

Precision Cosmology with Gaussian Processes

U. Alam, D. Bingham, SH, K. Heitmann, D. Higdon, T. Holsclaw,
L. Knox, E. Lawrence, H. Lee, C. Nakhleh, B. Sanso,
M. Schneider, M. White, B. Williams, ---

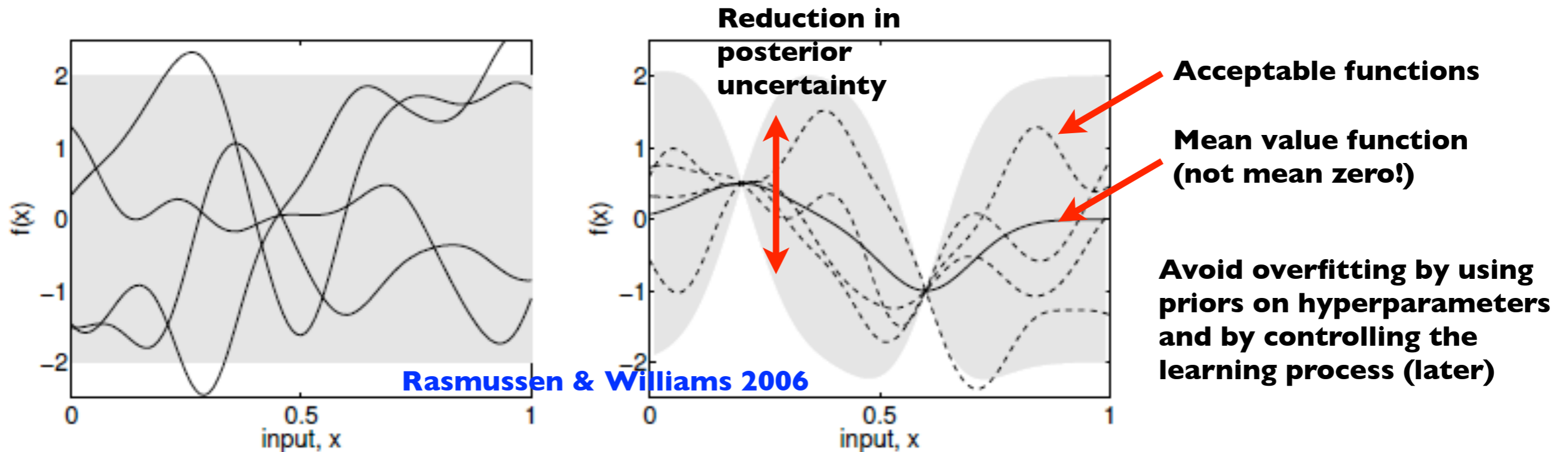
LA-UR 08-07921, 09-05888

- Precision cosmology requires dealing with (i) expensive state of the art simulations, (ii) large number of dimensions, (iii) regression (input-output relationships), (iv) estimation and control of errors, (v) regularizing and solving ill-posed inverse problems (given data, estimate model parameters)
- In solving regression and inverse problems (both of which may be considered as problems in Bayesian inference) one has to make choices about characteristic functions by either (i) restricting attention to a single type (linear) or class of functions (polynomials), or (ii) assign prior probabilities to classes of functions, with some considered more likely (due to smoothness, say)
- Gaussian Processes (GPs) provide a surprisingly computationally effective method with which to apply the latter approach; now becoming increasingly popular (in latest edition of Numerical Recipes)
- We have applied the GP in several places: (i) the COSMIC CALIBRATION framework (talk by Katrin), (ii) photo-z estimation, (iii) $w(z)$ reconstruction from S_n data, ---

Salman Habib, Benasque workshop, August 2010



Bayesian Approach: Basics



Prior distribution over random functions: global mean zero (although individual choices clearly are not mean-zero), variance assumed to be independent of x , 2-SD band in gray

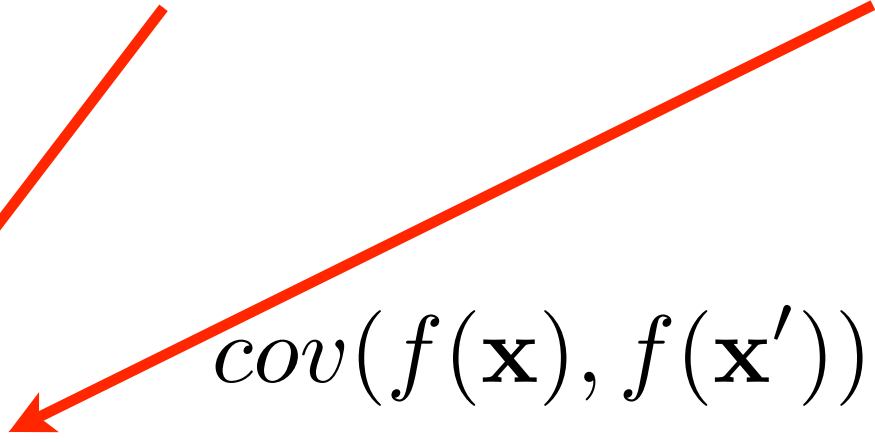


Posterior distribution conditioned on exact information at two x points, consider only those functions from the prior distribution that pass through these points

- **GPs are *nonparametric*, so there is no need to worry if the functions can fit the data (e.g., linear functions against nonlinear data), even with many observations still have plenty of candidate functions**
- **With GP models, the choice of prior distribution over random functions is essentially a statement of the properties of the initial covariance function, these properties can be specified in terms of a set of *hyperparameters*, using data to determine these defines the *learning* problem for the GP approach**

GP Modeling: Basics I

GPs are straightforward generalizations of Gaussian distributions over vectors to function spaces, and are specified by a mean function and a covariance function

$$\mathbf{f} = (f_1, \dots, f_n)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$


$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$
$$\text{cov}(f(\mathbf{x}), f(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}')$$

They have several convenient properties, of which the two most significant are

- **Marginalization yields a Gaussian distribution**

$$p(\mathbf{y}_a) = \int p(\mathbf{y}_a, \mathbf{y}_b) d\mathbf{y}_b$$

$$p(\mathbf{y}_a, \mathbf{y}_b) = \mathcal{N} \left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \right) \implies p(\mathbf{y}_a) = \mathcal{N}(\mathbf{a}, \mathbf{A})$$

GP Modeling: Basics II

- **Conditioning yields a new Gaussian distribution**

$$p(\mathbf{y}_a | \mathbf{y}_b) = \frac{p(\mathbf{y}_a, \mathbf{y}_b)}{p(\mathbf{y}_b)}$$

$$p(\mathbf{y}_a, \mathbf{y}_b) = \mathcal{N} \left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \right)$$

$$\implies p(\mathbf{y}_a | \mathbf{y}_b) = \mathcal{N}(\mathbf{a} + \mathbf{B}\mathbf{C}^{-1}(\mathbf{y}_b - \mathbf{b}), \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T)$$

The result also holds for conditioning with Gaussian errors. This property is important because it means that conditioning can be carried out “analytically”, without a brute force rejection algorithm being employed.

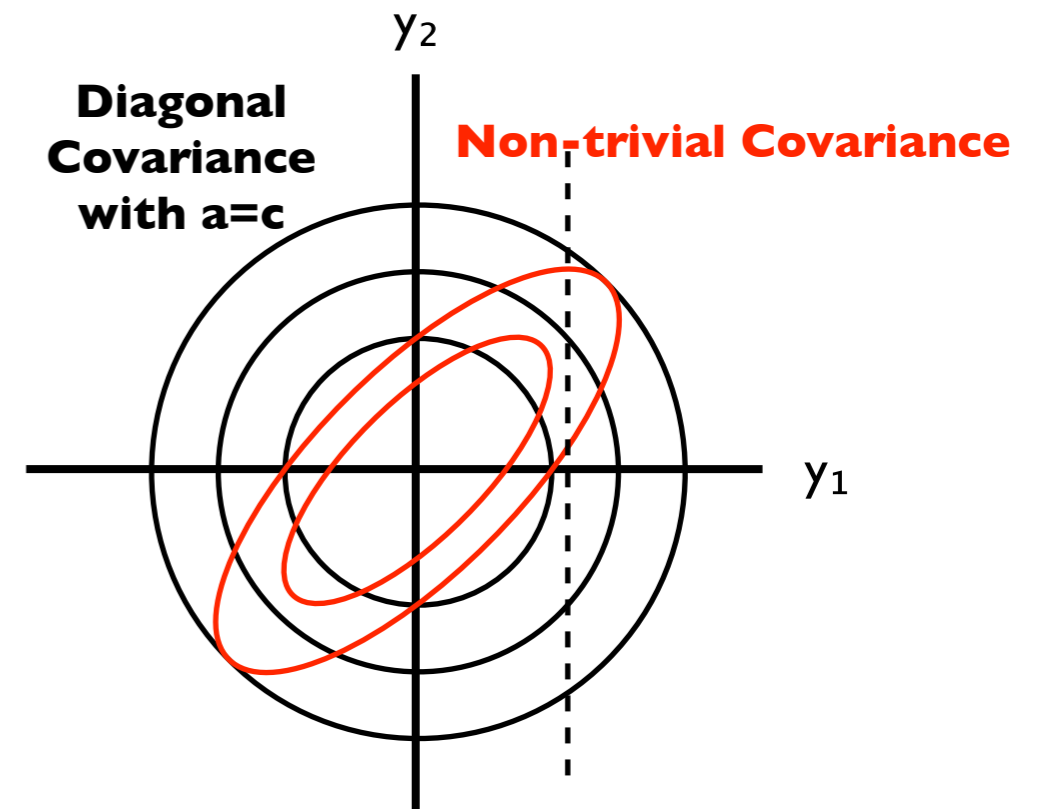
Note, however, that a *matrix inversion* is required for this step. This is one aspect of the “curse of dimensionality” in regression/inverse problems. Ideas on how to deal with this issue are at the cutting edge of current research.

GP Modeling: Basics III

Simple illustration of the conditioning formula in 2 dimensions for a mean-zero Gaussian process:

consts. absorbed in normalization,
since y_a is known

$$\begin{aligned}
 p(y_b|y_a, \Sigma) &= \frac{p(y_a, y_b|\Sigma)}{p(y_a|\Sigma)} \propto \exp -\frac{1}{2} \left[(y_a \ y_b) \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} y_a \\ y_b \end{pmatrix} \right] \\
 &= \exp -\frac{1}{2} [ay_a^2 + 2by_ay_b + cy_b^2] \propto \exp -\frac{1}{2} [2by_ay_b + cy_b^2] \\
 &\propto \exp -\frac{1}{2} \left[\left(y_b^2 + 2\frac{b}{c}y_ay_b + \frac{b^2}{c^2}y_a^2 \right) c \right] \\
 &= \exp -\frac{1}{2} \left[\frac{\left(y_b - \left(-\frac{b}{c}y_a \right) \right)^2}{c^{-1}} \right]
 \end{aligned}$$



Even though the joint distribution of y_1 and y_2 is mean-zero, the conditioned distribution of y_2 is not mean-zero, if the covariance matrix is not diagonal

The Covariance Function I

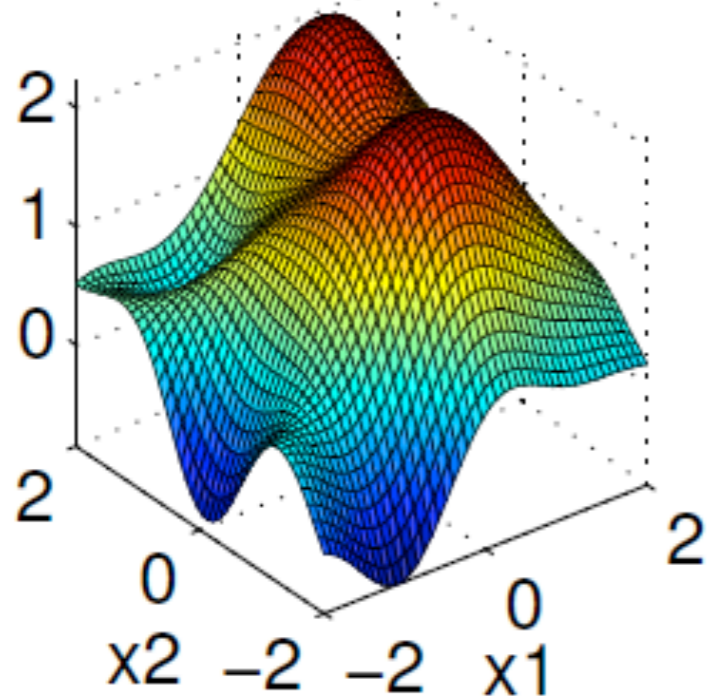
The (*symmetric, positive semi-definite*) covariance function is the key ingredient in GP modeling. Depending on the application, various choices of the covariance function are possible, both in terms of the *form* and the underlying *parameters*

