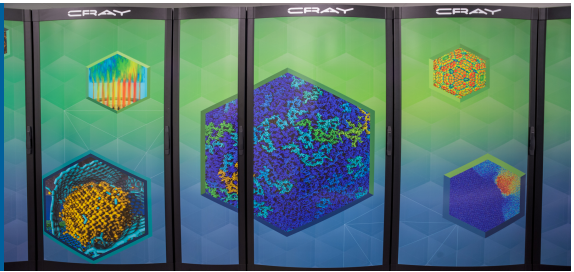
Argonne NATIONAL LABORATORY 75 1946–2021

Kernel Model Validation: How To Do It, And Why You Should Care



Carlo Graziani
Argonne National Laboratory

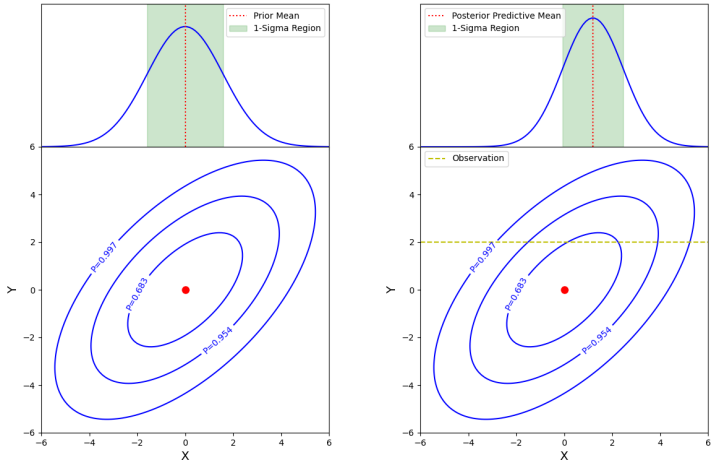
24 April 2025
DATAWORKS 2025
Alexandria, VA

Gaussian Process Models

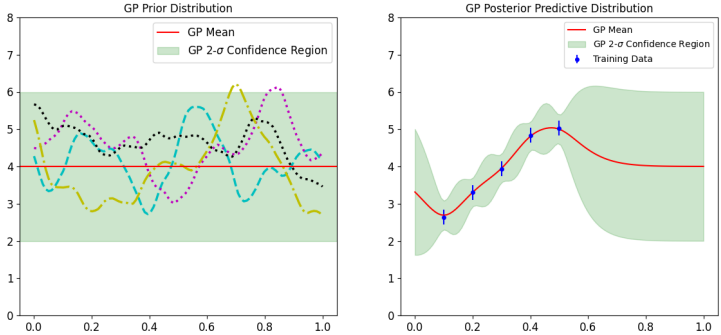
- *Gaussian Process Models* are a broad family of nonparametric methods within the broader family of *kernel methods*. GPs embed optimal reconstruction methods in a statistical context that allow one to reason quantitatively and consistently about model error.
- A Gaussian process may be thought of (informally) as an infinite-dimensional multivariate normal distribution over a space of functions.
- Just as a MVN is entirely characterized by its mean vector and a symmetric-positive-definite covariance matrix, a GP over a space Ω (e.g $\mathbb{R}^N$) is characterized by a mean *function* $\mu(\boldsymbol{x})$ and by a symmetric-positive-definite *covariance kernel* $K(\boldsymbol{x}, \boldsymbol{x}')$.
- The choice of $K(\cdot, \cdot)$ is in effect the choice of the space of functions that is sampled by the GP. It controls function properties such as continuity, order of differentiability, periodicity, scale, etc.

Why Are GPs Useful?

- GPs can be trained on noisy observations, to estimate hyperparameters embedded in $\mu(\boldsymbol{x})$ and $K(\boldsymbol{x}, \boldsymbol{x}')$.
- More importantly, GPs can make *posterior predictions* of function values at points $\boldsymbol{x}^*$ that have not yet been observed/acquired.
- These are inherently *probabilistic* predictions. They are embody not only an interpolation, but also the uncertainty in that interpolation.



