Simple Inference in Exploratory Data Analysis: SiZer

J. S. Marron

Operations Research and Indust. Eng.
Cornell University
&
Department of Statistics
University of North Carolina

Starting Request

Please ask questions (make comments) on the fly

- Keeps me in contact with you
- Somebody else may be wondering about that, too
- Creates time for ideas to percolate …
- Talk's organization allows adaptation of time

Organization

Section I: SiZer Introduction
Section II: A careful look under the hood
Section III: Examples (real data and simulated)
Section IV: Extensions (SiCon, 2d, Time Series, …)
Section V: Fun with scale space & Historical connections
Section VI: Concluding Thoughts

Organization, Section I

SiZer Introduction:

- settings: scatterplot smoothing and histograms
- Fossils data
- Incomes data
- Central Question:
  Which features are "really there"?
- Solution Part I, Scale Space
- Solution Part II, SiZer

Main Smoothing Settings

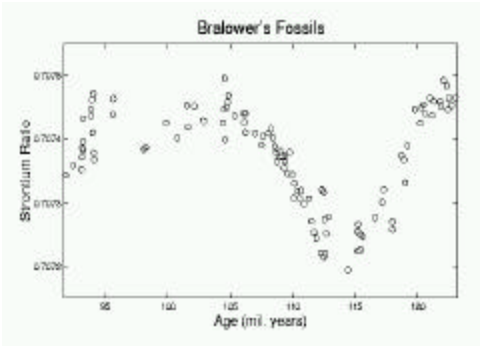
1. Scatterplot Smoothing (nonparametric regression)
2. Histograms (density estimation)

Main Setting 1: Scatterplots

Fossil Data

- from T. Bralower, Dept. Geological Sciences, UNC
- Strontium Ratio in fossil shells
- reflects global sea level
- surrogate for climate
- over millions of years

Show top of SiZer2Eg_Fossil.ps



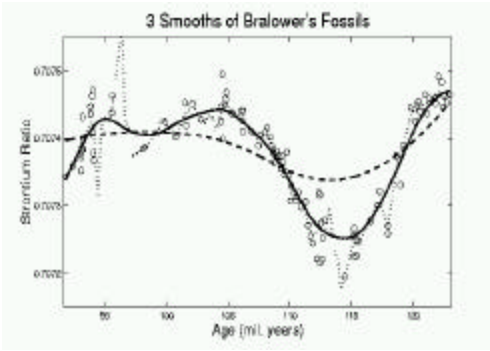
Main Setting 1: Scatterplot Smoothing

Smooths of Fossil Data (details given later)

- dotted line: undersmoothed (feels sampl'g variability)
- dashed line: oversmoothed (imp't features missed?)
- solid line: smoothed about right?

Central question: Which features are "really there"?

Show bottom of SiZer2Eg_Fossil.ps



Main Setting 2: Histograms

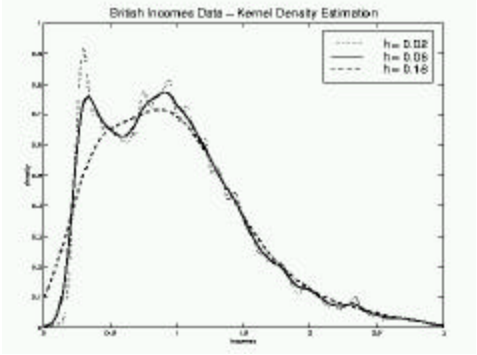
Family Income Data: British Family Expenditure Survey

- Distribution of Incomes
- ~ 7000 families
- "Smooth histograms" (details given later)
- Again under- and over- smoothing issues

Central question: Which features are "really there"?

(e.g. 2 modes?)

Show Sizer2Eg_Incomes.ps



Central Question

In Exploratory Data Analysis:

Which features are "really there"?

A rephrasing: What is "important underlying structure", as opposed to being "noise artifacts", or "attributable to sampling variability"?

Central Question (cont.)

In Exploratory Data Analysis:

Which features are “really there”?

