Designing Supercomputer Experiments

Trinity Mathematical Society

Trinity College, March 2008

Robert B. Gramacy

Statistical Laboratory, University of Cambridge

http://www.statslab.cam.ac.uk/~bobby

bobby@statslab.cam.ac.uk

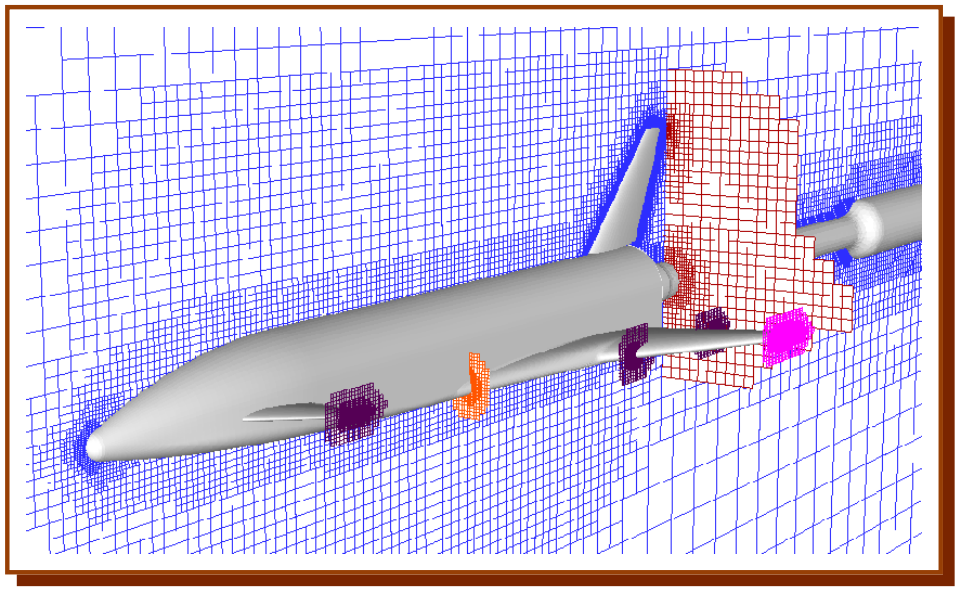
with Herbert K. H. Lee

Applied Math & Statistics Dept., UC Santa Cruz

Designing Computer Experiments

Real-world Application: CFD simulations of a proposed reusable NASA launch vehicle (Langley-Glide-Back Booster)

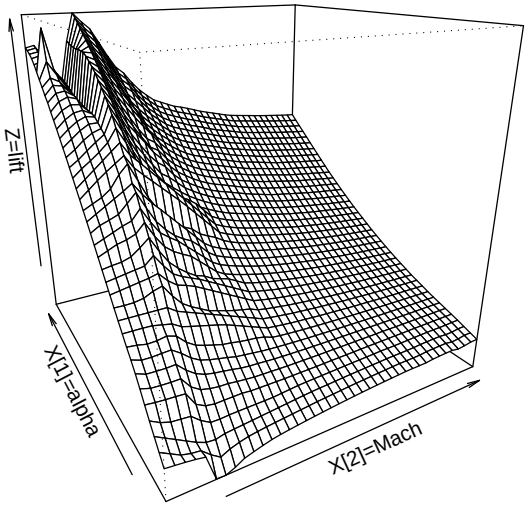
3 inputs:
side slip angle
Mach number
angle of attack

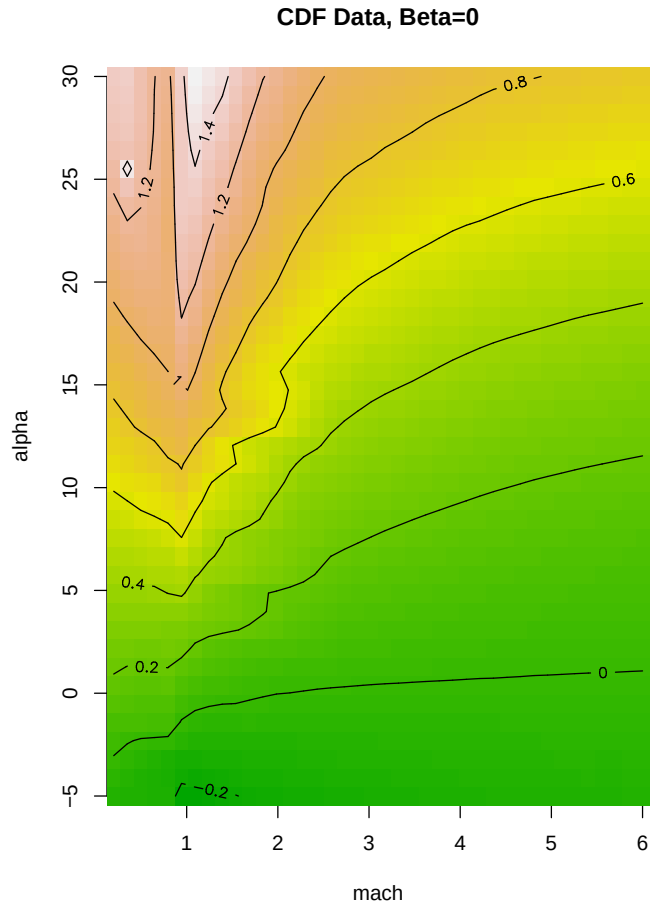


6 outputs:
lift
drag
pitching moment
side-force
yawing moment
rolling moment

- integrate inviscid Euler equations over mesh of 1.4×10^6 cells
- each input configuration takes 5-20 hours of CPU time

Motivation: example





- Exhibits typical characteristics of a computer experiment
  - not: stationary, linear, differentiable, continuous

Augmenting the State of the Art

Canonical model: Gaussian process (Santner, et al. 2003)

- conceptually simple non-parametric extension of LM
- flexible & non-linear

But:

- inference scales poorly with number of data points N
- strictly stationary

Augmenting the State of the Art

Canonical model: Gaussian process (Santner, et al. 2003)

- conceptually simple non-parametric extension of LM
- flexible & non-linear

But:

- inference scales poorly with number of data points N
- strictly stationary

Our idea: use trees to partition the input space like Bayesian treed LMs (Chipman et al., 2002) with GPs instead of LMs

- allows for modeling of non-stationary behavior
- predictive variance is location (region) dependent
- smaller covariance matrices ameliorate computational demands

Standard Linear Model (LM)

LM for n inputs (covariates) $\mathbf{x}$ of dimension p, and responses y

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + e_i, \qquad \text{where } e_i \sim N(0, \sigma^2), \qquad \text{for } i = 1, \ldots, n$$

- Collect $Y = (y_1, \ldots, y_n)^\top$, and $\mathbf{X} = [\mathbf{x}_1^\top \cdots \mathbf{x}_n^\top]$ and

  define the *likelihood*: $Y \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$

  Jeffrey's prior: $(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$

MLE: $\hat{\boldsymbol{\beta}} \sim (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y$, & $s^2 \sim \frac{1}{n-p} (Y - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (Y - \boldsymbol{X}\hat{\boldsymbol{\beta}})$

- *Bayes Rule* $p(\boldsymbol{\theta}|Y) \propto p(Y|\boldsymbol{\theta}) p(\boldsymbol{\theta})$ gives the posterior:

$$\boldsymbol{\beta}|\sigma^2, Y = N_p(\hat{\boldsymbol{\beta}}, (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2), \quad \& \quad \sigma^2|Y = \text{Inv-}\chi^2(n-p, s^2)$$

Gaussian Process (GP)

The Gaussian Process is *also* a LM (of sorts)

- Replacing $\mathbf{I}$ with $\mathbf{K}$

$$Y|\boldsymbol{\beta}, \sigma^2 \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad Z|\boldsymbol{\beta}, \sigma^2, \mathbf{K} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{K})$$

- $\mathbf{K}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$ is, e.g., an isotropic *correlation matrix*

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{d}\right\} + g\delta_{j,j}$$

- with *range* (or length–scale) parameter d
- and *nugget* (or measurement–error) parameter g
  - Range $d \to 0$ gives $\mathbf{K} \to (1+g)\mathbf{I}$; i.e., the LM

- Conditional on $\mathbf{K}$ through d and g, we have MLEs

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{K}^{-1} Z,$$

and

$$s^2 = \frac{1}{n-p} (Z - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top \mathbf{K}^{-1} (Z - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

- Numerical optimization required for d and g

Bayesian inference is similar

$$\boldsymbol{\beta} | \sigma^2, d, g, Z \sim N_p(\hat{\boldsymbol{\beta}}, (\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X})^{-1} \sigma^2)$$