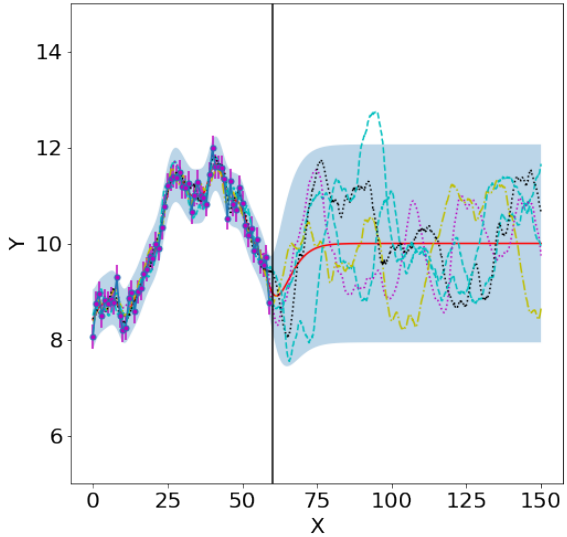
With a finite-dimensional MVN, evaluation/observation of some variables leads to a more informative conditional distribution over the remaining variables.



Similarly, observations of values of a function governed by a GP (or of any linear functional of such a function) results in an updated conditional mean function and covariance kernel, making probabilistic predictions in the rest of the space.

GP Applications

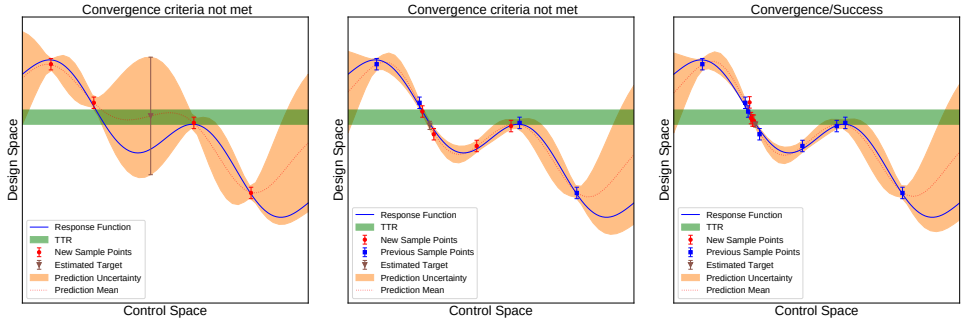
- The ability to make probabilistic predictions about *model* properties have made GPs a very popular tool for many data analysis/UQ applications:
  - Image reconstruction;
  - Surrogate modeling;
  - Output emulators for expensive simulators;
  - Black-box function optimization;
  - Optimal experimental design.
- However, while the predictive uncertainty output by a GP seems valuable, it is often difficult to say precisely what it means. What is the *calibration statement* of a GP model's 90% (say) predictive credible regions?

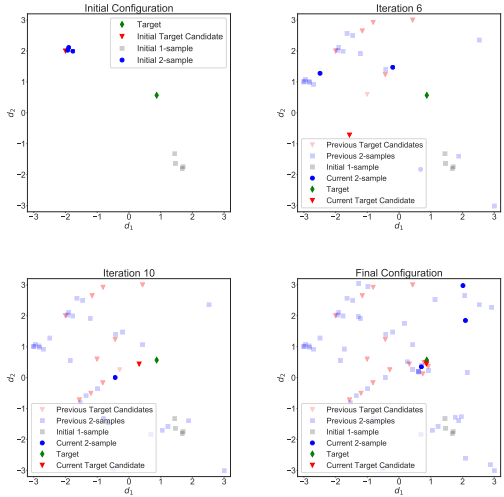


What *is* that functional (co-)variance thing, anyway? What is it good for?

Motivating Example: Targeted Adaptive Design

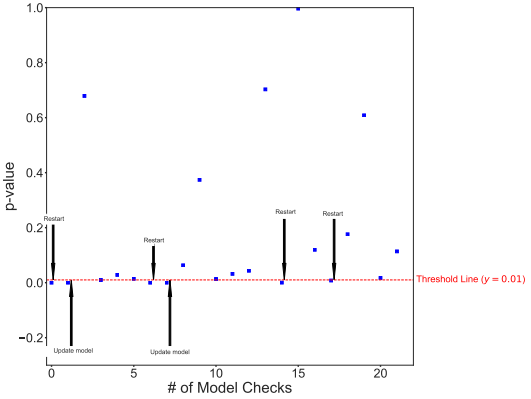
- Targeted Adaptive Design (Graziani & Ngom, SIAM/ASA JUQ 2024, arxiv.org/2205.14208) is a target optimization algorithm for discovering settings $\boldsymbol{x}$ of some complex experiment that result in a desired set of output features $\boldsymbol{F}$, within some tolerances, when the response $\boldsymbol{f}(\boldsymbol{x})$ is not known *a priori*, and must be determined by expensive experiments or simulations.



