SAND2023-05790C

Sandia National Laboratories

# Inverse prediction using functional data in a Bayesian framework

Audrey McCombs

K. Goode, K. Shuler, J. D. Tucker, A. Zhang, D. Ries

15th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2022)

18 December 2022

King's College London, UK

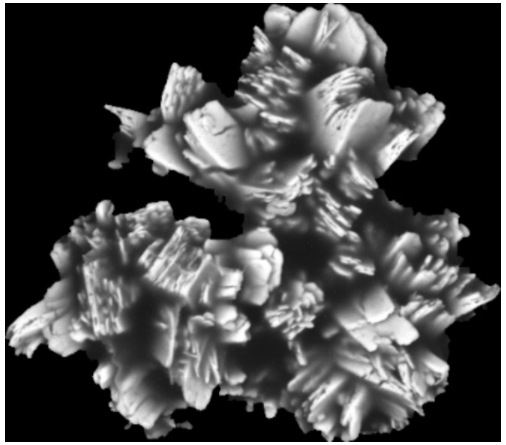


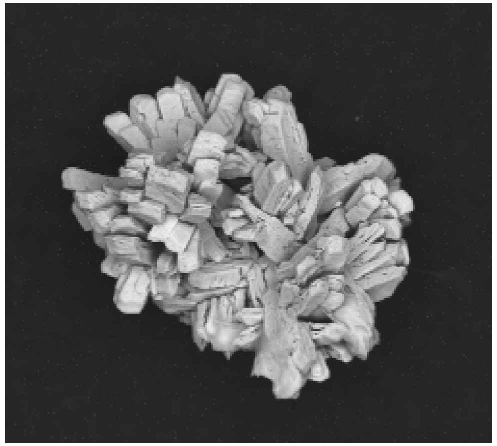


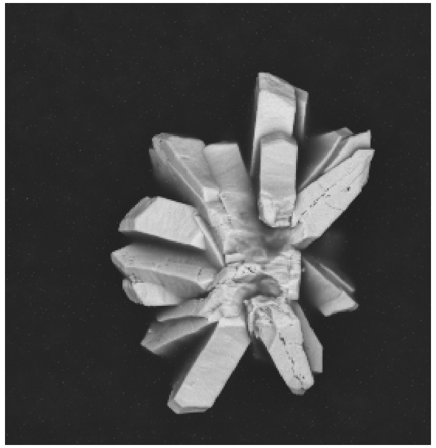
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