Confounding factor: level of smoothing

- Everything disappears with oversmoothing
- Spurious features appear from undersmoothing

Solution, Part I

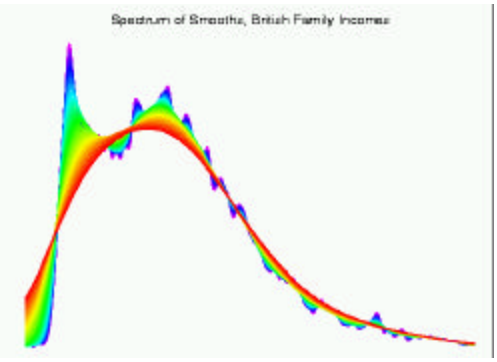
“Scale Space” – idea from Computer Vision

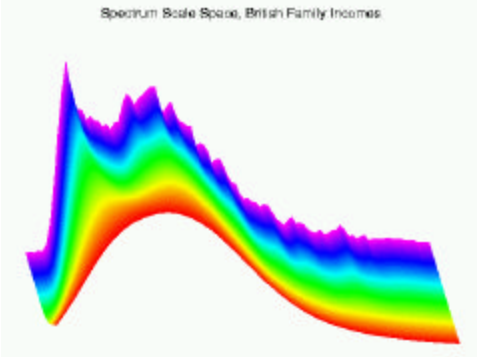
Background concept:

- Oversmoothing = “view from afar” (macroscopic)
- Undersmoothing = “zoomed in view” (microscopic)

Main idea: all smooths contain useful information,
so study “full spectrum” (i. e. all smoothing levels)

Show IncomesKDEspect.mpg, IncomesKDEspect.ps and IncomesKDEspect3d.ps





Solution, Part I (cont.)

“Scale Space” – from Computer Vision

Main idea: all smooths contain useful information,
so study “full spectrum” (i. e. all smoothing levels)

Note: this viewpoint makes
“data based bandwidth selection”
much less important (than I once thought….)

Solution, Part II

SiZer:

Significance of ZERo crossing of the derivative, in scale space

Combines:

- needed statistical inference
- novel visualization

To get: a powerful exploratory data analysis method

SiZer

Basic idea: a "bump" is characterized by:
an increase, followed by a decrease

Generalization: many "features of interest" captured by
sign of the slope of the smooth

SiZer Basis:
Statistical inference on slopes, over scale space

SiZer (cont.)

Visual presentation:
Color map over scale space:

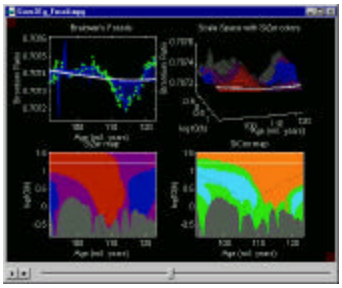
- Blue: slope significantly upwards (deriv. CI above 0)
- Red: slope significantly downwards (der. CI below 0)
- Purple: slope insignificant (deriv. CI contains 0)

SiZer – Fossil Data

Show Sizer2Eg_Fossil.mpg

Upper Left: Scatterplot, family of smooths, 1 highlighted
Upper Right: Scale space rep'n of family, with SiZer colors
Lower Right: SiZer map, more easy to view
Lower Left: SiCon map – will discuss later

Slider (in movie viewer) highlights different smoothing levels



SiZer – Fossil Data (cont.)

Oversmoothed: Decreases at left, not on right
Medium smoothed:
- Main valley significant, and left most increase
- smaller valley not statistically significant

Undersmoothed:
- "noise wiggles" not significant

Additional SiZer color: gray not enough data for inference

SiZer (cont.)

Common Question: which is "right"?
- decreases on left, then flat
- up, then down, then up again
- no significant features

Answer: All are "right", just different "scales of view",
i.e. "levels of resolution of data"

SiZer – Incomes data

Show Sizer2Eg_Incomes.mpg

Same format as above

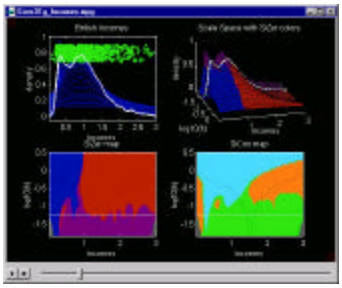
Oversmoothed: Only one mode

Medium smoothed: Two modes statistically significant

Confirmed by PhD dissert'n of H. P. Schmitz (U. Bonn)

Undersmoothed: many "noise wiggles", not significant

Again: all are "correct", just different "scales"



SiZer (cont.)