$$\sigma^2 | d, g, Z \sim \text{Inv-}\chi^2(n-p, s^2)$$

- Metropolis–Hastings is used to sample the posterior $p(d, g|Z)$

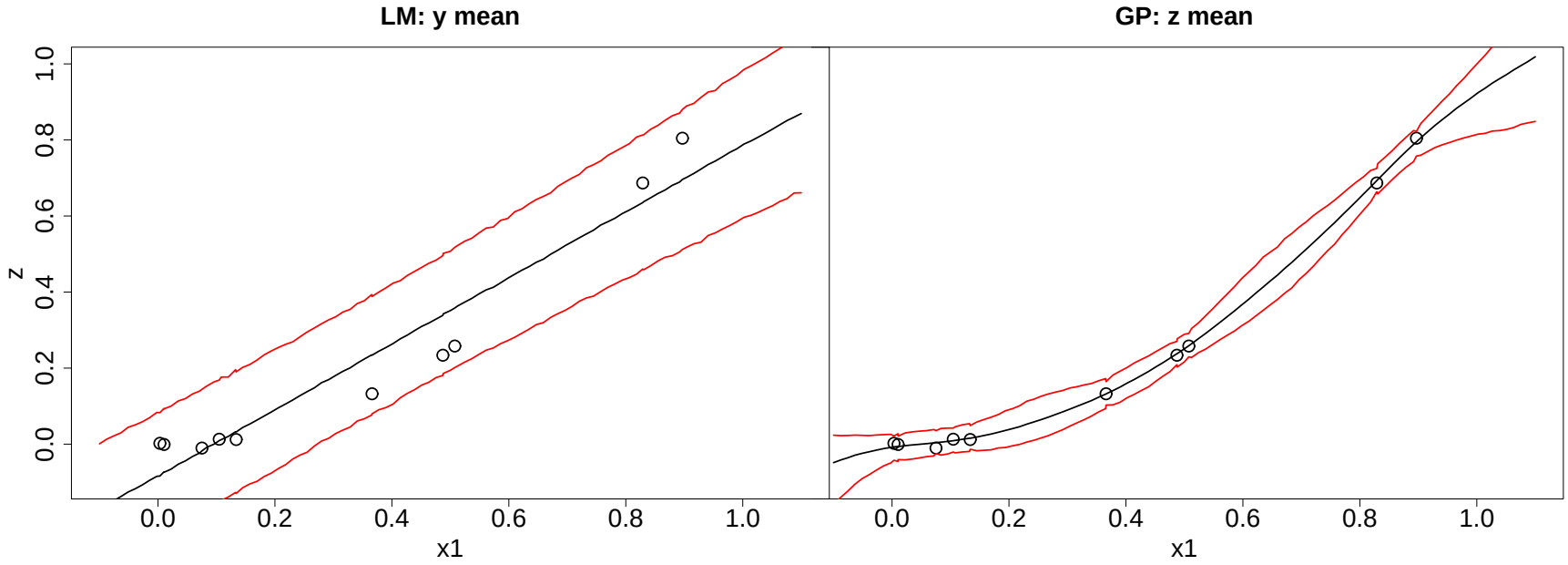
GP Prediction (or Kriging)

The predicted value of $z(\tilde{\mathbf{x}})$ is Normal with

$$\text{mean} \qquad \hat{y}(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}^\top \hat{\boldsymbol{\beta}} + \tilde{\mathbf{k}}(\tilde{\mathbf{x}})^\top \mathbf{K}^{-1}(Z - \mathbf{X}\hat{\boldsymbol{\beta}})$$

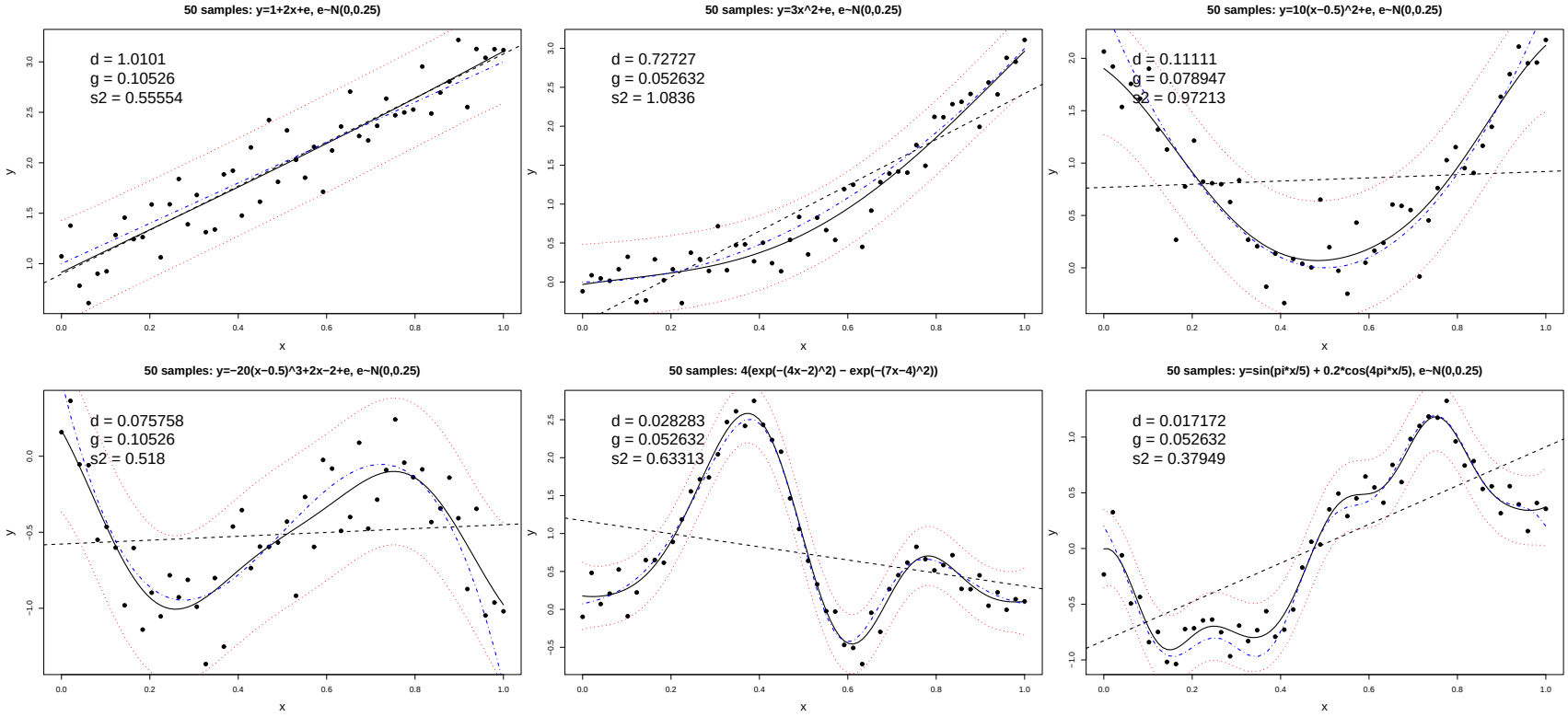
$$\text{variance} \qquad \hat{\sigma}(\tilde{\mathbf{x}})^2 = \sigma^2[\kappa(\tilde{\mathbf{x}}) - \mathbf{q}(\tilde{\mathbf{x}})^\top (\mathbf{K} + \mathbf{X}\mathbf{X}^\top)\mathbf{K}^{-1}\mathbf{q}(\tilde{\mathbf{x}})]$$

where $\kappa(\tilde{\mathbf{x}}) = K(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \tilde{\mathbf{x}}^\top \tilde{\mathbf{x}}$, $\mathbf{q}(\tilde{\mathbf{x}}) = \mathbf{k}(\tilde{\mathbf{x}}) + \mathbf{X}\tilde{\mathbf{x}}$, $\mathbf{k}_j(\tilde{\mathbf{x}}) = K(\tilde{\mathbf{x}}, \mathbf{x}_j)$, and $\mathbf{x}_j^\top$ is the j^{th} column of $\mathbf{X}$



The GP range parameter

MLE range (d) for various data:

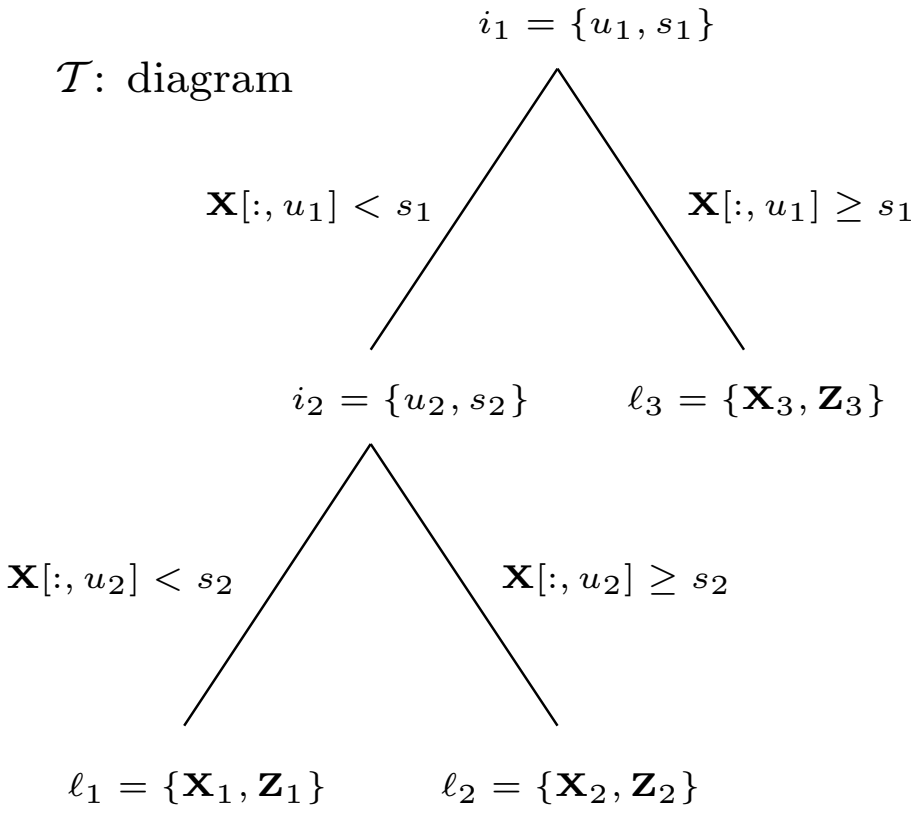


- However, recall that $d \to 0$ gives $\mathbf{K} \to (1+g)\mathbf{I}$

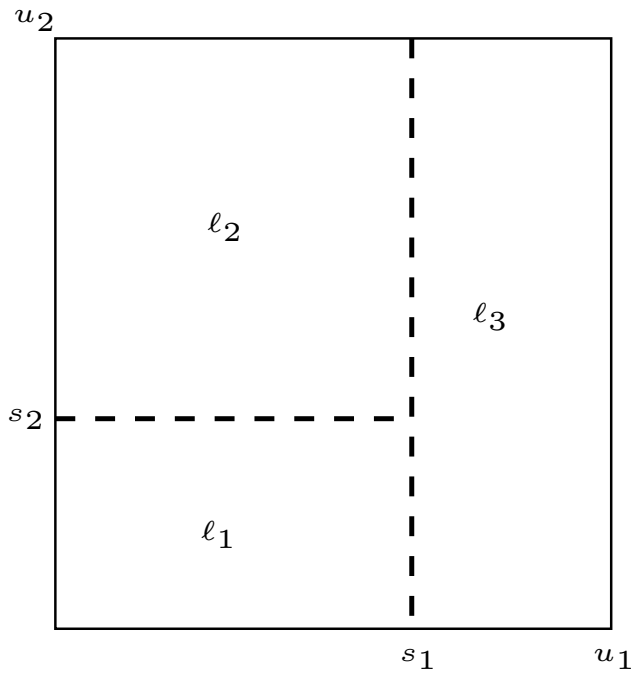
Tree Example

How a tree $\mathcal{T}$ recursively partitions the input space:

$\mathcal{T}$: diagram



$\mathcal{T}$: graphically



- prior for node η at depth $q_\eta \in \mathcal{T}$: $p_{\text{SPLIT}}(\eta, \mathcal{T}) = a(1 + q_\eta)^{-b}$

Hierarchical Model

Conditioning on tree $\mathcal{T}$ we have R regions: $\{r_\nu\}_{\nu=1}^R$

$$\mathbf{Z}_\nu|\boldsymbol{\beta}_\nu, \sigma^2_\nu, \mathbf{K}_\nu \sim N(\mathbf{F}_\nu\boldsymbol{\beta}_\nu, \sigma^2_\nu\mathbf{K}_\nu) \qquad \sigma^2_\nu \sim IG(\alpha_\sigma/2, q_\sigma/2)$$

$$\boldsymbol{\beta}_\nu|\sigma^2_\nu, \tau^2_\nu, \mathbf{W}, \boldsymbol{\beta}_0 \sim N(\boldsymbol{\beta}_0, \sigma^2_\nu\tau^2_\nu\mathbf{W}) \qquad \tau^2_\nu \sim IG(\alpha_\tau/2, q_\tau/2)$$

$$\boldsymbol{\beta}_0 \sim N(\boldsymbol{\mu}, \mathbf{B}) \qquad \mathbf{W}^{-1} \sim W((\rho\mathbf{V})^{-1}, \rho)$$

with $\mathbf{F}_\nu = (\mathbf{1}, \mathbf{X}_\nu)$, and

$$K_\nu(\mathbf{x}_j, \mathbf{x}_k) = K^*_\nu(\mathbf{x}_j, \mathbf{x}_k) + g_\nu\delta_{j,k}$$

for nugget g and *true* correlation K^* is separable:

$$K^*_\nu(\mathbf{x}_j, \mathbf{x}_k|\mathbf{d}_\nu) = \exp\{-\textstyle\sum_{i=1}^{m_X} |x_{ij} - x_{ik}|^{p_0}/d_{i\nu}\}$$

Tree MCMC

Conditional on a particular tree $\mathcal{T}$

- Gibbs samples for all GP parameters $\{\boldsymbol{\theta}_\nu\}_{\nu=1}^R|\mathcal{T}$
  - except $K(\cdot,\cdot|\mathbf{d}_\nu, g_\nu)$ requires MH

Sample from the joint posterior of $(\mathcal{T}, \boldsymbol{\theta})$

(Richardson & Green, 1997; Chipman et al., 2002)

- Average over $\mathcal{T}$ with reversible-jump MCMC (RJ-MCMC)
  - Tree operations: *grow*, *prune*, *change*, *swap*

Bayesian Tree *grow & prune*

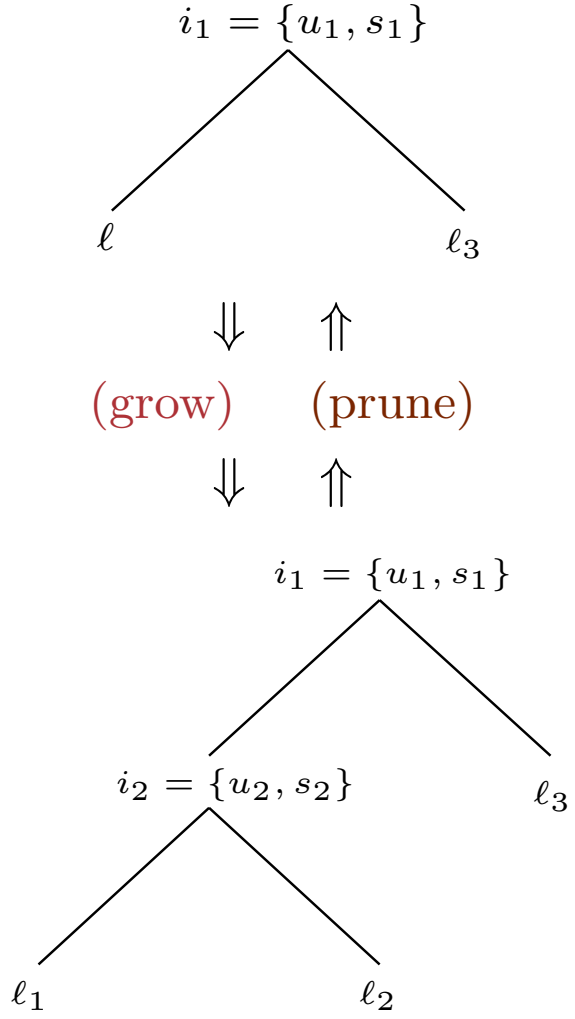
There is always a *leaf* to *grow* $\mathcal{T}_t$:

- randomly select a *leaf* node $\eta \in \ell(\mathcal{T}_t)$
- randomly select a splitting rule (u, s)
- create $\mathcal{T}^*$ from $\mathcal{T}_t$
  - split η in dimension u at location s
  - $\eta \in \ell(\mathcal{T}_t)$ becomes $(\eta_1, \eta_2) \in \ell(\mathcal{T}^*)$
- accept or reject the proposed $\mathcal{T}^*$
  - via the MH acceptance ratio:

$$\alpha = \min\left\{1, \frac{p(\mathcal{T}^*|Y)q(\mathcal{T}|\mathcal{T}^*)}{p(\mathcal{T}_t|Y)q(\mathcal{T}^*|\mathcal{T})}\right\}$$

  - set $\mathcal{T}_{t+1} = \mathcal{T}^*$ w.p. α; or $\mathcal{T}_{t+1} = \mathcal{T}_t$

The opposite of *grow* is *prune* $\mathcal{T}_t$:



Bayesian Tree *change & swap*

Splits are moved with *change* operations

- randomly select *internal* node $\eta \in i(\mathcal{T})$
  - η has split point (u, s), say
- propose a new $\eta^* \in \mathcal{T}^*$
  - by modifying the split location (u, s^*)

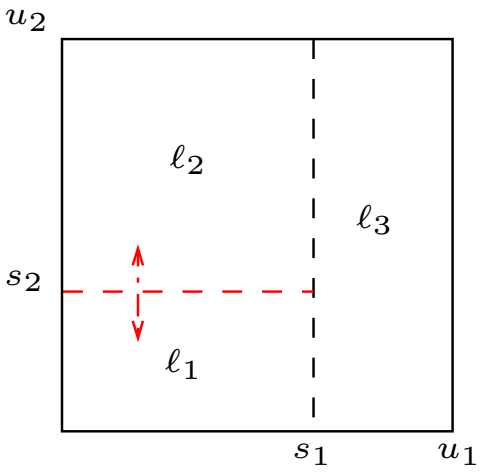
Or, switch the order of splits with *swap*

- randomly select a pair of *internal* nodes
  - $\eta_1, \eta_2 \in i(\mathcal{T})$ so that $\mathcal{P}(\eta_2) = \eta_1$
  - swap their order, thus proposing $\mathcal{T}^*$

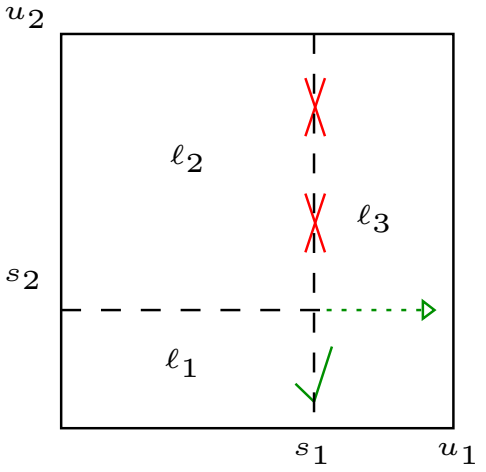
Accept or reject $\mathcal{T}^*$ with probability

$$\alpha = \min\left\{1, \frac{p(\mathcal{T}^*|Y)}{p(\mathcal{T}_t|Y)}\right\}$$

(change)



(swap)



Prediction (or Kriging)

The predicted value of $Y(\mathbf{x})$ at $\mathbf{x} \in r_\nu$ is Normal with

$$\text{mean} \qquad \hat{y}(\mathbf{x}) = \mathbf{f}^\top(\mathbf{x})\tilde{\boldsymbol{\beta}}_\nu + \mathbf{k}_\nu(\mathbf{x})^\top \mathbf{K}_\nu^{-1}(\mathbf{Z}_\nu - \mathbf{F}_\nu\tilde{\boldsymbol{\beta}}_\nu)$$

$$\text{variance} \qquad \hat{\sigma}(\mathbf{x})^2 = \sigma_\nu^2[\kappa(\mathbf{x},\mathbf{x}) - \mathbf{q}_\nu^\top(\mathbf{x})\mathbf{C}_\nu^{-1}\mathbf{q}_\nu(\mathbf{x})]$$

where $\mathbf{k}_\nu(\mathbf{x})$ is a $n_\nu-$vector with $\mathbf{k}_{\nu,j}(\mathbf{x}) = K_\nu(\mathbf{x},\mathbf{x}_j)$, $\mathbf{x}_j \in \mathbf{X}_\nu$