# Modeling Objectives

## Can we predict the processing conditions for this material?





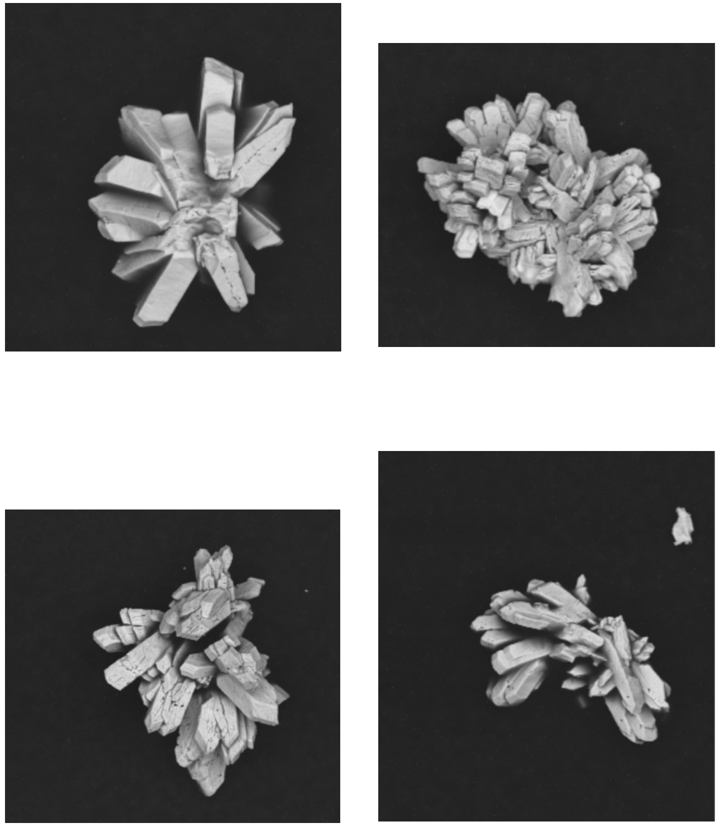


Scanning Electron Microscope (SEM) images of Pu particles

# Modeling Objectives

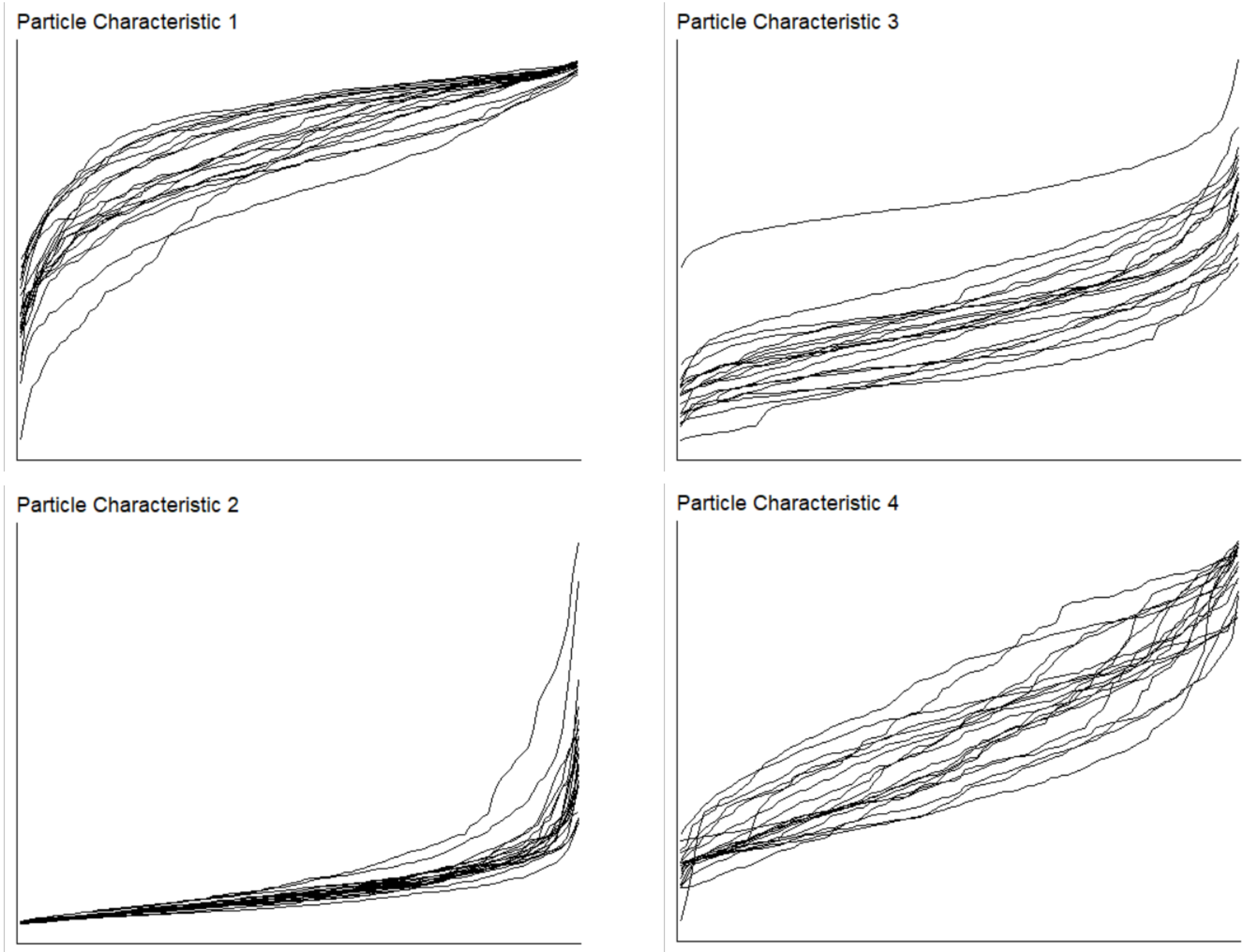
- 4 particle characteristics (e.g., color, texture, size)
- 3 process conditions (e.g., temperature, chemical characteristic, processing time)
- Given the measured particle characteristics, can we predict the exact conditions used to produce the material?
- Project framework:
  - Process Pu under known conditions
  - Measure resulting particle characteristics
  - Fit forward model
  - Inverse-predict test values, quantify uncertainty, and assess predictive accuracy
- Model framework combines:
  - Functional data analysis (FDA)
  - Inverse prediction
  - Seemingly unrelated regression (SUR)
  - Bayesian modeling

# Functional Data



Avg. 640 measurements of each particle characteristic per experimental run

## Empirical CDFs of four particle characteristics



# Functional Inverse Prediction (FIP) framework

1. Represent functional responses using basis functions

fPCA on empirical CDF

2. Fit forward model

Process Condition = *f*(Particle Characteristics) + ε
Determine Particle Characteristics to use in Step 3
Stepwise regression & LASSO

3. Fit inverse model

a) Particle Characteristics = *f*(Process Conditions) + ε
Seemingly Unrelated Regression
b) Predict Processing Conditions
P(Processing Conditions | Particle Characteristics)