Usefulness of SiZer in exploratory data analysis:

- Smoothing experts: saves time
- Smoothing beginners: avoids terrible mistakes
  - don't find things that "aren't there"
  - do find important features
- Directly targets critical scientific question:
  "is a deeper analysis worthwhile?"

Organization, Section II

SiZer: A careful look under the hood

- why not histograms?
- kernel density estimation
- local linear smoothing
- simultaneous inference
- bias issues
- why not confidence bands?

Why not histograms?

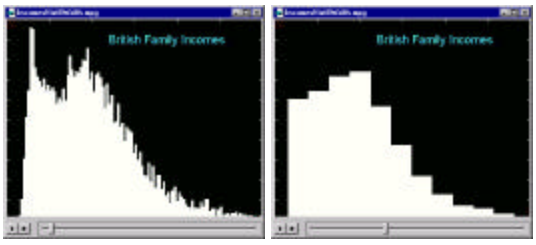
Incomes Data

Histogram Problem 1: Binwidth (well known)

Show IncomesHistBWidth.mpg

Undersmoothing vs. Oversmoothing

Major impact, as expected



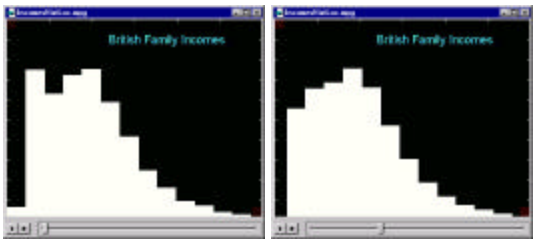
Why not histograms? (cont.)

Histogram Problem 2: Bin shift (less well known)

Show IncomesHistLoc.mpg

For *same binwidth*, can get much different impression,
by only "shifting grid location"

Serious impact: much less expected



Smooth Histograms

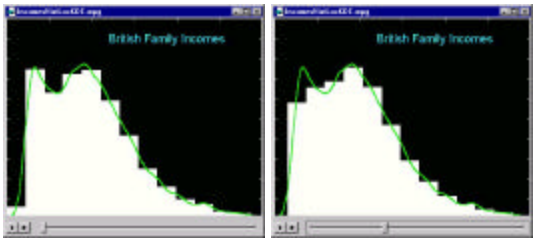
Solution to binshift problem:
essentially "average over all shifts"

Show IncomesHistLocKDE.mpg

- 1st peak all in one bin: bimodal
- 1st peak split between bins: unimodal

Smooth histogram provides understanding,
so should use for data analysis

Another name: Kernel Density Estimate



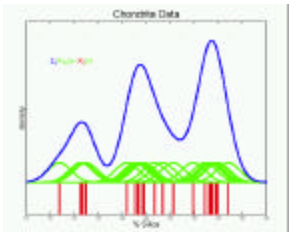
Kernel density estimation

View 1: Smooth histogram

View 2: distribute probability mass, according to data

Show EGkdeCombined.pdf

E.g. Chondrite data (from how many sources?)

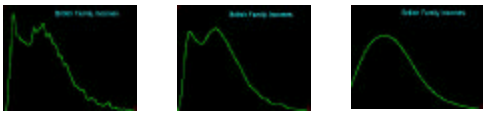


Kernel density estimation (cont.)

Central Issue: width of window, i.e. "bandwidth", h

Show IncomesKDE.mpg

Controls critical amount of smoothing



Old Approach: data based bandwidth selection

Jones, Marron and Sheather (1996), *JASA*, 91, 401-407.

New Approach: scale space (look at all of them)

Kernel density estimation (cont.)

Less Important Issue: shape of window

Personal Recommendation: Gaussian

- "Looks best"
- "Bump monotonicity" (discussed later)
- Can avoid apparent computational drawback
  (using fast "binned" implementation)

Scatterplot smoothing

There are many methods (most with fierce advocates):

- kernel / local polynomials
- smoothing splines
- B – splines (regression splines)
- orthogonal series (e.g. wavelets)

Scatterplot smoothing (cont.)