The *squared exponential* form is very common:

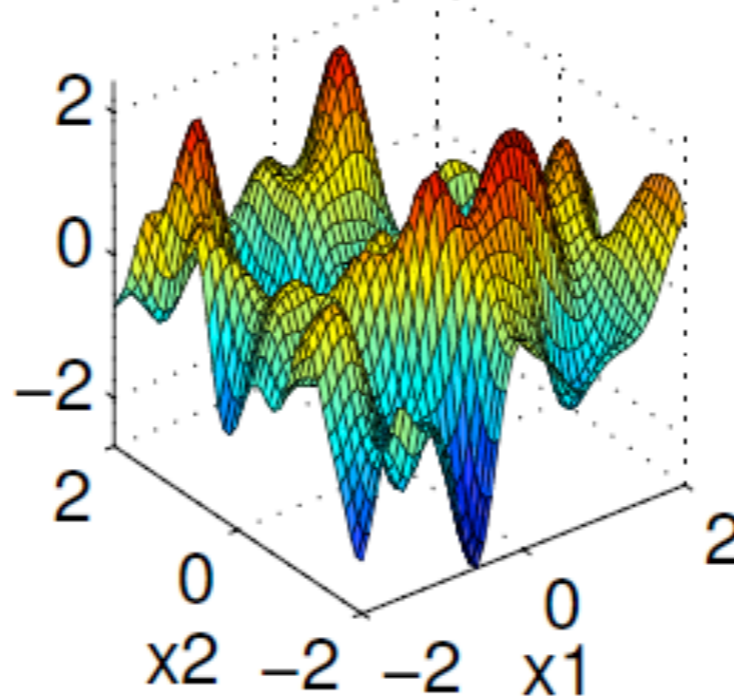
$$k_{SE}(r) = \exp\left(-\frac{r^2}{2l^2}\right)$$

here l defines a characteristic length scale; the realizations are infinitely differentiable (possibly unrealistic?)

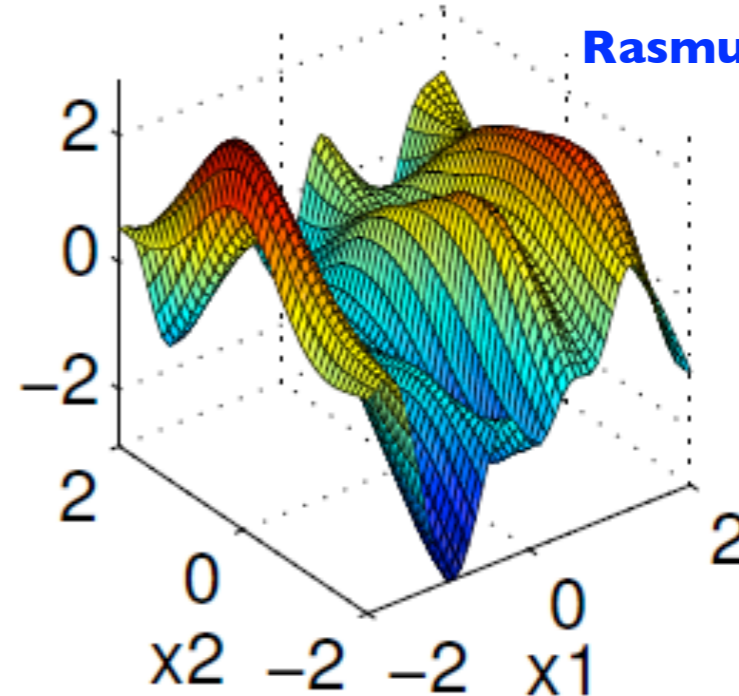
$$l_1 = l_2 = 1$$



$$l_1 = l_2 = 0.32$$



$$l_1 = 0.32, l_2 = 1$$



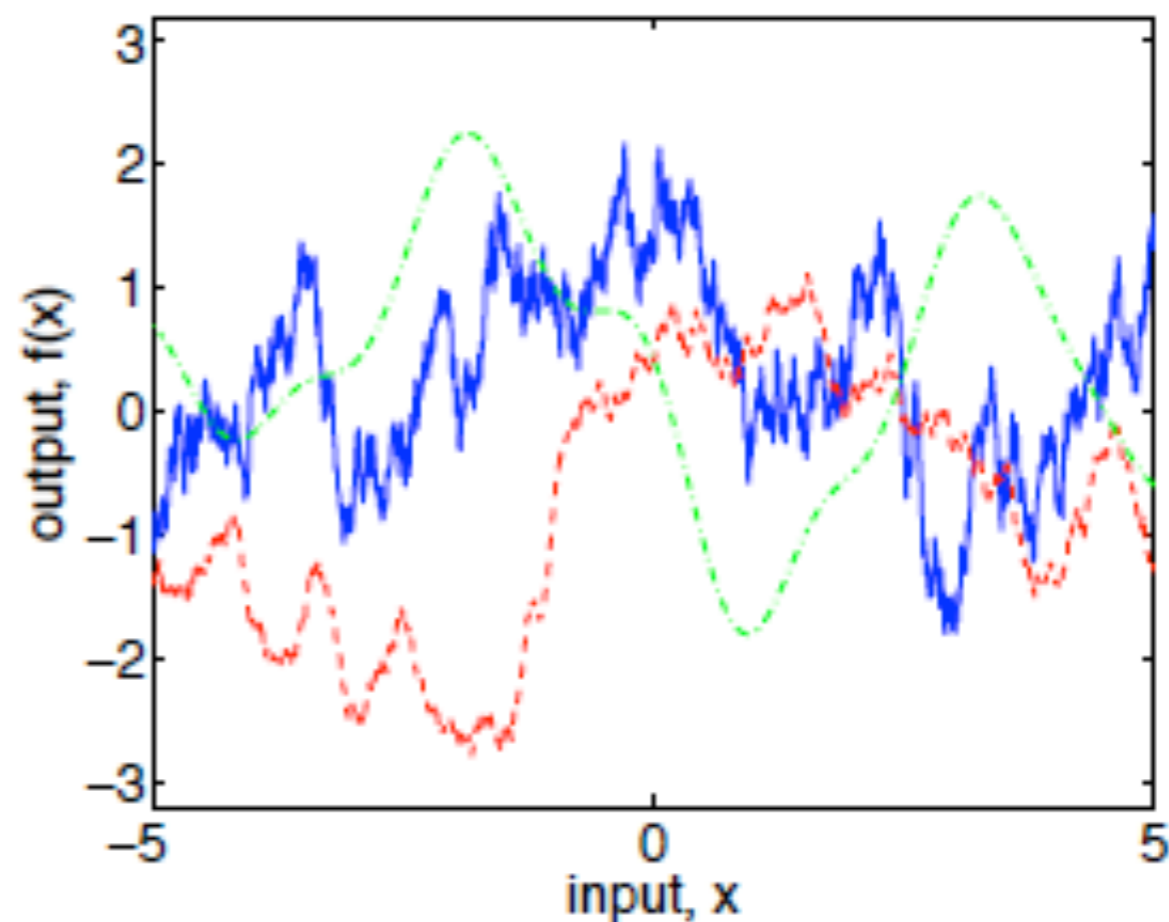
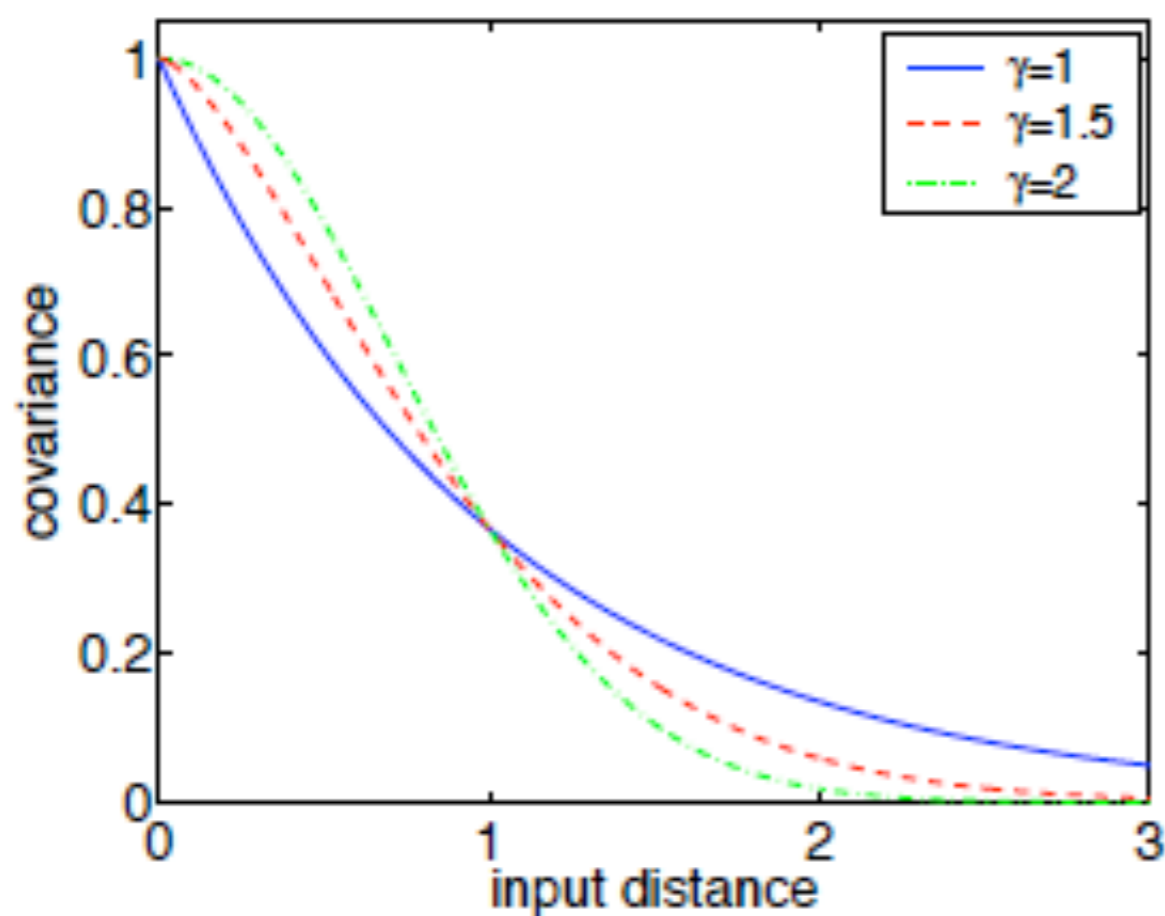
Rasmussen 2006

The Covariance Function II

The *Gamma-exponential* form

$$k_{GE}(r) = \exp\left(-\left(r/l\right)^\gamma\right), \quad 0 < \gamma \leq 2$$

corresponds to an Ornstein-Uhlenbeck process in one dimension for unit exponent, where it yields continuous but non-(MS) differentiable functions, except in the squared-exponential limit (Matern cov. fn. is smoother)



