SANDIA REPORT

SAND2007-6712xxxx
Unlimited Release
Printed October 2007

Computationally Efficient Bayesian Inference for Inverse Problems

Youssef M. Marzouk, Habib N. Najm, and Larry A. Rahn

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online ordering:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2007-6712xxxx
Unlimited Release
Printed October 2007

Computationally Efficient Bayesian Inference for Inverse Problems

Youssef M. Marzouk, Habib N. Najm, and Larry A. Rahn
Biological and Energy Sciences Center
Sandia National Laboratories
PO Box 969
Livermore, CA 94551 USA
ymarzou@sandia.gov (corresponding author)
hnnajm@sandia.gov, rahn@sandia.gov

Abstract

Bayesian statistics provides a foundation for inference from noisy and incomplete data, a natural mechanism for regularization in the form of prior information, and a quantitative assessment of uncertainty in the inferred results. Inverse problems—representing indirect estimation of model parameters, inputs, or structural components—can be fruitfully cast in this framework. Complex and computationally intensive forward models arising in physical applications, however, can render a Bayesian approach prohibitive. This difficulty is compounded by high-dimensional model spaces, as when the unknown is a spatiotemporal field.

We present new algorithmic developments for Bayesian inference in this context, showing strong connections with the forward propagation of uncertainty. In particular, we introduce a stochastic spectral formulation that dramatically accelerates the Bayesian solution of inverse problems via rapid evaluation of a surrogate posterior. We also explore dimensionality reduction for the inference of spatiotemporal fields, using truncated spectral representations of Gaussian process priors. These new approaches are demonstrated on scalar transport problems arising in contaminant source inversion and in the inference of inhomogeneous material or transport properties.

We also present a Bayesian framework for parameter estimation in stochastic models, where intrinsic stochasticity may be intermingled with observational noise. Evaluation of a likelihood function may not be analytically tractable in these cases, and thus several alternative Markov chain Monte Carlo (MCMC) schemes, operating on the product space of the observations and the parameters, are introduced.

Acknowledgements

This research was supported in part by an appointment to the Sandia National Laboratories Truman Fellowship in National Security Science and Engineering, sponsored by Sandia Corporation (a wholly owned subsidiary of Lockheed Martin Corporation) as operator of Sandia National Laboratories under US Department of Energy contract number DE-AC04-94AL85000. Y. M. Marzouk was a Truman Fellowship award recipient from October 2004 through September 2007. Support was also provided by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences.

Contents

1 Stochastic spectral methods for Bayesian inference in inverse problems 7
1.1 Introduction 7
1.2 Formulation 9
1.2.1 Bayesian inference for inverse problems 9
1.2.2 Polynomial chaos expansions 11
1.2.3 Efficient evaluation of the posterior: Sampling from the prior 12
1.2.4 Efficient evaluation of the posterior: Sampling from alternate distributions 13
1.2.5 Efficient evaluation of the posterior: Change of variables and MCMC 14
1.2.6 Decomposition of parameter space 16
1.3 Results 17
1.3.1 Source inversion under diffusive transport 17
1.3.2 Stochastic spectral solution of the forward problem 19
1.3.3 Posterior densities 20
1.3.4 Posterior sampling and speedup 21
1.3.5 Markov chain Monte Carlo 22
1.3.6 Choice of transformation and PC basis 23
1.4 Conclusions 26

2 Bayesian inference of spatial fields with Gaussian process priors 51
2.1 Introduction 51
2.2 Formulation 52
2.2.1 Gaussian processes 52
2.2.2 Karhunen-Loève expansion 53
2.2.3 Bayesian approach to inverse problems 55
2.2.4 Dimensionality reduction in inverse problems 56
2.2.5 Polynomial chaos acceleration 57
2.2.6 Gaussian processes, K-L expansions, RKHS, and regularization ... 61
2.3 Numerical implementations and results 62
2.3.1 Inverse problem setup 63
2.3.2 Grid-based inversion 64
2.3.3 Reduced-dimensionality inversion 67
2.3.4 Reduced-dimensionality inversion with polynomial chaos acceleration 70
2.3.5 Data length scales 73

3 MCMC schemes for parameter estimation in stochastic models 107
3.1 Forms of the posterior density 107
3.2 Markov chain Monte Carlo 109
3.3 Discussion and comparison to other schemes 110
3.4 An alternative robust scheme 111

References 113

Appendix: Reproducing kernel Hilbert spaces **121**

1 Stochastic spectral methods for Bayesian inference in inverse problems

1.1 Introduction

Inverse problems, broadly defined, arise from indirect observations of a quantity of interest [26, 50]. A physical system may be described by a *forward model*, which predicts some measurable features of the system given a set of parameters. The corresponding inverse problem consists of inferring these parameters from a set of observations of the features.

The simplicity of this definition belies many fundamental challenges. In realistic applications, data is almost always noisy or uncertain. Also, the forward model may have limitations on its predictive value; i.e, it may be an imperfect or imprecise model of the physical system. Furthermore, as highlighted in [50], inverse problems are often *non-local* and/or *non-causal*. In a forward model, solution values usually depend only on neighboring regions of space and affect only future values in time. Inverting these models, however, may (implicitly) require time-reversal or deconvolution. In mathematical terms, these properties render inverse problems ill-posed. No feasible parameters may match the observed data (existence), or a multiplicity of model parameters may fit the data (uniqueness). Small errors in measurement can lead to enormous changes in the estimated model (stability).

The Bayesian setting for inverse problems offers a rigorous foundation for inference from noisy data and uncertain forward models, a natural mechanism for incorporating prior information, and a quantitative assessment of uncertainty in the inferred results [74, 89]. Indeed, the output of Bayesian inference is not a single value for the model parameters, but a probability distribution that summarizes all available information about the parameters. From this *posterior* distribution, one may estimate means, modes, and higher-order moments, compute marginal distributions, or make additional predictions by averaging over the posterior.

Bayesian approaches to inverse problems have seen much recent interest [74, 7, 50], with applications ranging from geophysics [43, 69] and climate modeling [49] to heat transfer [97, 98]. In all of these applications, the primary computational challenge remains one of extracting information from the posterior density [66, 75]. Most estimates take the form of integrals over the posterior, which may be computed with asymptotic methods, deterministic methods, or sampling. Deterministic quadrature or cubature [25, 17] may be attractive alternatives to Monte Carlo at low to moderate dimensions, but Markov chain Monte Carlo (MCMC) [91, 11, 42] remains the most general and flexible method for complex and high-dimensional distributions. All of these methods, however, require evaluation of the likelihood or posterior at many values of the model parameters $\mathbf{m}$. In this setting, evaluating the likelihood requires solving the forward problem. With complex forward models, such as those described by partial differential equations, each single evaluation can be a computationally expensive undertaking [46]. For Monte Carlo simulations requiring 10^3–10^5 samples, the total cost of these forward evaluations quickly becomes prohibitive.

This report presents a new formulation designed to accelerate evaluation of Bayesian integrals and other characterizations of the posterior. We develop methods to substantially reduce the cost of evaluating the posterior density, based on a stochastic spectral reformulation of the forward problem. These methods have their roots in uncertainty quantification (UQ) using polynomial chaos (PC) expansions [40, 59, 21].

The efficient *forward* propagation of uncertainty—i.e., from model parameters to model predictions—is a central challenge of uncertainty quantification. A simple approach is Monte Carlo simulation: sampling known distributions of the model parameters to obtain statistics or density estimates of the model predictions. Again, each sample requires a solution of the forward model, and with complex models, this sampling approach is computationally intensive. A useful alternative is to employ spectral representations of uncertain parameters and field quantities, specifically polynomial chaos (PC) expansions for random variables and stochastic processes. The polynomial chaos [100, 14, 70, 73, 15, 40] was first defined by Wiener [100]; successive polynomial chaoses give rise to a functional basis consisting of Hermite polynomials of Gaussian random variables [13]. Ghanem & Spanos [40] describe the implementation of polynomial chaos in a finite element context. These stochastic finite element approaches have found numerous modeling applications, including transport in porous media [36], and solid [37, 38] or structural [39] mechanics. Le Maître *et al.* [59, 61] extended these techniques to thermo-fluid systems. Xiu *et al.* [104] used generalized polynomial chaos [103] for uncertainty quantification in fluid-structure interactions and in diffusion problems [102], while Debusschere *et al.* [20] used polynomial chaos to characterize uncertainty in electrochemical microfluid systems.

We will show that Bayesian estimation is intimately related to the forward propagation of uncertainty. In particular, using PC to propagate a wide range of uncertainty—e.g., prior uncertainty—through the forward problem and sampling the resulting spectral expansion enables a substantially more efficient Bayesian solution of the inverse problem. To this end, we employ an "intrusive" stochastic spectral methodology, in which polynomial chaos representations of the unknown parameters lead to a reformulation of the governing equations of the forward model. This process involves: (1) constructing PC expansions $\mathbf{g}(\boldsymbol{\xi})$ for each unknown parameter, according to probability distributions that include the support of the prior; (2) substituting these expansions into the governing equations and using Galerkin projection to obtain a coupled system of equations for the PC mode strengths; (3) solving this system; and (4) forming an expression for the posterior density based on the resulting PC expansions of forward model predictions, then exploring this posterior density with an appropriate sampling strategy. In this scheme, sampling can have negligible cost; nearly all the computational time is spent solving the system in step 3. Depending on model nonlinearities and the necessary size of the PC basis, this computational effort may be orders of magnitude less costly than exploring the posterior via direct sampling.

Other attempts at accelerating Bayesian inference in computationally intensive inverse problems have relied on reductions of the forward model. Wang & Zabaras [98] use proper orthogonal decomposition (POD) [10] and Galerkin projection to speed forward model calculations in a radiative source inversion problem. The empirical basis [86] used for model

reduction is pre-constructed using full forward problem simulations. The choice of inputs to these simulations—in particular, how closely the inputs must resemble the inverse solution—can be important [98]. Balakrishnan *et al.* [8] introduce a PC representation of the forward model in a groundwater transport parameter identification problem, but obtain the PC coefficients by collocation; again, this process depends on a series of "snapshots" obtained from repeated forward simulations.

In the statistical literature, under the headline of "Bayesian parameter calibration," Gaussian processes have been used extensively as surrogates for complex computational models [55]. These approaches treat the forward model as a black box, and thus require careful attention to experimental design and to modeling choices that specify the mean and covariance of the surrogate Gaussian process. A different set of approaches retain the full forward model but use reduced models to guide and improve the efficiency of MCMC. Christen & Fox [16] use a local linear approximation of the forward model to improve the acceptance probability of proposed moves, reducing the number of times the likelihood must be evaluated with the full forward model. This "approximate MCMC" algorithm is shown to yield the same stationary distribution as a standard Metropolis-Hastings chain. Higdon *et al.* [46] focus on the estimation of spatially-distributed inputs to a complex forward model. They introduce coarsened representations of the inputs and apply a Metropolis-coupled MCMC scheme [34] in which "swap proposals" allow information from the coarse-scale formulation, which may be computed more quickly, to influence the fine-scale chain. Efendiev *et al.* [24] also develop a two-stage MCMC algorithm, using a coarse-scale model based on multiscale finite volume methods to improve the acceptance rate of MCMC proposals. In contrast to the present formulation, however, all of the approaches cited above require repeated solutions of the full-scale deterministic forward model.

We will demonstrate our new formulation on a transient diffusion problem arising in contaminant source inversion, and compare the efficiency of the method and the accuracy of posterior estimates to direct evaluation of the posterior.

1.2 Formulation

1.2.1 Bayesian inference for inverse problems

Consider a forward problem defined as follows:

$$\mathbf{d} \approx \mathbf{G}(\mathbf{m}) \tag{1}$$

Here $\mathbf{m}$ is a vector of model parameters and $\mathbf{d}$ is a vector of observable quantities, or data. The forward model $\mathbf{G}$ yields predictions of the data as a function of the parameters. In the Bayesian setting, both $\mathbf{m}$ and $\mathbf{d}$ are random variables, and for the remainder of this chapter we will take these random variables to be real-valued. We use Bayes' rule to define a posterior probability density for the model parameters $\mathbf{m}$, given an observation of the data

$\mathbf{d}$:

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p_m(\mathbf{m})}{\int p(\mathbf{d}|\mathbf{m})p_m(\mathbf{m})\mathrm{d}\mathbf{m}} \tag{2}$$

In the Bayesian setting, probability is used to express knowledge about the true values of the parameters. In other words, prior and posterior probabilities represent *degrees of belief* about possible values of $\mathbf{m}$, before and after observing the data $\mathbf{d}$.

Data thus enters the formulation through the likelihood $p(\mathbf{d}|\mathbf{m})$, which may be viewed as a function of $\mathbf{m}$: $L(\mathbf{m}) \equiv p(\mathbf{d}|\mathbf{m})$. A simple model for the likelihood assumes that independent additive errors account for the deviation between predicted and observed values of $\mathbf{d}$:

$$\mathbf{d} = \mathbf{G}(\mathbf{m}) + \boldsymbol{\eta} \tag{3}$$

where components of $\boldsymbol{\eta}$ are i.i.d. random variables with density p_η. A typical assumption is $\eta_i \sim N(0, \sigma^2)$, in which case $p(\mathbf{d}|\mathbf{m})$ becomes $N(\mathbf{G}(\mathbf{m}), \sigma^2\mathbf{I})$. The likelihood is thus

$$L(\mathbf{m}) = \prod_i p_\eta\left(d_i - G_i(\mathbf{m})\right) \tag{4}$$

In this simple model, $\boldsymbol{\eta}$ may encompass both measurement error (e.g., sensor noise) and model error—the extent to which forward model predictions may differ from "true" values because of some unmodeled physics of the system.

Any additional information on the model parameters may enter the formulation through a suitably-defined prior density, $p_m(\mathbf{m})$. Prior models may embody simple constraints on $\mathbf{m}$ such as a range of feasible values, or may reflect more detailed knowledge about the parameters (shapes, correlations, smoothness, etc). In the absence of additional information, one may simply choose a prior that is *uninformative*.

Bayesian estimation typically gives rise to integrals over the posterior density:

$$I[f] = \int f(\mathbf{m})L(\mathbf{m})p_m(\mathbf{m})\mathrm{d}\mathbf{m} \tag{5}$$

The posterior expectation of a function f, for instance, is $\mathbb{E}_\pi f = I[f]/I[1]$. Though more sophisticated means of estimating such integrals will be discussed later, we note here that a conceptually simple method of obtaining posterior estimates is Monte Carlo sampling. If independent samples $\mathbf{m}^{(j)}$ can be drawn from the prior, then a Monte Carlo estimate of (5) is

$$\hat{I}_n[f] = \frac{1}{n}\sum_{j=1}^{n}\left[f\left(\mathbf{m}^{(j)}\right)\prod_i p_\eta\left(d_i - G_i(\mathbf{m}^{(j)})\right)\right] \tag{6}$$

If parameters ϕ_m of the prior density $p_m(\mathbf{m}|\phi_m)$ or parameters ϕ_η of the error model $p_\eta(\eta_i|\phi_\eta)$ are not known *a priori*, they may become additional objects for Bayesian inference. In other words, these *hyperparameters* may themselves be endowed with priors and estimated from data [74]:

$$p(\mathbf{m}, \phi_m, \phi_\eta|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m}, \phi_\eta)p_m(\mathbf{m}|\phi_m)p(\phi_\eta)p(\phi_m) \tag{7}$$

The resulting joint posterior over model parameters and hyperparameters may then be interrogated in various ways—e.g., by marginalizing over the hyperparameters to obtain $p(\mathbf{m}|\mathbf{d})$; or first marginalizing over $\mathbf{m}$ and using the maximizer of this density as an estimate of the hyperparameters; or by seeking the joint maximum *a posteriori* estimate or posterior mean of $\mathbf{m}$, ϕ_m, and ϕ_η [68, 74].

1.2.2 Polynomial chaos expansions

Let $(\Omega, \mathcal{U}, P)$ be a probability space, where Ω is a sample space, $\mathcal{U}$ is a σ-algebra over Ω, and P is a probability measure on $\mathcal{U}$. Also, let $\{\xi_i(\omega)\}_{i=1}^{\infty}$ be a set of orthonormal standard Gaussian random variables on Ω. Then any square-integrable random variable $X : \Omega \to \mathbb{R}$ has the following representation:

$$\begin{aligned} X(\omega) &= a_0\Gamma_0 + \sum_{i_1=1}^{\infty} a_{i_1}\Gamma_1(\xi_{i_1}) + \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1} a_{i_1 i_2}\Gamma_2(\xi_{i_1},\xi_{i_2}) \\ &+ \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1}\sum_{i_3=1}^{i_2} a_{i_1 i_2 i_3}\Gamma_3(\xi_{i_1},\xi_{i_2},\xi_{i_3}) + \cdots \end{aligned} \tag{8}$$

where Γ_p is the Wiener polynomial chaos of order p [100, 40, 53] This expansion may be re-written in a more compact form

$$X(\omega) = \sum_{k=0}^{\infty} \hat{a}_k \Psi_k(\xi_1, \xi_2, \ldots) \tag{9}$$

where there is a one-to-one correspondence between the coefficients and functionals in (8) and in (9) [40]. For the standard normal random variables ξ_i chosen above, orthogonality of successive Γ_p requires that the Γ_p be multivariate Hermite polynomials; both these and the corresponding Ψ_k may be generated from univariate Hermite polynomials by taking tensor products.

Of course, in computations it is not useful to represent a random variable with an infinite summation, and one truncates the expansion both in order p and in dimension n—i.e., by choosing a subset $\boldsymbol{\xi} = \{\xi_{\lambda_i}\}_{i=1}^{n}$ of the infinite set $\{\xi_i\}$, $\lambda_i \in \mathbb{N}$. The total number of terms P in the finite *polynomial chaos expansion*

$$X(\omega) = \sum_{k=0}^{P} x_k \Psi_k(\xi_1, \xi_2, \ldots, \xi_n) \tag{10}$$

is:

$$P + 1 = \frac{(n+p)!}{n!p!}. \tag{11}$$

Polynomial chaos (PC) expansions have been generalized to broader classes of orthogonal polynomials in the Askey scheme, each family resulting from a different choice of distribution for the ξ_i [103, 83]. For each of these choices, orthogonality of the polynomials $\Psi_k(\boldsymbol{\xi})$

with respect to the inner product on $L^2(\Omega)$ is guaranteed:

$$\begin{aligned}\langle \Psi_i \Psi_j \rangle &= \int \Psi_i(\boldsymbol{\xi}(\omega))\Psi_j(\boldsymbol{\xi}(\omega))\,dP(\omega) \\ &= \int \Psi_i(\boldsymbol{\xi})\Psi_j(\boldsymbol{\xi})\,w(\boldsymbol{\xi})\,d\boldsymbol{\xi} \\ &= \delta_{ij}\langle \Psi_i^2 \rangle \end{aligned} \tag{12}$$

where, in the second (Riemann) integral, $w(\boldsymbol{\xi})$ denotes the probability density of $\boldsymbol{\xi}$. This property can be used to calculate the truncated PC representation of a random variable $f \in L^2(\Omega)$ by projecting onto the PC basis:

$$\tilde{f}(\omega) = \sum_{k=0}^{P} f_k \Psi_k(\boldsymbol{\xi}), \quad f_k = \frac{\langle f(X)\Psi_k \rangle}{\langle \Psi_k^2 \rangle} \tag{13}$$

This orthogonal projection minimizes the error $\|f - \tilde{f}\|$ on the space spanned by $\{\Psi_k\}_{k=0}^{P}$, where $\|\cdot\|$ is the inner-product norm on $L^2(\Omega)$.

Suppose that the behavior of f can be expressed as $O(f,X) = 0$, where O is some deterministic operator and $X(\omega)$ is a random variable with a known PC expansion $X = \sum_i^P x_i \Psi_i$. Substituting PC expansions for f and X into this operator and requiring the residual to be orthogonal to Ψ_j for $j = 0 \ldots P$ yields a set of coupled, deterministic equations for the PC coefficients f_k:

$$\left\langle O\left(\sum_k^P f_k \Psi_k, \sum_i^P x_i \Psi_i\right)\Psi_j \right\rangle = 0, \quad j = 0 \ldots P \tag{14}$$

This Galerkin approach is known as "intrusive" spectral projection [61], in contrast to "non-intrusive" approaches in which the inner product $\langle f(X)\Psi_k \rangle$ is evaluated by sampling or quadrature, thus requiring repeated evaluations of $f(X)$ corresponding to different realizations of $\boldsymbol{\xi}$ [80].

In practice, we employ a pseudospectral construction to perform intrusive projections efficiently for higher powers of random variables, e.g. $f(X) = X^j, j \geq 3$, and have developed additional techniques for nonpolynomial functions f. These operations are incorporated into a library for "stochastic arithmetic," detailed in [21].

1.2.3 Efficient evaluation of the posterior: Sampling from the prior

We now connect ideas introduced in the two preceding sections to formulate a computationally efficient scheme for Bayesian inference. Suppose the model parameters have been endowed with a prior density $p_m(\mathbf{m})$. Knowing this density, one can construct a corresponding polynomial chaos expansion for each component m_i of the random vector $\mathbf{m}$:

$$m_i(\boldsymbol{\xi}) = \sum_{k=0}^{P} m_{ik} \Psi_k(\boldsymbol{\xi}) \tag{15}$$

where the dimension of $\boldsymbol{\xi}$ is at least equal to the dimension of $\mathbf{m}$.

Next, we introduce these PC expansions into the forward model and use Galerkin projection to obtain a PC representation for each component of the predicted data $G_i(\mathbf{m})$. Here $G_i(\mathbf{m})$ denotes the i-th component of $\mathbf{G}(\mathbf{m})$ and $\tilde{G}_i(\boldsymbol{\xi})$ is its (approximate) PC representation:

$$\tilde{G}_i(\boldsymbol{\xi}) = \sum_{k=0}^{P} d_{ik}\Psi_k(\boldsymbol{\xi}) \tag{16}$$

Now consider a generic integral over the unnormalized posterior density, given in (5). Drawing samples $\boldsymbol{\xi}^{(j)}$ from the distribution of $\boldsymbol{\xi}$ will yield samples of $\mathbf{m}$ from the prior, calculated according to (15). But the corresponding forward model prediction, $\mathbf{G}(\mathbf{m})$, can now be computed very cheaply, simply by substituting the same $\boldsymbol{\xi}^{(j)}$ into (16). In general, evaluating this $(P+1)$-term expansion will be significantly faster than solving the forward model for each sample. The prediction $\tilde{\mathbf{G}}(\boldsymbol{\xi})$ obtained in this fashion then replaces $\mathbf{G}(\mathbf{m})$ in the likelihood $L(\mathbf{m})$. For the simple likelihood proposed in (4), an n-sample Monte Carlo estimate of the integral becomes:

$$\hat{I}_n[f] = \frac{1}{n}\sum_{j=1}^{n}\left[f\left(\mathbf{m}(\boldsymbol{\xi}^{(j)})\right)\prod_i p_\eta\left(d_i - \tilde{G}_i(\boldsymbol{\xi}^{(j)})\right)\right] \tag{17}$$

1.2.4 Efficient evaluation of the posterior: Sampling from alternate distributions

Polynomial chaos reformulations of a Bayesian estimate need not be limited to expansions for which $\mathbf{m}$ has density p_m, as specified in (15), however. Consider a different set of PC expansions, $\mathbf{m} = \mathbf{g}(\boldsymbol{\xi})$, where each component $m_i = g_i$ is given by

$$m_i = g_i(\boldsymbol{\xi}) = \sum_{k=0}^{P} g_{ik}\Psi_k(\boldsymbol{\xi}) \tag{18}$$

Let the g_{ik} be chosen such that $\mathbf{g}(\boldsymbol{\xi})$ has probability density q, where the support of q includes the support of the prior. Drawing samples $\boldsymbol{\xi}^{(j)}$ from the distribution of $\boldsymbol{\xi}$ will now yield samples of $\mathbf{m}$ drawn from q. A Monte Carlo estimate of the integral in (5) now becomes:

$$\hat{I}_n[f] = \frac{1}{n}\sum_{j=1}^{n}\left[f\left(\mathbf{g}(\boldsymbol{\xi}^{(j)})\right)\prod_i p_\eta\left(d_i - \tilde{G}_i(\boldsymbol{\xi}^{(j)})\right)\frac{p_m\left(\mathbf{g}(\boldsymbol{\xi}^{(j)})\right)}{q\left(\mathbf{g}(\boldsymbol{\xi}^{(j)})\right)}\right] \tag{19}$$

If q is chosen to sample from regions where $|f(\mathbf{m})|L(\mathbf{m})p_m(\mathbf{m})$ is relatively large, then the estimate in (19) amounts to importance sampling [5, 27]. The variance of $f(\mathbf{m})L(\mathbf{m})p_m(\mathbf{m})/q(\mathbf{m})$, where $\mathbf{m} \sim q$, will be reduced by this sampling strategy and consequently the variance of the estimator $\hat{I}_n[f]$ will be reduced as well. In the present context, of course, evaluating the

likelihood of each sample via PC expansions is already inexpensive, so importance sampling may not yield significant gain. However, freedom in the choice of $\mathbf{g}(\boldsymbol{\xi})$ has a more immediate utility. If the prior distribution is such that it is difficult or inconvenient to write a PC expansion for $\mathbf{m}$ with density p_m, the ability to choose a density q that may be simpler than p_m ensures flexiblity in prior modeling.

1.2.5 Efficient evaluation of the posterior: Change of variables and MCMC

The essence of the two sampling schemes presented above is that the likelihood, when computed using PC expansions, becomes a function of $\boldsymbol{\xi}$ rather than of $\mathbf{m}$. Thus we sample $\boldsymbol{\xi}$ to generate samples of $\mathbf{m}$ from a specified distribution and simultaneously use $\boldsymbol{\xi}$ to compute the likelihood of each sample. Implicit in these schemes is a change of variables from $\mathbf{m}$ to $\boldsymbol{\xi}$, and it is fruitful to consider this change explicitly, as follows:

$$\begin{aligned} I[f] &= \int_{\mathcal{M}} f(\mathbf{m}) L(\mathbf{m}) p_m(\mathbf{m}) \mathrm{d}\mathbf{m} \\ &= \int_{\Xi_m} f(\mathbf{g}(\boldsymbol{\xi})) L(\mathbf{g}(\boldsymbol{\xi})) p_m(\mathbf{g}(\boldsymbol{\xi})) |\det D\mathbf{g}(\boldsymbol{\xi})| \, d\boldsymbol{\xi} \end{aligned} \tag{20}$$

Here, $D\mathbf{g}$ denotes the Jacobian of $\mathbf{g}$. Making this change of variables explicit imposes certain constraints on the transformation $\mathbf{g}$, namely (1) that $\Xi_m = \mathbf{g}^{-1}(\mathcal{M})$, the inverse image of the support of the prior, be contained within the range of $\boldsymbol{\xi}$, and (2) that $\mathbf{g}$ be a differentiable transformation from Ξ_m to $\mathcal{M}$ with a differentiable inverse (i.e., that $\mathbf{g}$ be a diffeomorphism from Ξ_m to $\mathcal{M}$). The first constraint is not new; indeed, in the preceding two sections, it is necessarily satisfied by a PC expansion that reproduces samples of $\mathbf{m}$ from the desired distribution, p_m or q. The latter constraint, however, limits $\mathbf{g}$ to rather simple PC expansions—for instance, linear transformations of $\boldsymbol{\xi}$. * But this limitation is not a great liability, as the transformed integral in (20) can now be evaluated by any suitable sampling scheme in $\boldsymbol{\xi}$-space—in particular, by Markov chain Monte Carlo (MCMC).

MCMC encompasses a broad class of methods that simulate drawing samples from the posterior [42, 29, 5], and thus can be used to directly estimate the posterior expectation of f:

$$\mathbb{E}_{\pi_m} f = \int f(\mathbf{m}) \pi_m(\mathbf{m}) \, \mathrm{d}\mathbf{m} \tag{21}$$

where π_m denotes the normalized posterior density of $\mathbf{m}$

$$\pi_m(\mathbf{m}) \propto L_m(\mathbf{m}) p_m(\mathbf{m}) \tag{22}$$

*This condition is not required by the sampling schemes in §§1.2.3–1.2.4. If, however, it *is* satisfied by $\mathbf{g}$ in §1.2.3 then we have

$$p_m(\mathbf{g}(\boldsymbol{\xi})) |\det D\mathbf{g}(\boldsymbol{\xi})| = w(\boldsymbol{\xi})$$

where w is defined in (12). An analogous condition holds true for q in §1.2.4, again when $\mathbf{g}$ is a diffeomorphism from Ξ_q to $\mathcal{M}$.

The subscript on L_m emphasizes that the likelihood is here a function of $\mathbf{m}$. The change of variables from $\mathbf{m}$ to $\boldsymbol{\xi}$ can be expressed compactly as

$$\mathbb{E}_{\pi_m} f = \mathbb{E}_{\pi_\xi} (f \circ \mathbf{g}) \tag{23}$$

where π_ξ is the posterior density in $\boldsymbol{\xi}$-space:

$$\pi_\xi(\boldsymbol{\xi}) \propto L_m(\mathbf{g}(\boldsymbol{\xi}))\, p_m(\mathbf{g}(\boldsymbol{\xi}))\, |\det D\mathbf{g}(\boldsymbol{\xi})| \tag{24}$$

As before, we would like to use the PC representation of $\mathbf{G}$ to accelerate evaluations of the posterior. We first introduce the following notational convention: any quantity computed by projection onto a finite PC basis—whether a forward model prediction, likelihood, or posterior density—is distinguished from its "direct" counterpart with a tilde. Thus, we seek samples from the surrogate posterior $\tilde{\pi}_\xi$:

$$\begin{aligned} \tilde{\pi}_\xi(\boldsymbol{\xi}) \propto & \quad \tilde{L}_\xi(\boldsymbol{\xi}) \quad p_m(\mathbf{g}(\boldsymbol{\xi}))\, |\det D\mathbf{g}(\boldsymbol{\xi})| \\ \propto & \quad \textstyle\prod_i p_\eta\left(d_i - \tilde{G}_i(\boldsymbol{\xi})\right) \quad p_m(\mathbf{g}(\boldsymbol{\xi}))\, |\det D\mathbf{g}(\boldsymbol{\xi})| \end{aligned} \tag{25}$$

The likelihood is now a function of $\boldsymbol{\xi}$. Since Metropolis-Hastings algorithms require knowledge only of the unnormalized posterior density, (25) is sufficient to simulate samples $\boldsymbol{\xi}^{(j)}$ from the posterior on $\boldsymbol{\xi}$-space. Eliminating b burn-in samples, the posterior expectation of f is estimated by an ergodic average:

$$\mathbb{E}_{\pi_m} f \approx \frac{1}{n-b} \sum_{j=b+1}^{n} (f \circ \mathbf{g}) \left(\boldsymbol{\xi}^{(j)}\right) \tag{26}$$

Note that an MCMC chain could just as easily be run on $\mathbf{m}$-space, simulating samples from the surrogate posterior $\tilde{\pi}_m \propto (\tilde{L}_\xi \circ \mathbf{g}^{-1}) \cdot p_m$. But this expression emphasizes why an invertible $\mathbf{g}$ must be used with MCMC; otherwise the argument to $\tilde{L}_\xi$ corresponding to a given chain position $\mathbf{m}^{(j)}$ would be ill-defined.

Advantages of MCMC over the simple Monte Carlo schemes in §§1.2.3–1.2.4 are several. In many applications, a well-designed MCMC algorithm can offer far better sampling efficiency than sampling from the prior or from some alternate distribution $q \neq \pi_m$, despite the fact that MCMC samples are serially correlated [82]. In the present context, however, the value of improved sampling efficiency is tempered by the fact that samples are rendered inexpensive by the PC reformulation. However, MCMC offers additional benefits. Because MCMC directly simulates the posterior, it is simple to extract marginal densities for individual components of $\mathbf{m}$ with the aid of kernel density estimation [30]. Also, MCMC eliminates the need to calculate the posterior normalization factor $I[1]$. Using (25) to evaluate the posterior, further particulars of the MCMC algorithm we adopt are essentially independent of the PC formulation and thus we reserve their presention for §1.3.5.

The choice of $\mathbf{g}$ and of the PC basis $\{\Psi_k(\boldsymbol{\xi})\}_{k=0}^P$, on the other hand, will have a crucial impact on the accuracy and cost of PC-reformulated posterior estimates and predictions. Because it is the initial PC representation of $\mathbf{m}$—its distribution defining the stochastic forward problem—$\mathbf{g}(\boldsymbol{\xi})$ will directly affect $\tilde{\mathbf{G}}$, as will the order and stochastic dimension of the PC basis used in Galerkin projections. The "surrogate" posterior $\tilde{\pi}$, obtained by replacing direct evaluation of the likelihood $L(\mathbf{m})$ with a likelihood written in terms of $\tilde{\mathbf{G}}(\boldsymbol{\xi})$, is then at the heart of any PC-induced errors in the three sampling schemes discussed above. Whether we write this posterior in terms of $\boldsymbol{\xi}$ (e.g., $\tilde{\pi}_\xi$) or $\mathbf{m}$ (e.g., $\tilde{\pi}_m$, if $\mathbf{g}$ is invertible), the difference between π and $\tilde{\pi}$ completely captures the impact of the polynomial chaos representation of forward model predictions on the inverse solution. We will explore the dependence of this posterior error on $\mathbf{g}$, p, and the distribution of $\boldsymbol{\xi}$—i.e., the type of PC basis—in §§1.3.3 and 1.3.6.

1.2.6 Decomposition of parameter space

When a component of the forward solution depends very steeply on an uncertain parameter, a PC basis of smooth, global polynomials may require increasingly high order to provide an accurate solution of the stochastic forward problem. In the diffusion-driven forward problem to be considered here, many terms will be required to represent a sharply-localized source with broad prior uncertainty in its location. In the limit, when solutions are *discontinuous* with respect to an uncertain parameter, e.g., in the neighborhood of a critical point, a global PC basis may be unsuitable [58].

Several methods have been proposed to address this difficulty [58, 60, 95] A PC basis of Haar wavelets was constructed in [58], then generalized to a multi-wavelet basis in [60]; both of these methods effectively resolved stochastic systems with multiple bifurcations. Computational efficiency of the latter scheme was improved by block-partitioning of the uncertain parameter space. We adopt a similar, but non-adaptive, partitioning scheme here. The support $\mathcal{M}$ of the prior, or equivalently the range of $\mathbf{m}$, is decomposed into N_b non-overlapping domains

$$\mathcal{M} = \bigcup_b^{N_b} \mathcal{M}^b, \quad \mathcal{M}^b \cap \mathcal{M}^{b'} = \emptyset \text{ if } b \neq b' \tag{27}$$

We seek a corresponding decomposition of the prior density $p_m(\mathbf{m})$, and thus of the Bayesian integral $I[f]$ in (5), as follows:

$$\begin{aligned} p^b(\mathbf{m}) &= \begin{cases} p_m(\mathbf{m}) & \mathbf{m} \in \mathcal{M}^b \\ 0 & \mathbf{m} \notin \mathcal{M}^b \end{cases} \\ p_m(\mathbf{m}) &= \sum_b^{N_b} p^b(\mathbf{m}) \end{aligned} \tag{28}$$

Note that the densities $p^b(\mathbf{m})$ are unnormalized on each domain $\mathcal{M}^b$, so that

$$I[f] = \sum_b^{N_b} I^b[f] = \sum_b^{N_b} \int f(\mathbf{m}) L(\mathbf{m}) p^b(\mathbf{m}) \mathrm{d}\mathbf{m} \tag{29}$$

This partitioning allows the construction of a separate polynomial expansion $\mathbf{g}^b(\boldsymbol{\xi})$ for $\mathbf{m}$ on each block, and thus a different version of the likelihood $\tilde{L}_\xi$ in each integral $I^b[f]$ contributing to (5). What does this imply for $\mathbf{g}^b$? Again, we identify three cases, corresponding to the sampling schemes in §§1.2.3–1.2.5. When sampling from the prior, we seek $\mathbf{g}^b$ so that $\mathbf{m}$ has density proportional to $p^b(\mathbf{m})$. Sampling $\boldsymbol{\xi}$ will then yield samples from the prior on each block. When sampling from an alternate distribution, e.g., from N_b densities $q^b(\mathbf{m})$, we simply require that the support of each q^b contain $\mathcal{M}^b$. In particular, the supports of each q^b need not be disjoint; the definition of $p^b(\mathbf{m})$ above ensures zero posterior density outside of each domain. Finally, in an MCMC scheme, we work directly with the partition of Ξ and allow the chain to transition from block to block according to the proposal distribution.

Le Maître *et al.* [60] provide criteria for adaptively refining the partitions based on the local variance of the solution. Here, because the inverse problem to be considered has a simple symmetry in both its priors and its computational domain, we will take a fixed partition of $\mathcal{M} = [0,1] \times [0,1]$ into four equal quadrants.

1.3 Results

We demonstrate the stochastic spectral formulation of Bayesian inference by inverting for the source field in a transient diffusion problem.

One practical context of this inverse problem lies in contaminant source inversion [3]. Given a sparse set of concentration measurements—from sensors scattered throughout some space, for instance—one would like to find the sources of a toxin that has spread through the ambient medium. Specific parameters to infer include the number of sources and their possibly time-dependent strengths and locations. Though convective transport can play a role in many practical source inversion problems, we will limit our attention here to a purely diffusion-driven inverse problem in order to focus on the demonstration and analysis of the new formulation. Diffusive source inversion problems themselves arise in the context of porous media flows [85] and heat conduction [9, 4, 57, 54, 96].

1.3.1 Source inversion under diffusive transport

We begin by defining the deterministic forward problem $\mathbf{G}(\mathbf{m})$, since this is the basis for the general Bayesian approach to inverse problems described in §1.2.1 and for the stochastic forward problem described in §1.2.5.

Consider a dimensionless diffusion equation on a square domain $S = [0,1] \times [0,1]$ with adiabatic boundaries:

$$
\begin{aligned}
\frac{\partial u}{\partial t} &= \nabla^2 u + \sum_{l=1}^{N} \frac{s_l}{2\pi\sigma_l^2} \exp\left(-\frac{|\chi_l - \mathbf{x}|^2}{2\sigma_l^2}\right)[1 - H(t - T_l)] \qquad (30) \\
\nabla u \cdot \hat{\mathbf{n}} &= 0 \text{ on } \partial S \\
u(\mathbf{x}, 0) &= 0
\end{aligned}
$$

The source term in (30) describes N localized sources, each one active on the interval $t \in [0, T_l]$ and centered at $\chi_l \in S$ with strength s_l and width σ_l. Note that the location, size, and shutoff time of each source enter the problem nonlinearly.

For the purposes of an initial demonstration and to allow direct visualization of the posterior, we restrict our attention to an inverse problem in two dimensions. Thus we fix $N = 1$; prescribe T, s, and σ; and leave the source location $\chi = (m_0, m_1)$ unknown. For any given value of $\mathbf{m}$, we solve the PDE in (30) using a finite difference method. The u-field is described on a uniform grid with spacing $h = 0.025$. Second-order centered differences are used to discretize the diffusion terms. Time integration is via an explicit, second-order-accurate, Runge-Kutta-Chebyshev (RKC) scheme [88] with $\Delta t = 0.002$. The number of substeps in the RKC scheme is automatically determined by stability constraints upon setting ε, the damping parameter that controls the extent of the stability region, to $2/13$ [93]. Numerical resolution studies were conducted to validate the present choices of h and Δt.

The forward model $\mathbf{G}(\mathbf{m})$ predicts the value of the field $u(\mathbf{x}, t)$ at specific locations and times. Below, unless otherwise specified, $\mathbf{G}$ will provide predictions on a uniform 3×3 grid covering the domain S at two successive times, $t = 0.05$ and $t = 0.15$. The *inverse problem* thus consists of inferring the source position from *noisy* measurements at these locations and times. We let independent zero-mean Gaussian random variables $\eta_i \sim N(0, \varsigma^2)$ express the difference between "real-world" measurements and model predictions, as specified in (3). In the examples below, we choose $\varsigma = 0.4$ unless otherwise specified. Priors simply constrain the source to lie in the domain S, i.e., $m_i \sim U(0,1)$. Again, in the interest of simplicity, we make no attempt at hyperparameter estimation for either the noise model or the prior distribution.

Figure 1 shows the u-field resulting from a representative value of $\mathbf{m}$: $(m_0, m_1) = (0.25, 0.75)$, with $T = 0.05$, $s = 0.5$, and $\sigma = 0.1$. Though the solution is initially peaked around the source, note that it flattens at the later time once the source is no longer active. As the measurement time $(t > T) \to \infty$, the inverse problem becomes increasingly ill-conditioned. Measurement noise will overwhelm any residual variation in the u-field resulting from the particular location of the source.

1.3.2 Stochastic spectral solution of the forward problem

The accuracy, and computational expense, of the stochastic spectral solution of the forward problem depend on the order of the PC basis used. For computational efficiency, the requisite PC order can be reduced by partitioning the domain and solving an independent forward problem, with a smaller range of input uncertainty, on each subdomain. Because the present problem has a simple symmetry in both its priors and its computational domain, we take a fixed partition of the prior support $\mathcal{M} = [0,1] \times [0,1]$ into four equal quadrants $\mathcal{M}^b$. On each of these quadrants, we prescribe a PC expansion $\mathbf{m} = \mathbf{g}^b(\boldsymbol{\xi})$ consisting of multivariate Legendre polynomials and uniformly distributed $\xi_i \sim U(-1,1)$. In particular, we choose PC coefficients such that each $\mathbf{g}^b(\boldsymbol{\xi})$ has a uniform probability density on $\mathcal{M}^b$ and zero probability density elsewhere. On each quadrant, $\mathbf{m}$ is thus distributed according to the (normalized) prior density $p^b(\mathbf{m})$ given in (28), e.g.,

$$\mathbf{g}^{b=1}(\boldsymbol{\xi}) = \begin{pmatrix} 1/4 \\ 1/4 \end{pmatrix} + \begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \tag{31}$$

and so on for $b = 2 \ldots 4$.

