Analysis of Computer Experiment Output Having Qualitative And Quantitative Input Variables

Gang Han, Dr. Thomas J. Santner, and Dr. William I. Notz

Department of Statistics, The Ohio State University

August 8, 2005
JSM 2005

With special thanks to Don Bartel and Tim Wright of the Joint Cornell-Hospital for Special Surgery Biomechanics Program

Outline

1. Goal and Motivating Example
2. A Bayesian Model
3. A Bayesian Method and Its Implementation
4. Prediction
5. An Example
6. Conclusion and Future Work

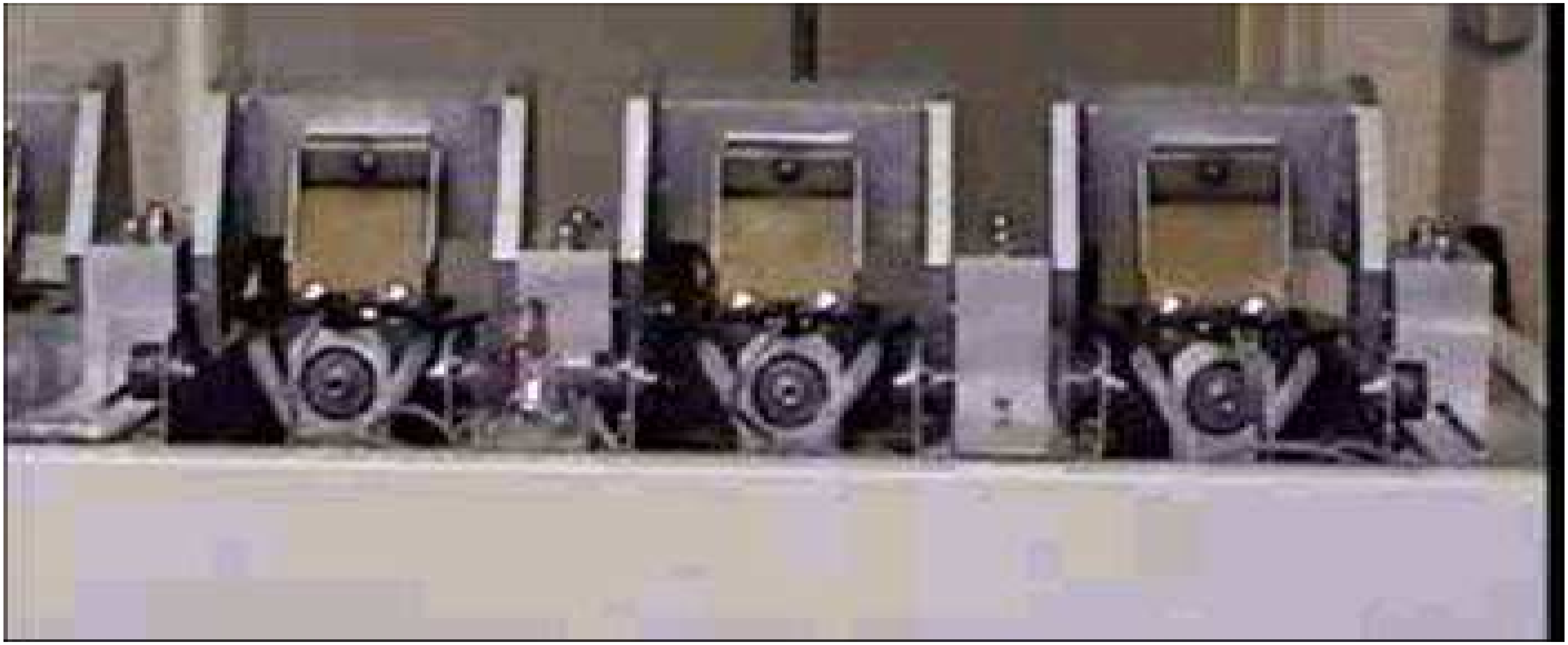
Goal And Motivating Example

- **Goal**
  1. Model computer experiment output when the inputs are both quantitative and qualitative
  2. Predict the unknown response

