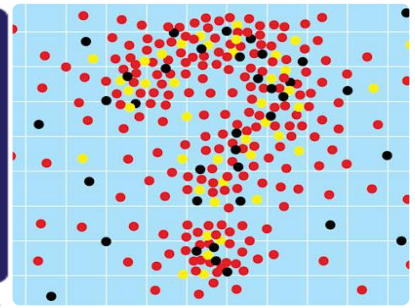
SAND2016-7347C

SANDIA STATISTICAL SCIENCES

Plan. Analyze. Inform.



Sandia National Laboratories

Exceptional service in the national interest

Selecting an Informative/Discriminating Multivariate Response for Inverse Prediction

Edward V. Thomas[1], John Lewis[1*]

Christine Anderson-Cook[2], Tom Burr[2], Michael S. Hamada[2], Adah Zhang[1]

[1] Sandia National Laboratories, Albuquerque, NM

[2] Los Alamos National Laboratory, Los Alamos, NM

*Presenting Author

Work funded by the Department of Homeland Security

National Technical Nuclear Forensics Center

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

Inverse Prediction

- Inferring the input factors (x) of models based on responses (y)
  - ***Classical calibration*** – inverting forward/causal models $y \approx f(x)$
  - ***Inverse calibration*** – $x \approx g(y)$ (e.g. regularization methods: PCR, PLSR)
- Motivation of this work - analysis of nuclear materials
  - Identify the processing conditions used to manufacture interdicted material based on physical and chemical measurements
- Focus here is on methods based on *classical calibration* (inverting forward models)
  - Experimentally or scientific-based models

Down-Selecting a Multivariate Response

- Motivation from nuclear forensics
  - Many types of physical/chemical measurements
- Constraints on the number of measurements:
  - Likely a limited amount of interdicted material
  - Destructive nature of some measurements
- Strategy to down-select an *informative/discriminating* subset of responses from a candidate set
  - Informative – precise predictions (small prediction variance)
  - Discriminating – effects of factors on the various responses are sufficiently dissimilar
- Depends on an assumed forward model for each of q responses related to p causal factors

$$Y_i = f_i(\beta_i; X) + \epsilon_i, \qquad i = 1, 2, \ldots, q$$

Y_i – i^{th} response, β_i – model parameters, X – factors, ϵ_i – mean zero error

Predicting X: Least Squares Accounting for Errors

- Estimate each model: $Y_i \approx f_i(\hat{\beta}_i; X), i = 1, \dots, q$
- A new observed multivariate response ($\boldsymbol{Y}^* = (Y_1^*, \dots, Y_q^*)^\top$) is used to predict unknown levels of factors X^*

$$Y_i^* = f_i(\beta_i, X^*) + \epsilon_i^* \text{ with}$$

- Goal: Find an "optimal" solution $\hat{X}^*$ such that $\hat{Y}_i^* \approx Y_i^*, i = 1, \dots, q$ where $\hat{Y}_i^* = f_i(\hat{\beta}_i, \hat{X}^*)$
- Prediction error at candidate solution $\hat{X}$: $d_i = \hat{Y}_i - Y_i^*$ where $\hat{Y}_i = f_i(\hat{\beta}_i, \hat{X})$

$$\hat{X}^* = \text{argmin}_X \, D^\top V^{-1} D$$

$$D = (d_1, \dots, d_q)^\top, \quad V = V(\hat{X}) = cov(D)$$

- Solved iteratively – requires $\hat{\beta}_i$, initial $\hat{X}$, and $\hat{V}(\hat{X})$

Estimating V

- To estimate V, first decompose d_i:

$$d_i = \lambda_i + \omega_i - \epsilon_i^* \text{ where,}$$

$$\lambda_i = f_i(\hat{\beta}_i, \hat{X}) - f_i(\beta_i, \hat{X}) \qquad \text{and} \qquad \omega_i = f_i(\beta_i, \hat{X}) - f_i(\beta_i, X^*)$$

- Interpretation of components of d_i
  - $\lambda_i = f_i(\hat{\beta}_i, \hat{X}) - f_i(\beta_i, \hat{X})$: error due to uncertainty in model parameters
  - $\omega_i = f_i(\beta_i, \hat{X}) - f_i(\beta_i, X^*)$: error due to uncertainty in the candidate solution $\hat{X}$
- Assuming properly specified models and unbiased solutions: $E(d_i) = 0$ and

$$V = V_\lambda(\hat{X}) + V_\omega(\hat{X}) + 2cov_{\lambda\omega}(\hat{X}) + V_\epsilon$$