The stochastic forward problem is then solved four times, once for each block of the prior support. On each block, we introduce the PC expansion $\boldsymbol{\chi} = \mathbf{m} = \mathbf{g}^b(\boldsymbol{\xi})$ into (30) with $N = 1$ and, using Galerkin projections and the same finite-difference/RKC scheme as in the deterministic problem, obtain a PC expansion for each prediction of the scalar field, $\tilde{G}_i^b(\boldsymbol{\xi})$. These predictions are random variables $u(\mathbf{x}_i, t_i, \boldsymbol{\xi}(\omega))$, giving the value of the scalar field at each measurement location and time $(\mathbf{x}_i, t_i)$.

Note that the source term in (30) may be factored into a time-dependent component and a stationary component:

$$\sum_{l=1}^{N} \frac{s_l}{2\pi\sigma_l^2} \exp\left(-\frac{|\boldsymbol{\chi}_l - \mathbf{x}|^2}{2\sigma_l^2}\right) [1 - H(t - T_l)] = \sum_{l=1}^{N} q(\mathbf{x}, \boldsymbol{\chi}_l)\, s(t, T_l) \tag{32}$$

The stationary component $q(\mathbf{x}, \boldsymbol{\chi})$ contains the exponential of a random variable, and its PC representation is thus expensive to compute. In the interest of efficiency, we evaluate this term once at each grid point $\mathbf{x}_{mn}$ and store the resulting array of PC expansions for use at each subsequent timestep. So that comparisons of computational cost remain fair, we also factor the source term when solving the deterministic forward problem, again pre-computing the stationary part; in other words, the Gaussian source profile is evaluated only once during time integration of (30) for a given source location.

Solutions of the stochastic forward problem may be interrogated in several ways. Figure 2 shows the predicted value of the scalar field at a particular measurement location and time, $u(x_{i=3}$=0.0, $y_{i=3}$=0.5, $t_{i=3}$=0.15), as a function of $\boldsymbol{\xi}$. This surface is a single component of the stochastic forward problem solution, $\tilde{G}_i(\boldsymbol{\xi})$ for $i = 3$, and is obtained with PC bases of increasing order (p=3, 6, and 9). Partition of the prior support into quadrants $b = 1 \ldots 4$ is indicated on each plot. Convergence is observed with increasing p. While there is no

guarantee of continuity or smoothness of the solution between neighboring blocks, both seem to be achieved at sufficient order. The pointwise error in these forward solutions is shown in Figure 3, again for $i = 3$. Here, error is defined as the difference between the PC expansion of the forward solution and the exact solution of the deterministic forward model $\mathbf{G}(\mathbf{m})$, with argument $\mathbf{m}$ corresponding to the appropriate quadrant b and value of $\boldsymbol{\xi}$:

$$\mathrm{err}_i^b(\boldsymbol{\xi}) = \tilde{G}_i^b(\boldsymbol{\xi}) - G_i(g^b(\boldsymbol{\xi})) \tag{33}$$

Once again, the error becomes negligible at sufficient order.

Since the input $\mathbf{m} = \mathbf{g}(\boldsymbol{\xi})$ to the forward model is a random variable, any forward model output $G_i(\mathbf{m})$ is also a random variable. The density of these forward model outputs is a useful diagnostic and may be estimated in one of two ways. A direct (and computationally expensive) method is to sample $\mathbf{m}$ and solve the forward problem for each sample, forming a normalized histogram from the resulting collection of forward model outputs. Alternatively, one can sample $\boldsymbol{\xi}$ and substitute it into the PC expansion $\tilde{G}_i(\boldsymbol{\xi})$, again forming a histogram of the resulting values. This process essentially weighs the surface response in Figure 2 according to the probability distribution of $\boldsymbol{\xi}$. The resulting density estimates for G_3 are shown in Figure 4. While a lower-order PC basis ($p = 3$) results in a poor density estimate, the probability density converges to its true shape—obtained by the direct method—as p increases.

The probability densities computed above represent the propagation of prior uncertainty through the forward problem. Accordingly, we may endow them with an additional interpretation. Figure 5 shows the probability density of u at a single measurement location (x=0.0, y=0.0) but at two successive times: t=0.05 and t=0.15. As in Figure 4, we observe convergence of the PC-obtained density to its "direct" counterpart with increasing order. But we also observe that the probability density in Figure 5(a) is significantly broader than in Figure 5(b). Under the prior uncertainty, the earlier-time measurement takes a wider range of values than the later-time measurement, and in this sense, the earlier measurement is more informative; it will allow the likelihood to discriminate more clearly among possible values of $\mathbf{m}$. In the inverse problem setting, this information may be useful in choosing when and where to collect data.

1.3.3 Posterior densities

We now examine solutions of the inverse problem using polynomial chaos. A noisy data vector $\mathbf{d}$ is generated by solving the deterministic forward problem for a "true" model $\mathbf{m} = (0.25, 0.75)$, then perturbing the value of u at each sensor location/time with independent samples of Gaussian noise $\eta_i \sim N(0, \varsigma^2)$.

Figure 6 shows contours of the posterior density conditioned on $\mathbf{d}$. Solid lines are obtained via direct evaluations of the forward problem—i.e., they represent the posterior π_m in (22)—while dashed lines represent the posterior density computed with PC expansions on four partitions of the support of the prior. These are computed on $\boldsymbol{\xi}$-space, using (24),

but since $\mathbf{g}$ is invertible it is simple to transform them back to $\mathbf{m}$-space: $\tilde{\pi}_m \propto (\tilde{L}_\xi \circ \mathbf{g}^{-1}) p_m$. Very close agreement between π_m and $\tilde{\pi}_m$ is observed with increasing order.

A quantitative assessment of the error in the posterior density is obtained by considering the Kullback-Leibler (KL) divergence of π_m from $\tilde{\pi}_m$:

$$D(\tilde{\pi}\|\pi) = \int_{\mathcal{M}} \tilde{\pi}(\mathbf{m}) \log \frac{\tilde{\pi}(\mathbf{m})}{\pi(\mathbf{m})} \mathrm{d}\mathbf{m} \tag{34}$$

Figure 7 plots $D(\tilde{\pi}_m\|\pi_m)$ for PC bases of increasing order p. In terms of KL divergence, we observe an exponential rate of convergence of the surrogate posterior $\tilde{\pi}_m$ to the true posterior.

1.3.4 Posterior sampling and speedup

Practical Bayesian computations must explore the posterior by sampling, and it is here that we expect the PC approach to achieve significant speedup over its direct counterpart. Figure 8 shows the computational time for Monte Carlo estimation of the posterior mean as a function of the number of samples n. Since $g^b(\boldsymbol{\xi})$ is chosen to have density proportional to $p^b(\mathbf{m})$, our Monte Carlo estimator uses samples from the prior, as described in §1.2.3. In other words, the posterior mean is evaluated using (17), with $\mathbb{E}_\pi \mathbf{m} = \hat{I}_n[\mathbf{m}]/\hat{I}_n[1]$. We use a 6th-order PC basis for the stochastic spectral forward solution and compare the computational cost to that of direct sampling.

Speedup over the direct method is quite dramatic. The initial cost of the PC approach is offset by the computation of stochastic forward solutions, but then grows very slowly. Indeed, the per-sample cost is three orders of magnitude smaller for PC evaluations than for direct evaluations, and thus for even a moderate number of samples the gain in efficiency is significant. The cost of the initial stochastic forward solutions is recouped for $n \approx 200$, and thereafter the computational time of direct sampling rapidly eclipses that of the PC approach. For more complex forward models, the ratio of these per-sample costs may widen and the cost of the stochastic forward solutions may be recouped at even smaller n.

Another measure of speedup is to compare the computational times required to achieve a certain error in the Monte Carlo estimate of $I[\mathbf{m}]$. The estimate $\hat{I}_n[f(\mathbf{m})]$ is a random variable with variance

$$\sigma^2\left(\hat{I}_n[f(\mathbf{m})]\right) = \frac{1}{n} \mathrm{Var}_{\mathbf{m}\sim p_m}\left[f(\mathbf{m})L(\mathbf{m})\right] \tag{35}$$

where the subscript $\mathbf{m} \sim p_m$ reflects the drawing of samples from the prior distribution. In practice, we estimate the variance of $\hat{I}_n$ from the Monte Carlo samples using the recursive formula given in [27]. The Monte Carlo "standard error" is then simply $\sigma(\hat{I}_n[\mathbf{m}])$. Figure 9 shows the error thus computed and normalized by $\hat{I}_n[\mathbf{m}]$, versus computational time, for both the direct and PC approaches. Since error decreases as $n^{-1/2}$, the number of samples required to reduce the error grows rapidly. Because PC-obtained samples are inexpensive,

however, very small relative errors are achievable at negligible cost. This certainly is not true for direct sampling, as the solid line in Figure 9 indicates.

1.3.5 Markov chain Monte Carlo

Next we demonstrate the use of MCMC to simulate samples from the surrogate posterior $\tilde{\pi}_\xi$ given in (25). Since the $\mathbf{g}^b$ chosen above are invertible linear transformations from Ξ to $\mathcal{M}^b$, the conditions following the change of variables in (20) are satisfied. Therefore, the posterior expectation of any function f of $\mathbf{m}$ can be computed in $\boldsymbol{\xi}$-space, with samples from $\tilde{\pi}_\xi$, according to (23) and (26).

We employ a random-walk Metropolis algorithm [42] for MCMC, in which the proposal distribution $q(\cdot|\cdot)$ is a bivariate normal centered on the current position of the chain. The standard deviation of the proposal distribution is $\sigma_q = 0.4$. Results showing the chain position over 10000 iterations are in Figure 10. Visual inspection suggests that the chain mixes—i.e., moves within the support of $\tilde{\pi}_\xi$—rapidly. The two-dimensional view in Figure 10(a) is reminiscent of the posterior contours in Figure 6; this is not surprising, as $\boldsymbol{\xi}$ is just a diagonal linear transformation of $\mathbf{m}$.

A Metropolis-Hastings algorithm, such as the random-walk Metropolis sampler used here, provides for the construction of a Markov chain with stationary distribution $\tilde{\pi}_\xi$. Under certain additional conditions [91, 92, 82], one can establish a central limit theorem for ergodic averages $\bar{f}_n$ [35]: †

$$\bar{f}_n = \frac{1}{n}\sum_{j=1}^{n} (f \circ \mathbf{g})\left(\boldsymbol{\xi}^{(j)}\right) \tag{36}$$

$$\sqrt{n}\left(\bar{f}_n - \mathbb{E}_{\pi_\xi}(f \circ \mathbf{g})\right) \xrightarrow{i.d.} N(0, \sigma_f^2) \tag{37}$$

where $\xrightarrow{i.d.}$ denotes convergence in distribution and

$$\sigma_f^2 = \mathrm{Var}_{\pi_\xi}(f \circ \mathbf{g}) + 2\sum_{s=1}^{\infty} \mathrm{Cov}_{\pi_\xi}\left[(f \circ \mathbf{g})(\boldsymbol{\xi}^{(0)}), (f \circ \mathbf{g})(\boldsymbol{\xi}^{(s)})\right] \tag{38}$$

The asymptotic variance in (38) thus reflects the correlation between successive samples. Even if a central limit theorem does not strictly hold in the form of (37), stronger correlations lead to larger variance of the MCMC estimate at a given n and thus less efficient sampling. We plot $\gamma(s)$, the empirical autocovariance at lag s, in Figure 11 for several

† For chains on continuous state spaces, uniform or geometric ergodicity provide for central limit theorems [91, 29]. Weaker conditions are sufficient to establish a law of large numbers. While uniform or geometric ergodicity have not been shown for Metropolis-Hastings samplers on general state spaces, many of these samplers are conjectured to be geometrically ergodic. The chain resulting from an independence sampler with bounded π/q is known to be uniformly ergodic [92].

random-walk Metropolis samplers, varying the scale parameter of the proposal distribution σ_q. If σ_q is too large, a great proportion of the proposed moves will be rejected, and the chain will not move very often. If σ_q is too small, most proposed moves will be accepted but the chain will move very slowly through the posterior support. Both of these situations are reflected in long correlations and poor mixing. With $\sigma_q = 0.4$, however, we observe that autocovariance decays relatively quickly with lag along the chain, consistent with the good mixing in Figure 10. We also show, in black, the autocovariance of a chain resulting from an independence sampler [91], using the prior as a proposal distribution. This sampler also appears to be relatively efficient at exploring the simple posterior here.

A useful feature of MCMC estimation is the ease of extracting marginal distributions for components of $\mathbf{m}$ or $\boldsymbol{\xi}$. This is performed with kernel density estimation:

$$\pi(\xi_i) = \frac{1}{n-b} \sum_{j=b+1}^{n} K\left(\xi_i | \boldsymbol{\xi}^{(j)}\right) \tag{39}$$

where $K\left(\xi_i | \boldsymbol{\xi}^{(j)}\right)$ is a density concentrated around $\xi_i^{(j)}$ [42]. Here we use a one-dimensional Gaussian kernel, $K = N(\xi_i^{(j)}, \sigma_k^2)$ with bandwidth $\sigma_k = 0.01$. Marginal distributions for the source coordinates, transformed back into $\mathbf{m}$-space, are shown in Figure 12. The kernel centers are points at the bottom of the figure.

1.3.6 Choice of transformation and PC basis

Using MCMC to explore the posterior offers seemingly considerable freedom in the choice of $\mathbf{g}$, and further, in the choice of PC basis. All that is required is that $\mathbf{g}$ be a diffeomorphism from a set Ξ_m to the support of the prior, where $\Xi_m \subseteq \Xi$ and Ξ is the range of $\boldsymbol{\xi}$. Then $\mathbf{g}(\boldsymbol{\xi})$ is used to solve the stochastic forward problem on the chosen PC basis and an expression for the surrogate posterior density $\tilde{\pi}$ is formed.

It is reasonable, however, to expect that these choices will influence the accuracy and cost of evaluating the posterior distribution. Increasing the order of the PC basis certainly improves the accuracy of forward problem solutions and reduces errors in the surrogate posterior, as we observed in §§1.3.2–1.3.3. But we must also address the larger question—essentially, what is the uncertainty that one should propagate through the forward problem?

We explore the impact of different transformations $\mathbf{g}$ and PC bases using the two-dimensional source inversion problem as before. We consider three "true" source locations, ranging from the center to the edge of the domain: $\mathbf{m} = (0.50, 0.50)$; $\mathbf{m} = (0.25, 0.75)$; and $\mathbf{m} = (0.10, 0.90)$. For each of these sources, we solve the deterministic forward problem, then perturb the value of u at each sensor with independent samples of Gaussian noise $\eta_i \sim N(0, \varsigma^2)$, thus generating three noisy data vectors $\mathbf{d}$ for inference. The sensor locations/times, the source strength and shutoff time, and the prior p_m are unchanged from previous sections.

For simplicity, we do not partition the prior support; i.e., we choose $b = 1$. To solve the inverse problem with polynomial chaos, we thus prescribe a single set of PC expansions $\mathbf{m} = \mathbf{g}(\boldsymbol{\xi})$. First, consider a Gauss-Hermite PC basis: bivariate Hermite polynomials $\Psi_k(\boldsymbol{\xi})$ with $\xi_i \sim N(0,1)$. We prescribe a PC expansion centered on $\mathbf{m} = (0.5, 0.5)$:

$$g_i(\boldsymbol{\xi}) = 0.5 + \sigma\xi_i; \;\; i = 1,2 \tag{40}$$

Since $\Xi = \mathbb{R}^2$, the inverse image of the support of the prior will be contained in Ξ for any σ.

Contours of the surrogate posterior density $\tilde{\pi}_m$ are shown in Figure 13 for $\sigma = 10^{-2}$ and a 6th-order PC basis. These are compared to the "direct" posterior π_m for each source location. While agreement of the posteriors is relatively good for a source in the center of the domain (Figure 13(a)), the accuracy of the surrogate posterior deteriorates rapidly as the source moves towards the upper-left corner. For $\mathbf{m}_{\text{true}} = (0.10, 0.90)$, the surrogate posterior is centered on an entirely different area of $\mathcal{M}$ than the direct posterior. This disagreement explicitly reflects errors in the forward problem solution $\tilde{\mathbf{G}}(\boldsymbol{\xi})$ for $\boldsymbol{\xi}$ corresponding to values of $\mathbf{m}$ that are close to the boundary—i.e., for $|\boldsymbol{\xi}| = O(1/\sigma)$.

Widening the density of $\mathbf{g}(\boldsymbol{\xi})$ improves agreements significantly. Figure 14 shows posterior contours for $\sigma = 10^{-1}$ and a 6th-order PC basis. Once again, the best agreement is obtained when the source is in the center of the domain, but reasonable overlap of the posteriors is achieved even for $\mathbf{m}_{\text{true}} = (0.25, 0.75)$. The mean of the surrogate posterior for $\mathbf{m}_{\text{true}} = (0.10, 0.90)$, while still misplaced, shows some improvement over the $\sigma = 10^{-2}$ case. Increasing the order of the PC basis to $p = 9$ sharpens agreement for all three source locations, as observed in Figure 15. However, the key trend—deteriorating agreement as the source moves closer to the boundary—is preserved.

It is instructive to compare these results to those obtained with a uniform-Legendre PC basis. We again use a single partition of the prior support, with a PC expansion centered on $\mathbf{m} = (0.5, 0.5)$: $g_i = 0.5 + 0.5\xi_i$, $\xi_i \sim U(-1,1)$. Results with $p = 9$ are shown in Figure 16. Compared to the Gauss-Hermite basis at the same order, we observe slightly poorer agreement for sources in the center of the domain, but a more consistent level of error as the source is moved towards the boundary. Agreement of the surrogate and direct posteriors for $\mathbf{m}_{\text{true}} = (0.10, 0.90)$ is substantially better with uniform-Legendre PC than with any of the Gauss-Hermite bases.

An explanation for these results may be found in the distribution of $\boldsymbol{\xi}$, which is of course a defining property of the PC basis. As noted above, error in the surrogate posterior at a particular value of $\mathbf{m}$ reflects error in the stochastic forward problem solution for $\boldsymbol{\xi} = \mathbf{g}^{-1}(\mathbf{m})$. Where is this error likely to occur? If, for simplicity, one ignores aliasing errors resulting from the pseudospectral construction [21], Galerkin projection of the forward problem outputs onto the PC basis minimizes $\|\mathbf{G}(\mathbf{g}(\boldsymbol{\xi})) - \tilde{\mathbf{G}}(\boldsymbol{\xi})\|_{L^2(\Omega)}$, where this inner-product norm is defined by the probability measure P on $(\Omega, \mathcal{U})$. Let $\boldsymbol{\xi}$ and P induce a probability distribution on $\mathbb{R}^n$. The Gaussian distribution weighs errors near the origin much more strongly than errors at large $|\boldsymbol{\xi}|$, whereas the uniform distribution weighs errors equally over the entire (finite) range of $\boldsymbol{\xi}$.

This weighing is consistent with the error trends in Figures 13–16, and its impact is described more extensively in Figure 17. Here we compute the Kullback-Leibler divergence of π from $\hat{\pi}$ for Gauss-Hermite bases of order $p = 6$ and 9, while varying the scale parameter σ in (40). Since $g_i(\boldsymbol{\xi}) \sim N(0.5, \sigma^2)$, the scale parameter controls the standard deviation of the input uncertainty to the forward problem. The Gauss-Hermite results exhibit a common dependence on σ for all three source locations. Posterior errors increase at small σ and large σ, but are minimized at intermediate values of σ, e.g., $\sigma = 10^{-1}$. If σ is very small, the input distribution is narrowly centered on $(0.5, 0.5)$ and the posterior distribution favors values of $\mathbf{m} = \mathbf{g}(\boldsymbol{\xi})$ that lie on the edges of this input distribution. Errors in the forward solution at these values of $\boldsymbol{\xi}$ receive little weight in the $L^2(\Omega)$ norm and thus lead to errors in the posterior. On the other hand, with large σ (e.g., $\sigma > 0.1$) the input distribution of $\mathbf{g}(\boldsymbol{\xi})$ broadens to include appreciable tails outside the square domain. While values of $\chi = \mathbf{m}$ outside the unit square S are perfectly feasible according to (30), the $L^2(\Omega)$ norm then penalizes errors in the stochastic forward solution for these values at the expense of errors inside the domain. As a result, posterior errors again increase.

In general, $p = 9$ results show smaller errors than those at $p = 6$. The errors at $p = 9$ and small σ with $\mathbf{m}_{\text{true}} = (0.50, 0.50)$ appear to be an exception, reflecting errors at very large $|\boldsymbol{\xi}|$. We find that the ratio of the $p = 9$ and $p = 6$ posterior divergences at a given σ varies with the realization of the measurement noise, and thus the trends in Figure 17 should be generalized to other data in a qualitative sense.

Figure 17 also compares Gauss-Hermite posterior divergences to those obtained with uniform-Legendre bases of the same order—shown as horizontal lines, since we fix $g_i(\boldsymbol{\xi}) \sim U(0,1)$ in the uniform-Legendre case. Again, these results show strong dependence on the source location. For a source (and thus a posterior density) in the center of the domain, it is possible to obtain lower posterior divergences with Gauss-Hermite PC at a given order than with uniform-Legendre PC. With an appropriately-chosen scale parameter, this is again possible for $\mathbf{m}_{\text{true}} = (0.25, 0.75)$ at $p = 6$, but the uniform-Legendre basis proves more accurate at higher order ($p = 9$). And for a source near the edge of the domain, the uniform-Legendre basis provides better accuracy than the Gauss-Hermite basis at all values of the scale parameter and both orders.

Implications of this exercise on the choice of transformation $\mathbf{g}(\boldsymbol{\xi})$ and on the choice of PC basis are as follows. One should avoid transformations or bases for which "true" value(s) of $\mathbf{m}$ correspond to values of $\boldsymbol{\xi}$ that have small probability. The true values of $\mathbf{m}$ are of course unknown *a priori*, so a useful guideline is to ensure that $\mathbf{g}(\boldsymbol{\xi})$ has distribution equal to the prior, approximately equal to the prior, or with longer tails than the prior. Then the posterior probability mass is likely to fall within regions where forward problem errors are penalized with sufficient weight in the $L^2(\Omega)$ norm. Of course, it is possible for very large quantities of data to overwhelm the prior. In this case, if it is apparent that the posterior is tending towards regions of $\boldsymbol{\xi}$ that have been accorded small probability—as in Figures 13–14(c), for instance—a new choice of $\mathbf{g}$ or PC basis would be indicated.

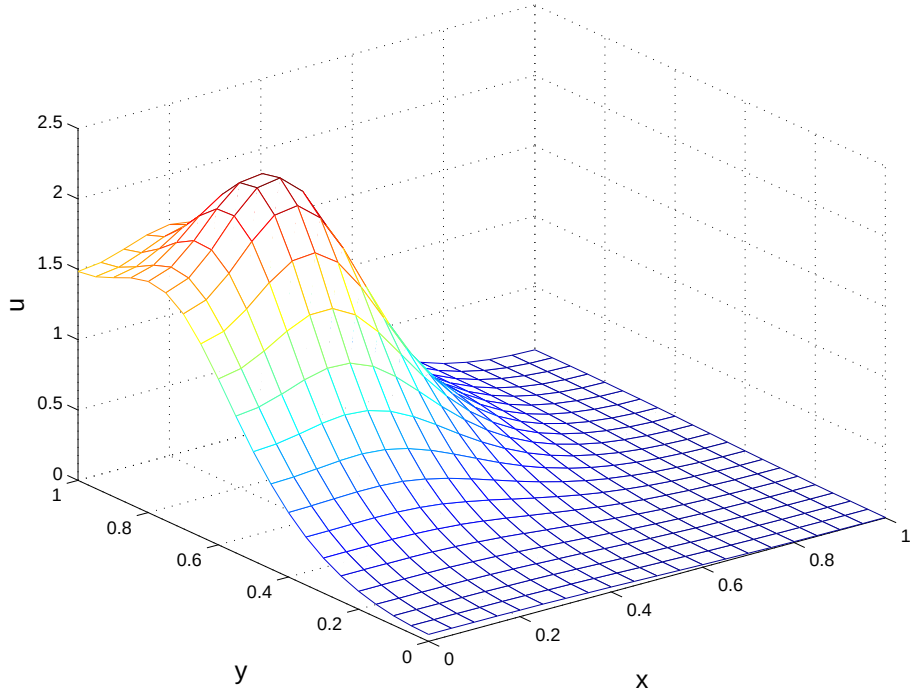
1.4 Conclusions

Bayesian inference provides an attractive setting for the solution of inverse problems. Measurement errors, forward model uncertainties, and complex prior information can all be combined to yield a rigorous and quantitative assessment of uncertainty in the inverse solution. Obtaining useful information from this posterior density—e.g., computing expectations or marginal distributions of the unknown parameters—may be a computationally expensive undertaking, however. For complex forward models, such as those that arise in inverting systems of PDEs, the cost of likelihood evaluations may render the Bayesian approach prohibitive.

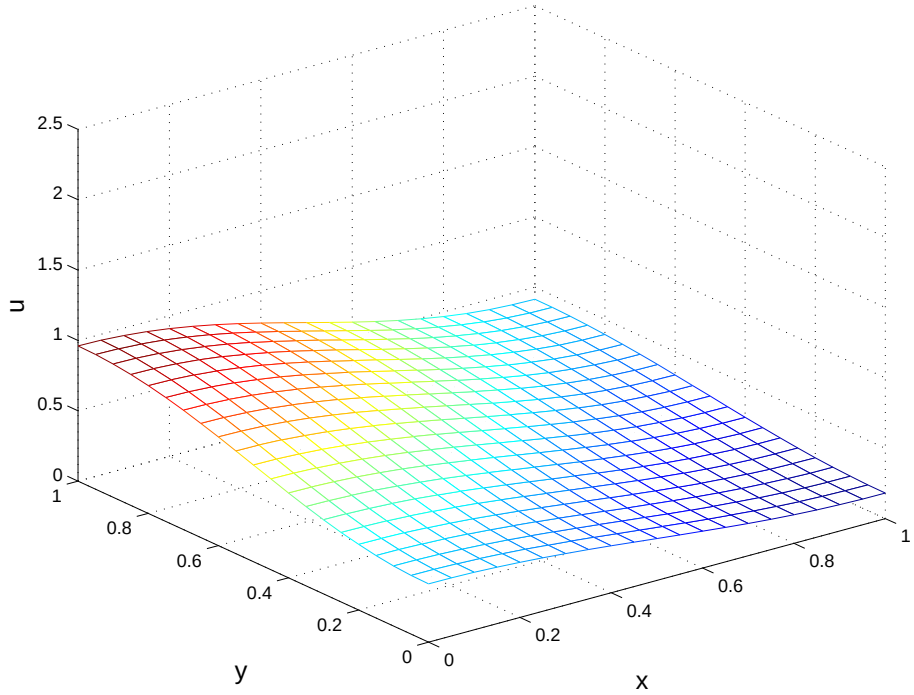
The theoretical developments in this report fundamentally accelerate Bayesian inference in computationally intensive inverse problems. We present a reformulation of the Bayesian approach based on polynomial chaos representations of random variables and associated spectral methods for efficient uncertainty propagation. Uncertain inputs that span the range of the prior define a stochastic forward problem; a Galerkin solution of this problem with the PC basis yields a spectral representation of uncertain forward model predictions. Evaluation of integrals over the unknown parameter space is then recast as sampling of the random variables $\boldsymbol{\xi}$ underlying the PC expansion, with significant speedup. In particular, we introduce three schemes for exploring the posterior: Monte Carlo sampling from the prior distribution, Monte Carlo sampling from an alternate distribution that includes the support of the prior, and Markov chain Monte Carlo in $\boldsymbol{\xi}$-space. Each of these schemes is compatible with partitioning of the prior support.

The new approach is demonstrated on a transient diffusion problem arising in contaminant source inversion. Spectral representation is found to reduce the cost of each posterior evaluation by three orders of magnitude, so that sampling of the PC-reformulated problem has nearly negligible cost. Error in the surrogate posterior decreases rapidly with increasing order of the PC basis; in the present case, convergence is exponentially fast. MCMC sampling of the posterior offers considerable freedom in choosing the PC basis and the initial transformation defining the stochastic forward problem, but a detailed exploration of posterior errors suggests guidelines for ensuring accuracy and computational efficiency.

Ongoing work will extend the polynomial chaos approach to more complex inverse problems. For instance, forward models with additional parametric uncertainty—parameters that may be marginalized in the posterior—should be quite amenable to PC acceleration. We also plan to explore stochastic spectral approaches to significantly higher-dimensional inverse problems, e.g., with spatially-extended input parameters. A further extension involves convective source inversion problems, with the associated challenges of spectral uncertainty propagation in nonlinear advection equations.

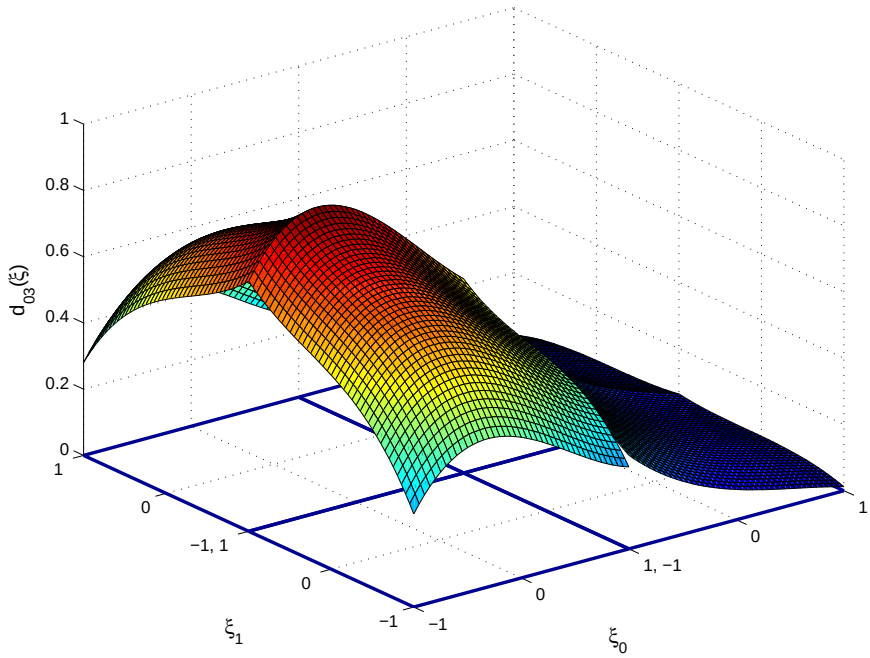


(a) $t = 0.05$

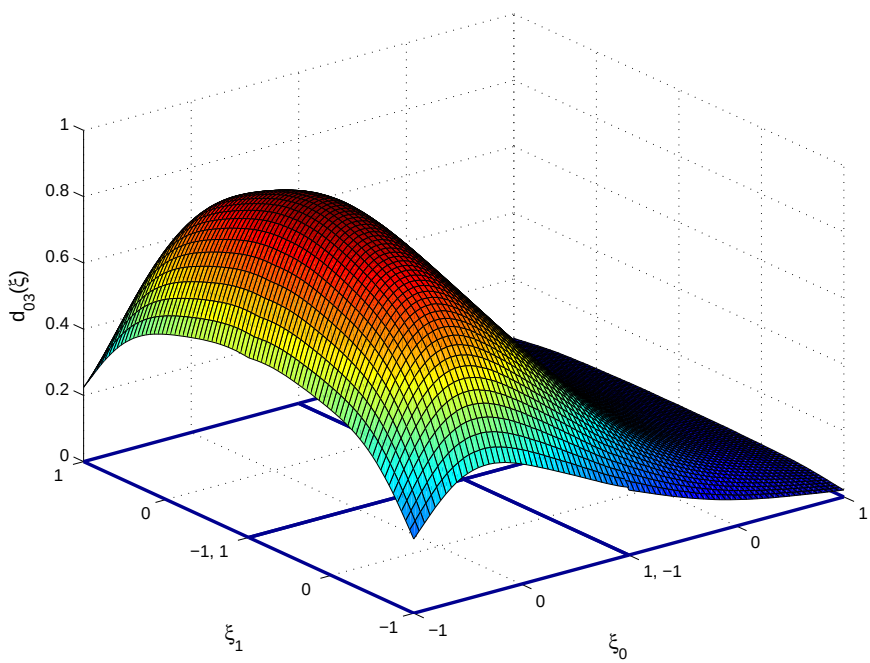


(b) $t = 0.15$

Figure 1. Scalar field u in the deterministic forward problem, for $(m_0, m_1) = (0.25, 0.75)$.

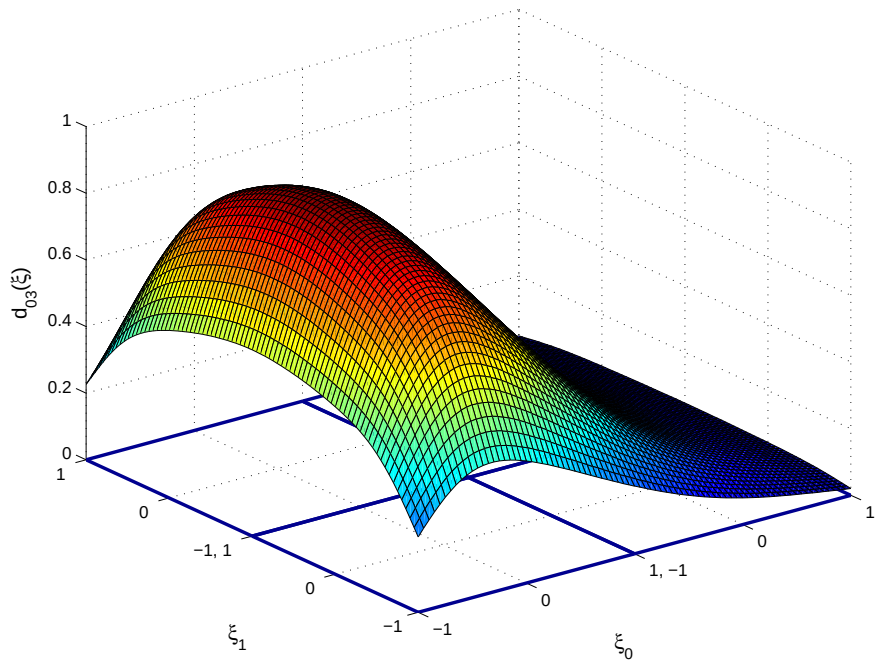


(a) $p = 3$



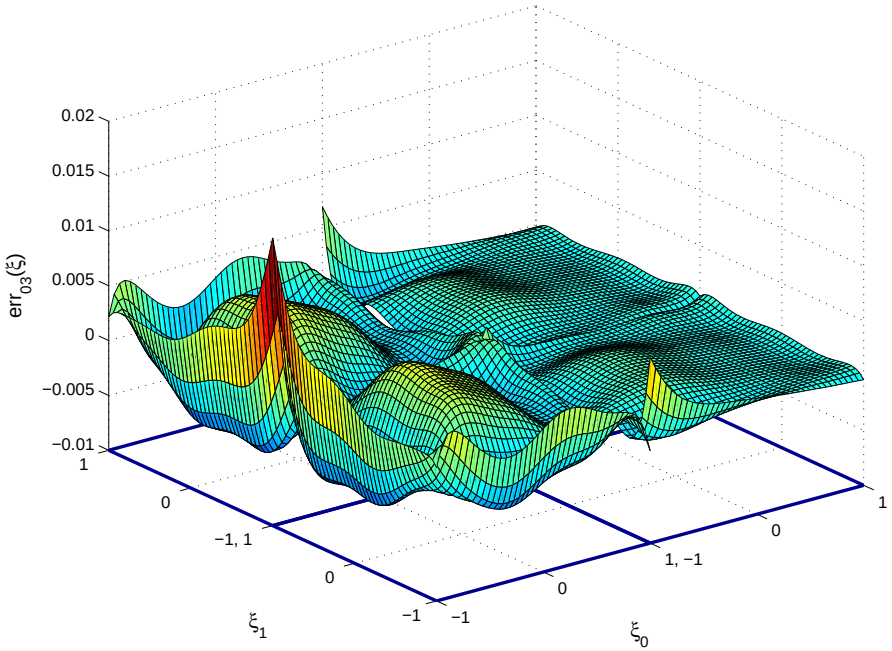
(b) $p = 6$

Figure 2. A single component of the stochastic forward problem solution, $\tilde{G}_3(\boldsymbol{\xi})$, shown as a surface response on four quadrants of prior support.

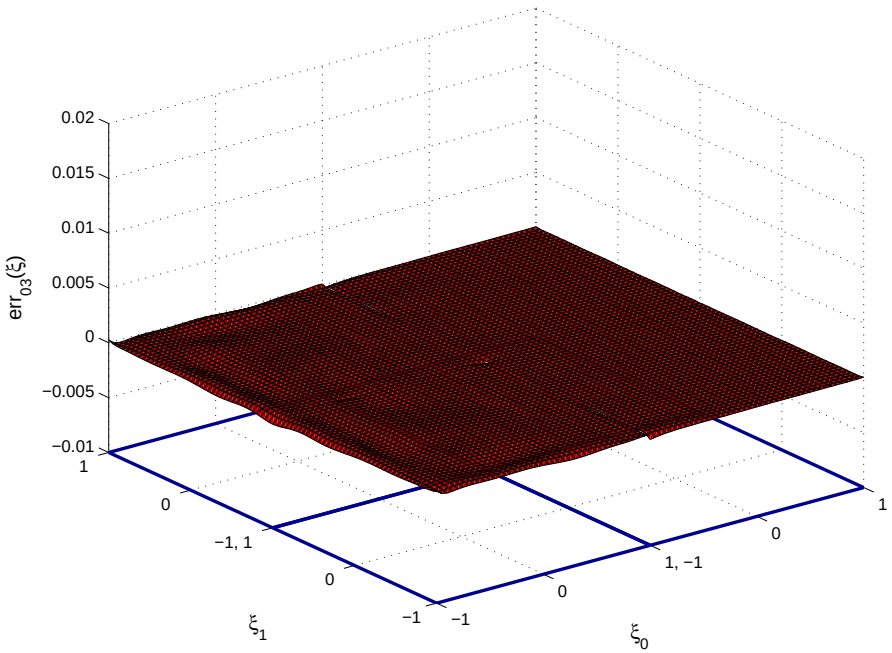


(c) $p = 9$

Figure 2. (cont.) Surface response.



(a) $p = 6$



(b) $p = 9$

Figure 3. Pointwise error in the solution to the stochastic forward problem, defined in (33).

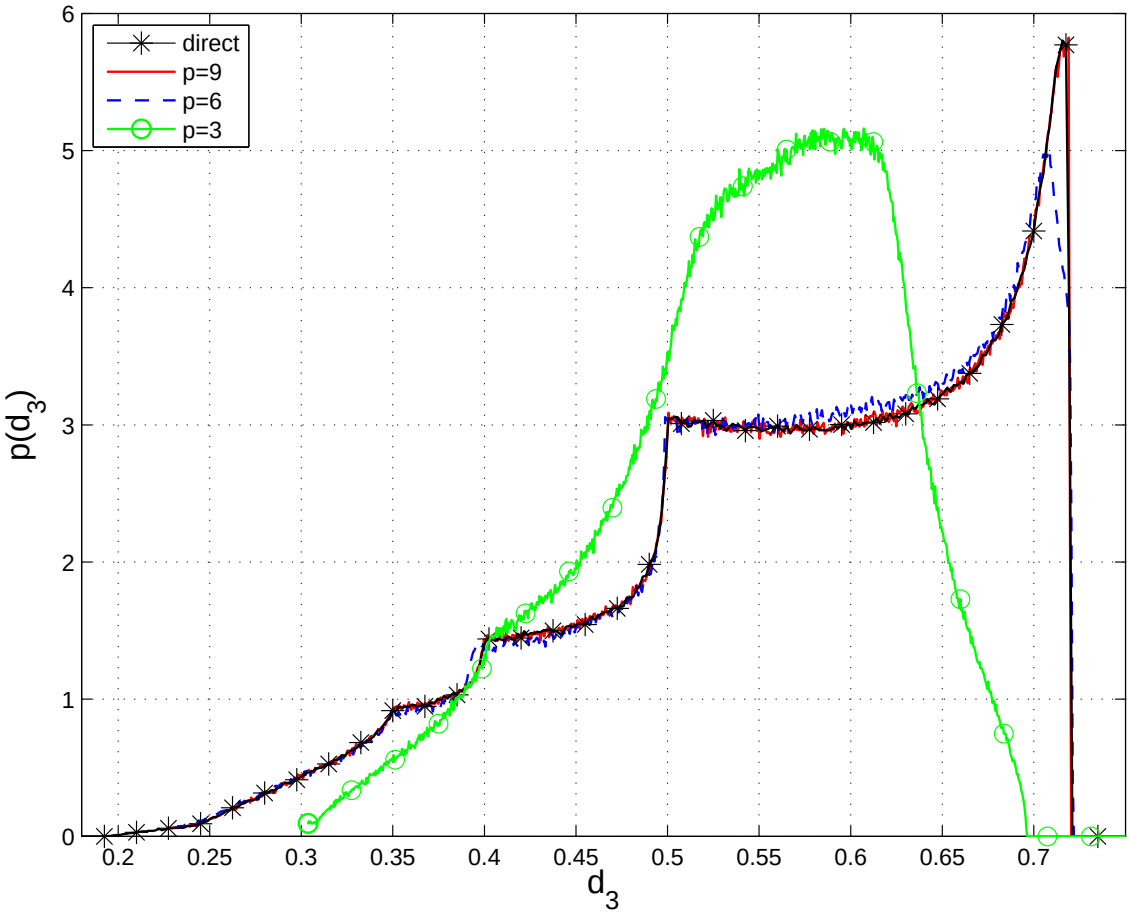
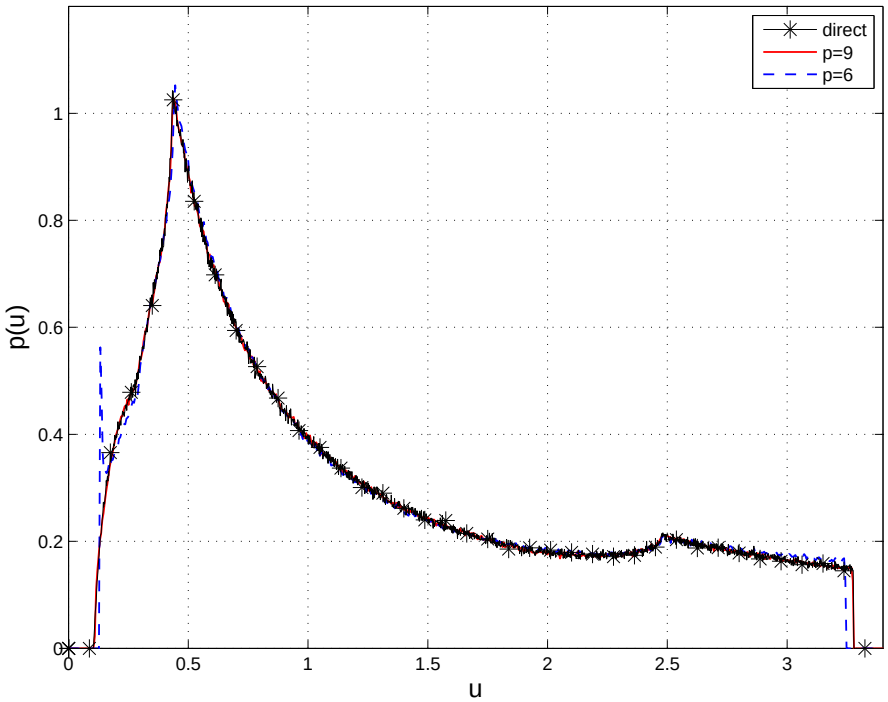
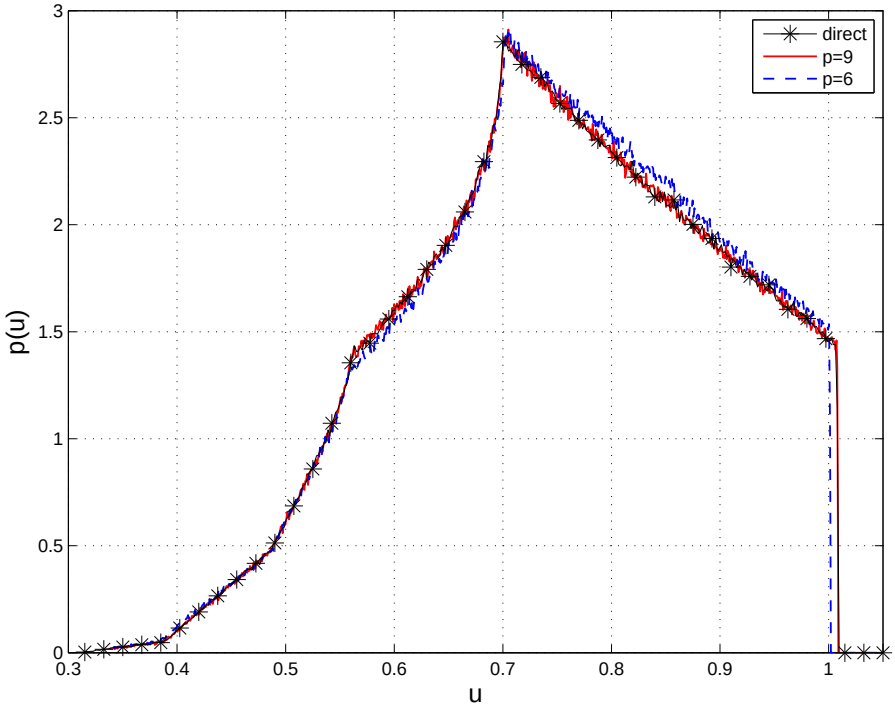


Figure 4. Probability density of a single measurement, $G_3(\mathbf{m}) \equiv u(x = 0.0, y = 0.5, t = 0.15)$, given prior uncertainty in the source location. Densities result from sampling PC expansions at varying order; compared to direct method.

