Nonparametric Inference in Cosmology: Dark Energy and Beyond

Christopher R. Genovese

Department of Statistics

Carnegie Mellon University

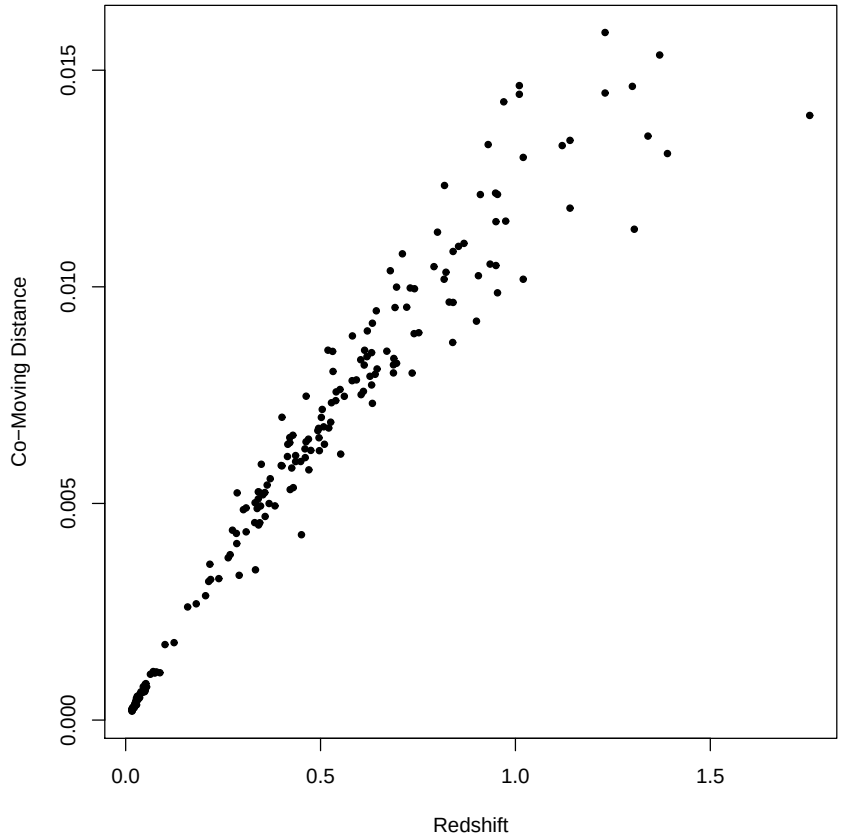
http://www.stat.cmu.edu/~genovese/

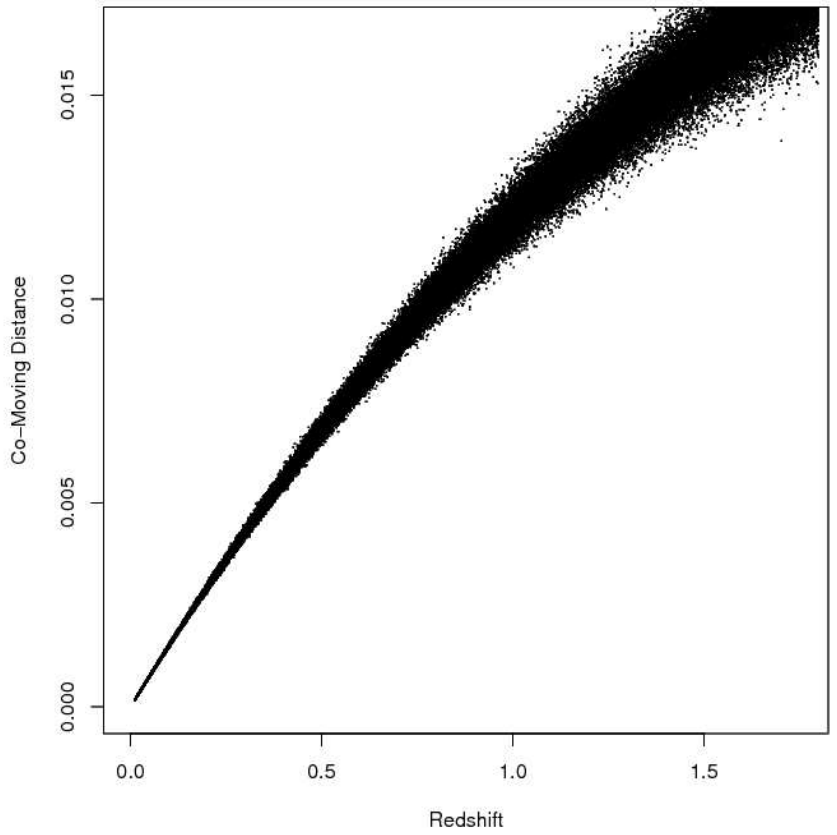
http://incagroup.org/

Collaborators:

Peter E. Freeman (Carnegie Mellon Statistics)
Larry Wasserman (Carnegie Mellon Statistics)
Robert C. Nichol (University of Portsmouth)
Christopher Miller (CTI/NOAO)
Isa Verdinelli (University of Rome/Carnegie Mellon)
Marco Perone-Pacifico (University of Rome)
Michael Wood-Vasey (University of Pittsburgh)

After the Flood . . .





(from Davis et al. 2007)

The Impacts of Massive Data Sets

- Even simple calculations become computationaly challenging.
- Increased power requires flexible models
- Can answer a wider range of questions.
- Move toward asymptotpia – large sample performance relevant to choosing procedures.
- Systematic errors dominate stochastic errors?
- ...

Three Approaches to Inference . . .

. . . in a massive data regime:

- Minimax/Frequentist Nonparametric
- Bayesian Nonparametric
- Statistical and Machine Learning

Road Map

1. An Overview of Nonparametric Regression
2. Inference for the Dark Energy Equation of State
3. Estimating Filaments

Road Map

1. An Overview of Nonparametric Regression
2. Inference for the Dark Energy Equation of State
3. Estimating Filaments

Parametric versus Nonparametric Models

- Parametric Case

  The model is a collection of probabiity distributions P_θ indexed by a parameter θ of fixed dimension.

  When the model is correct, MSE $= O(\frac{1}{\sqrt{n}})$

  But if model is wrong, then model bias will come to dominate with large n.

- Nonparametric Case

  The distributions P_θ are indexed by an infinite dimensional parameter, i.e., a function.

  Put another way, the effective dimension of the model increases with the number of data.

Why Nonparametric?

Goal: make sharp inferences about unknown functions with a minimum of assumptions.

A nonparametric approach is useful

1. when we don't have a well justified parametric (finite-dimensional) model for the object of interest.

2. when we have a reasonable parametric model but have enough data to go after even more detail.

3. when we can do as well (or better) more simply.

4. when we want to assess sensitivity to parametric model assumptions.

The Nonparametric Regression Problem

Observe data (X_i, Y_i) for $i = 1, \ldots, n$ where

$$Y_i = f(X_i) + \epsilon_i,$$

where $\mathsf{E}(\epsilon_i) = 0$ and the X_is can be fixed (x_i) or random.

Leading cases: 1. $x_i = i/n$ and $\mathsf{Cov}(\boldsymbol{\epsilon}) \equiv \Sigma = \sigma^2 I$.

2. X_i IID g and $\mathsf{Cov}(\boldsymbol{\epsilon}) \equiv \Sigma = \sigma^2 I$.

Key Assumption: $f \in \mathcal{F}$ for some infinite dimensional space $\mathcal{F}$.

Examples

1. Sobolev$_p(C)$: $\mathcal{F} = \{f\colon \int |f|^2 < \infty$ and $\int |f^{(p)}|^2 \leq C^2\}$
2. Lipschitz(A): $\mathcal{F} = \{f\colon |f(x) - f(y)| \leq A|x - y|,$ for all $x, y\}$

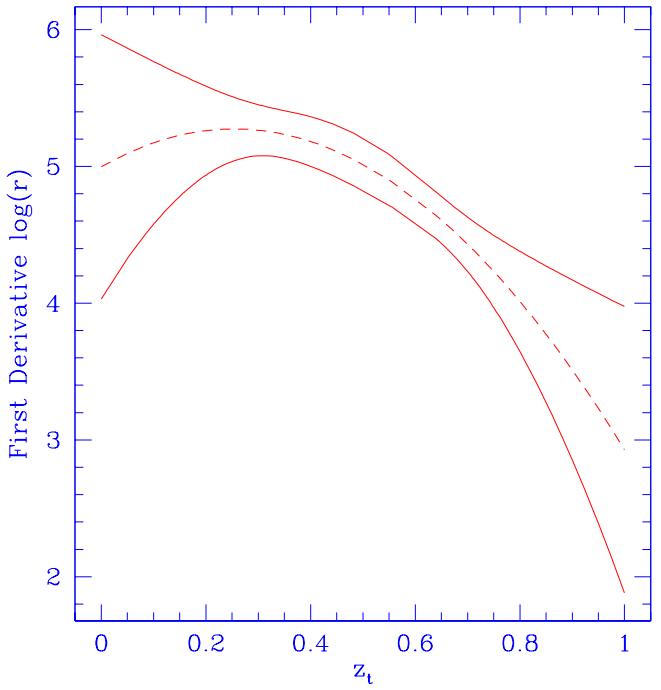
Goal: Make inferences about f or about specific features of f.

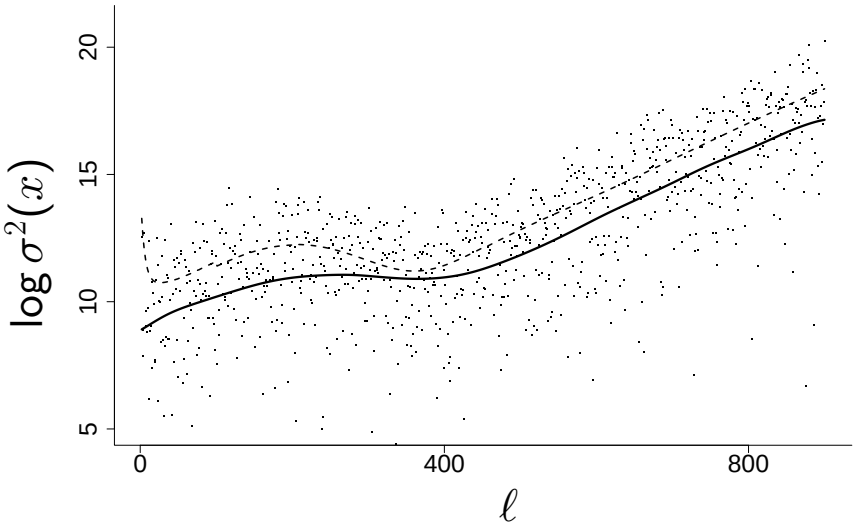
Variants of the Problem

- Inference for Derivatives of f
- Estimating Variance functions
- Regression in High dimensions
- Inverse Problems
- Inferences about specific functionals of f

Related Problems:

  - Density Estimation
  - Spectral Density Estimation





Rate-Optimal Estimators

Choose a performance measure, or risk function, e.g.,
$R(\widehat{f}, f) = \mathsf{E}\int(\widehat{f} - f)^2$ or $R(\widehat{f}, f) = \mathsf{E}|\widehat{f}(x_0) - f(x_0)|^2$

Want $\widehat{f}$ that minimizes worst-case risk over $\mathcal{F}$ (minimax).