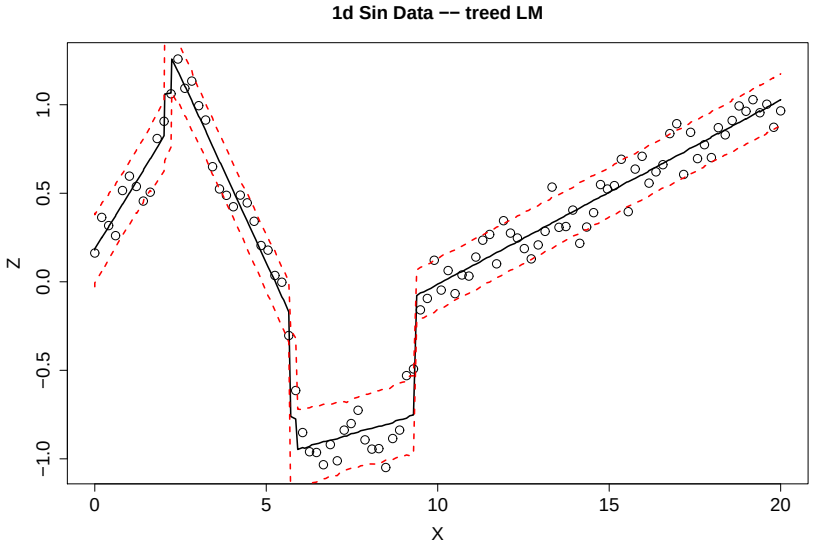
$$\mathbf{C}_\nu^{-1} = (\mathbf{K}_\nu + \mathbf{F}_\nu\mathbf{W}\mathbf{F}_\nu^\top/\tau_\nu^2)^{-1} \qquad \mathbf{q}_\nu(\mathbf{x}) = \mathbf{k}_\nu(\mathbf{x}) + \tau_\nu^2\mathbf{F}_\nu\mathbf{W}_\nu\mathbf{f}(\mathbf{x})$$

$$\kappa(\mathbf{x},\mathbf{y}) = K_\nu(\mathbf{x},\mathbf{y}) + \tau_\nu^2\mathbf{f}^\top(\mathbf{x})\mathbf{W}\mathbf{f}(\mathbf{y}) \qquad \mathbf{f}^\top(\mathbf{x}) = (1,\mathbf{x}^\top)$$

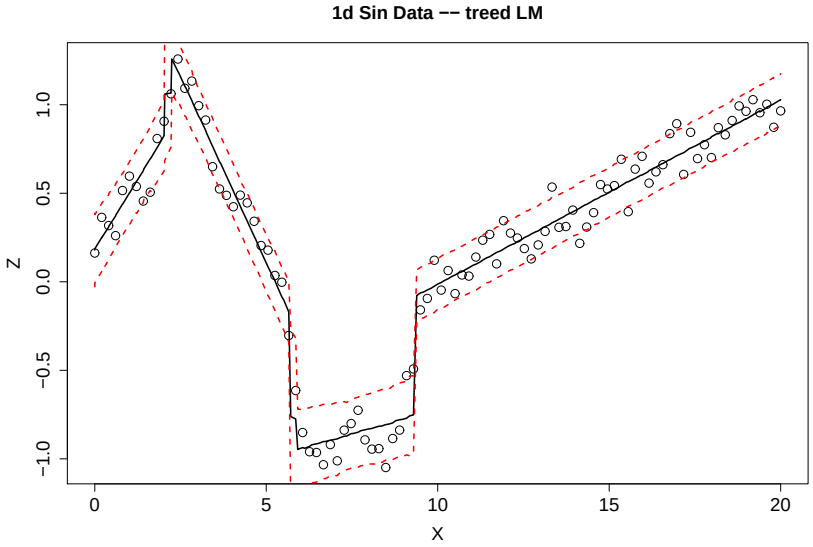
Expected reduction in squared error

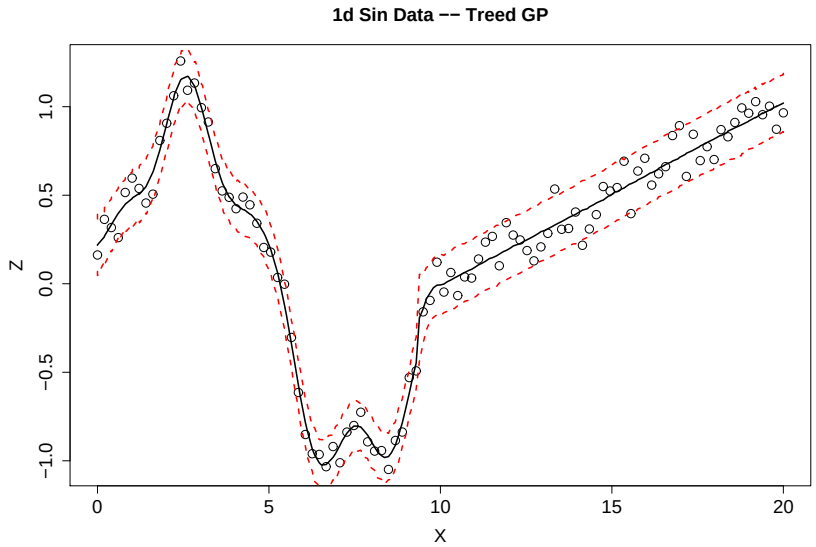
$$\Delta\hat{\sigma}_{\mathbf{y}}^2(\tilde{\mathbf{x}}) = \hat{\sigma}_{\mathbf{y}}^2 - \hat{\sigma}_{\mathbf{y}}^2(\tilde{\mathbf{x}}) = \frac{\left[\mathbf{q}_N^\top(\mathbf{y})\mathbf{C}_N^{-1}\mathbf{q}_N(\tilde{\mathbf{x}}) - \kappa(\tilde{\mathbf{x}},\mathbf{y})\right]^2}{\kappa(\tilde{\mathbf{x}},\tilde{\mathbf{x}}) - \mathbf{q}_N^\top(\tilde{\mathbf{x}})\mathbf{C}_N^{-1}\mathbf{q}_N(\tilde{\mathbf{x}})}$$

Treed GP on sine Data

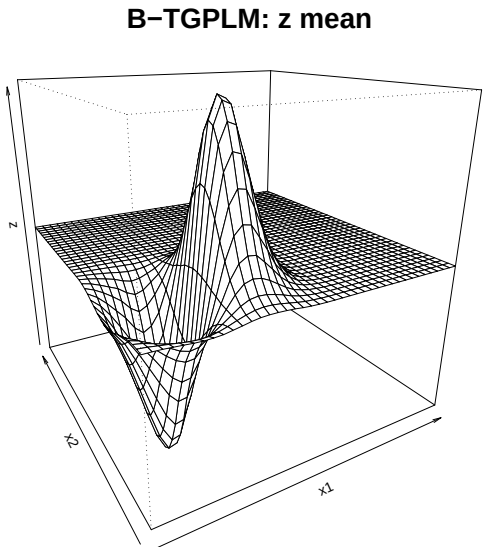


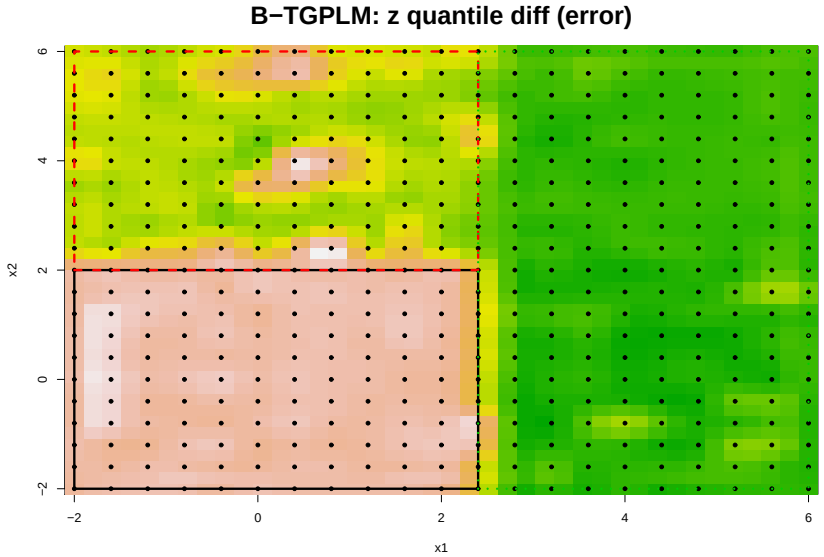
Treed GP on sine Data

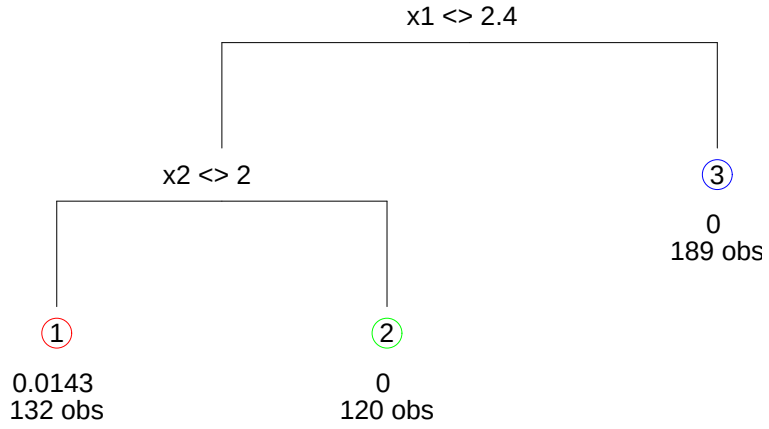




Treed GP on exp Example



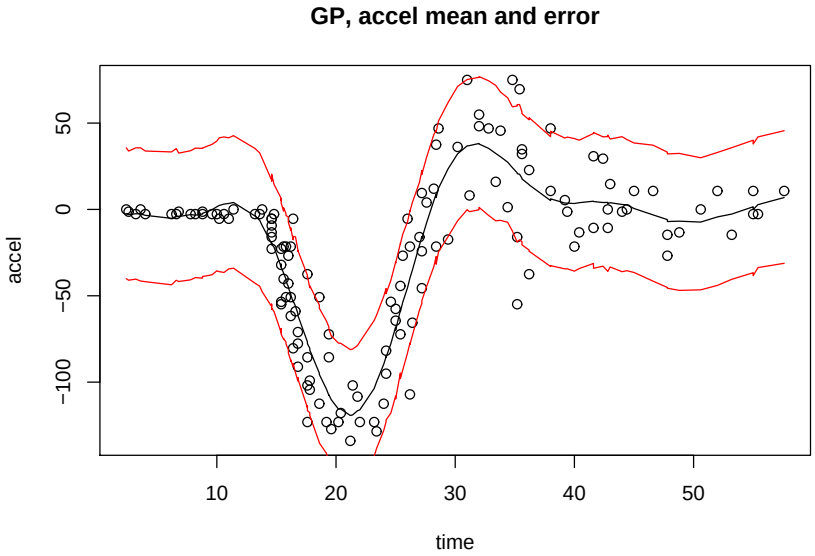




Motorcycle data:

(Silverman, 1985)

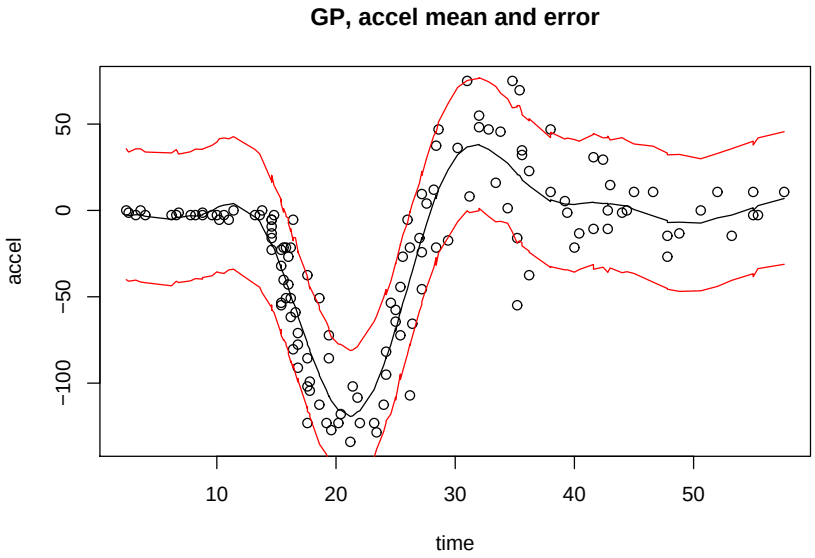
- non-stationary
- input-dependent noise

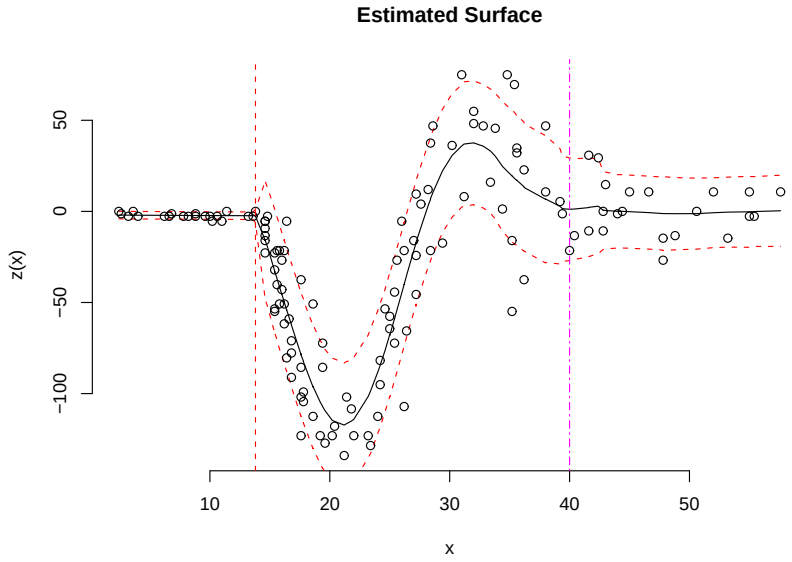


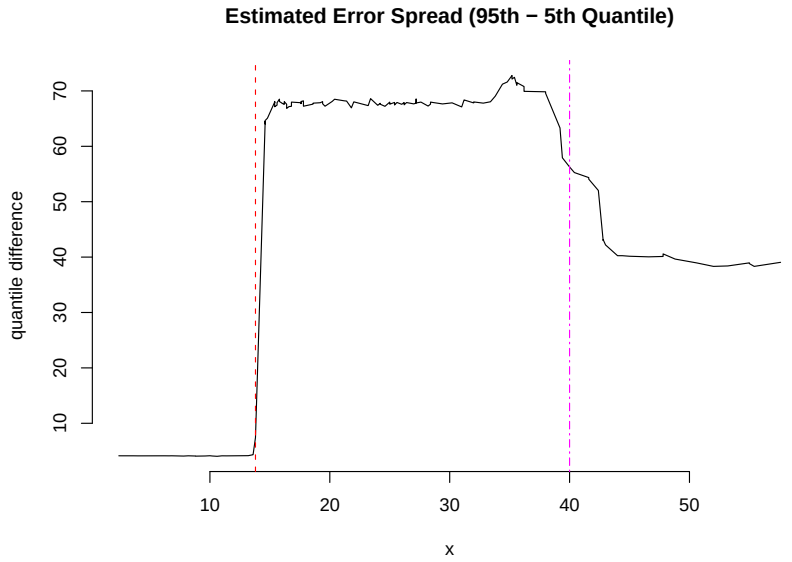
Motorcycle data

Motorcycle data:
(Silverman, 1985)

- non-stationary
- input-dependent noise







Bayesian adaptive sampling proceeds in trials ...

Current trial:

- Choose candidates: $\tilde{\mathbf{X}}$
- estimate model parameters $(\boldsymbol{\theta}, \mathcal{T})$ for data $\{\mathbf{X}_i, \mathbf{z}_i\}_{i=1}^N$
- Order candidates $\tilde{\mathbf{X}}$ by
  - ALM: maximize predictive error $\hat{\sigma}^2(\tilde{\mathbf{x}})$
  - ALC: maximize average expected reduction in predictive error

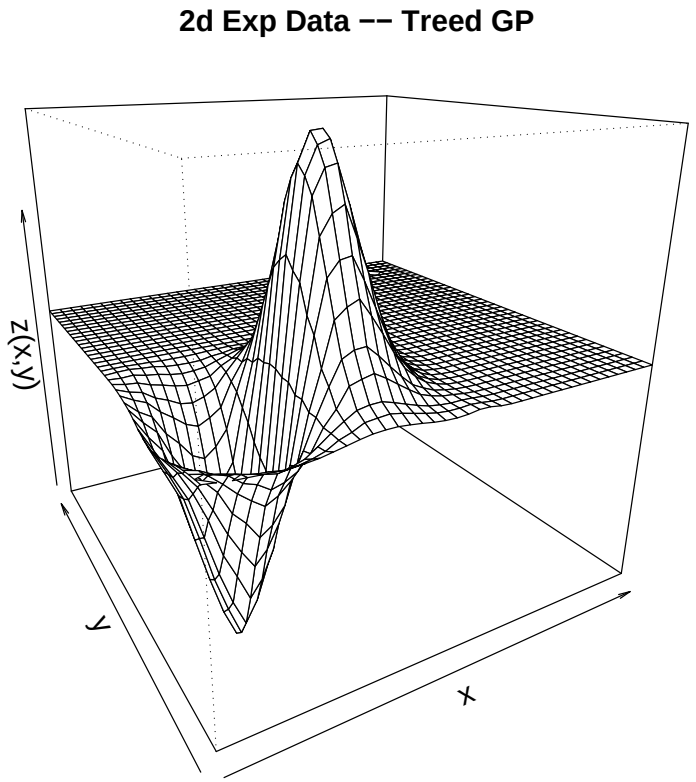
$$\Delta\hat{\sigma}^2(\tilde{\mathbf{x}}) = \int_D \Delta\hat{\sigma}^2_{\mathbf{y}}(\tilde{\mathbf{x}})\, d\mathbf{y}$$

- (repeat) begin next trial

Adaptive Sampling Comparison

2d exponential data:

$$y(\mathbf{x}) = x_1 \exp(-x_1^2 - x_2^2)$$



75 adaptive samples

model	as	rmse
btgp	alc	0.0035
btgp	alm	0.0037
bcart	alm	0.0090
bcart	alc	0.0093
btlm	alm	0.0099
btlm	alc	0.0115
bgp	alm	0.0352
bgp	alc	0.0493

