NONPARAMETRIC REGRESSION WITH MEASUREMENT ERROR: SOME RECENT PROGRESS

David Ruppert

Cornell University

www.orie.cornell.edu/~davidr

(These transparencies, preprints, and references available — link to “Recent Talks” and “Recent Papers”)

Work done jointly with

Scott Berry, Texas A & M

R. J. Carroll, Texas A & M

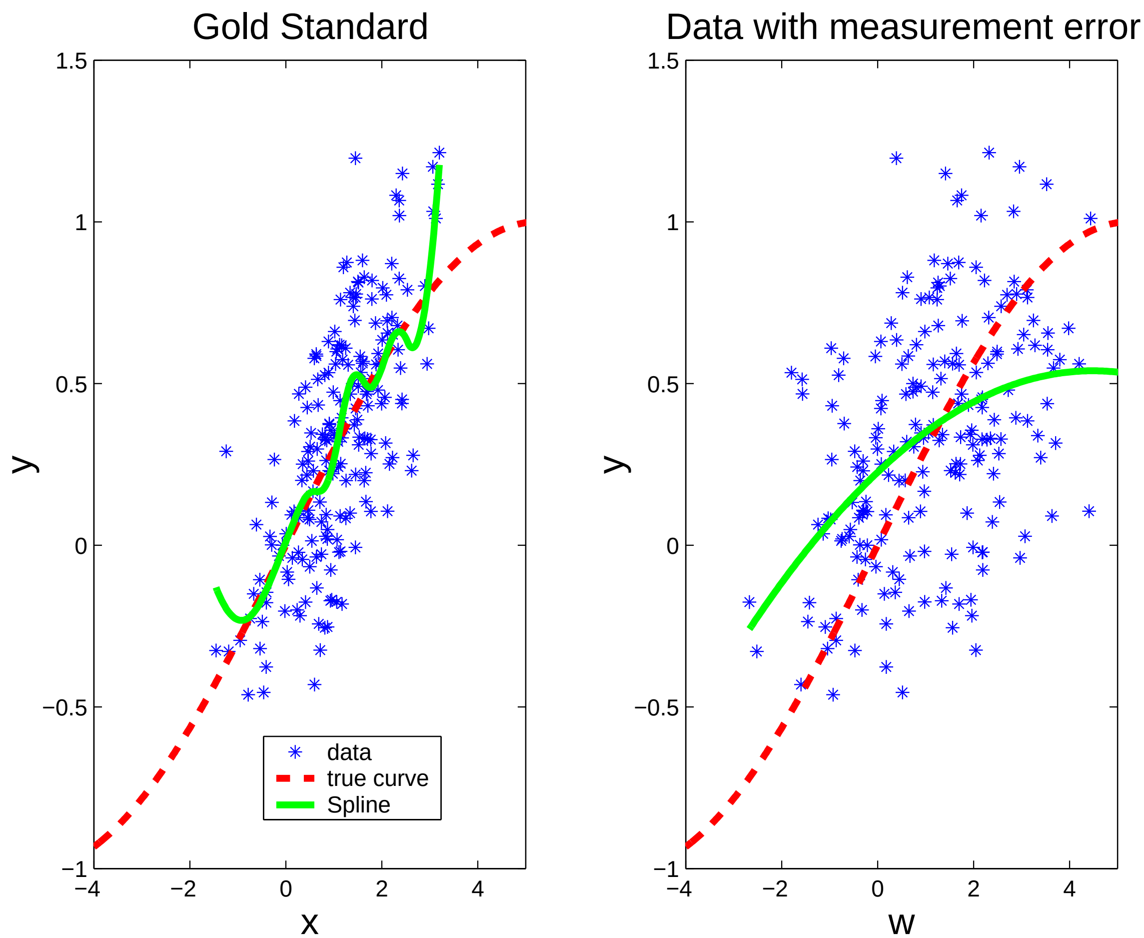
Jeff Maca, Novartis

John Staudenmayer, Cornell

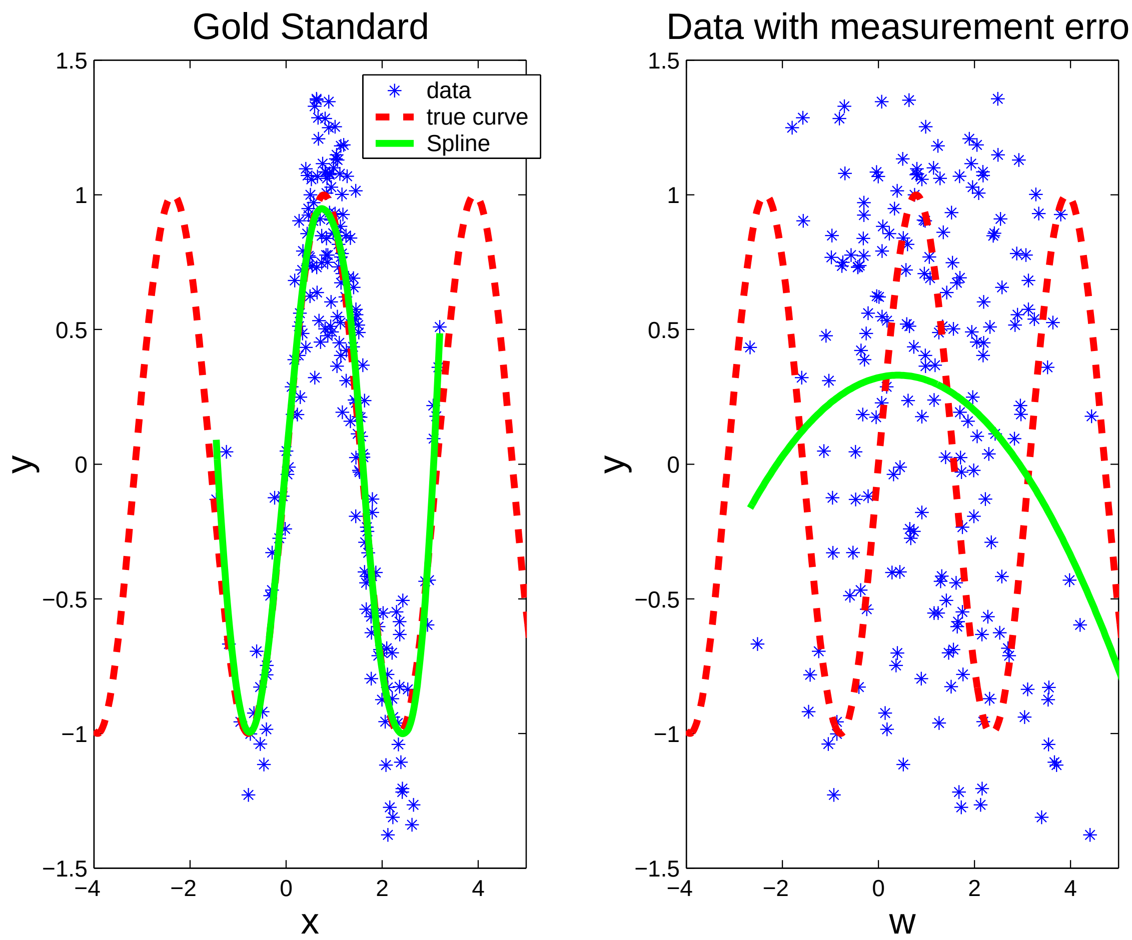
OUTLINE

- Statement of problem — nonparametric regression with measurement error
- Review of the currently available estimators
  - deconvolution kernels (Fan & Truong, 1993, *Annals*)
  - SIMEX (Carroll, Maca, Ruppert, 1999, *Biometrika*)
  - structural splines (Carroll, Maca, Ruppert, 1999)
- New Bayesian spline approach (Berry, Carroll, and Ruppert, 2000)
- Simulation results
- Examples
  - Simulated data
  - Clinical trial of a psychiatric medication

THE PROBLEM OF MEASUREMENT ERROR — ILLUSTRATION



THE PROBLEM OF MEASUREMENT ERROR — ILLUSTRATION



THE PROBLEM OF MEASUREMENT ERROR

- We are interested in nonparametric regression when the predictor X cannot be observed exactly.
  - The regression model is
  $$Y = m(X) + \epsilon$$
  where m is only known to be smooth
  - Observe
  $$Y \text{ and } W = X + U,$$
  where
    * $E(U|X) = 0$
    * var($U|X$) = σ_u^2
    * $U|X$ normally distributed
  - The normality is not important.
- Measurement error variance σ_u^2 is estimated from internal replicate data. (Observe W_{ij}, $j = 1, \ldots, n_i$.)