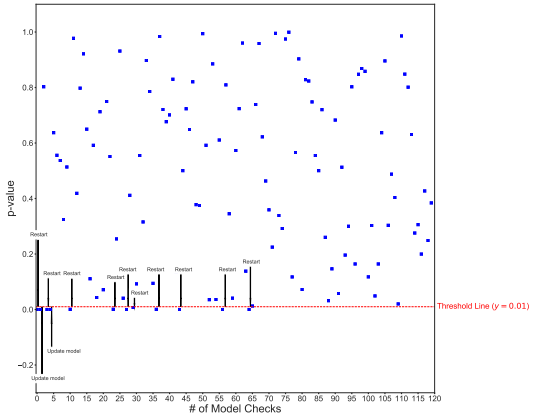


- TAD replaces the unknown response with a vector-valued GP surrogate model.
- It optimizes an acquisition function over a cloud of probe (“2”) points and a single target point, looking for a location where the GP surrogate predictive distribution at the target has an uncertainty that fits within the tolerance box.
- At the end of each iteration, it acquires the function value at the optimal probe and target points.

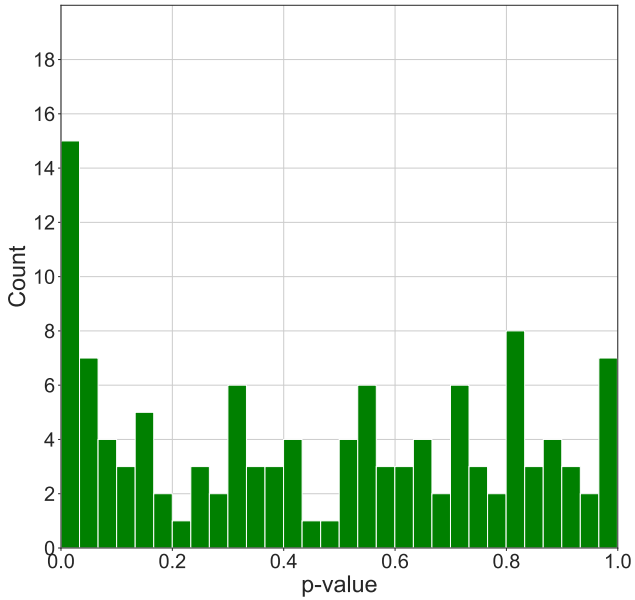
- Early versions of TAD suffered from convergence failures.
- The aquisition function, $E_{\boldsymbol{f}_2}(\boldsymbol{f}_T = \boldsymbol{F} | \boldsymbol{f}_1, \boldsymbol{x}_1, \boldsymbol{f}_2, \boldsymbol{x}_2)$ would at first increase in the direction of the solution, but eventually start leading away from it.
- An investigation showed that the model was making very bad *probabilistic* predictions of $\boldsymbol{f}(\boldsymbol{x})$ near the optimization solutions. The aquired data $\boldsymbol{f}_2$ did not look like plausible samples from the GP predictive distribution.
- The *Mahalanobis distance* $\chi^2 = (\boldsymbol{f}_2 - \boldsymbol{\mu}_{2,pred})^\top \boldsymbol{K}_{pred}^{-1} (\boldsymbol{f}_2 - \boldsymbol{\mu}_{2,pred})$ should have been distributed as χ^2 with DOF=$\mathrm{Dim}(\boldsymbol{f}_2)$. It was not.
- By monitoring χ^2, and adding flexibility to the covariance every time it reached implausible values, we obtained a convergent algorithm.



Convergence/Success



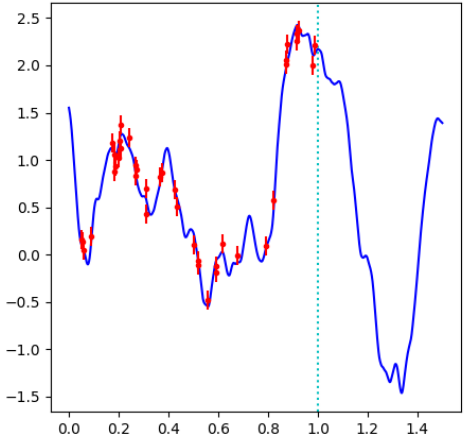
Convergence/No Solution



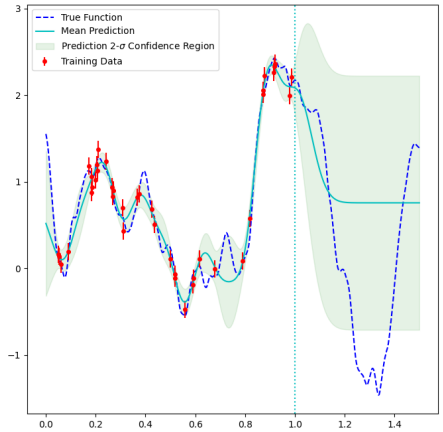
P-values from χ^2 distribution, no-solution case