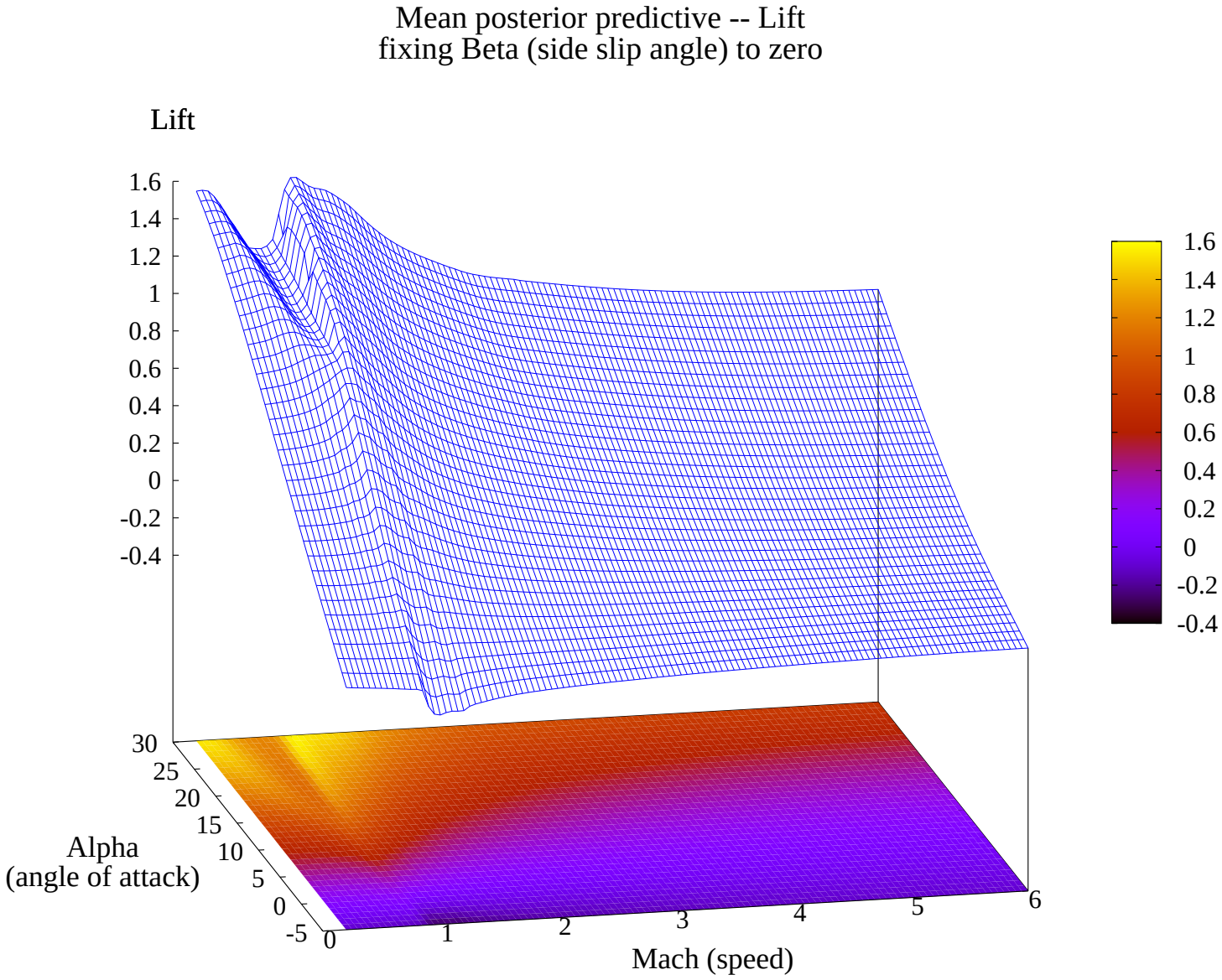
3-D CFD DATA

Adaptive sampling on the Langley-Glide-Back Booster (LGBB):

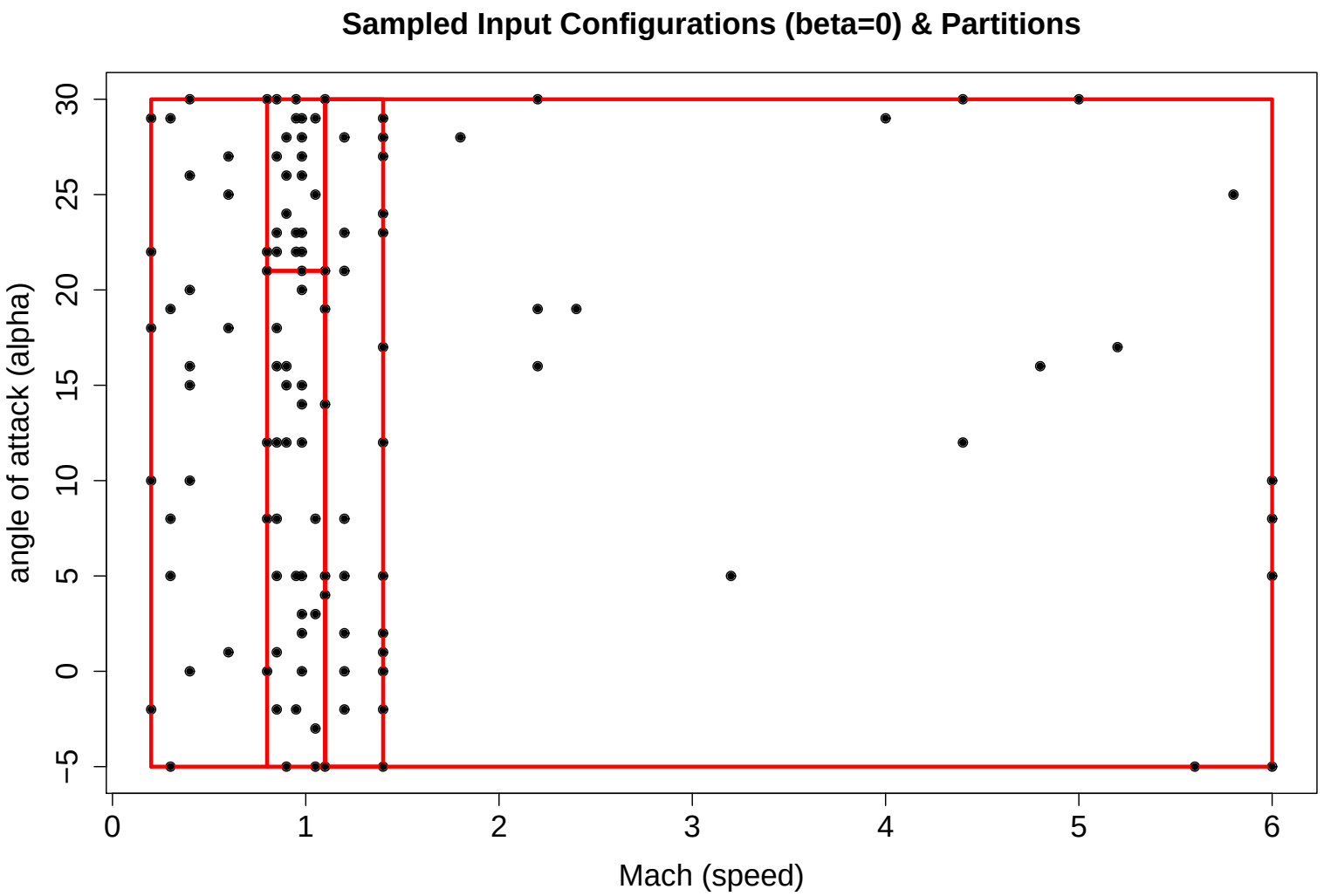
Interfacing with a NASA supercomputer:

- fit six independent treed GPs– one for each response (lift, drag, pitch, side-force, yaw, roll)
  - Six parallel treed GP modules
- design via expected reduction in predictive error (ALC)
  - pooled predictive quantiles across the six models

