Bayesian Model Averaging Kriging

Jize Zhang and Alexandros Taflanidis
Why use metamodeling techniques?

- High-fidelity **numerical simulators** are increasingly used in a variety of engineering applications
- Since they can be computationally expensive →

- Metamodeling techniques are adopted in many implementations (for example in UQ setting) to approximate the simulator with a **cheaper-to-run emulator**
- Popular choices include:
  - Support vector machine
  - Artificial neural networks
  - Polynomial Chaos
  - **Kriging (aka Gaussian Process Regression); Data-driven statistical emulator**
Kriging (Gaussian process) intro I

\( h(x) = f(x)^T \beta + z(x) \)

- The regression structure captures deterministic global trend, through basis functions:
  \[ f = [f_1, \cdots, f_{nf}] \]

- The residual GP dictates local deviation from trend with some chosen covariance kernel:
  \[ \text{cov}(z(x), z(x')) = \sigma^2 R\left(\frac{x - x'}{\theta}\right) \]

Output = Trend + GP residual

Select basis functions \( f \) and correlation kernel \( R \)

Tune Kriging parameters based on experiments (input X/output H pair)

i.e., optimize \((\beta^*, \sigma^2, \theta^*)\)
Kriging (Gaussian process) intro I

\[ h(x) = f(x)^T \beta + z(x) \]

Output = Trend + GP residual

Select basis functions \( f \) and correlation kernel \( R \)

Tune Kriging parameters based on experiments (input \( X \)/output \( H \) pair)

Formulate Kriging predictions (predictive mean and variance) for any new input \( x \)
Kriging (Gaussian process) intro II

Select basis functions \( f \) and correlation kernel \( R \)

Tune Kriging parameters based on experiments (input \( X \)/output \( H \) pair)

Formulate Kriging predictions (predictive mean and variance) for any new input \( x \)

Output prediction \( \sim p(h \mid x, X, H) \sim \mathcal{N}(\bar{h}(x), \sigma^2(x)) \)

\[
\bar{h}(x) = f(x)^T \beta_* + r_*(x)^T R_*^{-1} (H - F \beta_*)
\]

\[
\sigma^2(x) = \sigma_*^2 \left[ 1 + u^T (F^T R_*^{-1} F)^{-1} u - r_*(x)^T R_*^{-1} r_*(x) \right]
\]

where \( u = F^T R_*^{-1} r_*(x) - f(x) \)
Is regression structure important? I

Real model

Experiments

Kriging with \( f(x) = \{1\} \)

Kriging with \( f(x) = \{x_1 x_2, (x_2)^2\} \)

Kriging with \( f(x) = \{x_2^3\} \)

Kriging with \( f(x) = \{x_2^4\} \)
Is regression structure important? II

Contour plots of absolute errors (pinker indicates higher error)

Kriging with $f(x) = \{1\}$

Kriging with $f(x) = \{x_1, x_2, x_2^2\}$

Kriging with $f(x) = \{x_2^3\}$

Kriging with $f(x) = \{x_2^4\}$
Bayesian inference for regression structure selection

Question: how can we select a regression structure, conditioned on all available data? → Bayesian inference

Terminology: Model $M_k$ for Kriging with basis function vector $\mathbf{f}_k = \{f_1^k, f_2^k, \ldots\}$

Bayes rule:

$$P(M_k \mid \mathbf{X}, \mathbf{H}) \propto p(\mathbf{H} \mid \mathbf{X}, M_k) p(M_k)$$

Model evidence

Posterior probability of each model

Prior on each model
Bayesian model averaging prediction

Repeat for every model

Predict under model $M_k$

$$p(h | x, X, H, M_k)$$

$$p(h | x, X, H, \bigcup M_k) = \sum_{k=1}^{n_m} p(h | x, X, H, M_k) P(M_k | X, H)$$

$$P(M_k | X, H) \propto p(H | X, M_k) p(M_k)$$

Mixture weights = Posterior probability
Evidence evaluation

Evidence: \( P(H \mid X, M_k) \)

\[ = \int \int \int P(H \mid X, \beta_k, \sigma_k^2, \theta_k, M_k) \pi(\beta_k, \sigma_k^2, \theta_k \mid M_k) d\beta_k d\sigma_k^2 d\theta_k \]