- V_λ, V_ω can be estimated using first order approximations, can use residuals to estimate V_ϵ
- Simplifying assumptions: V_λ, V_ω, V_ϵ assumed diagonal, covariance 0.
- Solution considers the uncertainty in predicted response – switch role of X and β, related to "errors-in-variables" literature

Variance-Covariance of Prediction

- Assume forward models are 1) Continuous functions of the factors 2) Not highly non-linear
- First-order linear approximation to $Y_i^* = f_i(\beta_i, X^*)$ near X^*

$$Y_i^* = f_i(\beta_i, X^*) \approx f_i\left(\beta_i; \hat{X}^*\right) + \sum_{j=1}^{p} J_{ij}\left(\hat{X}_j^*\right)\left(\hat{X}_j^* - X_j^*\right), \text{ where } J_{ij}\left(\hat{X}_j^*\right) = \frac{\partial}{\partial x_j} f_i\left(\beta_i; \hat{X}_j^*\right).$$

- Locally linear regression of $\boldsymbol{Y}^*$ on $\hat{J}_{ij}(X^*)$ leads to an estimate of the covariance of $\hat{X}^*$

$$\hat{C}_{\hat{X}^*} = \left(\hat{J}^T\left(\hat{X}^*\right)\hat{V}^{-1}\hat{J}\left(\hat{X}^*\right)\right)^{-1}, \quad \text{where } \hat{J}_{ij}\left(\hat{X}^*\right) = \frac{\partial}{\partial x_j} f_i\left(\hat{\beta}_i; \hat{X}^*\right).$$

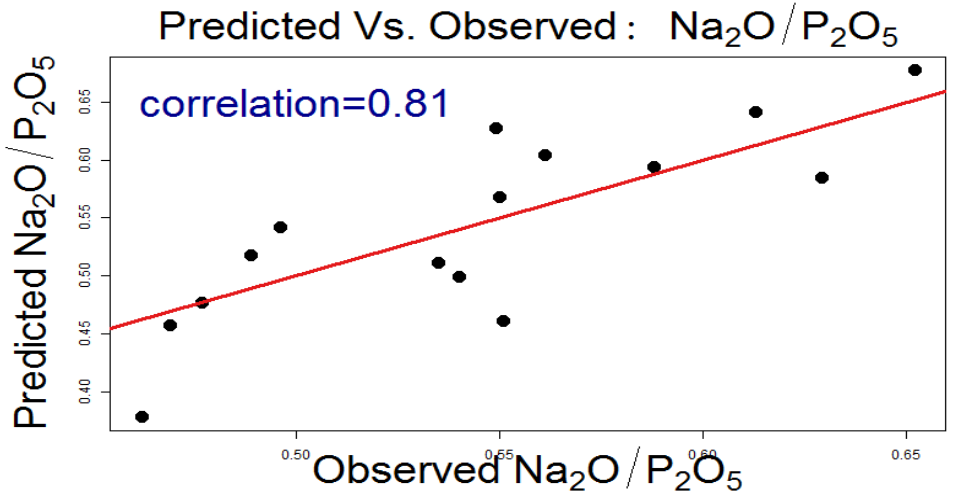
- Multivariate response is
  - *Informative* if diagonal elements are sufficiently small
  - *Discriminating* if the off diagonal elements are sufficiently small