- **Motivating example** *FEA computer model* of a 4-station *knee wear simulator*. There are **mixed inputs**.
  Some *qualitative input variables* are
  - *Mesh Density (FEA)*$\{$1,1.25,2$\}$
  - *Initial AP position (Simulator)*$\{$0,2,4,6$\}$
  - *Flexion Lag (FEA)* $\{$Yes,No$\}$

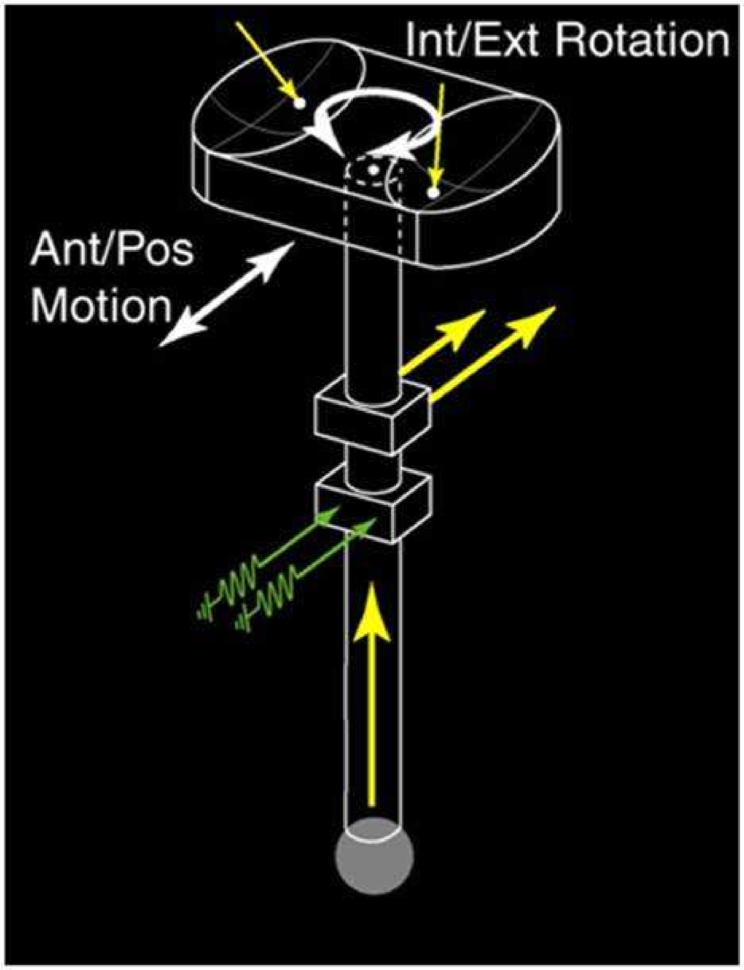
  Some *quantitative input variables* are
  - Spring length (FEA&Simulator)$\{$range:[24,48]$\}$
  - Spring stiffness (FEA&Simulator)$\{$range:[14.5,43.5]$\}$
  - Frequency (FEA)$\{$range:[0.3,1.4]$\}$