"Best" method is personal choice based on crit. such as:

- statistical efficiency
- computational efficiency
- simplicity
- interpretability

For further discussion: Marron (1996) in *Statistical Theory and Computational Aspects of Smoothing*, eds. W. Härdle and M. Schimek, 1-9.

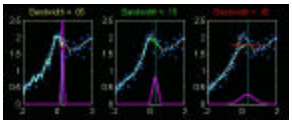
Scatterplot smoothing (cont.)

Personal Preference: local linear smoothing

Main idea: use kernel window to "determine neighborhood",
then "fit a line within the window"
then "slide window along"

Show NPRMovie1a.mpg

- Window Width again critical



Local linear smoothing (cont.)

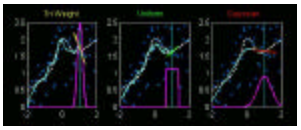
- Window shape much less important

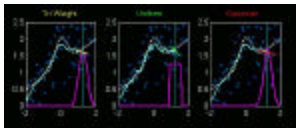
Show NPRMovie5b.mpg

- After “modding out window size”

Show NPRMovie5a.mpg

See: Marron and Nolan (1989) "Canonical kernels for density estimation", *Statist. Prob. Letters*, 7, 195-199.





Simultaneous inference in SiZer

Problem: for many independent Hypo. Tests,
just by chance some will reject (even under null)

I.e. “multiple comparisons problem”

For full map, simultaneous inference is very important

Simultaneous inference in SiZer (cont.)

Simple Approach: for each row

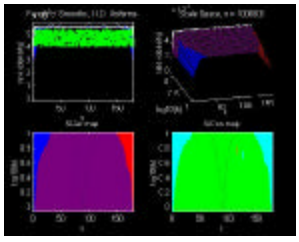
measure “Effective Sample Size” = # pts. in kern’l window

Then s = “# indep. Subsamples” = n / ESS

Do standard adjustment for “s independent CIs”

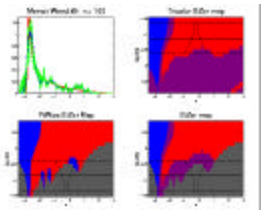
Result: looks good (only see boundary effects)

Show SiZerUnif1M.mpg



Simultaneous inference in SiZer (cont.)

Check effects of ignoring simultaneous inference



Show MW3TricolorSiZer.ps

Bias Issues

Classical Analysis of Smoothing:

Mean Squared Error = Variance + Squared Bias

Variance: big when undersmoothed

Squared Bias: big when oversmoothed

Temptation: Estimate bias, and recenter C. I.

Bias Issues (cont.)

Problem: too hard to estimate

(else could genuinely improve smoothing method)

(not possible via minimax lower bound theory)

Simulation verification:

Härdle and Marron (1991) *Ann. Statistics*, 19, 778-796

Solution: Scale Space philosophy from Computer Vision

Bias Issues (cont.)

Computer vision - scale space view of smoothing bias:

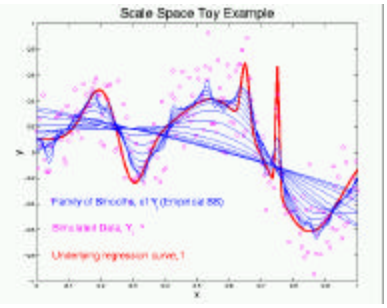
Not important, because reflects "unavailable information"

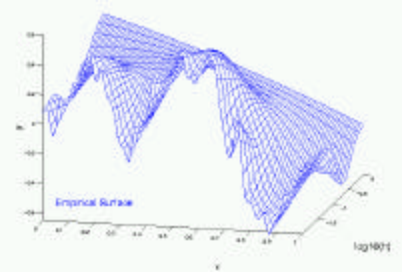
Show ScaSpaTalk1Combined.pdf

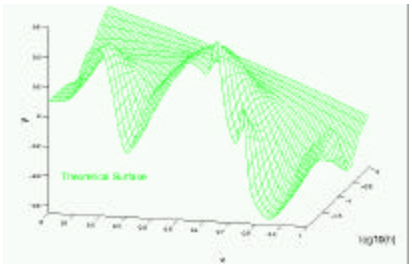
"Empirical scale space surface" is (unbiased) estimate of the "theoretical scale space surface".