4. Validate model

Leave-one-out cross-validation

# Simulated Functional Data

$t \in [-4, 4], n = 150$

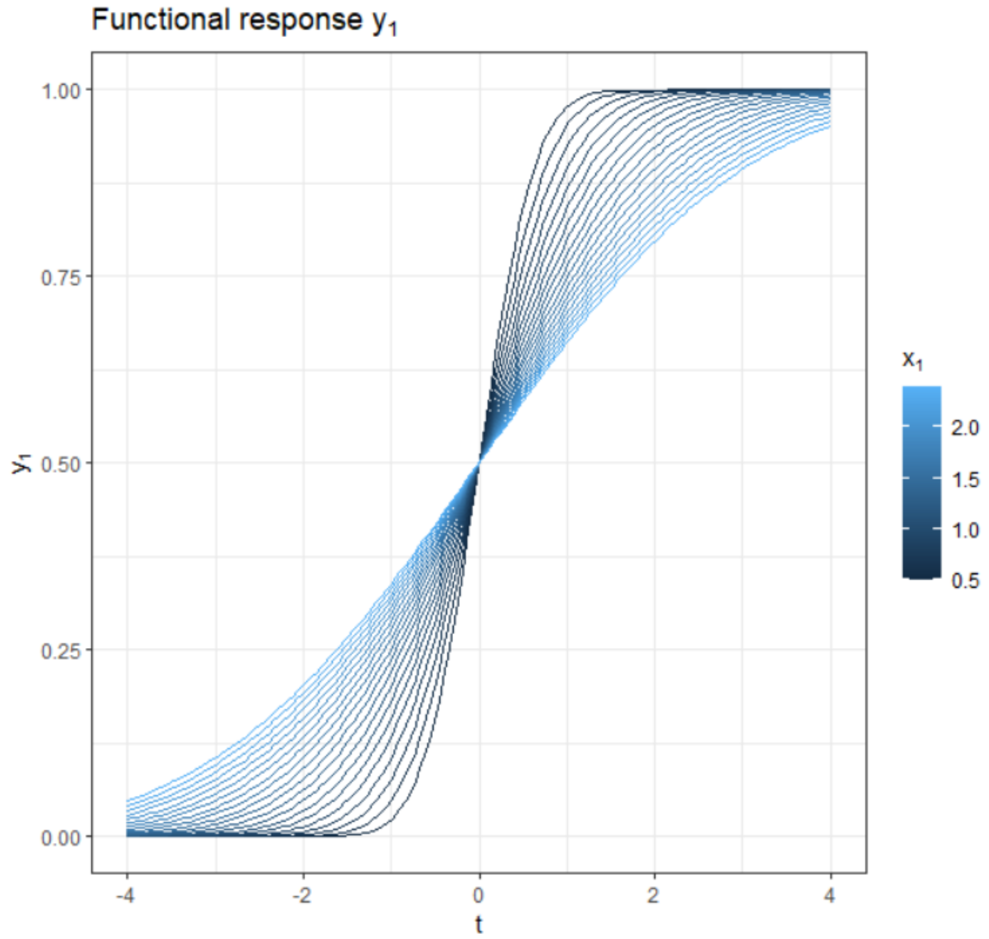
$x_1 \in [0.5, 2.4], n = 20$

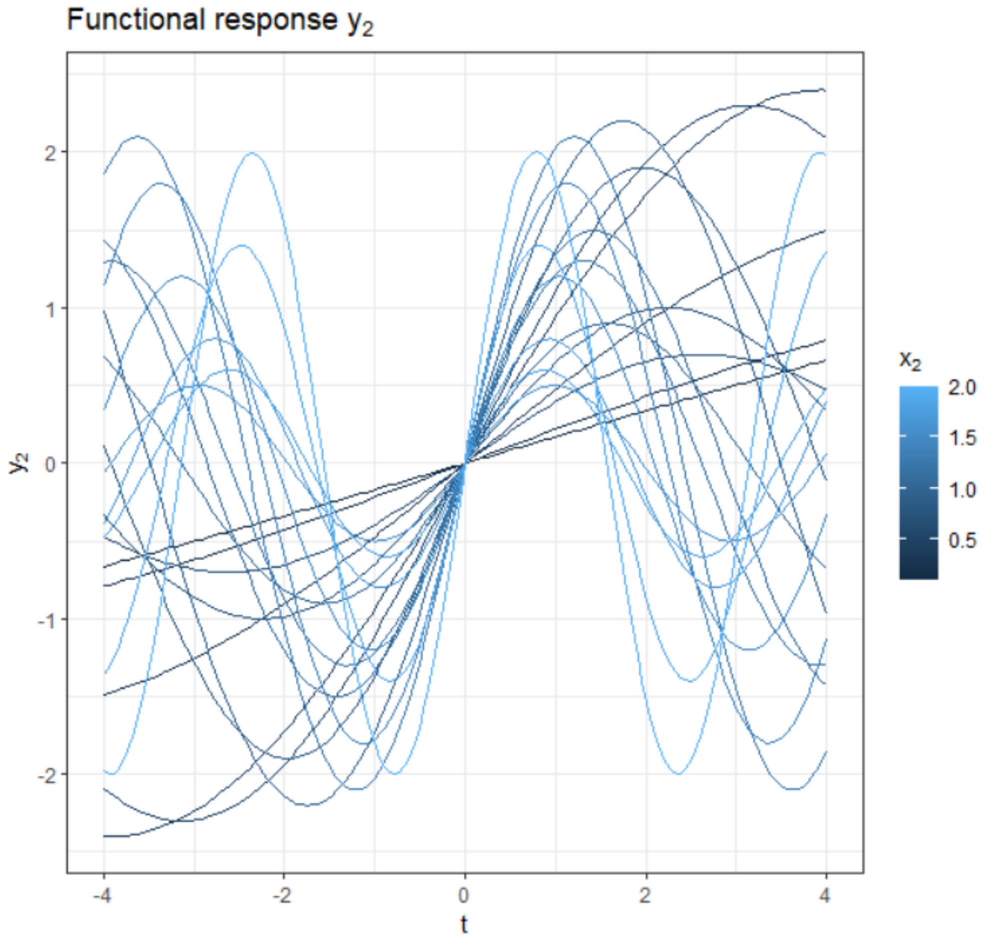
$x_2 \in [0.1, 2], n = 20$

$y_1 = \Phi_t(0, x_1)$

$y_2 = x_1 \sin(x_2 t)$

Use first two principal components (PCs) as response variables (4 total)



# Functional Inverse Prediction (FIP)

<u>Model:</u>

$$y_{q,i}^{k} = \beta_{0qk} + \beta_{1qk}x_{1,i} + \beta_{2qk}x_{2,i} + \epsilon_{q,i}^{k}$$

Principal Component $k \in \{1, 2\}$ from response variable $y_q, q \in \{1, 2\}$ and observation $i \in \{1, \ldots, 20\}$

<u>Bayesian Implementation:</u>

For inverse-predicting $x_{1,1}$ and $x_{2,1}$

$$\begin{bmatrix} \mu_{q,1}^{k} \\ \mu_{q,2}^{k} \\ \vdots \\ \mu_{q,n}^{k} \end{bmatrix} = \beta_{0qk} + \beta_{1qk}\begin{bmatrix} NA \\ x_{1,2} \\ \vdots \\ x_{1,n} \end{bmatrix} + \beta_{2qk}\begin{bmatrix} NA \\ x_{2,2} \\ \vdots \\ x_{2,n} \end{bmatrix}$$

$$y_{q,i}^{k} \sim N\left(\mu_{q,i}^{k}, \tau_{qk}\right)$$

Standard priors on $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$

Data: $\boldsymbol{X}, \boldsymbol{Y}$ matrices

Estimated: $\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\tau}$ matrices

