DISCUSSION ON

“*Bayesian Smoothing and Feature Selection using Variational Automatic Relevance Determination*”

By Zihe Liu, Diptarka Saha and Feng Liang
Presented by Feng Liang

O’Bayes Conference 2025
Stavros Niarchos Foundation Cultural Center
Athens, Greece

Discussant: Xenia Miscouridou
Dep of Mathematics and Statistics, University of Cyprus
Dep of Mathematics and IX Centre, Imperial College London

Πανεπιστήμιο Κύπρου
University of Cyprus

Imperial College London

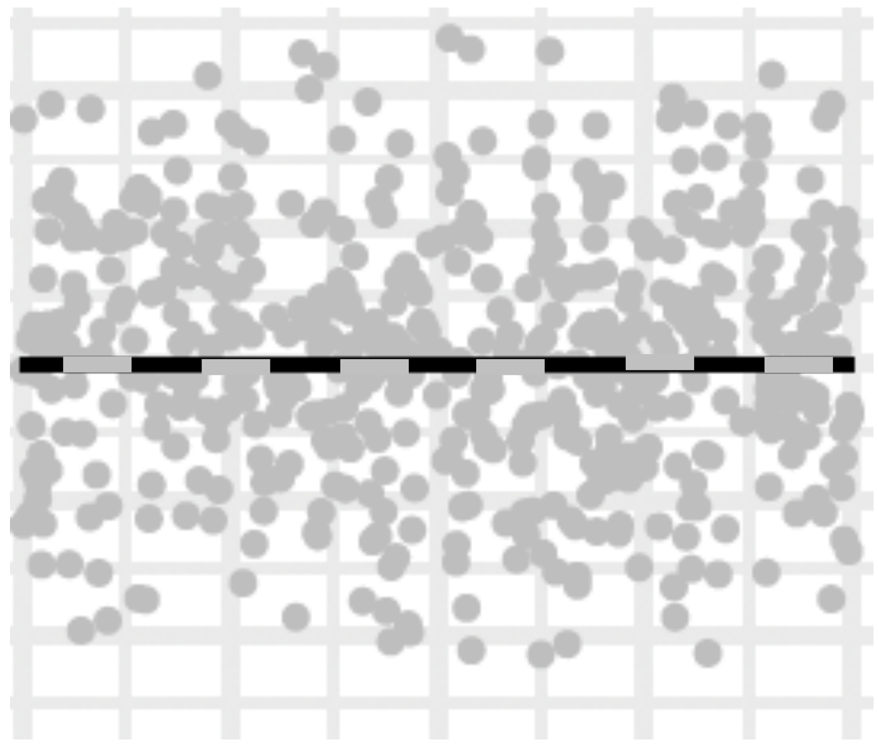
Problem Formulation

