n^γ resolution uniform RDP (we will refer to this as the *coarse resolution*, as opposed to the *fine resolution*, which has a total of n voxels). Note that each coarse resolution voxel contains $n^{1-\gamma}$ measurements. For each of these coarse resolution voxels let $x_c(i)$ be the average of the measurements falling into $V_c(i)$, therefore

$$x_c(i) \sim \mathcal{N}(\theta_c(i), n^{-(1-\gamma)}\sigma^2),$$

where $\{x_c(i)\}_i$ are statistically independent and

$$\theta_c(i) = \frac{1}{n^{1-\gamma}} \sum_{j:V(j)\subseteq V_c(i)} \theta(i).$$

We can evaluate the mean squared error at the coarse resolution, and it is given by

$$\begin{aligned} \text{MSE}_c &\equiv \mathbb{E}\left[\frac{1}{n^\gamma}\sum_{i=1}^{n}(\theta_c(i)-\widehat{\theta}_c(i))^2\right] \quad (4)\\ &= O(n^{-(1-\gamma)}(n^\gamma)^{-1/d}) \\ &= O(n^{-1+\gamma\frac{d-1}{d}}). \quad (5)\end{aligned}$$

This can be derived by noting that the variance of each $x_c(i)$ is $n^{-(1-\gamma)}\sigma^2$ and the MSE of the piecewise constant estimator at resolution n^γ decays like $O((n^\gamma)^{-1/d})$, for unit variance noise.

Denote the pruned RDP at coarse resolution by RDP_c. In the refinement step we consider a piecewise linear fit on the leafs of RDP_c that were not pruned (*i.e.*, the leafs of RDP_c that are at the deepest level), keeping all the other leafs unaltered, as shown in Figure 1(d). The main reasoning is that with very high probability (we will make this precise below), most voxels at the coarse resolution that do not intersect the boundary are going to be pruned, so we can use the unpruned voxels as a good indication for the presence of a boundary.

In the following we evaluate the asymptotic behavior of the MSE at the fine resolution for the two step procedure. Our analysis makes repeated use of the fact that the number of coarse voxels intersecting the boundary is $O(n^{\gamma\frac{d-1}{d}})$ (this follows from the assumption that the boundary is Hölder-2, and therefore the boundary set has box-counting dimension $d-1$). For each leaf in the pruned RDP_c we consider three situations, and analyze the impact on the overall MSE:

Case 1: *Leafs of RDP_c that do not intersect the boundary:* In this case, averaging the observations is the optimal solution and the MSE at the fine resolution behaves like (5). Therefore these leafs contribute $O(n^{-1+\gamma\frac{d-1}{d}})$ to the fine resolution MSE. This dictates our choice of γ; by choosing $\gamma = \frac{d}{d+1}$ we obtain the desired MSE rate of $O(n^{-2/(d+1)})$.

Case 2: *Leafs of RDP_c that were not pruned:* For leafs in this scenario we are going to perform a wedgelet fit (the refinement step) and the MSE decays exactly as if we were doing wedgelet fits everywhere. Therefore these leafs contribute $O(n^{-2/(d+1)})$ to the fine resolution MSE.

Case 3: *Leafs of RDP_c that were pruned, but intersect the boundary:* This scenario corresponds to a case were a voxel intersecting the boundary was somehow "erroneously" pruned to a larger leaf. Therefore this leaf is approximated with a constant, but contains a fragment of the boundary. For large enough n the function f in the hypercube corresponding to this leaf is composed of two constant regions. Because the voxel containing the boundary was pruned, we know that the volume of one of the two constant regions is small with respect to the total volume of the leaf hypercube, and behaves like $O(n^{-1/2+\gamma/2})$ (this follows from the fact that the squared bias at the coarse resolution in bounded by (5)). This yields an average bias squared at the fine resolution of order $O(n^{-1/2+\gamma(1/2-1/d)})$. Setting $\gamma = \frac{d}{d+1}$ (which is necessary to bound the Case 1 error) does not result in the desired fine resolution MSE rate of $O(n^{-2/(d+1)})$.