But typically must settle for achieving the optimal minimax rate of convergence r_n:

$$\inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} R(\widehat{f}_n, f) \asymp r_n^2$$

In infinite-dimensional problems, typically $r_n\sqrt{n} \to \infty$.

For example, $r_n = n^{-\frac{p}{2p+1}}$ on Sobolev$_p$.

Rate-optimal estimators exist for a wide variety of spaces and risk functions.

Adaptive Estimators

It's unsatisfying to depend too strongly on intangible assumptions such as whether $f \in \text{Sobolev}_p(C)$ or $f \in \text{Lipschitz}(A)$.

Instead, we want procedures that *adapt* to the unknown smoothness.

For example, $\widehat{f}_n$ is a *(rate) adaptive procedure* over the Sobolev_p spaces if, when $f \in \text{Sobolev}_p$,

$$\widehat{f}_n \to f \text{ at rate } n^{-p/2p+1}$$

without knowing p.

Adaptive estimators exist over a variety of function families and over a range of criteria.

Adaptive confidence sets? Harder.

Common Smoothing Methods

- (Quasi-) Linear Methods
  - Kernel and Local Polynomial Regression
  - Regularized Maximum Likelihood

    ($\widehat{f}_n = \arg\min_{\xi} \sum_{i=1}^{n} (Y_i - \xi(X_i))^2 + \lambda Q(f)$, e.g., smoothing splines)
  - Basis Decomposition (Assume $f = \Sigma_k \beta_k \phi_k$ for some basis ϕ_k.)
  - Sieves ($f \in \mathcal{M}_n$ for a sequence of models of increasing dimension.)
- Nonlinear Methods
  - Wavelet Shrinkage
  - Variable Bandwidth Kernels

    (Attempt to adapt spatially by changing smoothing over domain; appealing but hard.)
- Others
  - Scale-Space Methods (e.g., SiZer)

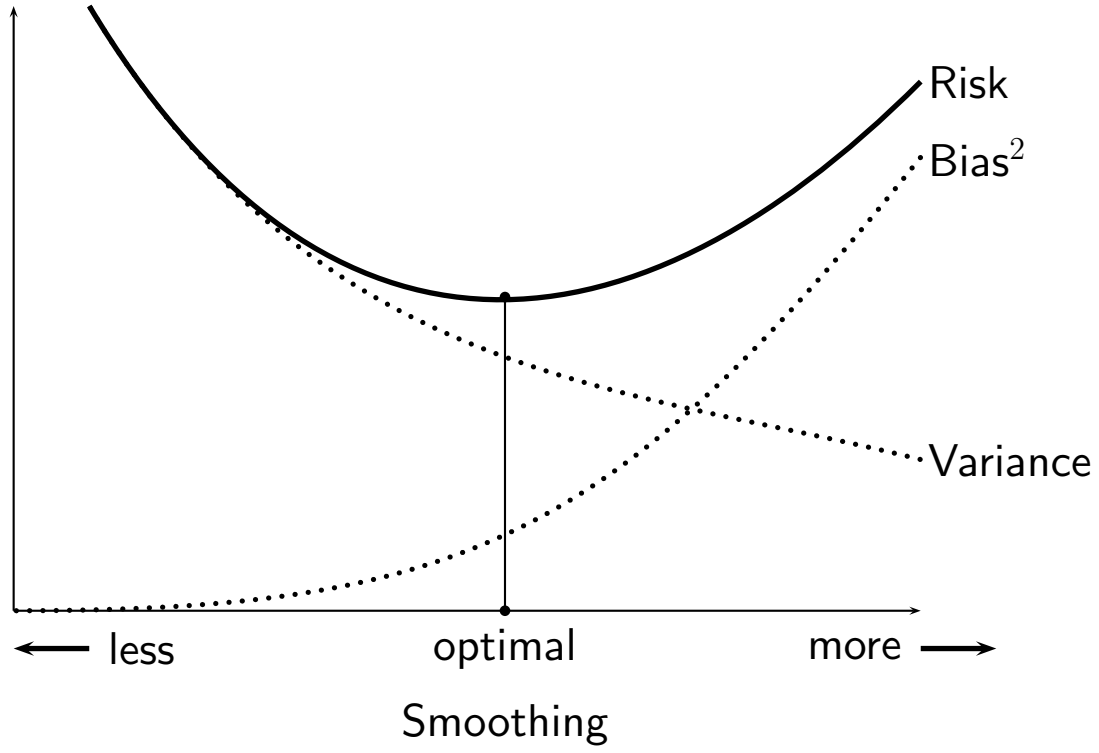
    (Consider all levels of smoothing simultaneously.)

The Bias-Variance Tradeoff

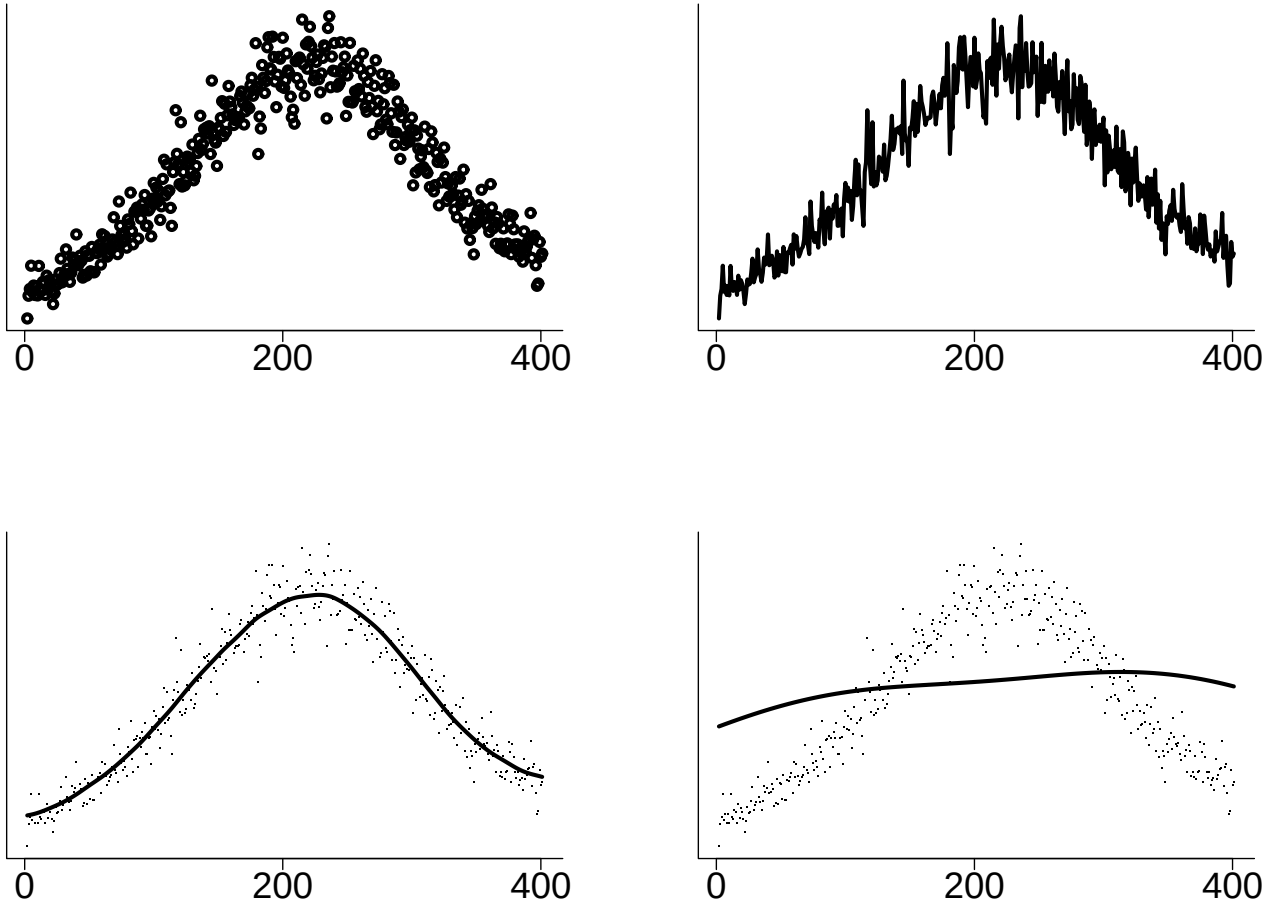
All of these methods introduce a tuning parameter that determines the complexity of the fit.

Key goal: Choose the correct level of smoothing.

$$R(\widehat{f}, f) = \mathsf{E} \int (f(x) - \widehat{f}(x))^2 \, dx = \int \mathsf{bias}^2(x) \, dx + \int \mathsf{variance}(x) \, dx$$



The Bias-Variance Tradeoff (cont'd)



Choosing the Smoothing Parameter

1. Plug-in

Asymptotically optimal bandwidths are often of the form

$h_* = C(f)r_n$ for $r_n \to 0$. Then use $\hat{h}_{\mathrm{PI}} = C(\widetilde{f})r(n)$ for a pilot estimate $\widetilde{f}$.

Do not perform well in general (Loader 1999).

2. Cross-Validation

Divide $\{1, \ldots, n\}$ into m disjoint subsets $S_1, \ldots, S_m$. Choose h to minimize

$$\widehat{\mathsf{PE}}(h) = \frac{1}{m} \sum_{\ell=1}^{m} \frac{1}{\#(S_\ell)} \sum_{i \in S_\ell} (Y_i - \widehat{f}^{[-\ell]}(X_i))^2,$$

where $\widehat{f}^{[-\ell]}$ uses *the original procedure* omitting all (X_i, Y_i) with $i \in S_\ell$.