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

Problem Formulation

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

VARIABLE SELECTION



True $f_j = 0$

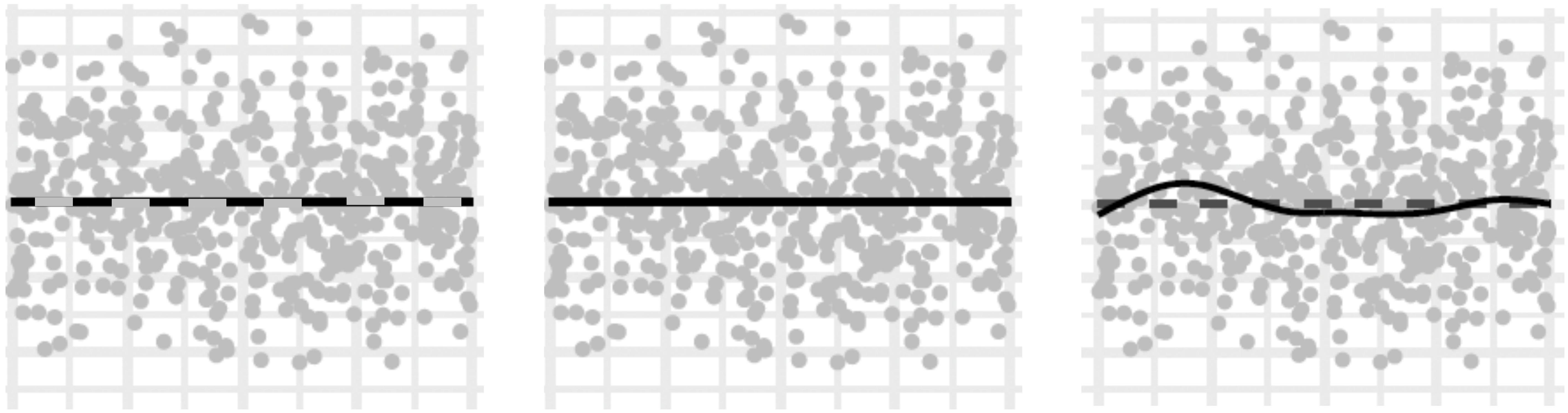
grey dots: data
black line: $\hat{f}_j$
dotted line: $f_j = 0$

Figures from Liu et al, 2025

Problem Formulation

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

VARIABLE SELECTION



True $f_j = 0$ | Good estimation | Bad estimation

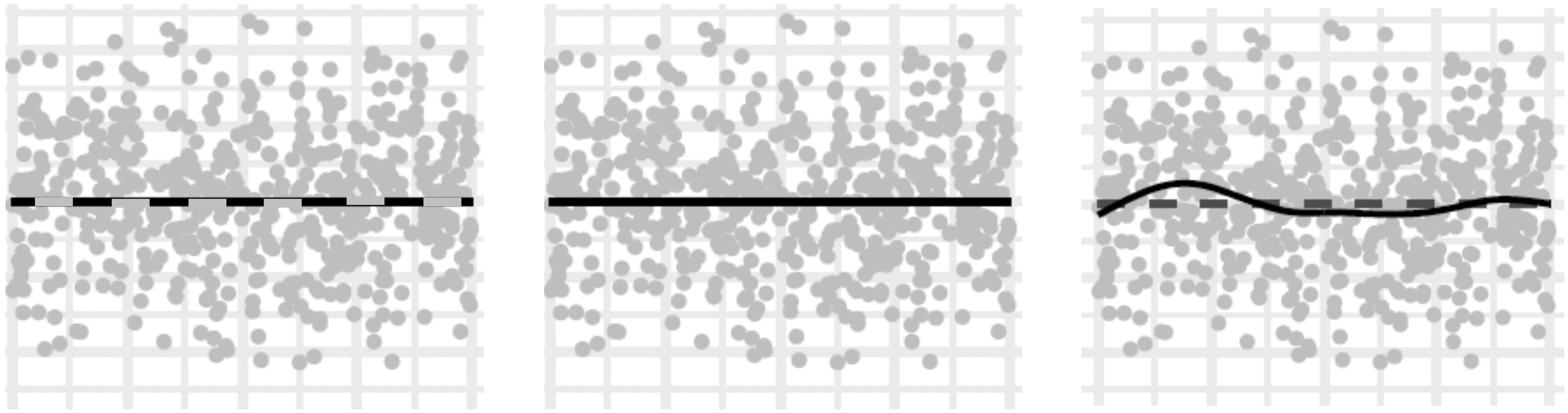
grey dots: data
black line: $\hat{f}_j$
dotted line: $f_j = 0$