Rasmussen & Williams 2006

Model Selection I

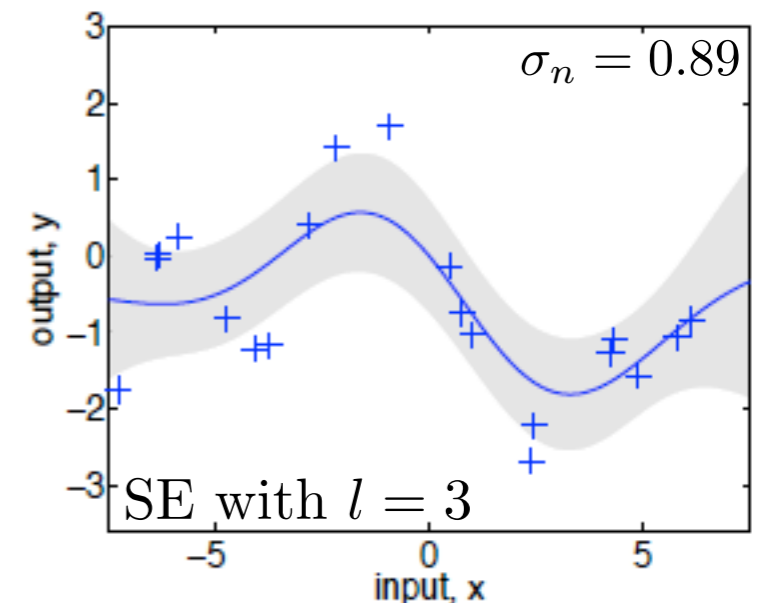
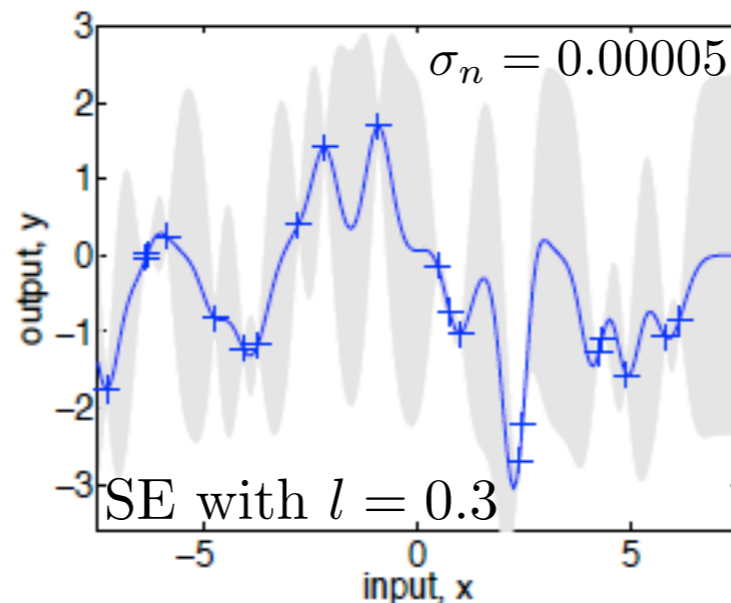
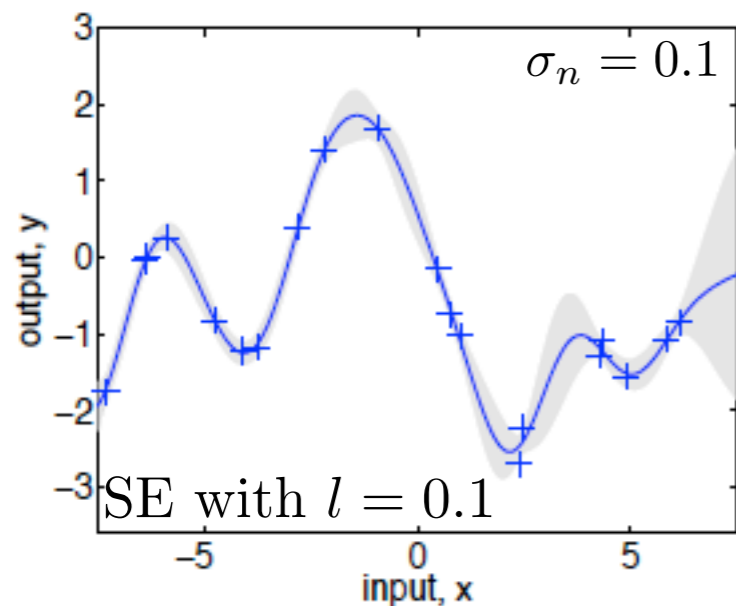
Model selection refers to the choice of covariance function and determining the hyperparameters (e.g., characteristic length scales) of the covariance functions. In a typical situation, one has access only to noisy versions of the GP function draws (“observations”) $y = f(\mathbf{x}) + \epsilon$. If the noise is IID (independent, identically distributed) with variance σ_n^2 , then the prior on the observations is

$$\text{cov}(y_i, y_j) = k(y_i, y_j) + \delta_{ij}\sigma_n^2$$

with the marginal likelihood

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)d\mathbf{f}$$

and since both the integrands are Gaussian, the (log) marginal likelihood can be written down immediately, this is very convenient as we shall see



Rasmussen & Williams 2006

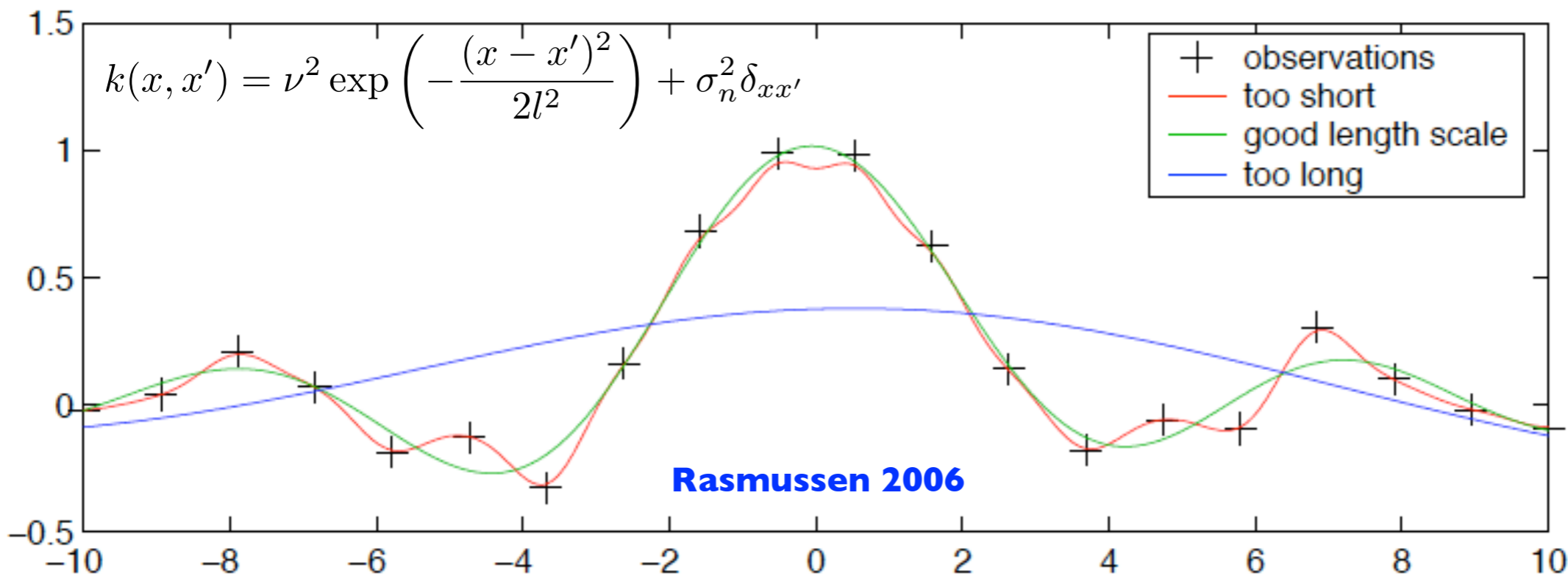
Model Selection II

The log marginal likelihood is ($K_y = K_f + \sigma_n^2 I$)

$$\log p(\mathbf{y} | X, \theta) = -\frac{1}{2} \mathbf{y}^T K_y^{-1} \mathbf{y} - \frac{1}{2} \log |K_y| - \frac{n}{2} \log 2\pi$$

noisy targets training inputs hyperparameters data fit term complexity penalty # of training inputs

The values of the hyperparameters are set by maximizing the marginal likelihood (e.g., using gradient-based optimizers)



Note: though a smaller value of the correlation length appears to provide an almost perfect fit, the marginal likelihood rejects it

Wrap-Up/Issues

Many additional issues show up in actual practice:

- **GPs can be applied directly to data or to weights of basis functions used to represent the data (e.g., to a Principal Components basis, more later)**
- **Robustness of results -- would prefer if answers were not too touchy as a function of choice of covariance function (usually the case)**
- **How good is the naive GP error theory in actual practice? How can one validate the procedure (more on hold-out tests and sub-sampling)? Important in cosmology applications with stringent error control requirements**
- **Use of weighted sampling and iterative procedures allowed within the GP approach, can be extended to covariances (Schneider et al 2008)**
- **Using fast surrogate models is a good way to build confidence in the GP approach and to optimize it**
- **Approximate methods to reduce the N^3 scaling due to the matrix inverse computation (e.g., compact support covariance functions)**
- **Prediction outside the fitted range of a GP is a bad idea**

Application I: $w(z)$ Reconstruction

How to approach the dark energy characterization problem?

(i) Show convincingly it's not a cosmological constant

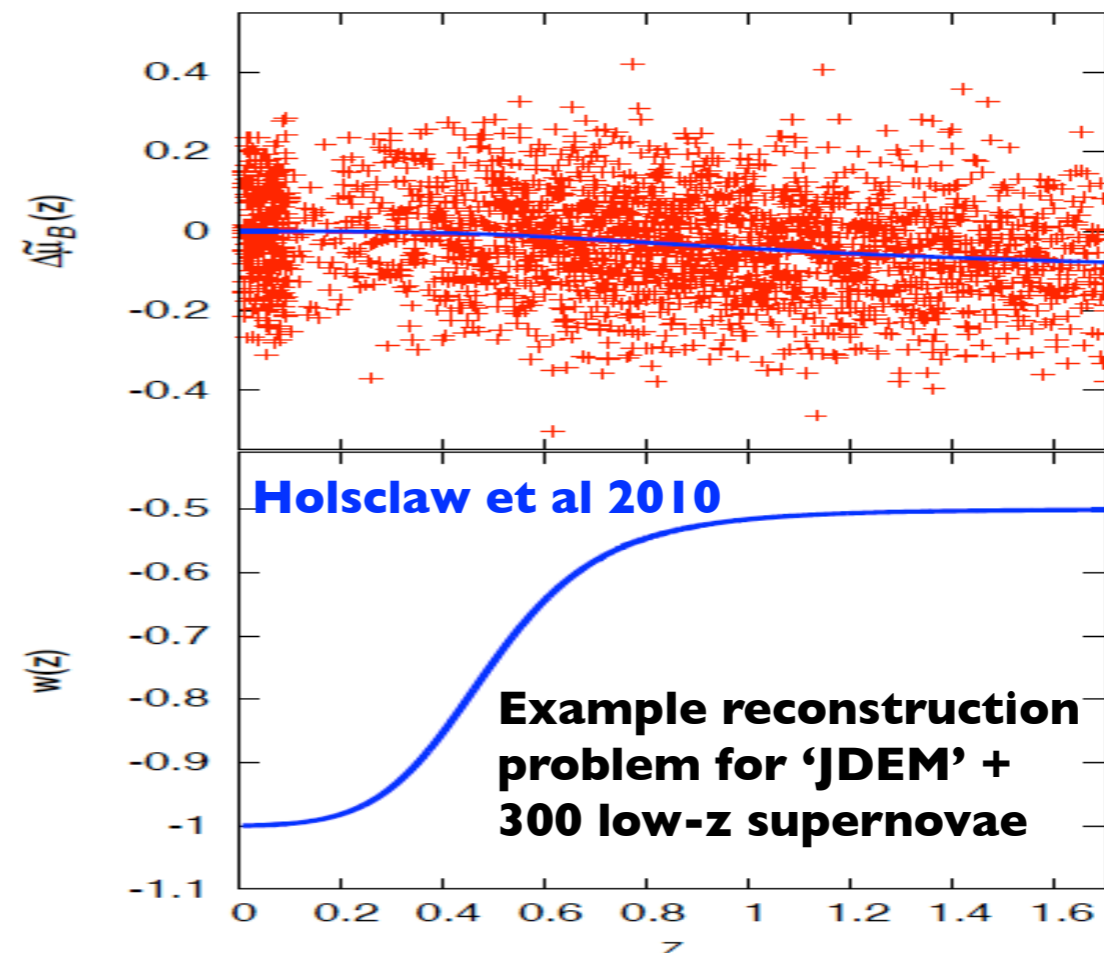
(ii) Given (i), try parameterized models or physically well-motivated ideas (ha!), worries about possible biases due to incompatibility with the data

(iii) Hypothesis testing to attempt to rule out classes of DE models (sort of the next step after (i))