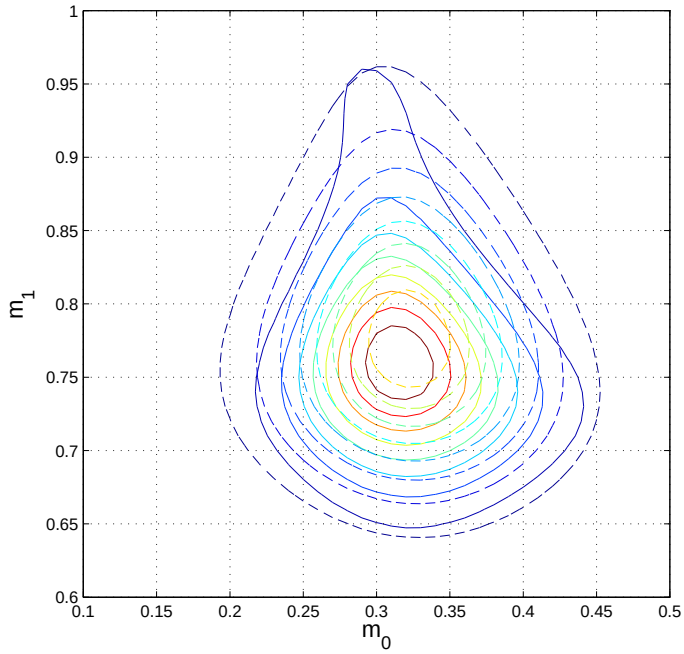


(a) $t = 0.05$

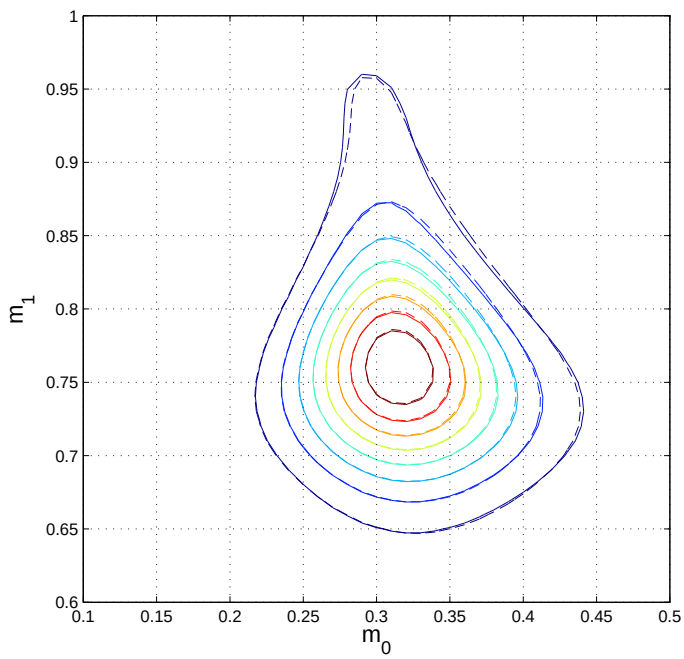


(b) $t = 0.15$

Figure 5. Probability density of the scalar $u(x = 0.0, y = 0.0)$ at two successive times, given prior uncertainty in the source location.

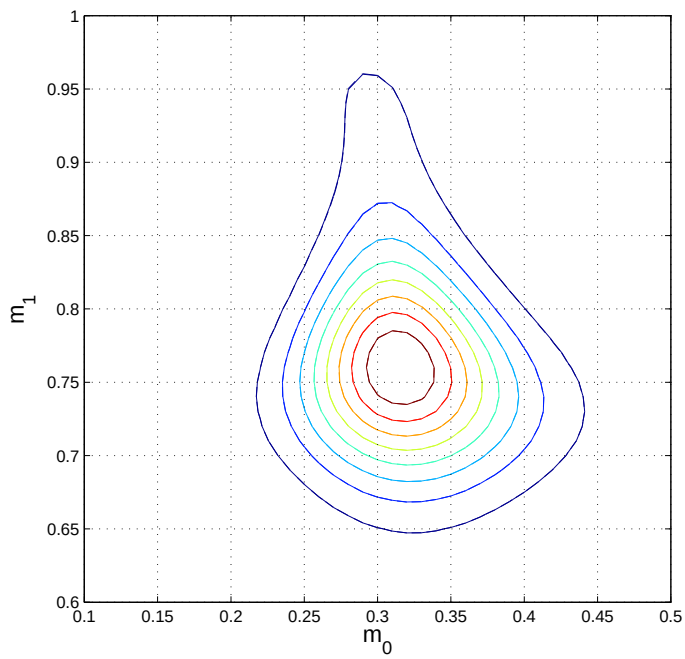


(a) $p = 3$



(b) $p = 6$

Figure 6. Contours of the posterior density of source location, $p(m_1, m_2|\mathbf{d})$. Solid lines are obtained via direct evaluations of the forward problem; dashed lines are obtained via polynomial chaos expansions.



(c) $p = 9$

Figure 6. (cont.) Contours of the posterior density of source location, $p(m_1, m_2|\mathbf{d})$. Solid lines are obtained via direct evaluations of the forward problem; dashed lines are obtained via polynomial chaos expansions.

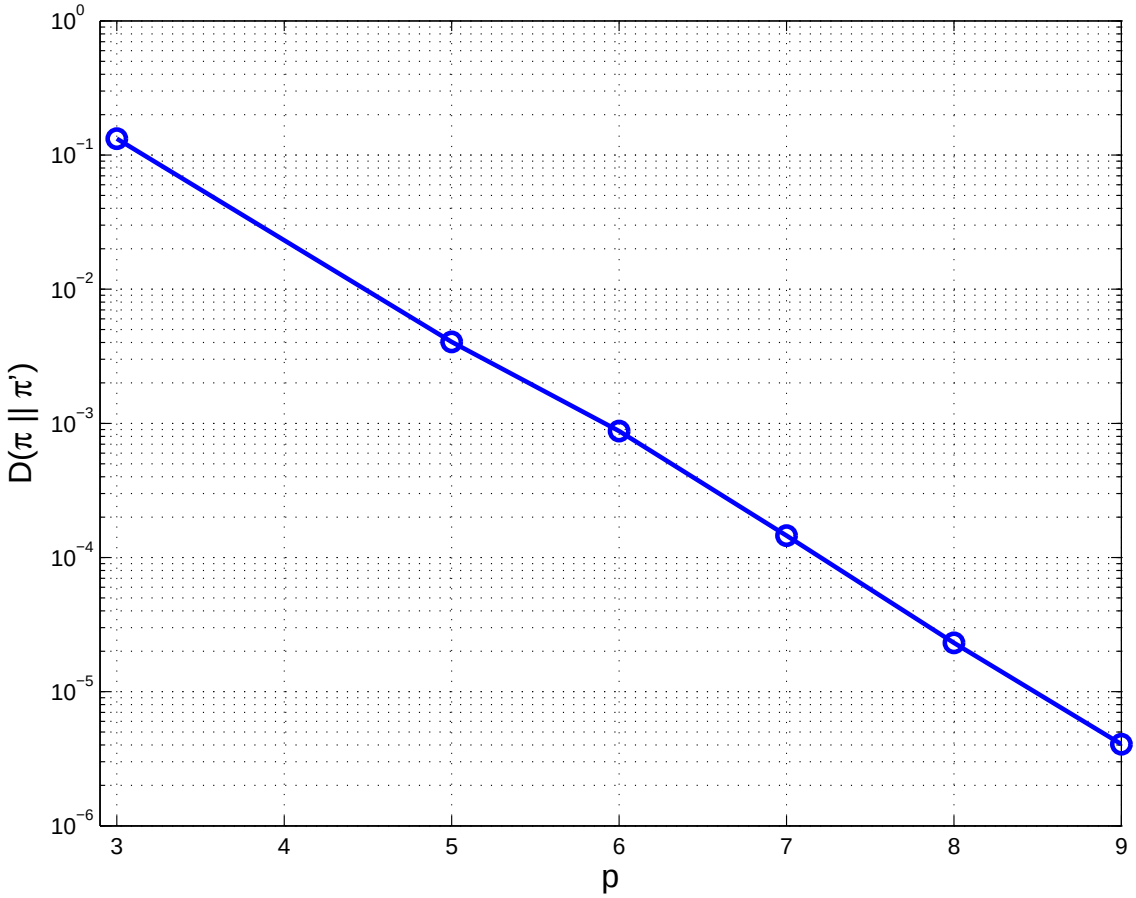


Figure 7. Kullback-Leibler distance $D(\tilde{\pi}||\pi)$ from the direct posterior to the PC-reformulated posterior, versus p.

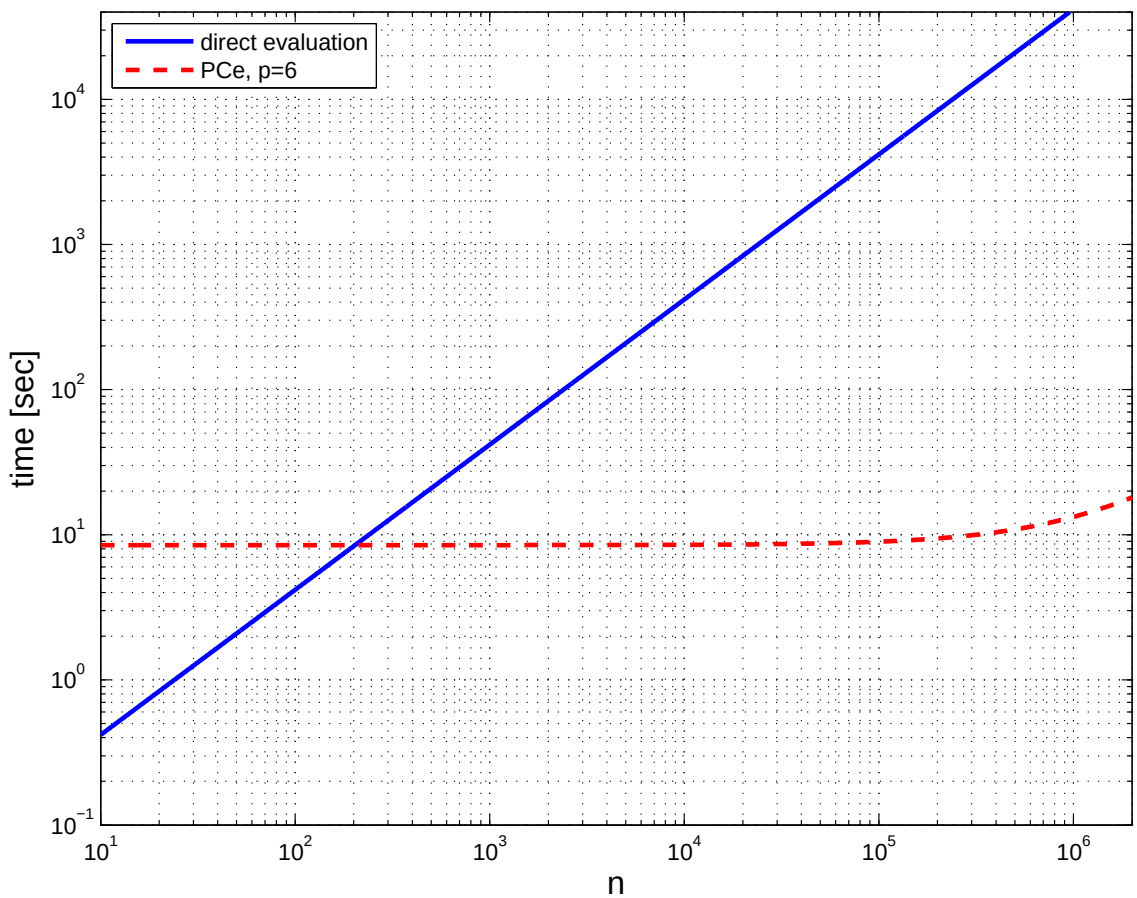


Figure 8. Computational time for Monte Carlo simulation of the posterior mean; direct evaluation versus PCe.

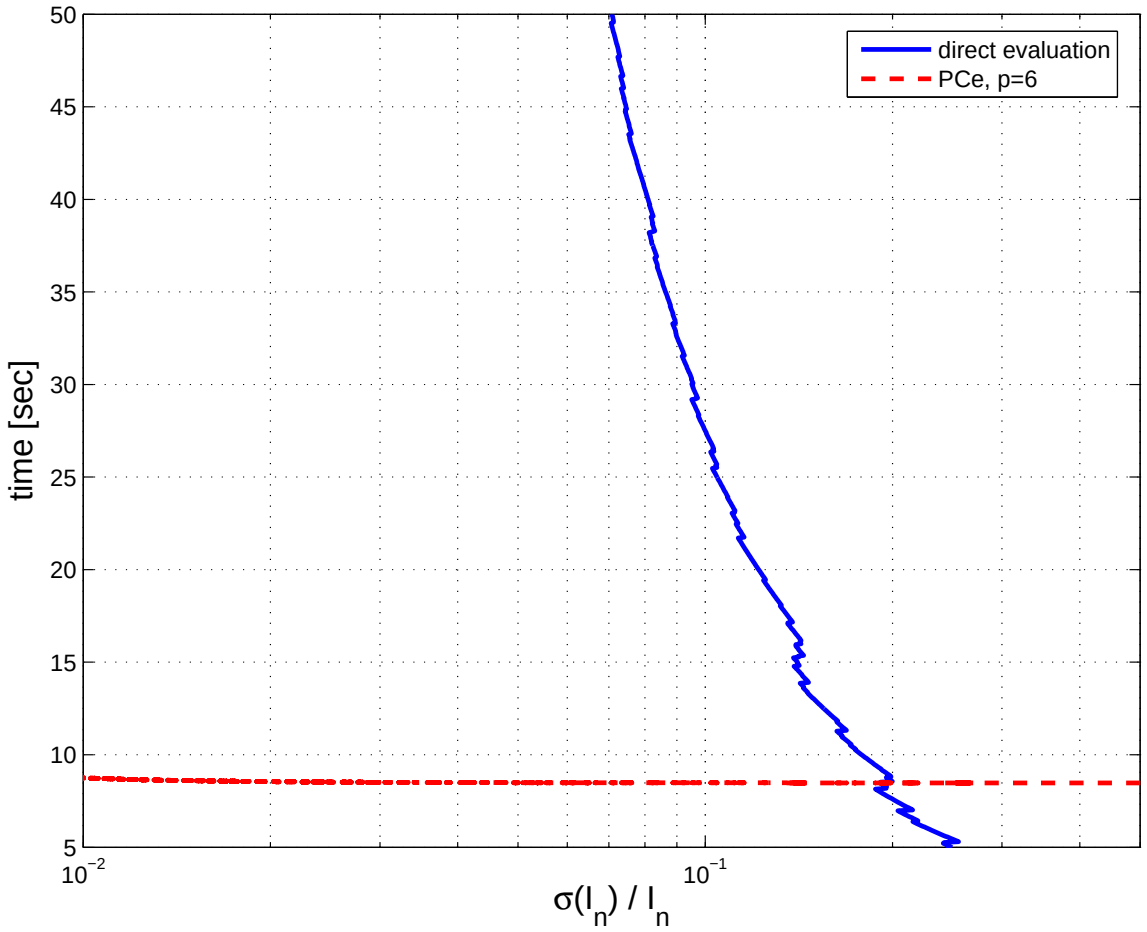
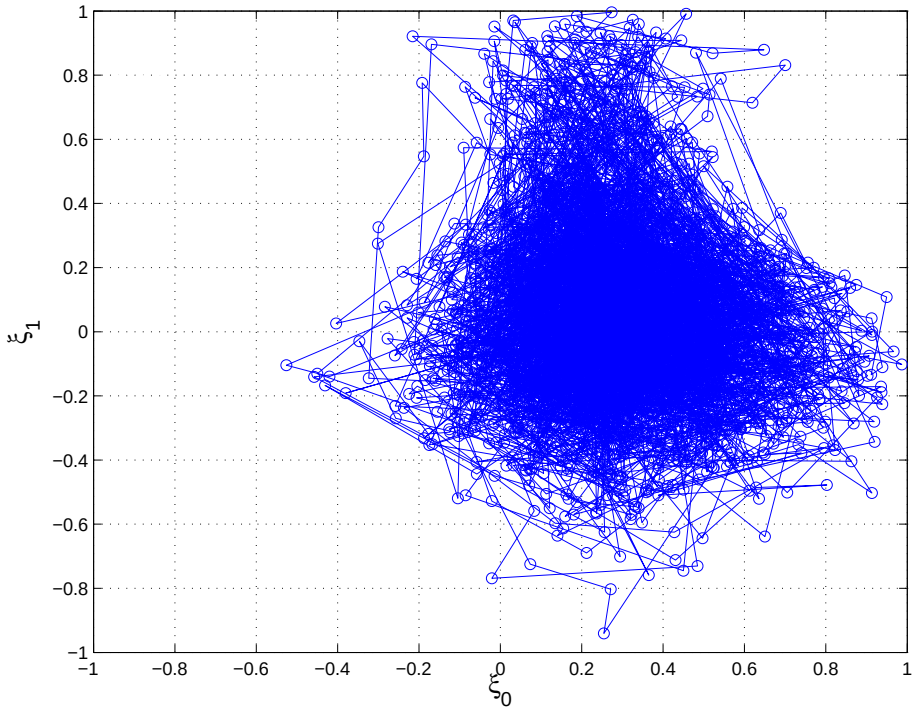
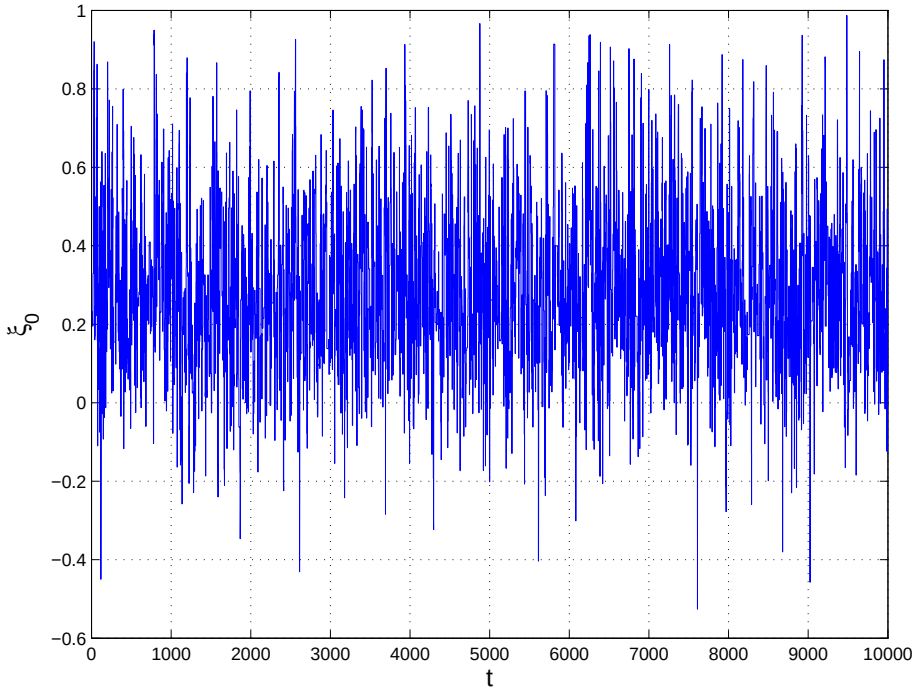


Figure 9. Relative error in Monte Carlo estimates of the posterior mean versus computational time, sampling from the prior; direct evaluation versus PCe.



(a) 2-D



(b) 1-D

Figure 10. MCMC chain in $\boldsymbol{\xi}$-space over 10,000 iterations; plotted in two dimensions and for a single component.

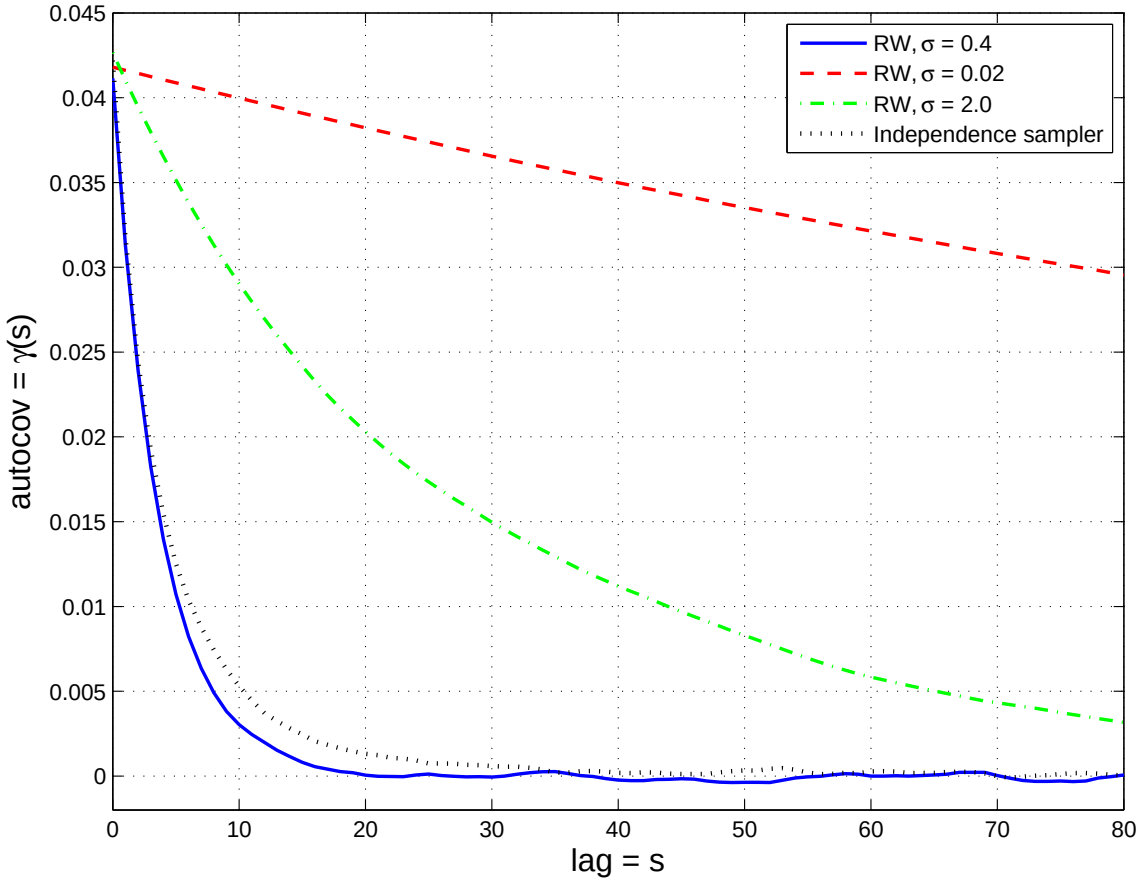
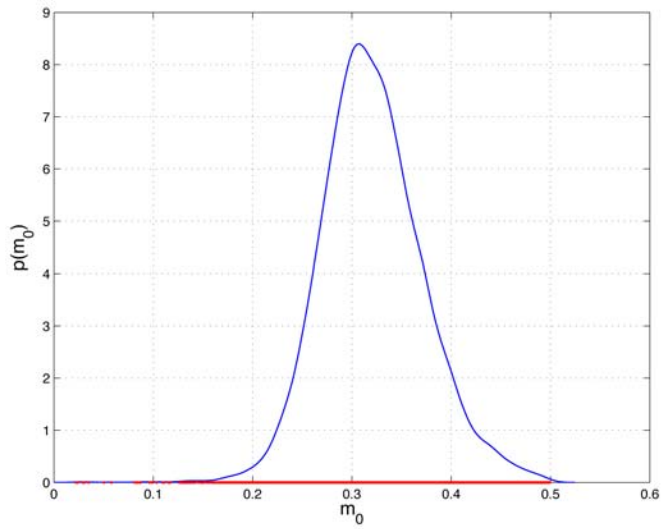
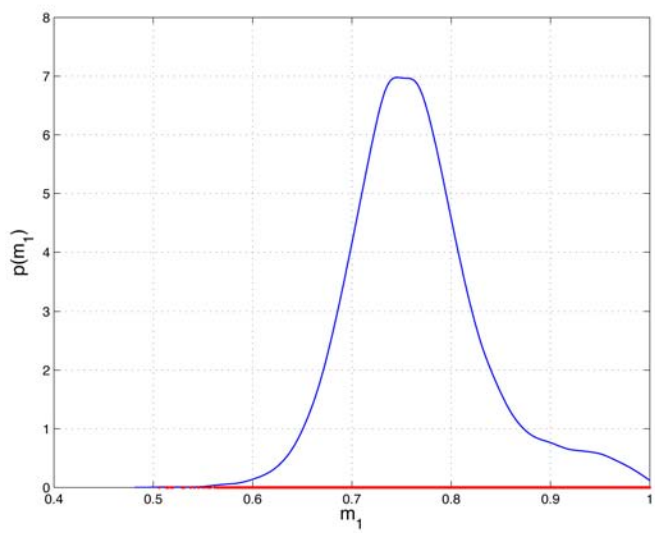


Figure 11. Autocovariance at lag s with various MCMC samplers.

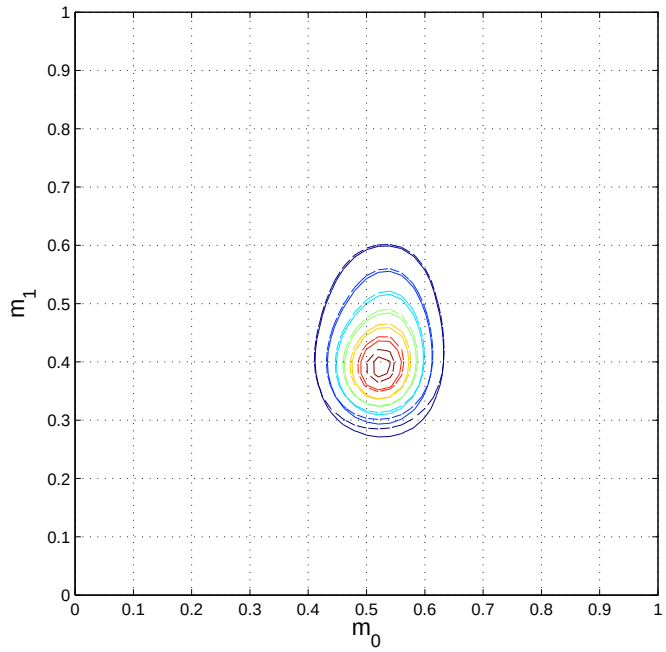


(a) m_0

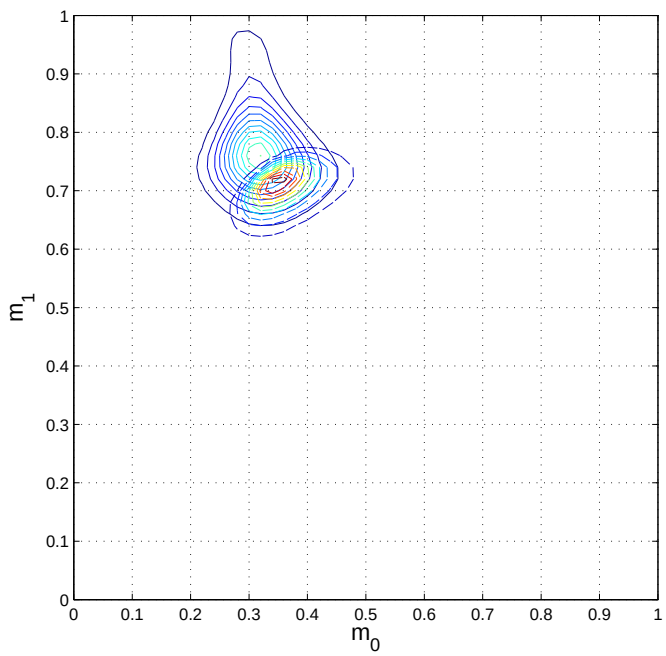


(b) m_1

Figure 12. Marginal distributions obtained with kernel density estimation.

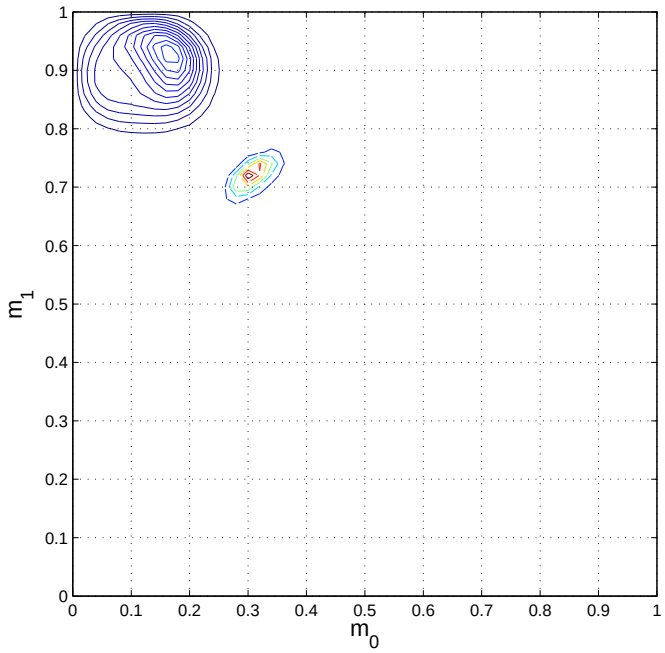


(a) $\mathbf{m}_{\text{true}} = (0.50, 0.50)$



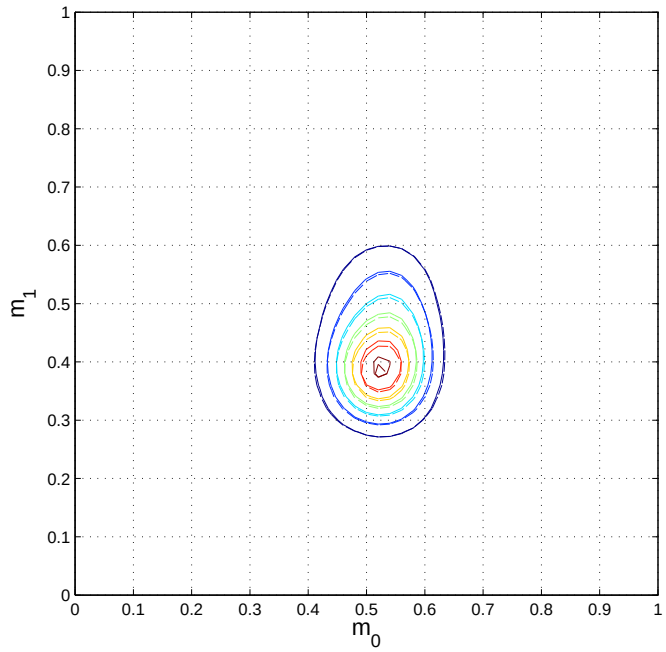
(b) $\mathbf{m}_{\text{true}} = (0.25, 0.75)$

Figure 13. Contours of posterior density. Gauss-Hermite PC, $p = 6$; $\sigma = 10^{-2}$. Solid lines are obtained with direct evaluations of the forward problem; dashed lines are obtained with polynomial chaos expansions.

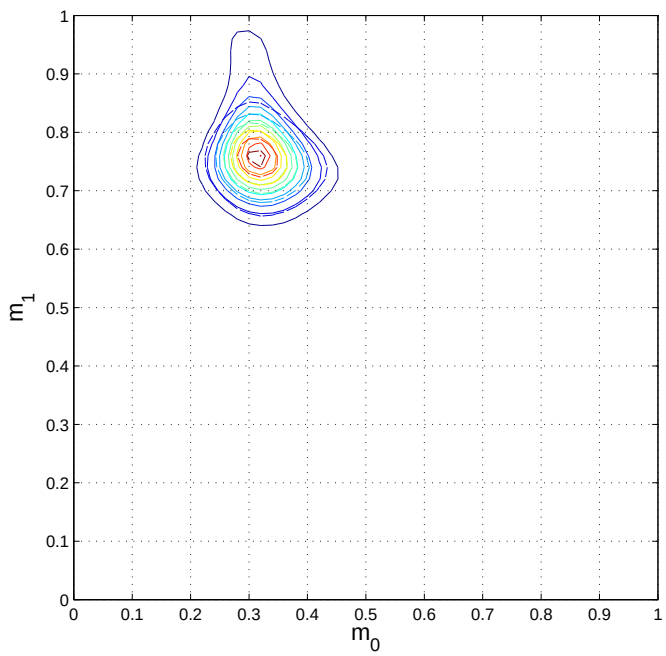


(c) $\mathbf{m}_{\text{true}} = (0.10, 0.90)$

Figure 13. (cont.) Contours of posterior density. Gauss-Hermite PC, $p = 6$; $\sigma = 10^{-2}$. Solid lines are obtained with direct evaluations of the forward problem; dashed lines are obtained with polynomial chaos expansions.

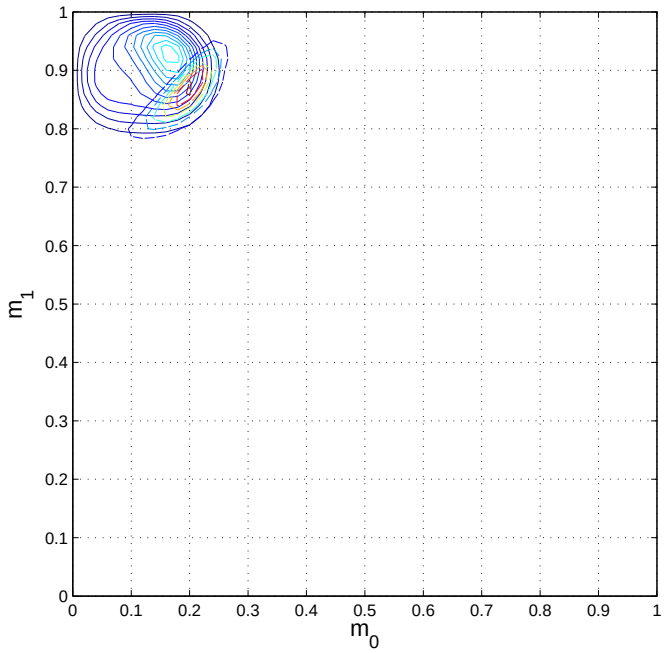


(a) $\mathbf{m}_{\text{true}} = (0.50, 0.50)$



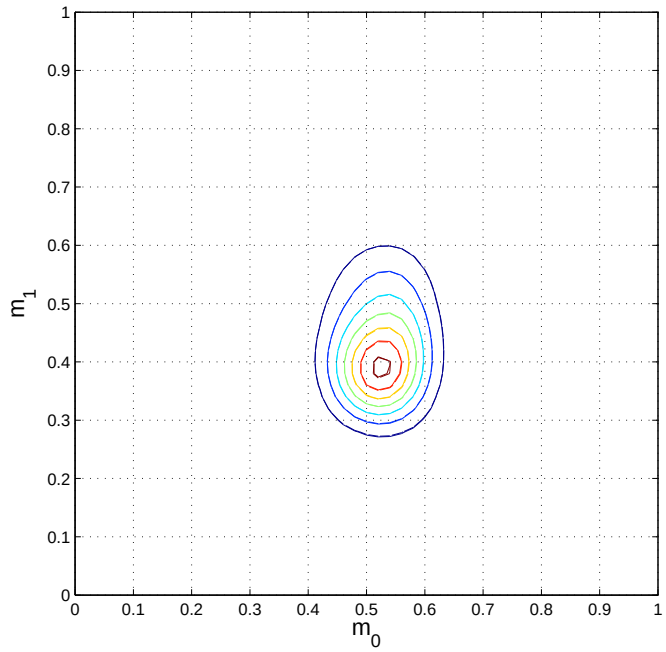
(b) $\mathbf{m}_{\text{true}} = (0.25, 0.75)$

Figure 14. Contours of posterior density. Gauss-Hermite PC, $p = 6$; $\sigma = 10^{-1}$. Solid lines are obtained with direct evaluations of the forward problem; dashed lines are obtained with polynomial chaos expansions.

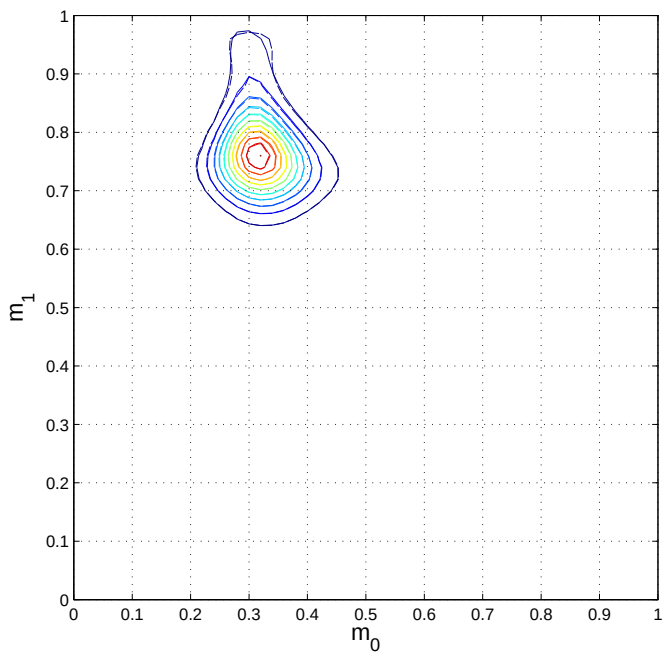


(c) $\mathbf{m}_{\text{true}} = (0.10, 0.90)$

Figure 14. (cont.) Contours of posterior density. Gauss-Hermite PC, $p = 6$; $\sigma = 10^{-1}$. Solid lines are obtained with direct evaluations of the forward problem; dashed lines are obtained with polynomial chaos expansions.