Figures from Liu et al, 2025

Problem Formulation

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

VARIABLE SELECTION



True $f_j = 0$ Good estimation Bad estimation

grey dots: data
black line: $\hat{f}_j$
dotted line: $f_j = 0$

METHODS

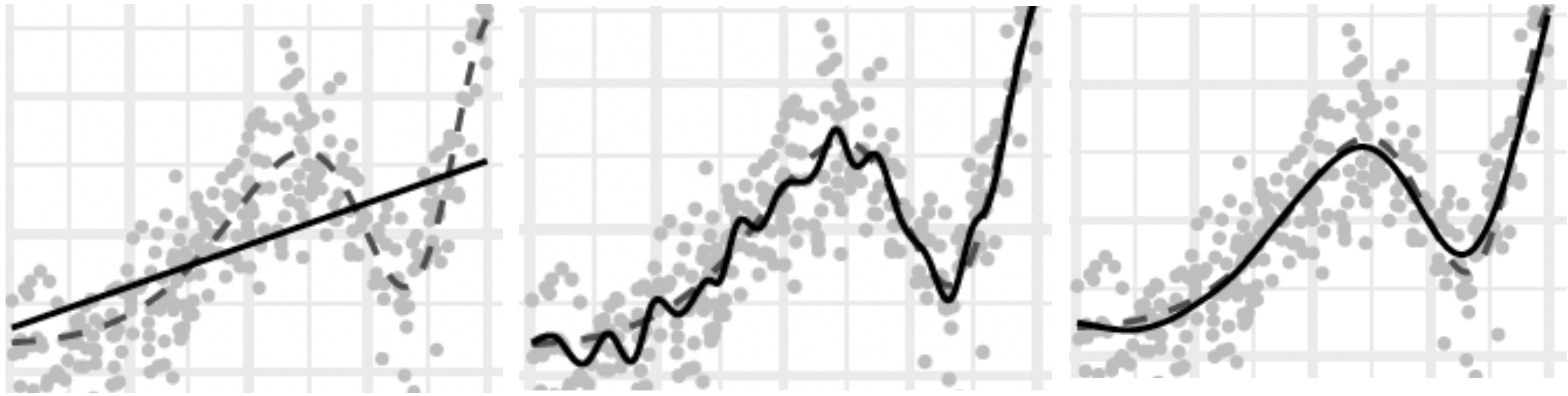
- gLasso-type penalty: COSSO, SPAM, GAMSEL
- BAYESIAN: spike and slab

Yuan and Lin, 2006; Lin and Zhang, 2006; Liu et al., 2007; Ravikumar et al., 2009; Chouldechova and Hastie, 2015; He and Wand, 2022; Fabian Scheipl and Kneib, 2012

Problem Formulation

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

SMOOTHING



Underfitting Overfitting Right smoothness!

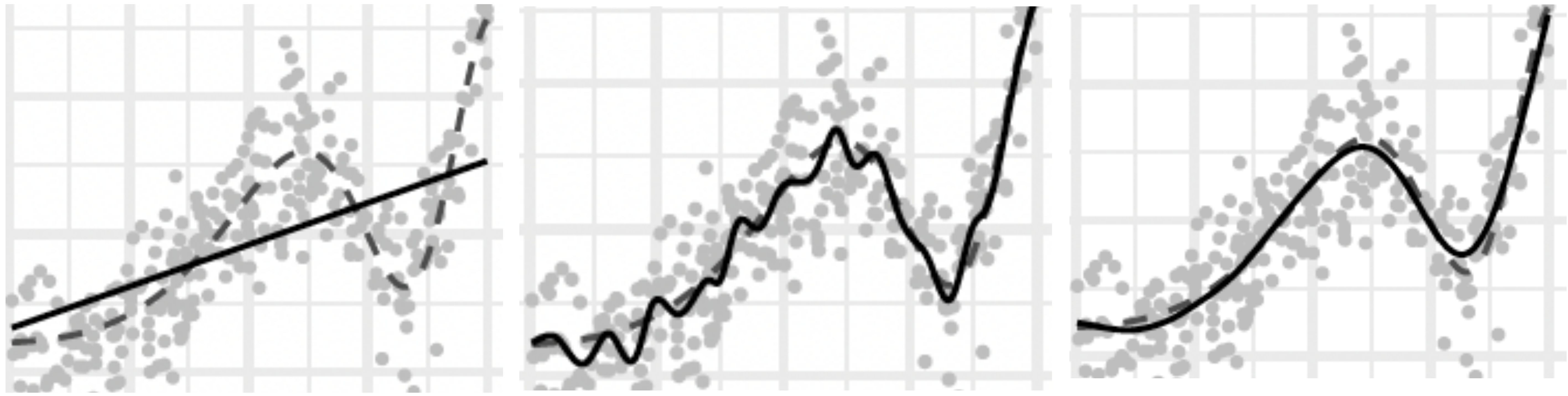
grey dots: data
black line: $\hat{f}_j$
dotted line: $f_j \neq 0$