3. Risk Estimation

Find $\widehat{R}(h)$ such that $\mathsf{E}\widehat{R}(h) = R(\widehat{f}_h, f)$. and choose $\hat{h}$ to minimize $\widehat{R}(h)$.

In many cases, can show that $\hat{h}$ approximates true minimizer.

Whatever the method, this is a hard problem.

Confidence Sets

- For inference, we need more than an estimator $\widehat{f}$;
  we need an assessment of uncertainty.

  We want to make claims relating to features of f:
  shape, magnitude, peaks, inclusion, derivatives.

- To do this, construct a $1-\alpha$ confidence set for f:

$$\text{a random set } \mathcal{C} \text{ such that } \mathsf{P}\{\mathcal{C} \ni f\} = 1-\alpha.$$

  $\mathcal{C}$ is the set of functions or vectors of the form $(f(x_1), \ldots, f(x_n))$.

- For nonparametric procedures, we prefer uniform coverage:

$$\sup_{f \in \mathcal{F}} \left| \mathsf{P}\{\mathcal{C}_n \ni f\} - (1-\alpha) \right| \to 0.$$

  This ensures that the coverage error depends only on n, not on f.

Road Map

1. An Overview of Nonparametric Regression
2. Inference for the Dark Energy Equation of State
3. Estimating Filaments

Preliminaries (for Statisticians)

- The Expanding Universe

  Scale factor $a(t)$ gives size of any region at time t relative to size at current time t_0.

  Redshift z is an observable shift in the wavelength of light from a distant object that is induced by the expansion of the universe.

$$1 + z = \frac{\lambda_{\rm obs}}{\lambda_{\rm emit}} = \frac{a(t_{\rm obs})}{a(t_{\rm emit})}.$$

  Hubble parameter $H(t) = \dfrac{\dot{a}(t)}{a(t)}$. ($H_0 = H(t_0)$ is the Hubble "constant".)

- Comoving distance

  Hubble's Law, $v = H_0 d$, defines a *comoving* reference frame in which an observer at rest sees galaxy recession velocities proportional to distance in all directions.

  The comoving distance, r, between two nearby comoving observers is the tape-measure distance between them divided by the scale factor at the time they are measured.

- The Distance-Redshift Relation

  The relationship between objects' distances and redshifts contains fundamental information about the Universe's geometry.

$$r(z) = \int_0^z \frac{{\rm d}z'}{H(z')}.$$

An Accelerating Universe

- Type Ia Supernovae (SNe) offer standard candles (at least standardizable candles) that can help probe the distance-redshift relation.

  Potential systematic errors are thought to be small.

- In 1998, two groups observing type Ia SNe detected an accelerating expansion. (Reiss et al. 1998, Perlmutter et al. 1998)

- Several independent lines of evidence (e.g., CMB, large-scale structure, X-ray clusters) support the hypothesis that the universe is both flat and not entirely composed of matter.

  Indeed, these studies suggest that the density of matter (dark and light) is $\Omega_{\mathrm{M}} \approx 0.25$ times the critical density ($\rho_{\mathrm{crit}} = 3H^2/8\pi G$).

- Taken together, current data strongly rule out a non-accelerating model in comparison to a simple accelerating model.

  This also suggests the existence of another field ρ_{DE} with density $\Omega_{\mathrm{DE}} \approx 0.75$ times the critical density.

Dark Energy

- Dark Energy is a smoothly-distributed energy density characterized by its negative pressure.

  It dominates the mass-energy density of the universe ($\sim 74\%$ versus $\sim 4\%$ for baryonic matter) and acts in opposition to gravity.

- We can quantify the evolution of dark energy using either the energy density $\rho_{\mathrm{DE}}(z)$ or the equation of state parameter $w(z)$.

  Let p_{DE} and ρ_{DE} be the dark energy pressure and energy density, then

  $$p_{\mathrm{DE}} = w\rho_{\mathrm{DE}}.$$

  Note: the equation of state parameter w is a function.

- In the special case where w is a constant w_0

  $$\rho_{\mathrm{DE}} \propto a^{-3(1+w_0)}.$$

  For the cosmological constant model, $w(z) \equiv -1$.

Type Ia Supernovae Data

We analyze data for 192 SNe Ia from Davis et al. 2007.
(See also Wood-Vasey et al 2007 and Reiss et al 2006.)

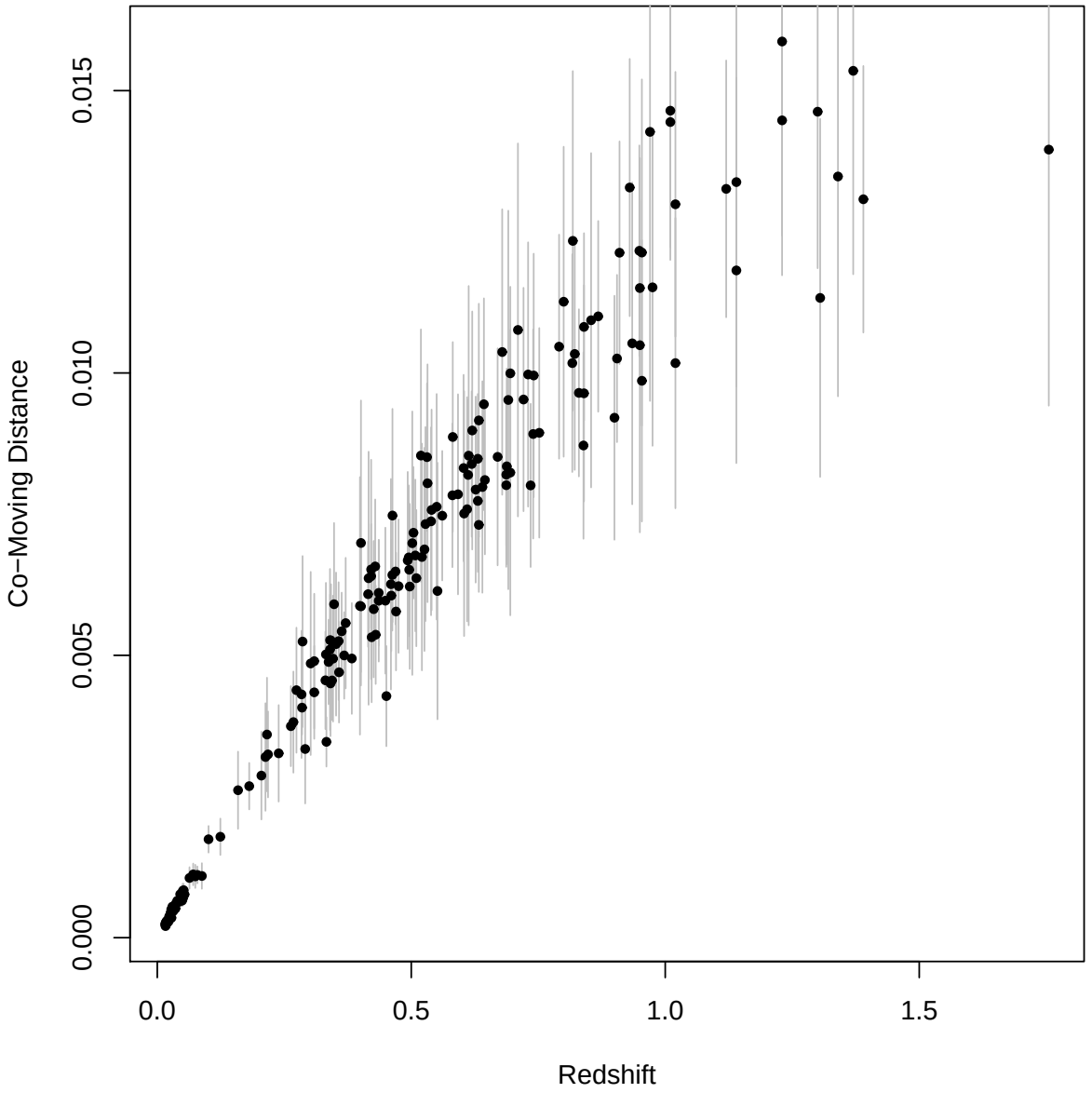
The data consist of

i) redshifts z,

ii) distance moduli $\mu = m - M$, and

iii) standard errors for each SN,
which we take as known.

Data (cont'd)



A Nonparametric Regression Problem

- After transformation and under mild assumptions, we can express the data as

$$Y_i = \log_{10} r(z_i) + \sigma_i \epsilon_i, \qquad i = 1, \ldots, n,$$

  where

  – the σ_i's are taken as known;

  – the ϵ_i's are mean zero, Gaussian;

  – $r(z_i)$ is the co-moving distance at redshift z_i; and

  – the data ordered are ordered by z_i rather than SN date.

- This gives a regression problem in terms of co-moving distance.

Inverse Problem Formulation

- Co-moving distance is related to the equation of state w by a non-linear operator

$$r(z) = H_0^{-1} \int_0^z ds \left[\Omega_{\mathrm{M}}(1+s)^3 + (1-\Omega_{\mathrm{M}})(1+s)^3 \, e^{-3\int_0^s \frac{-w(u)}{1+u}\, du} \right]^{-\frac{1}{2}}.$$

  where H_0 is the Hubble constant and Ω_{M} is the density of matter relative to the critical density.

- Specifically, the right-hand side defines an operator T such that

$$r = T(w; H_0, \Omega_{\mathrm{M}}).$$

- Inference for w from Y is thus a nonlinear inverse problem

$$Y_i = T(w; H_0, \Omega_{\mathrm{M}})(z_i) + \sigma_i \epsilon_i.$$

Fundamental questions

Statistical inferences for w can address several critical questions:

1. Are the data consistent with the cosmological constant model $w \equiv -1$?
2. If not, do the data require that w varies with time/redshift?
3. If so, how well can we estimate w?
4. Do the data rule out any competing cosmological models?

Determining if $w \equiv -1$ is currently the key task.

But ruling out competing models and obtaining sharp estimates of w will help constrain theoretical explanations of dark energy.

Consensus Findings

Studies to date, combining several types of observations, have developed a robust consensus around several findings:

1. There is strong evidence that the universe is accelerating.
2. Under the current theoretical framework, there is strong evidence that dark energy exists with $\Omega_{\rm DE} \approx 0.76$.
3. The best available estimates suggest $w \approx -1$. That is, the current data are well fit by the cosmological constant model.