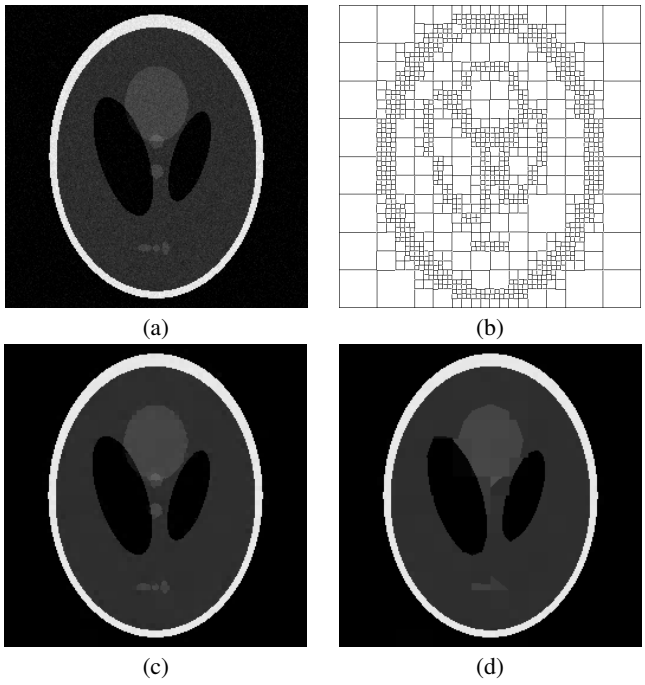
We propose a technique that overcomes this difficulty while incurring only minimal extra computational cost. Recall that, for large enough n, the function f in the hypercube corresponding to a prunned leaf is composed of two constant regions. Case 3 errors occur when the boundary of the two regions is closely aligned with the uniform RDP of resolution $n^\gamma/2^d$, consisting of $n^\gamma/2^d$ hypercubes; this is just one level of resolution coarser than the unpruned coarse RDP. An example Case 3 error is depicted in Figure 1(c). The basic idea is then to perform two preview stages – one as described above, and the other on a version of the partition which is shifted by one coarse resolution voxel in each coordinate. This ensures that the boundary is detected with high probability in one of the preview stages. A voxel erroneously pruned in one of the preview stages would not be pruned in the other preview stage. In the refinement stage we perform a wedgelet fit to any coarse resolution voxel that was left unpruned in either the first or second preview steps.

4.2. Computational Complexity

We will describe here the computational complexity associated with the two-dimensional case, as studied in the simulations; the extension to d dimensions is straightforward. First, recall that wedgelets are fit to all leafs of RDP_c which were not pruned, and since $\gamma = 2/3$, each of these hypercubes has sidelength $\ell = n^{-1/3}$ and contains $n^{1/3}$ pixels. Because $\delta \sim n^{-2/3}$, $(\ell/\delta)^2 \sim n^{2/3}$ wedgelet fits are evaluated at each of these leafs. Since there are $O(n^{1/3})$ such leafs, a total of $O(n)$ wedgelet fits must be calculated, resulting in an overall computational complexity of $O(n^{4/3})$. Note that the complexity can be reduced to $O(n^{7/6})$ operations by calculating wedgelet fits incrementally, as described above. Thus this coarse-to-fine approach is significantly faster that the traditional wedgelet analysis method.

5. SIMULATIONS

We demonstrate the effectiveness of the proposed method using the Shepp-Logan phantom brain image, commonly used in medical imaging simulations. The n noisy measurements, arranged on a 256×256 grid, is displayed in Figure 2(a); in this example, $\sigma^2 = 0.001$ and the mean pixel value is 0.15. For this simulation, the penalization weights are chosen according to the theory in 3. Under this scenario, $n^\gamma = 256^{4/3}$, and so the preview stage is initialized with an RDP of 4096 4×4 coarse resolution squares. The preview partition in Figure 2(b) demonstrates how the initial Haar estimate does not prune the initial coarse resolution squares in regions near the boundaries; after the preview stage pruning, the initial coarse resolution RDP of 4096 squares has been pruned back to a nonuniform RDP with only 1199 4×4 squares remaining after using the procedure described in Section 4.