Data likelihood

Model parameter prior

• For each model, its evidence calculation requires a probabilistic integration over the model’s parameter space (Kriging tunable parameters), which can be numerically cumbersome (needs sophisticated MCMC samplers) …
Evidence evaluation

Evidence: \( P(H \mid X, M_k) \)

\[
= \iiint P(H \mid X, \beta_k, \sigma_k^2, \theta_k, M_k) \pi(\beta_k, \sigma_k^2, \theta_k \mid M_k) d\beta_k d\sigma_k^2 d\theta_k
\]

- Data likelihood has a determined form, because of Kriging’s GP nature
- Model parameter prior
- The prior can be selected to result in analytical expression of the integral, eliminating numerical integration challenges
Prior elicitation $I$ [$\beta$]

\[
[\beta_k | \sigma^2, \theta] \sim \mathcal{N}(0, \max(n, n_k^2)\sigma^2(F_k^T F_k)^{-1})
\]

- This prior is similar in spirit from the extremely popular-used *benchmark g-prior* in linear regression context.
- The red component is adaptive to both data size $n$ and basis function size $n_k$.
- The blue component corresponds to the *Fisher’s information matrix* of the basis function values of the training data ($F_k$).
Prior elicitation II \([\sigma^2 \text{ and } \theta]\)

Prior selection

\[
[\beta_k \mid \sigma^2, \theta] \sim \mathcal{N}(0, \text{max}(n, n_k^2)\sigma^2(F_k^T F_k)^{-1})
\]

\[
[(\sigma_k^*)^2, \theta_k^*] = \arg \max_{\sigma_k^2, \theta_k} P(H \mid X, M_k, \sigma_k^2, \theta_k)
\]

*Empirical Bayesian* (EB) approach: the maximizer of the conditional evidence chosen as prior; leads to closed-form expression for evidence:

\[
\log P(H \mid X, M_k) = \log P(H \mid X, M_k, \sigma_k^2, \theta_k)
\]

\[
= -\frac{1}{2\sigma_k^2} H^T R_{\theta_k}^{-1} H + \frac{1}{2} H^T C_k H - \frac{1}{2} \log |\sigma_k^2 R_{\theta_k}| + \frac{1}{2} \log c_k \sigma_k^2 (F_k^T F_k)^{-1} + \frac{1}{2} \log |A_k| - \frac{n}{2} \log 2\pi
\]

Illustrative contour plots of conditional likelihood
Empirical Bayesian (EB) approach: the maximizer of the conditional evidence chosen as prior; leads to closed-form expression for evidence:

$$[\beta_k | \sigma^2, \theta] \sim \mathcal{N} \left( 0, \max(n, n_k^2) \sigma^2 (F_k^T F_k)^{-1} \right)$$

$$[(\sigma^*_k)^2, \theta^*_k] = \arg \max_{\sigma^2_k, \theta_k} P(H | X, M_k, \sigma^2_k, \theta_k)$$

Prior selection

Illustrative contour plots of conditional likelihood

But

Maximization is non-trivial (non-convex optimization problem)
Retain Gaussian prediction characteristics

\[ [\beta_k \mid \sigma_k^2, \theta] \sim \mathcal{N}\left(0, \max(n, n_k^2)\sigma_k^2(F_k^T F_k)^{-1}\right) \]

\[ ((\sigma_k^*)^2, \theta_k^*) = \arg\max_{\sigma_k^2, \theta_k} P(H \mid X, M_k, \sigma_k^2, \theta_k) \]

Not only evidence is analytically tractable

\[
\log P(H \mid X, M_k) = \log P(H \mid X, M_k, \sigma_k^2, \theta_k) \\
= -\frac{1}{2\sigma_k^2} H^T R_{\theta_k}^{-1} H + \frac{1}{2} H^T C_k H - \frac{1}{2} \log |\sigma_k^2 R_{\theta_k}| - \frac{1}{2} \log |c_k \sigma_k^2 (F_k^T F_k)^{-1}| - \frac{1}{2} \log |A_k| - \frac{n}{2} \log 2\pi
\]