See Frieman et al. 2008 for a comprehensive review.

Attention is therefore focused on future data sets, which promise to be much larger.

Approaches to Inference for Dark Energy

Many approaches have been used, but there have been three main threads:

1. Parametric family $w(z;\theta)$ mapped through forward problem and fit by ML or similar criterion.

   (e.g., Barboza and Alcaniz 2008, Liu et al. 2008, ...)

   The most common parameterizations are constant $w(z) = w_0$, linear in the scale parameter $w(z) = w_0 + w_1 z/(1+z)$, or piecewise constant.

2. Nonparametric fit to r and its derivatives and estimate w by the "reconstruction equation"

$$w(z) = \frac{H_0^2\Omega_m(1+z)^3 + \frac{2}{3}(1+z)r''(z)/(r'(z))^3}{H_0^2\Omega_m(1+z)^3 - 1/(r'(z))^2} - 1.$$

   (e.g., Daly et al. 2008)

3. Eschew the equation of state parameterization, re-expressing T in terms of the energy density $\rho_{\rm DE}$ (e.g., Wang and Mukherjee 2004) or estimating various kinematic parameters directly (e.g., Shapiro and Turner 2006).

(See also Huterer and Starkman 2003 and Saini et al. 2004 for another approach.)

Objectives and Contributions

Our goals:

1. Use the data to distinguish among competing cosmological models for dark energy.
2. Derive adaptive, rate-optimal estimators for w in the inverse problem.
3. Construct tight confidence bounds on w or features of w.

Key point: We have very little *a priori* information constraining the structure of w.

Contributions from this work (Genovese et al. 2009):

1. Nonparametric hypothesis tests that can distinguish among competing cosmological models with minimal assumptions about the form of w.
2. A framework for nonparametric estimation of w with assessment of uncertainty.

Shape Constraints

The structure of $T(w; H_0, \Omega_{\mathrm{M}})$ gives us several useful features:

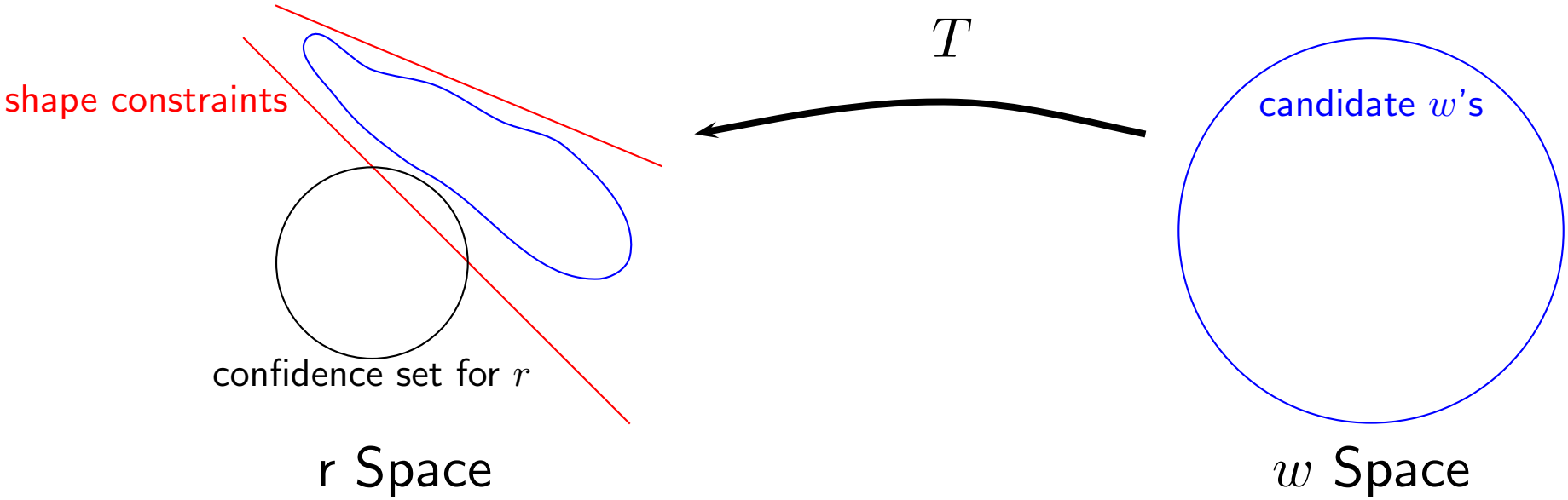
1. $r(0) = 0$
2. $(1+z)^{-3/2}/H_0 \leq r'(z) \leq (1+z)^{-3/2}/\sqrt{H_0^2\Omega_{\mathrm{M}}}$.

   In particular, r is monotone increasing.
3. When $w > -1/(1-\Omega_{\mathrm{M}})$, r is concave.
4. T is monotone decreasing in w.

These give surprisingly strong constraints on the inferences, and can be made stronger with additional assumptions.

Methods I: Hypothesis Testing

- These shape constrains make it possible to test among cosmological models without assuming a parametric form for w or computing a preliminary estimator.
- Basic idea: map a set of candidate w's into the r domain and test by inverting a nonparametric confidence set for r.



Methods I: Hypothesis Testing (cont'd)

- Applicable hypotheses including the following:
  - Simple equalities for w: $w = w_0$, (e.g., cosmological constant, defect)
  - Inequalities for w: $w_0 \leq w \leq w_1$, (e.g., quintessence, cardassian)
  - Inequalities for w': $w'_0 \leq w' \leq w'_1$,
  - Inclusion: $w \in V$ for a linear space V, and

  Hypotheses of these forms correspond directly to several cosmological models (e.g., Caldwell and Linder 2005; Zlatev, Wang, Steinhardt 1999; Freese and Lewis 2002).

- For example, thawing models in quintessence must satisfy

$$\frac{1+w(0)}{(1+z)^3} - 1 \leq w(z) \leq \frac{1+w(0)}{1+z} - 1,$$

  when $-1 \leq w \leq -0.8$. (e.g., Caldwell and Linder 2005)

Methods I: Hypothesis Testing (cont'd)

The basic method

0. Select a small $0 < \alpha < 1$.
2. Construct a $1 - \alpha$ confidence set $\mathcal{C}$ for the unknown vector $(r(z_1), \ldots, r(z_n))$.
3. Construct the set R_0 of vectors $(r_0(z_1), \ldots, r_0(z_n))$ where r_0 is a co-moving distance function produced by an equation of state consistent with the null hypothesis
4. Reject the null hypothesis if $\mathcal{C} \cap R_0 = \emptyset$.

In practice, the sets in Steps 1 and 2 need not be constructed explicitly, and the procedure can be made computationally efficient for a broad range of hypotheses.

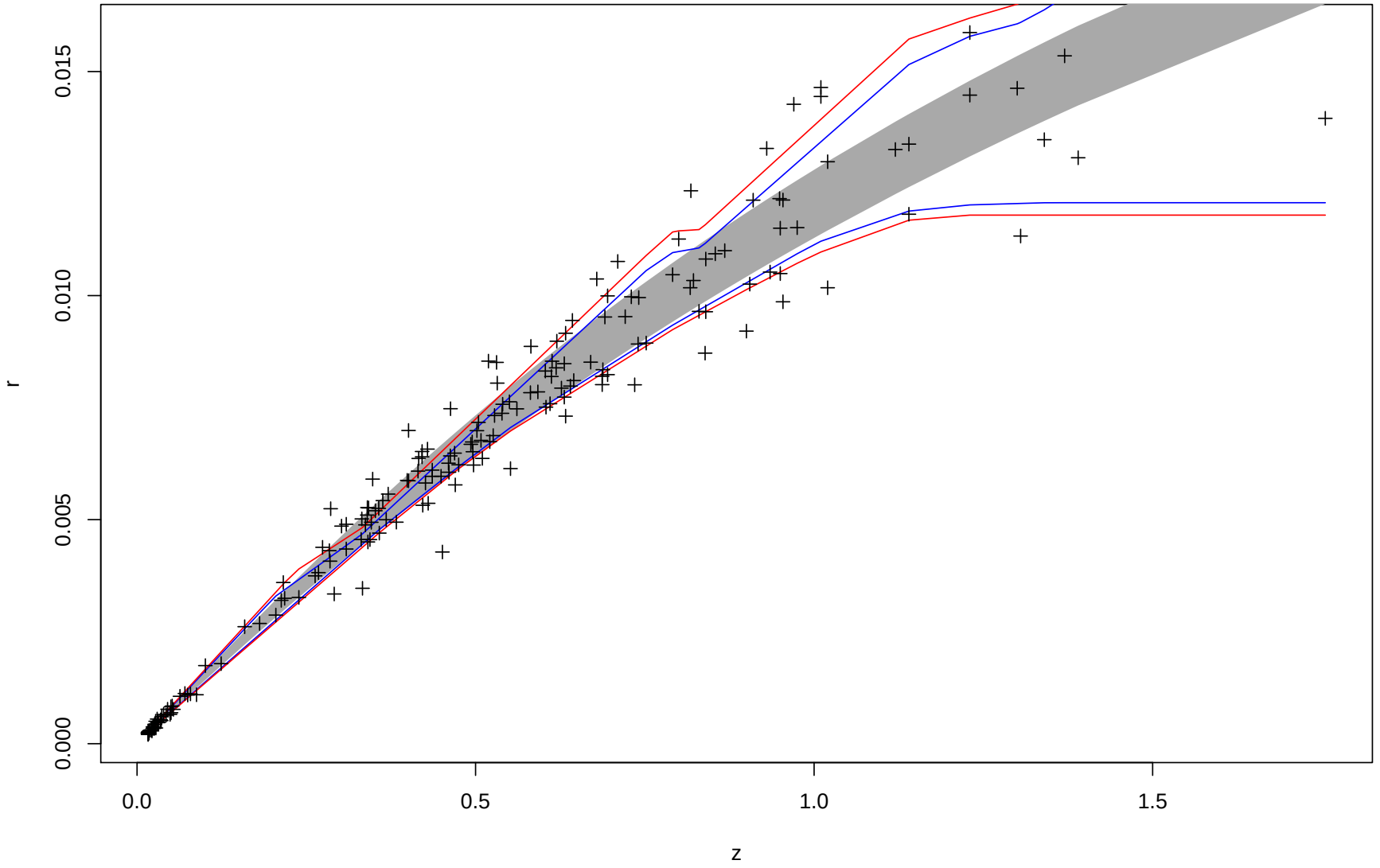
Methods I: Hypothesis Testing (cont'd)

To construct the confidence sets, we use procedures based on Davies et al. 2007 or Baraud 2004 that give much sharper bounds than the standard chi-squared confidence set.