Figures from Liu et al, 2025

Problem Formulation

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

SMOOTHING



Underfitting Overfitting Right smoothness!

grey dots: data
black line: $\hat{f}_j$
dotted line: $f_j \neq 0$

METHODS

- Smoothing splines with ridge type penalty
- BAYESIAN: Normal prior on the coefficients

Wahba, 1990; Wood, 2017, Figures from Liu et al, 2025

Variational Automatic Relevance Determination

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

Liu et al, 2025

Variational Automatic Relevance Determination

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

The proposed method

✓ performs both smoothing and variable selection

Liu et al, 2025

Variational Automatic Relevance Determination

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

The proposed method

✓ performs both smoothing and variable selection

✓ can classify a feature's contribution as linear, non linear or zero

Liu et al, 2025

Variational Automatic Relevance Determination

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

The proposed method

- ✓ performs both smoothing and variable selection
- ✓ can classify a feature's contribution as linear, non linear or zero
- ✓ can achieve exact sparsity

Liu et al, 2025

Variational Automatic Relevance Determination

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

The proposed method

- ✓ performs both smoothing and variable selection
- ✓ can classify a feature's contribution as linear, non linear or zero
- ✓ can achieve exact sparsity
- ✓ is tuned with a single hyper parameter

Liu et al, 2025

Variational Automatic Relevance Determination

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

The proposed method

✓ performs both smoothing and variable selection

✓ can classify a feature's contribution as linear, non linear or zero

✓ can achieve exact sparsity

✓ is tuned with a single hyper parameter

✓ is efficient

Bayesian Smoothing and Feature Selection using Variational Automatic Relevance Determination, Liu, Saha, Liang, 2025

Variational Automatic Relevance Determination

- Simultaneous smoothing and variable selection for additive models $f(x_1, \ldots, x_p) = f_1(x_1) + \ldots + f_p(x_p)$

The proposed method

- ✓ performs both smoothing and variable selection
- ✓ can classify a feature's contribution as linear, non linear or zero
- ✓ can achieve exact sparsity
- ✓ is tuned with a single hyper parameter
- ✓ is efficient
- ✓ outperforms other methods in accuracy in estimation and selection

Liu et al, 2025

Setup

Setup

n responses $\mathbf{y} = (y_1, \ldots, y_n)^T$
p predictors $\mathbf{x}_j = \{(x_{j1j}, \ldots, x_{nj})^T\}_{j=1}^{p}$
$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ iid $N(0, \sigma^2)$

$$\mathbf{y} = \sum_{j=1}^{p} f_j(\boldsymbol{x}_j) + \boldsymbol{\epsilon}$$

where

Each f_j is represented through a basis expansion of linear and nonlinear terms

$$f_j(x) = \beta_0 x + \sum_{k=1}^{d_j} \beta_{jk} h_{jk}(x)$$

Setup

n responses $\mathbf{y} = (y_1, \ldots, y_n)^T$
p predictors $\mathbf{x}_j = \{(x_{j1j}, \ldots, x_{nj})^T\}_{j=1}^{p}$
$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ iid $N(0, \sigma^2)$

$$\mathbf{y} = \sum_{j=1}^{p} f_j(\boldsymbol{x}_j) + \boldsymbol{\epsilon}$$

where

Each f_j is represented through a basis expansion of linear and nonlinear terms

$$f_j(x) = \beta_0 x + \sum_{k=1}^{d_j} \beta_{jk} h_{jk}(x)$$

$$\boxed{\text{We then get a matrix representation}}$$

$$\mathbf{y} = \sum_{j=1}^{2p} Z_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon}$$