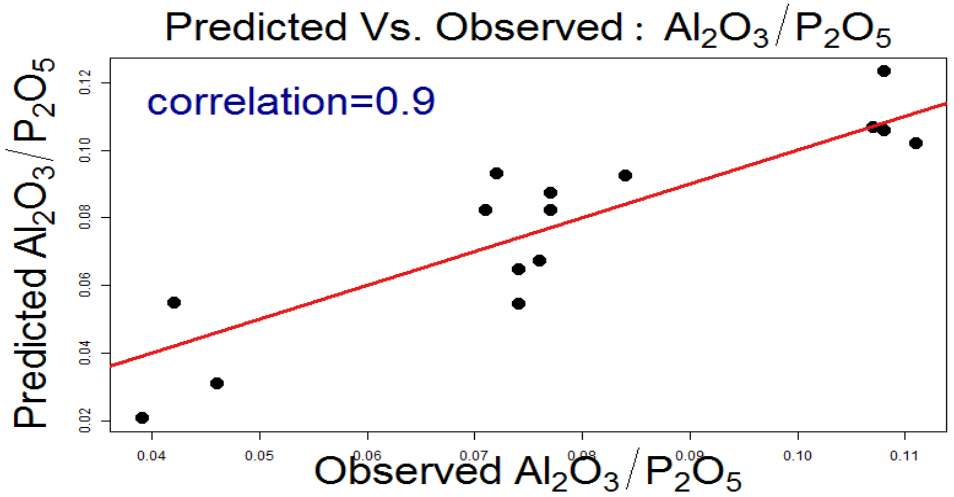
Glass Composition Example

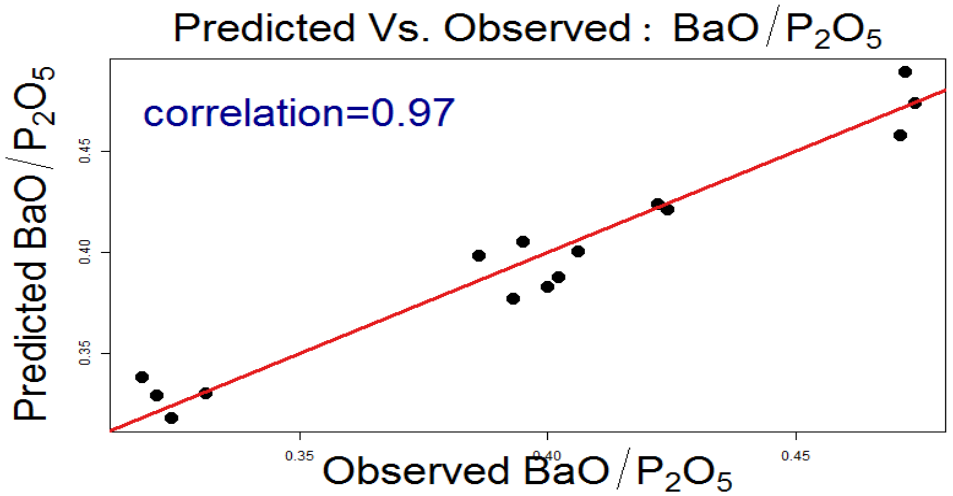
- Study to investigate how glass properties vary as a function of composition
  - Constituents are mole ratios: $X_1 = Na_2O/P_2O_5, X_2 = BaO/P_2O_5,\ X_3 = Al_2O_3/P_2O_5$
- ***Goal:*** Predict constituents based on six glass properties ($i = 1, \ldots, 6$)
- Use experimental data to estimate models (simple linear): $\hat{Y}_i = \hat{\beta}_{i0} + \hat{\beta}_{i1}X_1 + \hat{\beta}_{i2}X_2 + \hat{\beta}_{i3}X_3$
- Best forward models (by R^2 metric) are of **density** and **index of refraction**
  - Don't depend on Na_2O

Property: $i = 1, \ldots, 6$	$\hat{\beta}_0$	$\hat{\beta}_1\ (Na_2O)$	$\hat{\beta}_2\ (Ba0)$	$\hat{\beta}_3\ (Al_2O_3)$	$\hat{\sigma}_\varepsilon$	R^2
1. Coeff. of Thermal Expansion	**155.8 (6.1)**	**70.59(10.3)**	**----**	**-216.5(31)**	**3.12**	**0.86**
2. Softening Temperature	**392.7(15.5)**	**-104.7(24.6)**	**----**	**694.6(63)**	**5.73**	**0.93**
3. Glass Transition Temperature	**374.8(14.7)**	**-104.5(23.7)**	**----**	**412.1(66)**	**6.39**	**0.82**
4. Crystallization Temperature	**570.5(28.9)**	**-219.5(48.5)**	**----**	**709.8(147)**	**14.7**	**0.74**
5. Density	**2.534(0.022)**	**----**	**1.113(0.051)**	**0.484(0.119)**	**0.0119**	**0.97**
6. Index of Refraction	**1.498(0.003)**	**0.0097(0.004)**	**0.0834(0.005)**	**0.1036(0.0123)**	**0.00113**	**0.97**

Prediction of Glass Composition







- Multivariate response good for predicting $Ba0$ and Al_2O_3, not as good for Na_20
  - Strongest models don't depend on Na_20
  - Intuition: need strong forward models for inverse prediction
- Density is responsible for the precise predictions of BaO – Barium is very dense compared to other constituents