For the Davies et al. procedure,

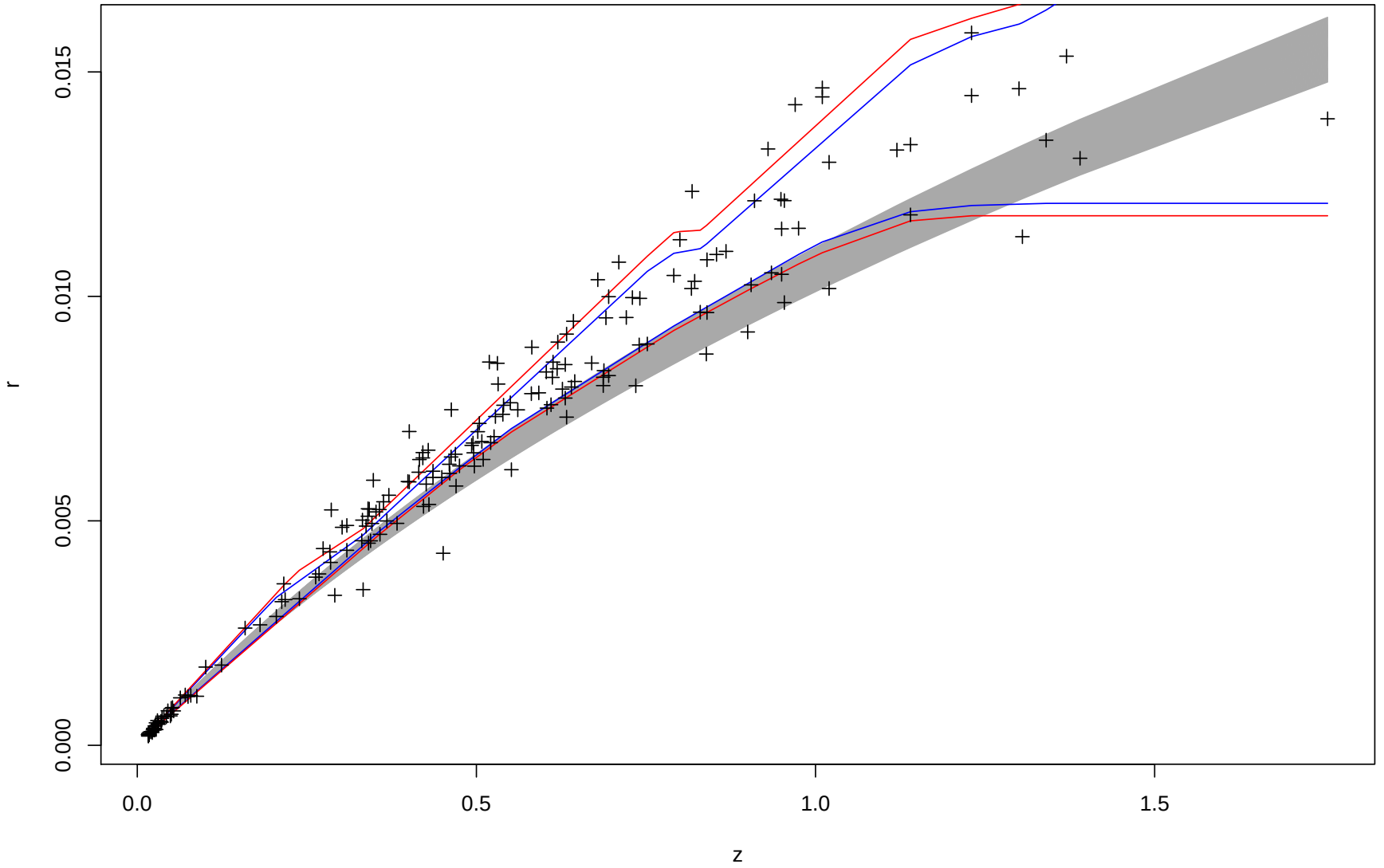
- Define the following for $1 \leq i \leq k \leq n$:
  - $f_i = \log_{10} r(z_i)$ and $f = (f_1, \ldots, f_n)$,
  - $I_{ik} = \{1 \leq j \leq n : \ z_i \leq z_j \leq z_k\}$, and
  - $T_{ik}(f) = \dfrac{1}{\sqrt{\#(I_{jk})}} \sum_{j \in I_{ik}} \dfrac{Y_j - f_j}{\sigma_j}$.
- Construct confidence set $\mathcal{T}$ for the $T_{ik}(f)$'s with linear or quadratic boundaries.
- The shape constraints $\mathcal{S}$, such as monotonicity and concavity, are also linear.
- $\mathcal{C} = \{\text{vectors } g \text{ such that } T_{ik}(g) \in \mathcal{T} \text{ and } g \in \mathcal{S}\}$
- In practice, easier to use confidence bands derived from $\mathcal{T}$ rather than $\mathcal{T}$ itself.

Results I: Hypothesis Testing



Thawing model — 95% — 68% — R_0

Results I: Hypothesis Testing (cont'd)



Results I: Hypothesis Testing (cont'd)

Model	Rejected At Level			
	32%	13%	5%	1%
Cosmological Constant	no	no	no	no
Frustrated Cosmic Strings	**yes**	**yes**	**yes**	no
Domain Walls	**yes**	no	no	no
Matter Dominated	**yes**	**yes**	**yes**	**yes**
Quintessence Thawing	no	no	no	no
Quintessence Freezing	no	no	no	no
Constant w	no	no	no	no

As expected, the cosmological constant model cannot be ruled out with current SNe data alone.

Current data do not allow sharp distinctions among the stronger models.

Note: These results are as good as what one gets under very optimistic assumptions (e.g., known parametric form).

Methods II: Estimation

To estimate w, again use the relation $r = T(w; H_0, \Omega_{\rm M})$.

For a given parameteric model for w,

$$w(z) = -\sum_j \beta_j \psi_j(z),$$

we get a nonlinear, parametric form for r:

$$r(z) = H_0^{-1} \int_0^z ds \left[\Omega_m(1+s)^3 + (1-\Omega_m)(1+s)^3\, e^{-3\sum_j \beta_j \widetilde{\psi}_j(s)}\right]^{-\frac{1}{2}},$$

where $\widetilde{\psi}_j(s) = \int_0^s \psi_j(u)/(1+u)du$.

This gives a likelihood over w, $\Omega_{\rm M}$, and H_0.

We can fit this efficiently, and the results automatically satisfy the shape constraints.

Methods II: Estimation (cont'd)

- But we want a nonparametric model. We use a sieve procedure:

  A. Devise a collection parametric models of increasing dimension $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \ldots$.

  B. Generate an estimator in each model.

  C. Then choose a model to to minimizing a measure of empirical risk (e.g., BIC).

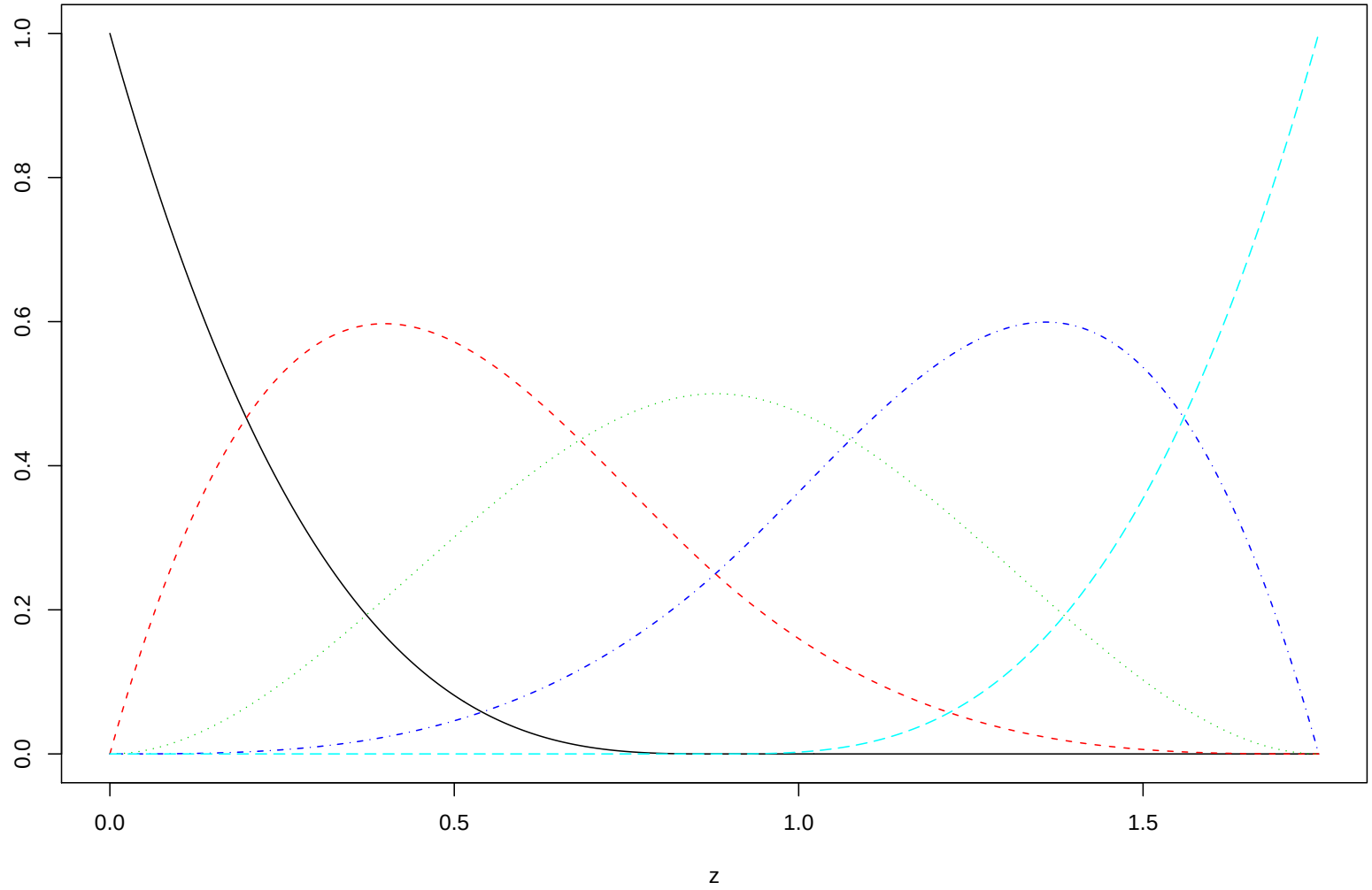
- We generate bootstrap confidence intervals for the parameters by resampling residuals and in turn confidence bands for w.

