ABSTRACT

Traditional methods of nonparametric function estimation (splines, kernels and especially wavelet filters) usually produce artificial features/spurious oscillations. Piecewise convex function estimation seeks to reliably estimate the geometric shape of the unknown function. We outline how piecewise convex fitting may be applied to signal recovery, instantaneous frequency estimation, surface reconstruction, image segmentation, spectral estimation and multivariate adaptive regression. Two distinct methodologies for shape-correct estimation are given. First, we propose a piecewise convex function estimation method that strongly penalizes additional inflection points and "efficiently" penalizes additional degrees of freedom. Second, a two-stage adaptive (pilot) estimator is described. In the first stage, the number and location of the change points are estimated without strong smoothing. In the second stage, a constrained smoothing spline fit is performed with the smoothing level chosen to minimize the MSE. The imposed constraint is that a single second-stage change point occurs in a region about each empirical change point of the first-stage estimate. This constraint is equivalent to requiring that the third derivative of the second-stage estimate has a single sign in a small neighborhood about each first-stage change point.

1. SIGNAL RECOVERY

Our basic tenet is that "Naturally occurring functions have very few inflection points and the fitted curve should preserve the geometric fidelity of the unknown function by having the same number of inflections." Consider a signal, \( y_i = g(t_i) + \epsilon_i \), measured at \( t_i = i\delta, i = 1 \ldots N \), where \( \{\epsilon_i\} \) are independent Gaussian random variables: \( \epsilon_i \sim N(0, \sigma^2) \). Let \( g(t) \) have \( K \) change points of \( \ell \)-convexity with change points \( x_1 \leq x_2 \leq \ldots \leq x_K \) if \( (-1)^{\ell-1}g^{(\ell)}(t) \geq 0 \) for \( x_k \leq t \leq x_{k+1} \). In practice, we take \( \ell = 2 \).

Our goal is to estimate \( g(t) \) while preserving the geometry of \( g \). A standard technique, kernel smoothers [11], estimates \( g(t) \) by a weighted local average: \( \hat{g}_h(t) = \frac{1}{h} \sum_{k=1}^{N} y_k \kappa \left( \frac{t_k - t}{h} \right) \), where \( h \) is the kernel halfwidth that determines the amount of smoothing. As \( h \) increases, the random error (variance) in \( \hat{g} \) decreases, while the systematic error (bias) increases. In [8], we have examined the geometric faithfulness of kernel smoothers in the limit as \( N \to \infty \) and \( h \to 0 \) with \( Nh = 1 \). The halfwidth that minimizes the mean square error scales as \( N^{-\frac{1}{2}} \) for \( g \in C^m[0, 1] \) provided that the kernel satisfies certain moment conditions. For \( \ell = m \) or \( m - 1 \), this halfwidth scaling, \( N^{-\frac{1}{2}} \), produces extra (artificial) \( \ell \)-change points with high probability. To eliminate artificial inflection points with probability \( \sim 1 \), the smoothing must be increased such that \( h \gg N^{-\frac{1}{2}} \ln N \). Thus, the level of smoothing required for geometric fidelity is large enough to degrade the MSE. In [8], we propose a two-stage estimator:

Stage 1: Smooth strongly with \( h_N N^{-\frac{1}{2}} \ln N \to \infty \). Denote the empirical \( \ell \)-change points by \( \hat{x}_k, k = 1 \ldots K \).

Stage 2: Perform a constrained smoothing spline fit by minimizing the penalized likelihood subject to the constraints that \( \hat{g}^{(\ell+1)}(t) \) does not change sign in the intervals, \([\hat{x}_k - \frac{1}{2} \sigma_k, \hat{x}_k + \frac{1}{2} \sigma_k]\). The \( k \)th empirical change point variance is \( \sigma_k^2 \equiv \left[ \frac{1}{h_{\text{stage}}^2} \int_{[\hat{x}_k - h_{\text{stage}}, \hat{x}_k + h_{\text{stage}}]} \kappa^2 \right] \), where \( \hat{g}_{\text{stage}} \), \( \hat{x}_k \) is the stage-1 estimate of \( g^{(\ell+1)} \) at \( \hat{x}_k \). \( c \) is a constant that depends on the kernel shape. The confidence interval parameter, \( \frac{1}{2} \alpha \), is the \( \alpha \)-quantile for the normal distribution.