Adaptive Sampling on LGBB: Lift

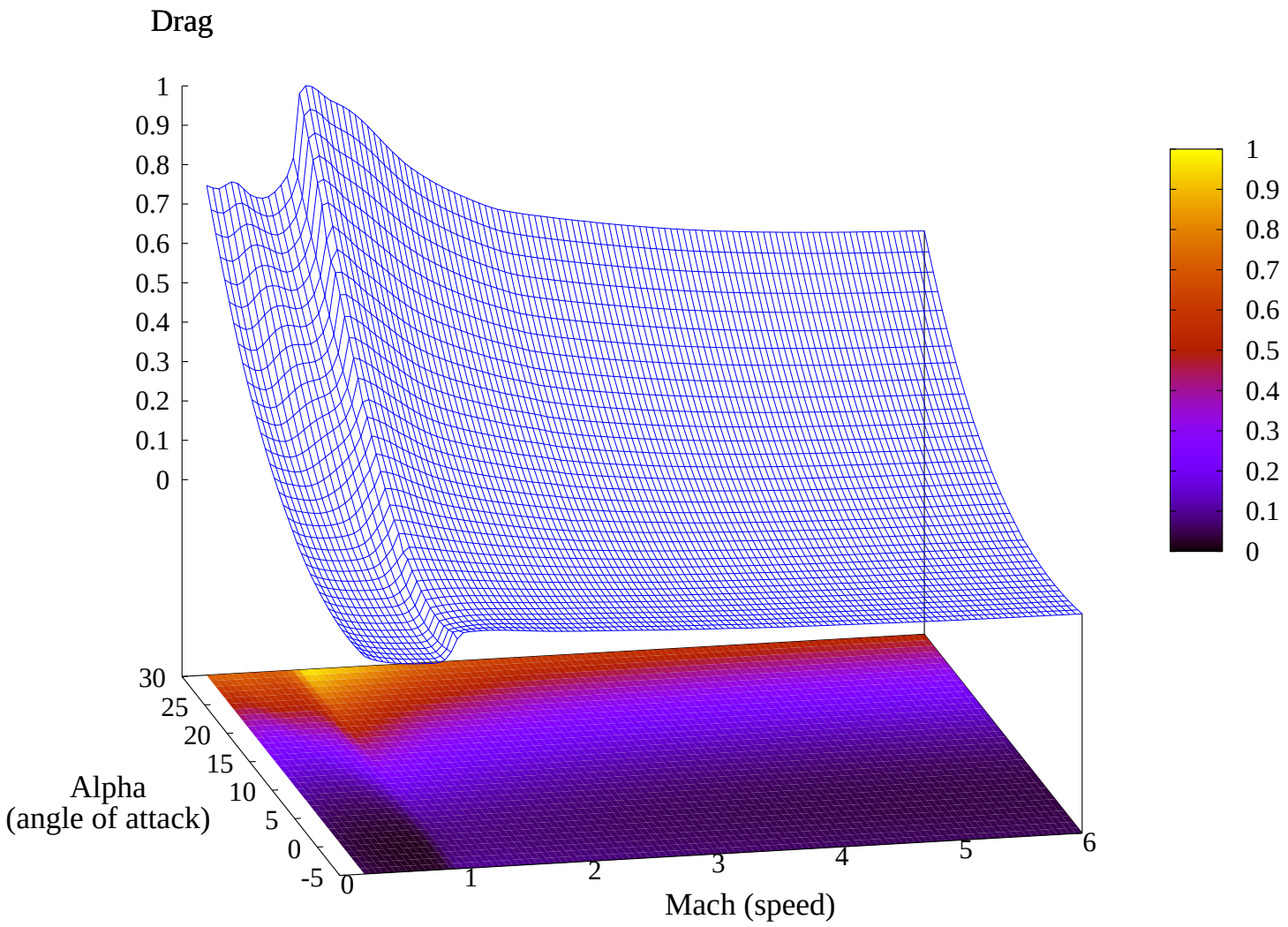


Adaptive Sampling on LGBB: Lift

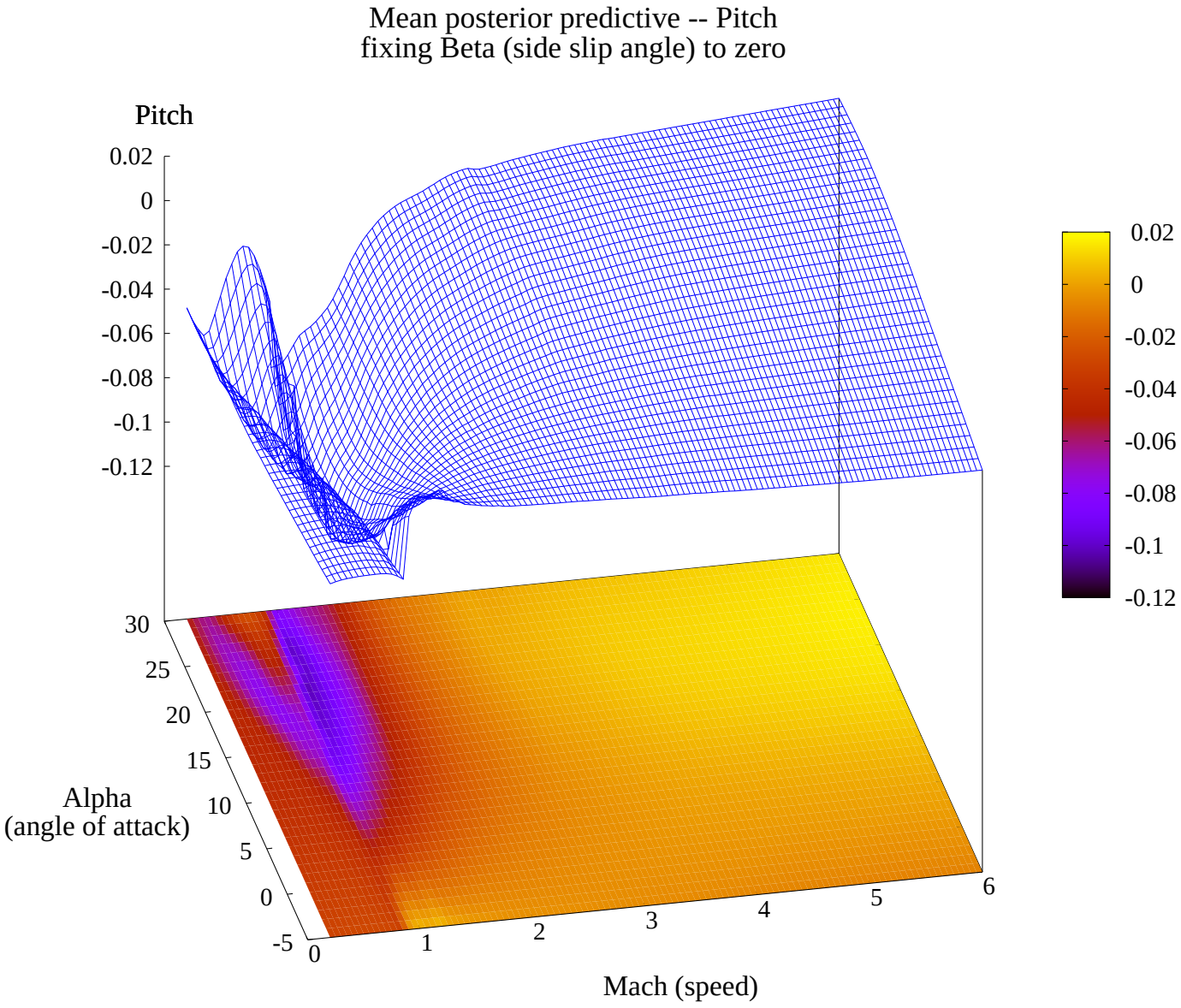


Adaptive Sampling on LGBB: Drag

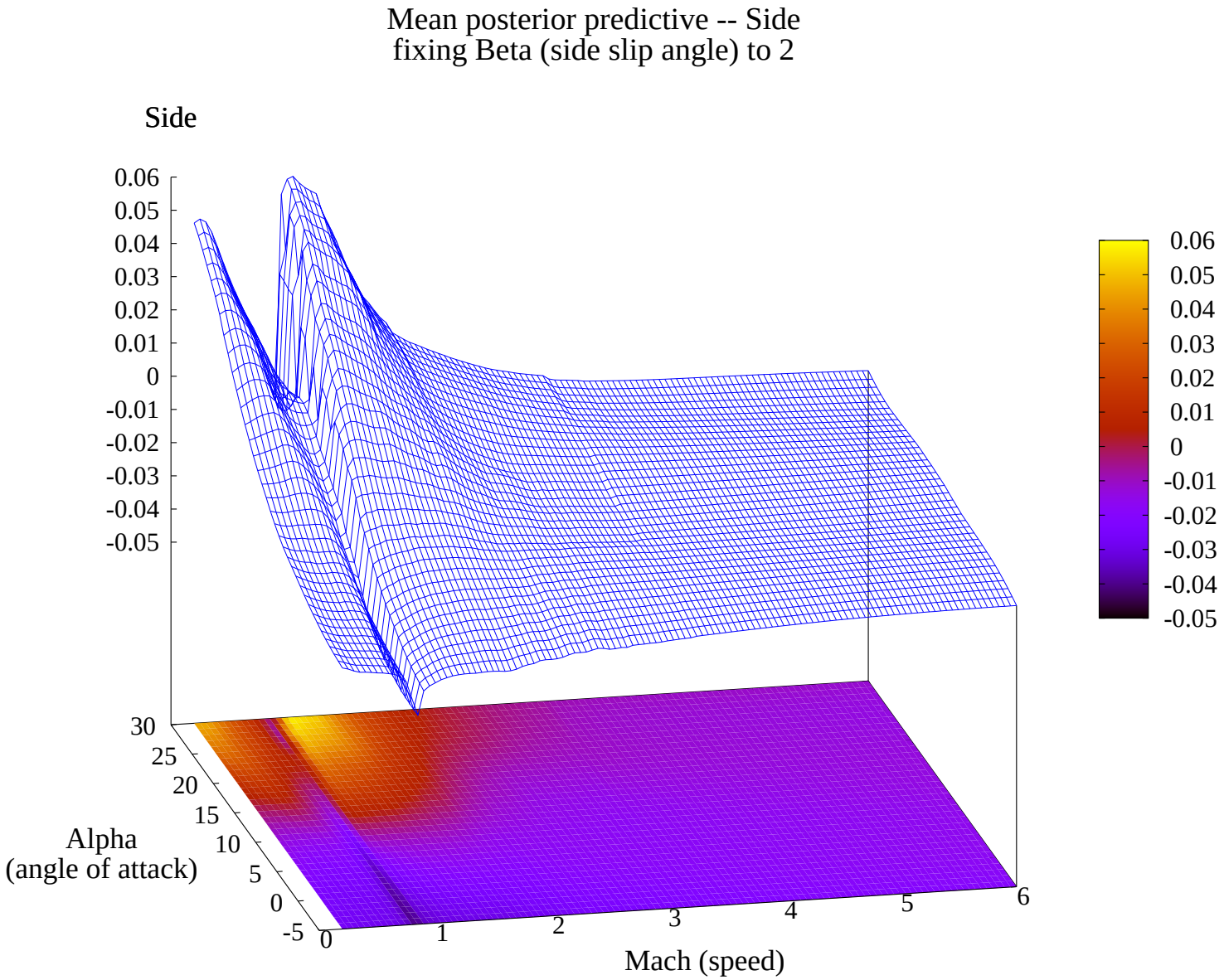
Mean posterior predictive -- Drag
fixing Beta (side slip angle) to zero