  This works effectively in practice and in simulations, but more theory is needed to obtain valid uniform confidence sets.

- A key criterion for the choice of the $\mathcal{M}_k$'s is that we want to get a reasonable fit in all models. We use a B-spline basis here.

Example Basis

Even the low dimensional models cover the whole space



Comment on Basis Selection

- Choice of basis in an inverse problem must strike a balance between the information passed by the forward operator and a parsimonious representation of the object (i.e., w).

- For instance, $T(w; \Omega_{\mathrm{M}}, H_0)$ poorly resolves structure at high z.

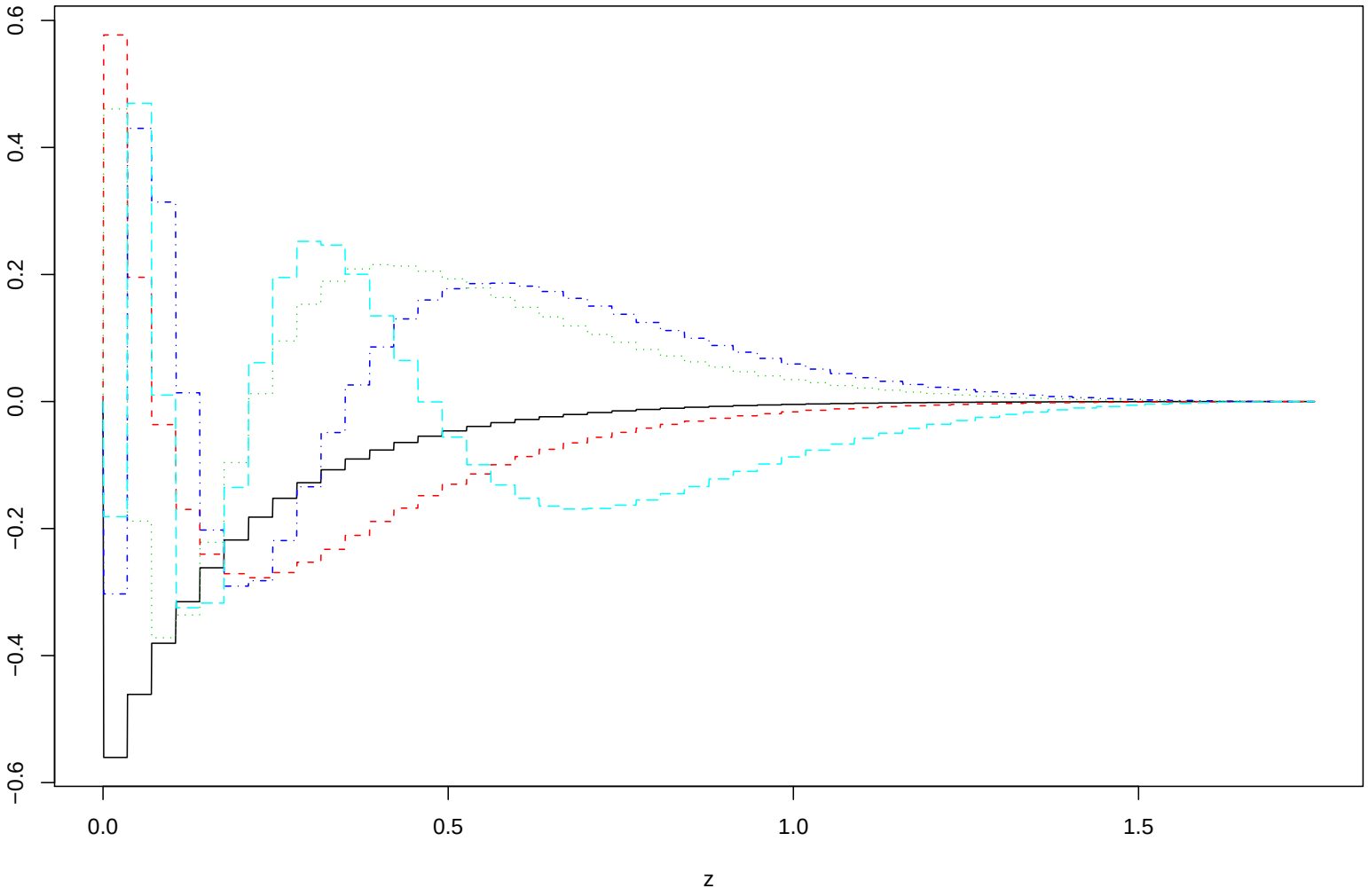
  And smooth w are certainly of interest.

- Example: Construct basis from the eigenvectors of the model's Fisher Information matrix and use K basis elements with largest eigenvalues.

  This basis requires large K to fit smooth models well, resulting in significant variance inflation.

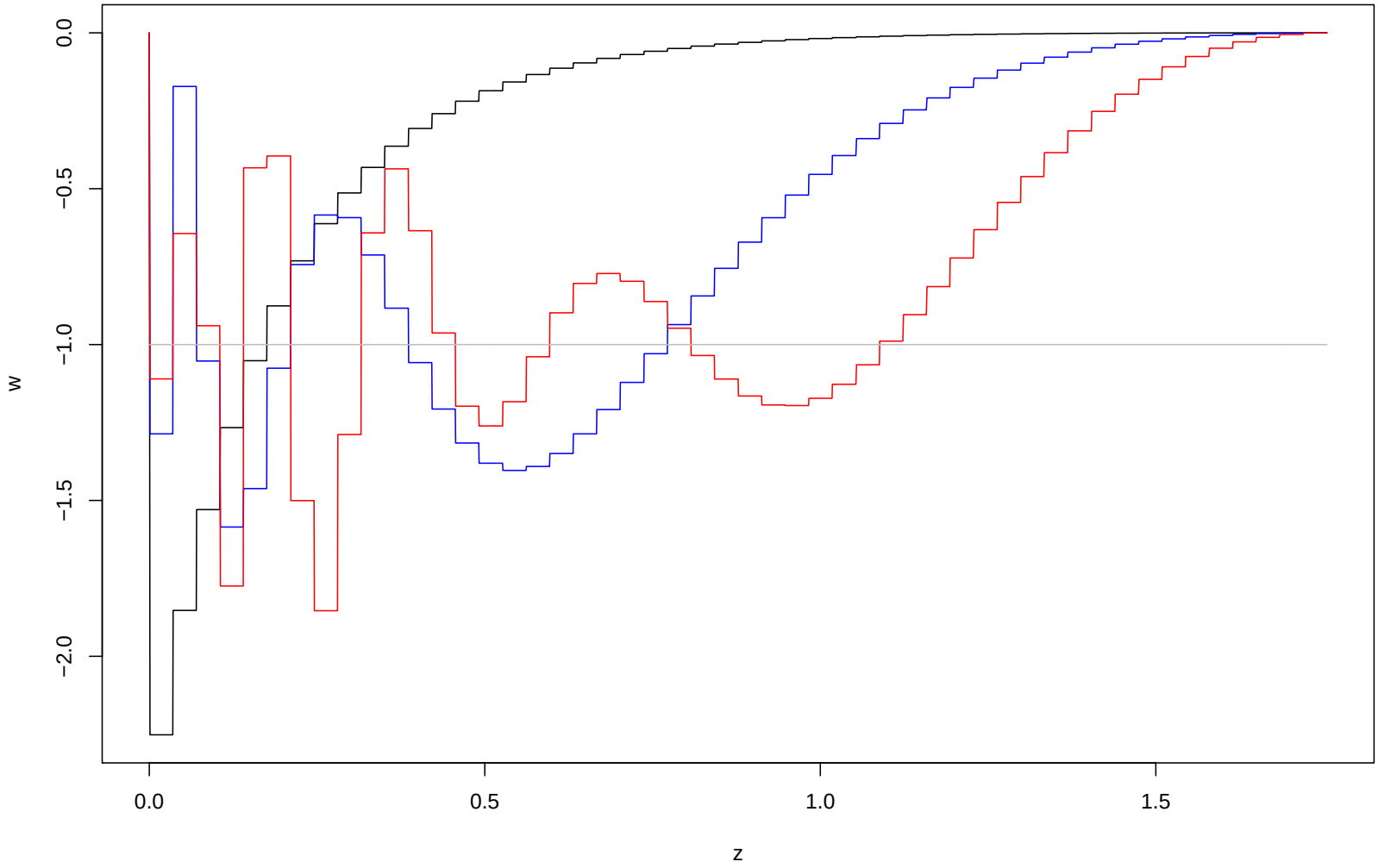
Comment on Basis Selection (cont'd)

Selected Fisher basis elements (index $1, 5, 10, 15, 20$):



Comment on Basis Selection (cont'd)

Corresponding best fits with $w \equiv -1$ and basis size $K = 1, 5, 10$.



Results II: Estimation

- BIC increases sharply with model dimension, leading to a constant w
-

Fit	Bootstrap BC_a 95% confidence intervals
$\widehat{w} \equiv -1.013 \pm 0.124$	$\widehat{w} \in [-1.262, -0.796]$
$\widehat{\Omega}_{\rm M} = 0.268 \pm 0.028$	$\widehat{\Omega}_{\rm M} \in [0.220, 0.324]$
$\widehat{H}_0 = 65.6 \pm 0.90^*$.	$\widehat{H}_0 \in [63.9, 67.4]^*$.

(*Note: H_0 has an arbitrary shift, so only relative uncertainty matters.)

- Power for distinguishing between the cosmological constant ($w = -1$) and even other constant w models is low.

- Uncertainty in $\Omega_{\rm M}$ is a primary driver of uncertainty in the results.

- Simulations show sustantial improvements in power with additional data. Combining with other data (e.g., LSS) improves power further.

Take Home Points/Next Steps

- Nonparametric methods are essential here.
  We know very little a priori about w.

