Presentations and Projects

Here is a list of topics suitable for presentation and a project. Alternatively, feel free to discuss with one of the instructors a topic of your own choice.

There are two tasks. First, you will make a short presentation (20 minutes + 5 minutes of questions) in class. The presentation is essentially describing the main ideas of the paper or software. Then you will continue with the same topic to carry out a project. It will not be presented, but a written report is due December 12 (electronic pdf please).

Some more details:

- Only one student per topic please. This applies across the three universities, not just within.
- There will be an editable spreadsheet where you can reserve a topic.
- There are separate lists from each instructor, reflecting their interests in getting particular questions answered. You may take a topic from any list; the grade will be assigned by your home-university instructor.
- You should hold a short planning meeting with your instructor before you start work.
- The time for presentation is very limited. Concentrate on *what* is achieved rather than the modelling details of *how* the authors implemented their method. You could, for instance, build your presentation around an example to illustrate the objectives and what was found. If one of the paper's examples is not useful for this purpose, simple examples can be found at `http://www.sfu.ca/~ssurjano/optimization.html`

  Technical comments should be limited to statements like, "The authors used GP models for ... (e.g., a variable, a component in a model)". There will not be time for the math details. Those details can go in your written project.

Derek's Suggested Topics

1. L. Pronzato and W. G. Müller, "Design of computer experiments: Space filling and beyond," *Statistics and Computing*, vol. 22, pp. 681–701, 2012
   - This paper outlines several design techniques for computer experiments. A reasonable question to ask is "does design matter?".
   - Using a selection of test functions from the Virtual Library of Simulation Experiments (`http://www.sfu.ca/~ssurjano/index.html`), conduct a simulation study to address the question "does design matter?".
   - Your project should consider at least two of the design strategies outlined in the paper. In addition, include a completely random design among the candidates.
   - To evaluate the designs, use a criterion (e.g., root mean square prediction error).
2. J. Oakley and A. O'Hagan, "Bayesian inference for the uncertainty distribution of computer model outputs," *Biometrika*, vol. 89, pp. 769–784, 2002
   - This paper considers the problem of estimating the distribution of computer model outputs when the input is a random variable.

- A project based on this paper would include a (i) description of the methodology and problem it aims to solve; (ii) an implementation of the methodology; (iii) an evaluation of the methodology using some test functions from the Virtual Library of Simulation Experiments (`http://www.sfu.ca/~ssurjano/index.html`).

3. A. O'Hagan, "Polynomial chaos: A tutorial and critique from a statistician's perspective," *Submitted to SIAM/ASA Journal of Uncertainty Quantification*, 2013. Available at http://www.tonyohagan.co.uk/academic/pub.html
   - This paper considers the estimation of the distribution of computer model outputs when the input is a random variable. In addition, it attempts to make connections with polynomial chaos expansions and the usual statistical approaches (see project 2).
   - A project based on this paper should (i) summarize the paper and its main findings; (ii) implement the non-intrusive polynomial chaos approach for at least test problems and evaluate the performance of the approach. Again the Virtual Library of Simulation Experiments could be used as a source of test functions.
4. D. Bingham, P. Ranjan, and W. J. Welch, "Design of computer experiments for optimization, estimation of function contours, and related objectives," in *Statistics in Action: A Canadian Outlook* (J. F. Lawless, ed.), pp. 109–124, Boca Raton, Florida: CRC Press, 2014
   - This paper considers several scenarios where sequential design is used in computer experiments.
   - A project based on this paper would include a (i) description of the methodology and problem it aims to solve; (ii) an implementation of the expected improvement approach for one scenario; (iii) an evaluation of the methodology using at least one test function from the Virtual Library of Simulation Experiments (`http://www.sfu.ca/~ssurjano/index.html`).
5. C. G. Kaufman, D. Bingham, S. Habib, K. Heitmann, and J. A. Frieman, "Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology," *Annals of Applied Statistics*, vol. 5, no. 4, pp. 2470–2492, 2011
   - This paper presents methodology for dealing with large ensembles of computer models. R code is available at Cari Kaufman's webpage (`http://www.stat.berkeley.edu/~cgk/`).
   - A project based on this paper would include a (i) description of the methodology and problem it aims to solve; (ii) an evaluation of the methodology for a moderately large computer experiment; (iii) a comparison of the GP in the same setting.
6. D. Higdon, J. Gattiker, B. Williams, and M. Rightley, "Computer model calibration using high-dimensional output," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 570–583, 2008
   - This paper introduces the multivariate calibration approach that Dave Higdon presented in his lecture.

- A project based on this paper would include a (i) description of the methodology and the problem it aims to solve; (ii) an evaluation of the methodology for a multivariate output.

7. D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global Optimization*, vol. 13, pp. 455–492, 1998
   - Read the first 17 pages of the paper and implement a Gaussian process. You should have done this by now.
   - Conduct a simulation study to investigate the relationship between sample sizes, the correlation parameters and dimension of the inputs.
   - If you had a fairly simple function is 20-d, how many design points are needed. What is the model is complex?