To motivate this two-stage estimator, consider the smoothing dilemma. If the smoothing level is optimized for MSE, there tends to be too many \( \ell \)-change points (wiggles). If the smoothing is chosen to eliminate the artificial change points (suppress wiggles), then the MSE suffers. Our two-stage estimate provides the best of both worlds! In [8], we show that this two-stage estimator achieves the optimal rate of convergence for functions in the Sobolev space \( W_{m,2} \), while suppressing artificial inflection points in the neighborhood where they are likely to occur.

In practice, we choose the second-stage smoothing parameter, \( \lambda_{\text{stage}} \), by generalized cross-validation while we scale the first-stage smoothing as \( h_{\text{stage}} = R(N)^{\alpha} \), where \( R(N) \equiv (\log(N)N^2)^{1/2} \) with \( \alpha = \frac{1}{2} - \frac{\ell}{2m} \). Unconstrained smoothing splines can be used in first stage with the smoothing parameter, \( \alpha \), scaled with the correspondence: \( \lambda = R_{\text{stage}}^{2/3} \), with \( R_{\text{stage}} \sim (\log N)^{3/2} N^{-\frac{1}{2}} \).

To impose the piecewise convex (PC) constraints, we recommend the simple iterative method of Villabos and Wahlba (VW) [15], i.e. we add pointwise constraints on the sign of \( g^{(m)}(z_m) \) with \( z_m \) chosen in the constraint regions. The goal is to select \( \{z_j, j = 1 \ldots M\} \) such that the constraints are satisfied everywhere even though they are imposed only at a finite number of points. An important advantage of our two-stage estimator is that we impose constraints only in small regions about \( \hat{x}_k \). Since the constraint regions are small, the number of \( z_j \) that are necessary in the VW scheme is small. Thus, our two-stage estimator can be interpreted as a pilot estimator to determine where to place the \( \{z_j\} \) in the VW scheme.

The VW implementation of the second-stage estimator reduces to a quadratic programming problem with linear in-
equality constraints. The number of constraints is bounded by \(m \cdot k + \# \text{ of data points in constraint regions}\). The dual formulation is easier to implement since positivity constraints are substituted for inequality constraints. The B-spline representation gives a banded structure to the programming problem. The quadratic programming minimization may be solved using either “active set” methods or interior point methods. Interior point methods are more easily modified to take advantage of the tridiagonal structure.

2. ROBUST ESTIMATION

Our asymptotic analysis of the statistics of false inflection points is limited to Gaussian errors and linear estimators. In practice, it is often advantageous to replace both the residual errors and the penalty function with more robust analogs: 
\[
\sum_{i=1}^{N} |y_i - \hat{g}(t_i)|^q + \lambda \int |\hat{g}^{(m)}(t)|^2 dt,
\]
where \(1 \leq q \leq 2\). Representation and duality theorems are given in [5] for \(q > 1\). When \(m = 1\), the stage-2 minimization reduces to a finite dimensional convex minimization in the dual formulation. A heuristic scaling shows that the effective halfwidth of the robustified function satisfies \(h_{\text{eff}} = \sqrt{\frac{\lambda}{\sum_{i=1}^{N} |y_i - \hat{g}(t_i)|^q}}\). The bias error scales as \(|g^{(m)}(t)| h_{\text{eff}}^m\) while the “variance” is proportional to \(1/Nh_{\text{eff}}^m\). The halfwidth that minimizes the mean square error (MSE) scales as \(h_{\text{opt}} \approx \sqrt{\frac{\lambda}{\sum_{i=1}^{N} |y_i - \hat{g}(t_i)|^q}}\), while the halfwidth to eliminate false change points of \(g^{(l)}\) with asymptotic probability one satisfies \(h_{\text{opt}} N^\frac{q-2}{2q} \rightarrow \infty\) for \(1 \leq q \leq 2\), the effective halfwidth of the robustified function automatically reduces the halfwidth in regions of large \(|g^{(m)}(t)|\) just like a variable halfwidth smoother, with \(q = 1\) minimizing the difference between \(h_{\text{eff}}\) and \(h_{\max}\).