- Current data are insufficient to conclusively distinguish among competing models. This should change with new data coming down the pipeline.

  Nonetheless, cosmological constant model fits the current data well.

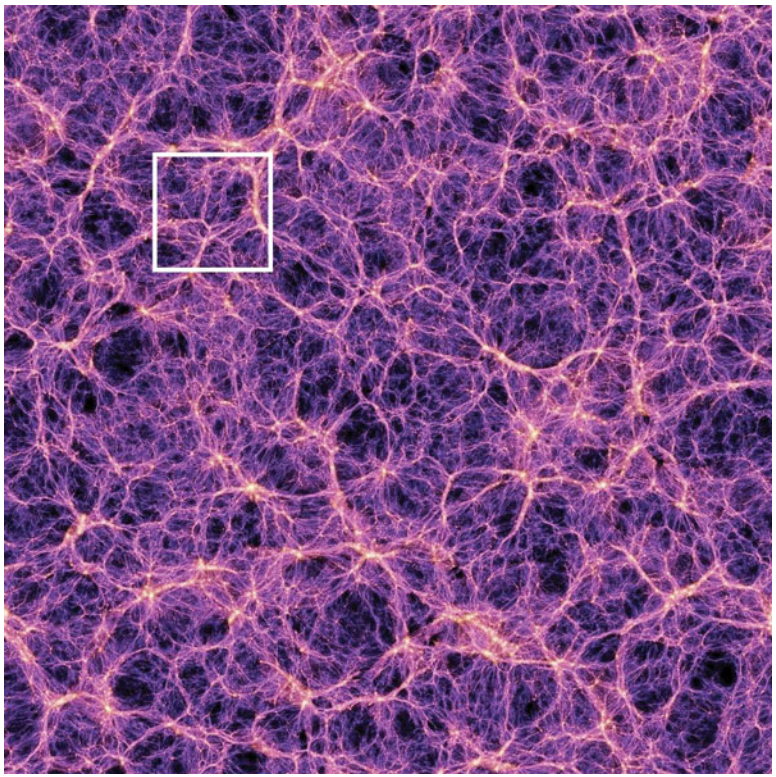
- Our methods can directly test a variety of interesting models with minimal assumptions and provide sharp nonparametric estimates for w.

- Next Steps: confidence procedures, theory, we are continuing analyze spatially distinct ensembles of SNe to test for different patterns of equation-of-state evolution. (collaborators: Peter Freeman, Larry Wasserman, Michael Wood-Vasey)

Road Map

1. An Overview of Nonparametric Regression
2. Inference for the Dark Energy Equation of State
3. Estimating Filaments

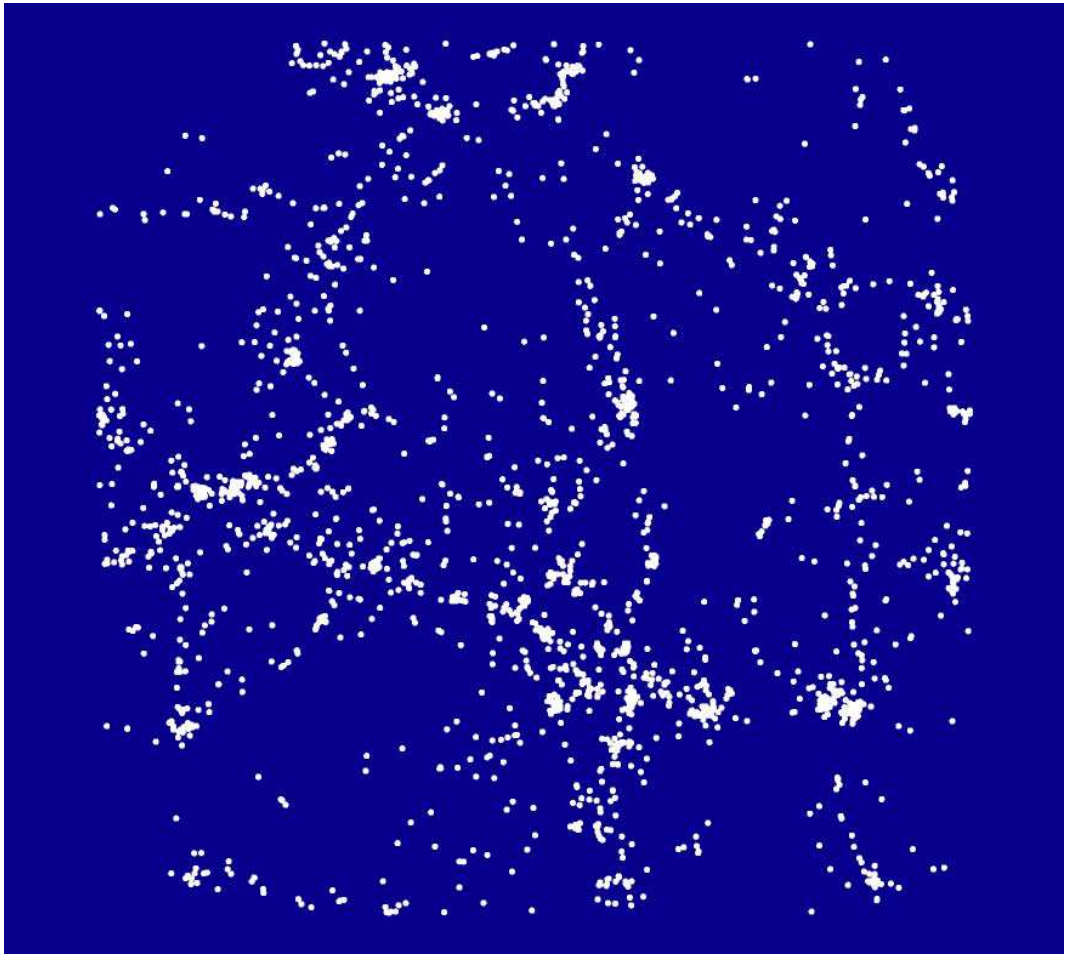
The Cosmic Web



simulated universe from Gnedin (2005)

Estimating the filaments gives a rich description of the universe's large-scale structure

A (small) Current Data Set



data from `www.mpa-garching.mpg.de/galform/agnpaper`

Model for Filaments

Filaments are curves embedded in a point process or random field.

For each $f_j : [0,1] \to \mathbb{R}^2$ in $\{f_1, \ldots, f_m\}$:

1. Draw $U_1, \ldots, U_{n_j}$ from a distribution H on $[0,1]$.
2. For $1 \leq i \leq n_j$, let

$$Y_{ij} = \begin{cases} f_j(U_i) + Z_{ij} & \text{with probability } \pi \\ \text{Uniform} & \text{with probabiity } 1 - \pi, \end{cases}$$

   where the Z_{ij} are drawn IID from a distribution F.

The filaments can be open, closed, or intersecting, though we will typically assume that it is not too rough.

Many Possible Methods

- Principal Curves (Hastie and Stuetzle 1989)
- Second Generation Principal Curves (Smola et al. 1999, Kegl et al. 2000)
- Correlations and Shape Statistics (cf. Martinez and Saar 2002)
- Skeleton Estimation (Novikov et al. 2006, Sousbie et al. 2006)
- Penalized Nonparametric ML (Tibshirani 1992)
- Manifold Learning (Diffusion Maps, ISOMAP, LLE)
- Beamlets (Xuo and Donoho)
- Combinatorial Curve Reconstruction (cf. Dey 2006, Cheng et al. 2004)
- Gradient-based Capture (GPVW 2008)
- Spin and Smooth (GPVW 2009)
- Order and Smooth (GPVW 2009)
- Medial Smoothing (GPVW 2009)

Filaments as a Measurement Error Problem

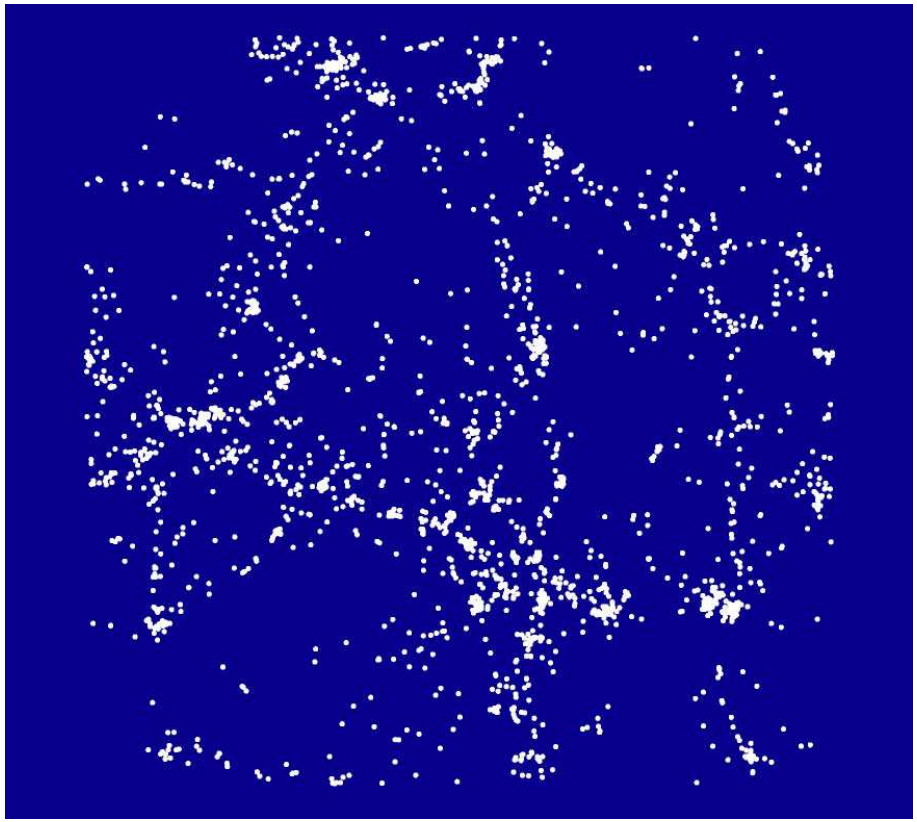
- Even when the filament is a simple function, the filament model is equivalent to a measurement error model, where we observe