(iv) Reconstruct $w(z)$ directly from data, very hard because of the double integral smoothing operator that must be inverted (smoothing data and then differentiating is a bad idea)

Simple case: Distance modulus for a spatially flat FRW cosmology

$$\mu_B(z) = 25 - 5 \log_{10}(H_0) + 5 \log_{10} \left\{ (1+z)c \int_0^z ds \left[\Omega_m (1+s)^3 + (1 - \Omega_m)(1+s)^3 \exp \left(3 \int_0^s \frac{w(u)}{1+u} du \right) \right]^{-1/2} \right\}$$



GP for $w(z)$ I

Assume a GP for the DE EOS parameter

$$w(u) \sim \mathcal{GP}(-1, K(u, u'))$$

Need to integrate over this in the expression for the distance modulus, where

$$y(s) = \int_0^s \frac{w(u)}{1+u} du$$

The integral of a GP is another GP, and assuming a gamma-exponential form of the covariance

$$y(s) \sim \mathcal{GP} \left(-\ln(1+s), \kappa^2 \int_0^s \int_0^{s'} \frac{\rho^{|u-u'|^\alpha} du du'}{(1+u)(1+u')} \right)$$

A joint GP for the two variables can be constructed:

$$\begin{bmatrix} y(s) \\ w(u) \end{bmatrix} \sim \mathcal{GP} \left(\begin{bmatrix} -\ln(1+s) \\ -1 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

GP for $w(z)$ II

where

$$\Sigma_{11} = \kappa^2 \int_0^s \int_0^{s'} \frac{\rho^{|u-u'|} du du'}{(1+u)(1+u')},$$

$$\Sigma_{22} = \kappa^2 \rho^{|u-u'|},$$

$$\Sigma_{12} = \Sigma_{21} = \kappa^2 \int_0^s \frac{\rho^{|u-u'|} du}{(1+u)}.$$

The mean for $y(s)$ given $w(u)$ is

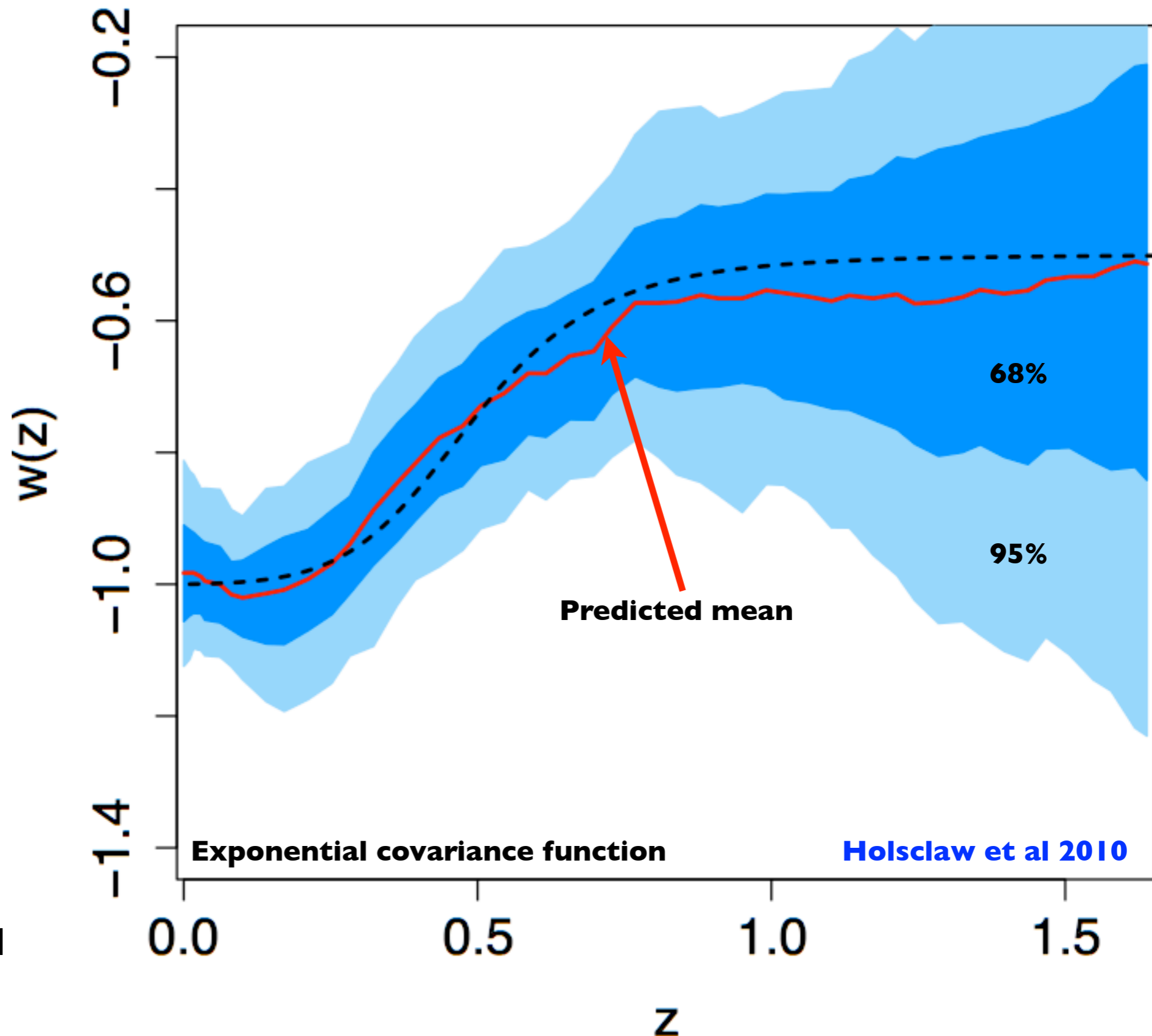
$$\langle y(s) | w(u) \rangle = -\ln(1+s) + \Sigma_{12} \Sigma_{22}^{-1} [w(u) - (-1)]$$

so the expensive double integral Σ_{11} does not have to be computed. Now that the GP model has been constructed one follows the procedure outlined earlier, ‘fits’ to the data, and extracts $w(z)$ (the details of the procedure are actually rather complicated and are given in a forthcoming paper, Holsclaw et al. 2010)

Depending on the assumptions made about the data, we find that smooth infinitely differentiable functions fit the current observations well, but that for simulated data we have to take a much smaller value for the power exponent in the covariance function

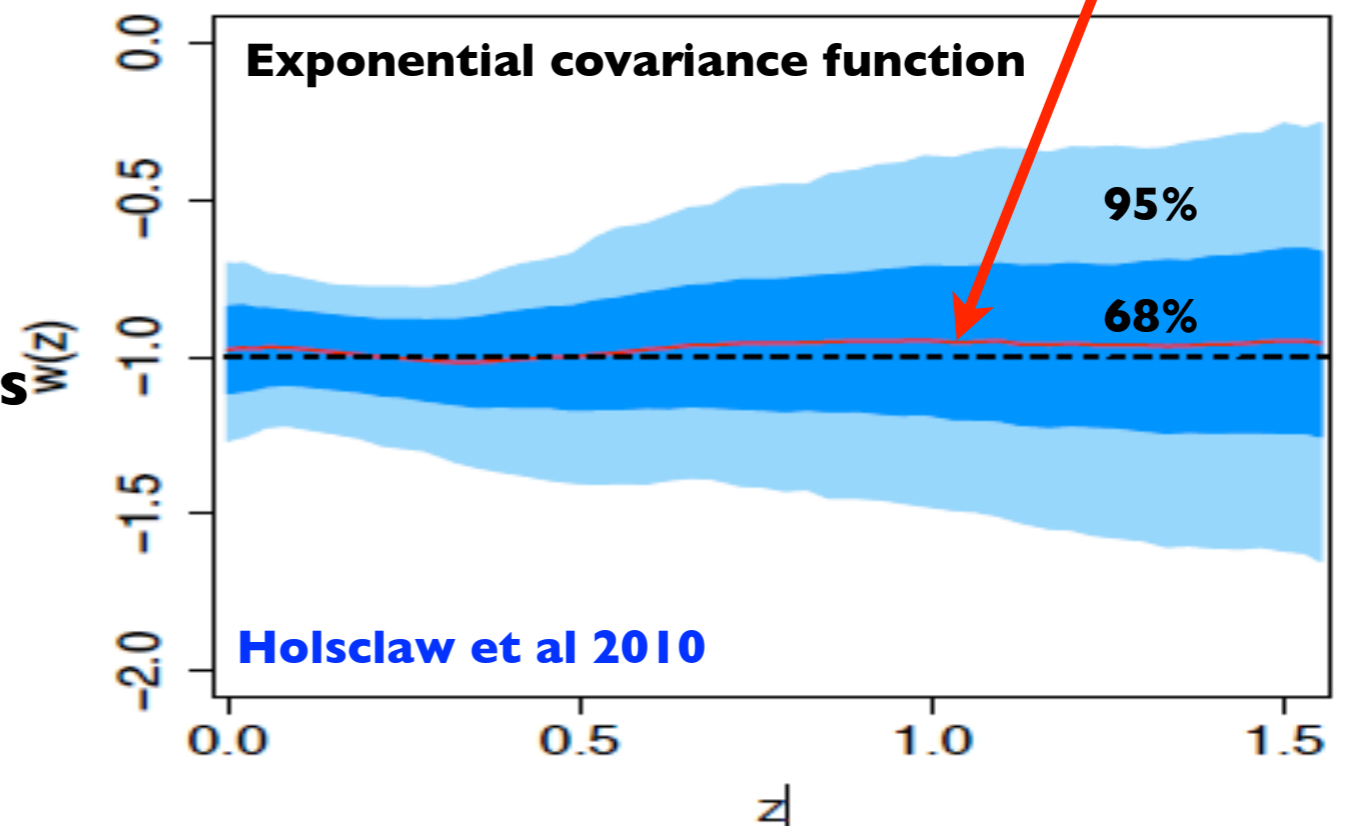
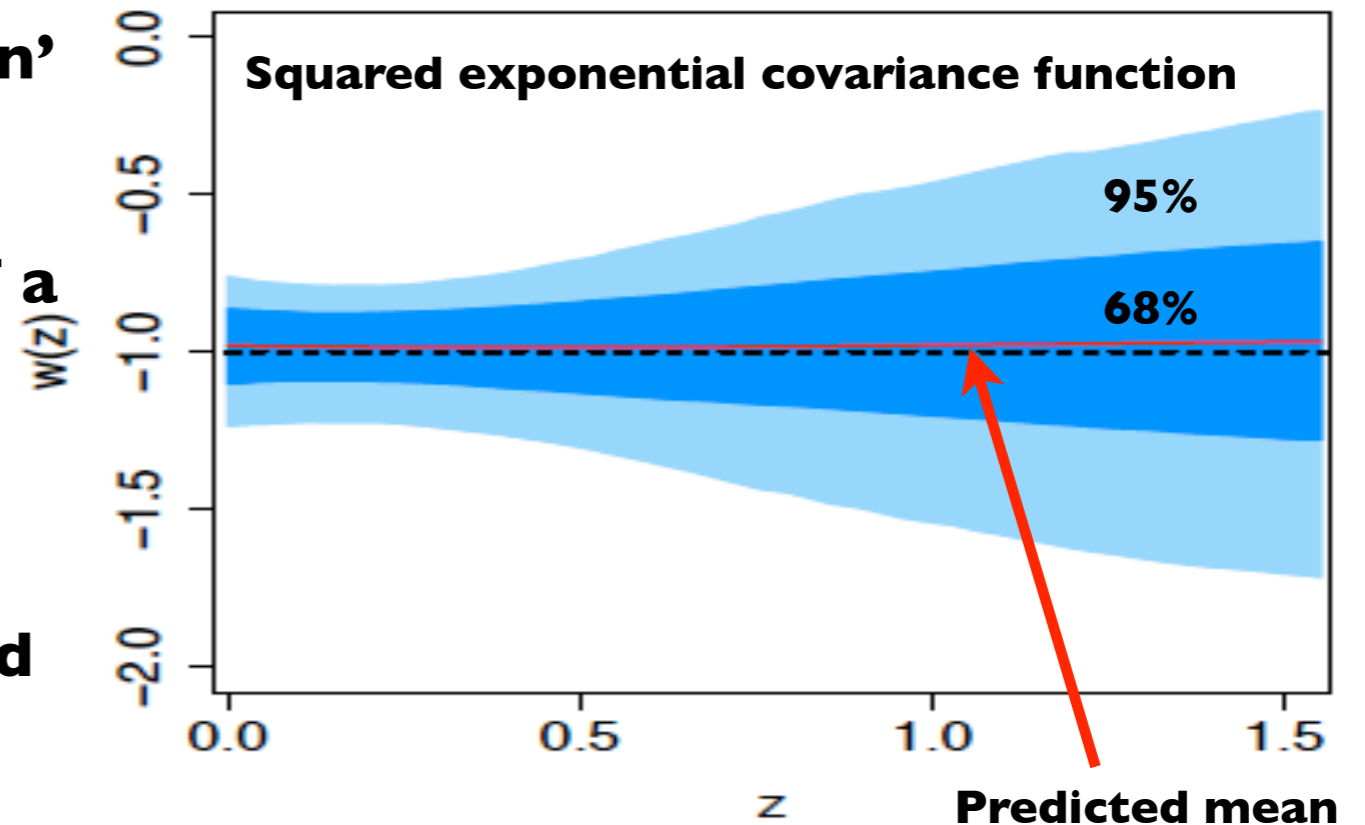
GP Reconstruction on 'Future' Data