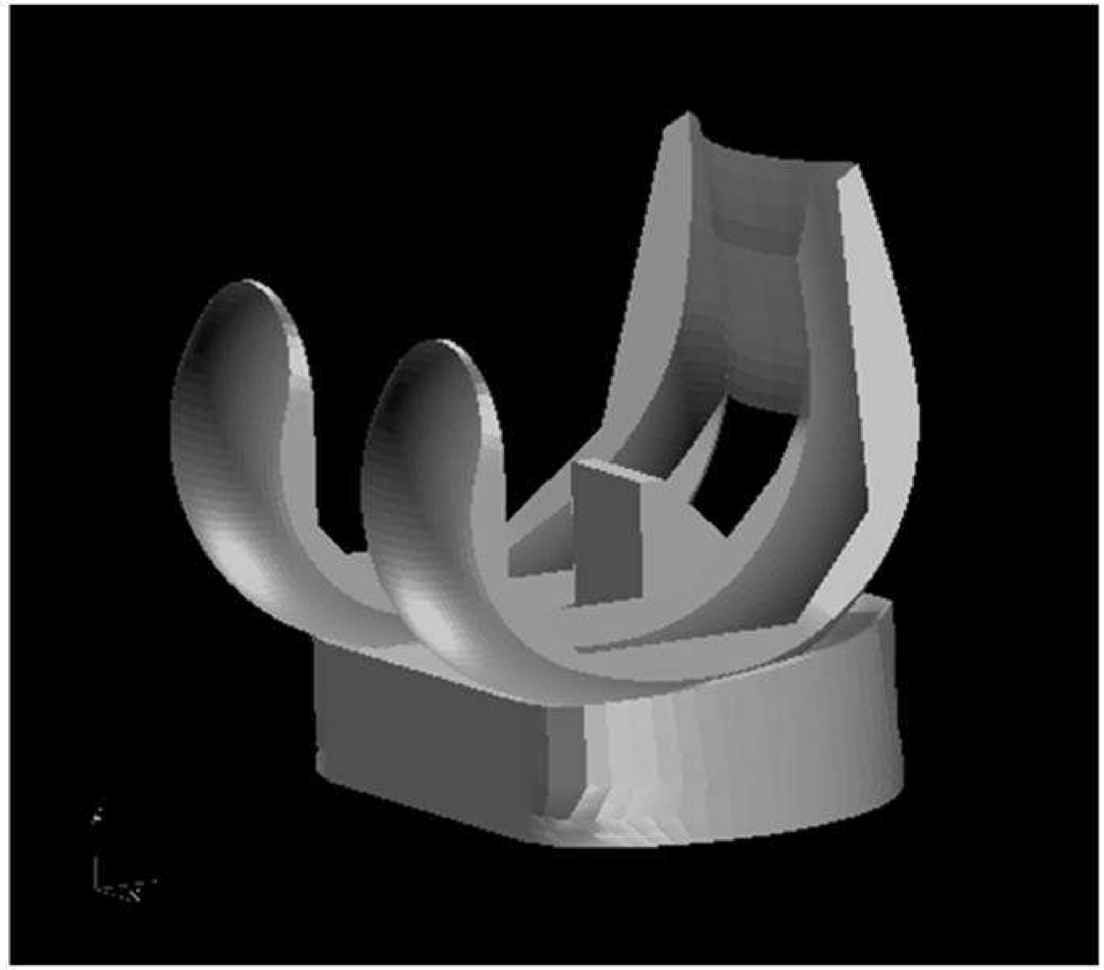
Knee Wear Simulator



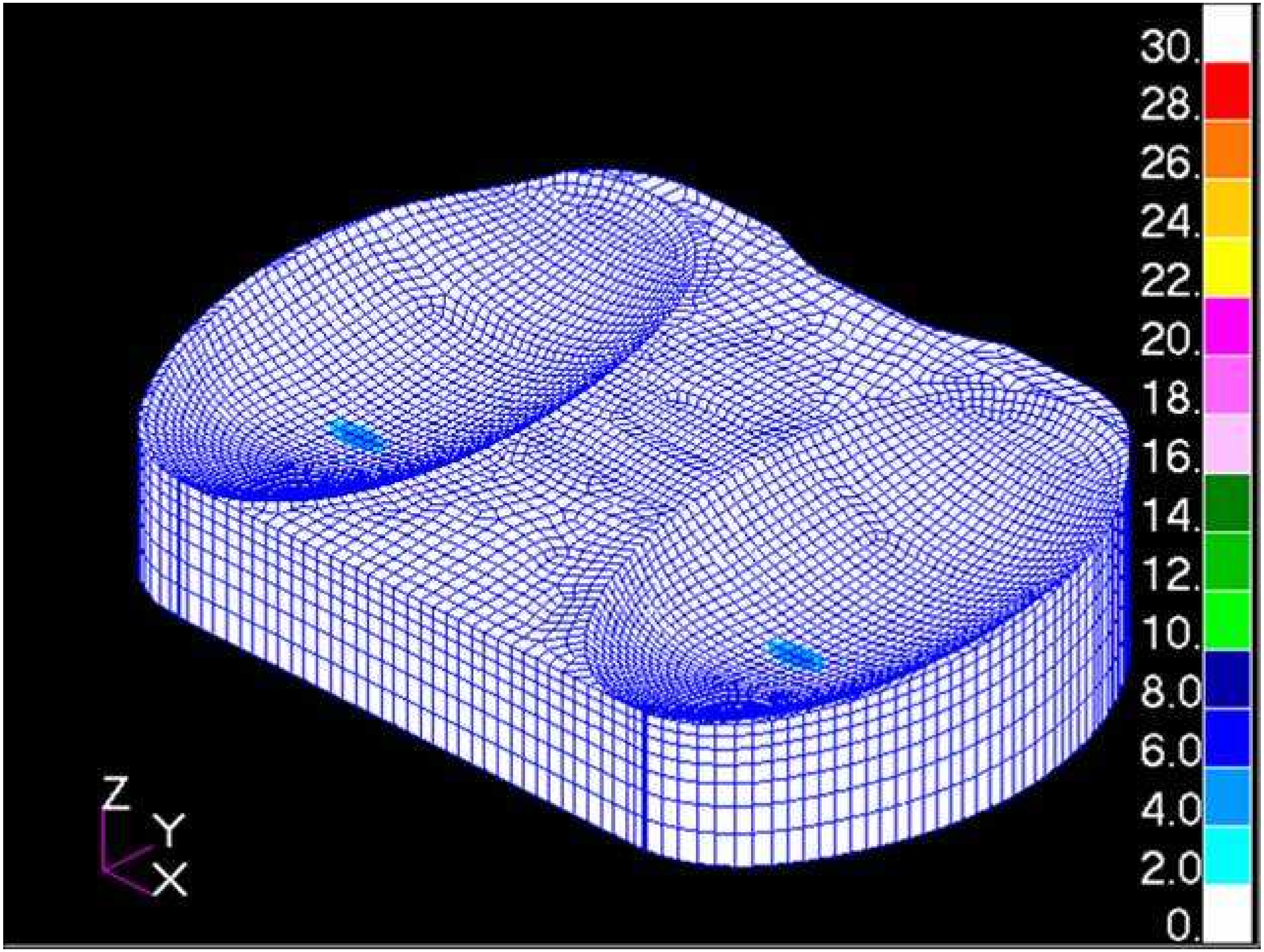
A 4 station device that mimics knee motion

Knee Wear Simulator





Mesh density of the FEA Simulator



A Bayesian Model

Suppose the inputs are $(i, \boldsymbol{x})$, where $i \in \{1, 2, ..., t\}$ is qualitative and $\boldsymbol{x} \in [0, 1]^d$ is quantitative. The output is the real-valued function $y(i, \boldsymbol{x})$.

1. **An Independence Model** Model $y(i, \boldsymbol{x})_x$ by t mutually independent stationary Gaussian Stochastic Processes. Fit a Bayesian model for each category separately.

2. **A Qualitative Quantitative Variable (QQV) Model**

   - Given $\{\boldsymbol{\rho}_1, ..., \boldsymbol{\rho}_t\}$, $\{\boldsymbol{\sigma}_1^2, ..., \boldsymbol{\sigma}_t^2\}$, and $\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_t\}_i$,

     $$Y(i, \boldsymbol{x}|\boldsymbol{\rho}_i, \beta_i, \sigma_i^2) = \beta_i + Z_i(\boldsymbol{x}).$$

     $Z_1(\boldsymbol{x}), ..., Z_t(\boldsymbol{x})$ are *independent* stationary Gaussian Processes with mean 0 and
     $cov(Z_i(\boldsymbol{x}_1), Z_i(\boldsymbol{x}_2)|\boldsymbol{\rho}_i, \sigma_i^2) = \sigma_i^2 R(\boldsymbol{x}_1 - \boldsymbol{x}_2|\boldsymbol{\rho}_i)$, where
     $R(\boldsymbol{h}|\boldsymbol{\rho}_i) = \prod_{j=1}^{d} \rho_{ij}^{h_j^2}, j = \{1, ..., d\}$.
     Here $\boldsymbol{\rho}_i = (\rho_{i1}, ..., \rho_{id})^T$, and $\boldsymbol{h} = (h_1, ..., h_d)^T$.