Pritams's Suggested Topics

1. G. M. Vernon, I. and R. Bower, "Galaxy formation: A Bayesian uncertainty analysis," *Bayesian Analysis*, vol. 5, pp. 619–669, 2010
   - This paper attempts to find appropriate values of a few parameters of a computer model called Galform (simulates the creation and evolution of approximately one million galaxies) which would lead to model output that can be compared with real world data. This has been referred to as the history matching problem.
   - Project : Summarize the paper; compare the results with a naive approach of minimizing $||\delta||$ for a toy function (I have a toy function, let me know if you need it).
2. M. T. Pratola, S. R. Sain, D. Bingham, M. Wiltberger, and E. J. Rigler, "Fast sequential computer model calibration of large nonstationary spatial-temporal processes," *Technometrics*, vol. 55, no. 2, pp. 232–242, 2013
   - The motivating problem in this article is a space-weather model with three design variables latitude, longitude, and time (fixed in this application) and three calibration parameters which have to be estimated. This calibration problem can be viewed as an inverse problem.
   - Project : Summarize the paper; compare the results with a naive approach of minimizing $||\delta||$ for a toy function (I have a toy function, let me know if you need it).
3. S. Montagna and S. Tokdar, "Computer emulation with non-stationary Gaussian processes," *arXiv:1308.4756*, 2013
   - Presents a new non-stationary covariance function for GP-based emulator for non-stationary processes. It claims to outperform TGP.
   - Project: Summarize the paper; compare it with BART (in BayesTree).
4. F. Liu and M. West, "A dynamic modelling strategy for Bayesian computer model emulation," *Bayesian Analysis*, vol. 4, pp. 393–412, 2009

- Presents a methodology for dynamic emulation of computer simulator with time series response.
- Project: Summarize the paper; compare it with the principal component based method in Higdon et al. (2008).
  D. Higdon, J. Gattiker, B. Williams, and M. Rightley, "Computer model calibration using high-dimensional output," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 570–583, 2008

5. C. Paciorek and M. Schervish, "Nonstationary covariance functions for Gaussian process regression," in *Advances in Neural Information Processing Systems 16* (S. Thrun, L. Saul, and B. Schölkopf, eds.), (Cambridge, MA), pp. 273–280, MIT Press, 2004

   - Presents a methodology for fitting non-stationary GP in computer experiments.
   - Project: summarize the paper; compare it with tgp

Will's Suggested Topics

1. T. Chen and J. Ren, "Bagging for Gaussian process regression," *Neurocomputing*, vol. 72, pp. 1605–1610, 2009

   - Model averaging (an ensemble of models) has received much research attention for data observed with random error. These authors apply combine bagging and GPs to model observational data from an industrial process.
   - Project: Adapt bagging with GP models to deterministic computer experiments (you won't want to sample with replacement). Try a few variants of your adaptation of bagging and try the method on say two examples: the G-protein computer experiment (easy to model) and the Wonderland experiment (hard to model).

2. GPM/SA software
   `http://www.stat.lanl.gov/source/orgs/ccs/ccs6/gpmsa/gpmsa.html`

   - The presentation should summarize the capabilities and limitations of the software.
   - Project: Run the software to demonstrate (some of) its capabilities using example data of your choice.

3. R. B. Gramacy and H. K. H. Lee, "Bayesian treed Gaussian process models with an application to computer modeling," *Journal of the American Statistical Association*, vol. 103, no. 483, pp. 1119–1130, 2008

   - The authors combine regression trees to divide the input space with GP fits in each node.
   - Try the method on the Wonderland application. Compare with a regression tree or random forest as commonly used in statistical learning (without a GP fit in each node).

4. M. S. Handcock and M. L. Stein, "A Bayesian analysis of kriging," *Technometrics*, vol. 35, no. 4, pp. 403–410, 1993

- This paper outlines a full Bayesian implementation of the use of a Gaussian Process for prediction, including a Bayesian treatment of uncertainty in estimating all parameters.
- Project: Use the same authors' priors and implement in `R` a full Bayesian MCMC implementation. Thus the project will build on your code in Assignment 2 for the likelihood function, etc. Run an example.

5. A. B. Owen, "Variance components and generalized Sobol' indices," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 1, pp. 19–41, 2013

   - Sobol' indices are measures of sensitivity of the output of computer code to its inputs. Owen reviews and extends them.
   - Project: Example 1 of Oakley and O'Hagan (2004) is a function to test methods of sensitivity analysis. Details for the example are available at
     `http://www.jeremy-oakley.staff.shef.ac.uk/psa_example.txt`
     Use the example to illustrate some of variations of Sobol' indices and compare with the functional ANOVA / visualization from a GP that we studied in class. You will need to screen the methods in Owen (2013) to identify some that can be readily tried. (Why or why not the various methods are easy to use would be useful content for your presentation.) I can help with code to do the GP functional ANOVA.

     J. E. Oakley and A. O'Hagan, "Probabilistic sensitivity analysis of complex models: a Bayesian approach," *Journal of the Royal Statistical Society B*, vol. 66, pp. 751–769, 2004

6. P. Ranjan, R. Haynes, and R. Karsten, "A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data," *Technometrics*, vol. 53, no. 4, pp. 366–378, 2011

   - The question of adding a nugget to reduce ill-conditioning has arisen several times during the course, and there is a large literature on the subject. This is the "go to" paper: it identifies when ill-conditioning will occur and provides an elegant solution if it does.
   - Project: Section 3.3 of Gramacy and Lee (2012) gives an example that raises another interesting question: can a nugget help for a computer model that is technically deterministic but has "random behaviour". Explore the example for various sample sizes and try fitting the following GP models: without a nugget; with nuggets of various fixed sizes; with the nugget optimized by MLE; and using the method in Ranjan et al. (2011). Compare the strategies via prediction accuracy. (All this can be done with software already available. Some analyses may fail due to ill-conditioning, but that is an interesting aspect of the results.)
     R. B. Gramacy and H. K. H. Lee, "Cases for the nugget in modeling computer experiments," *Statistics and Computing*, vol. 22, no. 483, pp. 713–722, 2012