3. WAVELET THRESHOLDING: A WIGGLE ENHANCER

Although fashionable, wavelet estimators usually produce very ugly/unphysical estimates with dozens of false inflection points. This unpleasant truth is never mentioned by wavelet protagonists. Wavelet algorithms for function estimation offer two advantages: 1) speed, the algorithms are often \(O(N \log(N))\), 2) asymptotic minimax optimality in a number of decision theoretic settings [2]. The speed arises from separability: each wavelet coefficient is estimated separately without regard to geometric fidelity. The asymptotic optimality theory assumes that the unknown function is an arbitrary member of a function space, which makes function fits with ten or twenty inflection points as reasonable as fits with no inflection points. Essentially, function spaces contain too many “unphysical” functions. We prefer the “common sense prior” that the function has only a few inflection points with high probability.

Both of these advantages of wavelets disappear when the more realistic assumption is made that the unknown function has only a small number of convexity change points. Wavelet thresholding has another intrinsic disadvantage when it comes to geometric fidelity: wavelets possess the complete oscillation property [1]. In contrast, B-splines have the antithetical and valuable property, total positivity: the number of sign changes of \(g(t)\) is less than the number of sign changes of the sequence of B-spline coefficients.

4. P.C. INFORMATION CRITERION

Information/discrepancy criteria are used to measure whether the improvement in the goodness of fit is sufficient to justify using additional free parameters. Both the number of free parameters and their values are optimized with respect to the discrepancy criterion, \(d\hat{g}(y_i)\). Let \(\hat{\sigma}^2\) be a measure of the average residual error: 
\[
\hat{\sigma}^2(y_i) = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{g}(t_i)|^2,
\]
or its \(L_1\) analog. Typical discrepancy functions are
\[
d^1(\hat{g}, \{y_i\}) = \hat{\sigma}^2 \sqrt{1 - \min(\gamma, p/N)^2} \quad \text{and} \quad (1)
\]
\[
d^2(\hat{g}, \{y_i\}) = \hat{\sigma}^2 [1 + \min(\gamma, p/N)]\sqrt{1 - \min(\gamma, p/N)^2} , \quad (2)
\]
where \(p\) is the number of free parameters in the fit. For \(\gamma_1 = 1\), \(d^1(\hat{f}, \{y_i\})\) is generalized cross-validation (GCV) which has the same asymptotic behavior as the Akaike information criterion. For \(\gamma_2 = 1\), \(d^2(\hat{f}, \{y_i\})\) is the Bayesian/Schwartz information criterion. For a nested family of models, \(\gamma_2 = 1\) is appropriate while \(\gamma_2 = 2\) corresponds to a nonnested family with \(2(N)^{\frac{1}{2}}\) candidate models at the \(h\)th level [2]. In very specialized settings in regression theory and time series, it has been shown that functions like \(d^2\) are asymptotically efficient while those like \(d^2\) are asymptotically consistent. In other words, using \(d^2\)-like criteria will asymptotically minimize the expected error at the cost of not always yielding the correct model. In contrast, the Bayesian criteria will asymptotically yield the correct model at the cost of having a larger expected error.