But also predictions retain Gaussian nature (analytically tractable, facilitate seamless integration with existing Kriging implementations)

\[
h(x) \mid X, H, M_k, \pi\{\beta_k, \sigma_k^2, \theta_k\} \sim \mathcal{N}\left(\bar{h}(x), \bar{\sigma}^2(x)\right)
\]

\[ \bar{h}(x \mid M_k) = f_k(x)^T \beta_{k_{BMA}} + r_{\theta_k}(x)^T R_{\theta_k}^{-1}(H - F_k \beta_{k_{BMA}}) \]

\[ \bar{\sigma}^2(x \mid M_k) = (\sigma_k^*)^2 \left[ 1 + u_k^T \left( (c_k)^{-1} F_k^T F_k + F_k^T R_{\theta_k}^{-1} F_k \right)^{-1} u_k - r_{\theta_k}^* (x)^T R_{\theta_k}^{-1} r_{\theta_k}(x) \right] \]
Computational implementation I

Repeat for every model

Predict under model $M_k$

Inefficient for large number of models $(n_m>>1)$

$p(h \mid x, X, H \bigcup M_k) = \sum_{k=1}^{n_m} p(h \mid x, X, H, M_k) P(M_k \mid X, H)$
Efficient averaging of predictions

Utilize *Occam’s Razor* principle, and only consider models whose plausibility is non-negligible [1]:

\[
\hat{\mathcal{M}} = \left\{ M_p \left| P(M_p \mid X, H) \geq \alpha \cdot \max_{M_k \in \mathcal{M}} P(M_k \mid X, H) \right. \right\}
\]

\[
p(h \mid x, X, H \bigcup M_k) = \sum_{M_p \in \hat{\mathcal{M}}} p(h \mid x, X, H, M_p) P(M_p \mid X, H)
\]

where \( \alpha \) controls the degree of restriction:

- \( \alpha = 0 \) : include all models;
- \( \alpha = 1 \) : include the most plausible one only.

Computational implementation II

Training data

$x$

$h$

Model (basis functions) \( M_1 \)

Evidence

Post prob.

Model (basis functions) \( M_2 \)

Evidence

Post prob.

Model (basis functions) \( M_k \)

Evidence

Post prob.

\( M_k \)

\[
P(H | X, M_k) = P(H | X, M_k, \sigma_k^2, \theta_k^*)
\]

where

\[
[(\sigma_k^*), \theta_k^*] = \arg \max_{\sigma_k^2, \theta_k} P(H | X, M_k, \sigma_k^2, \theta_k)
\]

Bayesian model class averaging

Each model’s evidence calculation requires optimization (EB parameters) \( \rightarrow \) intensive computation!
Model space exploration setting

We assume the pool of interested basis functions are polynomial functions up to some order $p$ (set to 2 commonly):

$$f_c = \left\{ \prod_{i=1}^{n_x} x_i^{p_i} \left| \sum_{i=1}^{n_x} p_i \leq p \right\} \right.$$  

$$= \left[ 1, x_1, \ldots, x_{n_x}, x_1^2, \ldots, x_{n_x}^2, x_1 x_2, \ldots, x_{n_x-1} x_{n_x}, \ldots, x_1^{p-1} x_2, \ldots, x_{n_x}^p \right]$$

- A popular setup because of its flexibility and physical interpretability

- Number of potential models is large, for example for $n_x=10$ dimensional problem, the combination to explore (for $p=2$) are $4 \times 10^{15}$!

- An efficient search implementation is necessary
Efficient model search principles

• *Occam’s Razor*: the model is ignored if its posterior probability is lesser than its simpler variants (those with basis functions as its subsets) [1].

• *Effect Heredity*: a basis function is allowed to appear only if all its parents (corresponding lower order ones) are already included [2].
  For example, the quadratic term \((x^2)\) is allowed only if the model already includes the corresponding constant and linear terms [1, \(x\)].

**Forward search from simpler to more complex models**

Start with ordinary Kriging (OK). Calculate its EB parameters through Eq. (16) and marginal likelihood \( P(M_0|X,H) \). Set active set \( \mathcal{A} = \{M_0\} \).

1st iteration (\( k = 1 \))

Select model \( M_p \) to correspond to the \( k \)th model in \( \mathcal{A} \) and obtain all allowable basis function additions (functions whose parents are included in \( M_p \)).