Methods For Kernel Model Validation

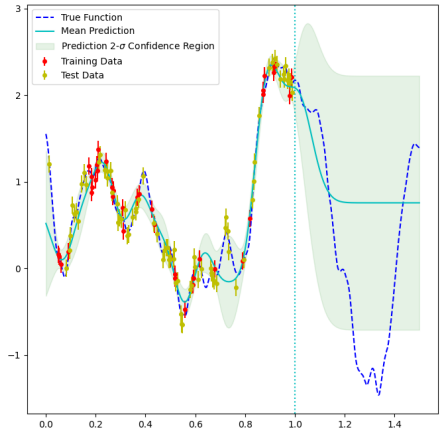
- The Mahalanobis χ^2 seems like a useful tool for determining whether the covariance kernel and the data are on speaking terms, if one has either a prediction/acquisition cycle or else data held out from training.
- What else could one do? Let's do some experiments.



- Blue line: a random function, sampled from a zero-mean GP with a Matern($\nu = 1.5$) covariance. This is a sample from a space of C^1 (once-differentiable) functions—quite rough.
- Red points: random data values sampled from the function at points $x < 1$, with added noise.

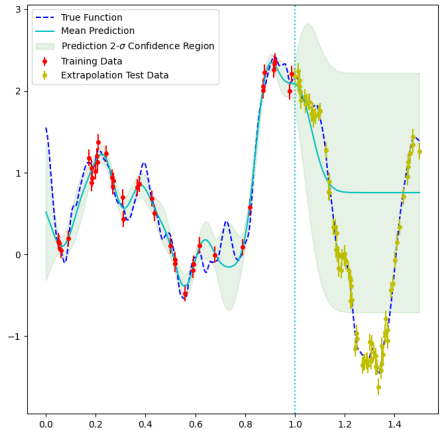


- Regression fit using GP model with squared-exponential covariance. The random functions assumed by the model are C^∞—very smooth.
- The band is a 2-sigma credible region on the function value.
- The kernel model is seriously mis-specified.
- The mean prediction looks OK, though...



- The 80 yellow points are data sampled from the true model, with the same additive noise as the training data, but held out from training.
- The Mahalanobis χ^2 for the held-out data is $\chi^2 = 129$ for DOF=80, a $P - value$ of 4.3×10^{-4}.
- Clearly, the predictive uncertainties are quite wrong with this model. They would be worthless for UQ.

What About Extrapolation?



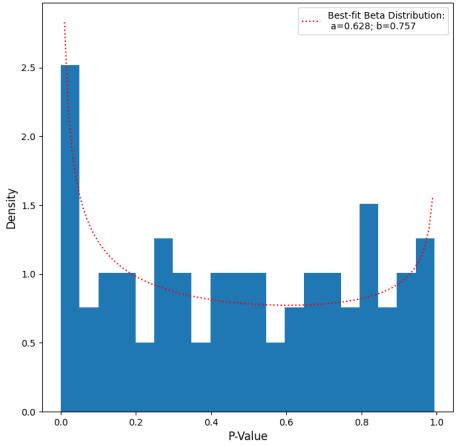
- The 80 yellow points are again data sampled from the true model, with the same additive noise as the training data. They all have $x > 1$, so they are outside the support of the training data.
- The Mahalanobis χ^2 for the held-out data is $\chi^2 = 141$ for DOF=80, a $P - value$ of 2.7×10^{-5}.
- So any UQ-bearing extrapolation using this model would also be worthless.

Exploring Fine-Grained Model Deviations

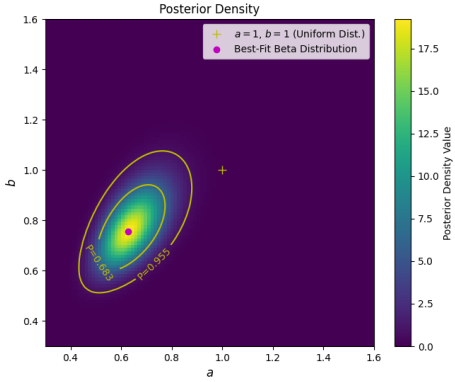
- The N held-out data points yielded a single measure, χ^2, which can indicate a problem, but is a coarse summary.
- A "good" χ^2 (P-value a reasonable sample from $U(0,1)$) might still conceal a bad model, if anomalously large residuals are compensated by anomalously small ones.
- By diagonalizing the predictive covariance, $\boldsymbol{K}_{pred}$, we can access individual standard normal residuals:

$$\boldsymbol{O}^\top \boldsymbol{K}_{pred} \boldsymbol{O} = \mathbf{diag}(\sigma_1^2, \ldots, \sigma_N^2) \quad ; \quad \boldsymbol{d} \equiv \boldsymbol{O}^\top (\boldsymbol{y} - \boldsymbol{\mu}_{pred}) \quad \Longrightarrow \; d_k \sim \mathcal{N}(0, \sigma_k).$$