Our goal is to consistently select the number of convexity change points and efficiently estimate the model subject to the change point restrictions. Therefore, we propose the following new discrepancy criterion:
\[
P C I C = \hat{\sigma}^2(\hat{g}, \{y_i\}) \left[1 + \gamma_2 K \ln(N)/N \right]^{\frac{1}{2}} \left[1 - \gamma_2 p/N \right] , \quad (3)
\]
where \(K\) is the number of convexity change points and \(p\) is the number of free parameters. PCIC stands for Piecewise Convex Information Criterion. In selecting the positions of the \(K\) change points, there are essentially \(2^{(N)^{\frac{1}{2}}}\) possible combinations of change point locations if we categorize the change points by the nearest measurement location. Thus, our default values are \(\gamma_1 = 1\) and \(\gamma_2 = 2\).

We motivate PCIC: to add a change point requires an improvement in the residual square error of \(O(\sigma^2 \ln(N))\), which corresponds to an asymptotically consistent estimate. If the additional knot does not increase the number of change points, it will be added if the residual error decreases by \(\gamma_1 \hat{\sigma}^2\). Presently, PCIC is purely a heuristic principle. We conjecture that it consistently selects the number of change points and is asymptotically efficient within the class of methods that are asymptotically consistent with regards to convexity change points.

5. ADAPTIVE KNOT PLACEMENT

Adaptive regression splines determine the knot locations by minimizing a discrepancy criterion. In [5], Friedman proposed a new criterion, \(d_N(\hat{g}, \{y_i\})\), that resembles the Akaike criterion, but penalizes the degrees of freedom (DoF) for the knot locations more than the other DoF:
\[
d^N(\hat{g}, \{y_i\}) = \hat{\sigma}^2 \sqrt{1 - \min(\gamma, p + 3p_{\text{DoF}})/N^2} , \quad (4)
\]
where \( p_{kn} \) is the number of knots in the fit and \( \hat{p} \) is the number of other free parameters. The heuristic argument for \( d^0(\hat{g}, \{u_i\}) \) is that the DoF for knot placement are more effective at reducing the predictive uncertainty relative to a given change in \( \hat{g}^2 \) than DoF for the spline coefficients. Although this motivation for \( d^0 \) is weak and needs further analysis, we give the corresponding Friedman correction for the PCIC:

\[
PCTIC = \hat{g}^2(\hat{g}, \{u_i\}) \left[ \frac{1 + \gamma_2 K \ln(N)/N}{1 - (\gamma_1 \hat{p} + 3p_{kn})/N^2} \right],
\]

where \( K \) is the number of convexity change points, and our default values are \( \gamma_1 = 1 \) and \( \gamma_2 = 2 \).

6. INSTANTANEOUS FREQUENCY AND TIME-FREQUENCY REPRESENTATIONS

We consider a signal whose frequency is being slowly modulated in time. The canonical example is a "chirp": \( y(t) = \cos(\omega_0 t^2 + 2bt) \), where the instantaneous frequency is \( \omega_0(0) = a + 2bt \). In [7], we proposed using the WKB (eikonal/geometric optics) representation for signals whose amplitude and frequency are being slowly modulated in time:

\[
y(t) = A(t) \cos \left( \int_0^t \omega(s) ds \right) + \epsilon_t,
\]

where \( \epsilon_t \) is white noise and \( \delta \) is a small parameter. The characteristic time scale for amplitude and frequency modulation is \( 1/\delta \). Since the unknown functions, \( A(t) \) and \( \omega(t) \), are one-dimensional, we estimate them nonparametrically instead of evaluating a two-dimensional time-frequency representation. In the time-frequency plane, we represent the signal as the curve, \( A(t, \omega(t)) \). In [7], we propose estimating \( A(t) \) and \( \omega(t) \) using data adaptive kernel smoothers. The instantaneous frequency corresponds to a first derivative estimate. Since we expect the phase to be piecewise convex, we replace the adaptive kernel estimate of \( \cos(\hat{\omega}(t)) \) and \( A(t) \), with PC fitting. The circular statistics require a penalized likelihood function of the form:

\[
\sum_{i=1}^{n} \left\{ \frac{y_i}{A(t_i)} - \cos \left( \int_0^{t_i} \hat{\omega}(s) ds \right) \right\}^2 + \lambda \int \left[ \hat{\omega}''(t)^2 + [\omega''(t)]^2 \right] dt.
\]

Eq. (7) may also be used for a smoothing spline fit without PC constraints. The first term in Eq. (7) differs from Katkovnik [4] by placing \( A(t) \) in the denominator instead of the numerator.