Construct the model \( (M_p^{+1}) \) with one added basis function. Calculate the posterior probability \( P(M_p^{+1} | X, H) \).

Elimination

\[ P(M_p^{+1} | X, H) \leq P(M_p | X, H) \]

- yes: Eliminate \( M_p^{+1} \)
- no: Retain \( M_p^{+1} \)

Repeat for each allowable basis function addition \( f_p^{+1} \).

Construct the final set of models by screening across \( \mathcal{A} \)

\[ \mathcal{M} \rightarrow \{M_p \in \mathcal{A} | P(H | X, M_p)P(M_p) \geq \alpha \cdot \max_{M_i \in \mathcal{A}} P(H | X, M_k)P(M_k) \} \]

Weights for each model are

\[ w_k = \frac{P(M_k | X, H)}{\sum_{p=1}^{\hat{n}_m} P(M_p | X, H)} \]

Is there a \( k+1 \) model in \( \mathcal{A} \)?

Enrich active set:

\[ \mathcal{A} \rightarrow \mathcal{A} + M_p^{+1} \]

Explore if more complex model satisfies Occam and heredity principles. If yes explore around it (consider even more complex models).
Intermediate elimination

- We can obtain (no computational cost) a guess for a new (complex) $M_p^{+1}$ model’s posterior probability using the known EB parameter of the parent $M_p$ simpler one:

$$P(M_p^{+1} \mid X, H) \approx P(M_p^{+1} \mid X, H, \sigma_p^2, \theta_p)$$

- Intermediate elimination: ignore the complex model without the need of calculating its EB parameter, if its approximated posterior probability is very unpromising (significantly below simpler one).

$$P(M_p^{+1} \mid X, H, \sigma_p^2, \theta_p) \leq \lambda P(M_p \mid X, H, \sigma_p^2, \theta_p)$$

- Only if the model passes intermediate elimination, we identify its actual EB parameters to examine actual Occam’s razor.
Forward and stepwise search II

1st iteration ($k = 1$)

- Construct the model ($M_p^{+1}$) with one added basis function.
- Approximate the posterior probability $\tilde{p}(M_p^{+1} \mid X, H)$
- Check $\tilde{p}(M_p^{+1} \mid X, H) \leq \lambda p(M_p \mid X, H)$
  - Yes: Eliminate $M_p^{+1}$
  - No: Retain $M_p^{+1}$

Repeat for each allowable basis function addition $f_p^{+1}$

Construct the final set of models by screening across $\mathcal{A}$

$$\mathcal{M} \rightarrow \left\{ M_p \in \mathcal{A} \mid P(H \mid X, M_p)P(M_p) \geq \alpha \cdot \max_{M_i \in \mathcal{A}} P(H \mid X, M_i)P(M_i) \right\}$$

Weights for each model are

$$w_k = \frac{P(M_k \mid X, H)}{\sum_{p=1}^{n_m} P(M_p \mid X, H)}$$

Termination

- Is there a $k+1$ model in $\mathcal{A}$?
  - Yes: Enrich active set: $\mathcal{A} \rightarrow \mathcal{A} + M_p^{+1}$
  - No: Eliminate $M_p^{+1}$

Obtain the EB parameters for $M_p^{+1}$ through Eq. (16)

and the actual posterior probability $P(M_p^{+1} \mid X, H)$

Repeat for each retained model $M_p^{+1}$

**Intermediate elimination**

Select model $M_p$ to correspond to the $k$th model in $\mathcal{A}$ and obtain all allowable basis function additions (functions whose parents are included in $M_p$)

Start with ordinary Kriging (OK). Calculate its EB parameters through Eq. (16) and marginal likelihood $P(M_0 \mid X, H)$. Set active set $\mathcal{A} = \{M_0\}$.
Comparison methods

**Blind Kriging (BK):** forwardly selects the next ‘best’ basis function to add until the cross validation error decreases [3].

**Dynamic Kriging (DK):** formulate the basis function selection as an integer programming problem, and seeks the combination that minimizes the process variance with Genetic Algorithm [4].


Comparison methods

**Blind Kriging (BK):** forwardly selects the next ‘best’ basis function to add until the cross validation error decreases [3].