- Prior distribution
  - $\boldsymbol{\rho}_1, ..., \boldsymbol{\rho}_t$ are *independently* and *identically* distributed. For any $j \in \{1, ..., d\}$, $\rho_{1j}, ..., \rho_{tj}$ are *independently and identically* distributed as $Beta(\alpha_j, \gamma_j)$, where α_j and γ_j are known.
  - $\sigma_1^2, \sigma_2^2, ..., \sigma_t^2$ are *independently and identically* distributed. $\frac{1}{\sigma_i^2}$ is of $Gamma(\alpha, \gamma)$, where α and γ are known.
  - $\beta_1, ..., \beta_t$ are *independently and identically* distributed, $[\beta]$ is proportional to 1.

Notice that $y(1, \boldsymbol{x}), ..., y(t, \boldsymbol{x})$ in the QQV model are conditionally independent, but unconditionally dependent on each other.

A Bayesian Method and Its Implementation

Idea We use the Monte Carlo Markov Chain method, the Metropolis Hastings Sampler, to draw the values of each parameter from the joint posterior distribution and to predict $y(i_0, \boldsymbol{x}_0)$, the unknown output corresponding to the inputs $(i_0, \boldsymbol{x}_0)$.

For a single generic parameter δ

a. Initialize δ^1

b. Given δ^n, generate $\delta^\star$ from a symmetric distribution $f(. \mid \delta^n)$, i.e. a distribution satisfying $f(\delta^n \mid \delta^\star) = f(\delta^\star \mid \delta^n)$

c. Generate

$$\delta_{new} = \begin{cases} \delta^\star, & \text{with probability } \alpha \\ \delta^n, & \text{with probability } 1-\alpha \end{cases}$$

$$\alpha = min\{1, \frac{\pi(\delta^\star|\widetilde{Y})}{\pi(\delta^n|\widetilde{Y})}\}.$$

Prediction

Goal: Predict $Y_0 = Y(i_0, \boldsymbol{x}_0)$ given output

$$\boldsymbol{Y}^n = (Y(i_1, \boldsymbol{x}_1), Y(i_2, \boldsymbol{x}_2), ..., Y(i_n, \boldsymbol{x}_n))$$

Basis of prediction

$$[\boldsymbol{Y}^n | \boldsymbol{\rho}, \boldsymbol{\beta}, \Sigma_{yn}] \sim N\left(\boldsymbol{F}\boldsymbol{\beta}, \Sigma_{yn}\right)$$

and

$$\left[\begin{pmatrix} Y_0 \\ \boldsymbol{Y}^n \end{pmatrix} | \boldsymbol{\rho}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2\right] \sim N\left(\begin{pmatrix} \boldsymbol{f}_0^T \\ \boldsymbol{F} \end{pmatrix} \boldsymbol{\beta}, \begin{pmatrix} \Sigma_{y0} & \Sigma_{0n}^T \\ \Sigma_{0n} & \Sigma_{yn} \end{pmatrix}\right)$$

- The predictor

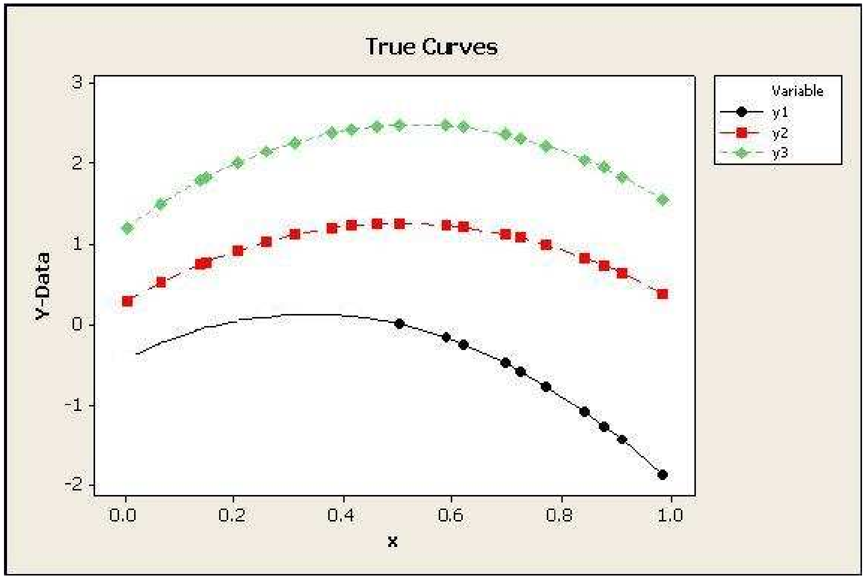
$$\rightarrow \tilde{Y}(i_0, \boldsymbol{x}_0) = E\{Y(i_0, \boldsymbol{x}_0|\boldsymbol{Y}^n)\} = E\{E(Y(i_0, \boldsymbol{x}_0)|\boldsymbol{Y}^n, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2)\}$$