Setup

n responses $\mathbf{y} = (y_1, \ldots, y_n)^T$
p predictors $\mathbf{x}_j = \{(x_{j1j}, \ldots, x_{nj})^T\}_{j=1}^p$
$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ iid $N(0, \sigma^2)$

$$\mathbf{y} = \sum_{j=1}^{p} f_j(\boldsymbol{x}_j) + \boldsymbol{\epsilon}$$

where

Each f_j is represented through a basis expansion of linear and nonlinear terms

$$f_j(x) = \beta_0 x + \sum_{k=1}^{d_j} \beta_{jk} h_{jk}(x)$$

$\boxed{\text{We then get a matrix representation}}$

$$\boldsymbol{y} = \underbrace{Z_1\boldsymbol{\beta}_1 + \ldots + Z_p\boldsymbol{\beta}_p}_{\text{p nonlinear terms}} + \underbrace{Z_{p+1}\boldsymbol{\beta}_{p+1} + \ldots + Z_{2p}\boldsymbol{\beta}_{2p}}_{\text{p linear terms}} + \boldsymbol{\epsilon}$$

Setup

n responses $\mathbf{y} = (y_1, \ldots, y_n)^T$
p predictors $\mathbf{x}_j = \{(x_{j1j}, \ldots, x_{nj})^T\}_{j=1}^p$
$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ iid $N(0, \sigma^2)$

$$\mathbf{y} = \sum_{j=1}^{p} f_j(\boldsymbol{x}_j) + \boldsymbol{\epsilon}$$

where

Each f_j is represented through a basis expansion of linear and nonlinear terms

$$f_j(x) = \beta_0 x + \sum_{k=1}^{d_j} \beta_{jk} h_{jk}(x)$$

Estimating the β_j can classify linear, non linear and zero contributions

$$\boldsymbol{y} = \underbrace{Z_1\boldsymbol{\beta}_1 + \ldots + Z_p\boldsymbol{\beta}_p}_{\text{p nonlinear terms}} + \underbrace{Z_{p+1}\boldsymbol{\beta}_{p+1} + \ldots + Z_{2p}\boldsymbol{\beta}_{2p}}_{\text{p linear terms}} + \boldsymbol{\epsilon}$$

Setup

n responses $\mathbf{y} = (y_1, \ldots, y_n)^T$
p predictors $\mathbf{x}_j = \{(x_{j1j}, \ldots, x_{nj})^T\}_{j=1}^p$
$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ iid $N(0, \sigma^2)$

$$\mathbf{y} = \sum_{j=1}^{p} f_j(\boldsymbol{x}_j) + \boldsymbol{\epsilon}$$

where

Each f_j is represented through a basis expansion of linear and nonlinear terms

$$f_j(x) = \beta_0 x + \sum_{k=1}^{d_j} \beta_{jk} h_{jk}(x)$$

Estimating the β_j can classify linear, non linear and zero contributions

$$\boldsymbol{y} = \underbrace{Z_1\boldsymbol{\beta}_1 + \ldots + Z_p\boldsymbol{\beta}_p}_{\text{p nonlinear terms}} + \underbrace{Z_{p+1}\boldsymbol{\beta}_{p+1} + \ldots + Z_{2p}\boldsymbol{\beta}_{2p}}_{\text{p linear terms}} + \boldsymbol{\epsilon}$$

Q1*:*
What about large p?
In relation to feasibility, efficiency, and theoretical results

Setup

$$\boldsymbol{y} = \underbrace{Z_1\boldsymbol{\beta}_1 + \ldots + Z_p\boldsymbol{\beta}_p}_{\text{p nonlinear terms}} + \underbrace{Z_{p+1}\boldsymbol{\beta}_{p+1} + \ldots + Z_{2p}\boldsymbol{\beta}_{2p}}_{\text{p linear terms}} + \boldsymbol{\epsilon}$$