**Dynamic Kriging (DK):** formulate the basis function selection as an integer programming problem, and seeks the combination that minimizes the process variance with Genetic Algorithm [4].

**BMAK:** with efficient model averaging and search strategies;

**Full-BMAK:** proposed approach without efficient strategies [only applicable to low-dimensional problems];

**BMSK:** same as BMAK, but only use the most plausible model to predict (i.e. model selection variant)
Benchmark problems I

**Branin:** two-dimensional problem with moderate nonlinearity

**Griewank:** two-dimensional problem with high nonlinearity
Benchmark problems II

**Borehole:** engineering problem modelling the water flow rate through a borehole (Harper et al. 1983)

**Half-Car:** realistic engineering problem calculating the road-holding statistics of a half-car nonlinear model riding on a rough road (Zhang et al., 2016).

<table>
<thead>
<tr>
<th></th>
<th>Dimension</th>
<th>Number of experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Branin</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>Griewank</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>Borehole</td>
<td>8</td>
<td>24</td>
</tr>
<tr>
<td>Half-Car</td>
<td>19</td>
<td>60</td>
</tr>
</tbody>
</table>
# Prediction correlation coefficient

<table>
<thead>
<tr>
<th></th>
<th>Branin</th>
<th>Griewank</th>
<th>Borehole</th>
<th>Half-Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full BMAK</td>
<td>0.925</td>
<td>0.946</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>BMAK</td>
<td>0.915</td>
<td>0.949</td>
<td>0.996</td>
<td>0.917</td>
</tr>
<tr>
<td>BMSK</td>
<td>0.923</td>
<td>0.945</td>
<td>0.995</td>
<td>0.890</td>
</tr>
<tr>
<td>BK</td>
<td>0.915</td>
<td>0.917</td>
<td>0.983</td>
<td>0.862</td>
</tr>
<tr>
<td>DK</td>
<td>0.897</td>
<td>0.932</td>
<td>0.914</td>
<td>0.692</td>
</tr>
</tbody>
</table>

*Reported*: the average prediction correlation over 50 independent runs

Higher value \(\rightarrow\) Better prediction
More comprehensive performance assessment

Half-car example

Correlation coefficient

RMS error

Prediction interval coverage

Better performance

Better performance

Better performance

ρ

RMSE (N)

P_c (%)
Performance for UQ predictions

Reported: the Kolmogorov–Smirnov measure between the true and prediction CDF over 50 independent runs

Smaller discrepancy $\rightarrow$ Better performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Borehole</th>
<th>Half-Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMAK</td>
<td>0.0144</td>
<td>0.0337</td>
</tr>
<tr>
<td>BMSK</td>
<td>0.0154</td>
<td>0.0408</td>
</tr>
<tr>
<td>BK</td>
<td>0.0283</td>
<td>0.0571</td>
</tr>
<tr>
<td>DK</td>
<td>0.0818</td>
<td>0.0724</td>
</tr>
</tbody>
</table>
Computational cost

<table>
<thead>
<tr>
<th>Comp. cost</th>
<th>Branin</th>
<th>Griewank</th>
<th>Borehole</th>
<th>Half-Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuning (EB optimization)</td>
<td>3.17</td>
<td>3</td>
<td>20.5</td>
<td>7.37</td>
</tr>
<tr>
<td>Predictions (model averaging)</td>
<td>3.14</td>
<td>3</td>
<td>16.7</td>
<td>7.25</td>
</tr>
<tr>
<td>Model space size</td>
<td>63</td>
<td>63</td>
<td>$3.5 \times 10^{13}$</td>
<td>$1.6 \times 10^{63}$</td>
</tr>
</tbody>
</table>

Conventional Kriging cost = 1 for both tuning and predictions

BMAK’s computational cost is at most one magnitude greater than conventional Kriging with deterministic regression structure
Conclusion

• A Bayesian Model Averaging formulation for Kriging (BMAK) is proposed to systematically treat regression structure uncertainties.

• A data-driven prior is proposed, combining the benchmark g-prior and the Empirical Bayes prior; Selection lends itself to analytical tractability of problem.

• Efficient model averaging and search strategy is proposed for problems with large model space.

• Results demonstrates the advantage of BMAK in both analytical and engineering problems.
Thank You!