# Simulated Data Results: FIP

$$y^k_{q,i} = g(\boldsymbol{X_q \beta_q}) + \epsilon^k_{q,i}$$

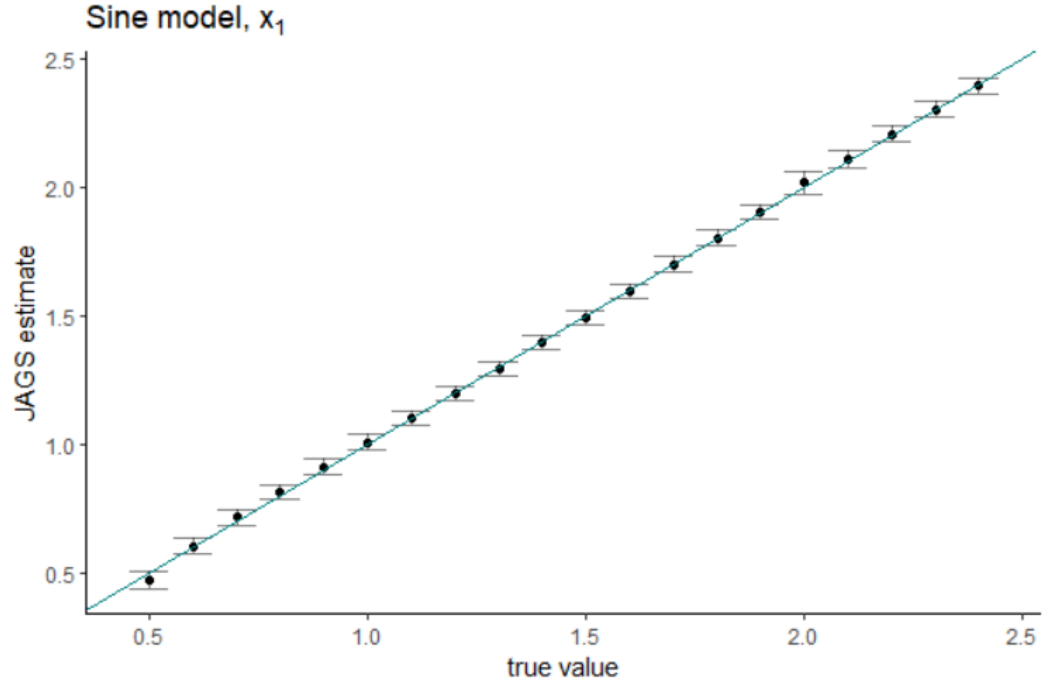
Linear model: $\beta_0 + \beta_1 x_1 + \beta_2 x_2$

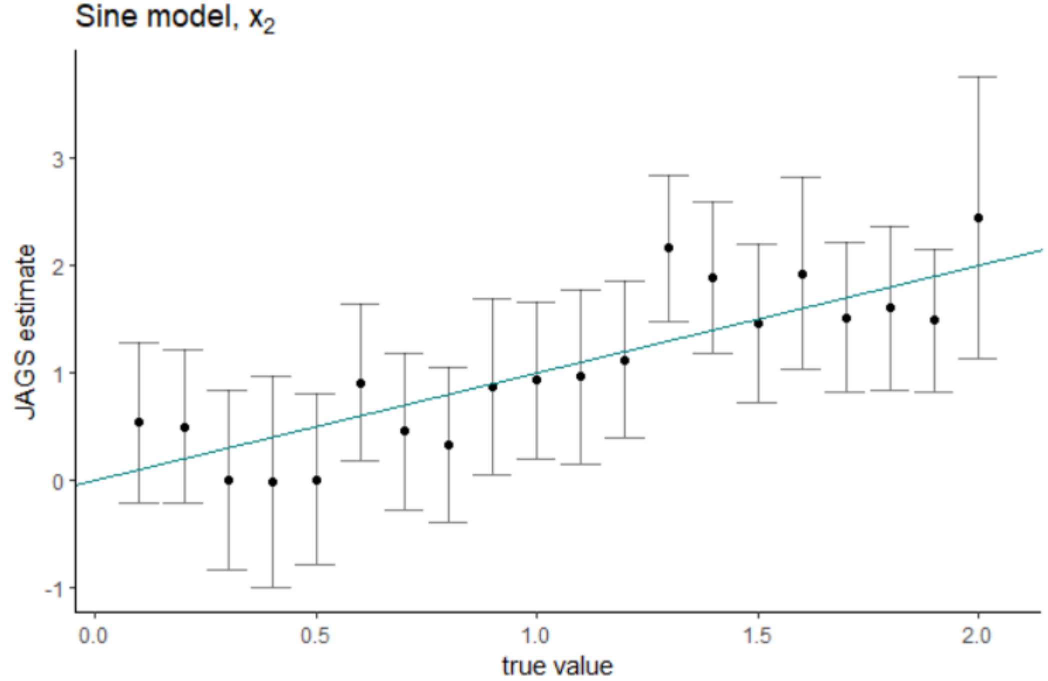
Interaction model: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

Quadratic model: $\beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2$

Sine model: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \sin(x_2)$

| Model | RMSE $x_1$ | RMSE $x_2$ |
|---|---|---|
| Linear | 0.044 | 0.479 |
| Interaction | 0.021 | 1.559 |
| Quadratic | 0.011 | 0.398 |
| Sine | 0.011 | 0.392 |



# Seemingly Unrelated Regression (SUR)

- Generalization of simple linear regression
- Looks like multiple regression
  - Response is $n \times Q$ matrix $\boldsymbol{Y}$ composed of $\boldsymbol{Q}$ response variables $\boldsymbol{y}_q, q = 1, \ldots Q$
- Each response variable $\boldsymbol{y}_q$ has its own regression equation
  - Possibly (usually) different predictors, different regression functions (e.g. linear, quadratic) associated with each regression equation
- Error terms across regression equations are allowed to be correlated
  - For response vectors $\boldsymbol{y}_q$and $\boldsymbol{y}_{\boldsymbol{r}}$
    - $cor(y_{qi}, y_{qj}) = 0$ Observations within a response are independent
    - $cor(y_{qi}, y_{ri}) \neq 0$ Observations across responses can be correlated

# Seemingly Unrelated Regression

$$\boldsymbol{y}_q = \boldsymbol{X}_q \boldsymbol{\beta}_q + \boldsymbol{\epsilon}_q$$

$\boldsymbol{y}_q$ is $n \times 1$

$\boldsymbol{X}_q$ is $n \times p_q$

$\boldsymbol{\beta}_q$ is $p_q \times 1$

$\boldsymbol{\epsilon}_q$ is $n \times 1$

$p_q$ is the number of predictors in regression equation $q$

## Stacked

$$\underset{nQ\times 1}{\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_Q \end{bmatrix}} = \underset{nQ\times \sum_Q p_q}{\begin{bmatrix} \boldsymbol{X}_1 & \cdots & 0 \\ 0 & \ddots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \boldsymbol{X}_Q \end{bmatrix}} \underset{\sum_Q p_q \times 1}{\begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_Q \end{bmatrix}} + \underset{nQ\times 1}{\begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_Q \end{bmatrix}}$$