- Using 'JDEM' simulated data mocking up smooth, but devious DE EOS histories, we can check if the GP model can correctly reconstruct them
- Results are encouraging as can be seen here with an 'extreme' quintessence model
- Smoother DE EOS histories are recovered well
- GP model readjusts mean starting from -1 to -0.7



GP Reconstruction on Current Data

- Using results from the ‘Constitution’ dataset Hicken et al (2009) and WMAP7 priors, the GP-based reconstruction finds no evidence of a deviation from the cosmological constant
- The GP methodology allows the integration of multiple datasets and sources of information within an overall Bayesian framework, work on adding CMB and BAO data is almost complete
- The GP, as used here, has many useful features: (i) the data is not massaged in any way, (ii) robustness of results to variations of GP hyperparameters can be easily tested, (iii) degeneracies are automatically found during the fitting process



Application II: Cosmic Calibration

Define problem:
identify data,
parameters and ranges,
outputs of interest,
codes

Design simulation campaign
over parameter ranges

Do 64, 128, ..., runs of
simulation code(s)

Statistical code (GPM)

Response surface
for simulation code

Calibration
distributions

Model
inadequacy

Predictive
distributions

Cosmic calibration, an
interlocking five-step process:

- Determine **optimal** simulation campaign
- Run simulations at specified parameter values
- Estimate response surface (**emulation**)
- Combine with observations via **MCMC** to determine parameters -- the **calibration** process
- Make new **predictions** with the calibrated emulator

Observed data

Heitmann et al 2006, Habib et al 2007

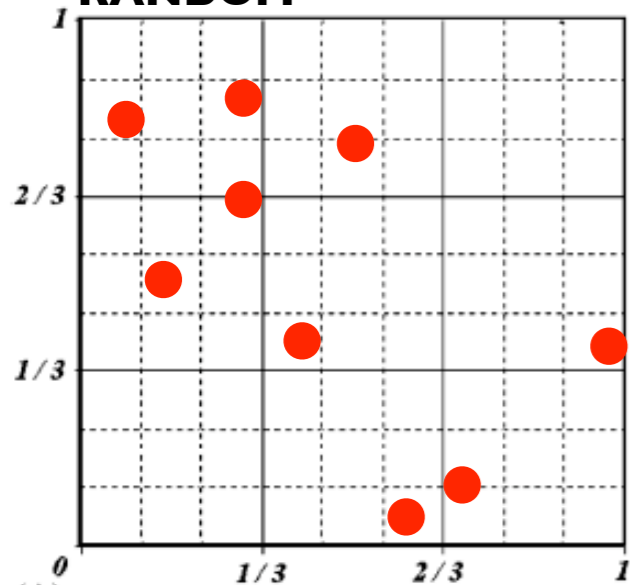
The Process (e.g., Coyote Universe project)

- **Decide if a certain calibration problem is feasible, think through number of simulations, error restrictions, number of variables, etc. (e.g., much easier to do a very large number of CMB runs versus, say, cluster physics runs)**
- **Statisticians generate sampling scheme (days?)**
- **Theorists and statisticians test simple surrogate model to check that the overall strategy will work (e.g., use Halofit to generate $P(k)$, build GP-based emulator, and do error tests)**
- **Iterate sampling strategy until satisfied that errors are controlled to the levels required**
- **Theorists run sufficiently accurate simulations and generate outputs (**months/year(s)**), statisticians twiddle thumbs (or play around with subset of output for quality control tests)**
- **Statisticians generate emulator (day), emulator reduces each forward model evaluation time from hours/days to fraction of a second**
- **Run MCMC with emulator against data to obtain posterior fits (hours)**
- **If required make predictions for other observables (trivial)**

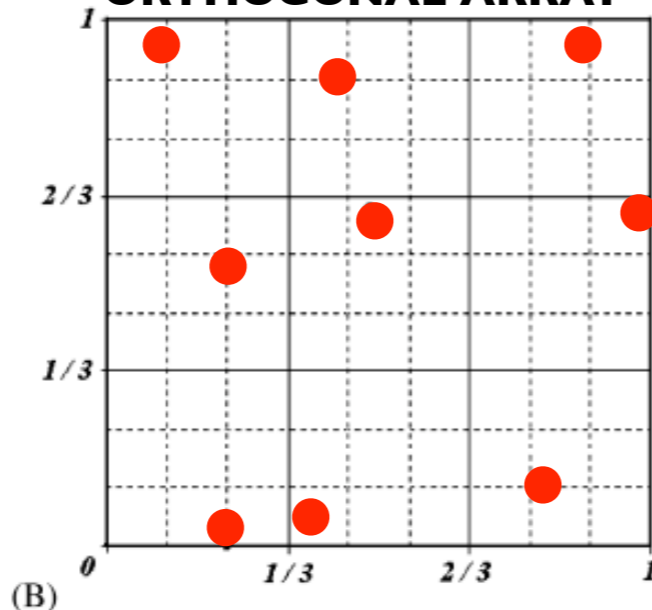
Sampling Designs

Step I

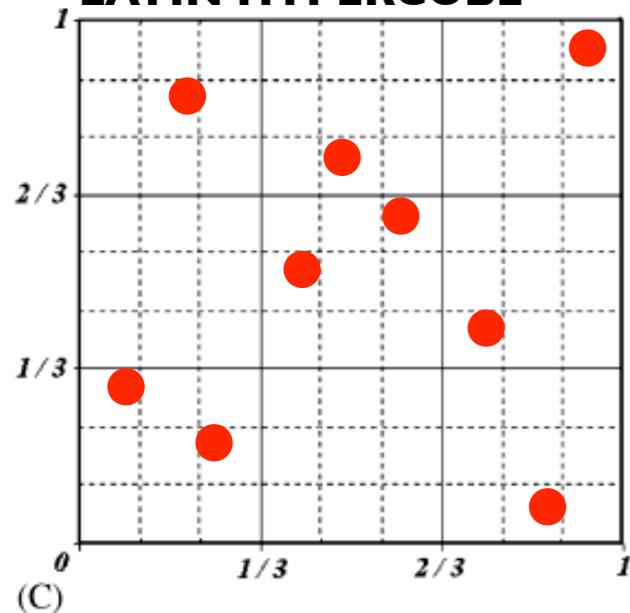
RANDOM



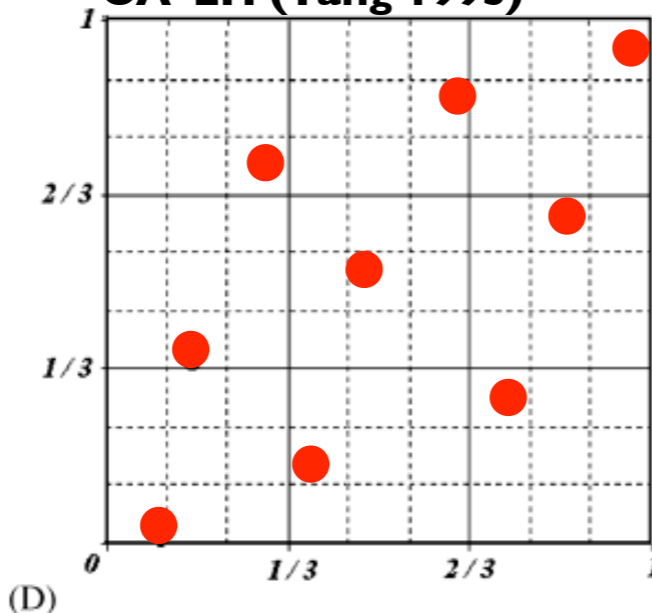
ORTHOGONAL ARRAY



LATIN HYPERCUBE

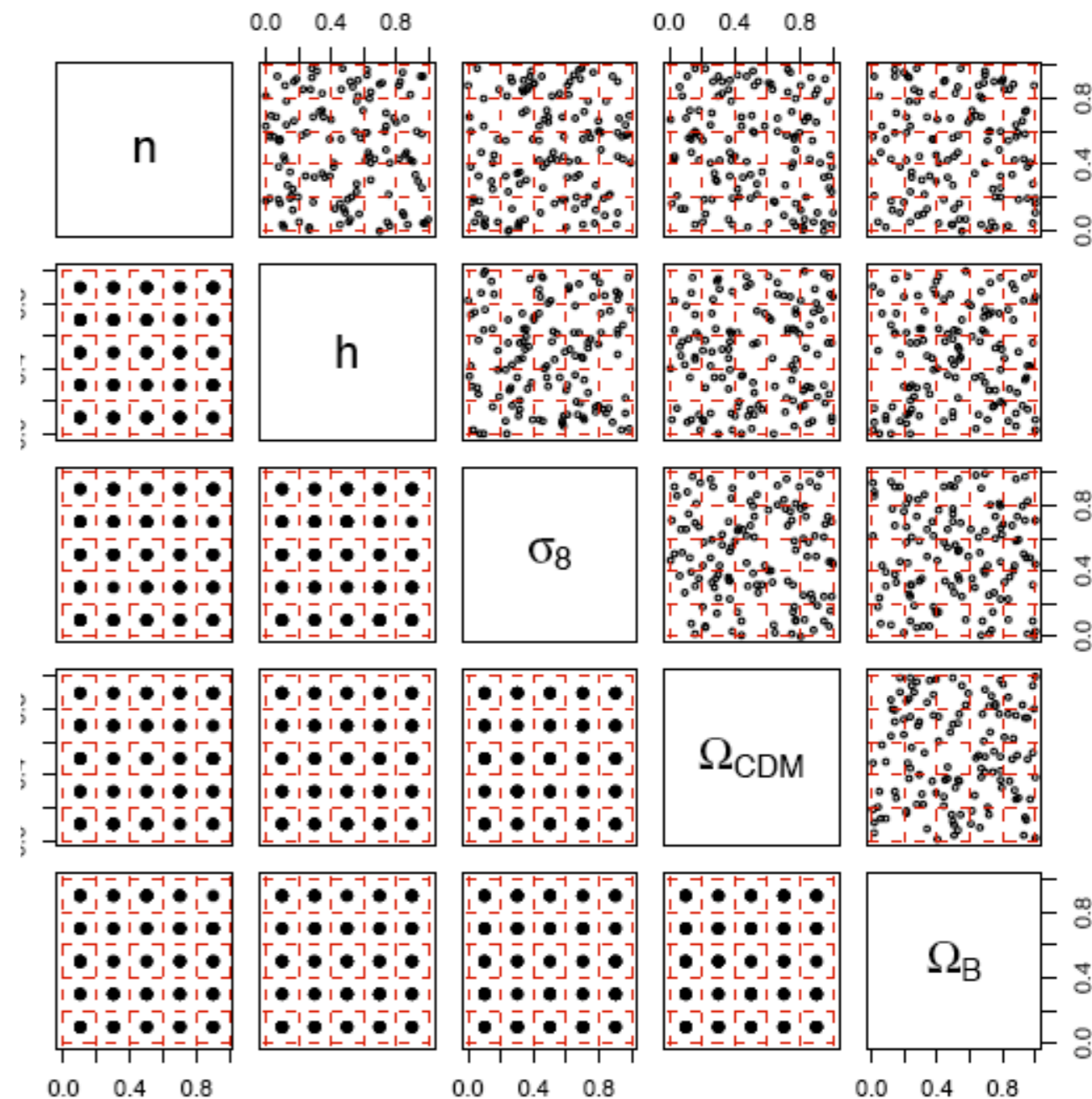


OA-LH (Tang 1993)



Sandor & Andras 2003

Strive for “equidistribution” property over the sampling space, best approach when ignorant of functional variation, well-suited to GPMs.