Comparing Different Subsets of Responses

- Use $\hat{C}_{\hat{X}^*} = \left(\hat{J}^T(\hat{X}^*)\hat{V}^{-1}\hat{J}(\hat{X}^*)\right)^{-1}$ to estimate average prediction variance for different subsets of the multivariate response

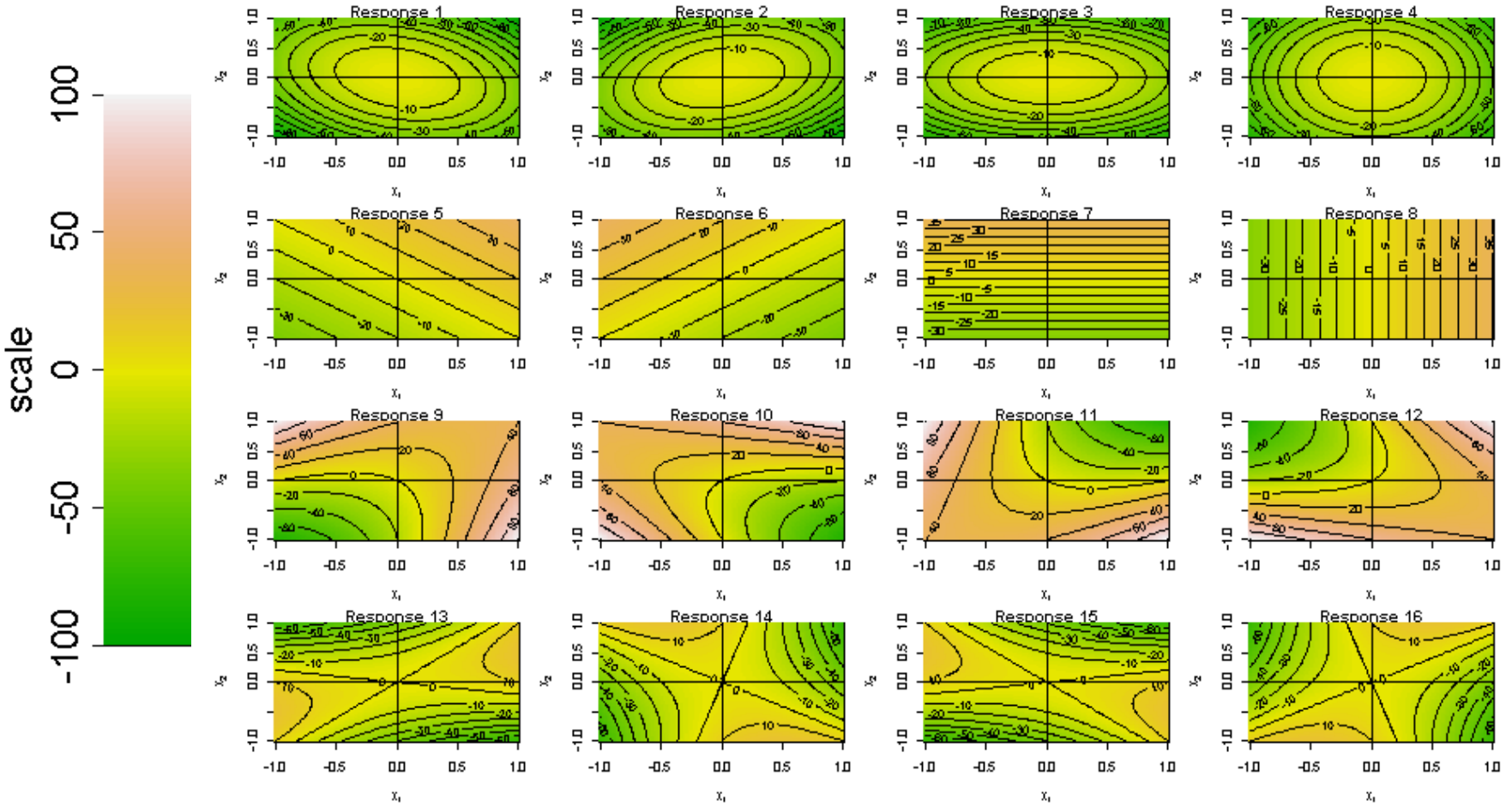
Subset	$\sqrt{Var_{avg}(X_1)}$ (Na_2O)	$\sqrt{Var_{avg}(X_2)}$ (BaO)	$\sqrt{Var_{avg}(X_3)}$ (Al_2O_3)
All Responses	0.08	0.013	0.018
Excluding Density	0.08	0.02	0.019
Excluding Index of Ref.	0.08	0.013	0.019