Each theoretical curve is:

"all can have at given level of resolution (i.e. scale)"







Why not Confidence Bands?

Reason 1: Bands don't capture "variability in curves"

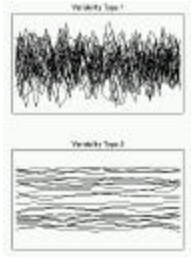
Show WhyNotCIs1.ps

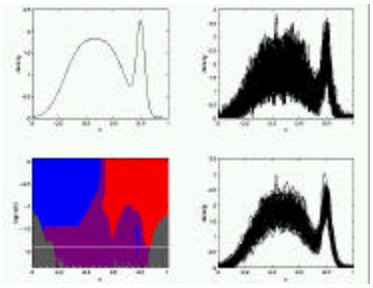
Reason 2: Properly adjusted bands too wide

From Hall (1992) *The Bootstrap and Edgeworth Expansion*, Springer.

Show WhyNotCIs2.ps

Bands more conservative than SiZer (since bias adjust'd)





Organization, Section III

SiZer Examples (real data and simulated)

- Stamps Data
- Mollusks Data
- Dust Data
- Simulations
- Chondrite Data
- Stock Prices (online)

SiZer Examples – Stamps Data

Stamp Thicknesses of Hidalgo Stamp (Mexico, 1800's)

Show Sizer2Eg_Stamps.mpg

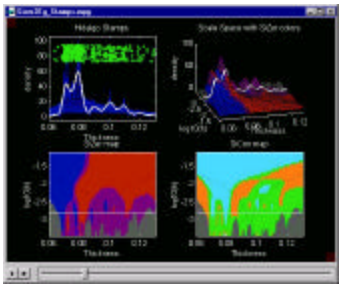
How many sources produced the paper?

Answers in literature: "More than 1" to "suggested 10"
7 is most common

SiZer: 3 for sure, suggestion of 4^{th} and 5^{th}

Lesson: SiZer is conservative, compared to "mode tests"

SiZer Examples – Mollusks Data



Data from Matthew Campbell

When were there massive mollusk extinctions?

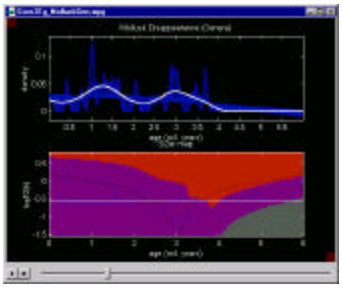
Smooth of “last times of appearance” of mollusk genera

Show Sizer2Eg_MolluskGen.mpg

SiZer shows only “overall decrease”

(can’t find “bumps” suggested by smooth)

SiZer Examples – Mollusks Data (cont.)



Clear need: more data (to sharpen inference)

Problem: these gathered over more than 100 years

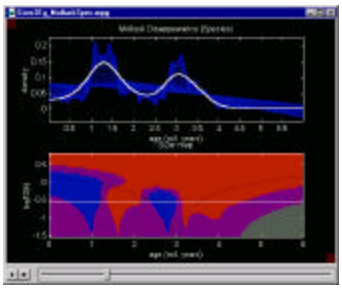
Solution: Go to species level

Show Sizer2Eg_MolluskSpec.mpg

Now two bumps nicely significant

- correlate with known major climatic change

SiZer Examples – Dust Data



Density of Dust Particle Sizes:

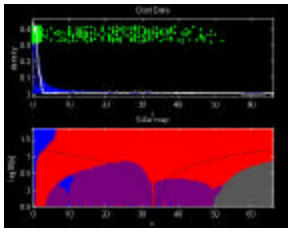
Show Sizer2Eg_Dust.mpg

Moderate Scales:

- Tight dist’n of small sizes
- Spread dist’n of larger sizes
- Valley between

Small Scales: “fringe of small significant features”

caused by heavy data rounding



SiZer Examples – Normal Mixture # 9

Truth: Two big modes, one small one in center

n = 100 SiZer : not enough info to find any mode

Show SiZer2Eg_NM9n100.mpg

n = 1000 SiZer : can find two big modes

Show SiZer2Eg_NM9n1000.mpg

n = 10000 SiZer : all 3 modes are very clear

Show SiZer2Eg_NM9n10000.mpg

Recall SiZer finds "structure really diff'nt from noise"

SiZer Examples – Normal Mixture # 15

Truth: Three fat modes, three narrow modes

n = 100, 1000, 10000 SiZer : similar lessons to n incre'g

Show SiZer2Eg_NM15n100.mpg, SiZer2Eg_NM15n1000.mpg & SiZer2Eg_NM15n10000.mpg

Note: full scale space is important, since different features appear at different scales

Interesting approach to "local bandwidth choice":

Draw "bandwidth function curve on SiZer map"

(Pieces are all there, but not done yet)

SiZer Examples – Chondrite Data

Show: Sizer2Eg_Chondrite.mpg

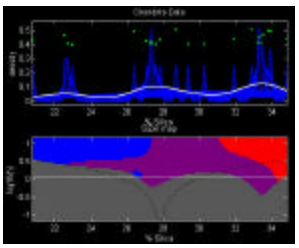
Lesson: not always enough info to find structure

3 modes not found here

Mode tests can be better by focussing on modes

SiZer is "omnibus type" test,

which broadly spreads power, at some cost



SiZer Examples – Finance Data

"Tick Data" – instantaneous prices of a stock

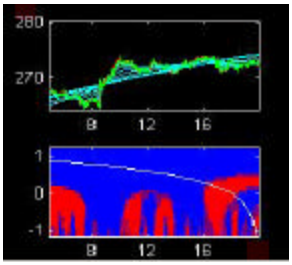
Imitation of "on line" view

Show: SiZerStockPrice2.mpg

Want to "predict trends" at various scales,

Use right edge, and white curve to indicate "time range"

- "quadruple point"
  (of scale based increase – decrease)
- "colors flop" as overall trend shifts



Organization, Section IV

SiZer Extensions:

- SiCon, significant curvature
- 2d: Significance in Scale Space
- Time Series
- Jumps
- Other models (censor, hazard est., gen. l'hood,…)
- Smoothing Splines

SiZer Extensions: SiCon, significant curvature

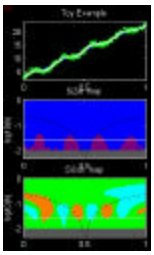
Idea: study "curvature", not "slopes"

orange for concave downwards

cyan for convex upwards

green for not significant

Show SiconToyEG.mpg



Found "cluster of shortcut runners"

Show MarathonTimesHalf.ps

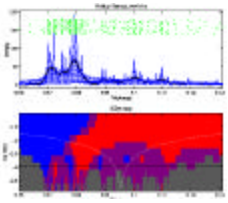
(not SiZer), note dissipated later…

Show MarathonTimesFull.ps

SiZer Examples – Stamps Data (revisited)

Finer grid: fine scale shows discretization effect

Show Sizer2Eg_StampsFineRes.ps

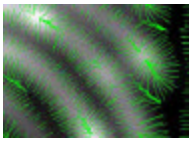


SiZer Extensions: 2d: Significance in Scale Space

Major challenge: what to look at?

- red – purple – blue solid regions?
- what is "up" and "down"?

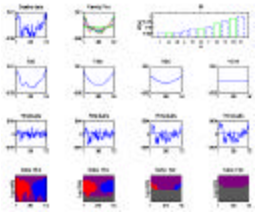
Show SSS1FIG1.EPS, SSS1FIG2.MPG, sss1fig3.mpg, sss1fig6b.mpg



SiZer Extensions: Time Series

Major challenge: what is "trend" vs. "serial correlation"?