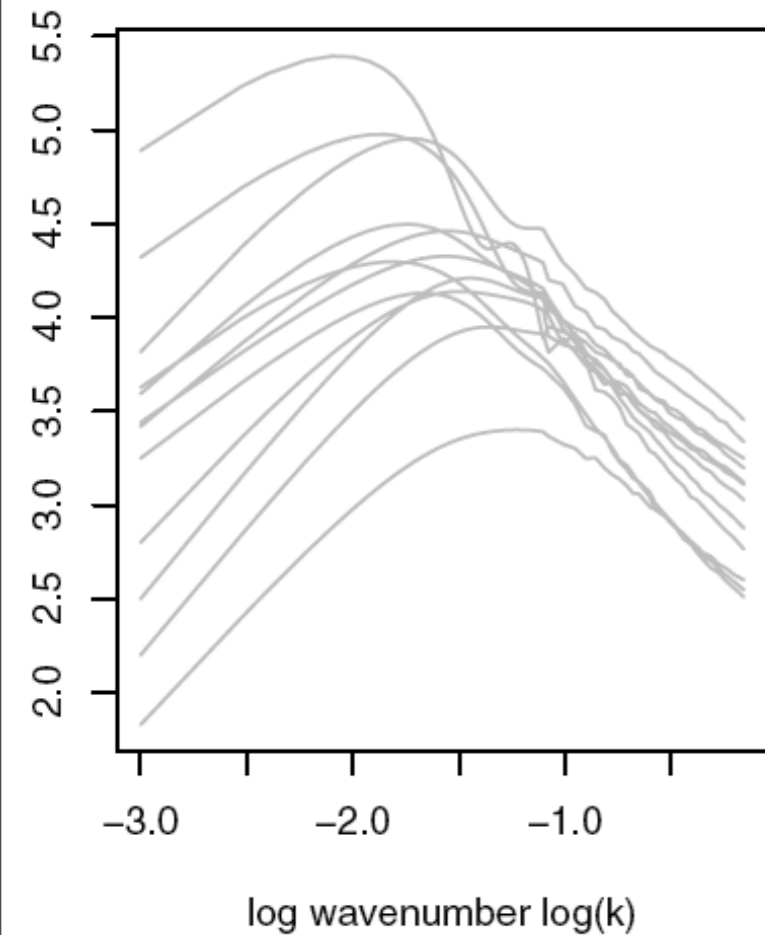
Practical 128 point, 5-level, strength-2-based design
[level=#variable slices, strength=(lower) dimension to be sampled, #columns=#variables, #rows=#trials]

Basis Representation of Simulated Spectra

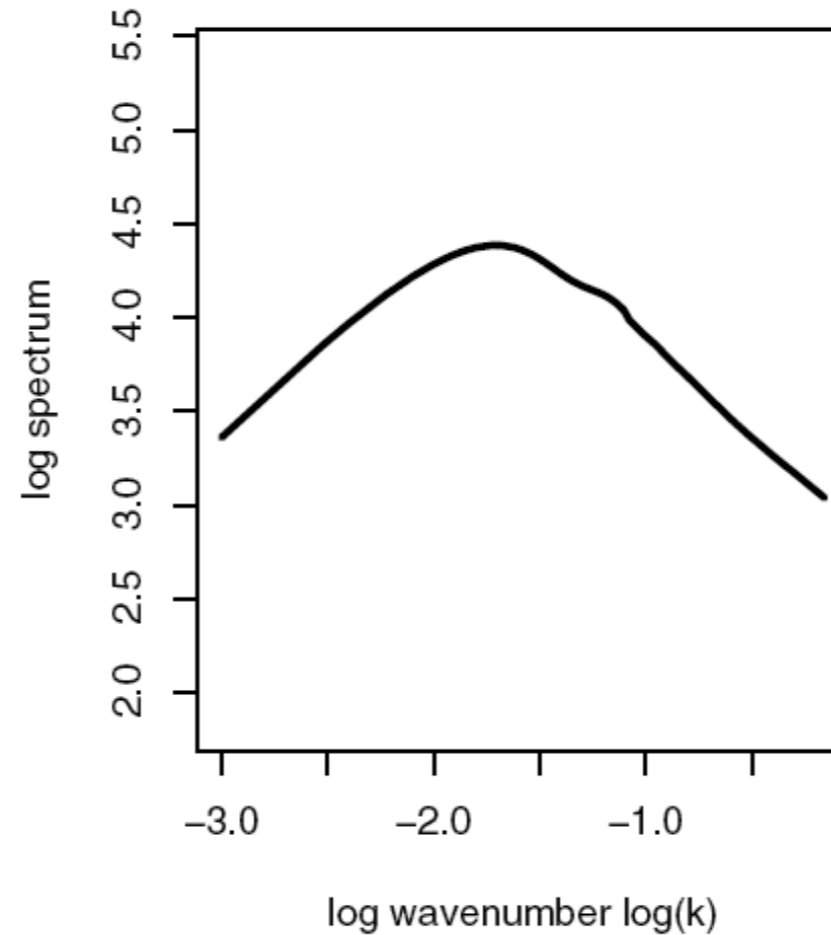
Step II

P(k) example

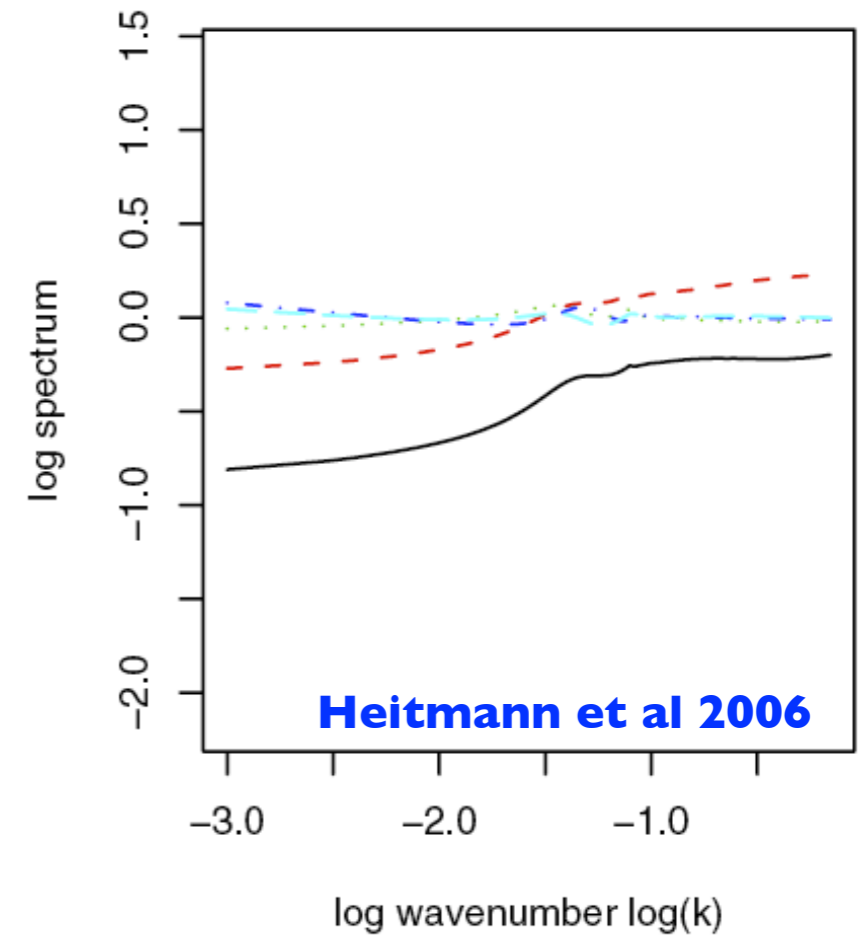
SIMULATIONS



MEAN



FIRST 5 PCs



Mean-adjusted Principal Component Representation

$$\eta(k; \theta) = \sum_{i=1}^{p_\eta} \phi_i(k) w_i(\theta) + \epsilon,$$

**COSMOLOGICAL/MODELING
PARAMETERS**

**PC BASIS
FUNCTIONS**

GP WEIGHTS

$$\theta \in [0, 1]^{p_\theta}$$

**STANDARDIZED
PARAMETER
DOMAIN**

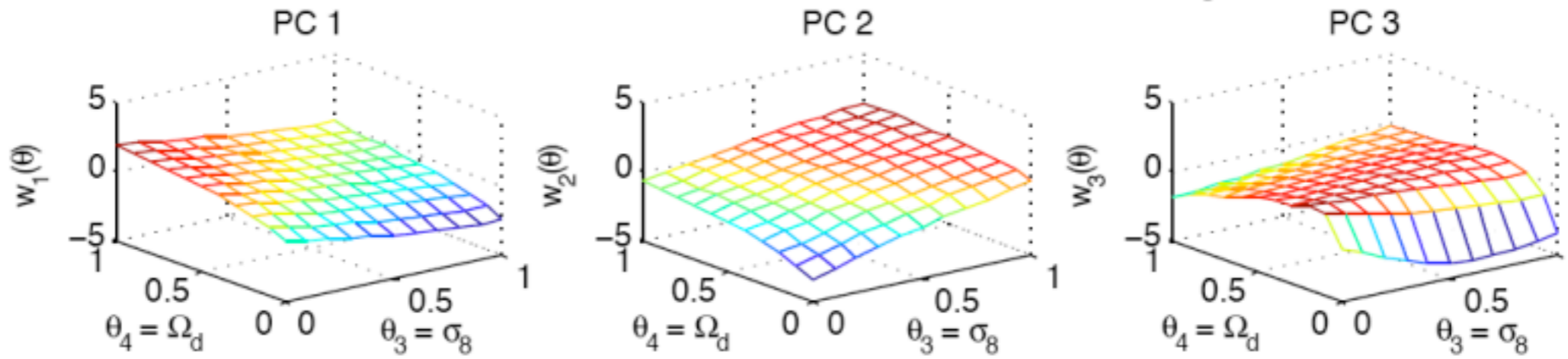
Gaussian Process Modeling

Step III

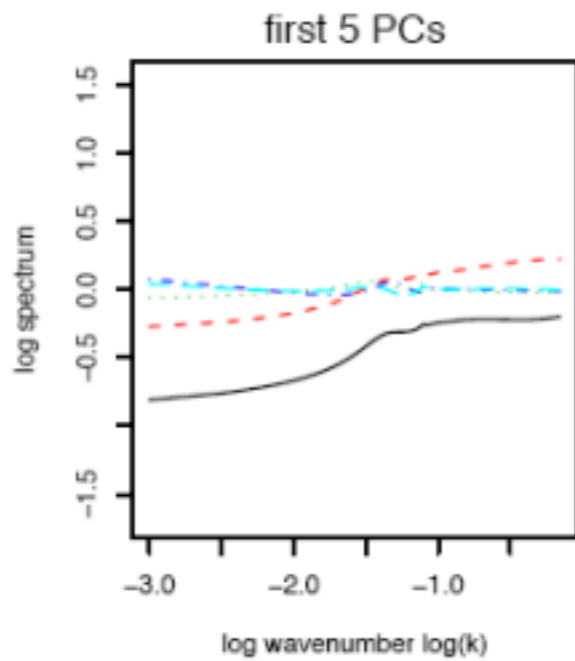
Gaussian process (GP) models are used to estimate the weights $w_j(\theta)$ at untried settings