- Excluding density results in ~1.5 times increase in the root prediction variance of X_2
- Multivariate response is less informative for predicting X_2 if density is excluded
- Excluding index of refraction is not as detrimental – despite a good forward model

Further Investigation: 16 Known Response Surfaces

Goal: Choose a subset of the 16 response surfaces that is ***informative*** (small prediction variance) and ***discriminating*** (sufficiently dissimilar shapes) for prediction of X_1 and X_2

16 response surfaces



1-4: peaks

5-8 : hillsides

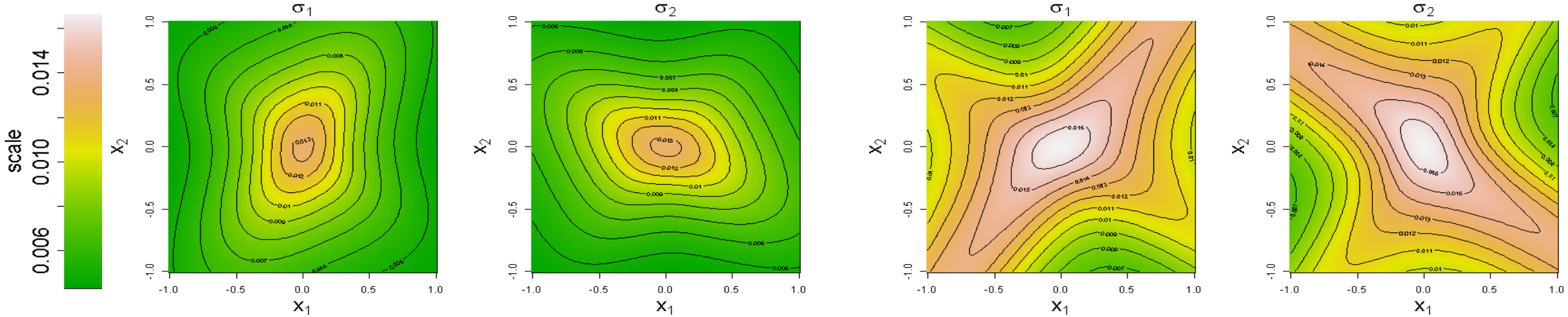
9-12: rising ridges

13-16: saddles

Prediction St. Dev. Across Design Space

- Analytical results using $\hat{C}_{\hat{X}^*} = \left(\hat{J}^T(\hat{X}^*)\hat{V}^{-1}\hat{J}(\hat{X}^*)\right)^{-1}$
- Two candidate sets of responses: $S = \{1,2,\ldots,16\}$ and $S = \{9,10,11,12\}$
- $\sigma_1 = sd(\hat{x}_1^*), \sigma_2 = sd(\hat{x}_2^*)$

$S = \{1,2,\ldots,16\}$ $S = \{9,10,11,12\}$



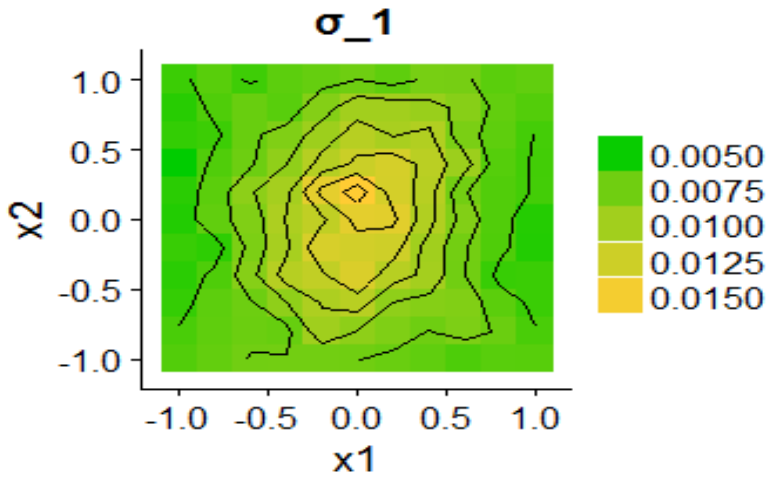
- Value depends on X^*. Smaller standard deviation across design space when using all 16 responses
- Relative increase using just four responses is small across the design space