THE PROBLEM OF MEASUREMENT ERROR, CONT.

- Measurement error occurs in a wide variety of problems.
  - Measuring nutrient intake
  - Measuring airborne lead exposure
  - Measuring blood pressure
  - C_{14} dating
- The effects of measurement error are:
  - biased estimates of the regression curve
  - increase in the perceived variability about the regression line.
- Other than the work of Fan and Truong (1993, *Annals*), there had been little done on nonparametric regression with measurement error until
  - Carroll, Maca, and Ruppert (1999, *Biometrika*) (**CMR**) and
  - Berry, Carroll, Ruppert (2000, submitted) (**BCR**) — available at `www.orie.cornell.edu/~davidr`.

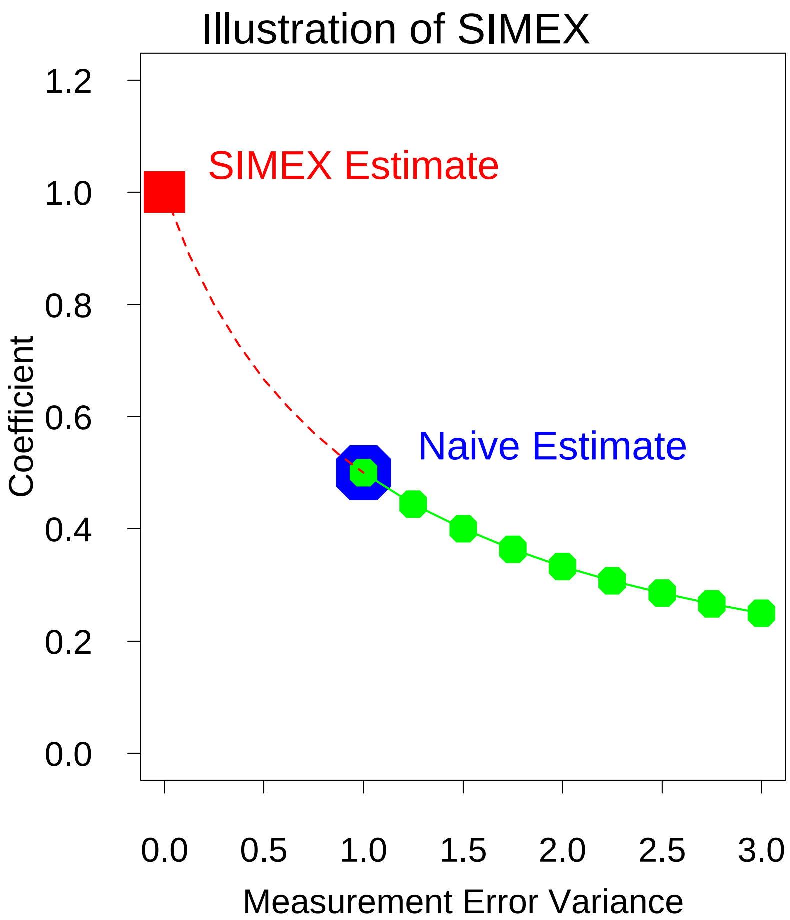
REVIEW OF CURRENT ESTIMATORS

- Globally consistent nonparametric regression by deconvolution kernels (Fan and Truong, 1993, *Annals*)
  - does not work so well
    - Fan & Truong show very poor asymptotic rates of convergence
    - we have simulations showing poor finite-sample behavior
  - no methodology for bandwidth selection or inference
- Standard measurement error method: SIMEX
  - functional — no assumptions on $[X]$
  - very general — can be applied to nearly any measurement error problem, parametric or nonparametric
- Structural Spline
  - Regression splines for basic regression model
  - Mixtures of normals for covariate density model
  - Emphasis is on flexible parametric modeling, not nonparametric modeling. (I believe there is little or no difference in practice.)