(a) (b) (c) (d)

Fig. 2. Shepp-Logan phantom. (a) 256×256 noisy measurements, $\sigma^2 = 0.001$, MSE $= 0.0115$. (b) Preview partition. (c) Shepp-Logan phantom estimate formed by fitting one wedgelet or constant to each of the unpruned squares from the preview stage; MSE $= 0.000504$. (d) Shepp-Logan phantom estimate using standard wedgelet, MSE $= 0.00163$.

The final estimate after the refinement stage is displayed in Figure 2(c); recall that this requires $O(n^{4/3})$ operations to compute. This estimate can be compared to a standard wedgelet decomposition, as seen in Figure 2(d). This requires $O(n^{7/3})$ operations; *i.e.*, a factor of $O(n)$ more operations than the proposed method. These estimates were calculated on a 667 MHz PowerPC G4 with 768 MB of memory running Mac OS 10.2.8; on this machine, the standard wedgelet estimate was computed in 665 seconds and the coarse-to-fine estimate was computed in 37.4 seconds. This is an excellent example of how the proposed method performs as well as a standard wedgelet estimate in terms of both MSE and visual quality with significant computational savings.

6. FINAL REMARKS

In this paper we study the estimation of piecewise constant functions, where the different constant regions are separated by Hölder-2 smooth boundaries. Techniques for this were previously developed by Donoho [1]. In this work we build on the class of models described in [1], but instead of performing a computationally demanding search over a large dictionary of models, we proceed in a sequential fashion, using a two stage process, where in the first stage we select a subset of image models, and in the second stage we make a final model selection. While this is a greedy procedure, it has desirable features, such as low computational cost, and we prove that it is asymptotically optimal. Although in this paper we present a specific problem, where the computational cost drives our choice of the sequential approach, there are other scenarios and problems that can benefit from the same kind of sequential approaches. For example, in estimation problems in sensor networks [9], a sequential approach is used so save valuable communication resources. Currently we are working on generalizations of the above sequential procedures for other estimation and classification problems.

7. REFERENCES

[1] D. Donoho, "Wedgelets: Nearly minimax estimation of edges," *Ann. Statist.*, vol. 27, pp. 859 – 897, 1999.

[2] E. Candès and D. Donoho, "Curvelets: A surprisingly effective nonadaptive representation for objects with edges," To appear in Curves and Surfaces, L. L. Schumaker et al. (eds), Vanderbilt University Press, Nashville, TN.

[3] D. Donoho and X. Huo, "Beamlets and multiscale image analysis," Tech. Rep., Stanford University, 2001, Available at http://www-stat.stanford.edu/∼donoho/reports.html.

[4] G. Blanchard and D. Geman, "Hierarchical testing designs for pattern recognition," Tech. Rep., Universite Paris-Sud, 2003, Available at http://www.cis.jhu.edu/publications/papers_in_database /GEMAN/seqtesting.pdf.

[5] A. P. Korostelev and A. B. Tsybakov, *Minimax theory of image reconstruction*, Springer-Verlag, New York, 1993.

[6] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1983.

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

[8] E. Kolaczyk and R. Nowak, "Multiscale likelihood analysis and complexity penalized estimation," submitted to Annals of Stat. August, 2001. Available at http://www.ece.wisc.edu/ nowak/msla.pdf.

[9] R. Willett, A. Martin, and R. Nowak, "Backcasting: A new approach to energy conservation in sensor networks," To appear in the third international symposium of Information Processing in Sensor Networks (IPSN'04), April 2004.