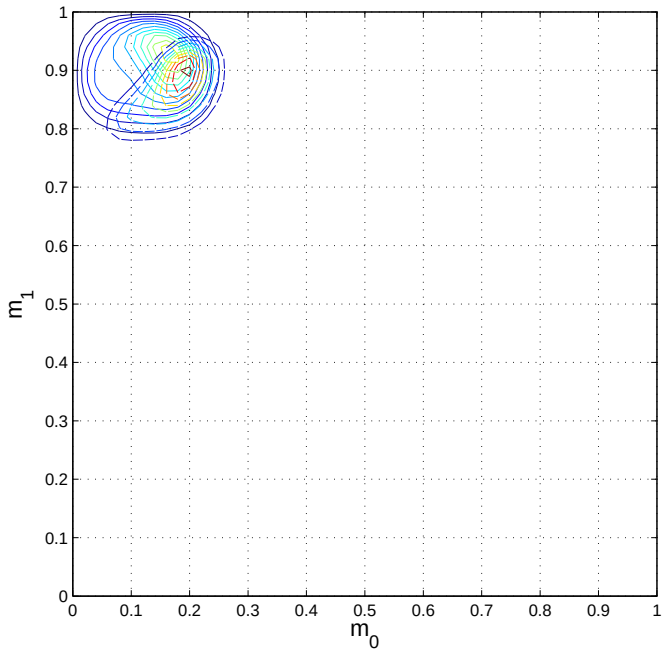


(a) $\mathbf{m}_{\text{true}} = (0.50, 0.50)$



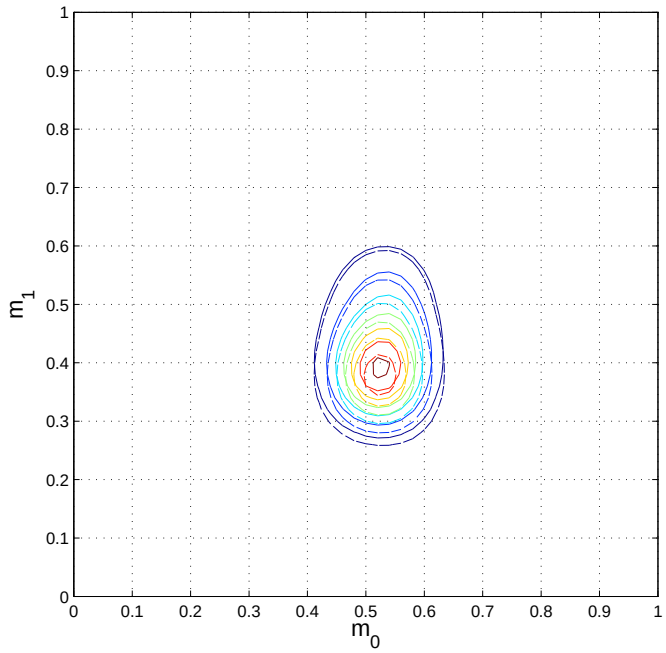
(b) $\mathbf{m}_{\text{true}} = (0.25, 0.75)$

Figure 15. Contours of posterior density. Gauss-Hermite PC, $p = 9$; $\sigma = 10^{-1}$. Solid lines are obtained with direct evaluations of the forward problem; dashed lines are obtained with polynomial chaos expansions.

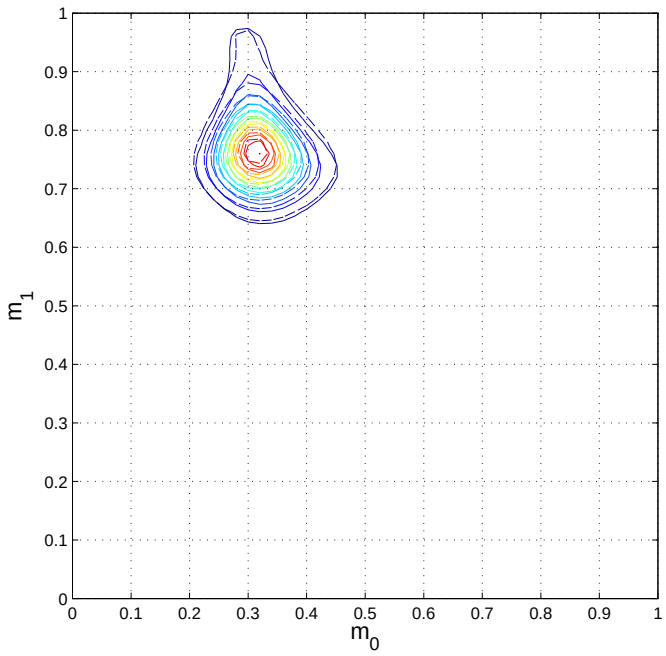


(c) $\mathbf{m}_{\text{true}} = (0.10, 0.90)$

Figure 15. (cont.) Contours of posterior density. Gauss-Hermite PC, $p = 9$; $\sigma = 10^{-1}$. Solid lines are obtained with direct evaluations of the forward problem; dashed lines are obtained with polynomial chaos expansions.

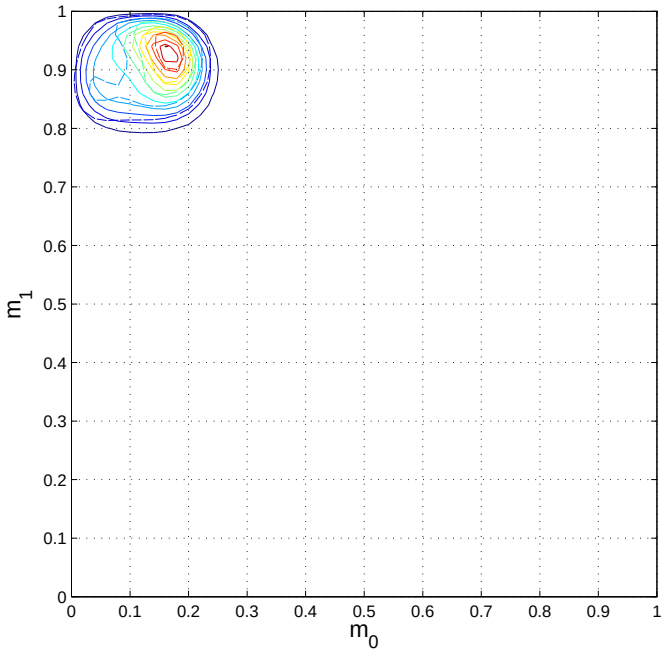


(a) $\mathbf{m}_{\text{true}} = (0.50, 0.50)$



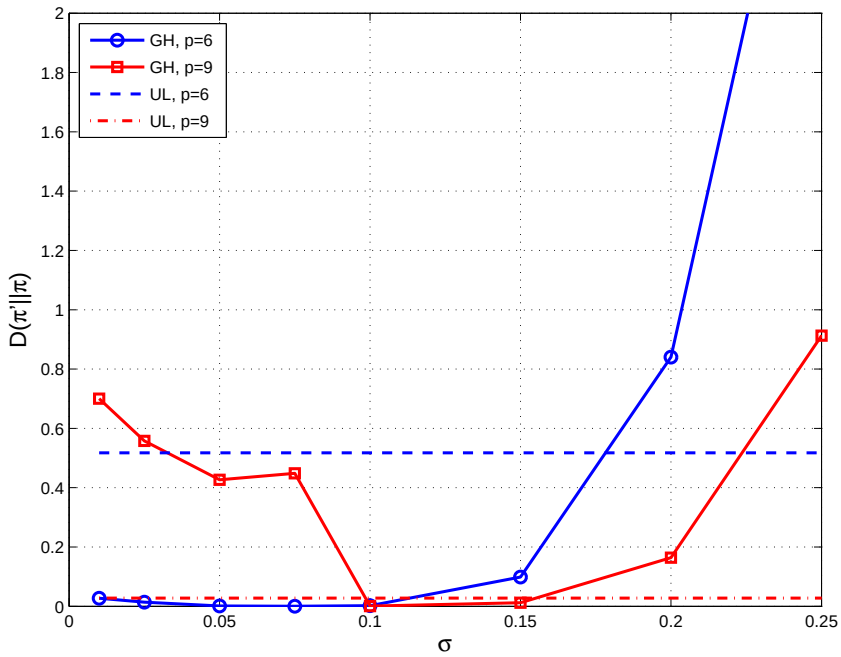
(b) $\mathbf{m}_{\text{true}} = (0.25, 0.75)$

Figure 16. Contours of posterior density. Uniform-Legendre PC, $p = 9$. Solid lines are obtained with direct evaluations of the forward problem; dashed lines are obtained with polynomial chaos expansions.

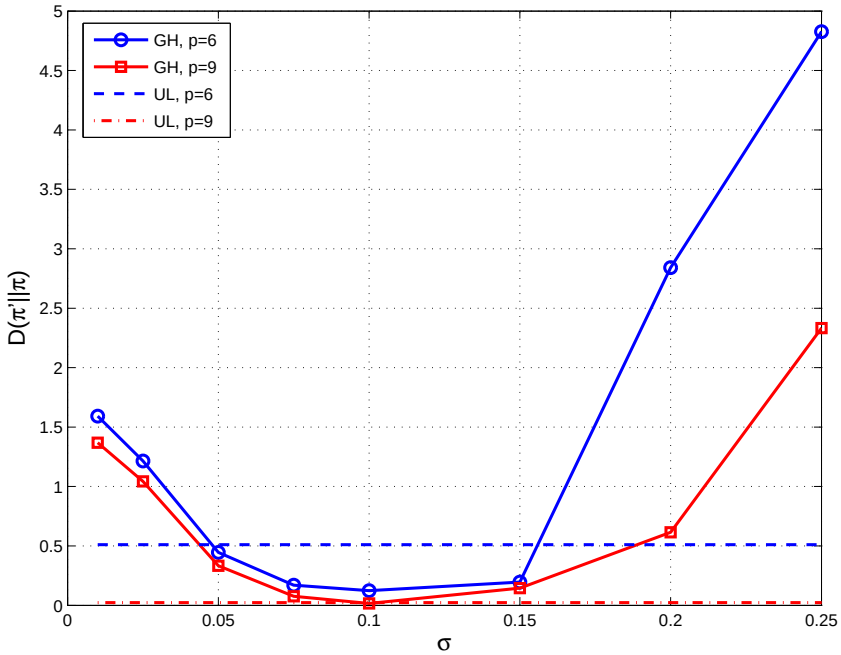


(c) $\mathbf{m}_{\text{true}} = (0.10, 0.90)$

Figure 16. (cont.) Contours of posterior density. Uniform-Legendre PC, $p = 9$. Solid lines are obtained with direct evaluations of the forward problem; dashed lines are obtained with polynomial chaos expansions.

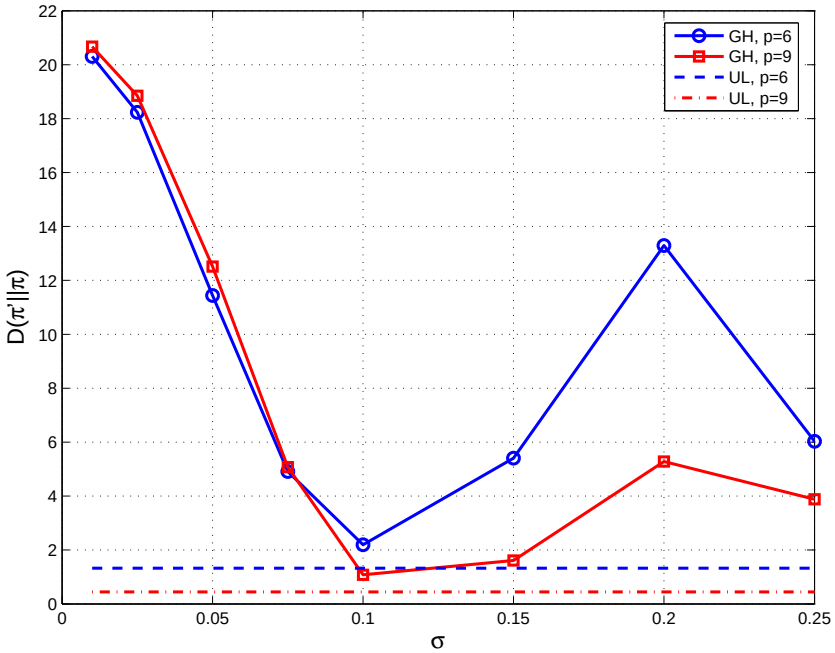


(a) $\mathbf{m}_{\text{true}} = (0.50, 0.50)$



(b) $\mathbf{m}_{\text{true}} = (0.25, 0.75)$

Figure 17. $D(\tilde{\pi}\|\pi)$ with Gauss-Hermite and uniform-Legendre PC at varying orders and source locations. Scale parameter σ is the standard deviation of input uncertainty to the forward problem in the Gauss-Hermite case.



(c) $\mathbf{m}_{\text{true}} = (0.10, 0.90)$

Figure 17. (cont.) $D(\tilde{\pi}\|\pi)$ with Gauss-Hermite and uniform-Legendre PC at varying orders and source locations. Scale parameter σ is the standard deviation of input uncertainty to the forward problem in the Gauss-Hermite case.

2 Bayesian inference of spatial fields with Gaussian process priors

2.1 Introduction

This chapter extends the the stochastic spectral methodology of §1 to inverse problems whose solutions are unknown functions—i.e., spatial or temporal fields. In doing so, we also explore dimensionality reduction in the Bayesian formulation of inverse problems, and the dependence of dimensionality on both the prior and the data. Inverse problems involving fields are vital to applications ranging from geophysics to medical imaging. Spatial fields may correspond to inhomogeneous material properties, such as permeabilities, diffusivities, or densities, or may represent distributed source terms in transport equations.

Estimating fields rather than parameters typically increases the ill-posedness of the inverse problem, since one is recovering an infinite-dimensional object from finite data. Obtaining physically meaningful results requires the injection of additional information on the unknown field—i.e., regularization [90]. A standard Bayesian approach is to employ Gaussian process (GP) or Markov random field (MRF) priors [62, 50, 99]. Most studies then explore the value of the field on a finite set of grid points [51]; the dimension of the posterior is tied to the discretization of the field. This recipe presents difficulties for stochastic spectral approaches, however, as the size of a PC basis does not scale favorably with dimension [40]. Moreover, with any degree of smoothness, the value of the field at each grid point hardly represents an independent direction.

Ideally, one should employ a representation that reflects how much information is truly required to capture variation among realizations of the unknown field. To this end, we introduce a Karhunen-Loève (K-L) expansion based on the prior random process, transforming the inverse problem to inference on a truncated sequence of weights of the K-L modes. Other recent work has also employed K-L expansions in the context of statistical inverse problems. Li & Cirpka [64] emphasize the role of K-L expansions in enabling geostatistical inversion on unstructured grids. Efendiev *et al.* [24] use K-L expansions to parameterize the log-permeability field in their two-stage MCMC scheme, and introduce constraints among the weights in order to match known values of the permeability at selected spatial locations. In contrast to [64], we use a fully Bayesian approach, generating exact conditional realizations from a non-Gaussian posterior.

A more fundamental distinction of the present work is that we combine a K-L representation of the prior process with spectral methods for uncertainty propagation. In particular, the Karhunen-Loève representation of a scaled Gaussian process prior *defines* the uncertainty that is propagated through the forward model with a stochastic Galerkin scheme. The deterministic forward model, originally specified by a (system of) partial differential equations, is thus replaced by stochastic PDEs; numerical approaches to such systems, in which random fields appear as boundary conditions or coefficients, have seen extensive development [61, 28]. Uncertainty propagation yields a polynomial approximation of the

forward operator over the support of the prior. This approximation then enters a reduced-dimensionality surrogate posterior, which we explore with MCMC. The overall scheme avoids repeated forward simulations, and the computational cost per MCMC iteration becomes negligible compared to the cost of a full forward solution.

We demonstrate our scheme with a nonlinear forward model, and develop a fully Bayesian treatment of the problem in which hyperparameters describing the prior covariance are estimated simultaneously with the unknown field in a joint posterior distribution. We explore convergence both with respect to the number of K-L modes and the order of the PC basis. We also examine the efficiency of MCMC, quantify the limiting distribution of the K-L modes, and explore the impact of data resolution on the approach to this distribution. To place the present Bayesian formulation in broader context, we recall connections between the K-L expansion and regularization penalties in the reproducing kernel Hilbert space (RKHS) norm corresponding to the prior covariance.

2.2 Formulation

2.2.1 Gaussian processes

Let $(\Omega, \mathcal{U}, P)$ be a probability space, where Ω is a sample space, $\mathcal{U}$ is a σ-algebra over Ω, and P is a probability measure on $\mathcal{U}$. Also, let $D \subset \mathbb{R}^n$ be a bounded spatial domain. If $M(\mathbf{x}) : \Omega \to \mathbb{R}$ is a $\mathcal{U}$-measurable mapping for every $\mathbf{x} \in D$, then $M : \Omega \times D \to \mathbb{R}$ is a random field. $M(\mathbf{x}, \omega)$, for $\omega \in \Omega$, can thus be seen as a collection of real-valued random variables indexed by $\mathbf{x} \in D$. Alternatively, one can view $M(\cdot)$ as a random variable taking values in $\mathbb{R}^D$, the space of all real-valued functions on D. Though our presentation will focus on 'random fields' (typically signifying processes indexed by a spatial coordinate), the developments below are applicable to processes indexed by time or by both time and space.

If, for any $n \geq 1$ we have

$$(M(\mathbf{x}_1), \ldots, M(\mathbf{x}_n)) \overset{i.d.}{=} (M(\mathbf{x}_1 + \mathbf{s}), \ldots, M(\mathbf{x}_n + \mathbf{s})) \tag{41}$$

where $\overset{i.d.}{=}$ denotes equality in distribution, $\mathbf{s}$ is a spatial shift, and $\{\mathbf{x}_i, \mathbf{x}_i + \mathbf{s}\}_{i=1}^n \in D$, then M is said to be stationary [45]. If in addition, all finite-dimensional distributions of M are multivariate normal, then M is a *stationary Gaussian random field*, or simply a stationary Gaussian process (GP). Let $\mathbf{M}_{(n)} = (M(\mathbf{x}_1), \ldots, M(\mathbf{x}_n))$ denote the restriction of M to a finite set of indices. Then the characteristic function of $\mathbf{M}_{(n)}$ is [45]

$$\phi_M(\boldsymbol{\lambda}) \equiv \mathbb{E}\left[\exp\left(i\boldsymbol{\lambda}^T \mathbf{M}_{(n)}\right)\right] = \exp\left(i\boldsymbol{\lambda}^T \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\lambda}^T \boldsymbol{\Sigma} \boldsymbol{\lambda}\right), \quad \boldsymbol{\lambda} \in \mathbb{R}^n \tag{42}$$

where the mean is spatially invariant, $\boldsymbol{\mu} \equiv \mu \mathbf{1}_n$, and entries of $\boldsymbol{\Sigma}$ are values of the covariance

function C:

$$\begin{aligned}\Sigma_{ij} &= C(\mathbf{x}_i, \mathbf{x}_j) \\ &\equiv \mathrm{Cov}\left[M(\mathbf{x}_i), M(\mathbf{x}_j)\right] = \mathbb{E}\left[(M(\mathbf{x}_i) - \mu)\left(M(\mathbf{x}_j) - \mu\right)\right] \quad (43) \\ &= \tilde{C}(\mathbf{x}_i - \mathbf{x}_j) \quad (44)\end{aligned}$$

Gaussian processes have finite second moments; that is, $M(\mathbf{x}) \in L^2(\Omega)$ for every $\mathbf{x}$ [47]. If Σ is invertible, the finite-dimensional density of order n of the Gaussian process is then

$$p(\mathbf{m}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{m} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{m} - \boldsymbol{\mu})\right) \quad (45)$$

where $\mathbf{m} = (m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n))$. If we further restrict C to depend only on the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, that is we put $\tilde{C}(\mathbf{d}) = f(\|\mathbf{d}\|)$, then the stationary GP is called *isotropic* [84]. It is common to specify the covariance function with scale and range parameters θ_1 and θ_2 respectively [62]:

$$\tilde{C}(\mathbf{d}) = \theta_1 \rho\left(\frac{\|\mathbf{d}\|}{\theta_2}\right) \quad (46)$$

Here $\rho(\cdot)$ is a correlation function, positive definite with $\rho(0) = 1$ [46], e.g., $\rho(d) = e^{-d}$ or $\rho(d) = e^{-d^2}$.

Gaussian processes are extensively employed as priors in Bayesian inference [84]. In particular, conceiving of the GP as a prior over functions motivates Gaussian process regression [101], also known as kriging in spatial statistics [19]; further applications include classification, with ties to support vector machines and other kernel methods [78]. Depending on the covariance kernel, realizations of a Gaussian process may be smooth or periodic, or for non-stationary kernels, capture certain trends [33]. Gaussian process priors can thus inject regularity by assigning low probability to fields with undesirable properties. See §2.2.6 for a more formal discussion of Gaussian process priors and regularization.

2.2.2 Karhunen-Loève expansion

Let $M(\mathbf{x}, \omega)$ be a real-valued random field with finite second moments, mean $\mu(\mathbf{x})$, and a covariance function that is continuous on $D \times D$, with D bounded. Then M has the following representation, termed a Karhunen-Loève (K-L) expansion [65]:

$$M(\mathbf{x}, \omega) = \mu(\mathbf{x}) + \sum_{k=1}^{\infty} \sqrt{\lambda_k} c_k(\omega) \phi_k(\mathbf{x}) \quad (47)$$

In general, this equality holds pointwise and in the mean square sense; that is, convergence is in $L^2(\Omega)$ for each $\mathbf{x} \in D$. If M is Gaussian and almost surely continuous, then con-

vergence is uniform over D with probability one [2].[‡] λ_k and $\phi_k(\mathbf{x})$ are eigenvalues and eigenfunctions of the linear operator corresponding to the covariance kernel C:

$$\int_D C(\mathbf{x}_1, \mathbf{x}_2)\phi_k(\mathbf{x}_2)d\mathbf{x}_2 = \lambda_k \phi_k(\mathbf{x}_1) \tag{48}$$

By the assumptions on M, the covariance kernel is symmetric and positive semidefinite, and thus by Mercer's theorem we have [45, 18]

$$C(\mathbf{x}_1, \mathbf{x}_2) = \sum_{k=1}^{\infty} \lambda_k \phi_k(\mathbf{x}_1)\phi_k(\mathbf{x}_2) \tag{49}$$

where the eigenfunctions $\phi_k(\mathbf{x})$ are continuous and form a complete orthonormal system in $L^2(D)$. The random variables $c_k(\omega)$ are uncorrelated with zero mean and unit variance:

$$\mathbb{E}c_k = 0, \;\; \mathbb{E}\left[c_j c_k\right] = \delta_{jk} \tag{50}$$

These variables are in general non-Gaussian

$$c_k(\omega) = \frac{1}{\sqrt{\lambda_k}} \int_D \left(M(\mathbf{x}, \omega) - \mu(\mathbf{x})\right) \phi_k(\mathbf{x}) d\mathbf{x} \tag{51}$$

but if M is also a Gaussian process, the c_k are Gaussian and independent, $c_k \sim N(0,1)$.

The K-L expansion is *optimal* in the following sense. Of all possible orthonormal bases for $L^2(D)$, the $\{\phi_k(\mathbf{x})\}$ satisfying (48) minimize the mean-squared error in a finite linear representation of $M(\cdot)$ [40]. That is, they minimize

$$\int_{\Omega \times D} \left(M(\mathbf{x}, \omega) - \mu(\mathbf{x}) - \sum_{k=1}^{K} \sqrt{\lambda_k} c_k(\omega) \phi_k(\mathbf{x}) \right)^2 dP(\omega) d\mathbf{x} \tag{52}$$

for any $K \geq 1$. As a result, the K-L expansion is an extremely useful tool for the concise representation of stochastic processes. It has close analogues in data reduction (i.e., principal components analysis), model reduction (proper orthogonal decomposition) [86], and linear algebra (SVD). If $M(\cdot)$ is approximated by a K-term K-L expansion,

$$M_K(\mathbf{x}, \omega) = \mu(\mathbf{x}) + \sum_{k=1}^{K} \sqrt{\lambda_k} c_k(\omega) \phi_k(\mathbf{x}) \tag{53}$$

the covariance function of M_K is simply

$$C_K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{k=1}^{K} \lambda_k \phi_k(\mathbf{x}_1)\phi_k(\mathbf{x}_2) \tag{54}$$

[‡]Sufficient conditions for the continuity of Gaussian processes are detailed in Adler [2]. Abrahamsen [1] suggests that any Gaussian process on compact $D \in \mathbb{R}^n$ with a continuous mean and a continuous and "reasonable" covariance function will satisfy these conditions. Covariance functions that provably yield a.s. continuous Gaussian processes include Gaussian, exponential, spherical, Matérn, spline, and polynomial kernels, along with numerous others [2, 84].

which converges uniformly to (49) as $K \to \infty$ [2]. In particular, the total variance or "energy" of M_K is

$$\int_D \mathbb{E}\left[M_K(\mathbf{x},\omega) - \mu(\mathbf{x})\right]^2 d\mathbf{x} = \int_D C_K(\mathbf{x},\mathbf{x}) d\mathbf{x} = \sum_{k=1}^{K} \lambda_k \tag{55}$$

following from the orthonormality of the $\{\phi_k(\mathbf{x})\}$.

2.2.3 Bayesian approach to inverse problems

Bayesian approaches to inverse problems have seen much recent interest [74, 7, 50], with applications ranging from geophysics [43, 69] and climate modeling [49] to heat transfer [97, 98]. We review this approach briefly below; for more extensive introductions, see [50, 89, 74].

Consider a forward problem defined as follows:

$$\mathbf{d} \approx \mathbf{G}(\mathbf{m}) \tag{56}$$

Here $\mathbf{m}$ is a vector of model parameters or inputs and $\mathbf{d}$ is a vector of observable quantities, or data; for simplicity, we let both be real-valued and finite-dimensional. The forward model $\mathbf{G}$ yields predictions of the data as a function of the parameters. In the Bayesian setting, $\mathbf{m}$ and $\mathbf{d}$ are random variables. We use Bayes' rule to define a posterior probability density for $\mathbf{m}$, given an observation of the data $\mathbf{d}$:

$$p(\mathbf{m}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m}) p_m(\mathbf{m}) \tag{57}$$

In the Bayesian paradigm, probability is used to express knowledge about the true values of the parameters. In other words, prior and posterior probabilities represent *degrees of belief* about possible values of $\mathbf{m}$, before and after observing the data $\mathbf{d}$.

Data thus enters the formulation through the likelihood function $p(\mathbf{d}|\mathbf{m})$, which may be viewed as a function of $\mathbf{m}$: $L(\mathbf{m}) \equiv p(\mathbf{d}|\mathbf{m})$. A simple model for the likelihood assumes that independent additive errors account for the deviation between predicted and observed values of $\mathbf{d}$:

$$\mathbf{d} = \mathbf{G}(\mathbf{m}) + \boldsymbol{\eta} \tag{58}$$

where components of $\boldsymbol{\eta}$ are i.i.d. random variables with density p_η. The likelihood then takes the form

$$L(\mathbf{m}) = p_{\boldsymbol{\eta}}\left(\mathbf{d} - \mathbf{G}(\mathbf{m})\right) = \prod_i p_\eta\left(d_i - G_i(\mathbf{m})\right). \tag{59}$$

Additional information on the model parameters may enter the formulation through the prior density, $p_m(\mathbf{m})$. Prior models may embody simple constraints on $\mathbf{m}$, such as a range of feasible values, or may reflect more detailed knowledge about the parameters, such as correlations or smoothness. In the absence of additional information, one may choose a

prior that is *uninformative*. Here we will focus on Gaussian process priors, which for finite-dimensional $\mathbf{m}$ take the form of (45).

If parameters ϕ_m of the prior density $p_m(\mathbf{m}|\phi_m)$ or parameters ϕ_η of the error model $p_\eta(\eta_i|\phi_\eta)$ are not known *a priori*, they may become additional objects for Bayesian inference. In other words, these *hyperparameters* may themselves be endowed with priors and estimated from data [74]:

$$p(\mathbf{m}, \phi_m, \phi_\eta|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m}, \phi_\eta) p_m(\mathbf{m}|\phi_m) p(\phi_\eta) p(\phi_m) \tag{60}$$

The resulting joint posterior over model parameters and hyperparameters may then be interrogated in various ways—e.g., by marginalizing over the hyperparameters to obtain $p(\mathbf{m}|\mathbf{d})$; or first marginalizing over $\mathbf{m}$ and using the maximizer of this density as an estimate of the hyperparameters; or by seeking the joint maximum *a posteriori* estimate or posterior mean of $\mathbf{m}$, ϕ_m, and ϕ_η [68, 74]. In the present study, we will introduce hyperparameters to describe aspects of the prior covariance.

2.2.4 Dimensionality reduction in inverse problems

Now consider an inverse problem in which the unknown quantities comprise a real-valued field $M(\mathbf{x})$. In a computational setting, this field and the forward model must be discretized. If $M(\mathbf{x})$ can be adequately represented on a finite collection of points $\{\mathbf{x}_i\}_{i=1}^n \in D$, then we can write both the prior and posterior densities in terms of $\mathbf{m} = (M(\mathbf{x}_1), \ldots, M(\mathbf{x}_n))$. That is, we can directly apply the Bayesian formulation described in the preceding section and explore the posterior density of $\mathbf{m}$ with Markov chain Monte Carlo (MCMC) [42]. The vector $\mathbf{m}$ will likely be high-dimensional, however. High dimensionality not only renders MCMC exploration of the posterior more challenging and costly, but taxes the polynomial chaos formulation we introduce below to accelerate evaluations of the posterior density [72].

Instead of exploring the value of $M(\mathbf{x})$ on each of n index points, we appeal to the K-L expansion. Let $M(\mathbf{x})$ be endowed with a Gaussian process prior with mean $\mu(\mathbf{x})$ and covariance kernel $C(\mathbf{x}_1, \mathbf{x}_2)$; we denote this as $M \sim \mathcal{GP}(\mu, C)$. Introduce the corresponding K-term K-L representation of $M(\mathbf{x})$ (53), with eigenvalues λ_k and eigenfunctions $\phi_k(\mathbf{x})$ satisfying (48). In general, $M(\mathbf{x}, \omega)$ is approached pointwise in mean square (and therefore in distribution) by $M_K(\mathbf{x}, \omega)$ as $K \to \infty$. For M a.s. continuous (see §2.2.2), realizations $M(\mathbf{x}, \omega)$ can be uniformly approximated as closely as desired by $M_K(\mathbf{x}, \omega)$—implying a corresponding realization $\mathbf{c}(\omega) \equiv (c_1(\omega), \ldots, c_K(\omega))$—with probability one. Updating distributions of M, by conditioning on the data, is thus equivalent to updating the joint distribution of the mode strengths c_k. We emphasize this viewpoint by writing $M_K(\mathbf{x}, \omega) = M_K(\mathbf{x}, \mathbf{c}(\omega)) = M_K(\mathbf{c})$, parameterizing M by the vector of weights $\mathbf{c}$. Components c_k are independent under the Gaussian process prior, with $c_k \sim N(0, 1)$. We thus

truncate the K-L expansion at K terms and write a posterior density for $\mathbf{c}$:

$$\begin{aligned} p(\mathbf{c}|\mathbf{d}) \propto & \quad p(\mathbf{d}|\mathbf{c}) \quad \prod_{k=1}^{K} p(c_k) \\ \propto & \quad p_{\eta}\left(\mathbf{d} - \mathbf{G}\left(M_K(\mathbf{c})\right)\right) \quad \prod_{k=1}^{K} \exp\left(-c_k^2/2\right) \end{aligned} \tag{61}$$

The inverse problem has been transformed to an inference problem on the weights c_k of a finite number of K-L modes. Note that the spatial discretization of $M(\mathbf{x})$ and of the forward model is now independent of the dimension of the posterior distribution. Here we have assumed the prior covariance to be completely known, thus ignoring hyperparameters in the expression for the posterior; we will relax this assumption in §2.3.

Truncating the K-L expansion in this context amounts to using a "modified" prior covariance kernel given by (54). Since the eigenvalues λ_k decay—exponentially fast for a smooth covariance kernel [28], algebraically fast in other cases—a small number of terms may be sufficient to capture almost all of the prior covariance. The linear operator corresponding to the modified covariance kernel now has finite rank; $\phi_k(\mathbf{x})$ that are not eigenfunctions of this operator (48) cannot contribute to the inverse solution. The impact of this truncation will be explored in §2.3.3.

2.2.5 Polynomial chaos acceleration

MCMC exploration of the reduced-dimensionality posterior (61) still requires repeated solutions of the forward model, once for each proposed move of the Markov chain. While dimensionality reduction may reduce the number of such evaluations, depending on the details of the MCMC sampler, it is desirable to avoid repeated forward solutions altogether.

Our previous work [72] introduced methods for accelerating Bayesian inference in this context, by using stochastic spectral methods to propagate prior uncertainty through the forward problem. These methods effectively create a "surrogate" posterior containing polynomial chaos (PC) representations of the forward model outputs. This density may be evaluated orders of magnitude more quickly than the "direct" posterior containing the full forward problem. Here, we will use the Gaussian process prior on M (and thus the prior distribution on $\mathbf{c}$) to define an appropriate stochastic forward problem. The K-L expansion of M ensures that we have chosen a concise representation of the prior uncertainty that is yet suitable for reconstructing inverse solutions. Beginning with polynomial chaos expansions, elements of this approach are described below.

Forward propagation of uncertainty:

Once again let $(\Omega, \mathcal{U}, P)$ be a probability space on which we define a random process $X : \Omega \to \mathbb{R}^D$ with index set $D \subseteq \mathbb{R}^N$. Also, let $\{\xi_i(\omega)\}_{i=1}^{\infty}$ be i.i.d. standard normal random

variables on Ω. Then any square-integrable X has the following representation:

$$\begin{aligned} X(\omega) &= a_0\Gamma_0 + \sum_{i_1=1}^{\infty} a_{i_1}\Gamma_1(\xi_{i_1}) + \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1} a_{i_1 i_2}\Gamma_2(\xi_{i_1},\xi_{i_2}) \\ &+ \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1}\sum_{i_3=1}^{i_2} a_{i_1 i_2 i_3}\Gamma_3(\xi_{i_1},\xi_{i_2},\xi_{i_3}) + \cdots \end{aligned} \tag{62}$$

where Γ_p is the Wiener polynomial chaos of order p [100, 40, 53] and the $a_{i_1 i_2 \ldots}$ may be functions on D. This expansion can be re-written in a more compact form

$$X(\omega) = \sum_{k=0}^{\infty} \hat{a}_k \Psi_k(\xi_1,\xi_2,\ldots) \tag{63}$$

where there is a one-to-one correspondence between the coefficients and functionals in (62) and in (63) [40]. For the standard normal ξ_i chosen above, orthogonality of successive Γ_p requires that the Γ_p be multivariate Hermite polynomials; both these and the corresponding Ψ_k may be generated from univariate Hermite polynomials by taking tensor products.

Of course, in computations it is not useful to retain infinite summations, and one truncates the expansion both in order p and in dimension n—i.e., by choosing a subset $\boldsymbol{\xi} = \{\xi_{\lambda_i}\}_{i=1}^n$ of the infinite set $\{\xi_i\}$, $\lambda_i \in \mathbb{N}$. The total number of terms P in the finite *polynomial chaos expansion*

$$X(\omega) = \sum_{k=0}^{P} x_k \Psi_k(\xi_1,\xi_2,\ldots,\xi_n) \tag{64}$$

is:

$$P+1 = \frac{(n+p)!}{n!p!}. \tag{65}$$

Polynomial chaos (PC) expansions have been generalized to broader classes of orthogonal polynomials in the Askey scheme, each family resulting from a different choice of distribution for the ξ_i [103, 83]. For each of these choices, orthogonality of the polynomials $\Psi_k(\boldsymbol{\xi})$ with respect to the inner product on $L^2(\Omega)$ is maintained:

$$\begin{aligned} \langle \Psi_i \Psi_j \rangle &= \int \Psi_i(\boldsymbol{\xi}(\omega))\Psi_j(\boldsymbol{\xi}(\omega))\,dP(\omega) \\ &= \int \Psi_i(\boldsymbol{\xi})\Psi_j(\boldsymbol{\xi})\rho(\boldsymbol{\xi})d\boldsymbol{\xi} \\ &= \delta_{ij}\langle \Psi_i^2 \rangle \end{aligned} \tag{66}$$

where $\rho(\boldsymbol{\xi})$ denotes the probability density of $\boldsymbol{\xi}$. This property can be used to calculate the truncated PC representation of a random variable $f \in L^2(\Omega)$ by projecting onto the PC basis:

$$\tilde{f}(\omega) = \sum_{k=0}^{P} f_k \Psi_k(\boldsymbol{\xi}), \quad f_k = \frac{\langle f(X)\Psi_k \rangle}{\langle \Psi_k^2 \rangle} \tag{67}$$

This orthogonal projection minimizes the error $\|f - \tilde{f}\|_2$ on the space spanned by $\{\Psi_k\}_{k=0}^P$, where $\|\cdot\|_2$ is the inner-product norm on $L^2(\Omega)$.

Suppose that the behavior of f can be expressed as $O(f,X) = 0$, where O is some deterministic operator and X is a random variable or process with a known PC expansion $X = \sum_{i=0}^P x_i \Psi_i(\boldsymbol{\xi})$. Substituting PC expansions for f and X into this operator and requiring the residual to be orthogonal to Ψ_j for $j = 0 \ldots P$ yields a set of coupled, deterministic equations for the PC coefficients f_k:

$$\left\langle O\left(\sum_k^P f_k \Psi_k, \sum_i^P x_i \Psi_i\right) \Psi_j \right\rangle = 0, \quad j = 0 \ldots P \tag{68}$$

This Galerkin approach is known as "intrusive" spectral projection [40, 61], in contrast to "non-intrusive" approaches in which the inner product $\langle f(X)\Psi_k \rangle$ is evaluated by sampling or quadrature, thus requiring repeated evaluations of $f(X)$ corresponding to different realizations of $\boldsymbol{\xi}$ [39, 80].

In practice, we employ a pseudospectral construction to perform intrusive projections efficiently for higher powers of random variables, e.g. $f(X) = X^j, j \geq 3$, and have developed additional techniques for nonpolynomial functions f. These operations are incorporated into a library for "stochastic arithmetic," detailed in [21].

Stochastic spectral formulation of Bayesian inference:

In [72] we described three accelerated schemes for computing posterior estimates, all based on spectral solutions of a stochastic forward problem: Monte Carlo sampling from the prior distribution, importance sampling, and MCMC. Here we focus on the latter case. The essential idea is to construct a stochastic forward problem whose solution approximates the deterministic forward model over the support of the prior.

Let us begin with (i) a finite-dimensional representation of the unknown quantity that is the object of inference, and (ii) a prior distribution on the parameters of this representation. For instance, if the unknown quantity is a field $M(x)$ endowed with a Gaussian process prior, the finite representation may be a truncated K-L expansion with mode strengths $\mathbf{c}$ and priors $c_i \sim N(0,1)$. The Bayesian formulation in §2.2.4 describes the inverse solution in terms of the posterior density of $\mathbf{c}$, which includes evaluations of the forward model $\mathbf{G}(M_K(\cdot))$. For simplicity, we shall abbreviate $\mathbf{G} \circ M_K$ as $\mathbf{G_c}$; inputs to this model are parameterized by $\mathbf{c}$. Also, let $\mathcal{C}$ denote the support of the prior.

Now define a random vector $\check{\mathbf{c}} = \mathbf{g}(\check{\boldsymbol{\xi}})$, each component of which is given by a PC expansion

$$\check{c}_i = g_i(\check{\boldsymbol{\xi}}) = \sum_{k=0}^P g_{ik} \Psi_k(\check{\boldsymbol{\xi}}) \tag{69}$$

This vector will serve as input to $\mathbf{G_c}$, thus specifying a *stochastic forward problem*. Recall that the distribution of $\check{\boldsymbol{\xi}}$ (e.g., standard normal) and the polynomial form of Ψ (e.g.,

multivariate Hermite) are intrinsic properties of the PC basis. We do not require that $\mathbf{g}$ be chosen such that $\check{\mathbf{c}}$ is distributed according to the prior on $\mathbf{c}$. Rather, we require only (1) that $\Xi_c = \mathbf{g}^{-1}(\mathcal{C})$, the inverse image of the support of the prior, be contained within the range of $\check{\boldsymbol{\xi}}$, and (2) that $\mathbf{g}$ be a diffeomorphism from Ξ_c to $\mathcal{C}$.

Next, using Galerkin projection to solve the stochastic forward problem, we obtain a PC representation for each component of the model output. Here G_i is the i-th component of $\mathbf{G_c}$, and $\tilde{G}_i(\check{\boldsymbol{\xi}})$ is its PC representation:

$$\tilde{G}_i(\check{\boldsymbol{\xi}}) = \sum_{k=0}^{P} d_{ik} \Psi_k(\check{\boldsymbol{\xi}}) \tag{70}$$

The forward prediction $\tilde{\mathbf{G}}$ obtained in this fashion is a function of $\check{\boldsymbol{\xi}}$, and is a polynomial chaos approximation of $\mathbf{G_c}(\mathbf{g}(\check{\boldsymbol{\xi}}))$. Note that both of these quantities are random variables, since $\check{\boldsymbol{\xi}}$ is a random variable. But $\tilde{\mathbf{G}}$ can also be evaluated with a deterministic argument[§]; in this sense, $\tilde{\mathbf{G}}$ is a polynomial approximation of the deterministic forward model $\mathbf{G_c} \circ \mathbf{g}$.

We would like to use this approximation to replace $\mathbf{G_c}$ in the likelihood function $L(\mathbf{c}) \equiv p_{\boldsymbol{\eta}}(\mathbf{d} - \mathbf{G_c}(\mathbf{c}))$:

$$L(\mathbf{g}(\boldsymbol{\xi})) \approx \tilde{L}(\boldsymbol{\xi}) \equiv p_{\boldsymbol{\eta}}\left(\mathbf{d} - \tilde{\mathbf{G}}(\boldsymbol{\xi})\right) \tag{71}$$

Implicit in this substitution is the change of variables $\mathbf{c} = \mathbf{g}(\boldsymbol{\xi})$, i.e., from the input parameterization of $\mathbf{G_c}$ to the input parameterization of $\tilde{\mathbf{G}}$, enabled because $\mathbf{g}$ satisfies conditions (1) and (2) above.