SIMEX

- The SIMEX method is due to Cook & Stefanski (1995, *JASA*).
  - The theory is in Carroll, et al. (1996, *JASA*)
  - Also see Carroll, Ruppert, and Stefanski (1995, *Measurement Error in Nonlinear Models*)
- SIMEX has been previously applied to parametric problems.
  - makes no assumptions about the true X's. (*Functional*)
  - results in estimators which are *approximately* consistent, i.e., consistent at least to order $O(\sigma_u^6)$.
- Here is the method, defined via a graph.

SIMEX, ILLUSTRATED



SIMEX

- **CMR** applied the SIMEX to nonparametric regression.
- **CMR** have asymptotic theory in the local polynomial regression (LPR) context.
  - The estimators have the usual rates of convergence.
  - They are approximately consistent, to order $O(\sigma_u^6)$.
- An asymptotic theory with rates seems very difficult for splines
  - but, simulations in **CMR** indicate that SIMEX/splines works a little *better* than SIMEX/kernel
  - problem seems due to undersmoothing
- Staudenmayer (2000, Cornell PhD thesis) is looking at bandwidth selection for SIMEX/LPR.
  - With better bandwidth selection, SIMEX/LPR is competitive with other methods.

STRUCTURAL MODELING

- We can estimate $E(Y|W)$ from our data — use ordinary smoothing.
- But, this is related to the desired function $m(X)$ by

$$E(Y|W) = E\left\{m(X)|W\right\} = \int m(x) f(x|W) dx.$$

- If we had a convenient form for $m(X)$, say $m(X;\boldsymbol{\beta})$, and if we knew $[X|W]$, then we could estimate $m(X;\boldsymbol{\beta})$ by minimizing over the data

$$\sum_{i=1}^{n} \left\{Y_i - \int m(x;\boldsymbol{\beta}) f(x|W_i)\, dx\right\}^2.$$

  – similar to regression calibration which uses the approximation

$$\int m(x;\boldsymbol{\beta}) f(x|W_i)\, dx \approx m\Big(\int x\, f(x|W_i)\, dx\,; \boldsymbol{\beta}\Big)$$

- We need two things to make this work:
  – convenient flexible form for $m(x;\boldsymbol{\beta})$ (must be parametric but flexible enough to be nonparametric for all intents and purposes)
  – convenient flexible distribution for X.

WHAT IF WE HAVE A DISTRIBUTION FOR X?

- Suppose however that we knew $[X]$, and that $[U|X]$ is normal.
  - We in particular know $[X|W]$**!**
  - Could we get anywhere?
- Consider regression splines of order J with K knots:

$$E(Y|X) = m(x;\beta) := \sum_{j=0}^{J} \beta_j X^j + \sum_{j=1}^{K} \beta_{j+J}(X - \xi_j)_+^J$$

- The terms $\xi_1, ..., \xi_K$ are knots.

REGRESSION SPLINES

- Recall from the previous slide:

$$E(Y|X) = m(X;\boldsymbol{\beta}) := \sum_{j=0}^{J} \beta_j X^j + \sum_{j=1}^{K} \beta_{j+J}(X-\xi_j)_+^J$$

- If we know $[X,U]$, and therefore $[X|W]$, then in the observed data we have

$$E(Y|W) = E(m(X;\boldsymbol{\beta})|W)$$
$$= \sum_{j=0}^{J} \beta_j E(X^j|W) + \sum_{j=1}^{K} \beta_{j+J} E\{(X-\xi_j)_+^J|W\}$$

  – This is just a linear model in the β's **!!!!**

  – There are many methods to fit such splines

- The key remaining issue: the joint distribution of X and U.

  – **CMR** used a mixtures of normals for $[X]$ and Gibbs sampling to estimate the parameters.

    ∗ This is an extension to measurement error of an idea of Roeder & Wasserman (*JASA*, 1997).

FULLY BAYESIAN MODEL

What's New?

Answer: Fully Bayesian MCMC method

- In **BCR**
- Uses splines
  - smoothing or penalized
  - P-splines in this talk
- Structural
  - X_i are iid normal
  - but seems robust to violations of normality
- Smoothing parameter is automatic
- Inference adjusts for the data-based smoothing parameter and for measurement error
  - all automatic

FULLY BAYESIAN MODEL — PARAMETERS

- **Regression Model**
  $Y_i = m(x_i; \boldsymbol{\beta}) + \epsilon_i$
  - $m(x_i; \boldsymbol{\beta})$ is a P-spline
  - ϵ_i iid $N(0, \sigma_\epsilon^2)$
- **Measurement Error Model**
  $W_{ij} = X_i + U_{ij}$ where U_{ij} iid $N(0, \sigma_u^2)$
- **Structural Model**
  X_i iid $N(\mu_x, \sigma_x^2)$
- **Parameters:** $\boldsymbol{\beta}, \sigma_e^2, \sigma_u^2, \mu_x, \sigma_x^2$
- **Priors**
  - $\boldsymbol{\beta}$ is $N\Big(0, (\gamma \boldsymbol{K})^{-1}\Big)$ where $\boldsymbol{K}$ is known. [$\alpha := \gamma\sigma_e^2$ is the smoothing parameter.]
  - γ is Gamma(A_γ, B_γ)
  - σ_e^2 is Inv-Gamma(A_e, B_e)
  - σ_u^2 is Inv-Gamma(A_u, B_u)
  - μ_x is $N(d_x, t_x^2)$
  - σ_x^2 is Inv-Gamma(A_x, B_x)
- **Hyperparameters**: $A_e, B_e, A_u, B_u, A_x, B_x, d_x, t_x^2, A_\gamma, B_\gamma$
  - all fixed at values making the priors noninformative
    - E.g., $t_x^2 = 10^6$.