## Covariance structure

$cov(\epsilon_{qi}, \epsilon_{qj}) = 0 \; for \; i \neq j$ observation

$cov(\epsilon_{qi}, \epsilon_{ri}) \neq 0 \; for \; q \neq r$ regression

$$\underset{Q\times Q}{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1Q} \\ \sigma_{21} & \sigma_{22} & & \sigma_{2Q} \\ \vdots & & \ddots & \vdots \\ \sigma_{Q1} & \sigma_{Q2} & \cdots & \sigma_{QQ} \end{bmatrix}, \sigma_{qr} = cov(\epsilon_{qi}, \epsilon_{ri}) \, \forall i$$

$$\Omega = \Sigma \otimes I_n$$

$nQ \times nQ$ block matrix whose blocks are diagonal matrices

# Simulated Data with Correlated Errors

## Covariance structure

$cov(\epsilon_{qi}, \epsilon_{qj}) = 0 \ for \ i \neq j$ observation

$cov(\epsilon_{qi}, \epsilon_{qj}) = 1 \ for \ i = j$ observation

$cov(\epsilon_{qi}, \epsilon_{ri}) = 0.9 \ for \ q \neq r$ regression

$$\Sigma = \begin{bmatrix} \sigma_{11} = 1 & \sigma_{12} = 0.9 \\ \sigma_{12} = 0.9 & \sigma_{22} = 1 \end{bmatrix}$$

$$\Omega = \Sigma \otimes I_{20}$$

## Uncorrelated simulated data

$t \in [-4, 4], n = 150$

$x_1 \in [0.5, 2.4], n = 20$

$x_2 \in [0.1, 2], n = 20$

$y_1 = \Phi_t(0, x_1)$

$y_2 = x_1 \sin(x_2 t)$

## Correlated simulated data

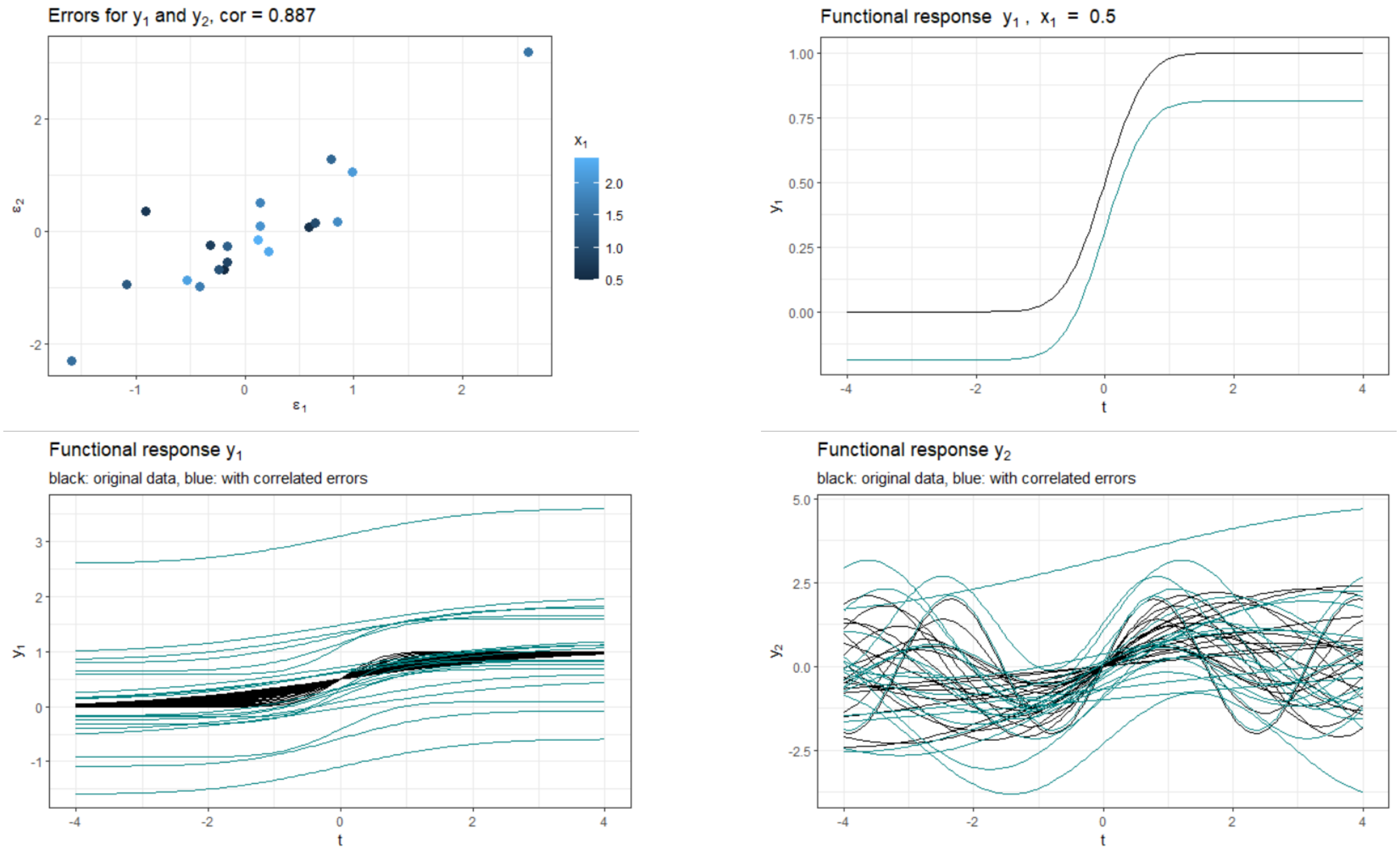
$\epsilon_q \sim \mathrm{MVN}(\mathbf{0}, \Omega)$

$\boldsymbol{y}_1^* = \boldsymbol{y}_1 + \boldsymbol{\epsilon}_1$

$\boldsymbol{y}_2^* = \boldsymbol{y}_2 + \boldsymbol{\epsilon}_2$

Use first two PCs as response variables (4 total)