Group ARD Prior

$$\boldsymbol{y} = \underbrace{Z_1\boldsymbol{\beta}_1 + \ldots + Z_p\boldsymbol{\beta}_p}_{\text{p nonlinear terms}} + \underbrace{Z_{p+1}\boldsymbol{\beta}_{p+1} + \ldots + Z_{2p}\boldsymbol{\beta}_{2p}}_{\text{p linear terms}} + \boldsymbol{\epsilon}$$

$\boxed{\text{Prior } p(\beta)}$

$$\left(\boldsymbol{\beta}_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(0, r_j^2 I_{d_j})$$

smoothness and
sparsity parameter

Mackay, 1995

Variational Inference

$$\boldsymbol{y} = \underbrace{Z_1\boldsymbol{\beta}_1 + \ldots + Z_p\boldsymbol{\beta}_p}_{\text{p nonlinear terms}} + \underbrace{Z_{p+1}\boldsymbol{\beta}_{p+1} + \ldots + Z_{2p}\boldsymbol{\beta}_{2p}}_{\text{p linear terms}} + \boldsymbol{\epsilon}$$

Prior $p(\beta)$

$$\left(\boldsymbol{\beta}_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(0, r_j^2 I_{d_j})$$

smoothness and
sparsity parameter

Variational $q(\beta)$

$$\left(\boldsymbol{\beta}_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(\mu_j, \Phi_j)$$

Mackay, 1995

Variational Inference

$$\boldsymbol{y} = Z_1\boldsymbol{\beta}_1 + \ldots + Z_p\boldsymbol{\beta}_p + Z_{p+1}\boldsymbol{\beta}_{p+1} + \ldots + Z_{2p}\boldsymbol{\beta}_{2p} + \boldsymbol{\epsilon}$$

p nonlinear terms

p linear terms

Prior $p(\beta)$

$$\left(\boldsymbol{\beta}_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(0, r_j^2\, I_{d_j})$$

Q2:

Implications of a different covariance matrix?

smoothness and
sparsity parameter

Variational $q(\beta)$

$$\left(\beta_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(\mu_j, \Phi_j)$$

Mackay, 1995; Neal, 1996

α-Variational ELBO

$$\boldsymbol{y} = \underbrace{Z_1\boldsymbol{\beta}_1 + \ldots + Z_p\boldsymbol{\beta}_p}_{\text{p nonlinear terms}} + \underbrace{Z_{p+1}\boldsymbol{\beta}_{p+1} + \ldots + Z_{2p}\boldsymbol{\beta}_{2p}}_{\text{p linear terms}} + \boldsymbol{\epsilon}$$

$\boxed{\text{Prior } p(\beta)}$

$$\left(\boldsymbol{\beta}_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(0, r_j^2 I_{d_j})$$

$\boxed{\text{Variational } q(\beta)}$

$$\left(\beta_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(\mu_j, \Phi_j)$$

$$L(\mu, \Phi, r^2) = -E_q\left[\log p\left(y \,|\, \beta_1, \ldots, \beta_{2p}\right)\right] + \tilde{\alpha}\, KL\left(q \,||\, p\right)$$

Yang et al, 2020; Higgins et al, 2017

α-Variational ELBO

$$\boldsymbol{y} = \underbrace{Z_1\boldsymbol{\beta}_1 + \ldots + Z_p\boldsymbol{\beta}_p}_{\text{p nonlinear terms}} + \underbrace{Z_{p+1}\boldsymbol{\beta}_{p+1} + \ldots + Z_{2p}\boldsymbol{\beta}_{2p}}_{\text{p linear terms}} + \boldsymbol{\epsilon}$$

Prior $p(\beta)$

$$\left(\boldsymbol{\beta}_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(0, r_j^2 I_{d_j})$$

Variational $q(\beta)$

$$\left(\beta_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(\mu_j, \Phi_j)$$

$$L(\mu, \Phi, r^2) = -E_q\left[\log p\left(y\,|\,\beta_1, \ldots, \beta_{2p}\right)\right] + \tilde{\alpha}\,KL\left(q\,||\,p\right)$$

Q3:
Other forms of modifications on KL or on other divergences?