GIBBS SAMPLING

- Iterate through $\boldsymbol{\beta}, \sigma_e^2, \sigma_u^2, \sigma_x^2, \mu_x, \gamma, X_1, \ldots, X_n$.
- Generate each one conditional on the current values of the others.
- All steps except one are easy, either gamma, inverse-gamma, or normal
  - E.g.,

$$
\begin{aligned}
[\boldsymbol{\beta}|\text{other parameters}, \boldsymbol{Y}, \boldsymbol{W}] &\sim \mathrm{Normal} \\
\mathrm{Mean} &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \gamma\boldsymbol{K})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{Y} \\
\mathrm{Cov} &= \sigma_e^2(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \gamma\boldsymbol{K})^{-1}.
\end{aligned}
$$

    * Here $\boldsymbol{X}$ is one of the “other parameters”
    * Essentially we’re fitting a spline to the imputed X’s and the observed Y’s
    * Estimate of $\boldsymbol{\beta}$ is

$$
(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \gamma\boldsymbol{K})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{Y}
$$

averaged over γ and $\boldsymbol{X}$.

GIBBS SAMPLING

- The exception to the sampling being quick and easy is that a Metropolis-Hastings step is needed for $X_1, \ldots, X_n$.

$$
\begin{aligned}
&[X_i | \mu_x, \sigma_x^2, \boldsymbol{\beta}, \sigma_e^2, \sigma_u^2, \boldsymbol{Y}, \mathbf{W}] \\
&\quad \propto \exp\Big(-\frac{1}{2\sigma_u^2} \sum_{j=1}^{m_i} (W_{ij} - X_i)^2 \\
&- \frac{1}{2\sigma_\epsilon^2} \{Y_i - m(X_i; \boldsymbol{\beta})\}^2 - \frac{1}{2\sigma_x^2} (X_i - \mu_x)^2 \Big).
\end{aligned}
$$

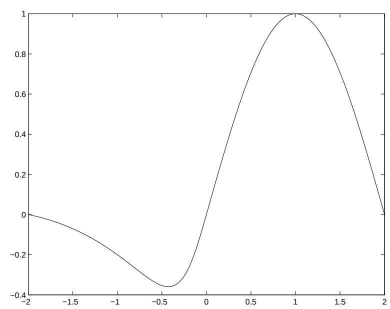
SIMULATIONS

The six cases were considered. $n_i \equiv 2$ in each case.

Case 1 The regression function is

$$m(x) = \frac{\sin(\pi x/2)}{1 + 2x^2\{\text{sign}(x) + 1\}}.$$

with $n = 100$, $\sigma_\epsilon^2 = 0.3^2$, $\sigma_u^2 = 0.8^2$, $\mu_x = 0$ and $\sigma_x^2 = 1$.



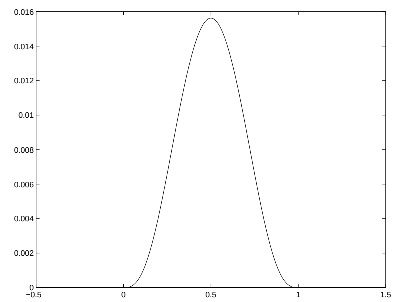
Case 2 Same as Case 1 except $n = 200$.

Case 3 A modification of Case 1 above except that $n = 500$.

Case 4 Case 1 of **CMR** so that

$$m(x) = 1000x_+^3(1-x)_+^3,$$

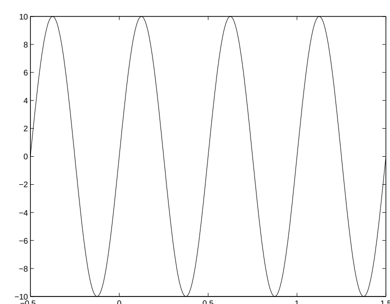
$x_+ = xI(x > 0)$, with $n = 200$, $\sigma_\epsilon^2 = 0.0015^2$, $\sigma_u^2 = (3/7)\sigma_x^2$, $\mu_x = 0.5$ and $\sigma_x^2 = 0.25^2$.



Case 5 A modification of Case 4 of **CMR** so that

$$m(x) = 10\ \sin(4\pi x),$$

with $n = 500$, $\sigma_\epsilon^2 = 0.05^2$, $\sigma_u^2 = 0.141^2$, $\mu_x = 0.5$ and $\sigma_x^2 = 0.25^2$.



Case 6 The same as Case 1 above except that X is a normalized chi-square(4) random variable. (Tests robustness against violation of the structural assumptions.)

Mean Squared Bias $\times 10^2$						
Method	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Naive	5.59	4.92	5.21	1,108	3,733	4.83
Bayes	**0.78**	**0.38**	1.04	17.4	468	1.74
Structural, 5 knots	1.38	0.62	**0.46**	3.7	838	**1.47**
Structural, 15 knots	1.44	0.60	0.66	**3.3**	**226**	1.75
Mean Squared Error $\times 10^2$						
Method	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Naive	6.91	5.57	5.38	1,155	3,793	5.77
Bayes	**2.84**	**1.56**	**1.47**	**195**	1,031	**2.69**
Structural, 5 knots	8.17	3.82	1.73	217	2,032	7.27
Structural, 15 knots	9.90	5.40	1.85	237	**799**	6.94

Table 1: Results based on 200 Monte Carlo simulations for each case. SIMEX was not included in the table — it was not among the best estimators.

EXAMPLE — SIMULATED

- $Y = \sin(2X) + \epsilon$
- X is $N(1, 1)$
- $\sigma_u = 1$
- $\sigma_e = 0.15$
- $n = 201$
- $n_i = 2$ for all i
- 15 knot quadratic P-splines
- 2,000 iterations of Gibbs. First 667 deleted (burn-in period).