We write the change of variables in terms of the posterior expectation of an arbitrary function f:

$$\mathbb{E}_{\pi_c} f = \mathbb{E}_{\pi_\xi}(f \circ \mathbf{g}) \tag{72}$$

where $\pi_c \equiv p(\mathbf{c}|\mathbf{d})$ is the posterior density on $\mathbf{c}$-space, and π_ξ is the corresponding posterior density on $\boldsymbol{\xi}$-space:

$$\pi_\xi(\boldsymbol{\xi}) \propto L(\mathbf{g}(\boldsymbol{\xi}))\, p_c(\mathbf{g}(\boldsymbol{\xi}))\, |\det D\mathbf{g}(\boldsymbol{\xi})| \tag{73}$$

Here, $D\mathbf{g}$ denotes the Jacobian of $\mathbf{g}$ and p_c is the prior density of $\mathbf{c}$. Eliminating the forward model from the likelihood function via (71) finally yields the "surrogate" posterior density $\tilde{\pi}_\xi$:

$$\begin{aligned} \tilde{\pi}_\xi(\boldsymbol{\xi}) \propto & \quad \tilde{L}(\boldsymbol{\xi}) \quad p_c(\mathbf{g}(\boldsymbol{\xi}))\, |\det D\mathbf{g}(\boldsymbol{\xi})| \\ \propto & \quad p_{\boldsymbol{\eta}}\left(\mathbf{d} - \tilde{\mathbf{G}}(\boldsymbol{\xi})\right) \quad p_c(\mathbf{g}(\boldsymbol{\xi}))\, |\det D\mathbf{g}(\boldsymbol{\xi})| \end{aligned} \tag{74}$$

This distribution may be explored with any suitable sampling strategy, in particular MCMC. Evaluating the density for purposes of sampling may have negligible cost; nearly all the

[§]In this exposition we have used $\check{}$ to identify the random variables $\check{\mathbf{c}}$ and $\check{\boldsymbol{\xi}}$ in order to avoid confusion with deterministic arguments to probability density functions, e.g., $\mathbf{c}$ and $\boldsymbol{\xi}$ below. Elsewhere, we will revert to the usual notational convention and let context make clear the distinction between the two.

computational time may be spent in intrusive spectral projection, obtaining the PC expansions in (70). Depending on model nonlinearities, the necessary size of the PC basis, and the number of posterior samples required, this computational effort may be orders of magnitude less costly than exploring the posterior via direct sampling. Accuracy of the surrogate posterior depends the order and family of the PC basis, as well as on the choice of transformation **g**—for instance, whether the distribution of $\check{\mathbf{c}}$ assigns sufficient probability to regions of $\mathcal{C}$ favored by the posterior on **c**. A detailed discussion of these issues can be found in [72].

2.2.6 Gaussian processes, K-L expansions, RKHS, and regularization

There are important connections between Gaussian process priors and regularization penalties in the corresponding reproducing kernel Hilbert space (RKHS) norm. These connections can be understood in terms of the spectral expansion of the covariance kernel, and it is useful to review these relationships in the present context.

The definition and properties of reproducing kernel Hilbert spaces are briefly recalled in the Appendix. It is natural to think of a positive definite reproducing kernel K as a *covariance* kernel, and indeed any Gaussian process can be associated with a RKHS. Let $X(t), t \in \mathcal{T}$ be a centered Gaussian process with covariance kernel K. If K has more than a finite number of non-zero eigenvalues, realizations of $X(t)$ are almost surely *not* in the corresponding RKHS $H(K)$ [94].[¶] However, there exists an isometry between the two. In particular, let $\mathcal{H}$ be the Hilbert space spanned by $X(t)$: $\mathcal{H} = \text{span}\{X_t, t \in \mathcal{T}\}$ with inner product $\langle Z_i, Z_j \rangle = \mathbb{E}[Z_i Z_j]$ for $Z_i, Z_j \in \mathcal{H}$. It can be shown that $\mathcal{H}$ is isometrically isomorphic to $H(K)$ [94, 2].

Bayesian estimates with Gaussian process priors may lie in the corresponding RKHS, however [94, 56]. Consider the case of an inverse problem with Gaussian process prior $\mathcal{GP}(0,K)$ on the unknown function f. Details of the likelihood function and forward model are unimportant here. For simplicity, we assume that the prior covariance is completely known. Let f_i denote the projection of f onto the i-th eigenfunction ψ_i:

$$f_i = \int_D f(s)\psi_i(s)ds \tag{75}$$

where ψ_i and λ_i are, as usual, eigenfunctions and eigenvalues of the linear operator corresponding to K. According to (51), the prior distribution on each f_i is $N(0,\lambda_i)$. Then the posterior probability of the function f, $\pi(f)$, has the form:

$$\begin{aligned} J = -\log\pi(f) &= \text{log-likelihood} + \sum_i^{\infty} \frac{f_i^2}{2\lambda_i} + \text{const} \\ &= \ldots + \frac{1}{2}\|f\|^2_{H(K)} + \ldots \end{aligned} \tag{76}$$

[¶]As an example of a GP whose realizations are not in the RKHS, consider standard Brownian motion. Sample paths are nowhere differentiable with probability one, but members of the RKHS are differentiable almost everywhere, with square-integrable derivatives.

J is thus a RKHS-norm penalized cost functional, in which the Gaussian process prior provides the regularization penalty. Minimizing J to obtain the MAP estimate of f is equivalent to finding a Tikhonov-regularized solution to the inverse problem, with $\operatorname{argmin} J = f_{\mathrm{MAP}} \in H(K)$. Changing the prior covariance kernel amounts to changing the RKHS norm and thus the nature of the regularization penalty.

Moreover, there is an equivalence between the RKHS regularization functional $\|f\|_{H(K)}$ and a standard $L^2(D)$-norm penalty $\|Lf\|_2$ containing the differential operator L: the reproducing kernel K is the Green's function of the operator L^*L, where L^* denotes the adjoint of L [89, 84]. Thus a Gaussian kernel leads to a penalty on derivatives of all orders; the exponential covariance kernel penalizes the square of the function value $f(s)$ and its derivative $\dot{f}(s)$ (i.e., a Sobolev H^1 norm); and the covariance kernel of Brownian motion, $K(s,t) = \min(s,t)$, leads to a penalty on the MAP estimate's squared derivatives.

Finally, we note that the present scheme of K-L based inversion (§2.2.4) recalls the "weight space" view of Gaussian process regression, in that we find weights on the set of feature vectors ϕ_k [78] implied by the Gaussian process prior. Truncation to a finite subset of these features might seem at odds with the usual use of Gaussian processes in nonparametric statistical models. However, some features are more important than others. More precisely, results in §2.3.3 will show that the weights of feature vectors beyond some limit $k > k^\star$ are unchanged by conditioning on the data—suggesting that it is reasonable, in the present inverse problem context, to limit conditioning on the data to a finite subset of weights. The remaining features are still "active" in the sense that they too may be seen as contributing to the posterior, but their contributions do not change from the prior.

2.3 Numerical implementations and results

We explore the accuracy and efficiency of our dimensionality reduction approach by estimating inhomogeneous diffusivity fields in a transient diffusion problem. We pursue these inverse problems both with and without the added step of solving the stochastic forward problem to construct a surrogate posterior (§2.2.5). In particular, we consider a diffusion equation on the unit interval $D = [0,1]$ with adiabatic boundaries:

$$\begin{aligned}
\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\nu(x)\frac{\partial u}{\partial x}\right) &+ \sum_{i=1}^{N} \frac{s_i}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{|l_i - x|^2}{2\sigma_i^2}\right)[1 - H(t - T_i)] \qquad (77)\\
\left.\frac{\partial u}{\partial x}\right|_{x=0} = \left.\frac{\partial u}{\partial x}\right|_{x=1} &= 0\\
u(x, t=0) &= 0
\end{aligned}$$

The source term in (77) involves N localized sources, each active on the interval $t \in [0, T_i]$ and centered at $l_i \in D$ with strength s_i and width σ_i, $i = 1 \ldots N$. Source parameters are

prescribed, and we infer $\nu(x)$ from noisy measurements of the u-field at a finite set of locations and times. This problem can be considered a prototype for the inverse estimation of an inhomogeneous conductivity field or any analogous material or transport property, such as the permeability field in a porous medium [62, 76].

2.3.1 Inverse problem setup

The transient diffusion equation above may be cast as a forward model that predicts the value of the field at specific locations and times. Taking the diffusivity to be uniformly bounded away from zero, $\nu(x) > \nu_0 > 0$ with $\nu_0 \equiv 0.1$, we define the log-diffusivity $M(x) \equiv \log[\nu(x) - \nu_0]$; this function is the input to the forward model. We evaluate the field at mn points $\{u(x_i, t_j) : 1 \leq i \leq m, 1 \leq j \leq n\}$. The "sensor locations" $\{x_1 \ldots x_m\}$ are uniformly spaced on D, including the endpoints, and the measurement times $\{t_1 \ldots t_n\}$ are uniformly spaced on an arbitrary time interval. Below, unless otherwise specified, we will use $m = 13$ sensors, $n = 9$ measurement times, and $N = 3$ sources. The source locations are staggered with respect to the sensors; i.e., they are placed at $l_i \in \{0.25, 0.50, 0.75\}$. We prescribe identical strengths $s_i = 100$, shutoff times $T_i = 0.01$, and widths $\sigma_i^2 = 10^{-2}$ for all three sources. Measurements take place over the time interval $t \in [0.01, 0.03]$.

The u-field is described on a uniform grid with spacing $h = 1/48$. Second-order centered differences are used to discretize the diffusion terms. Time integration is via an explicit, second-order-accurate, Runge-Kutta-Chebyshev (RKC) scheme [88] with $\Delta t = 10^{-4}$. For any input $\nu(x)$, the number of substeps in the RKC scheme is automatically determined by stability constraints upon setting ε, the damping parameter that controls the extent of the stability region, to $2/13$ [93]. Numerical resolution studies were conducted to validate the present choices of h and Δt.

Note that the forward model is nonlinear in $\nu(x)$. Consider the simple case of a uniform diffusivity, $\nu(x) = \bar{\nu}$. Figure 18 shows the resulting forward maps, from $\log[\bar{\nu} - \nu_0]$ to u, at a single measurement location and two successive times. The measurement location, $x^* = 1/6$, is adjacent to a source at $x = 1/4$. For very small diffusivities, the scalar introduced by the source does not diffuse towards the sensor in appreciable quantity, and hence u is small; in this regime, the magnitude of the scalar field rises with t and with $\bar{\nu}$. At very large diffusivities, the scalar introduced by all N sources rapidly diffuses towards all the sensors and the u-field quickly becomes uniform, approaching $u = \sum_i s_i T_i = 3$ as $\bar{\nu} t \to \infty$. For intermediate diffusivities, the measured value of u may decrease with rising $\bar{\nu}$: the scalar field, locally peaked at the nearby source, flattens as the diffusivity increases, until the influence of the remaining sources is felt at sufficiently high $\bar{\nu}$, raising the local value of u once again. The behavior of analogous forward maps in the case of nonuniform $\nu(x)$ is expected to be even more complicated.

The *inverse problem* thus consists of inferring $M(x) \equiv \log[\nu(x) - \nu_0]$ from *noisy* measurements of $u(x_i, t_j)$. In the Bayesian setting, we provide statistical information about the measurement process and about our prior knowledge of $M(x)$. We let independent zero-

mean Gaussian random variables $\eta_i \sim N(0, \varsigma^2)$ express the difference between "real-world" measurements and model predictions, as specified in (3). In the examples below, we choose $\varsigma = 0.1$. We endow $M(x)$ with a zero-mean Gaussian process prior $M \sim \mathcal{GP}(0, C)$, where C is a stationary Gaussian covariance kernel:

$$C_\theta(x_1, x_2) = \tilde{C}_\theta(|x_1 - x_2|) = \theta \exp\left(\frac{|x_1 - x_2|^2}{2L^2}\right). \tag{78}$$

For simplicity, we assume that the correlation length L is known; in practical applications, an estimate of L will often be available [62, 76]. We do not, on the other hand, presume to know the scale θ of the prior covariance. Adopting a fully Bayesian approach, we let θ be a hyperparameter endowed with a conjugate inverse gamma hyperprior, $\theta \sim IG(\alpha, \beta)$ [32, 31]:

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-\alpha-1} \exp\left(-\frac{\beta}{\theta}\right). \tag{79}$$

In the examples below, we fix the shape parameter $\alpha = 1$ and the scale parameter $\beta = 1$. This yields a proper but long-tailed prior for θ, with undefined mean and variance. The magnitude of the prior covariance θ joins the remaining parameters describing $M(x)$ in the joint posterior density; we can then marginalize over θ to obtain a posterior describing $M(x)$ alone (see §1.2.1). Note that when considering MAP estimates of M conditioned on θ, the ratio ς^2/θ is akin to the regularization parameter appearing in deterministic inversion; thus we are effectively estimating the strength of the regularization when conditioning on the data [50].

We will solve the inverse problem for four different "target profiles" $M(x)$: a simple linear profile, a sinusoidal profile, a profile randomly drawn from the Gaussian process prior with $L = 0.3, \theta = 1.0$, and a well-shaped profile. Plots of these target $M(x)$ are provided in the next section. For each profile, a noisy data vector $\mathbf{d}$ is generated by solving the deterministic forward problem with the target log-diffusivity, then perturbing the resulting value of u at each sensor location/time with independent samples of Gaussian noise $\eta_i \sim N(0, \varsigma^2)$. To avoid an "inverse crime" [50], we generate the mn values of $u(x_i, t_j)$ by solving the forward problem at a much higher resolution than that used in the inversion, i.e., with $h = 1/408$ and a correspondingly finer Δt.

2.3.2 Grid-based inversion

We begin with a straightforward full-dimensional Bayesian approach to the inverse problem, as described at the start of §2.2.4. Let $M(x)$ be represented on a finite collection of points $\{x_i\}_{i=1}^n \in D$; an obvious choice with adequate resolution is the collection of grid points used to discretize the forward model, uniformly spaced on the unit interval with $x_{i+1} - x_i = h$. Then we can write both the prior and posterior densities in terms of

$\mathbf{m} = (M(x_1), \ldots, M(x_n))$:

$$
\begin{aligned}
p(\mathbf{m}, \theta | \mathbf{d}) &\propto p(\mathbf{d}|\mathbf{m}) \times p(\mathbf{m}|\theta) \times p(\theta) \\
&\propto p_{\eta}(\mathbf{d} - \mathbf{G}(\mathbf{m})) \times \theta^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\mathbf{m}^T \Sigma_\theta^{-1} \mathbf{m}\right) \times p(\theta) \\
&\propto \exp\left(-\frac{[\mathbf{d} - \mathbf{G}(\mathbf{m})]^T [\mathbf{d} - \mathbf{G}(\mathbf{m})]}{2\varsigma^2}\right) \\
&\quad \times \theta^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\mathbf{m}^T \Sigma_\theta^{-1} \mathbf{m}\right) \times \theta^{-\alpha-1} \exp\left(-\frac{\beta}{\theta}\right)
\end{aligned} \tag{80}
$$

where $(\Sigma_\theta)_{ij} \equiv C_\theta(x_i, x_j)$.

Directly applying a Metropolis-Hastings algorithm to this posterior is not likely to be successful, however. Simple proposal distributions for $\mathbf{m}$, such as normal distributions centered at the current position of the chain, generate candidate points with very low acceptance probabilities—even when applied component-at-a-time [76]. These proposals do not account for correlations among neighboring components of $\mathbf{m}$. We surmount this issue with a change of variables, using the Cholesky factorization of the prior covariance matrix, using $\theta = 1$: $\Sigma_{(\theta=1)} = \mathbf{L}\mathbf{L}^T$. If $\mathbf{z}$ is vector of n i.i.d. standard normal random variables, then $\forall\theta$,

$$
\mathbf{m} = \sqrt{\theta}\mathbf{L}\mathbf{z} \tag{81}
$$

will have a zero-mean multivariate normal distribution with covariance Σ_θ. (Multiplication by $\mathbf{L}$ is analogous to, in the continuous case, generating samples of a Gaussian process by convolution with white noise [63].) Thus the $N(\mathbf{0}, \Sigma_\theta)$ prior distribution on $\mathbf{m}$ reduces to a Gaussian prior on $\mathbf{z}$ with *diagonal* covariance, $N(\mathbf{0}, \mathbf{I})$. Equivalently, we can write $\mathbf{m} = \mathbf{L}\mathbf{z}$ and let the scale parameter θ control the prior variance of $\mathbf{z} \sim N(\mathbf{0}, \theta\mathbf{I})$, thus reparameterizing the posterior density as follows:

$$
\begin{aligned}
p(\mathbf{z}, \theta | \mathbf{d}) &\propto \exp\left(-\frac{[\mathbf{d} - \mathbf{G}(\mathbf{L}\mathbf{z})]^T [\mathbf{d} - \mathbf{G}(\mathbf{L}\mathbf{z})]}{2\varsigma^2}\right) \\
&\quad \times \theta^{-\frac{n}{2}} \exp\left(-\frac{\mathbf{z}^T\mathbf{z}}{2\theta}\right) \times \theta^{-\alpha-1} \exp\left(-\frac{\beta}{\theta}\right)
\end{aligned} \tag{82}
$$

We use a Metropolis-Hastings algorithm to simulate samples from this distribution [42]. For the scale parameter θ, we apply Gibbs updates: the full conditional $p(\theta|\mathbf{z}, \mathbf{d})$ is proportional to $IG(\alpha + n/2, \beta + (\sum_{i=1}^n z_i^2)/2)$ [32], so we sample directly from this distribution with acceptance probability 1. For the remaining parameters $\mathbf{z}$, we use single-component random-walk Metropolis updates: each proposal distribution $q(\cdot|\cdot)$ is a univariate normal centered on the current position of the chain. It may be possible to increase the efficiency

of this sampler by using single-component updating for the first few components of $\mathbf{z}$ and block updating for the higher-index, less important components [62], but we do not pursue such fine-tuning here. MCMC yields a series of samples $\{(\mathbf{z}^{(s)}, \theta^{(s)})\}$, which are easily transformed to $\{(\mathbf{m}^{(s)}, \theta^{(s)})\}$. From these samples, we can estimate posterior expectations (e.g., means, variances, higher moments), extract marginal densities $p(M(x_i)|\mathbf{d})$, and estimate quantiles of the marginal distributions.

Figures 19–22 show the results of grid-based inversion with each of the four target profiles. In each case, we have assumed a value for the prior correlation length appropriate to the target: $L = 1.0$ for the linear profile, $L = 0.2$ for the sinusoidal profile, $L = 0.1$ for the well-shaped profile, and $L = 0.3$ for the profile corresponding to a random draw from the GP prior. Figures 19(a)–22(a) show the posterior mean and standard deviation, along with five samples from each posterior distribution. In part (b) of these figures, we extract one-dimensional marginal distributions of $M(x)$ at each grid point x_i, then plot the median and 5% and 95% quantiles of the distributions. Even though statistical dependence among different spatial locations has been marginalized away, these profiles reflect an envelope of uncertainty in the inverse solution. In all four cases, uncertainty in the log-diffusivity is greatest near the boundaries, with some additional rise near the center of the domain. Unsurprisingly, shorter prior correlation lengths result in larger uncertainties overall. All of the results presented here are based on 6×10^5 MCMC samples; we find negligible change in the estimated moments and quantiles with further iterations.

In three of the four cases (the linear, sinusoidal, and random-draw targets), the posterior mean and median are good estimates of the true profile; the true log-diffusivity is generally contained within the credibility intervals bounded by the marginal quantiles. Mismatch with the true profile may be ascribed to limited sensor resolution (in space and in time), noise in the data, and the interaction of these conditions with the physics of the forward model. In the fourth case (Figure 22, i.e., the well-shaped target), the inferred profile is smoother than the true profile. While the location of the well ($0.4 < x < 0.7$) may be surmised from the posterior, the true profile does not fall entirely within the marginal quantiles. Here, information encoded in the prior is actually *inconsistent* with the well-shaped log-diffusivity. Even with a small correlation length, a GP prior with a Gaussian covariance encodes significant smoothness, assigning very small probability to sharp variations. The posterior distribution reflects this belief in the character of the log-diffusivity profile. To obtain more accurate reconstructions and credibility intervals in this case, the prior distribution must be chosen more carefully. Tarantola [89] suggests that if discontinuities are expected, their geometric properties should enter explicitly into the parameterization of the inverse problem. One may also construct *structural priors*, typically Gaussian but not isotropic or stationary, that encode the location and geometry of non-smooth features [50, 52, 12].

Since the full posterior is a distribution on $n+1$-dimensional space, it contains much more information than can be shown in Figures 19–22. Consider, for instance, the change in the covariance of $M(x)$ from the prior to the posterior. Computing $\mathrm{Var}[\mathbf{m}] = \mathrm{Cov}\left[M(x_i), M(x_j)\right]$ requires marginalizing over the hyperparameter θ. The prior marginal, with density $p(\mathbf{m}) =$

$\int_0^\infty p(\mathbf{m}|\theta)p(\theta)\mathrm{d}\theta$, is a multivariate t-distribution; its covariance can be obtained analytically as $\beta\Sigma/(\alpha-1)$ for $\alpha > 1$. The posterior covariance is estimated numerically from the MCMC samples. Figure 23(a) shows the prior covariance with $L = 0.3$; in this case only, we put $\alpha = 3$ and $\beta = 2$ so the magnitude of the marginal prior covariance is well-defined. Figure 23(b) shows the corresponding posterior covariance, again with $\alpha = 3$ and $\beta = 2$, conditioned on the noisy data vector used to infer the random-draw target in Figure 20. The posterior covariance clearly reflects a nonstationary process, and its overall scale is more than an order of magnitude smaller than the prior covariance. The diagonal of the posterior covariance is analogous to the square of the standard deviation in Figure 20(a). Decay of the covariance away from the diagonal reflects the character of spatial variation (around the mean, up to second order) in the log-diffusivity profiles comprising the inverse solution.

It is important to note that, because the forward operator G is nonlinear, the posterior distributions shown here (whether 1-D marginals or full joint distributions) are not in general Gaussian or even symmetric.

2.3.3 Reduced-dimensionality inversion

Now we pursue a reduced-dimensionality solution of the inverse problem by exploring the posterior distribution of the weights c_k of a finite number of K-L modes, as described in §2.2.4. First, we must determine the eigenfunctions and eigenvalues appearing in the K-L expansions. For the Gaussian covariance kernel (78) on $D = [0,1]$, there is no analytical solution for the spectral decomposition of the corresponding integral operator. Instead, we solve the integral equation (48) numerically, using the Nystrom method [77] with a Gauss-Legendre quadrature rule and a LAPACK solver for the first K eigenvalues and eigenvectors of the resulting real symmetric matrix.

The hyperparameter θ is treated as in the previous section. The scale of the covariance kernel does not affect the eigenfunctions $\phi_k(x)$; it simply multiplies the eigenvalues λ_k. Thus, we can compute the K-L expansion of $M(x) \sim \mathcal{GP}(0, C_\theta)$ while fixing $\theta = 1$, and let the hyperparameter control the prior variance of the random variables c_k, $\mathbf{c} \sim N(\mathbf{0}, \theta\mathbf{I})$. The posterior density in (61) is re-written as follows:

$$\begin{aligned} p(\mathbf{c},\theta|\mathbf{d}) \propto & \quad p(\mathbf{d}|\mathbf{c})p(\mathbf{c}|\theta)p(\theta) \\ \propto & \quad \exp\left(-\frac{(\mathbf{d}-\mathbf{G}(M_K(\mathbf{c})))^T(\mathbf{d}-\mathbf{G}(M_K(\mathbf{c})))}{2\varsigma^2}\right) \\ & \times\theta^{-\frac{K}{2}}\exp\left(-\frac{\mathbf{c}^T\mathbf{c}}{2\theta}\right)\times\theta^{-\alpha-1}\exp\left(-\frac{\beta}{\theta}\right) \end{aligned} \tag{83}$$

where the forward model $\mathbf{G}$ now maps functions $M: D \to \mathbb{R}$, representing the log-diffusivity, to $\mathbb{R}^{mn}$. $M_K(\mathbf{c})$ denotes the K-term K-L expansion (53) evaluated at $\mathbf{c}$:

$$M_K(x;\mathbf{c}) = \sum_{k=1}^{K}\sqrt{\lambda_k}c_k\phi_k(x) \tag{84}$$

Figures 24–30 show the results of K-L-based inversion with each of the four target profiles. As in the previous section, part (a) of each figure shows the posterior mean and standard deviation along with randomly-chosen posterior realizations, while part (b) overlays the 1-D posterior median and quantiles with the true log-diffusivity. Results obtained with a sufficiently large number of K-L modes K become indistinguishable from the grid-based inverse solutions. As expected, shorter prior correlation lengths require a larger number of K-L modes for accurate inverse solutions. Consider the remainder of the total *prior* variance integrated over the domain D (55), i.e., $1-\sum_{k=1}^{K}\lambda_k$, shown in Figure 31. This quantity decays exponentially fast with increasing K, reflecting the decay of the eigenvalues of the Gaussian covariance kernel (78), but eigenvalues corresponding to large-L kernels decay more quickly than those corresponding to small L. Since the distributions of c_k are altered by conditioning on $\mathbf{d}$, the relative importance of the K-L modes changes in the posterior, but still decays at larger index. Figure 32 compares MCMC estimates of the posterior moments of $M(x)$, obtained via grid-based inversion, to MCMC-estimated posterior moments of $M_K(x)$ obtained via K-L-based inversion with varying numbers of modes. In particular, we compute the L^2 distance between estimates of the posterior mean, $\|\widehat{\mu}(M_K(x))-\widehat{\mu}(M(x))\|_2=\left(\int_D|\widehat{\mu}(M_K(x))-\widehat{\mu}(M(x))|^2dx\right)^{1/2}$, and the L^2 distance between estimates of the posterior standard deviation, $\|\widehat{\sigma}(M_K(x))-\widehat{\sigma}(M(x))\|_2$. Differences between these posterior estimates at first fall rapidly with increasing K, but then plateau. The plateau region appears at smaller K for the large-L cases (e.g., the line profile) and at larger K for the small-L cases (e.g., the well profile), and reflects the fact that differences between moments of the grid and K-L-based inverse solutions eventually become comparable to the variability of the MCMC estimates themselves. Indeed, each realization of a Markov chain yields slightly different estimates of the posterior mean and standard deviation, and differences among these realizations account for continued "jitter" in the plateau regions. To illustrate the spread in these estimates, we have plotted results from additional realizations of the MCMC chain at $K=6$, 8, and 10, for the random-draw target. (Each realization corresponds to a distinct choice of random seed.) Differences in the magnitude of the plateau region associated with each target profile reflect the fact that variance of an MCMC estimate is dependent on the variance of the actual quantity being estimated [81]; the posterior associated with the well-shaped target, for instance, shows much larger variances than the posterior associated with the linear target.

Further insight into the contribution of each K-L mode to the inverse solution is obtained by examining boxplots of the posterior marginals of the mode strengths. In particular, we consider marginal densities of $\sqrt{\lambda_k}c_k$, the *scaled* contribution of each K-L mode. The K-L eigenfunctions multiplied by these factors each have an L^2 norm of unity, and thus the relative importance of each eigenfunction—e.g., each mode's contribution to the mean and spread of the posterior—is captured by the boxplots in Figure 33. Results are reported for the random-draw and well-shaped targets. The horizontal line at the center of each box marks the median of the posterior marginal $p(\sqrt{\lambda_k}c_k|\mathbf{d})$; the extent of each box marks the 25% and 75% quantiles of the posterior marginal; and the vertical lines span the entire range of the MCMC samples. The importance of each mode does not decrease strictly with k. For instance, K-L mode $\phi_4^{L=0.3}(x)$ contributes more to the posterior of the random-draw target than $\phi_2^{L=0.3}(x)$ and $\phi_3^{L=0.3}(x)$; with the well-shaped target, mode $\phi_9^{L=0.1}(x)$ contributes

more to the posterior than $\phi_7^{L=0.1}(x)$. At sufficiently large index, however, the [exponential] decrease of the λ_k takes over: variances of the mode strengths decrease and the medians tend towards zero.

Spatial correlations are also well-reproduced by the reduced-dimensionality inverse solution. Consider contours of the posterior covariance $\mathrm{Cov}\,[M(x_1), M(x_2)]$, shown in Figure 34. Solid lines are obtained via the grid-based inversion described in §2.3.3, while dashed lines represent the posterior covariance computed with K K-L modes. Very close agreement is observed with increasing K. This result may be somewhat surprising, as $\phi_k(x)$ are not eigenfunctions of the posterior covariance and thus not an optimal basis for posterior in the sense of a K-L representation. Nonetheless, a modest number of these modes is able to capture the posterior covariance.

Eigenfunctions aside, the ability to reproduce the posterior covariance also depends on the emergence of correlations in the joint posterior of the K-L mode strengths c_k. Figure 35 shows all of the one- and two-dimensional posterior marginals of $(\theta, \mathbf{c})$ conditioned on the random-draw target. Significant correlations are apparent among the lower-indexed modes, and between these modes and the hyperparameter θ. Higher-indexed modes, on the other hand, appear uncorrelated—and based on the shape of their 2-D marginals, mutually independent. We examine the 1-D marginals of these modes more closely in Figure 36. The solid lines are conditional densities $p(c_k|\mathbf{d}, \theta)$ extracted from the posterior via kernel density estimation and plotted for different values of θ. The dashed lines are the corresponding conditional prior densities $p(c_k|\theta)$; recall that these are simply $c_k \sim N(0, \theta)$. The posterior densities of c_6, shown in Figure 36(a), are shifted and somewhat narrower than their priors. (Much more dramatic changes from the prior are observed for c_1 through c_5.) Conditional posteriors of c_8 in Figure 36(b), on the other hand, match the prior conditionals quite closely. Similarly close matching is observed for modes c_9, c_{10}, and so on. This pattern leads us to conjecture that, at sufficiently large m, the posterior distribution of K-L modes $c_{k\geq m}$ approaches the prior:

$$p(c_m, c_{m+1}, \ldots | \mathbf{d}, \theta) \rightarrow \prod_{k \geq m} p(c_k|\theta) \tag{85}$$

This conjecture is consistent with the decay of λ_k at large k; since higher-index modes have smaller λ_k, these modes should have less impact on the predicted $u(x,t)$ and on the data $\mathbf{d}$. Absent conditioning on the data, these modes revert to being independent and conditionally Gaussian. §2.3.5 revisits this issue in the case of coarser data.

A practical benefit of using K-L modes to compute the inverse solution is more efficient MCMC sampling. Figure 37(a) plots $\gamma(s)/\gamma(0)$, the empirical autocorrelation at lag s, for several components of the Markov chain used to explore $p(\mathbf{z}, \theta|\mathbf{d})$ (i.e., grid-based inversion) and several components of the chain exploring $p(\mathbf{c}, \theta|\mathbf{d})$ (K-L-based inversion). In both cases, the noisy data vector $\mathbf{d}$ is obtained from the random-draw target. With the grid-based posterior, lower-index z_i multiply columns nearer the left side of the Cholesky factor $\mathbf{L}$, which contain a larger number of non-zero entries (specifically, $n-i+1$). These modes mix less efficiently than their larger-k counterparts, even though we have individually tuned the proposal width of each single-component Metropolis update. Mode strengths c_k and the

hyperparameter θ of the K-L-based posterior show much more rapid decay of their autocorrelations, reflecting improved mixing. In Figure 37(b), we transform the c_k and z_i into realizations of $M(x)$ and compare the chain autocorrelations at particular spatial locations. With the grid-based posterior, mixing improves toward the right side of the physical domain D, as the value of the solution at larger x is influenced by a larger number of modes z_i—and in particular, better-mixing modes. The autocorrelations of K-L-based $M(x_i)$ still decay more rapidly, however.

2.3.4 Reduced-dimensionality inversion with polynomial chaos acceleration

We now construct a stochastic forward problem whose solution captures the output of the deterministic forward model over the support of the prior distribution, and use this solution to formulate a surrogate posterior distribution. The resulting scheme is intended to achieve an accelerated, reduced-dimensionality Bayesian solution of the inverse problem, as described in §2.2.5.

We begin with the K-L representation of the log-diffusivity field and the hierarchical priors on c_k derived in the previous section. Following the notation in §2.2.5, introduce the scaling transformation

$$\mathbf{c} = \mathbf{g}(\boldsymbol{\xi}) = \varpi\boldsymbol{\xi}. \tag{86}$$

This transformation is equivalently a first-order Gauss-Hermite PC expansion, where $\boldsymbol{\xi}$ is a K-vector of i.i.d. standard normal random variables; the expansion defines the uncertainty that we propagate through the forward model. Since the prior distribution of each c_k has support over the entire real line, as does the $N(0,1)$ distribution of each ξ_k, we have considerable freedom in choosing $\varpi > 0$; any choice of ϖ will map the range of $\boldsymbol{\xi}$ onto the range of $\mathbf{c}$ [72]. The choice is particularly open since θ, the prior variance of c_k, is itself unknown. Here we will fix $\varpi^2 = 0.5$, which is the mode of the hyperprior $p(\theta)$.

Together, (77), (84), and (86) define a stochastic forward problem. The input is the Gaussian random field $M(x) \equiv \log[\nu(x,\boldsymbol{\xi}(\omega)) - \nu_0]$ represented with a truncated K-L expansion, and the outputs are random variables $u(x_i,t_j,\omega)$ giving the value of the scalar field at each measurement location and time. We write these random variables in terms of their PC expansions $u(x_i,t_j,\omega) = \sum_l u_l^{ij}\Psi_l(\boldsymbol{\xi}(\omega))$, and collect them in a vector $\tilde{\mathbf{G}}(\boldsymbol{\xi}) \equiv (u(x_1,t_1;\boldsymbol{\xi}),\ldots,u(x_m,t_n;\boldsymbol{\xi}))$. Solving the stochastic forward problem—i.e., using Galerkin projection to compute the coefficients u_l^{ij}—requires transforming the input log-diffusivity into an actual diffusivity $\nu(x,\omega)$. The Gauss-Hermite PC representation of this log-normal process may be evaluated analytically [40, 67]. Recall that the multivariate polynomial functionals comprising the PC basis are given by the tensor product of one-dimensional polynomials $\psi_i(\xi)$, here Hermite polynomials of order i. Each multivariate polynomial Ψ_l is associated with a multi-index $\boldsymbol{\alpha}^l = (\alpha_1^l,\ldots,\alpha_K^l) \in \mathbb{N}^K$, where $\sum_{k=1}^K \alpha_k \leq p$:

$$\Psi_l(\boldsymbol{\xi}) = \prod_{k=1}^{K} \psi_{\alpha_k^l}(\xi_k) \tag{87}$$

The diffusivity $\nu(x,\omega) = \nu_0 + \exp(M_K(x,\boldsymbol{\xi}(\omega))$ then has PC coefficients

$$\begin{aligned}
\nu_{(l=0)}(x) &= \nu_0 + e^{\sigma^2(x)/2} \\
\nu_{(l\geq 1)}(x) &= e^{\sigma^2(x)/2} \prod_{k=1}^{K} \frac{\left[\varpi\sqrt{\lambda_k}\phi_k(x)\right]^{\alpha_k^l}}{\sqrt{\alpha_k^l!}}
\end{aligned} \tag{88}$$

with[‖]

$$\sigma^2(x) = \varpi^2 \sum_{k=1}^{K} \lambda_k \phi_k^2(x) \tag{89}$$

This PC expansion (88) is introduced into the transient diffusion equation (77). Using a pseudospectral stochastic Galerkin scheme coupled with the same finite-difference spatial discretization and RKC time integrator as in the deterministic problem, we obtain PC expansions for the outputs of interest $\tilde{\mathbf{G}}(\boldsymbol{\xi})$.

The surrogate posterior density may then be written in terms of $\boldsymbol{\xi}$:

$$p(\boldsymbol{\xi},\theta|\mathbf{d}) \propto \tilde{L}(\boldsymbol{\xi}) p_{\mathbf{c}|\theta}(\mathbf{g}(\boldsymbol{\xi})|\theta)\,|\det(D\mathbf{g}(\boldsymbol{\xi}))|\,p(\theta) \tag{90}$$

$$\begin{aligned}
&\propto p(\mathbf{d}|\boldsymbol{\xi}) p_{\boldsymbol{\xi}|\theta}(\boldsymbol{\xi}|\theta) p(\theta) \\
&\propto \exp\left(-\frac{\left(\mathbf{d}-\tilde{\mathbf{G}}(\boldsymbol{\xi})\right)^T\left(\mathbf{d}-\tilde{\mathbf{G}}(\boldsymbol{\xi})\right)}{2\varsigma^2}\right) \\
&\quad \times \left(\frac{\theta}{\varpi^2}\right)^{-\frac{K}{2}} \exp\left(-\frac{\boldsymbol{\xi}^T\boldsymbol{\xi}}{2\theta/\varpi^2}\right) \times \theta^{-\alpha-1}\exp\left(-\frac{\beta}{\theta}\right)
\end{aligned} \tag{91}$$

MCMC sampling from this posterior proceeds as in the previous two sections, except that the full conditional $p(\theta|\boldsymbol{\xi},\mathbf{d})$ used for Gibbs updates is now $IG(\alpha+K/2, \beta+\varpi^2(\sum_{k=1}^{K}\xi_k^2)/2)$.

A useful diagnostic of the stochastic forward solution's fidelity is the probability density of the forward model outputs $u(x_i,t_j,\boldsymbol{\xi})$. Fixing the number of terms K in the K-L expansion of $M(x)$, these densities may be estimated in one of two ways. A direct (and computationally expensive) method is to sample $\mathbf{c}$ and solve the forward problem for each sample, forming a histogram or kernel density estimate from the resulting collection of forward model outputs. Alternatively, one can sample $\boldsymbol{\xi}$ and substitute it into the PC expansion $\tilde{G}(\boldsymbol{\xi})$, again forming a histogram of the resulting values. Figure 38 shows such estimates at two measurement locations and times. While a lower-order PC basis ($p=2$) produces a somewhat poor density estimate, the probability density quickly converges to its true shape—obtained by the direct method—as p increases. Reasonable agreement is obtained

[‖]Note that as $K\to\infty$, we have $\sigma^2(x)/\varpi^2 \to 1$.

even at $p = 4$. Also, note that these densities are not log-normal; their shapes reflect the nonlinearity of the forward maps from $\nu(x)$ to $u(x_i, t_j)$.

Figures 39–41 show inverse solutions corresponding to the random-draw target profile. We fix the number of K-L modes to $K = 6$, since this value provided accurate results in §2.3.3, and vary the order p of the PC basis used to solve the stochastic forward problem. Each of these figures represents MCMC samples from the surrogate posterior (91) that are transformed into realizations of $M(x)$. Even at low order, the posterior mean, standard deviation, and quantiles are not far from their direct counterparts; at $p = 4$ and $p = 6$ these summaries of the posterior distribution are visually indistinguishable from the profiles in Figure 26. Figure 42 quantifies differences between the posterior means/standard deviations obtained with 6 K-L modes and direct forward problem solutions, and those obtained with 6 K-L modes and sampling of the surrogate posterior. Again, we plot the L^2 norm of the differences between MCMC estimates of these quantities: $\|\widehat{\mu}\left(M_K^p(x)\right) - \widehat{\mu}\left(M_K(x)\right)\|_2$ and $\|\widehat{\sigma}\left(M_K^p(x)\right) - \widehat{\sigma}\left(M_K(x)\right)\|_2$. The difference in posterior mean estimates drops more than an order of magnitude from $p = 2$ to $p = 4$ and continues to fall towards $p = 6$. At both $p = 4$ and $p = 6$, we plot results from four separate realizations of the MCMC chain on $\tilde{\pi}_\xi$ (corresponding to different random seeds) in order to illustrate the variability of the MCMC estimates. In this regime, as in the case of K-convergence in §2.3.3, the distance between estimated means of the direct and surrogate posteriors becomes comparable to the standard deviations of the MCMC estimates themselves. (Both Figure 42 and 32(a) show plateaus around 10^{-3} for the random-draw target.) Differences between the estimates of posterior standard deviation show similar dependence on p.

The surrogate posterior accurately captures spatial correlations among possible values of the inverse solution. Figure 43 shows contours of the posterior covariance with varying p: solid lines correspond to the direct posterior, while dashed lines represent the surrogate posterior. Very close agreement is observed at $p = 4$, and this agreement improves further at $p = 6$.

Of course, the ultimate goal of introducing the stochastic forward problem and surrogate posterior is greater computational efficiency. Significant speedups were obtained with a similar approach in [72] for inverse estimation of parameters in PDEs. In the present context, even though the inverse solution is a spatial field, the pattern of computational costs is the same as in [72]. The initial cost of the scheme is offset by the computation of stochastic forward solutions, but then grows very slowly, because the cost per MCMC iteration is orders of magnitude smaller for the surrogate posterior (91) than for direct solutions of the transient diffusion equation (83). Table 1 summarizes the cost at each stage of three representative calculations: inferring the random-draw target by the methods of §2.3.2, §2.3.3, and §2.3.4. For a fixed number of MCMC iterations, K-L based inversion with $K = 6$ is approximately one order of magnitude faster than grid-based inversion, because fewer posterior evaluations are required per MCMC iteration in the former case (using the single-component MCMC sampler). It may be possible to design more efficient MCMC methods for both cases, perhaps updating blocks of components at a time with suitably-shaped proposals, but reducing the chain dimension will inevitably enable greater efficiency [23] and

reduce the number of posterior evaluations required. We also note that comparing the grid-based and K-L based solutions at a fixed number of MCMC iterations is not truly fair to the K-L parameterization, since the $(\mathbf{c}, \theta)$ chain mixes more rapidly than the $(\mathbf{z}, \theta)$ chain (see Figure 37). The Monte Carlo error obtained with 2×10^5 MCMC samples from $p(\mathbf{c}, \theta|\mathbf{d})$ is thus matched at a *larger* number of MCMC samples from $p(\mathbf{m}, \theta|\mathbf{d})$.

The third row of Table 1 shows even greater speedups, entirely independent of the MCMC implementation. The majority of the computational time in this case is spent on the stochastic forward solve. Yet we emphasize that, because it depends only on the prior and forward model, the stochastic forward solve may be performed "offline" before introducing any data. Afterwards, sampling is inexpensive. Here, the cost per posterior evaluation and per MCMC iteration is 1.8 orders of magnitude smaller than for the direct K-L based posterior. Including the time for the stochastic forward solve, inference via exploration of the surrogate reduced-dimensionality posterior is 2.3 orders of magnitude faster than exploration of the direct full-dimensional posterior, with negligible loss of accuracy.

2.3.5 Data length scales

It is useful to consider the behavior of the inverse solution as one coarsens the length scale on which data is collected. For instance, what if observations of the scalar field $u(x,t)$ were limited to the boundaries of the domain, $x = 0.0$ and $x = 1.0$? We still take noisy measurements at 9 successive times, spaced uniformly over the interval $t \in [0.01, 0.03]$. Figure 44 shows the inverse solution corresponding to the sinusoidal target, obtained with a larger number of K-L modes ($K = 10$) and direct forward problem solutions as described in §2.3.3. Contrast these results with those in Figure 28 or 21. The two-sensor results show much greater variability, particularly in the center of the domain, compared to their 13-sensor counterparts. The posterior mean and median still appear sinusoidal, along with the majority of the posterior realizations. All of these profiles are closest to the true solution near the domain boundaries. Asymmetry in the standard deviation profile may be ascribed to asymmetry in the realizations of the sensor noise perturbing observations of $u(x,t)$.