- The variance of the predictor

$$\rightarrow var(Y_0|\boldsymbol{Y}^n) =$$
$$var\{E(Y_0|\boldsymbol{Y}^n, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2)\} + E\{var(Y_0|\boldsymbol{Y}^n, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2)\}$$

An Example

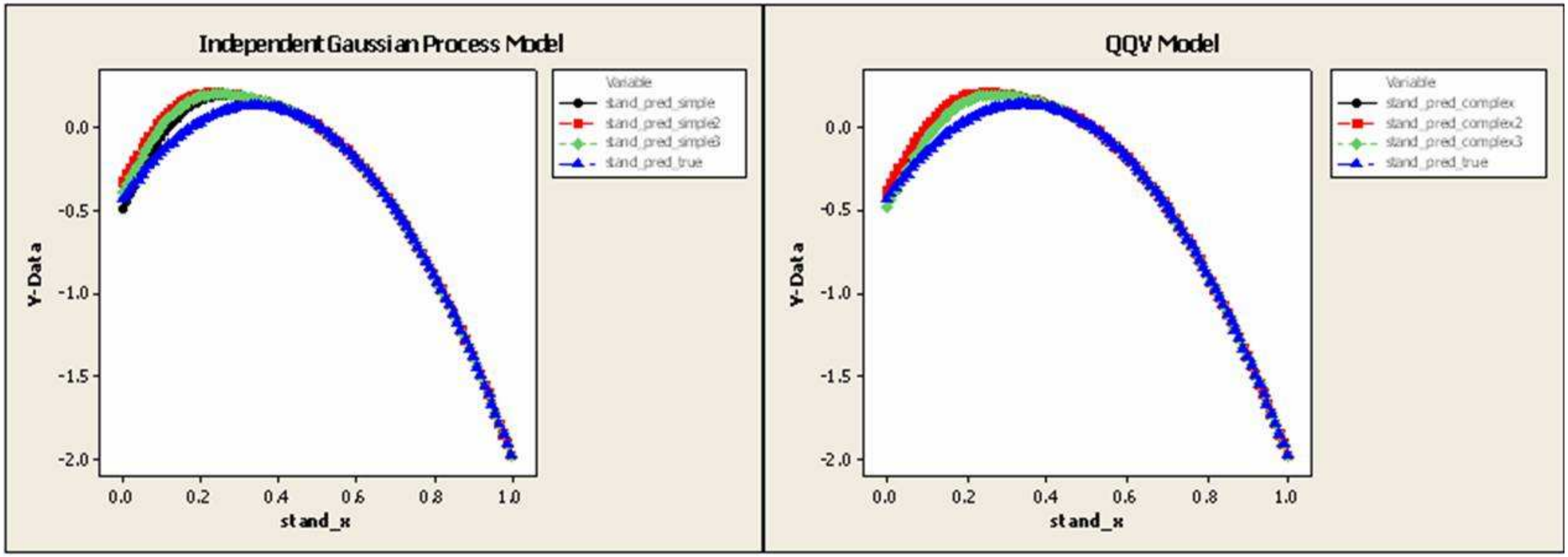
We look at a simple example. Suppose we have one qualitative and one quantitative input. The qualitative input identifies one of three quadratic curves, each a function of the quantitative input(x).