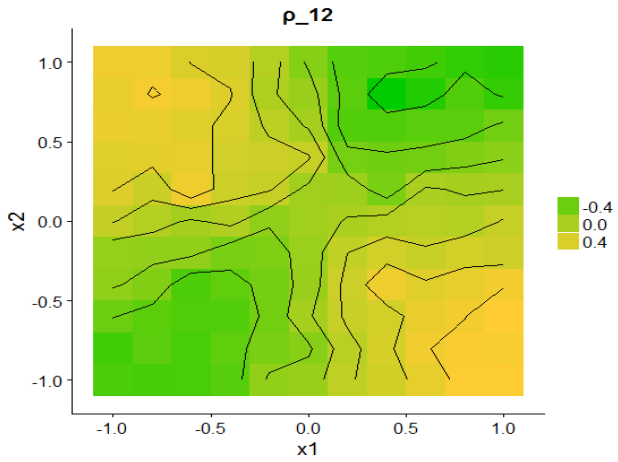
Simulation Results compared to Analytical

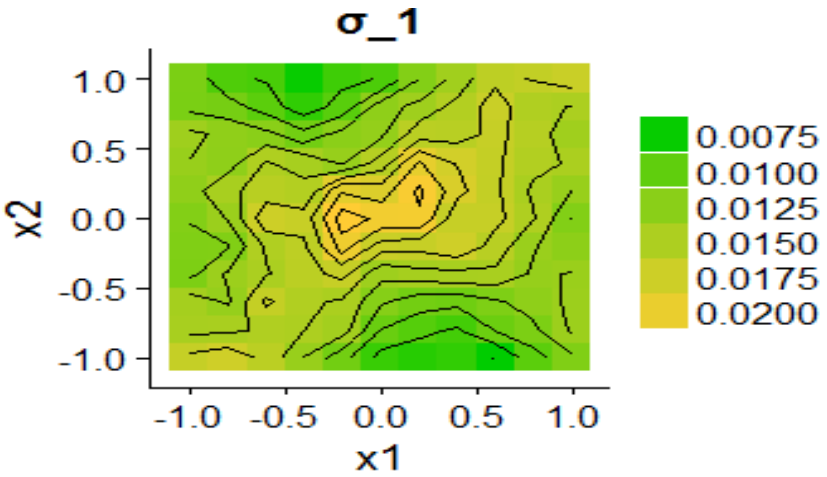
Qualitative agreement when computing prediction variance using simulations