$$\hat{\eta}(\theta; k) = \sum_{j=1}^{p_\eta} w_j(\theta) \phi_j(k)$$

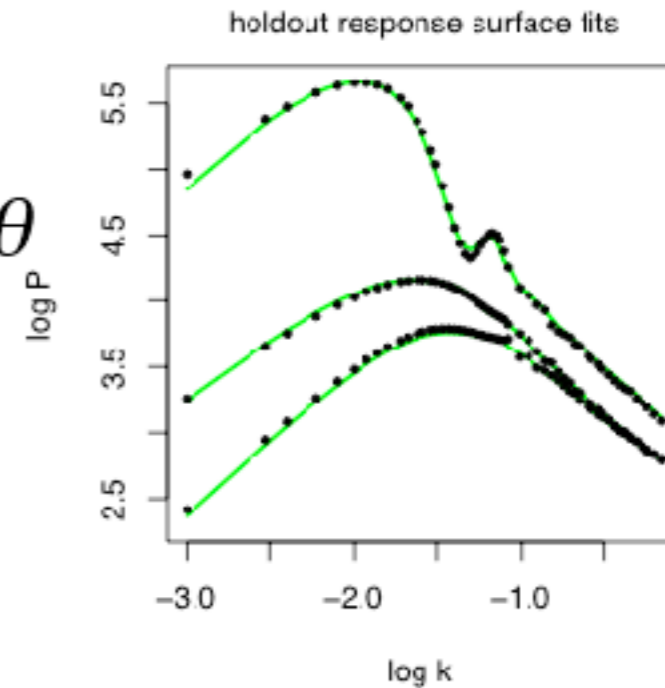
GP models $w_j(\theta)$



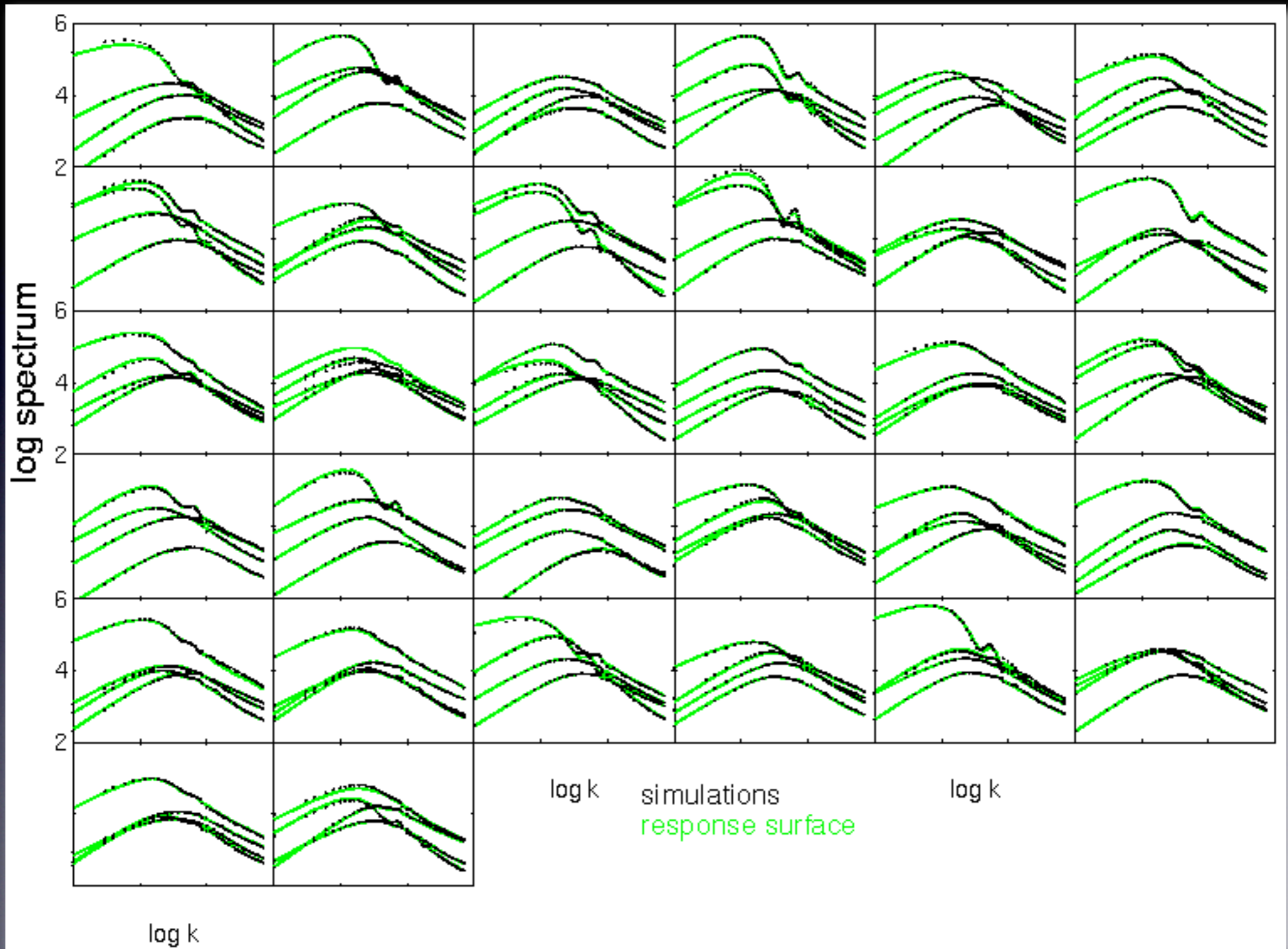
Bases $\phi_j(\theta)$



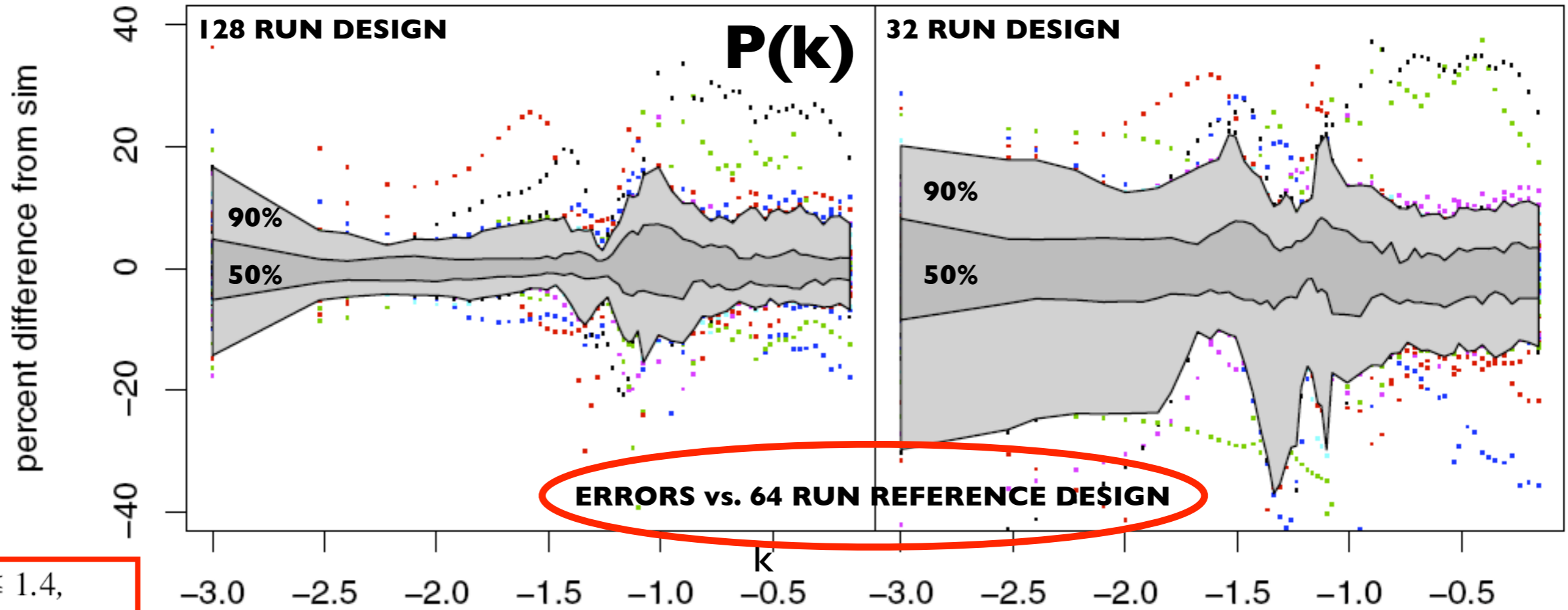
Prediction at new θ



HOLDOUT TEST

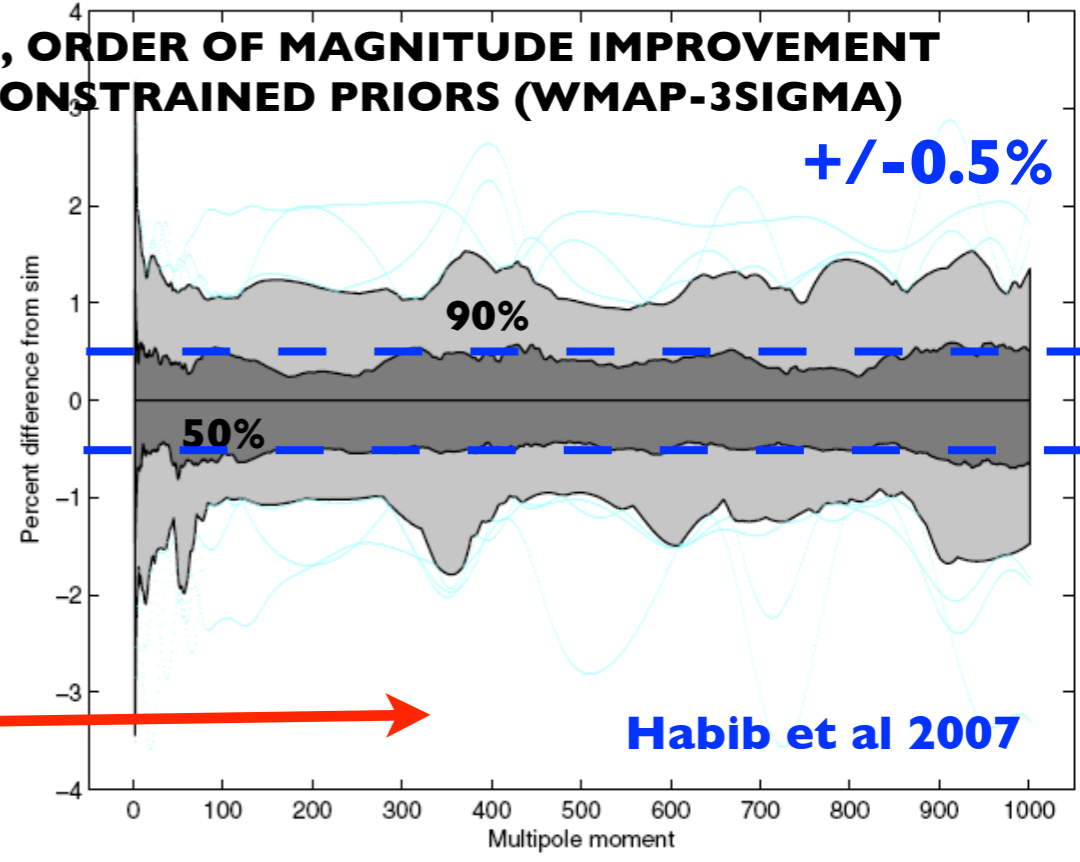
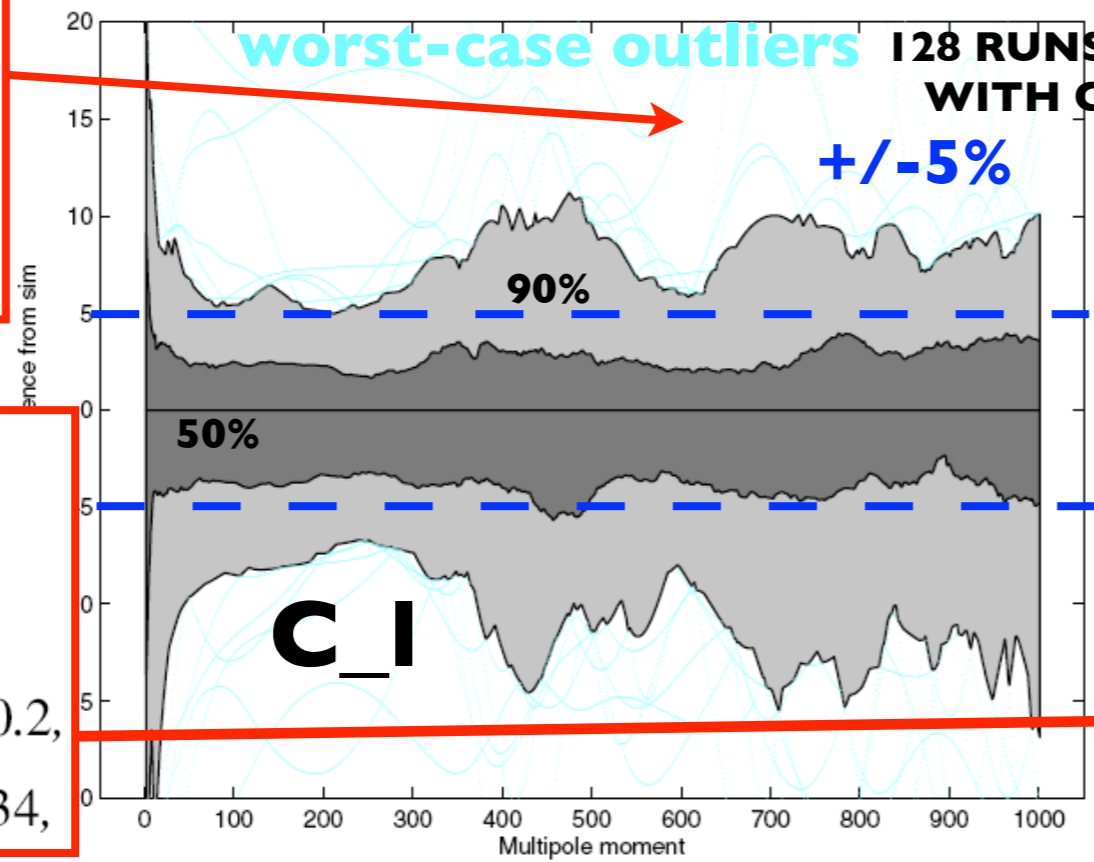