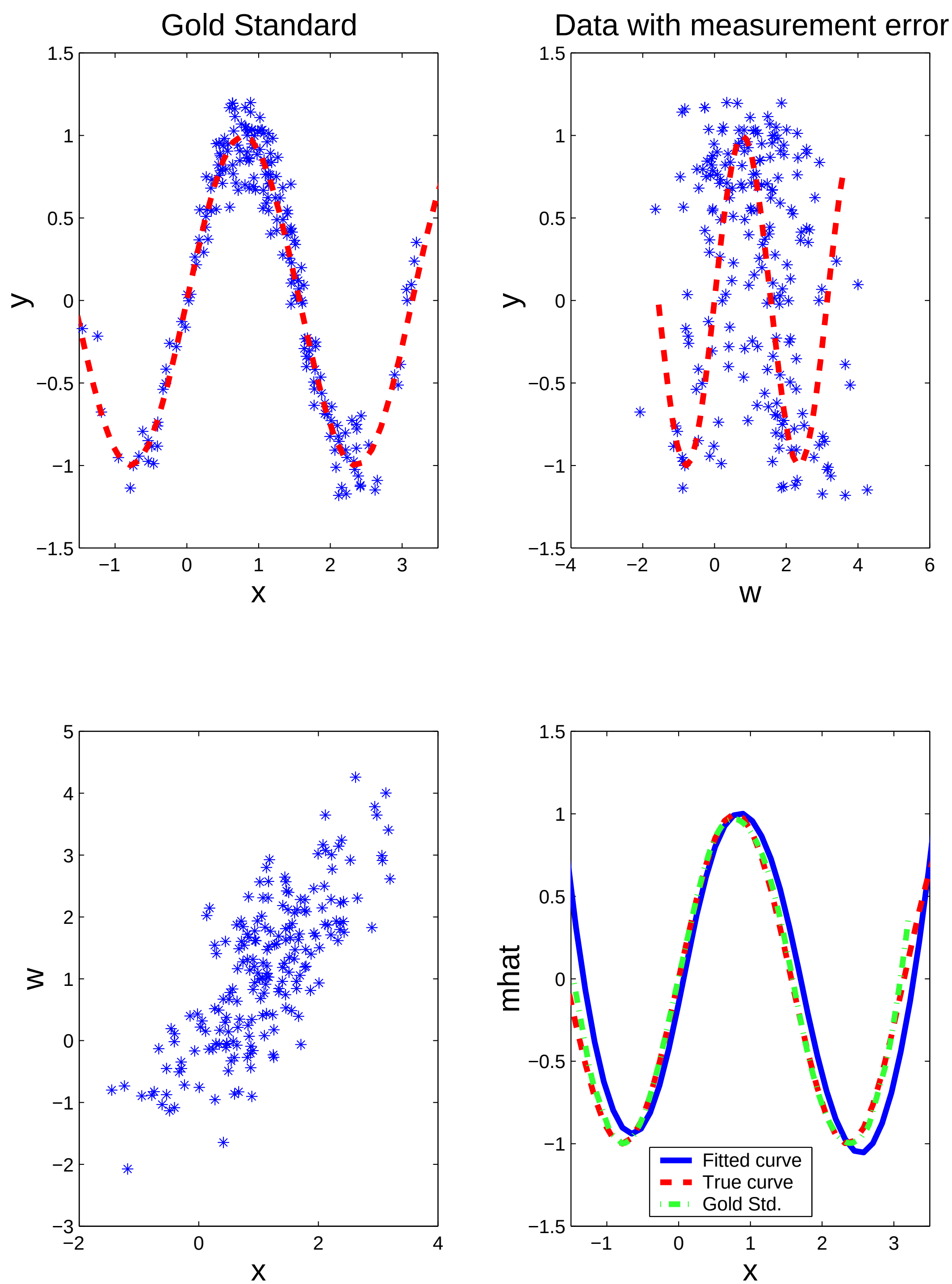


Figure 1: Simulated Data.