$S = \{1,2,\ldots,16\}$ $S = \{9,10,11,12\}$

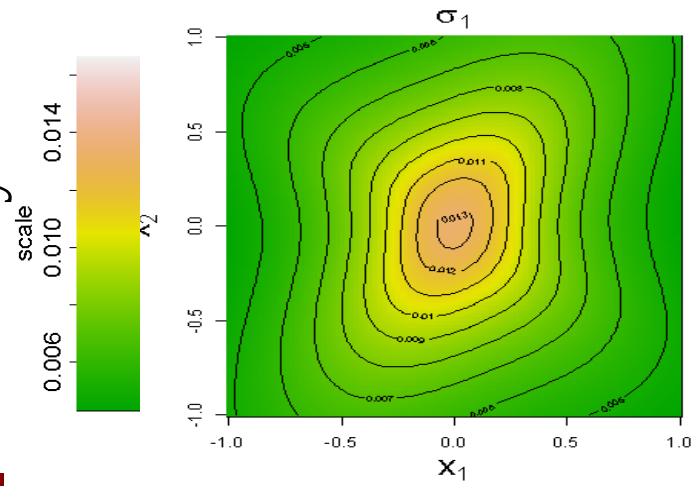
Simulation

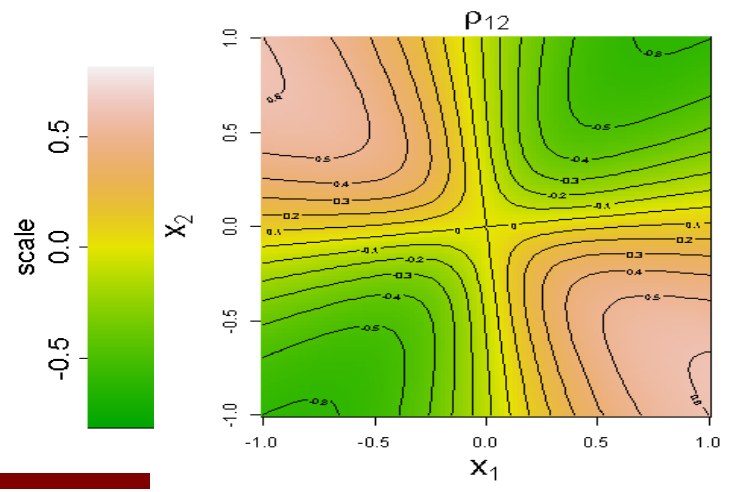


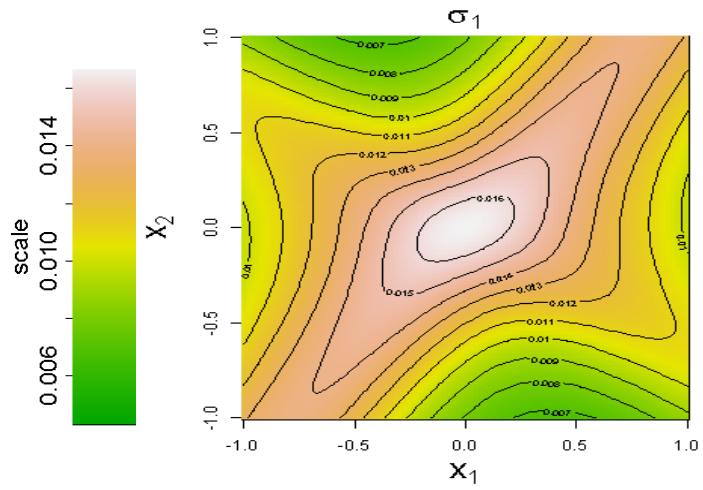




Analytical







Quantifying Differences in Prediction Variance

Subset	$\sqrt{Var_{avg}\ \hat{X}_1^*}$	$\sqrt{Var_{avg}\ \hat{X}_2^*}$
{1,2,…,16}	0.0075	0.0075
{7,8}	0.0286	0.0286
{3,7,9,13}	0.0291	0.0154
{9,10,11,12}	0.0121	0.0121

~ 4x larger (for $\hat{X}_1^*$)

~ 1.6x larger for ¼ of responses

- Set {9,10,11,12} is a good choice for prediction across the space of interest if constraints exist in obtaining new measurements
- Responses in this set complement each other well – i.e. steep contours are present in one or more of the responses throughout the range of interest

Summary

- Described a method for assessing a multivariate response's usefulness in inverse prediction
- Accounts for multiple sources of uncertainty: model parameter uncertainty, solution uncertainty, measurement error
- Derived first order approximation to the covariance of prediction $\hat{C}_{\hat{X}^*} = \left(J^T(\hat{X}^*)\hat{V}^{-1}J(\hat{X}^*)\right)^{-1}$
  - *Informative* and *discriminating* if the elements of $\hat{C}_{\hat{X}^*}$ are sufficiently small
    - Level of collinearity in the Jacobian J is an indicator of discriminating ability (as well as level of redundancy)
- Method can be used to down-select responses from a candidate set when constraints exists in measuring new observations (e.g. nuclear forensic applications)
  - Ideal combination: responses with strong difference across input space, and several responses with different shaped relationships
- Concentrated on empirical models derived from well designed experiments
- Methods can be applied to scientific models – may need numerical estimation of partial derivatives

Continuing Research

- Different objective functions – robust predictions
- Bayesian methods incorporating prior information on X^* to down-select multivariate response
- Assess the degree of match between new samples and data observed in the experiment
  - Experiment is highly controlled. Future sample don't come from these experiments. Are the predictions reliable?
  - Use predictions across several methods to assess the degree of match?

Method	$\sqrt{Var_{avg}\ \hat{X}_1^*}$ w/ 16 responses	$\sqrt{Var_{avg}\ \hat{X}_1^*}$ w/ 4 responses
Forward with LS	**0.0075**	**0.0125**
PCR	**0.014**	**0.017**
PLSR	**0.014**	**0.017**

Estimates of root average prediction variance based on simulations: $n = 27, \sigma = 1$, and 100 repetitions.

One repetition : generate data from the response surfaces using 3 – level full factorial with 3 replicates, predict simulated data on a grid spanning the design space