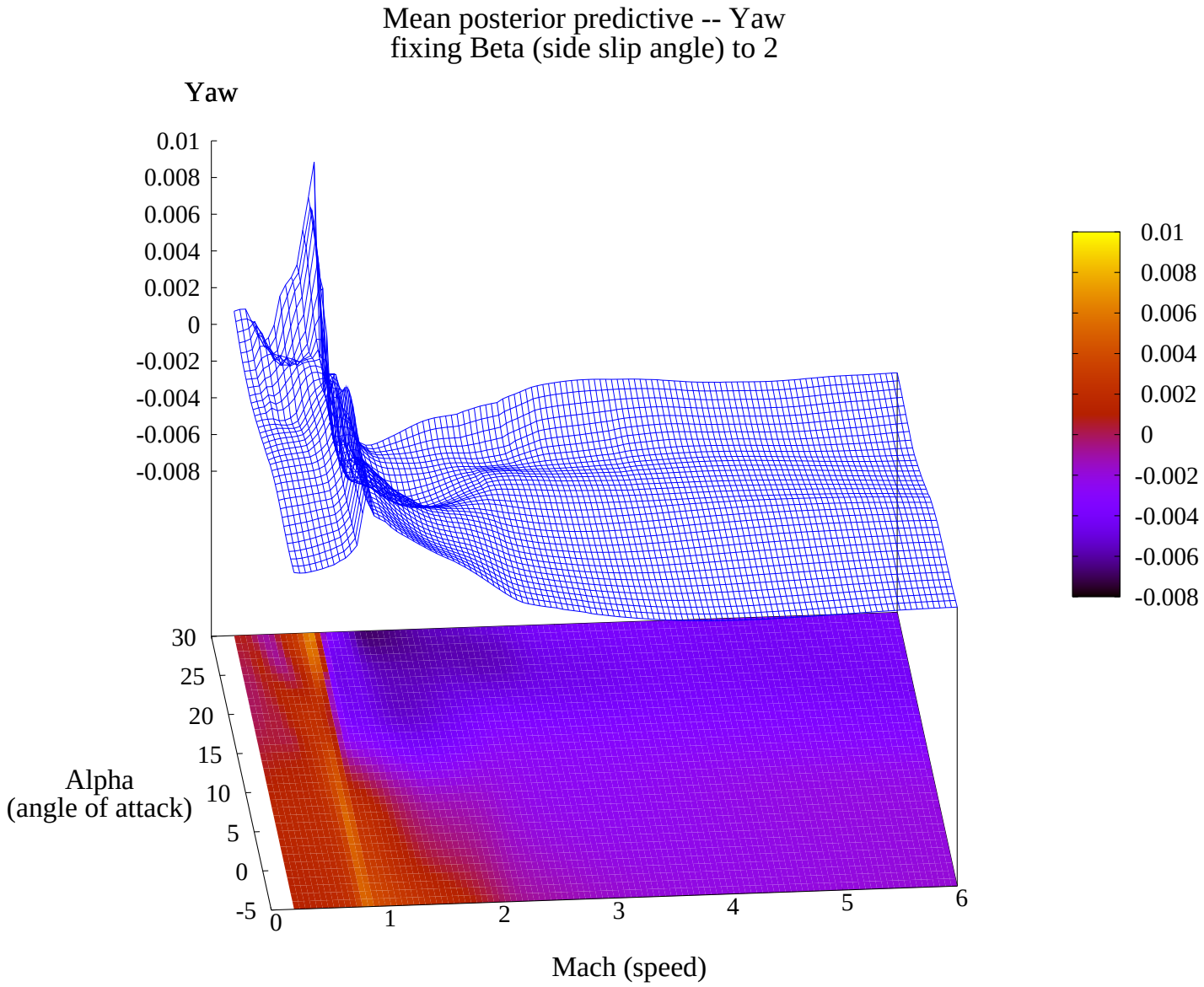
Adaptive Sampling on LGBB: Pitch



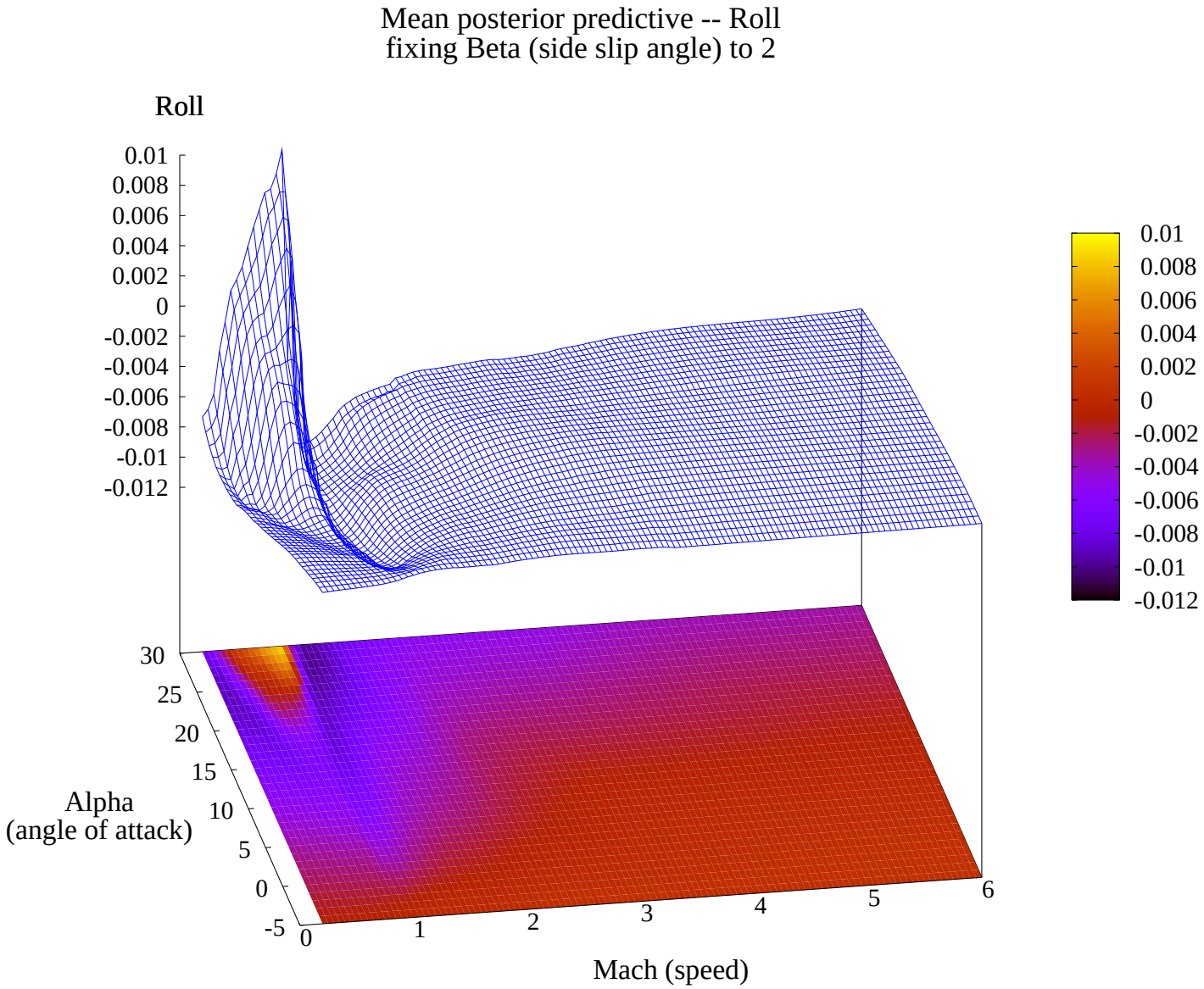
Adaptive Sampling on LGBB: Side



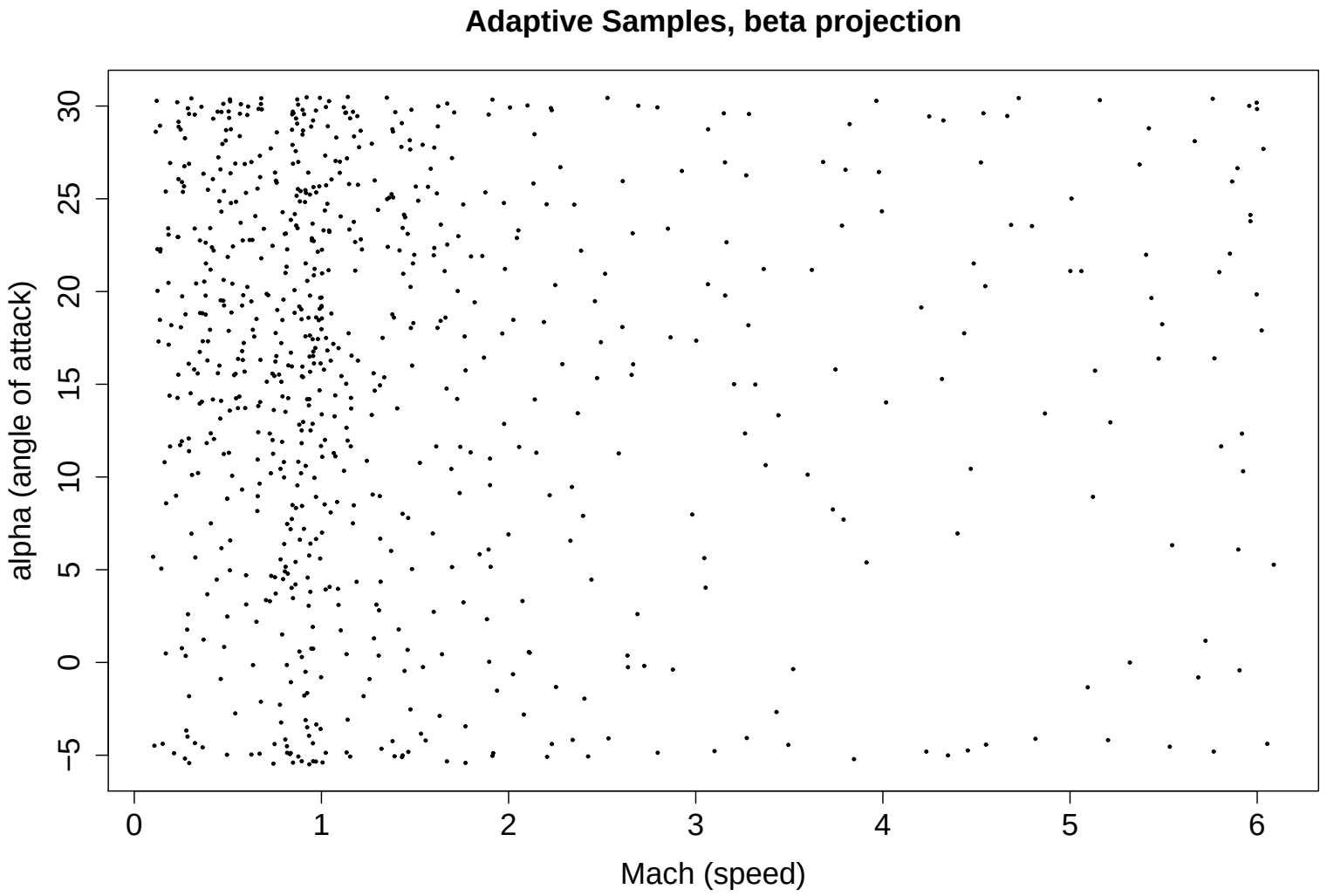
Adaptive Sampling on LGBB: Yaw



Adaptive Sampling on LGBB: Roll



Adaptive Sampling on LGBB



780 adaptive samples, compared to more than 3,250

Implementation & R package

Implementation & computing details

- `C++` (trees) and `C` (GPs) with `LAPACK/BLAS`
- `Pthreads` for (shared memory) parallelization:

R package called `tgp` on CRAN

- Implements all model combinations and special cases
  - blm, bgp, btlm, btgp, bgpllm, btgpllm
- Adaptive sampling: ALM, ALC, etc.