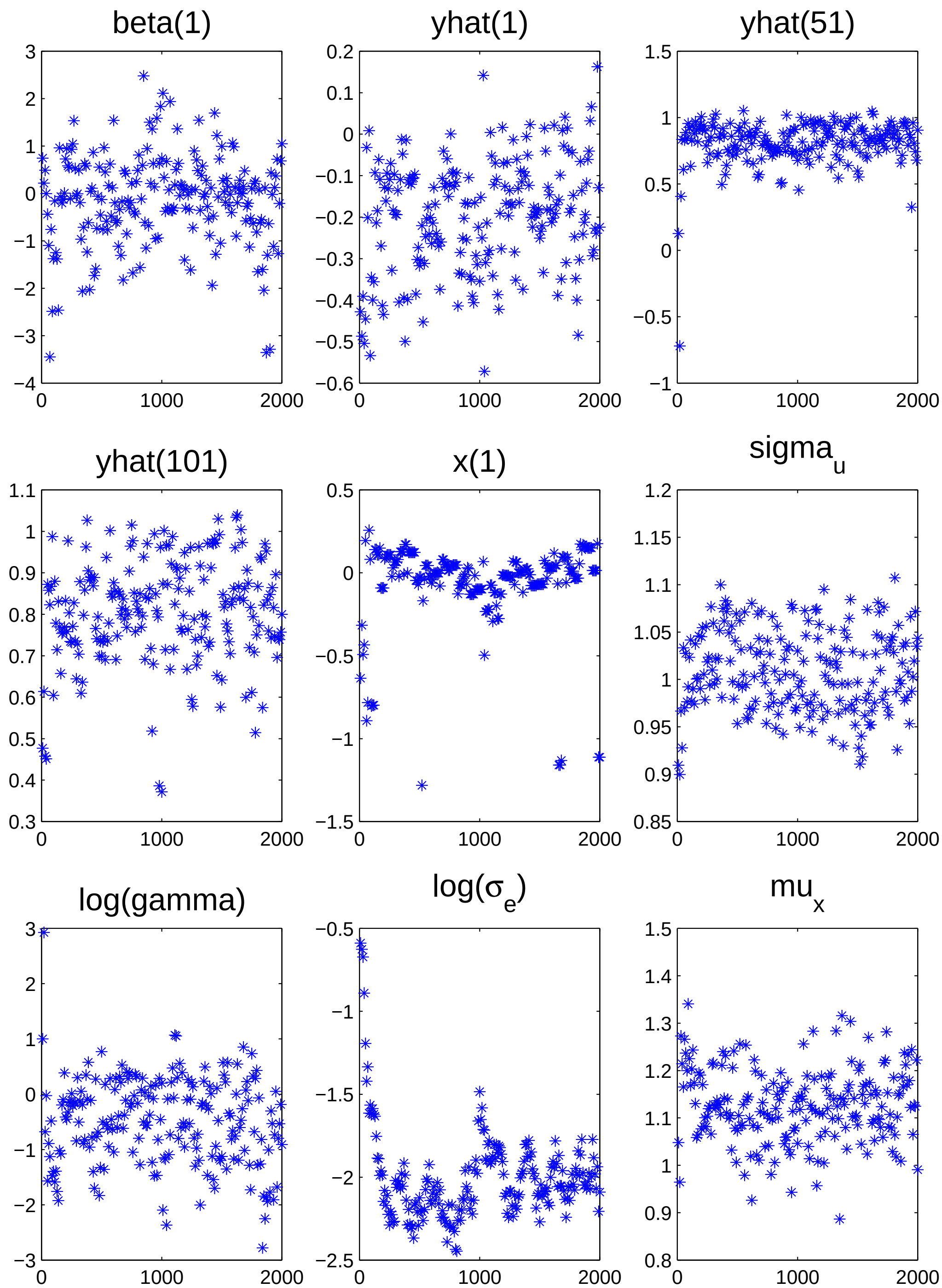


Figure 2: Simulated Data. Results of Gibbs Sampling. Every twentieth iteration plotted. **Note:** $X(1) = -1.5$ and $\overline{W}(1) = -0.8$. Also, $\log(\sigma_e) = -1.9$.

EXAMPLE — SIMULATED

What does the Bayes approach work so well? Here's my explanation:

Bayes uses all possible information to estimate X and, especially, $m(X)$.

- $\left\| m(X) - E\{m(X) | \mathbf{W}, \boldsymbol{Y}, \text{other param.}\} \right\|$

  $\approx \left\| m(X) - \text{ave}\{m(\widehat{X})\} \right\| = 2.47$

- $\left\| m(X) - m(E\{X | \mathbf{W}, \boldsymbol{Y}, \text{other param.}\}) \right\|$

  $\approx \left\| m(X) - m(\text{ave}\{\widehat{X}\}) \right\| = 4.67$

- $\left\| m(X) - m(E(X | \overline{W})) \right\| = 10.25$

- $\left\| m(X) - m(\overline{W}) \right\| = 12.36$

EXAMPLE — CLINICAL TRIAL

- Study of a psychiatric medication.
- Treatment and control group.
- Evaluation at baseline (W) and at end of study (Y).
  - smaller values $\rightarrow$ more severe disease
  - scale is a combination of self-report and clinical interview so there is considerable measurement error
  - it is believe that $\sigma_u^2 \approx 0.35$.
- We are interested in $\Delta(X) := m(X) - X = E(Y - W|X)$.
- Preliminary Wilcoxon test found a highly significant treatment effect.
- *Question*: How does the treatment effect depend upon the baseline value?

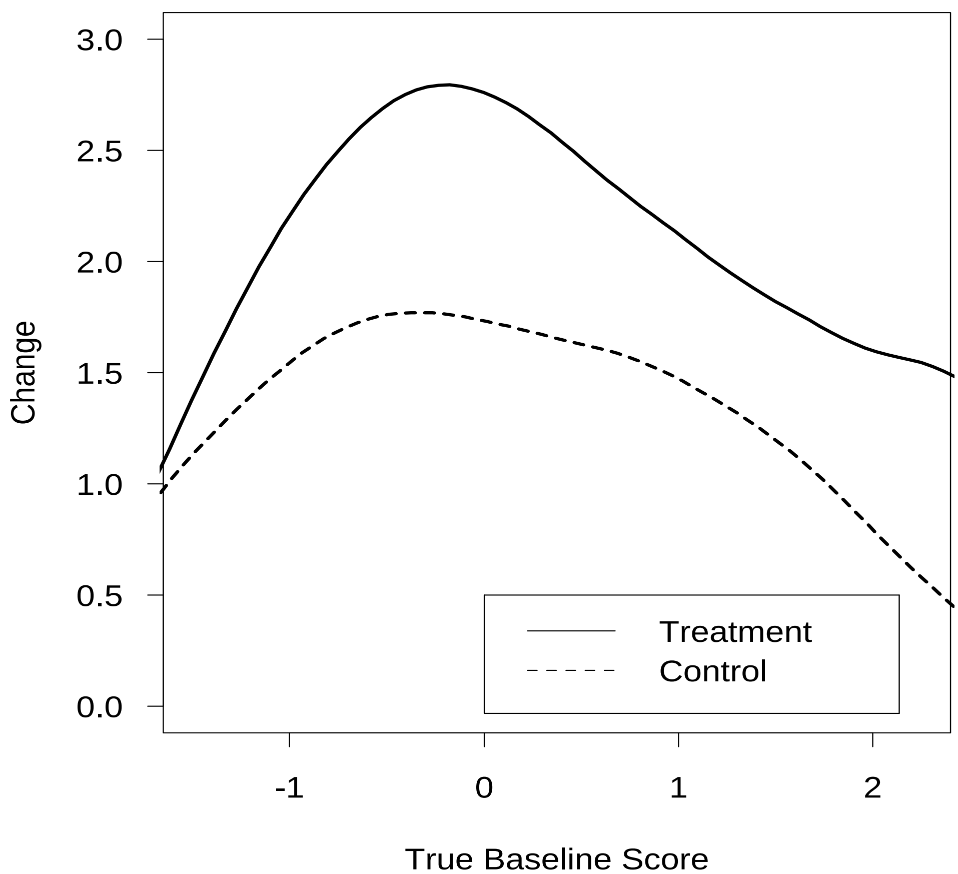


Figure 3: Estimate of the function $\Delta(x) = m(x) - x$ for the control group and the treatment group in the example. (**Note:** Positive change is an improvement.)