Yang et al, 2020; Higgins et al, 2017

Coordinate Descent

$\boxed{\text{Prior } p(\beta)}$

$$\left(\boldsymbol{\beta}_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(0, r_j^2\, I_{d_j})$$

$\boxed{\text{Variational } q(\beta)}$

$$\left(\beta_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(\mu_j, \Phi_j)$$

$$L(\mu, \Phi, r^2) = -\,E_q\left[\log p\left(y\,|\,\beta_1, \ldots, \beta_{2p}\right)\right] + \tilde{\alpha}\, KL\left(q\,||\,p\right)$$

Coordinate Descent

Learn $\left(\Phi_j, \mu_j, r_j^2\right)_{j=1}^{2p}$

- Can simplify into a univariate problem on r_j^2
- Grid search to find the optimal values

Prior $p(\beta)$

$$\left(\boldsymbol{\beta}_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(0, r_j^2 I_{d_j})$$

Variational $q(\beta)$

$$\left(\beta_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(\mu_j, \Phi_j)$$

$$L(\mu, \Phi, r^2) = -E_q\left[\log p\left(y \,|\, \beta_1, \ldots, \beta_{2p}\right)\right] + \tilde{\alpha}\, KL\left(q\,||\,p\right)$$

Coordinate Descent

Learn $\left(\Phi_j, \mu_j, r_j^2\right)_{j=1}^{2p}$

- Can simplify into a univariate problem on r_j^2
- Grid search to find the optimal values

Q4: *Other way than grid search? In relation to the feasibility in high dimensions.*

Prior $p(\beta)$

$$\left(\boldsymbol{\beta}_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(0, r_j^2\, I_{d_j})$$

Variational $q(\beta)$

$$\left(\beta_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(\mu_j, \Phi_j)$$

$$L(\mu, \Phi, r^2) = -E_q\left[\log p\left(y\,|\,\beta_1, \ldots, \beta_{2p}\right)\right] + \tilde{\alpha}\, KL\left(q\,||\,p\right)$$

Coordinate Descent

Learn $\left(\Phi_j, \mu_j, r_j^2\right)_{j=1}^{2p}$

- Can simplify into a univariate problem on r_j^2
- Grid search to find the optimal values

Q4: *Other way than grid search? In relation to the feasibility in high dimensions.*

Prior $p(\beta)$

$$\left(\boldsymbol{\beta}_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(0, r_j^2 I_{d_j})$$

Variational $q(\beta)$

$$\left(\beta_j\right)_{j=1}^{2p} \overset{ind}{\sim} N(\mu_j, \Phi_j)$$

Q5: *How practical is tuning α?*

$$L(\mu, \Phi, r^2) = -E_q\left[\log p\left(y \,|\, \beta_1, \ldots, \beta_{2p}\right)\right] + \tilde{\alpha}\, KL\left(q \,||\, p\right)$$

Note: $\alpha = \tilde{\alpha}\sigma^2$

Lastly

Experiments with synthetic and real-world datasets demonstrating

- effectiveness of VARD in feature selection and individual smoothing
- capacity to differentiate nonlinear, linear, and zero functions
- estimation accuracy
- competing performance to other methods

Lastly

Experiments with synthetic and real-world datasets demonstrating

- effectiveness of VARD in feature selection and individual smoothing
- capacity to differentiate nonlinear, linear, and zero functions
- estimation accuracy
- competing performance to other methods

Q6*:*
Where does VARD stand in comparison to deep learning methods? How can it compare?

Congratulations to the authors!

Thank you for your attention!