7. SPECTRAL ESTIMATION

Consider a stationary time series, \( \{x(t)\} \), with an unknown spectral density, \( S(f) \). A standard method to estimate \( S(f) \) is to multiply the data by a spectral taper, compute the windowed periodogram, and then smooth the windowed periodogram or its logarithm with a data-adaptive kernel smoother or smoothing splines. In [12-14], this estimator is improved by replacing the single spectral window with a family of orthonormal spectral windows. The multiple window estimate reduces the broad band bias while making a variance stabilizing transformation of the log-periodogram. When we use the sinusoidal tapers of [12], \( v^i(t) = \sqrt{\frac{2}{N+1} \sin \left( \frac{kt}{N+1} \right)} \), the multi-window estimate reduces to

\[
\hat{S}_{MW}(f) = \frac{\Delta}{K} \sum_{k=1}^{K} \left| y(f + k\Delta) - y(f - k\Delta) \right|^2 + \lambda \int \left| \hat{\omega}''(t)^2 + [\omega''(t)]^2 \right| dt
\]

where \( \Delta = 1/(2N + 2) \). The difference, \( y(f + k\Delta) - y(f - k\Delta) \), results in sideloobe cancellation. From [13], we recommend \( K = (N/2)^{1/2} \). Instead of kernel smoothing \( ln[\hat{S}_{MW}(f)] \), we now advocate using the two-stage piecewise convex fitting procedure. Piecewise convex fitting should prove particularly advantageous to spectral estimation because it will suppress the \( 1/N \) oscillations that arise from discrete time sampling.

8. ADDITIVE MODELS, PROJECTION PURSUIT AND MARS

Many classes of models attempt to fit multi-dimensional functions as sums of one-dimensional functions. In each case, we replace the standard nonparametric estimation methods with piecewise convex fitting. Additive growth curve models [9] fit models of the form \( g(t, x_1, \ldots, x_m) = f_0(t) + \sum_{i=1}^{m} f_i(t) x_i \), where the \( f_i \) are determined by smooth splines (old method) or PC fitting (new method). The back fitting algorithm (corresponding to the Gauss-Seidel iteration) may be used to PC fit the \( f_i \) iteratively. Similar remarks apply to projection pursuit [3]. In MARS (multivariate adaptive regression splines), the model is a sum of products: \( f(x_1, x_2, \ldots, x_m) = \sum_{j=1}^{M} g_j(x_j) h_j(x_j) \). The knots are placed adaptively, in our case, using the piecewise convex information criterion (3).

9. IMAGE SEGMENTATION

Image segmentation divides a digital picture into similar regions for further processing. The image is assumed to be piecewise constant and the goal is to determine the the boundaries of the regions. The Mumford-Shah (MS) algorithm [5] estimates the region boundaries by minimizing the sum of the residual square fit error plus a penalty term proportional to the length of the boundary. If the boundary is parameterized as \( x(s) \), the penalty term is the total variation of \( x(s) \). We modify the MS algorithm by replacing the total variation penalty with a piecewise convex constrained fit using a robustified penalty function such as the \( L_1 \) integral of the boundary curvature. When using the PCIC (3), we use the arclength divided by the grid spacing as a proxy for \( N \), the number of data points.

10. RESPONSE SURFACE ESTIMATION

We seek to estimate a smooth function, \( g(x_1, x_2) \), given \( N \) noisy measurements. The two-dimensional analog of PC fitting is to divide the plane into regions where the Gaussian curvature (or more simply \( \Delta f \)) has a single sign. The boundaries between regions of positive and negative Gaussian curvature are free boundaries that we require to be PC. The piecewise convex information criterion may be applied in a straightforward manner, but the numerical optimization of the nonsmooth functional is challenging.