# Simulated Data with Correlated Errors



# Correlated Simulated Data Results: FIP with SUR

Correlated errors

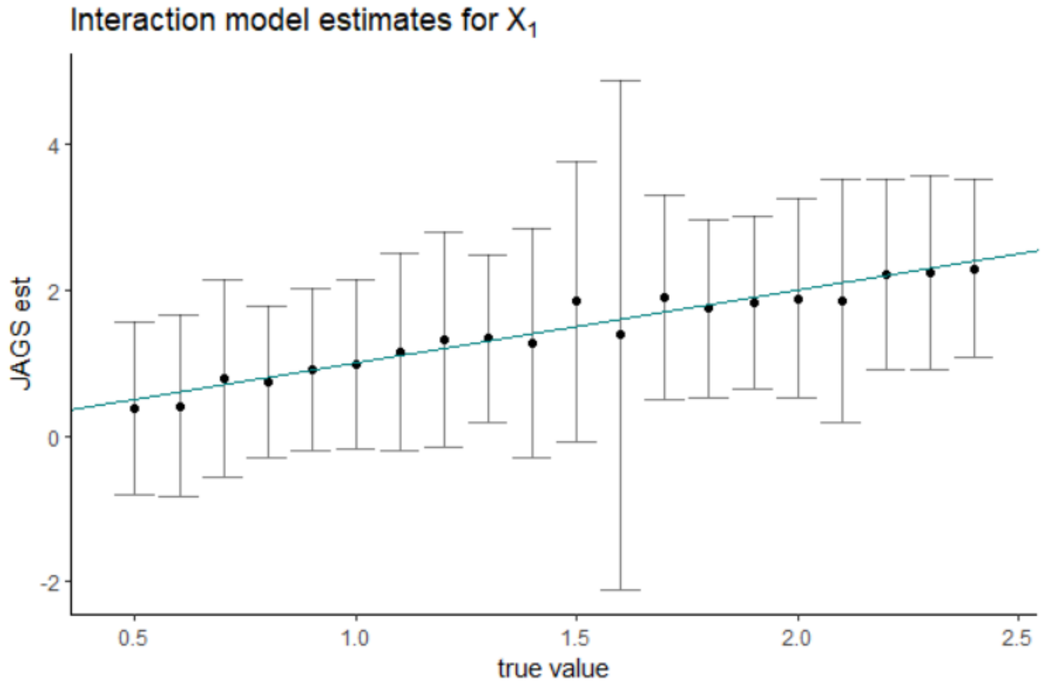
$$y^k_{q,i} = g(\boldsymbol{X_q \beta_q}) + \epsilon^k_{q,i}$$

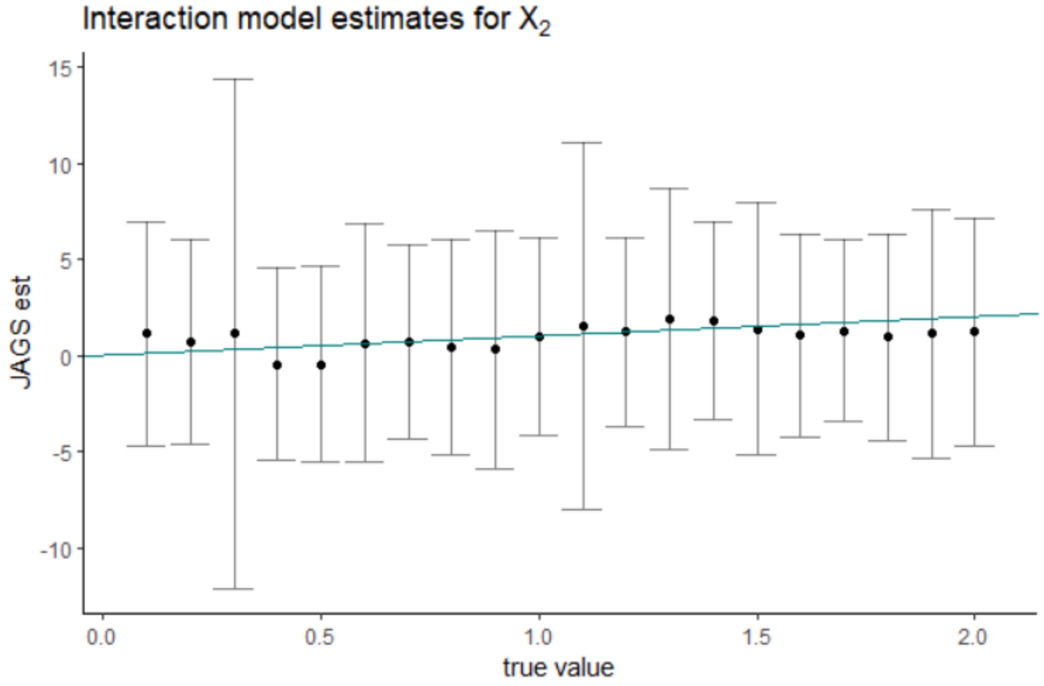
Interaction model: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

Quadratic model: $\beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2$

Sine model: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \sin(x_2)$

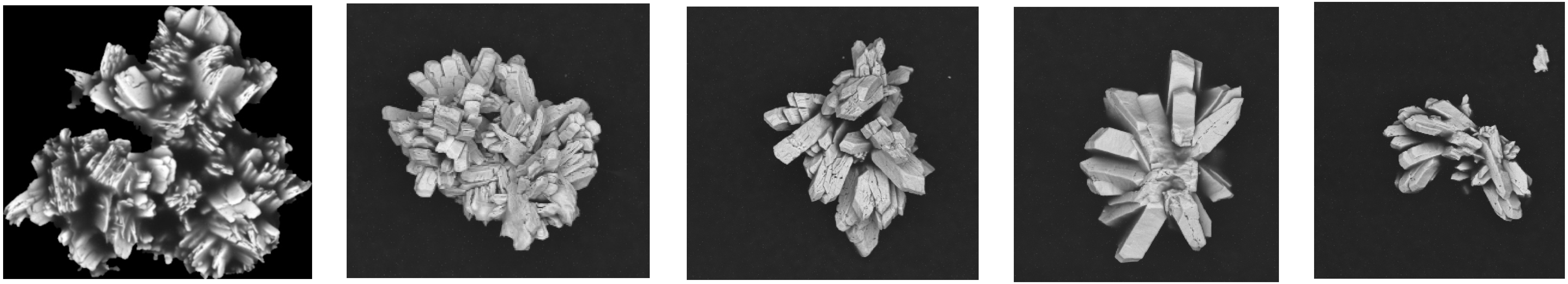
| Model | RMSE $x_1$ | RMSE $x_2$ |
|---|---|---|
| Interaction | 0.141 | 0.602 |
| Quadratic | 0.171 | 0.782 |
| Sine | 0.171 | 0.893 |