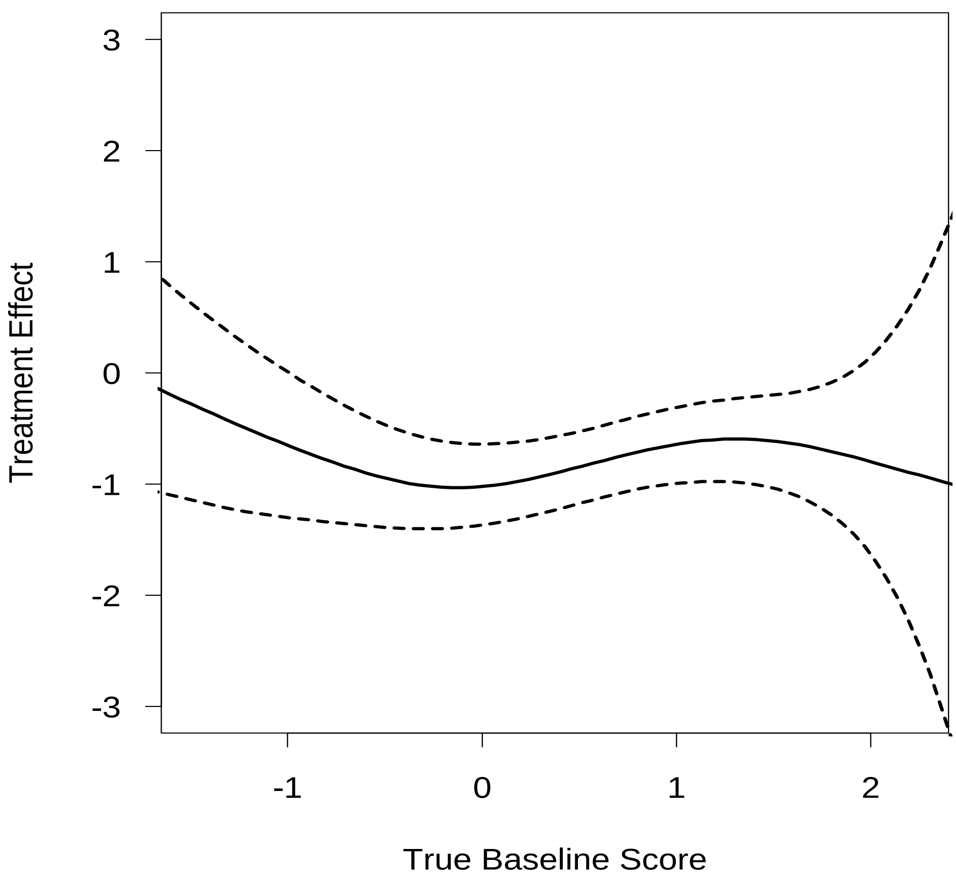


Figure 4: Estimate (solid) of the **difference** of the function $\Delta(x) = m(x) - x$ **between the treatment group and the control group** in the example (control-treatment) with 90% pointwise credible intervals (dashed).