Show StrikesEggs, PanelSiZerSine.ps, DepSiZerSine.ps, DepSiZerDeaths.ps

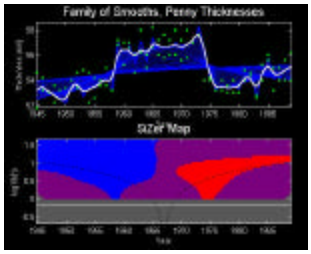


SiZer Extensions: Jumps

Idea: "jumps" (discont's) have signature in SiZer map

Analyze math'ly , and "invert" to give "jump indicator"

Show SZJPenny.mpg, SZJBlocks.ps & SZJBlocksJump.ps



SiZer Extensions: Other smoothing contexts

- Censored Data
- Hazard Estimation
- Generalized Likelihood
- Length Biased Estimation

SiZer Extensions: Smoothing Splines

Idea: alternate smoother based on "regularization"

Show SmoothingSplinesFossils.mpg

- smoothing parameter ✍ "scale space"
- Adapted SiZer gives important inference
- finds different features from local linear

Could show SiZerSSgoodSS.eps, SiZerSSbadLL.eps, SiZerSSbadSS.eps, SiZerSSgoodLL.eps

Organization, Section V

Fun with scale space & historical connections

- Heat equation and smoothing
- "Bump monotonicity" of Gaussian kernel
- Connection to the "mode tree"

Important reference:

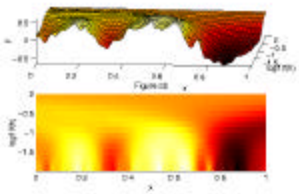
Lindeberg (1994) *Scale space theory in computer vision*, Kluwer: Boston.

Heat equation and smoothing

Paradigm from Image Processing:

Understand smoothing via heat equation

Show HeatEqnColors.eps



"Bump monotonicity" of Gaussian kernel

Statistics: "Silverman's Theorem":

Gaussian Kernel implies "bump monotonicity"

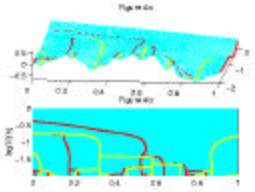
Converse? Well known in scale space theory

1. Axioms ✍ Heat Equation (unique solution)
2. Total Positivity & Semi-Group Property

Connection to the "mode tree"

Minnotte and Scott (1993) *JCGS*, 2, 51-68.

Show ModeTree. eps



"Bump monotonicity" of Gaussian kernel (cont.)

What about the Cauchy kernel?

- Semigroup property
- Minnotte and Scott didn't find non-monotonicity
- Above theories say "no"
- verified with 3 point example

Show CauchyNonVarDim.eps

Organization, Section VI

Concluding Thoughts

- Review usefulness of SiZer
- Contact Information
- Acknowledgements
- Want to try SiZer yourself?

SiZer Windup

Usefulness of SiZer in exploratory data analysis:

- Smoothing experts: saves time
- Smoothing beginners: avoids terrible mistakes
  - don't find things that "aren't there"
  - do find important features
- Directly targets critical scientific question:
  "is a deeper analysis worthwhile?"

Contact Information:

J. S. Marron

Usual mailing address:	8/1/01 – 5/31/01:
Department of Statistics	Dept. Op. Res. & Ind. Eng.
University of North Carolina	Cornell University
Chapel Hill, NC 27599-3260	Ithaca, NY

Email: marron@stat.unc.edu

Web Page: http://www.stat.unc.edu/faculty/marron.html

Published Papers

Main SiZer paper:

Chaudhuri & Marron (1999) *JASA*, 94, 807-823

Main Scale Space paper:

Chaudhuri & Marron (2000) *Ann. Stat.,* 28, 408-428

Variations:

to appear

This talk

PDF version, and graphics (.ps , .pdf and .mpg):

http://www.unc.edu/depts/statistics/postscript/papers/marron/ASAContEd/

Want to try SiZer yourself?

Matlab version:

http://www.stat.unc.edu/faculty/marron/marron_software.html

JAVA version (demo, beta): Follow the SiZer link from the Wagner Associates home page:

http://www.wagner.com/www.wagner.com/SiZer/

Show SiZerDownload.html

Acknowledgements:

Core research on SiZer (and variations) was supported by:

NSF Grants DMS-9504414 & DMS-9971649

The JAVA and C development of SiZer, was done by Molly Megraw of Daniel H. Wagner and Associates, Inc.

http://www.wagner.com/

That development, and this presentation, was supported by NIH SBIR Grant # 1 R43 RR16089-01