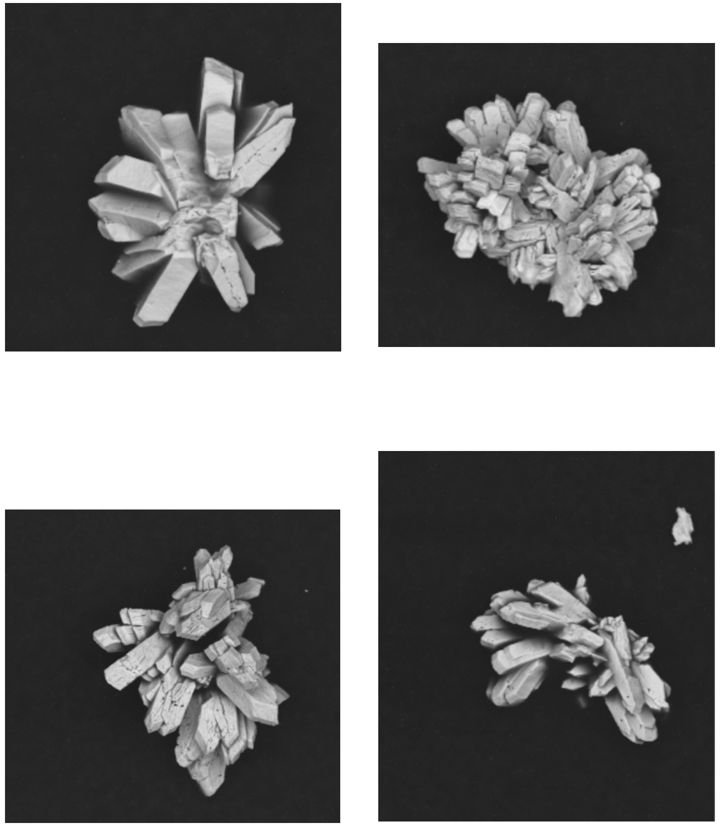
# Can we predict the processing conditions for this material?

- 4 particle characteristics (e.g., color, texture, size)
- 3 process conditions (e.g., temperature, chemical characteristic, processing time)