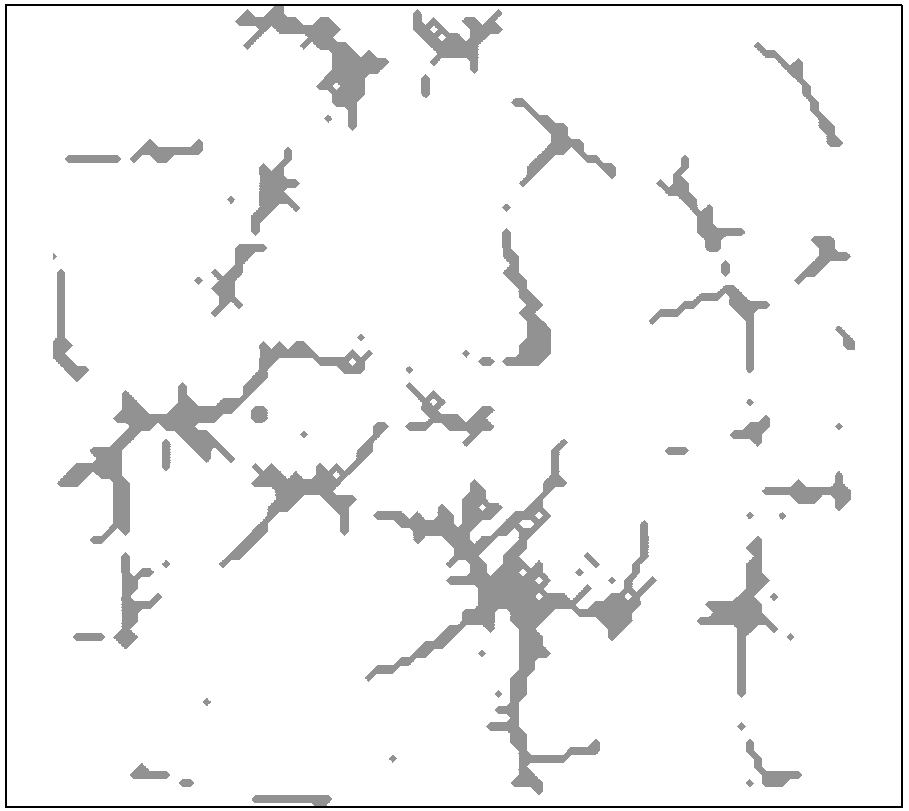
$$Y_i = f(\widetilde{X}_i) + \epsilon_i$$
$$X_i = \widetilde{X}_i + \delta_i.$$

- This tells us that the noise distribution F is critical.
  - When F is Gaussian, the best possible rate of convergence for a *single filament with no background* is **logarithmic** – very poor.
  - When F is Uniform, better rates are possible, but even then simple smoothing incurs significant bias. More care is required.
- Put simply: filament estimation is a subtle problem.

New Method 1: Capturing

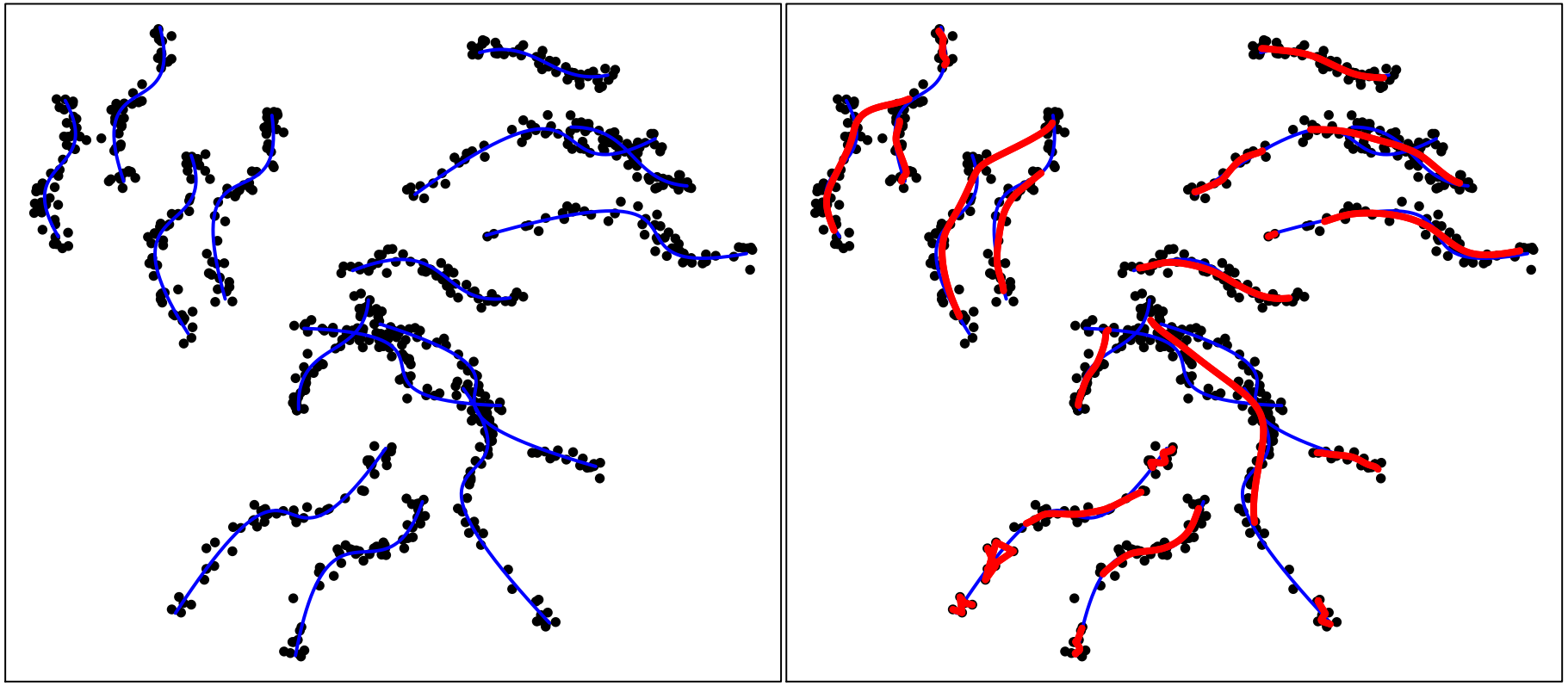
Can construct a function $\widehat{p}_n$ whose level sets capture the filaments with high probability (GPVW 2008).





New Method 2: Vector Quantization

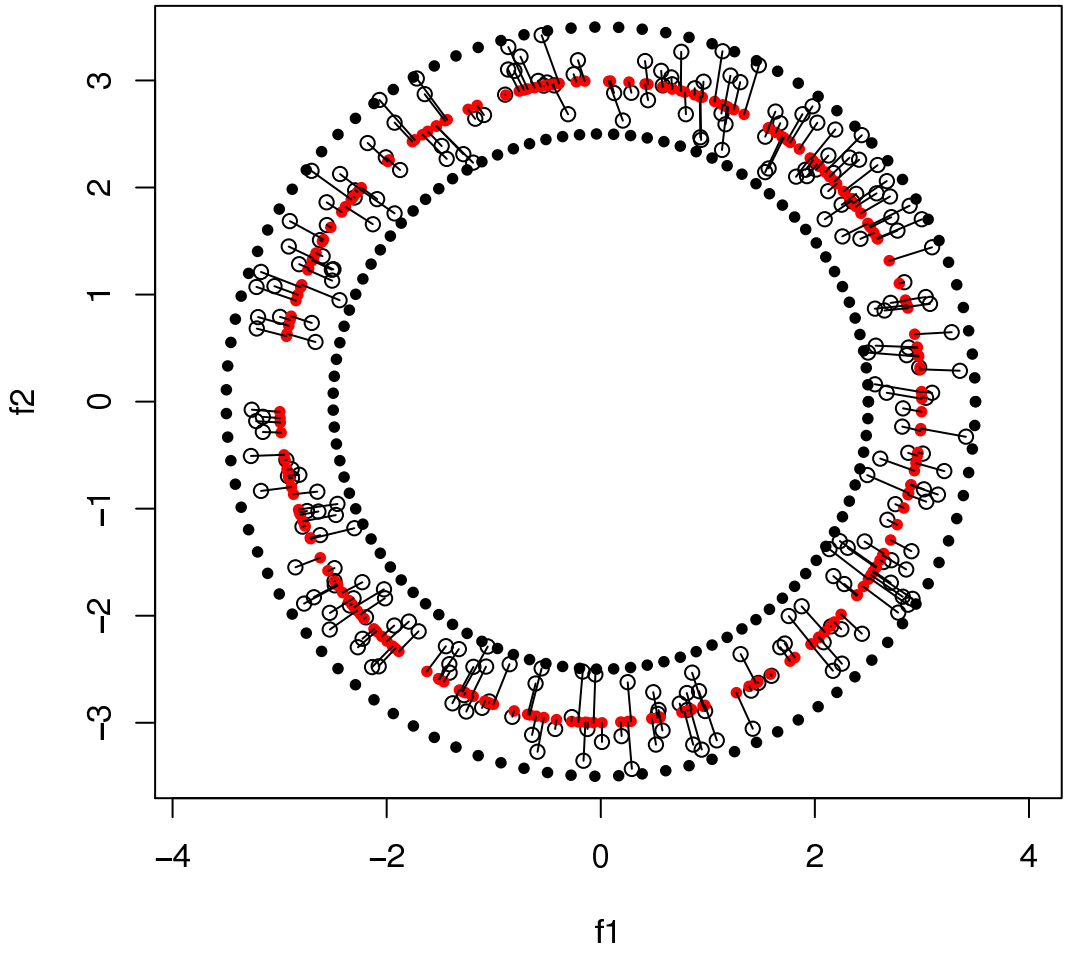
Partition the space into single filament cells, rotate to a principal axis, and smooth.



New Method 3: Medial Smoothing

With Uniform noise, estimate the support of the data density and estimate the medial axis by bias correcting a gradient estimate of the density.

Gives $n^{-1/2}$ rate (?).



Take Home Points/Next Steps

- Filament estimation is a surprisingly difficult problem, even in the single filament case.
- Consistency requires an extra step in addition to smoothing.
- Nonparametric rates are useful for selecting good procedures. Small-sample performance does not distinguish procedures well.
- These methods scale well for large data sets.
- Next Steps: refine theory, incorporate background
  (collaborators: Marco Perone-Pacifico, Isa Verdinelli, Larry Wasserman)

Acknowledgements

This work partially supported by NSF Grant ACI 0121671 and NIH Grant 1 R01 NS047493-01.