- The d_k are IID. Their CDFs should be IID $U(0,1)$, if the model is good.

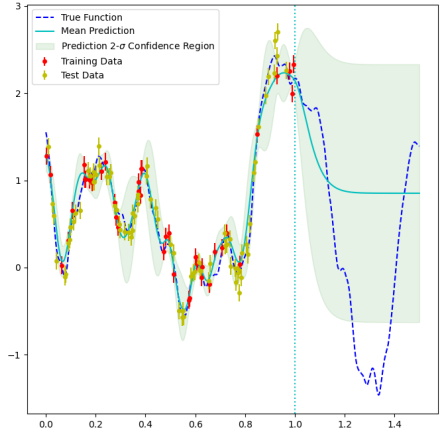


- These are the IID cumulative probabilities corresponding the earlier 80 testing points.
- A likelihood fit of a Beta distribution density ($\pi_\beta(p; a, b) \propto p^{a-1}(1-p)^{b-1}$) yields a best-fit model with “horns” near $p = 0$ and $p = 1$.
- There is an excess of both too-small and too-large residuals!

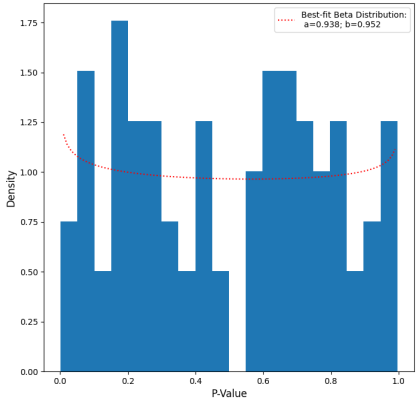


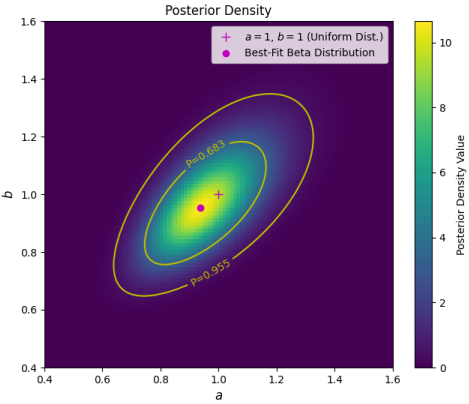
- This is a Bayesian posterior over the Beta distribution parameters a, b.
- The values $a = 1$, $b = 1$ that correspond to a $U(0, 1)$ distribution are very unlikely in the posterior: the iso-posterior contour that crosses this point contains a probability $P = 1 - 1.3 \times 10^{-4}$.

Let's Try With A Better Model



- This is the posterior predictive from a GP model with a Matern($\nu = 2.5$) covariance kernel.
- Note that this model is *still* somewhat mis-specified, since it samples C^2 functions instead of C^1 functions.
- But it seems to be doing better than the squared-exponential model on the held-out data: $\chi^2 = 85.8$ for DOF=80, which gives a respectable P-value of 0.307.





The residual analysis confirms that this model makes trustworthy probabilistic predictions, despite its known misspecification.

Higher Dimensions

- 1-D scalar function domains are relatively simple to model.
- With higher-dimensional input and/or output, covariance kernel choices embrace not only kernel functional form, but also spatial correlation among dimensions.
- Factored covariances, isotropic covariances etc. are often *very* bad at modeling real data (e.g. the TAD example).
- The methodologies presented here can still serve as a guide to model (in-)adequacy in such cases.

Summary

- It is in fact possible to ascribe quantitative meaning to GP predictive credible regions. The key is testing with data held out from training.
- This is hardly ever done in literature: generally mean-squared error is felt to be an reasonable measure of model adequacy. It isn't.
- GPs furnish *probabilistic* predictions, not just predictive means. If these distribution are wrong, the UQ consequences cannot be trusted. This can be verified.
- GP practitioners should *always* verify their predictive distributions.