Examining the posterior distribution of $M(x)$ does not complete the story, however. In particular, the posterior distributions of the K-L mode strengths c_k exhibit interesting features when conditioned on coarser data. Figure 45 shows boxplots of the marginal posteriors of c_k and θ, contrasting 13-sensor and 2-sensor inference of the sinusoidal target. First, we note that the hyperparameter θ and the mode strengths c_k—particularly the lower-index c_k—have narrower posterior distributions in the data-rich case. As discussed in §2.3.3, higher-index modes approach a zero-mean limiting distribution, but crucially, this tendency is observed at much lower k in the 2-sensor case. To further elucidate this point, consider the matrix of one- and two-dimensional posterior marginals, shown in Figure 46 for two sensors. The marginal distributions of $c_6, c_7, \ldots$ do indeed appear quite similar in shape and range, and moreover, correlations among the modes weaken at larger k, becoming quite negligible for $k > 6$. We examine the limiting distributions quantitatively as in §2.3.3, plotting posterior conditionals $p(c_6|\mathbf{d}, \theta)$ in Figure 47. The posterior conditionals of c_6 in

the 13-sensor case are far from the corresponding Gaussian prior conditionals $p(c|\theta)$, but in the the 2-sensor case $p(c_6|\mathbf{d},\theta)$ matches the Gaussian prior $N(0,\theta)$ quite closely for various values of θ. Even closer matching is observed at higher k. These observations lead us to expand upon the conjecture of (85), by suggesting that the posterior distribution of K-L modes $c_{k\geq m}$ approaches the prior at *smaller* index m as the length scale of the data is coarsened.

Implications of this statement are numerous. One possibility is that when the data is scarce relative to the complexity of the model (understood here as the spatial complexity of $\log\nu(x)$ as constrained by the prior), dimensionality reduction based on the prior may be improved upon. For instance, in the present case of 2-sensor inversion of the sinusoidal profile, the fraction of the posterior standard deviation contained in modes 6 and higher (integrated over D) is 14.1%. These modes are thus important to the posterior distribution, but if their joint distribution is essentially unchanged by conditioning on the data, the corresponding c_k could be removed from the inference process. Adding unconditioned realizations of $\sum_{k=6}^{\infty}\sqrt{\lambda_k}c_k\phi_k(x)$ to posterior realizations of $M_{K=5}(x)$ would then yield samples from the full posterior.

	time for stochastic forward solve	time per posterior evaluation	time per MCMC iteration	TOTAL time for inversion (2×10^5 samples)
grid-based; $n = 49$	·	5.05×10^{-3}	2.48×10^{-1}	49542
K-L; $K = 6$	·	5.21×10^{-3}	3.43×10^{-2}	6863
K-L and polynomial chaos; $K = 6$, $p = 4$	137	9.2×10^{-5}	5.50×10^{-4}	248

Table 1. Computational times, in seconds, for inference of the "random-draw" target profile (e.g., Figure 20) by three different methods.

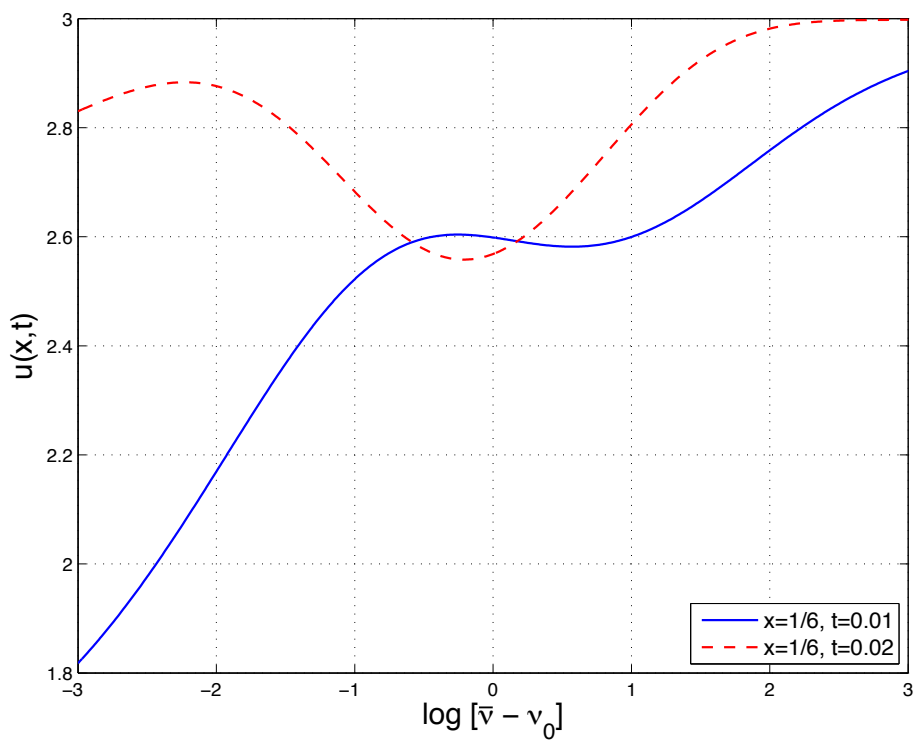
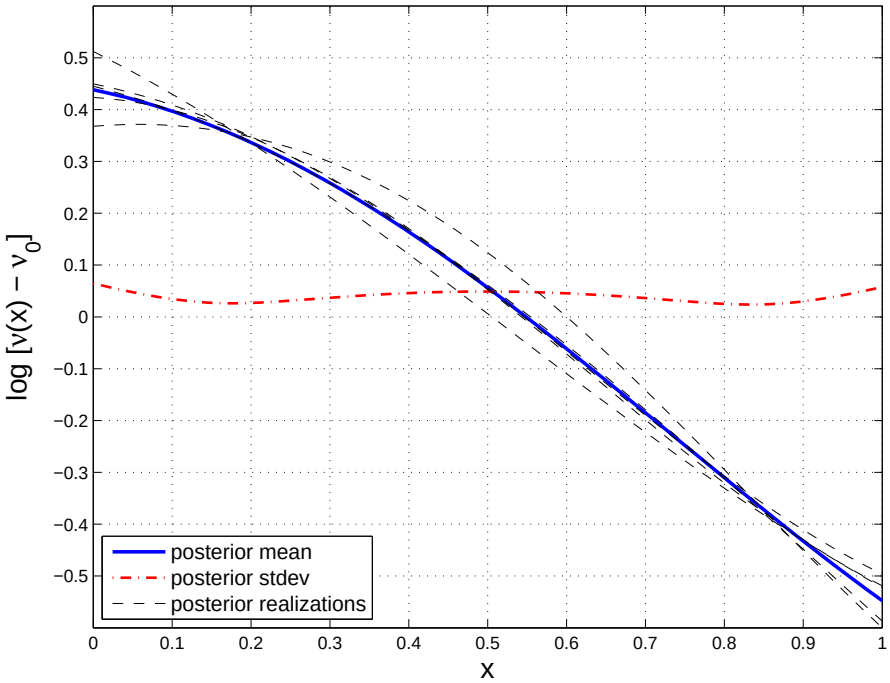
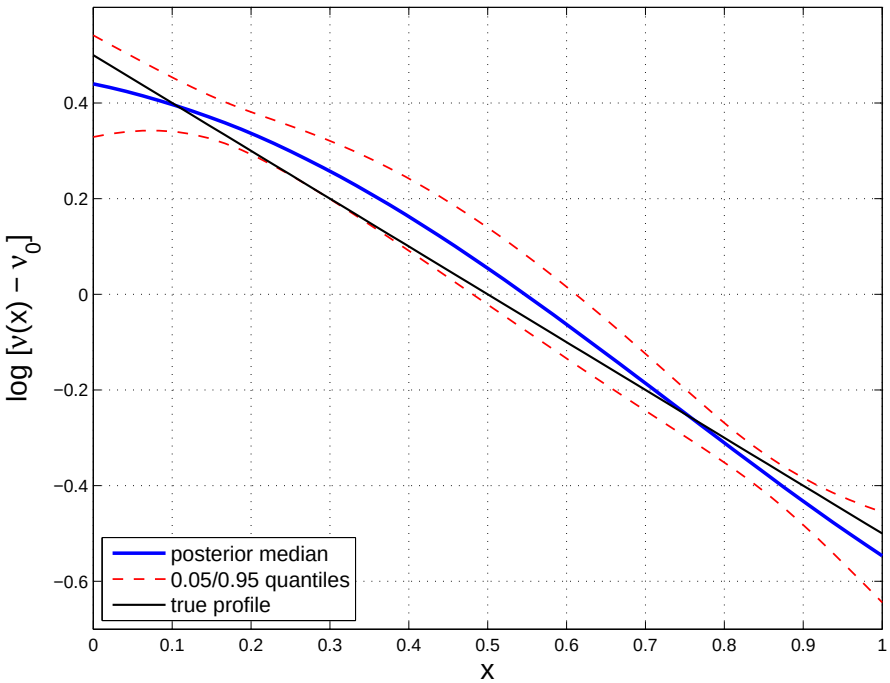


Figure 18. Simple forward map, from a uniform diffusivity $\bar{\nu}$ to the value of the scalar field $u(\bar{\nu}; x, t)$ at two successive times. Source parameters are given in §2.3.1.

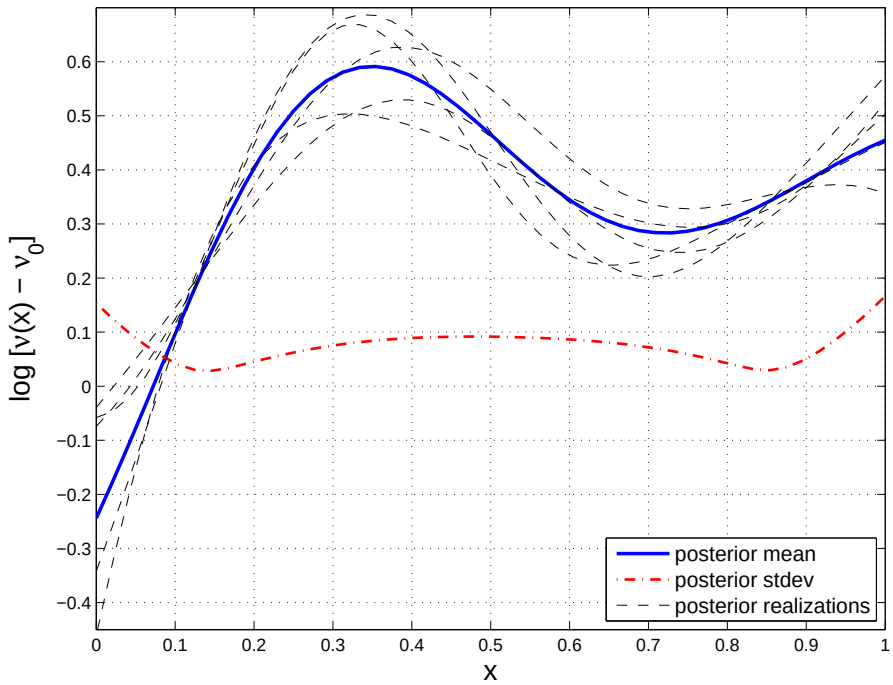


(a) Mean, standard deviation, and five posterior realizations.

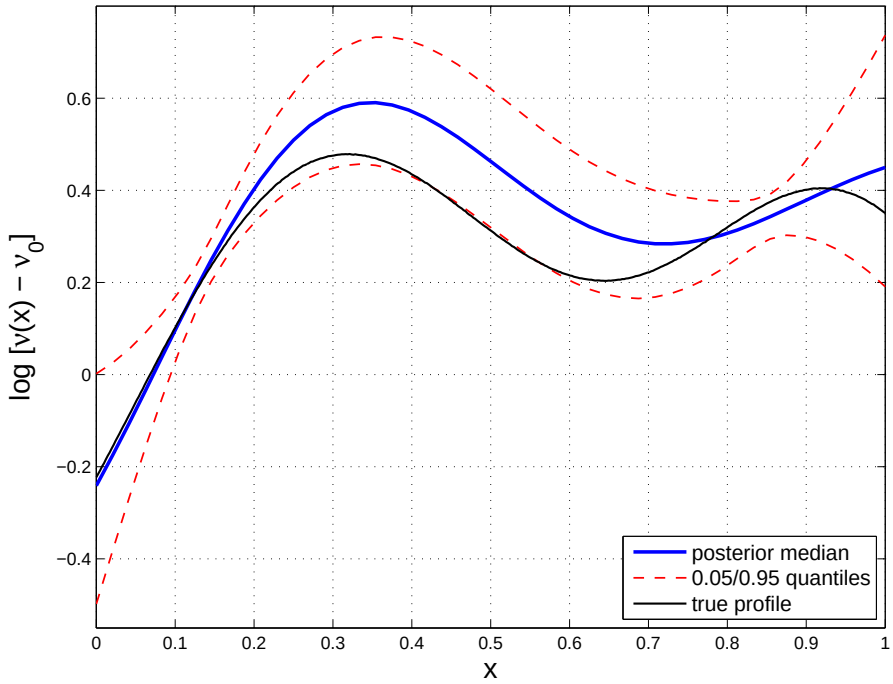


(b) Median, 1-D credibility intervals, and true profile.

Figure 19. Grid-based inversion of a linear log-diffusivity profile.

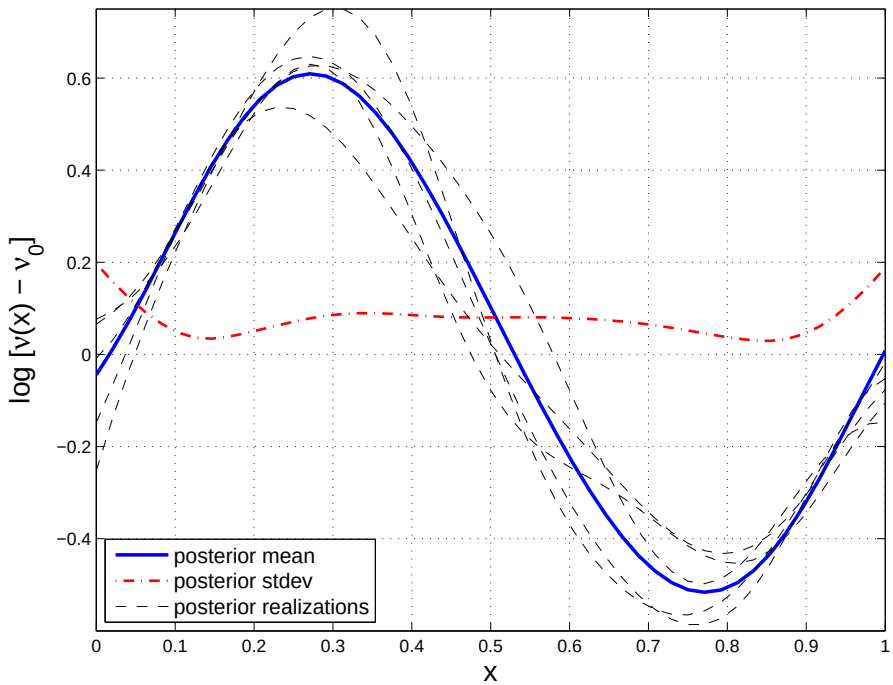


(a) Mean, standard deviation, and five posterior realizations.

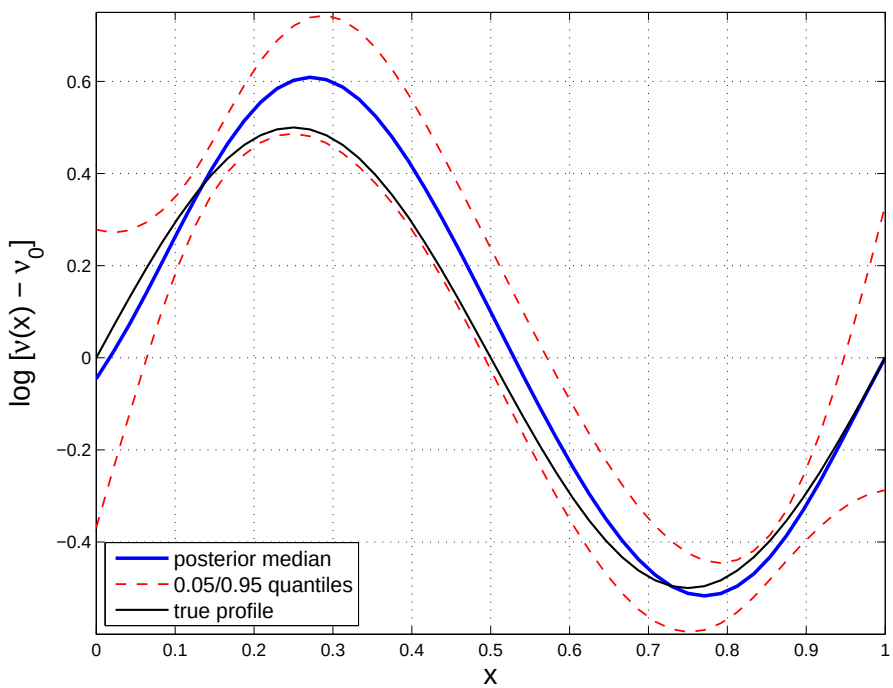


(b) Median, 1-D credibility intervals, and true profile.

Figure 20. Grid-based inversion of a log-diffusivity profile randomly drawn from a Gaussian process with $L = 0.3$ and scale parameter $\theta = 1.0$.

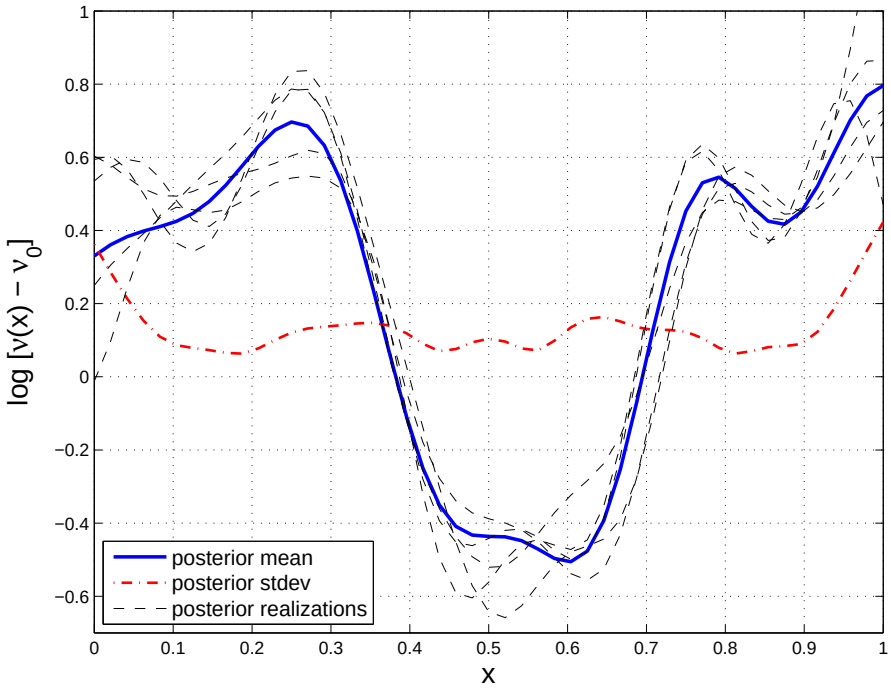


(a) Mean, standard deviation, and five posterior realizations.

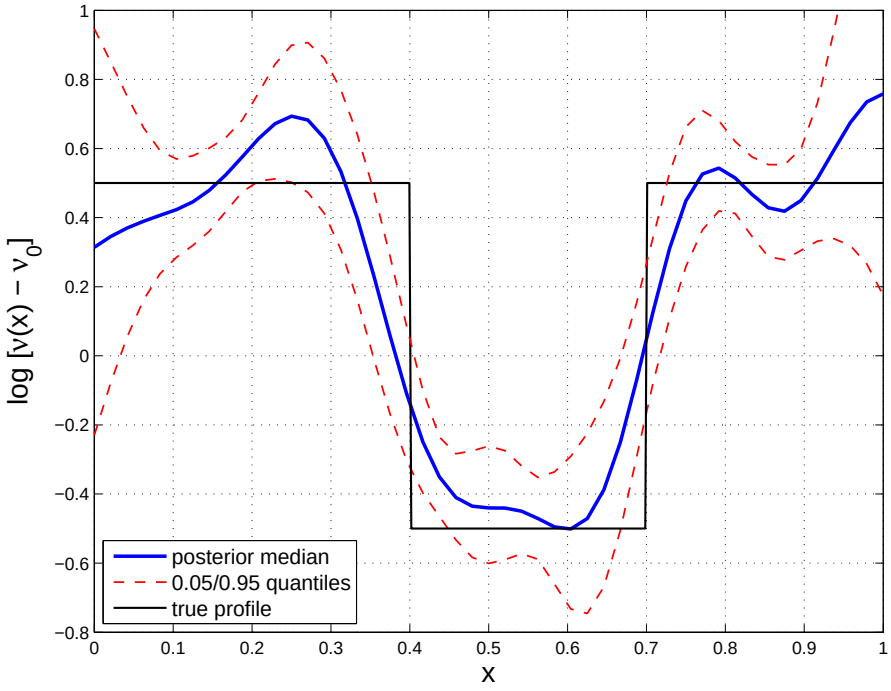


(b) Median, 1-D credibility intervals, and true profile.

Figure 21. Grid-based inversion of a sinusoidal log-diffusivity profile.

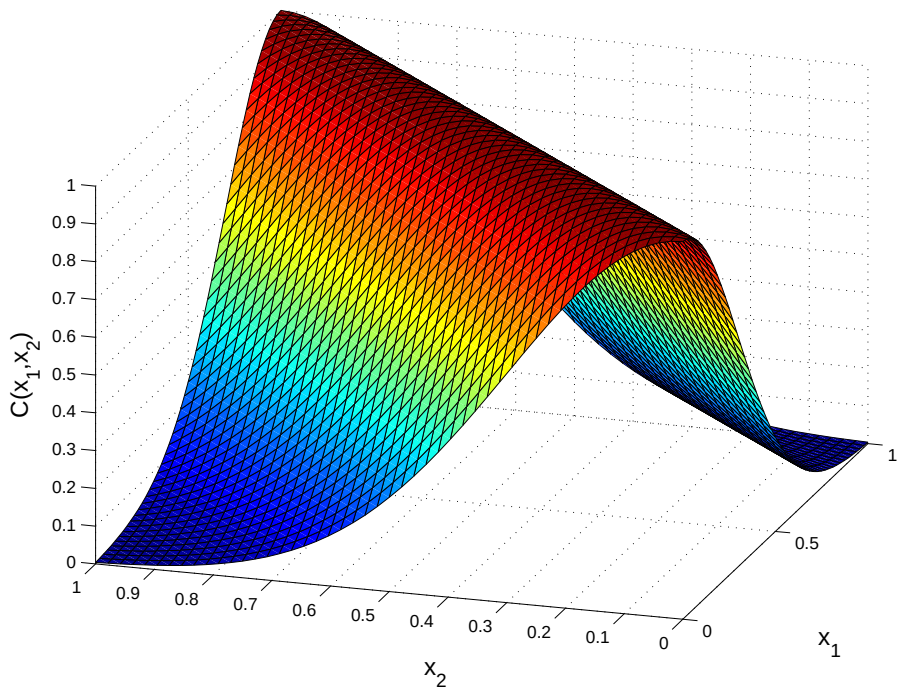


(a) Mean, standard deviation, and five posterior realizations.

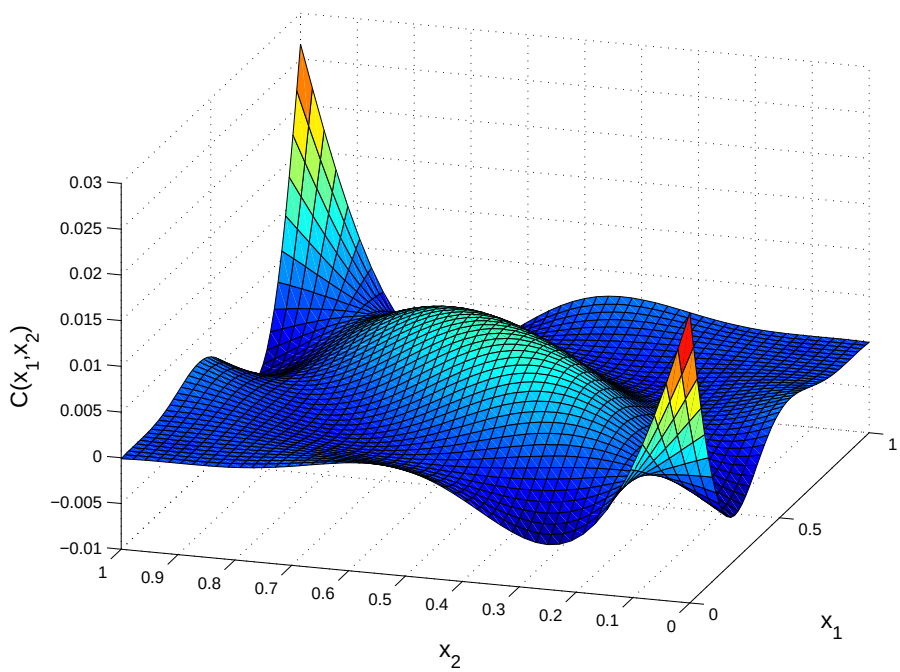


(b) Median, 1-D credibility intervals, and true profile.

Figure 22. Grid-based inversion of a well-shaped log-diffusivity profile.

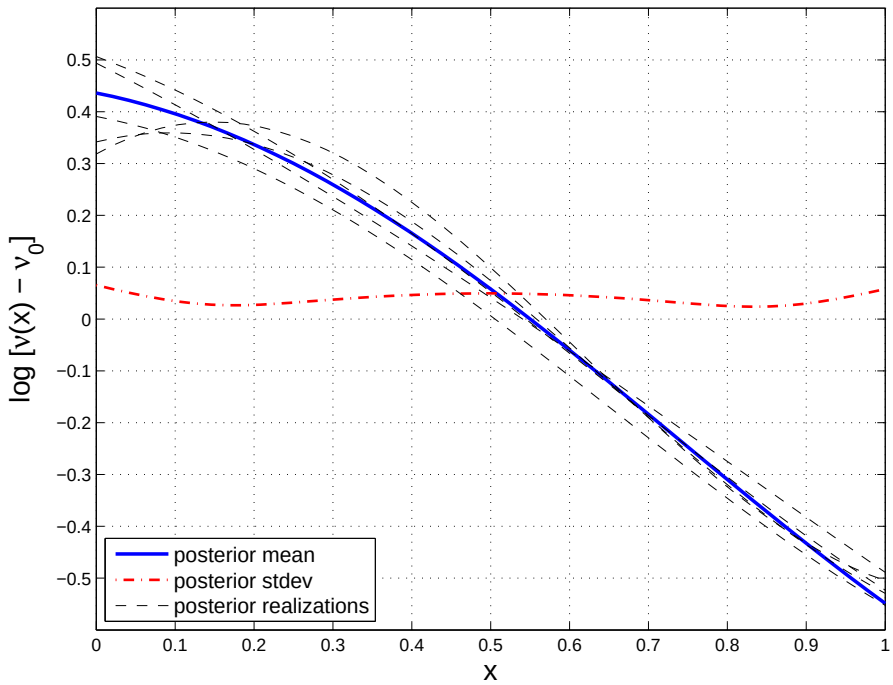


(a) Prior covariance, $L = 0.3$.

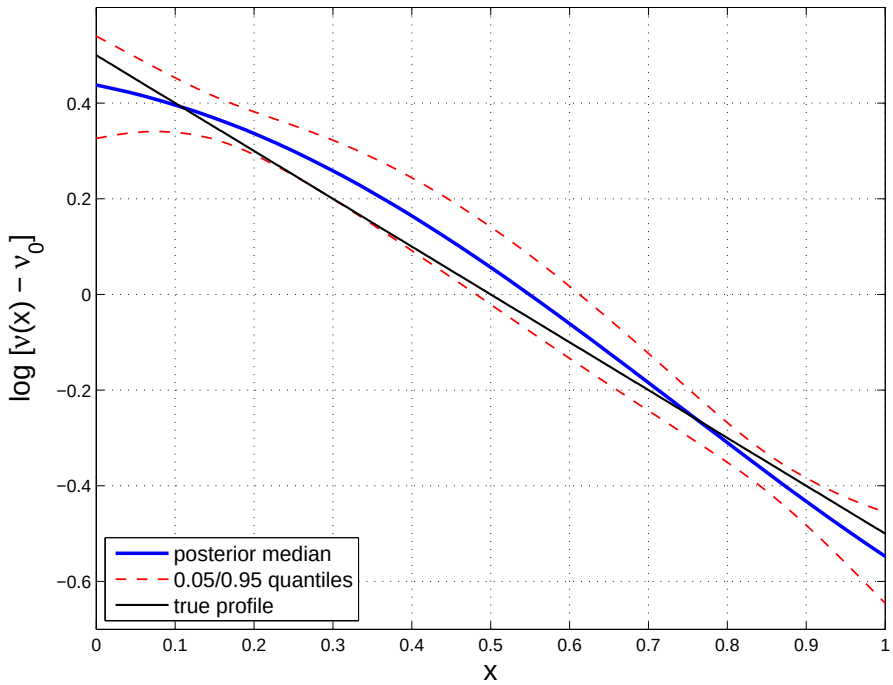


(b) Posterior covariance with the random-draw profile.

Figure 23. Change in the covariance of $M(x)$ from the prior to the posterior, for inference of the random-draw target. Both (a) and (b) reflect marginalization over the scale parameter θ. In this example only, the hyperprior is $\theta \sim IG(3,2)$.

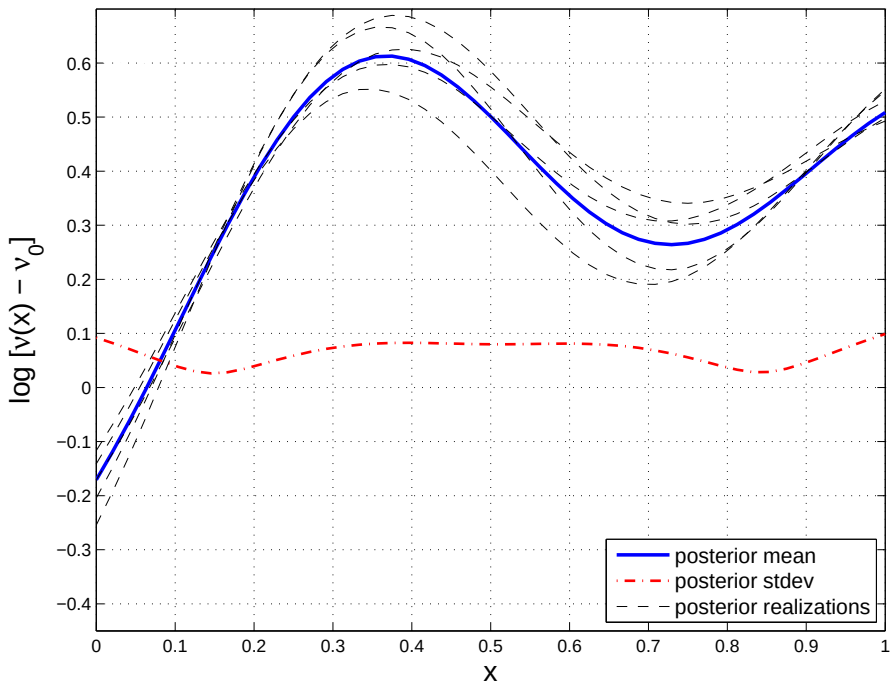


(a) Mean, standard deviation, and five posterior realizations.

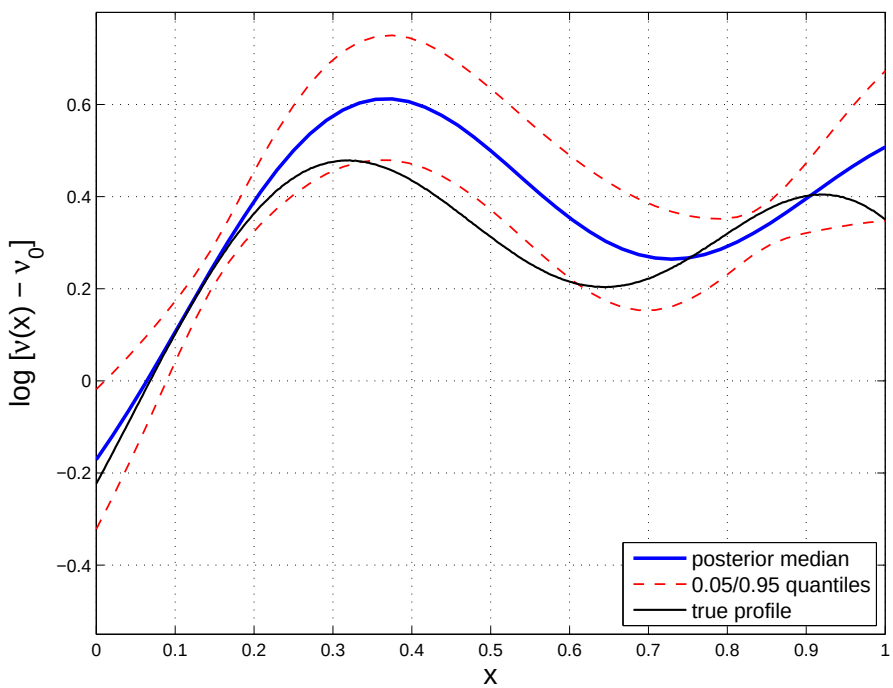


(b) Median, 1-D credibility intervals, and true profile.

Figure 24. K-L-based inversion of the linear log-diffusivity profile, $K = 4$.

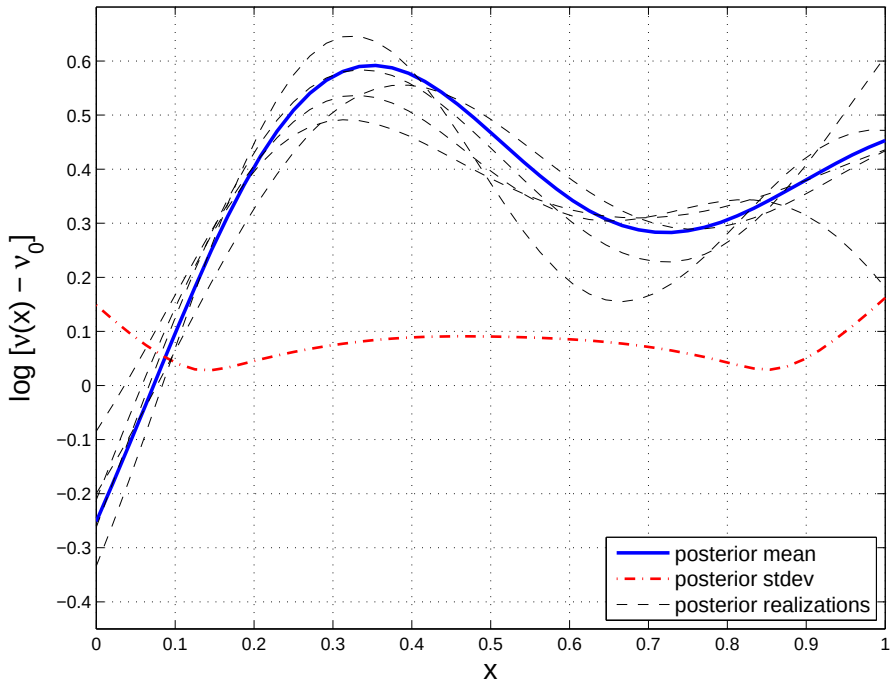


(a) Mean, standard deviation, and five posterior realizations.

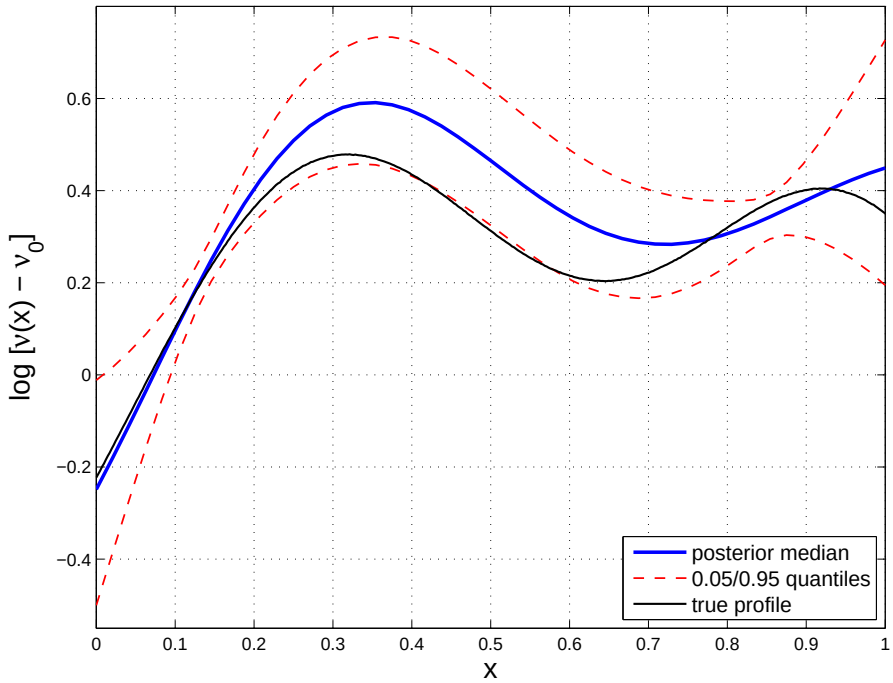


(b) Median, 1-D credibility intervals, and true profile.

Figure 25. K-L-based inversion of the random-draw log-diffusivity profile, $K = 4$.

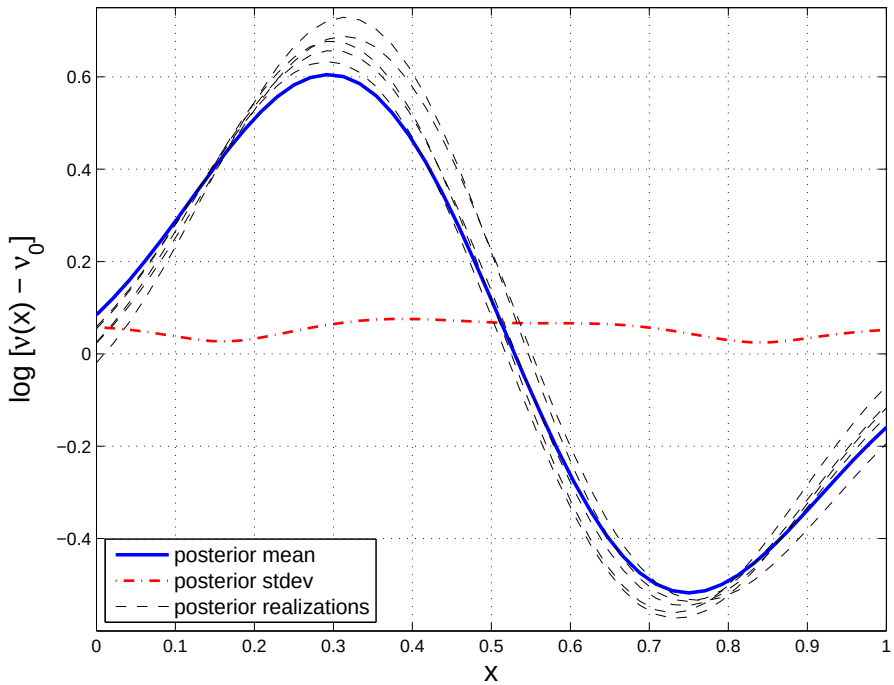


(a) Mean, standard deviation, and five posterior realizations.

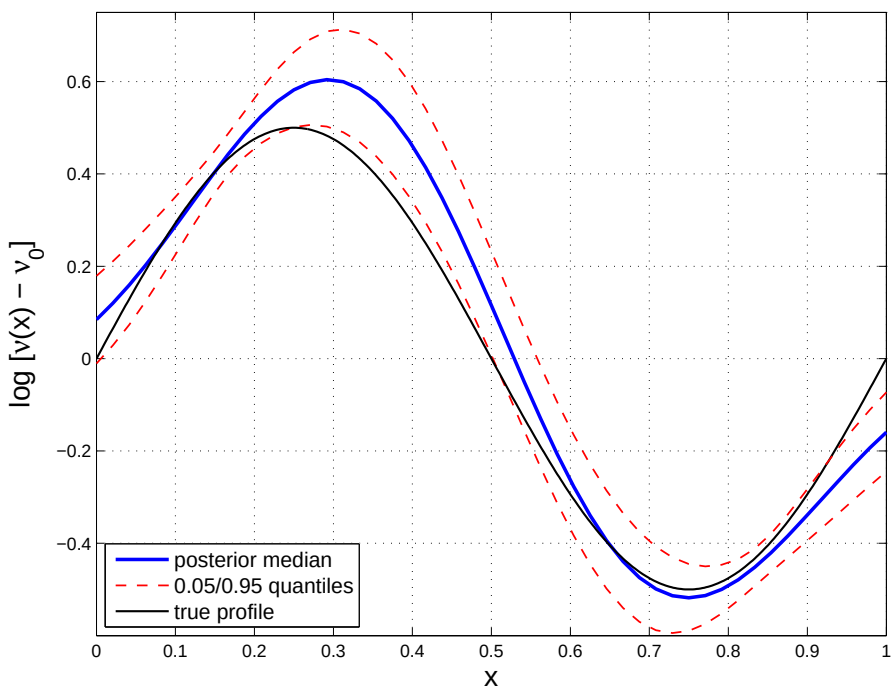


(b) Median, 1-D credibility intervals, and true profile.

Figure 26. K-L-based inversion of the random-draw log-diffusivity profile, $K = 6$.