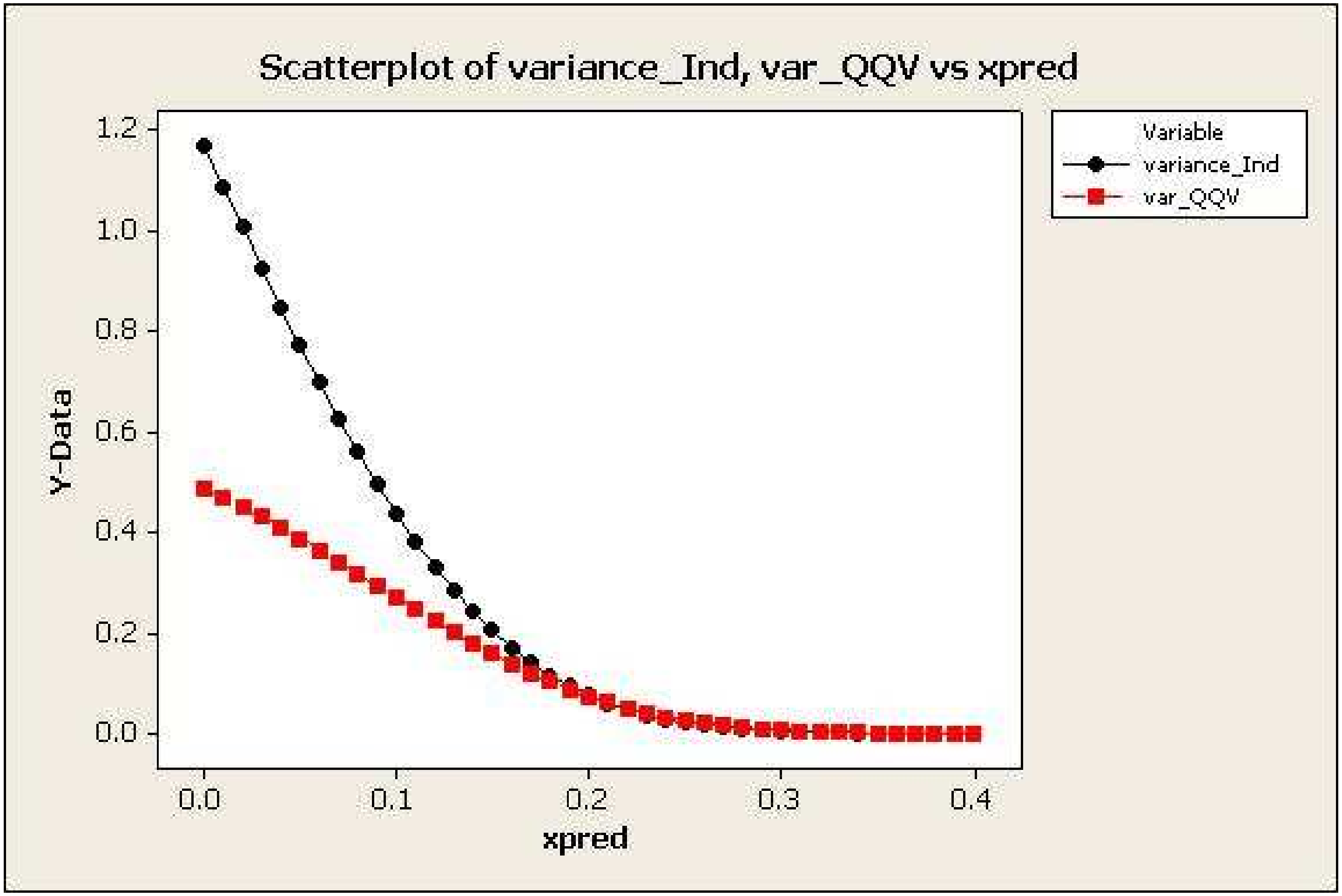
We observe the first two curves at 21 equally spaced points from 0 to 20, and the third curve only on the points 10 to 20.

Goal: predict $y(3, x)$ over $x \in [0, 1]$.

For point prediction, the results were comparable when we fit three independent curves versus our QQV model.



But the QQV model has smaller prediction variance.



Conclusion and Future work

Conclusion

- The point predictions using either model are close.
- The QQV model has smaller model uncertainty than the Independent model. This means a curve in the QQV model can effectively borrow information from other curves.

Future work

- Develop software for arbitrary numbers of quantitative and qualitative variables
- Additional Empirical Studies
- Optimization for outputs having quantiitative and qualitative inputs

The slides of this talk can be found at
http://www.stat.ohio-state.edu/$\sim comp\_exp/2005JSM/slide\_JSM2005.pdf$.