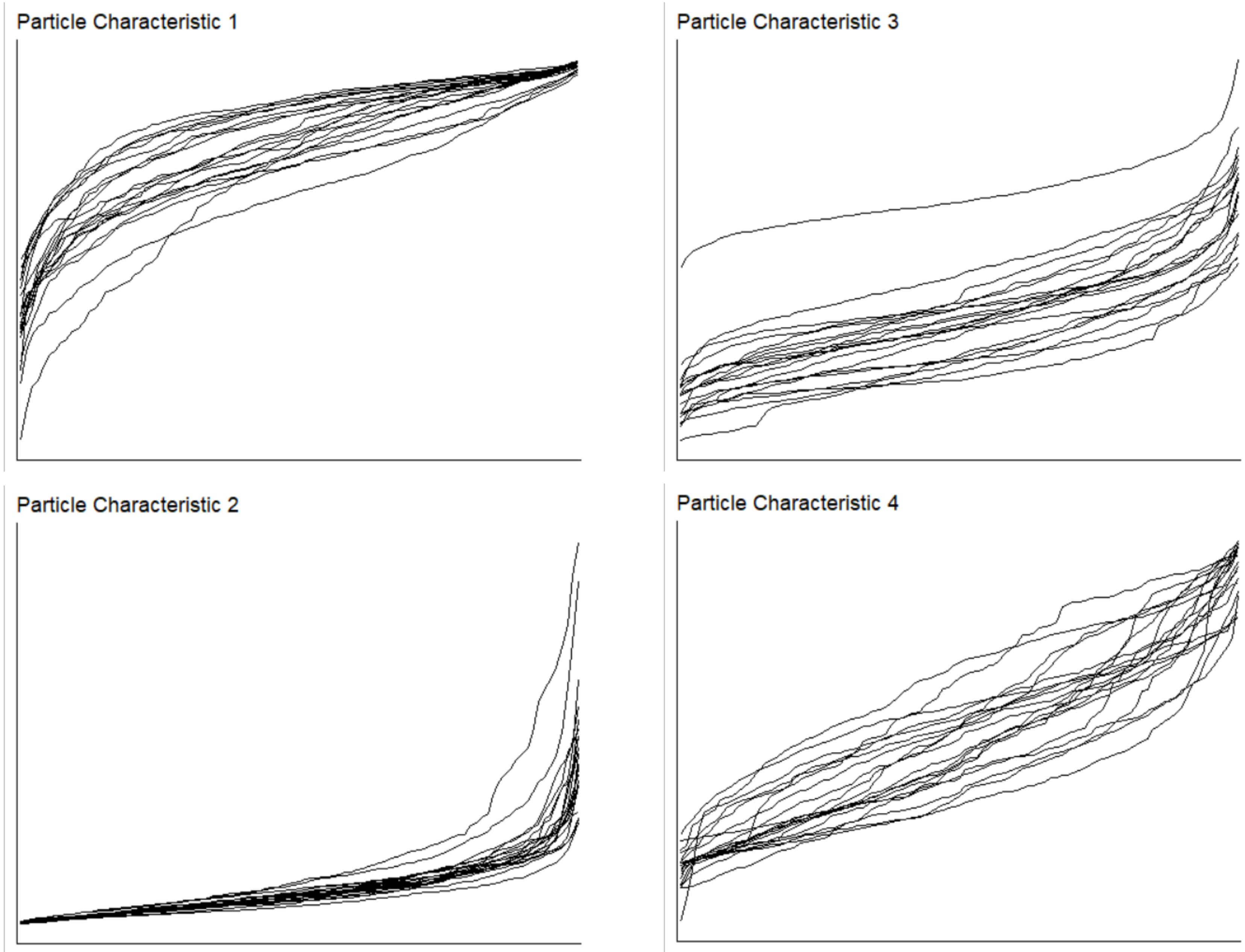
SEM images of Pu particles

# Functional Data



Avg. 640 measurements of each particle characteristic per experimental run

## Empirical CDFs of four particle characteristics



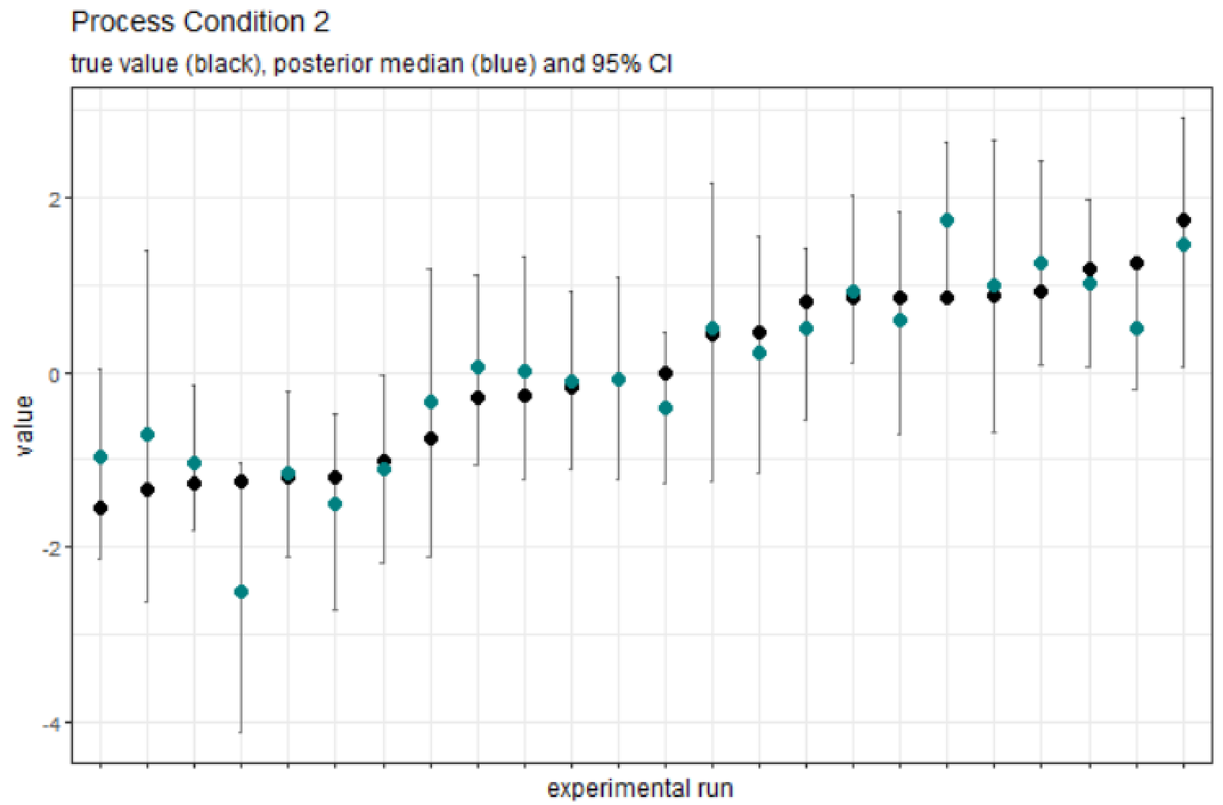
# Implementation

- Forward model: Process Condition = *f*(Particle Characteristics) + ε
  - Identify Particle Characteristics associated with each Process Condition
  - Stepwise regression and LASSO
  - e.g., Particle Characteristics 2 & 3 identified as important for Process Condition 1
- Inverse Model: Particle Characteristics = *f*(Process Conditions) + ε
  - Bayesian implementation of SUR
  - Y matrix: PCs of Particle Characteristics
  - X Matrix: Process Conditions
  - “Masking” matrix
- Four inverse models:
  - Linear
  - Quadratic
  - 2 interaction models

# SUR results

RMSEs for best model. A value of 1 indicates parity with mean-only model.

| Model | Process Cond. 1 | Process Cond. 2 | Process Cond. 3 |
|---|---|---|---|
| Linear | 0.996 | 0.589 | 0.958 |
| Quadratic | 0.974 | 0.525 | 0.883 |
| Interaction (sparse) | 1.014 | 0.592 | 1.097 |
| Interaction (full) | 1.162 | 0.450 | 0.987 |



# Conclusions

- Functional inverse prediction (FIP) framework successfully extended to Seemingly Unrelated Regression (SUR) in a Bayesian context
- MCMC interpolates missing data to perform inverse prediction
- Identification of "important" predictors in forward model (LASSO and stepwise regression) is somewhat subjective. Stepwise regression seemed to do a better job with our data.
- Results on simulated data are promising, although key simplifying assumption is important
  - Further work needed to develop framework for generating correlation matrix $\Omega$ in a functional context
- Results on actual data are more limited, although when it works, it works well.

# Acknowledgements

We would like to thank the National Technical Nuclear Forensics Center (NTNFC) within the Countering Weapons of Mass Destruction (CWMD), formally the Domestic Nuclear Detection Office (DNDO) of the Department of Homeland Security (DHS) for funding this work. In particular, we thank Sandra Gogol, Margaret Goldberg, and the Plutonium Expert Panel for their leadership in making this data set development possible. We'd like to additionally acknowledge the past champions and those critical in laying the foundation of this effort including but not limited to: Jeff Morrison, Ed Thomas, Tom Burr, and John Lewis. Thanks for all your support and guidance

# References


- D. Ries, J. Lewis, A. Zhang, C. Anderson-Cook, Wilkerson, G. M., Wagner, and J. Gravelle. Utilizing distributional measurements of material characteristics from SEM images for inverse prediction. Journal of Nuclear Materials Management, XLVII(1):37–46, 2019.
- Srivastava, Virendra K.; Giles, David E.A. (1987). Seemingly unrelated regression equations models: estimation and inference. New York: Marcel Dekker.
- James Ramsay and Bernard Silverman. Functional Data Analysis. Springer, 2005.
- John R Lewis, Adah Zhang, and Christine M Anderson-Cook. Comparing multiple statistical methods for inverse prediction in nuclear forensics applications. Chemometrics and Intelligent Laboratory Systems, 175:116–129, 2018.