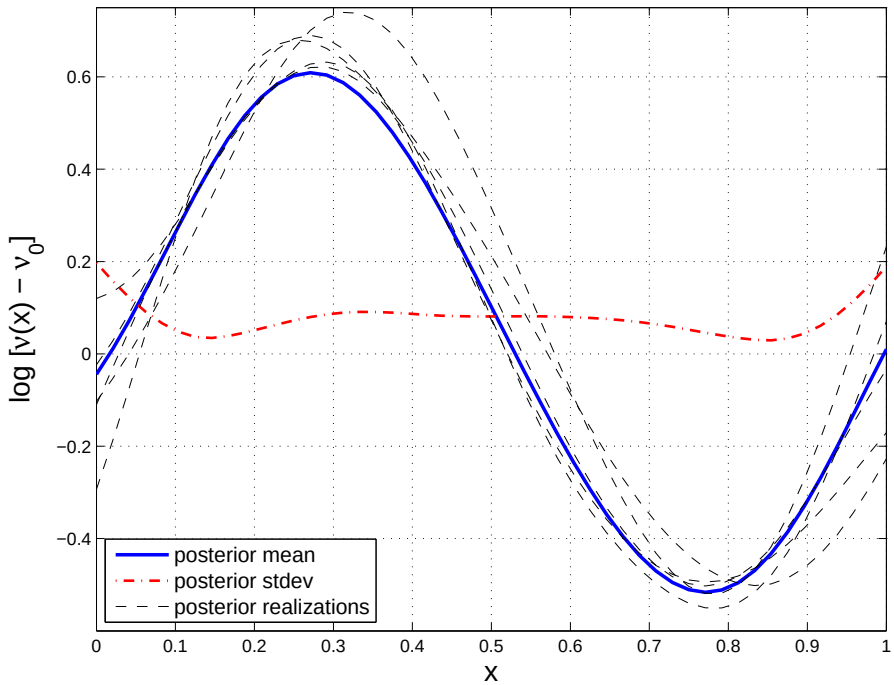


(a) Mean, standard deviation, and five posterior realizations.

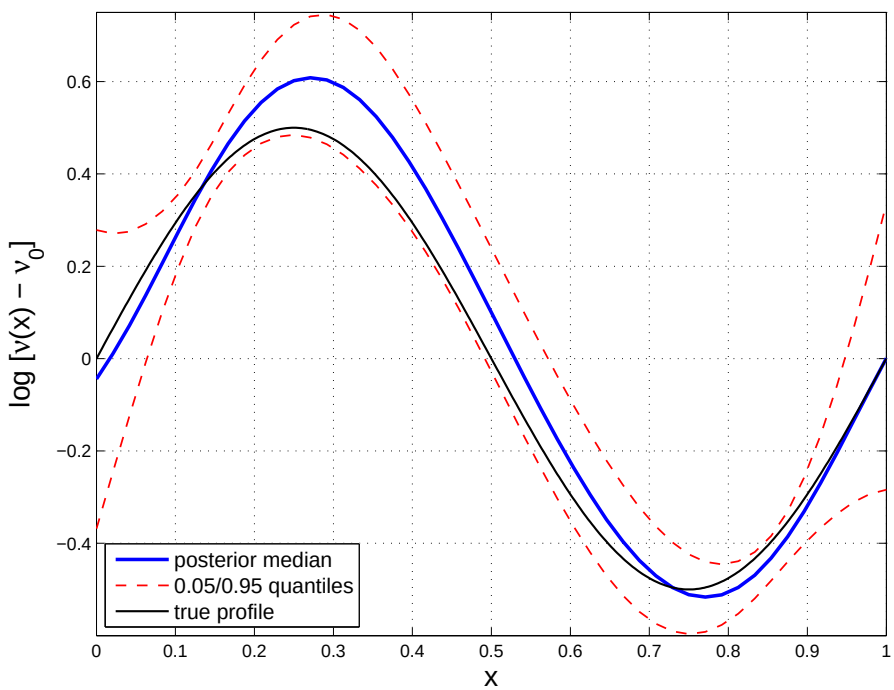


(b) Median, 1-D credibility intervals, and true profile.

Figure 27. K-L-based inversion of the sinusoidal log-diffusivity profile, $K = 4$.

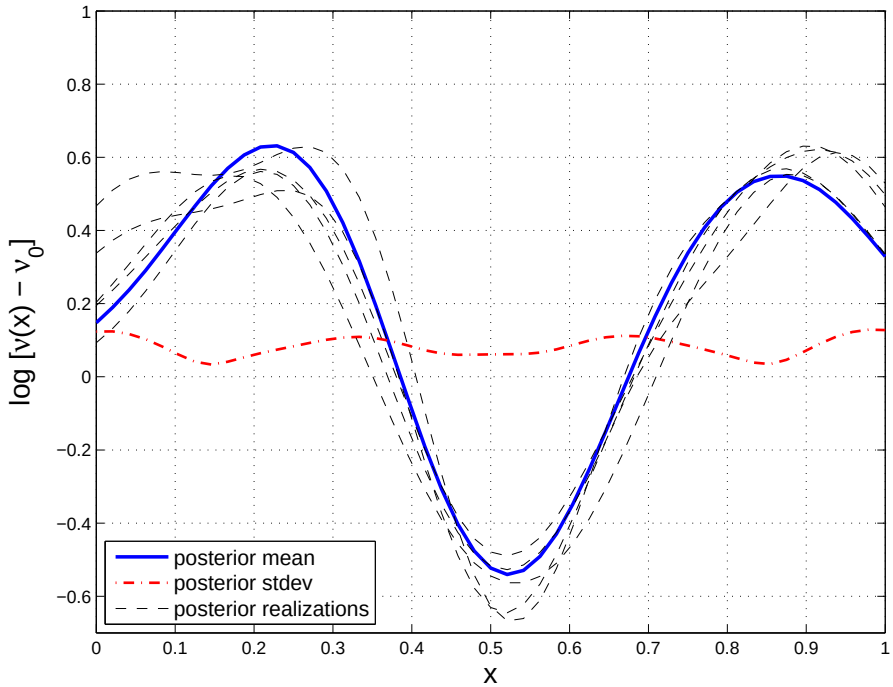


(a) Mean, standard deviation, and five posterior realizations.

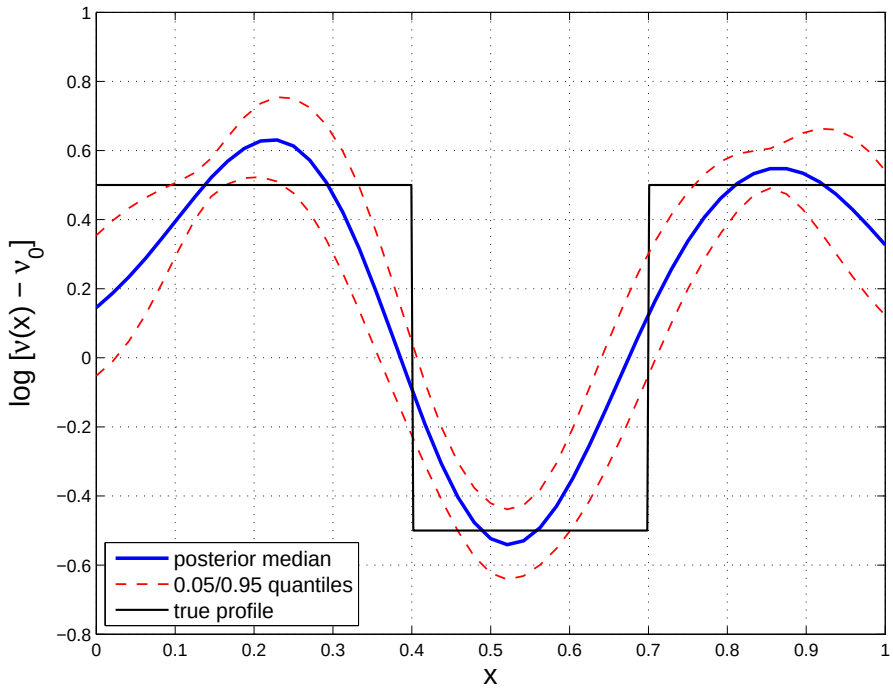


(b) Median, 1-D credibility intervals, and true profile.

Figure 28. K-L-based inversion of the sinusoidal log-diffusivity profile, $K = 10$.

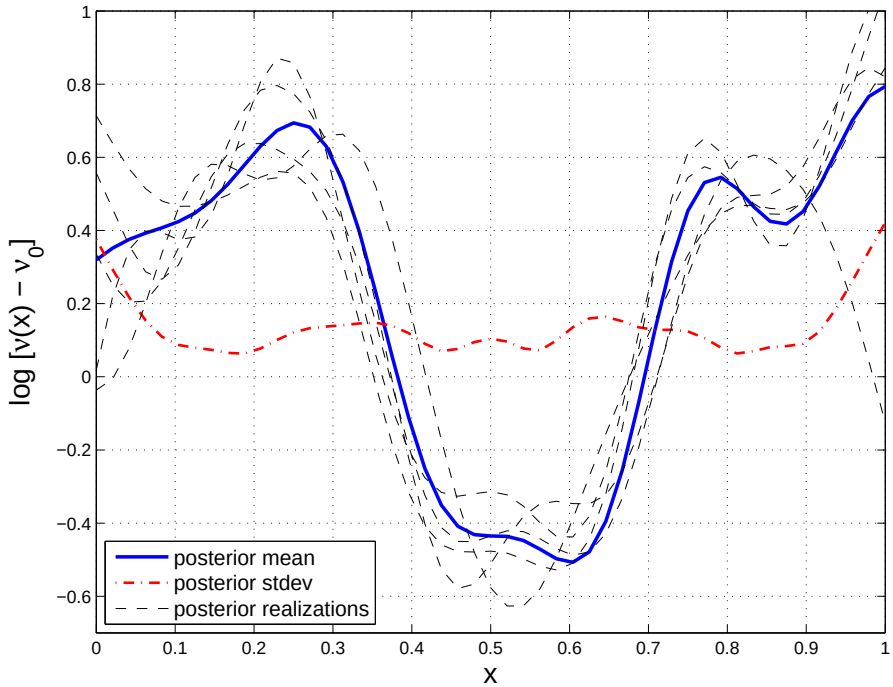


(a) Mean, standard deviation, and five posterior realizations.

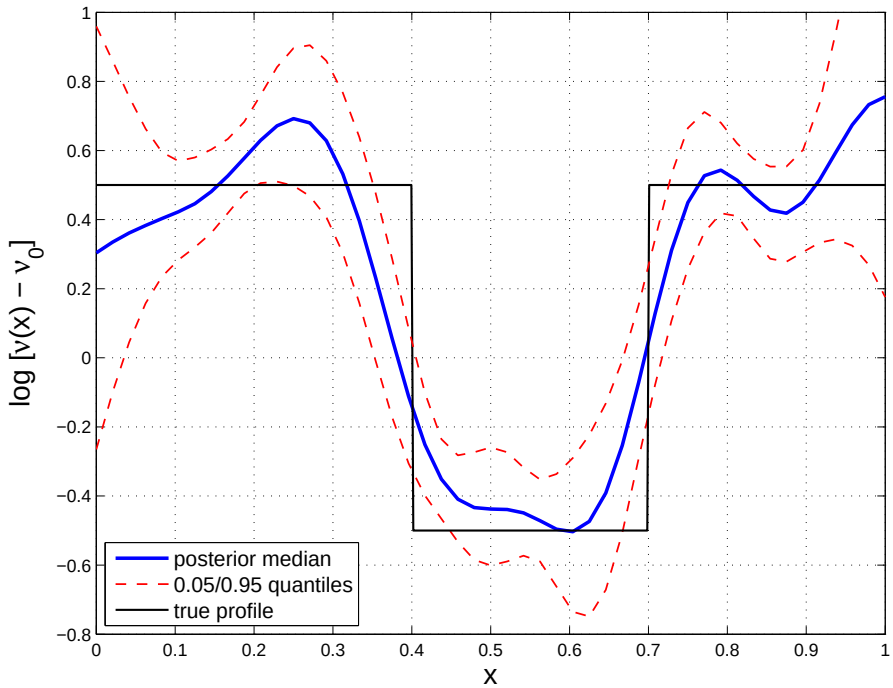


(b) Median, 1-D credibility intervals, and true profile.

Figure 29. K-L-based inversion of the well-shaped log-diffusivity profile, $K = 7$.



(a) Mean, standard deviation, and five posterior realizations.



(b) Median, 1-D credibility intervals, and true profile.

Figure 30. K-L-based inversion of the well-shaped log-diffusivity profile, $K = 15$.

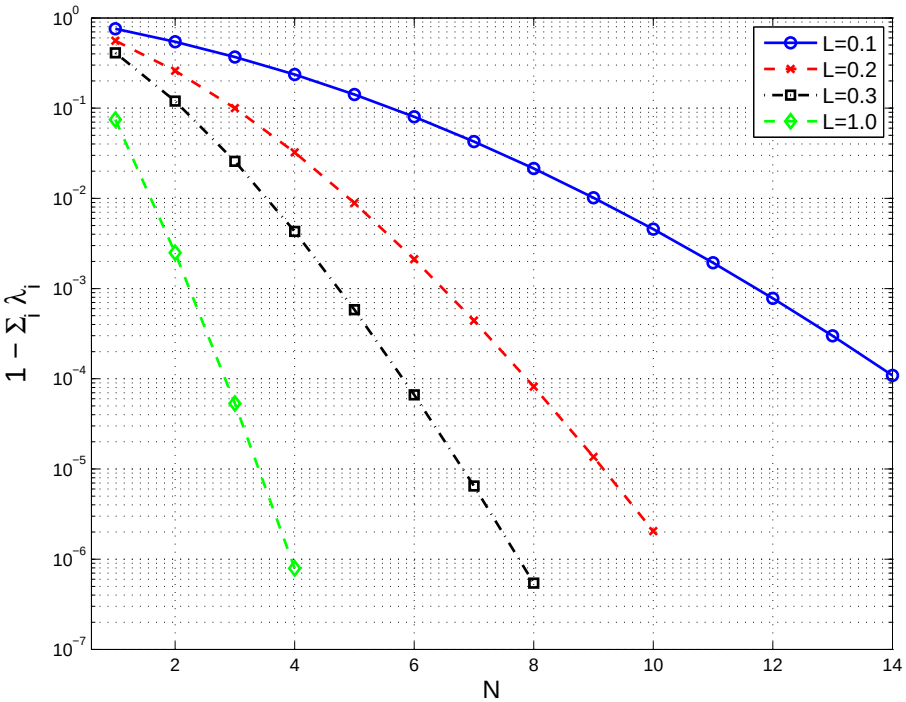
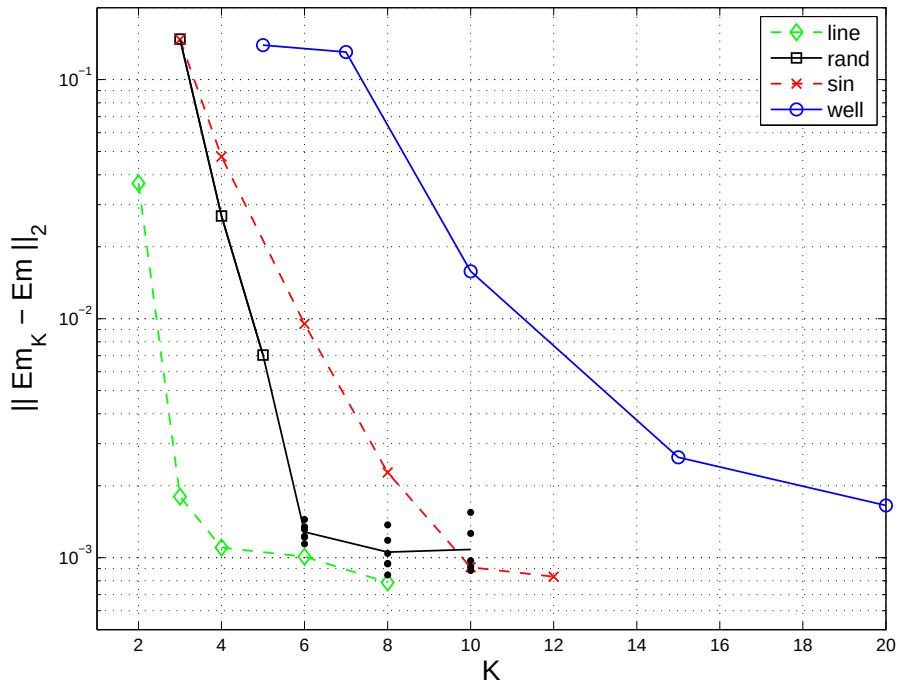
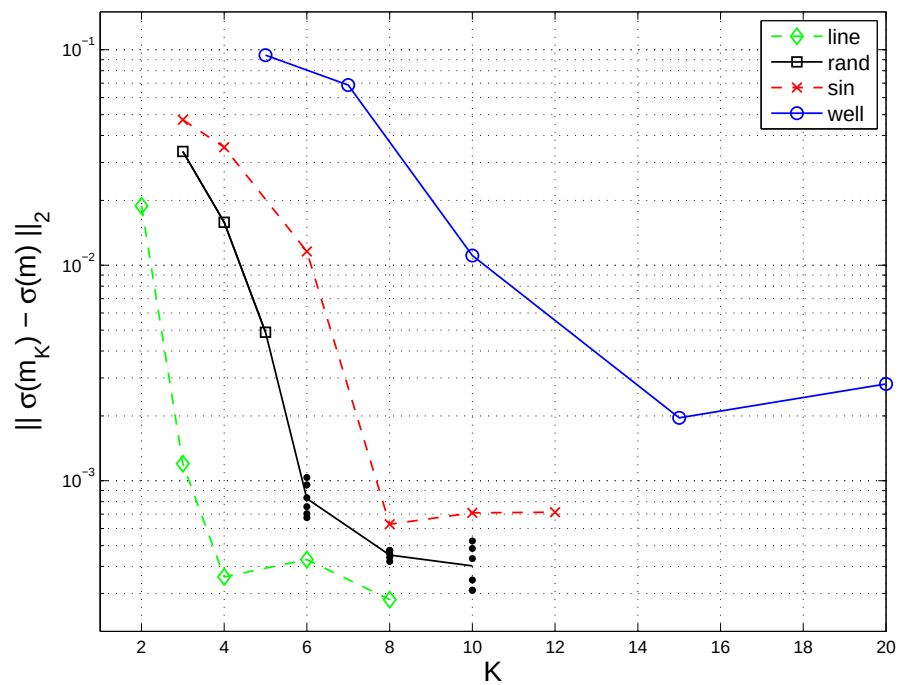


Figure 31. Decay of K-L eigenvalues with different prior correlation lengths L. Vertical axis shows the missing fraction of the prior variance, $1 - \sum_i^K \lambda_i$, versus K.

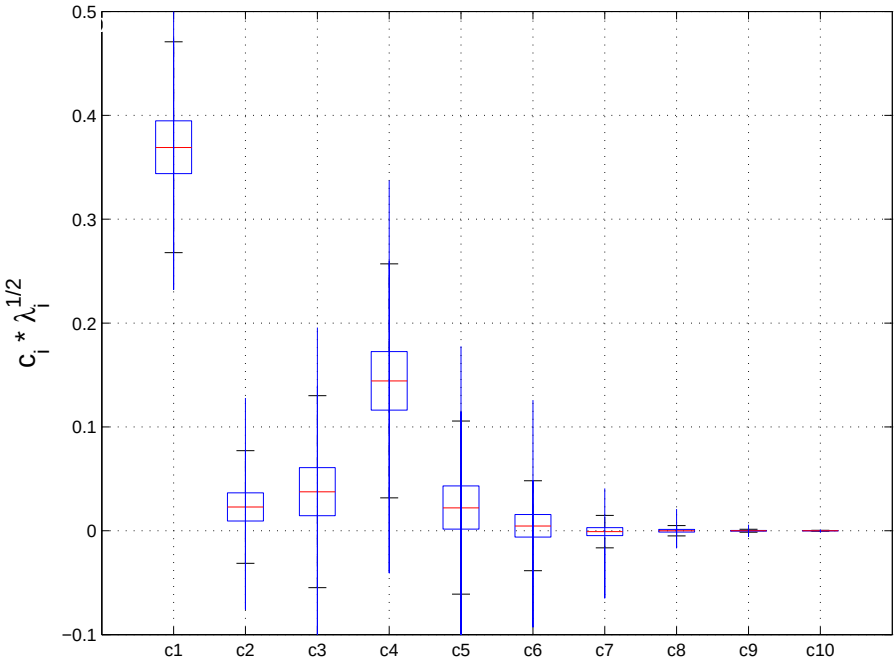


(a) $L^2(D)$-norm of differences in the posterior mean.

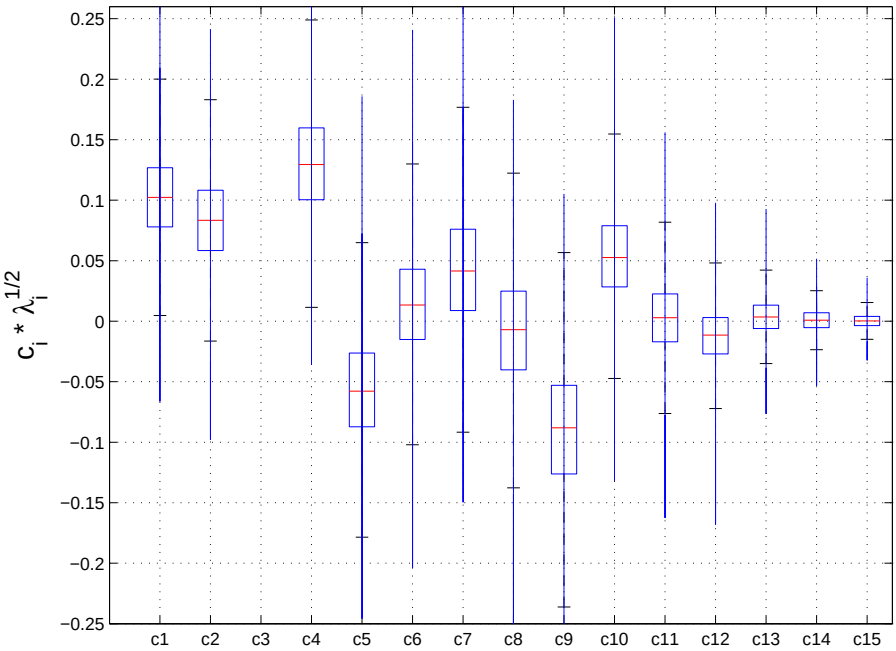


(b) $L^2(D)$-norm of differences in the posterior standard deviation.

Figure 32. Differences in posterior moments computed via grid-based inversion and K-L based inversion, versus K. At $K = 6$, 8, and 10, repeated symbols (•) correspond to multiple MCMC simulations from the posterior of the random-draw target.

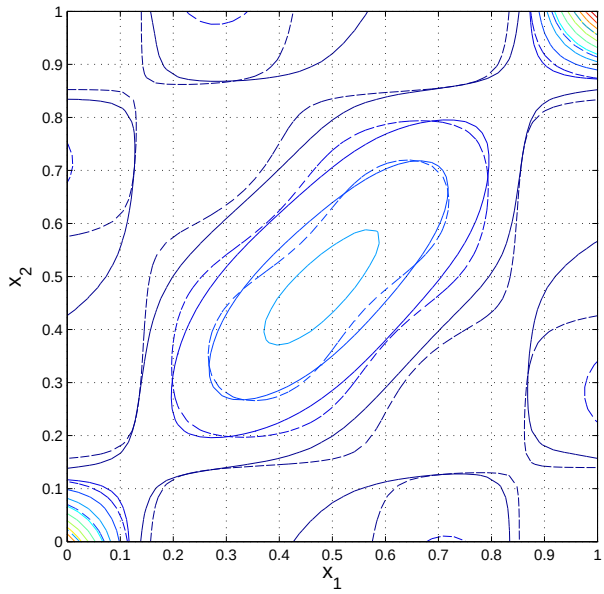


(a) Random-draw target.

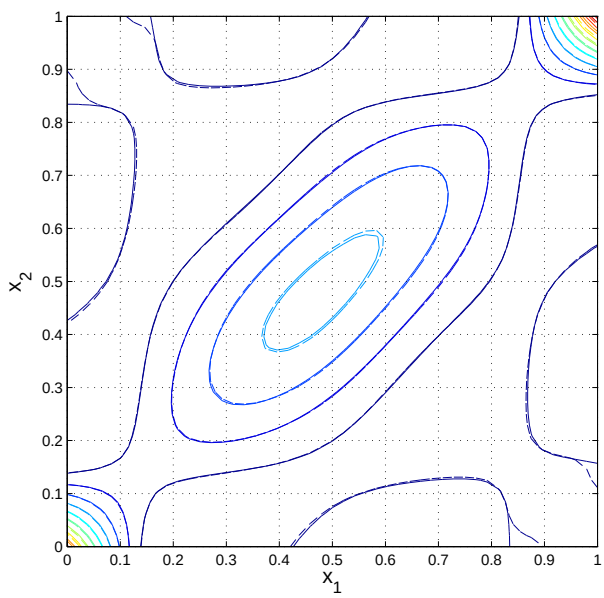


(b) Well-shaped target.

Figure 33. Boxplot of the posterior marginals of the K-L mode strengths c_i, scaled by $\sqrt{\lambda_i}$.

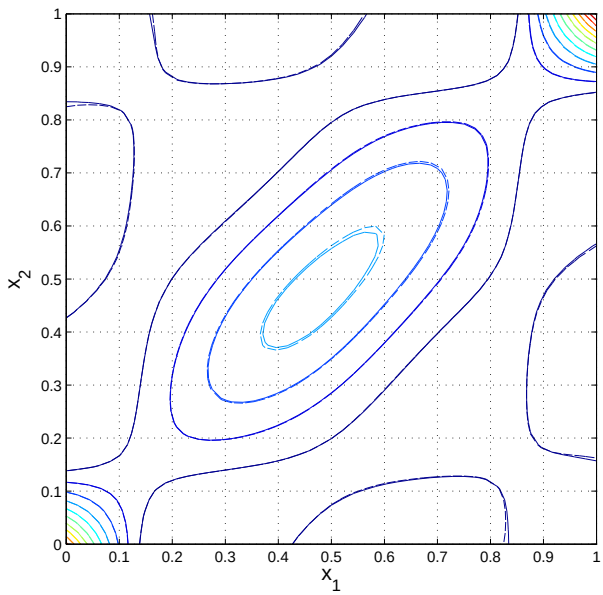


(a) $K = 4$



(b) $K = 6$

Figure 34. Contours of the posterior covariance, $\mathrm{Cov}\,[M(x_1), M(x_2)]$. Solid lines are obtained via grid-based inversion; dashed lines are obtained via reduced-dimensionality inversion with K K-L modes. All are for inference of the random-draw target.



(c) $K = 8$

Figure 34. (cont.) Contours of the posterior covariance, $\mathrm{Cov}\,[M(x_1), M(x_2)]$. Solid lines are obtained via grid-based inversion; dashed lines are obtained via reduced-dimensionality inversion with K K-L modes.

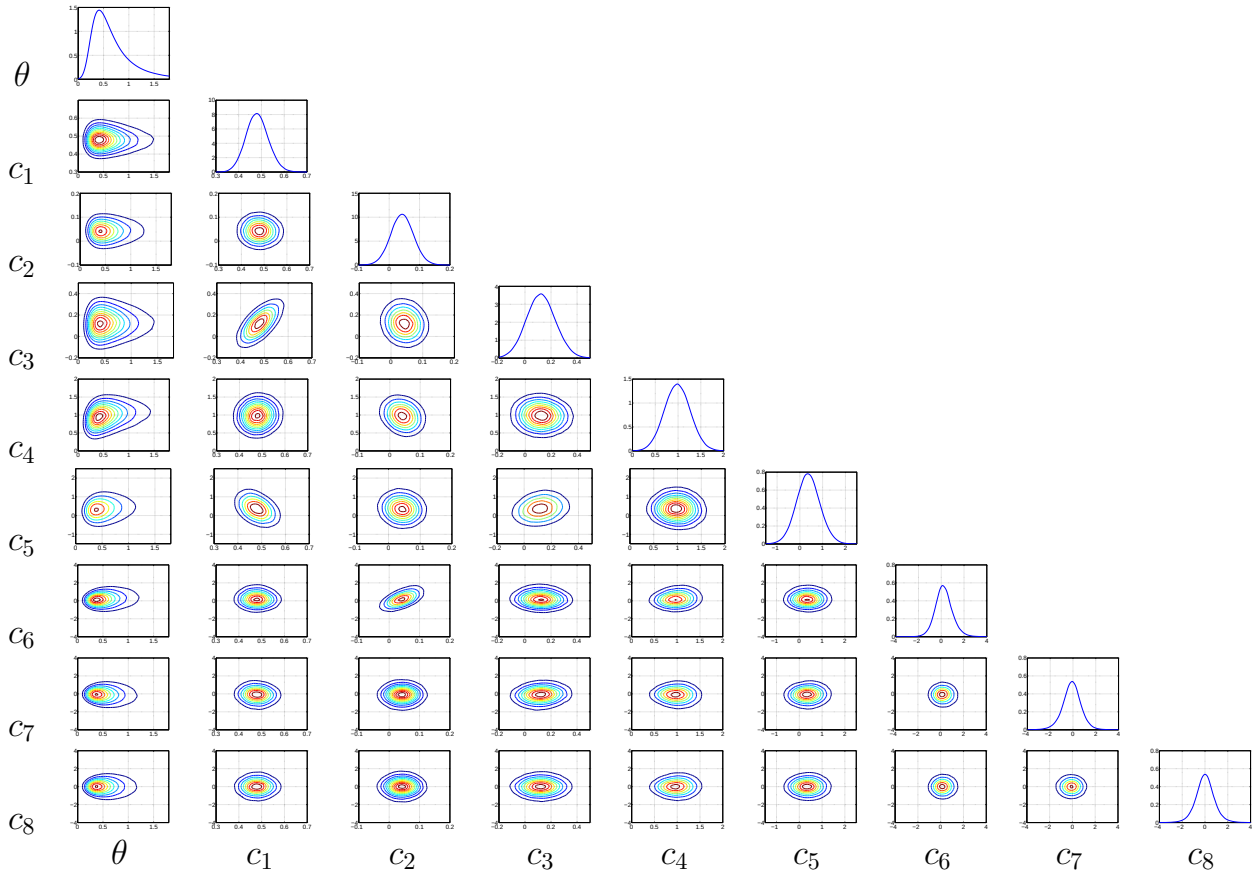
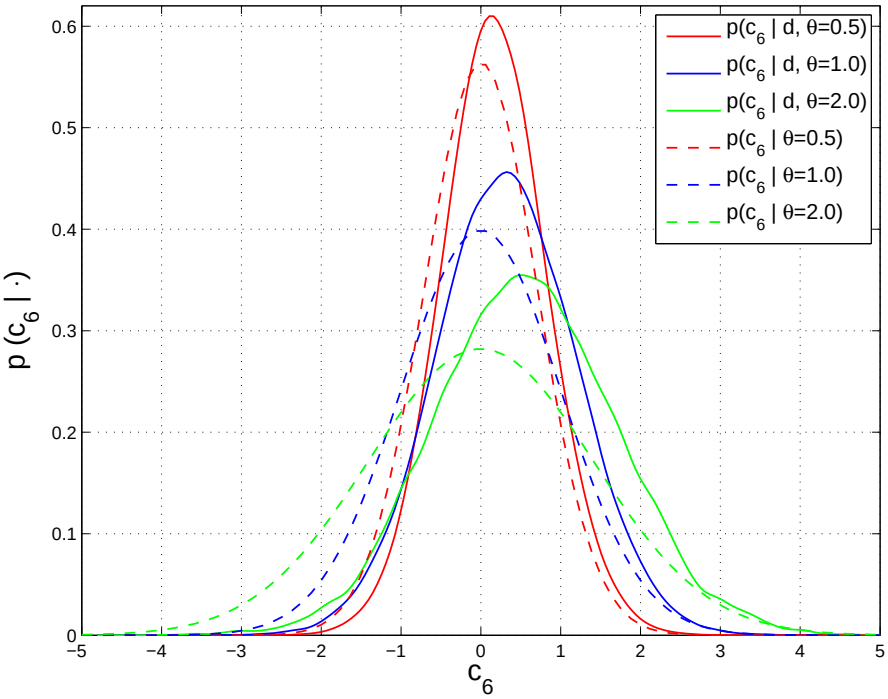
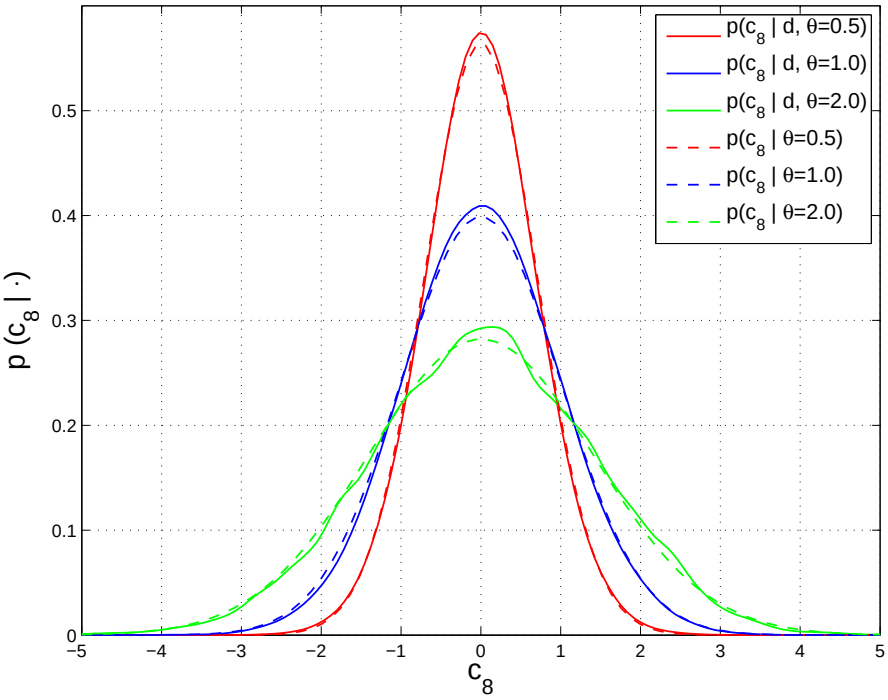


Figure 35. 1-D and 2-D posterior marginals of the K-L mode strengths c_i and the hyperparameter θ, for inference of the random-draw target.

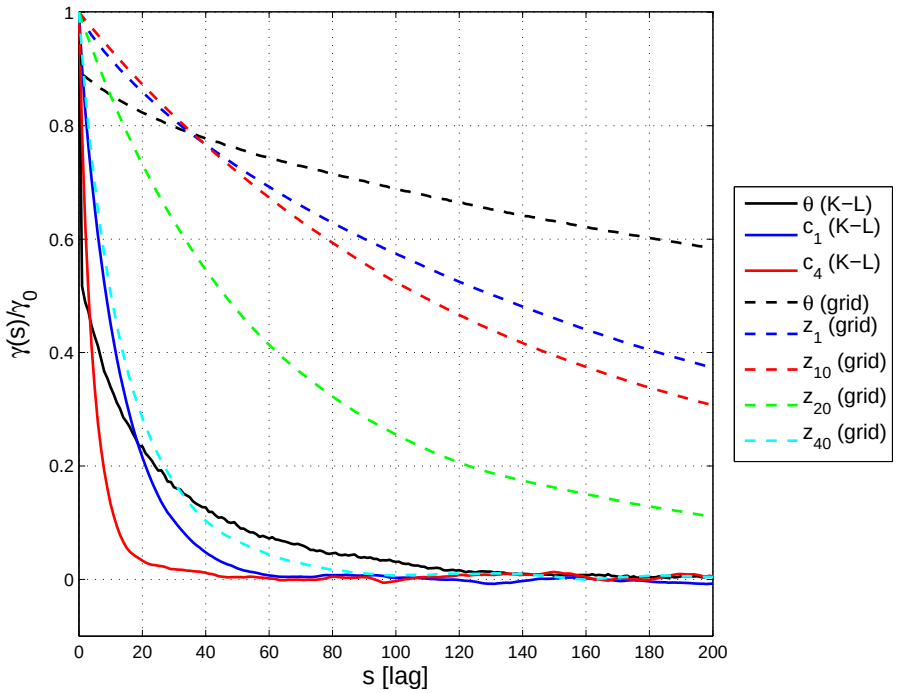


(a) c_6 prior and posterior conditional distributions

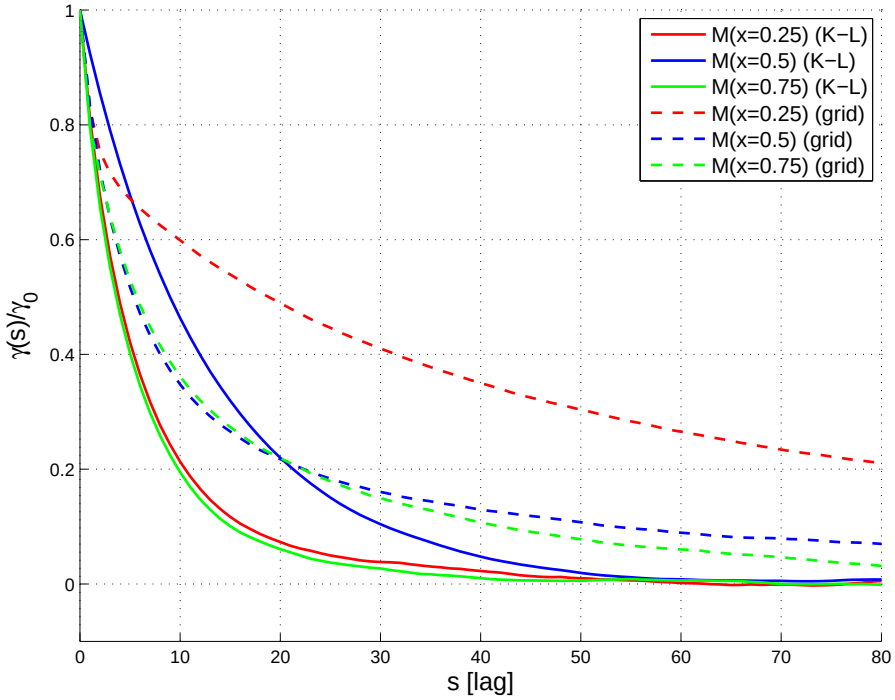


(b) c_8 prior and posterior conditional distributions

Figure 36. Limiting distributions of K-L modes, 13 sensors; inference of the random-draw target.

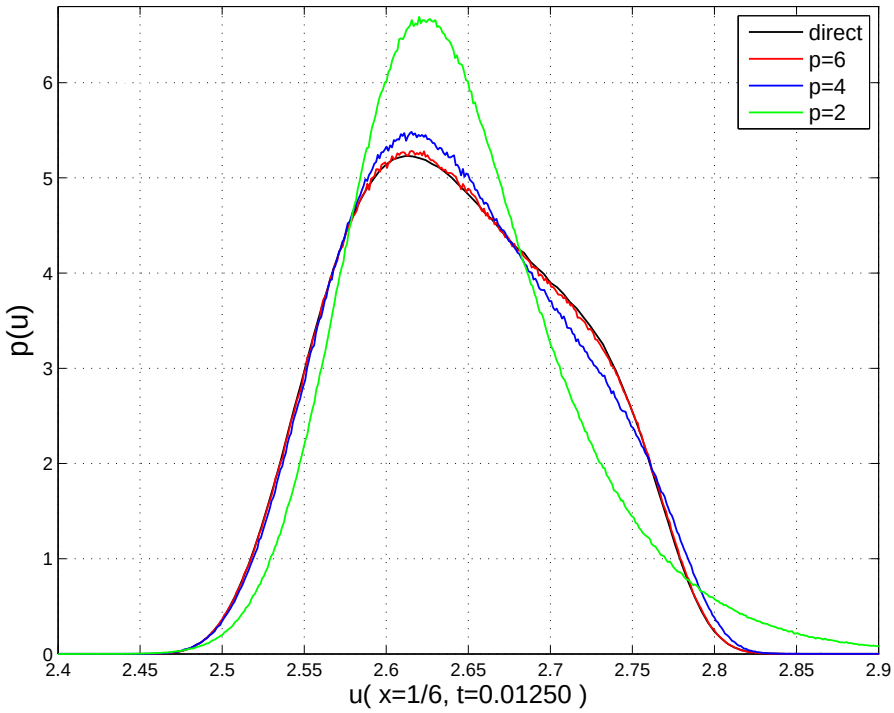


(a) Chain variables.

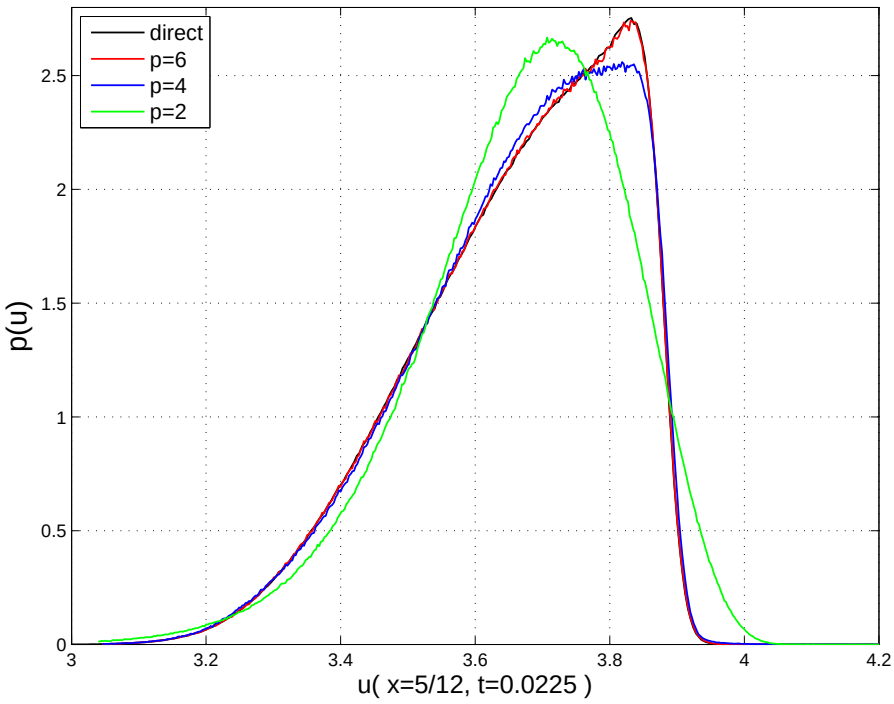


(b) $M(x_i)$.