More on Convergence: Post Hold-Out



$0.8 \leq n \leq 1.4,$
 $0.5 \leq h \leq 1.1,$
 $0.6 \leq \sigma_8 \leq 1.6,$
 $0.05 \leq \Omega_{\text{CDM}} \leq 0.6,$
 $0.02 \leq \Omega_b \leq 0.12.$

$0.85 \leq n \leq 1.25,$
 $0.6 \leq h \leq 0.9,$
 $0.6 \leq \sigma_8 \leq 1.2,$
 $0.06 \leq \Omega_{\text{CDM}} h^2 \leq 0.2,$
 $0.018 \leq \Omega_b h^2 \leq 0.034,$



Results: CMB + P(k)

(simulated data plus 128/128 runs, 6 parameters)

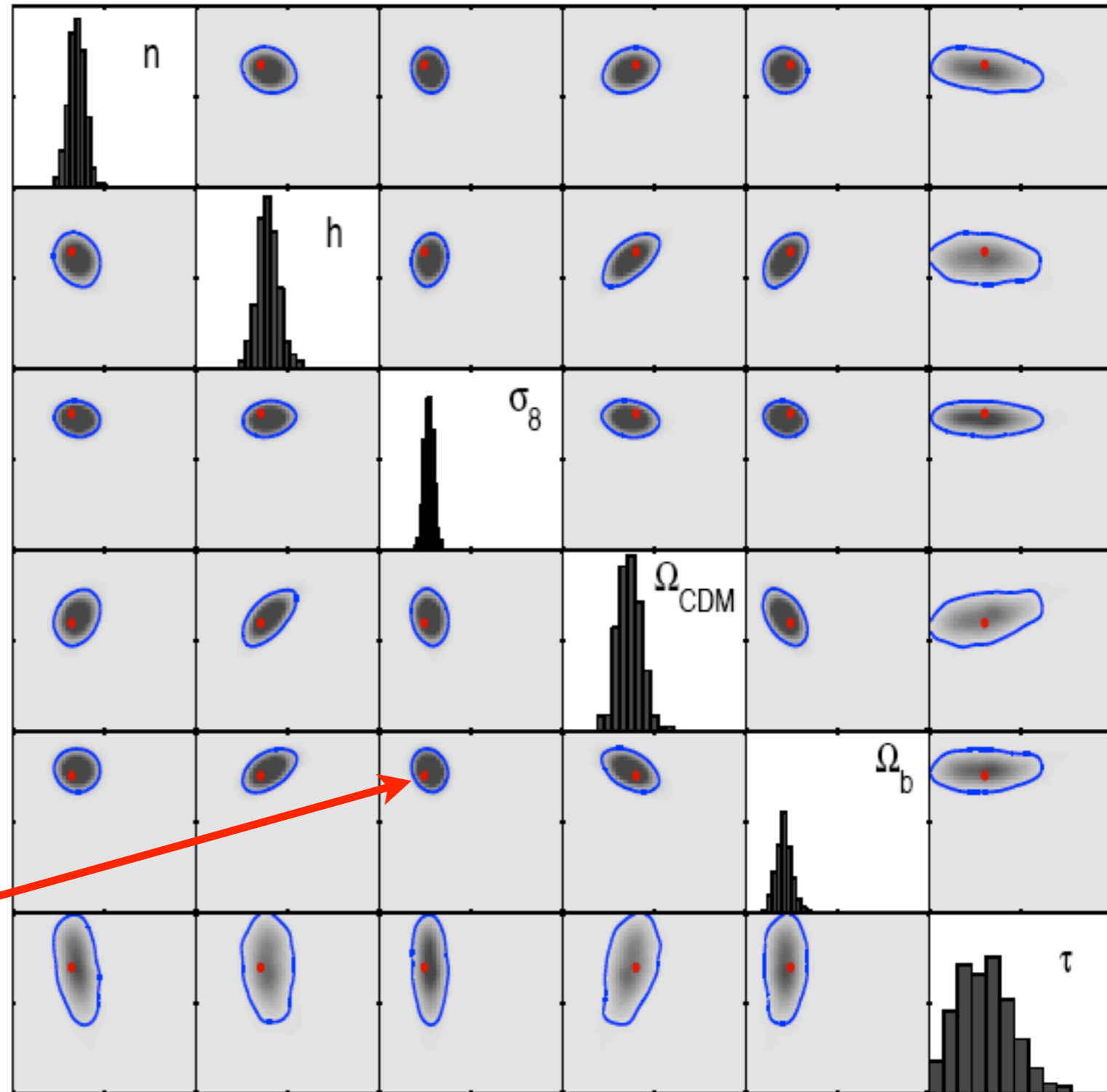
Step IV

Estimate parameters: explore posterior distribution via MCMC taking emulation errors into account.

Framework simultaneously handles C_l and P(k) emulation; other inputs can be easily added. Here P(k) simulated data was “SDSS main sample”.

Very good results from a small number of base simulations.

Target points correspond to the values at which the simulated data were generated.



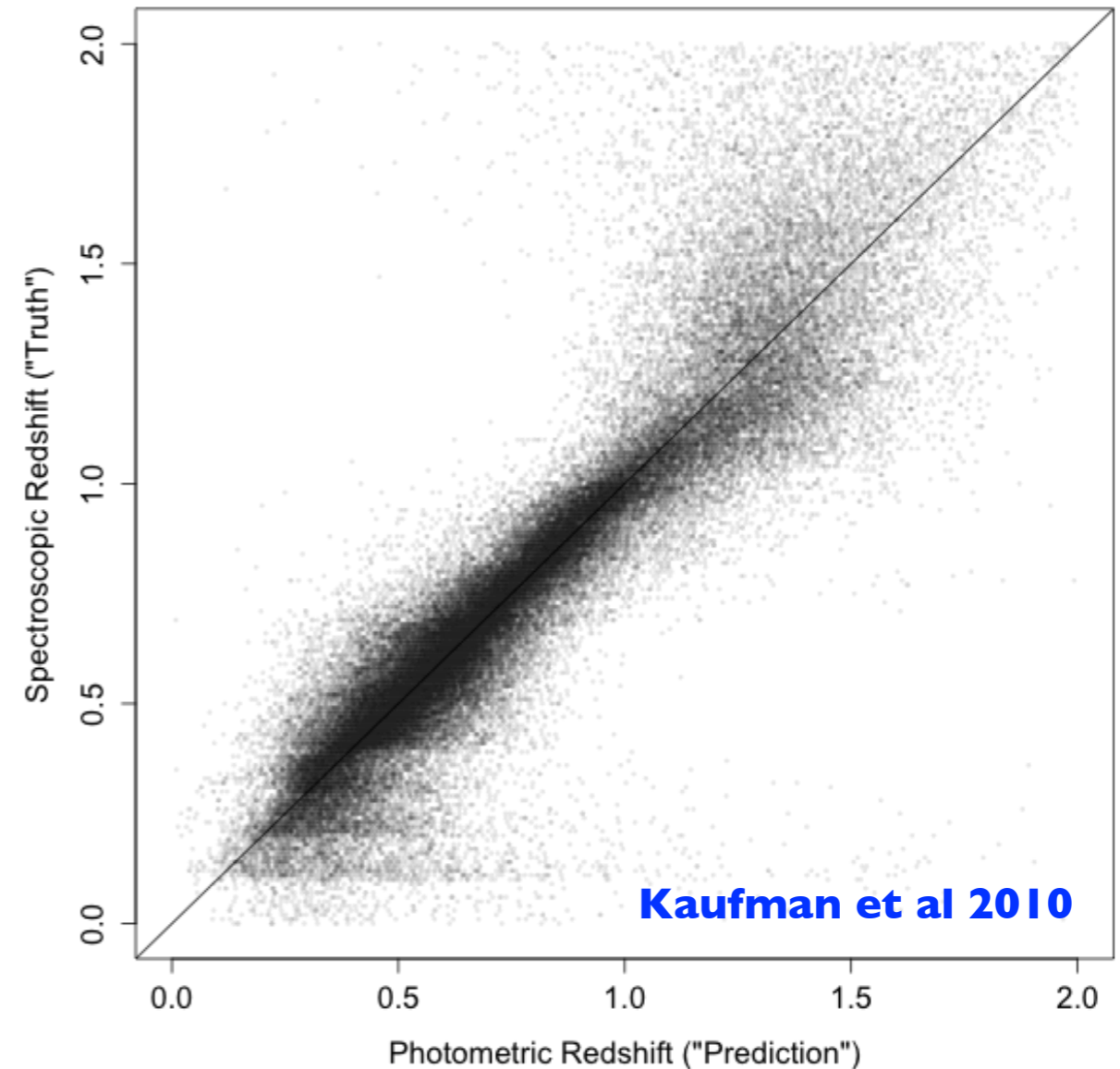
Application III: GPs and Photo-Zs

As a final example, we mention using GPs for photo-z estimation. This is a complex and difficult problem and we have approached it by modifying the standard GP technology in several ways:

(i) used covariance functions with compact support, so that sparse matrix algorithms can be employed

(ii) the correlation range in each dimension is varied and a constraint is imposed on these ranges to enforce a minimum level of sparsity in the covariance matrix

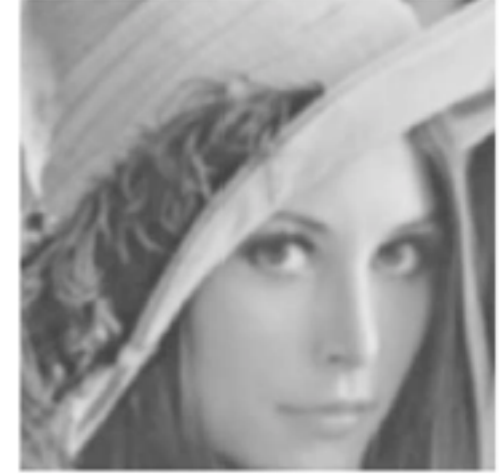
(iii) a (regression) model is proposed for the GP mean, rather than assuming it to be a scalar, the resulting decrease in the correlation length offsets some of the loss of performance in using a compactly supported covariance



Preliminary results from a subsample of a simulated DES dataset

GP Resources

- www.GaussianProcess.org
- **C.E. Rasmussen & C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006, available at www.GaussianProcess.org/gpml**
- **D.J.C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2003**
- **D. Higdon et al in the Oxford Handbook of Applied Bayesian Analysis, edited by A. O'Hagan & M. West, Oxford University Press, 2010**
- **M. Kennedy & A. O'Hagan, Bayesian Calibration of Computer Models (with discussion), J. Roy. Stat. Soc. 68, 425 (2001)**



Proof that progress occurs in interpolation methods (Bleau, Thevenaz, & Unser 2004)