If instead a pilot estimator is to be used, we suggest using a penalty function of the form \( \lambda \int \Delta^{1/2} d^2 \), where we
require that $\lambda_1^{2m} \gg \log(N) N^{-\frac{1}{2m}}$ for the first stage of the fit. This scaling is heuristic since the statistics of false zeros of $\Delta^{2\ell+2} \hat{g}$ are unknown, as is the critical scaling of the smoothing parameter that avoids extra regions of incorrectly specified curvature. In the second stage, we suggest imposing constraints on the sign of $\partial_{\alpha\beta} \Delta^{2\ell+2} \hat{g}$ near the first-stage convexity boundaries. The stage-2 convexity boundaries, $(x(s), y(s))$, are free and need to be fitted using a penalty function plus PC constraints on the curve shape. The numerical implementation appears tricky with a need for some elliptic analog of front tracking. Overall, the two-dimensional problem appears very challenging from both theoretical and numerical perspectives.

11. EVOLUTIONARY SPECTRA

We consider a nonstationary stochastic process: $x_t = \int A(\omega,t) dZ(\omega)$, where $dZ(\omega)$ is a stochastic process with independent spectral increments $E[dZ(\omega_1)dZ(\omega_2)] = \delta(\omega - \omega')|d\omega|$. The representation is nonunique for Gaussian processes since $A(\omega,t)$ corresponds to a square root of the covariance matrix. To resolve the nonuniqueness, we require that $A(\omega,t)e^{i\omega t}$ correspond to the Fourier transform of the positive definite square root of the covariance matrix. When $A(\omega,t)$ evolves slowly in time, the evolutionary spectrum is $S(\omega,t) = \left| A(\omega,t) \right|^2$.

Let $\lambda_f$ be the characteristic frequency scale length and $\tau$ be the characteristic time scale of $A$: $A(f/\lambda_f, t/\tau)$, with the sampling rate $= 1$. In [6], we present an asymptotic expansion $(\tau \lambda_f > 1)$ of the mean square error in estimating $S(f/\lambda_f, t/\tau)$. We begin by evaluating the multi-window analog of the log-spectrogram on a two-dimensional time frequency lattice. The bias error is minimized by using a window length of $N_w \sim \sqrt{\tau/\lambda_f}$. In [6], we estimate $\hat{S}(\omega,t)$ using a two-dimensional cross-product kernel smoother on the log-multi-window spectrogram. The optimal halfwidths, $h_1$ & $h_2$, scale as $h_1/h_2 \sim \sqrt{\tau/\lambda_f}$ and $h_1/h_t \sim (\tau^2 \lambda_f^2)^{-1/3}$, where $h_t$ and $h_2$ are the halfwidths in the $t$ and $f$ directions.

We now advocate replacing kernel smoothing with PC fitting the log multi-windowed spectrogram. This method should eliminate the spurious $1/N_w$ oscillations which occur due to the discrete sampling.

12. SUMMARY

We have described two different nonparametric estimation methods that can exclude spurious oscillations with negligible sacrifice of fit quality. Pilot estimators apply heavy smoothing to determine the correct shape and then relax the smoothing subject to shape restrictions. The piecewise convex information principle is a heuristic that attempts to be asymptotically consistent in the number of inflection points and asymptotically efficient in the number of free parameters. A number of applications of piecewise convex function estimation have been outlined. This PC-constrained methodology can be used to solve inverse problems.

In our analysis, we have imposed positivity/negativity constraints on $\hat{g}^{(\ell+k)}(t)$ with $\ell + k \leq m$. However the estimates of $\hat{g}^{(m)}(t)$ (both constrained and unconstrained) are very unphysical. We believe that future work should concentrate on imposing piecewise convex constraints on the highest order derivative, $g^{(m)}(t)$.

13. REFERENCES