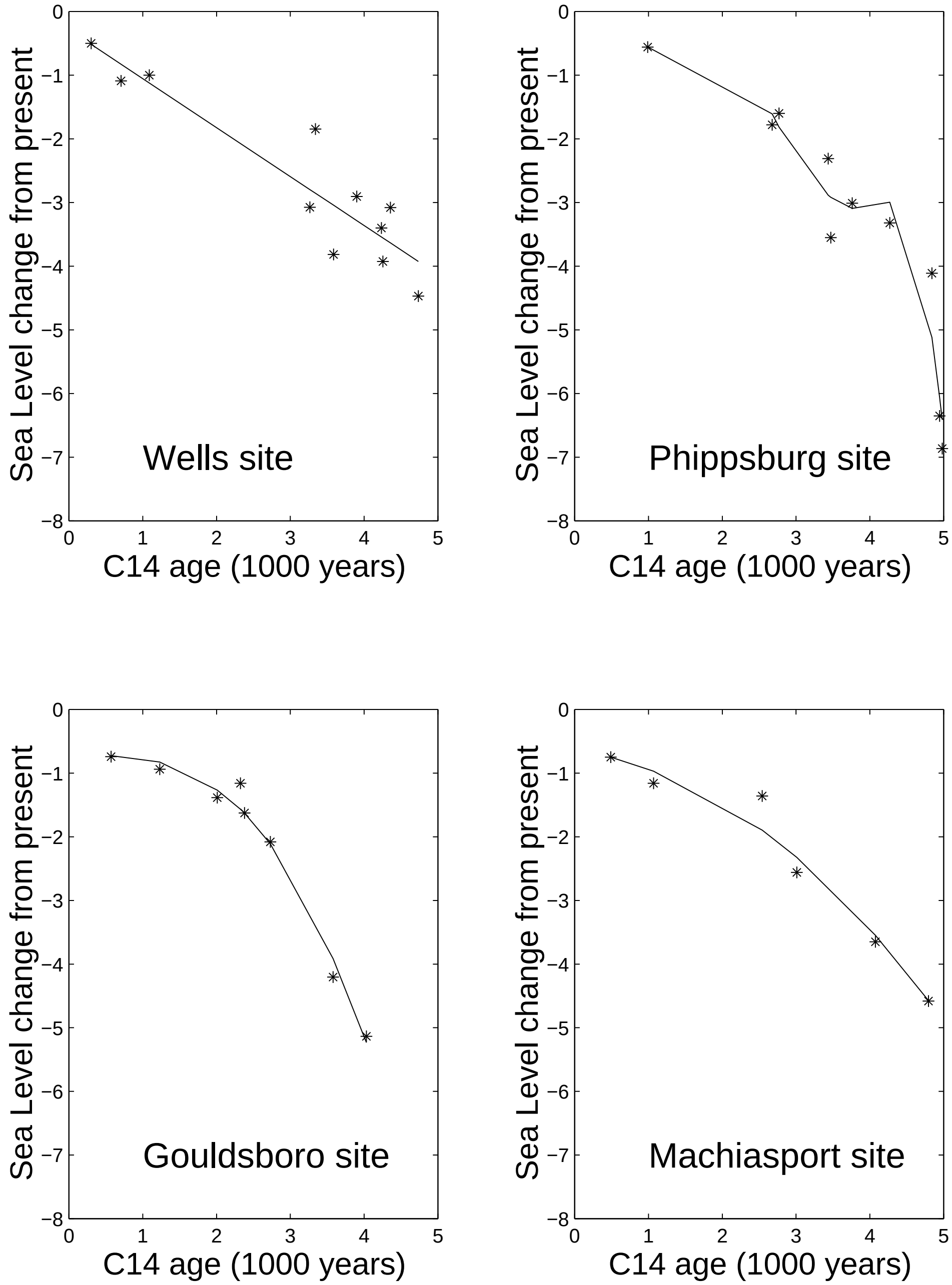
EXAMPLE - ^{14}C DATING AND SEALEVEL

- Core samples taken from four salt marshes in Maine
- $X =$ true age (or true C_{14} age?) of sample
- Only one W measured per core sample, but there is a standard deviation also report.
- $Y =$ sea level as determined by micro fossils in sample
  - expressed as a deviation from present
- Preliminary data analysis suggests:
  - Nonlinear $X - Y$ relationship
  - Site effects, which might be modeled as linear:

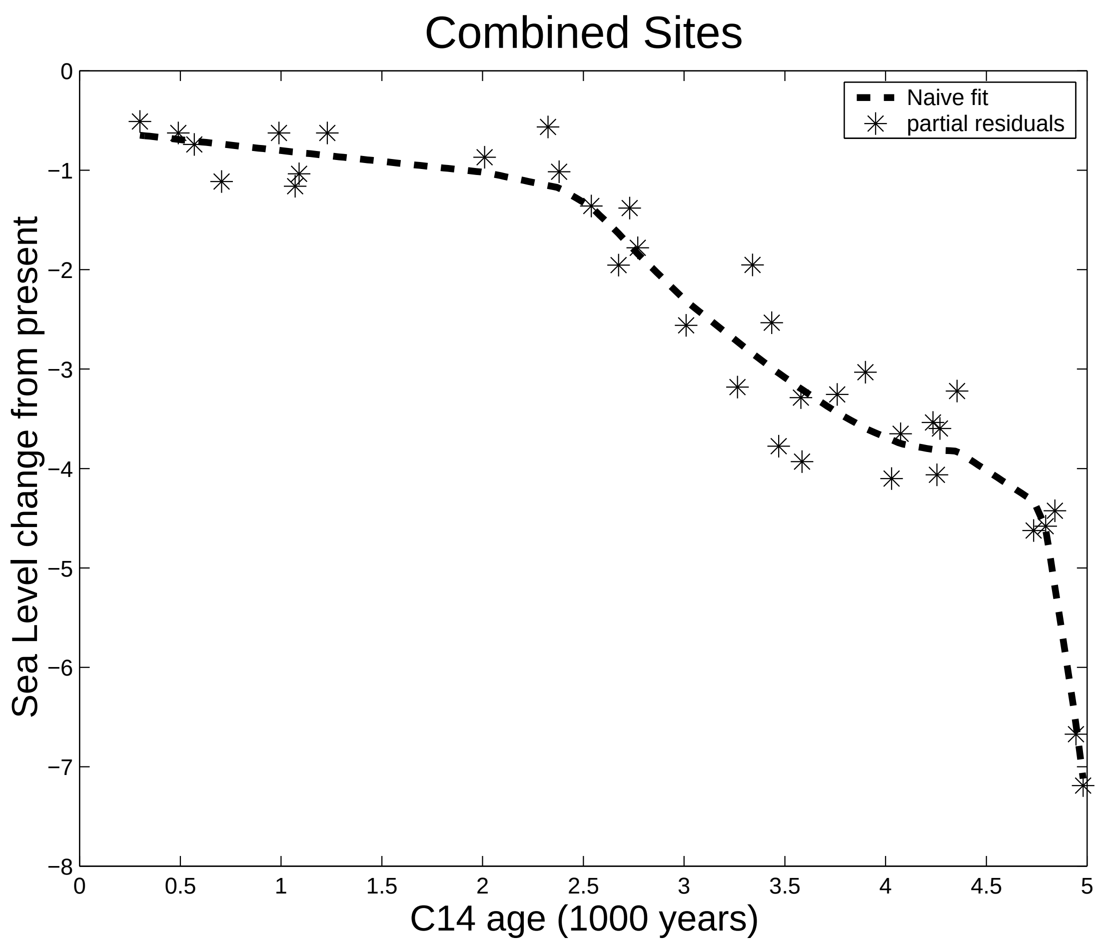
$$Y = m(X) + \sum_{j=1}^{3} \beta_j X * I(\text{site} = j) + \epsilon_i.$$

- Methodology can be applied w/o much change.

EXAMPLE - ^{14}C DATING AND SEALEVEL



EXAMPLE - ^{14}C DATING AND SEALEVEL



DISCUSSION

- With the work of **CMR** and **BCR** we now have reasonably efficient estimators for nonparametric regression with measurement error.
  - SIMEX (LPR and splines) — in **CMR**
  - (Flexible) Structural splines — in **CMR**
  - Fully Bayesian (hardcore structural) — in **BCR**
- With **BCR** we have a methodology that
  - automatically selects the amount of smoothing
  - estimates the unknown X's
  - allows inference that takes account of the effects of smoothing parameter selection and measurement error
- Most efficient methods appear to be structural, though SIMEX may be competitive
  - hardcore structural methods seem reasonably robust