Figure 37. Autocorrelation at lag s of components of the MCMC chain, and of field values $M(x)$ parameterized by the MCMC chain variables.

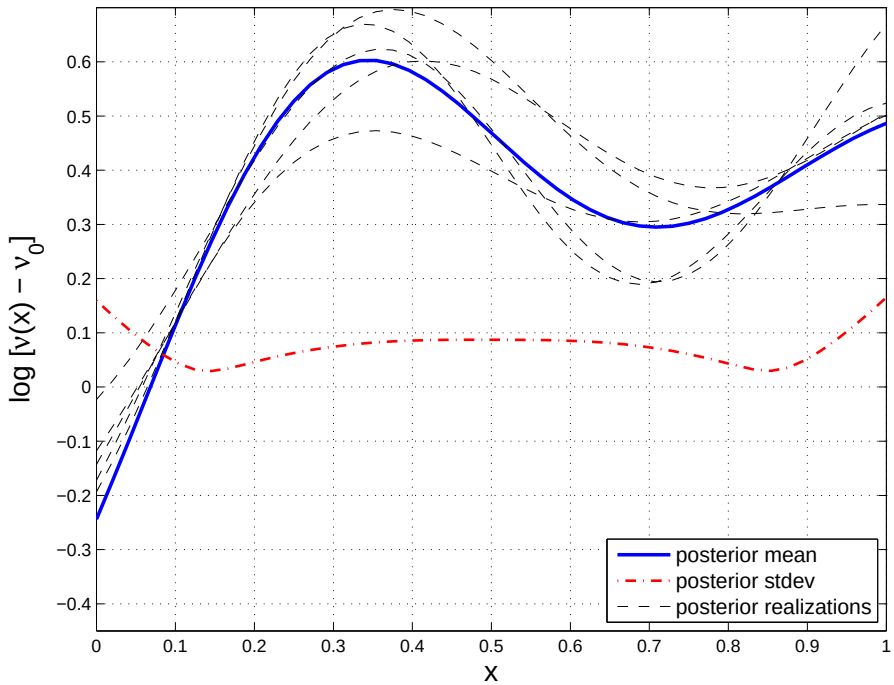


(a) $u(x = 1/6, t = 0.0125)$.

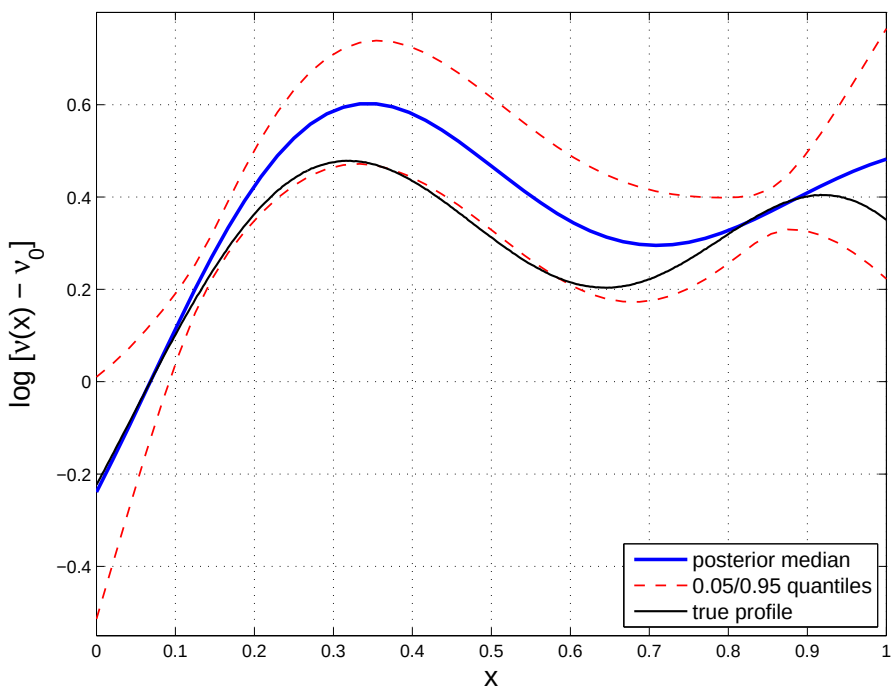


(b) $u(x = 5/12, t = 0.0225)$.

Figure 38. Probability density of the scalar $u(x,t)$ at two measurement locations and times, given input uncertainty in the diffusivity field. Here, $\nu(x)$ is the log-normal random field defined in §2.3.4.

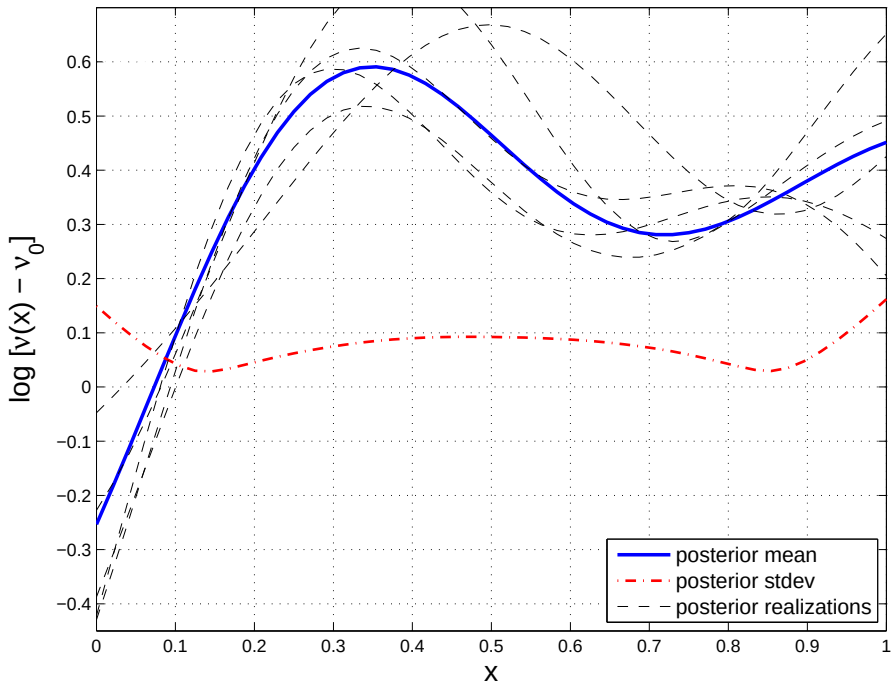


(a) Mean, standard deviation, and five posterior realizations.

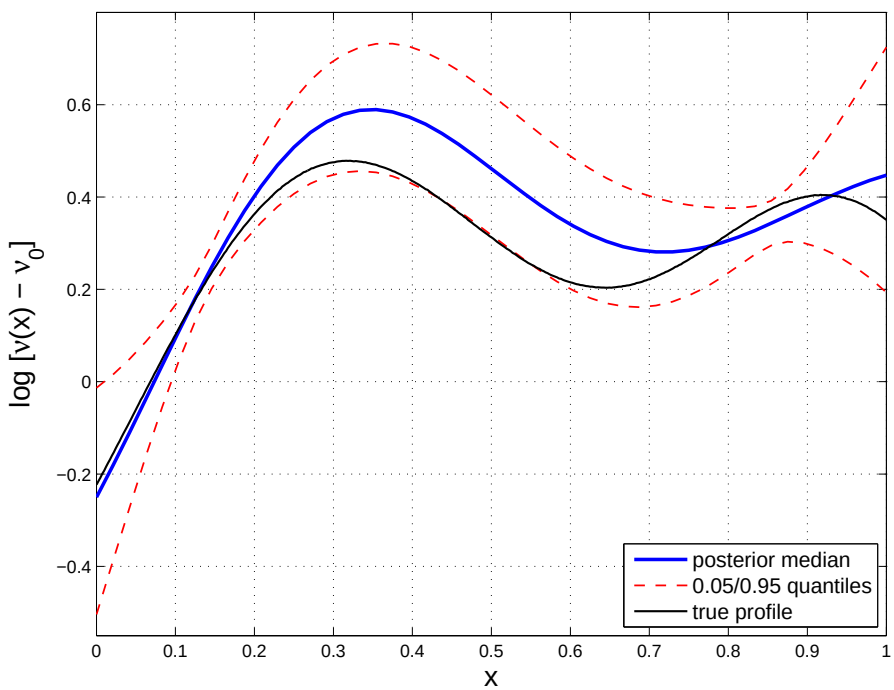


(b) Median, 1-D credibility intervals, and true profile.

Figure 39. Inversion of the random-draw log-diffusivity profile using K-L and polynomial chaos, $K = 6$, $p = 2$.

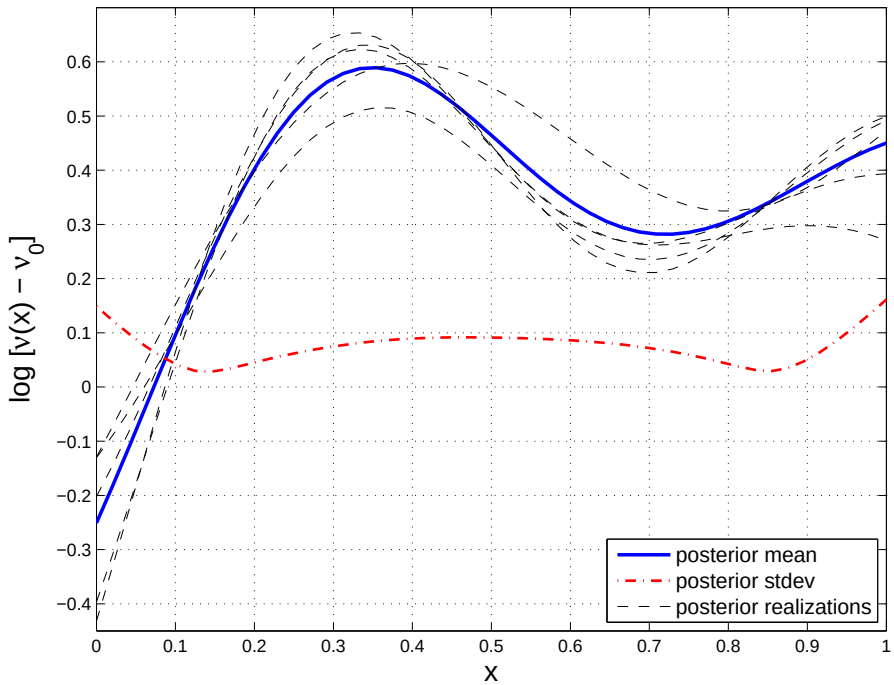


(a) Mean, standard deviation, and five posterior realizations.

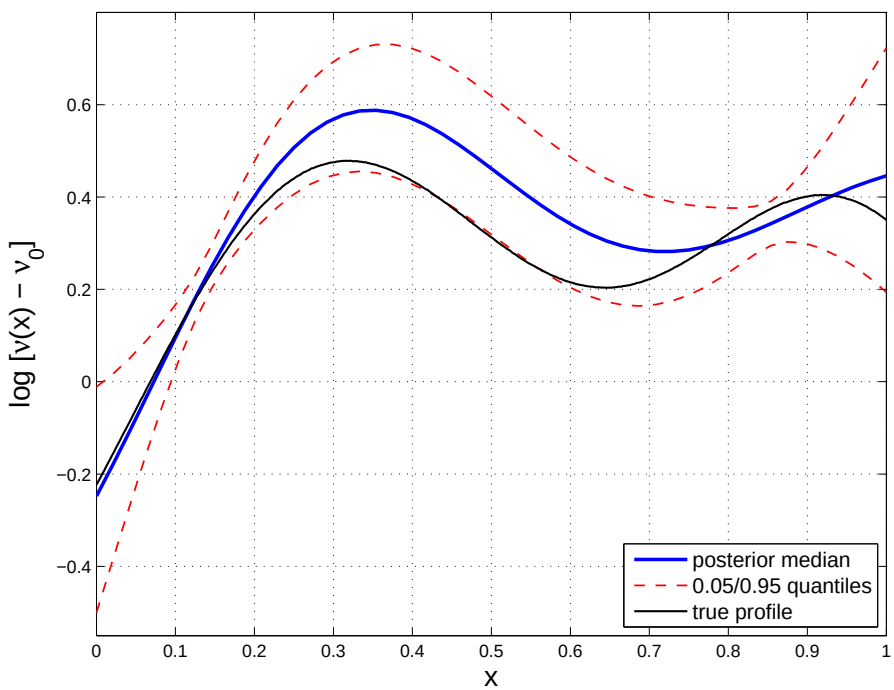


(b) Median, 1-D credibility intervals, and true profile.

Figure 40. Inversion of the random-draw log-diffusivity profile using K-L and polynomial chaos, $K = 6$, $p = 4$.



(a) Mean, standard deviation, and five posterior realizations.



(b) Median, 1-D credibility intervals, and true profile.

Figure 41. Inversion of the random-draw log-diffusivity profile using K-L and polynomial chaos, $K = 6$, $p = 6$.

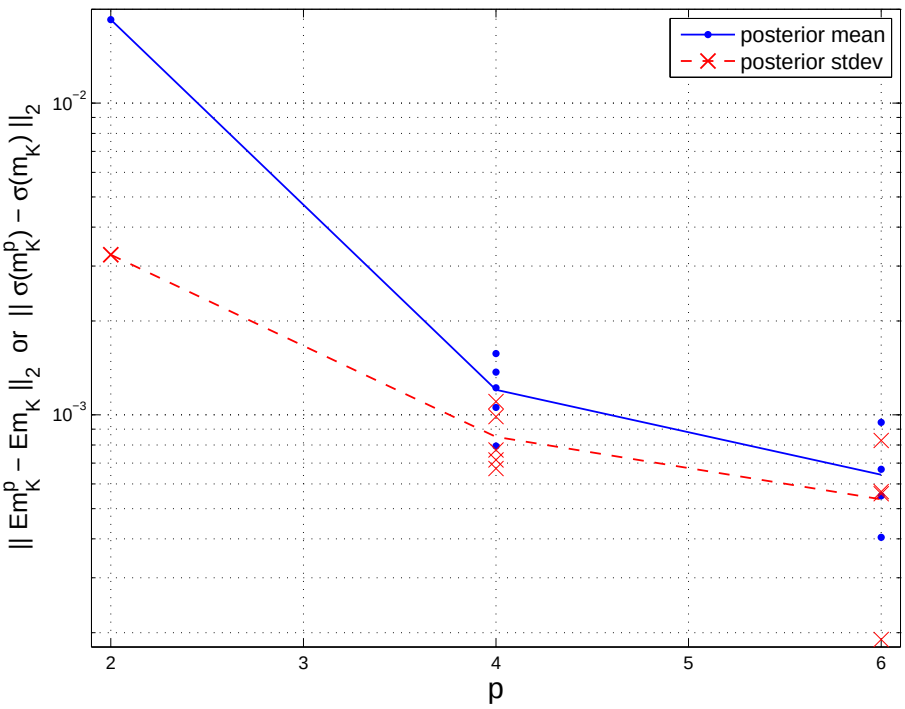
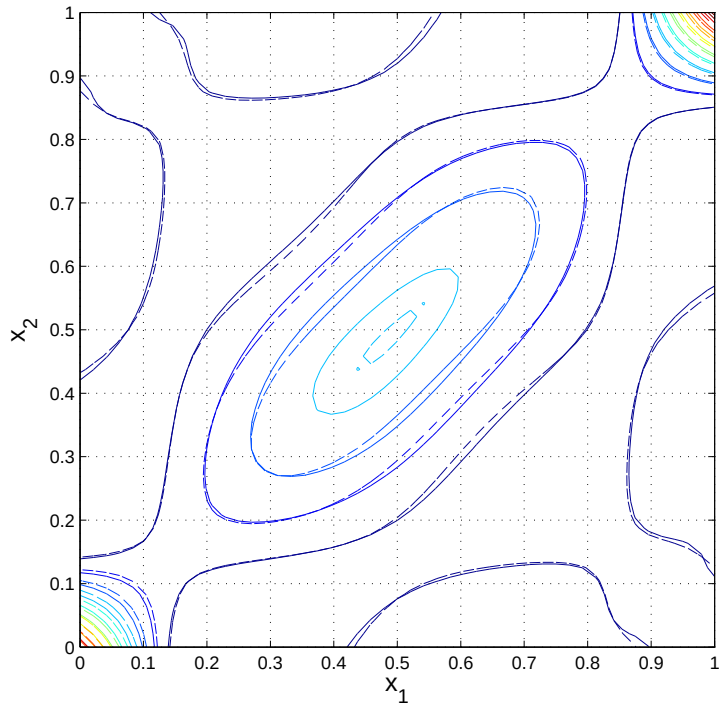
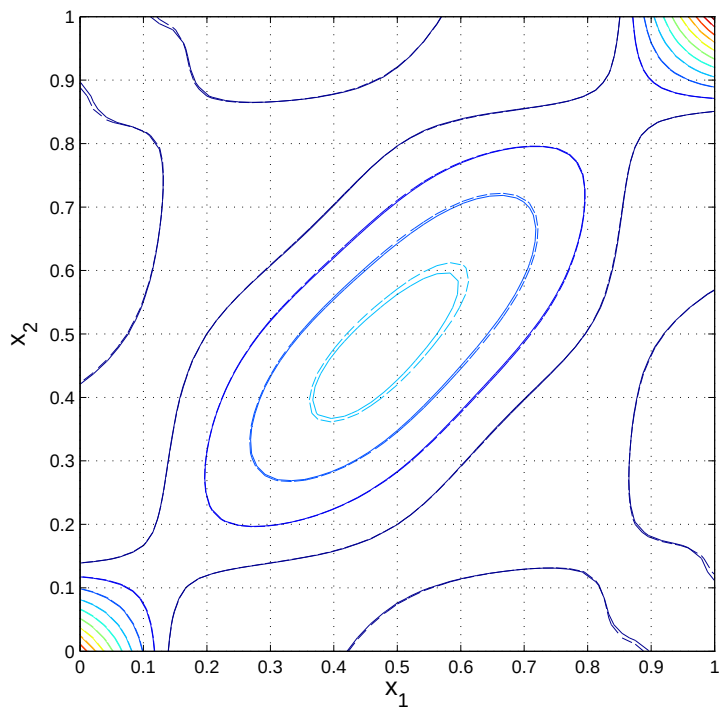


Figure 42. Differences between means/standard deviations of $M(x)$ computed via the "direct" posterior (83) and via the surrogate posterior (91), versus p. All results are for inversion of the random-draw target with 6 K-L modes. At $p = 4$ and $p = 6$, repeated symbols correspond to multiple MCMC simulations from the surrogate posterior.

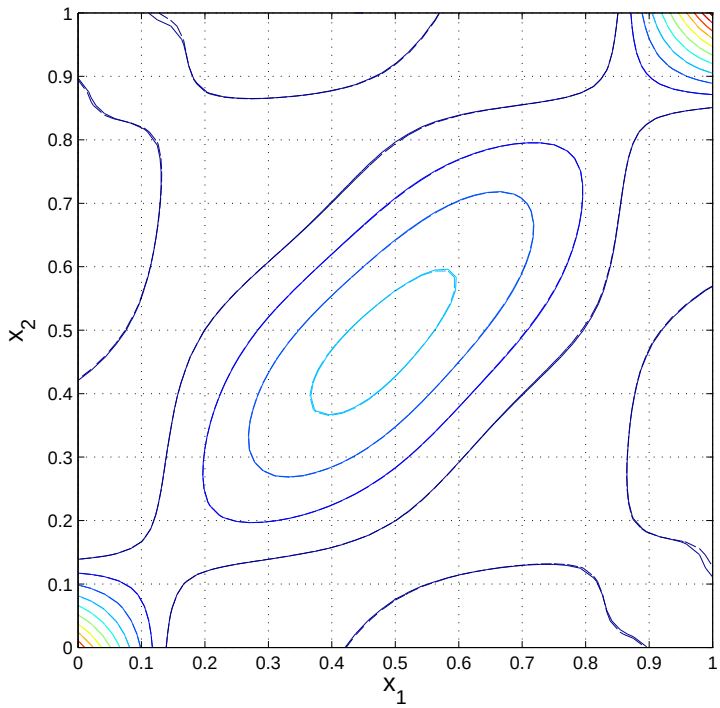


(a) $p = 2$



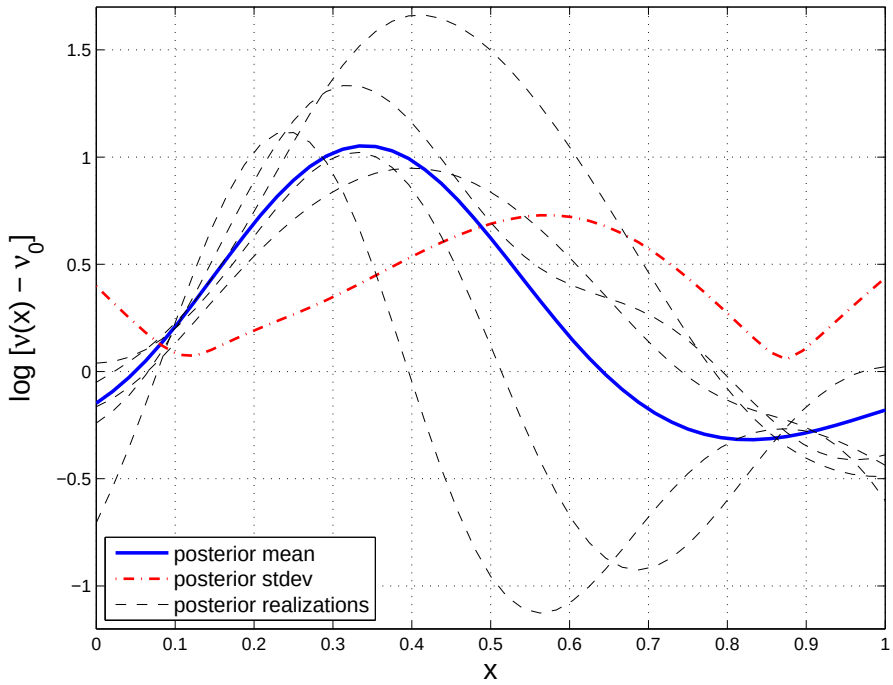
(b) $p = 4$

Figure 43. Contours of the posterior covariance, $\mathrm{Cov}\,[M(x_1), M(x_2)]$. Solid lines are obtained via direct forward problem solutions; dashed lines are obtained via evaluation of the surrogate posterior. All are for inference of the random-draw target with 6 K-L modes.

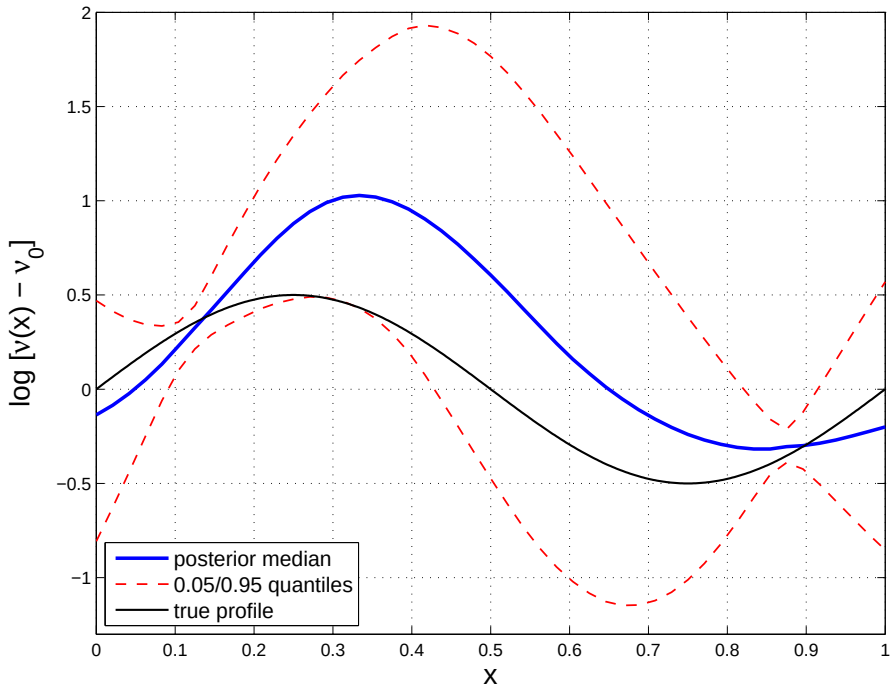


(c) $p = 6$

Figure 43. (cont.) Contours of the posterior covariance, $\mathrm{Cov}\,[M(x_1), M(x_2)]$. Solid lines are obtained via direct forward problem solutions; dashed lines are obtained via evaluation of the surrogate posterior.

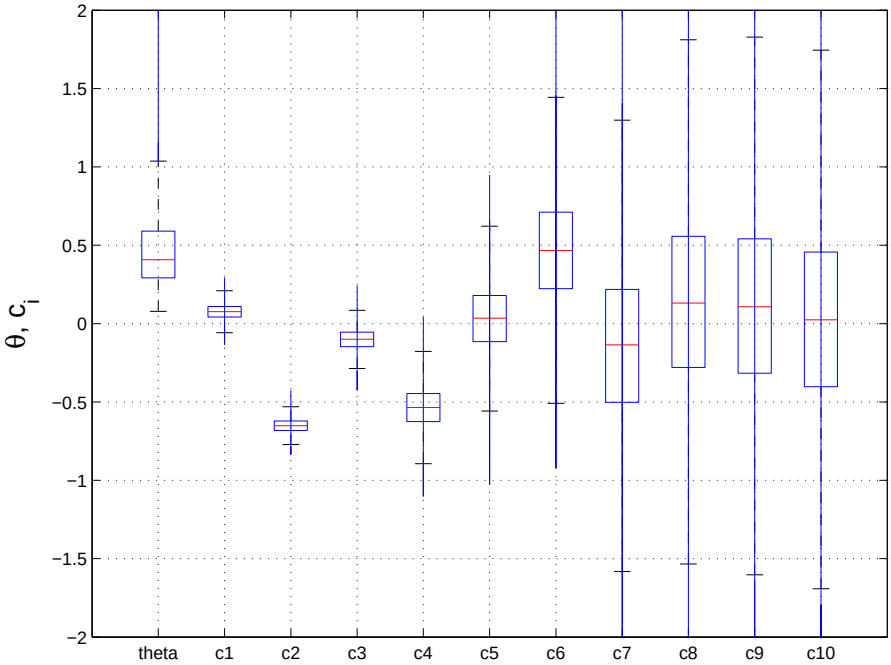


(a) Mean, standard deviation, and five posterior realizations.

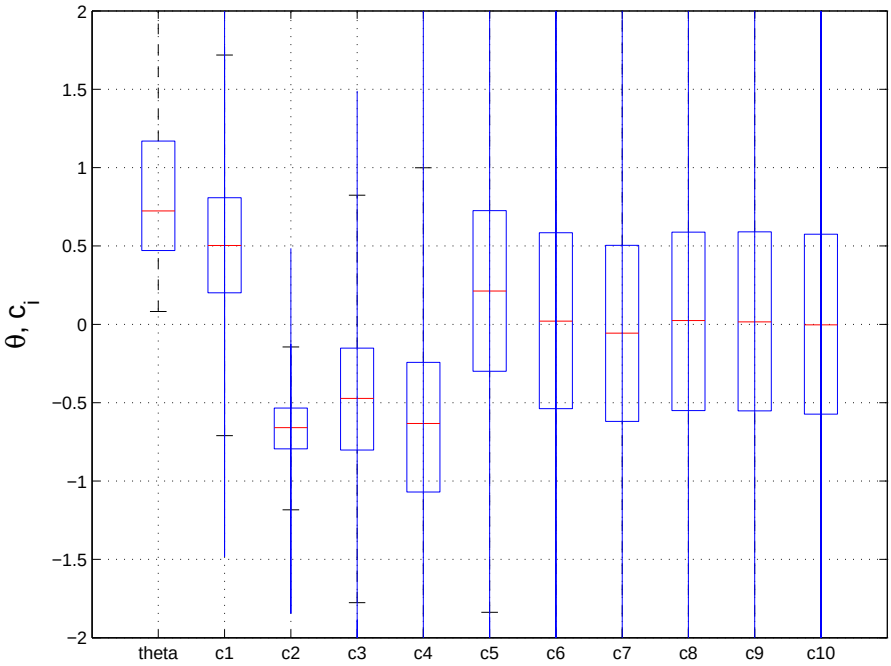


(b) Median, 1-D credibility intervals, and true profile.

Figure 44. Inversion of the sinusoidal log-diffusivity profile with only 2 sensors (at $x = 0.0$ and $x = 1.0$).



(a) 13 sensors



(b) 2 sensors

Figure 45. Boxplot of the posterior marginals of the K-L mode strengths c_i, sinusoidal target.

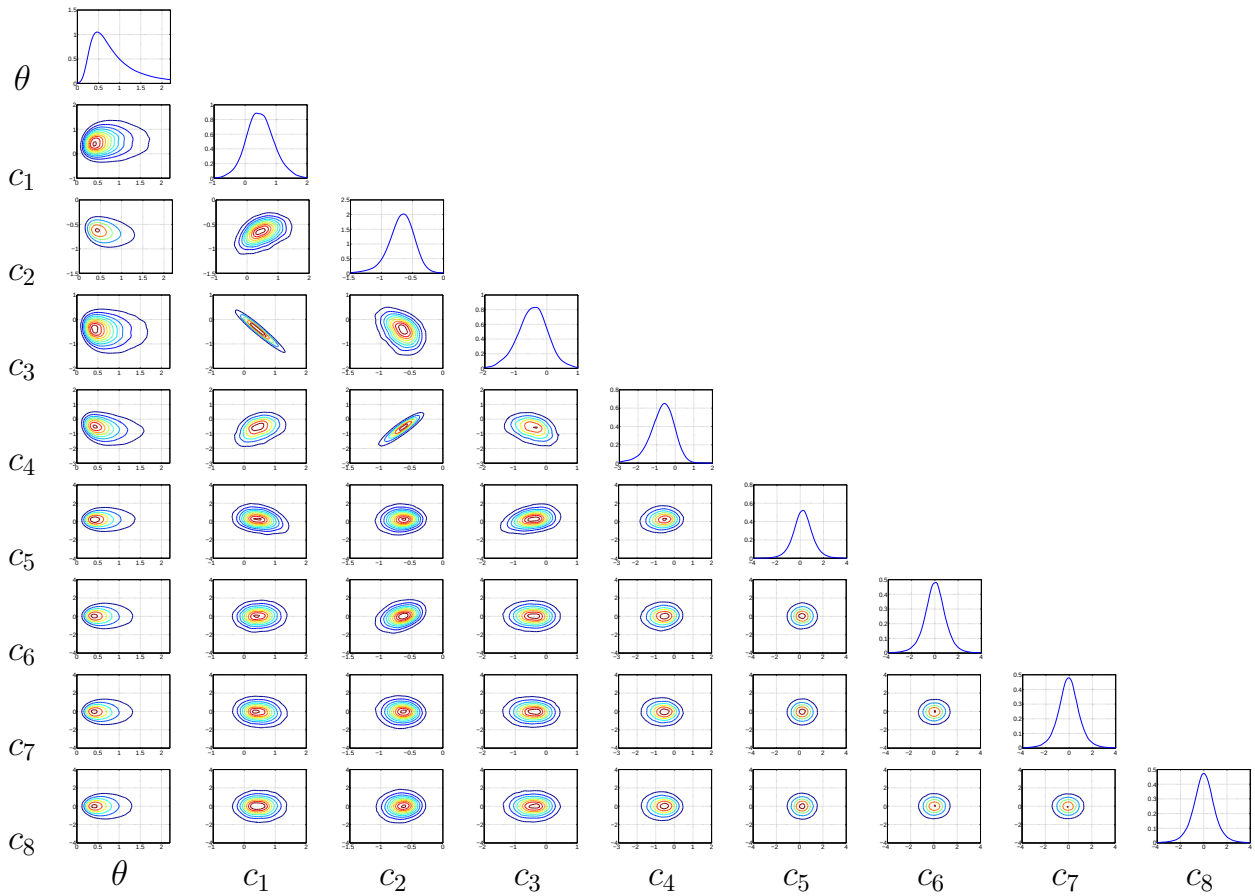
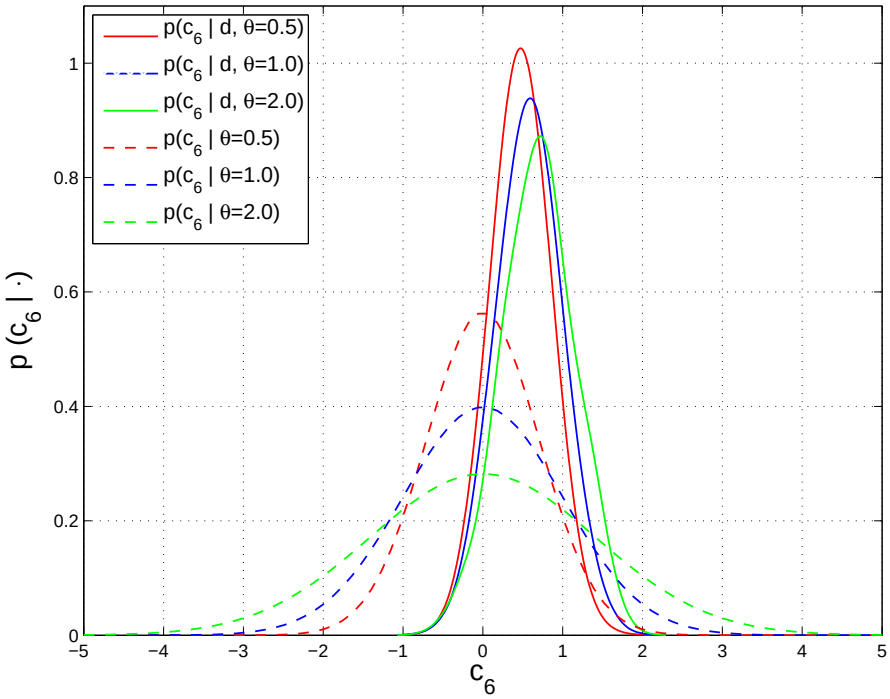
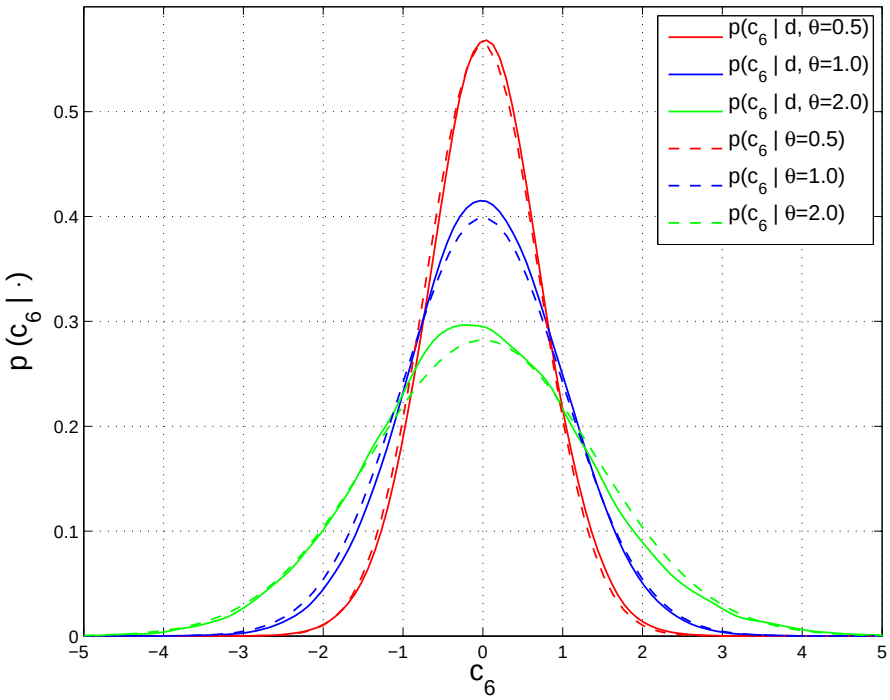


Figure 46. 1-D and 2-D posterior marginals of the K-L mode strengths c_i and the hyperparameter θ; sinusoidal target profile with 2 sensors.



(a) 13 sensors



(b) 2 sensors

Figure 47. Prior and posterior conditional distributions of K-L mode c_6; inference of the sinusoidal target.

3 MCMC schemes for parameter estimation in stochastic models

Stochastic models describe many natural phenomena and engineered systems. Often one would like to estimate parameters of these models from a few realizations of the model output, perhaps observed incompletely and with noise.

A simple example is a polynomial chaos expansion $X = \sum_k x_k \Psi_k(\boldsymbol{\xi})$, whose coefficients x_k are to be estimated from realizations of the random variable X. A more complex example is a stochastic reaction network, where chemical interactions between limited populations of molecules result in intrinsic variability. Species populations comprise a continuous-time Markov process on a discrete state space, exactly simulated with Gillespie's stochastic simulation algorithm. A subset of the interacting chemical species may be observed at a finite set of times and with noise; from these observations, we seek estimates of the reaction propensities and associated uncertainties.

Bayesian statistics, in principle, provides a means of parameter estimation in these models. Exploring the posterior distribution of the model parameters, conditioned on the data, requires repeated evaluations of the likelihood function. Yet in many stochastic models, the likelihood is not available analytically. One may obtain samples of the model outputs for a fixed value of the parameters $\mathbf{m}$, but evaluating the probability density $p(\mathbf{d}|\mathbf{m})$, e.g., via kernel density estimation [44], requires a large number of model realizations $\mathbf{d}^{(i)}$. At each candidate value of $\mathbf{m}$, a new set of forward realizations must be computed. In the context of MCMC, this approach nests a density estimation problem within each calculation of the transition probability, resulting in a scheme that may be feasible only for simple stochastic models that are inexpensive to sample [41]. This chapter explores some more elegant and efficient alternatives.

3.1 Forms of the posterior density

Tarantola [89] (see in particular §1.5) formulates very general expressions for the likelihood function in inverse problems, distinguishing the contribution of observational uncertainty from that of uncertainty or inherent variability in model predictions. We review his exposition here. For simplicity, let the model parameters $\mathbf{m}$ and the outputs or observations $\mathbf{d}$ be real-valued and finite-dimensional: $\mathbf{m} \in \mathcal{M} = \mathbb{R}^m$ and $\mathbf{d}, \mathbf{d}_{\text{obs}} \in \mathcal{D} = \mathbb{R}^d$. With this assumption, we will avoid some of the idiosyncratic language in [89] (e.g., "conjunctions" and "disjunctions" of probabilities) and place the present formulation in a more traditional Bayesian setting.

First, introduce some notation. Let $p_M(\mathbf{m})$ denote the prior density of $\mathbf{m}$, encapsulating any information on the model parameters that is obtained independently of the present measurements. Let the density $p_D(\mathbf{d})$ represent information about the model outputs $\mathbf{d}$ obtained through the measurement process. That is, we don't just have data points; any

measurement yields a "state of information" about what the corresponding model output might be, and this information is encapsulated in p_D. A few examples will make this clear. Suppose that $\mathbf{d}$ is observed through an instrument that introduces additive errors $\boldsymbol{\epsilon}$

$$\mathbf{d}_{\text{obs}} = \mathbf{d} + \boldsymbol{\epsilon} \tag{92}$$

and these errors are independent of the actual value of $\mathbf{d}$. If $\boldsymbol{\epsilon}$ has a density $p_{\boldsymbol{\epsilon}}$, then

$$p_D(\mathbf{d}) = p_{\boldsymbol{\epsilon}}(\mathbf{d}_{\text{obs}} - \mathbf{d}) \tag{93}$$

In particular, if $\boldsymbol{\epsilon}$ is normally distributed, $\varepsilon \sim N(0, \mathbf{C})$, then p_D will give $\mathbf{d} \sim N(\mathbf{d}_{\text{obs}}, \mathbf{C})$. Alternatively, consider the case of perfect measurements: observation of $\mathbf{d}$ through a perfect instrument always yields the exact value of the model output. Then

$$p_D(\mathbf{d}) = \delta(\mathbf{d} - \mathbf{d}_{\text{obs}}) \tag{94}$$

In general, the statistics of the measurement process may be expressed with a conditional density $p(\mathbf{d}_{\text{obs}}|\mathbf{d})$.

Tarantola suggests that forward models, whether deterministic or stochastic, can be expressed in a very general way—as a joint probability density $\varphi(\mathbf{d}, \mathbf{m})$. This is perhaps too general; the physical law or model itself may not say much about the marginal distribution of $\mathbf{m}$. So we can rewrite the forward model as $\varphi(\mathbf{d}, \mathbf{m}) = \varphi(\mathbf{d}|\mathbf{m})\mu_M(\mathbf{m})$ where μ_M is the homogeneous probability density on $\mathbb{R}^m$—i.e., a constant. The forward model thus is a conditional density; for any value of the model parameters $\mathbf{m}$, we predict some density for $\mathbf{d}$. Deterministic forward models $\mathbf{d} = \mathbf{G}(\mathbf{m})$ are a special case, with $\varphi(\mathbf{d}|\mathbf{m}) = \delta(\mathbf{d} - \mathbf{G}(\mathbf{m}))$.

Now we combine (i) prior information on $\mathbf{m}$, (ii) observational information on $\mathbf{d}$, and (iii) the probabilistic model relating $\mathbf{m}$ to $\mathbf{d}$, to obtain *a posteriori* information on $\mathbf{m}$. Tarantola expresses this information as a joint density on the model parameters and outputs:**

$$\pi(\mathbf{m}, \mathbf{d}) \propto \varphi(\mathbf{d}|\mathbf{m}) p_D(\mathbf{d}) p_M(\mathbf{m}) \tag{95}$$

We then marginalize over $\mathbf{d}$ to obtain the posterior density of $\mathbf{m}$ alone:

$$\begin{aligned} \pi(\mathbf{m}) &\propto p_M(\mathbf{m}) \int_{\mathcal{D}} \varphi(\mathbf{d}|\mathbf{m}) p_D(\mathbf{d}) \mathrm{d}\mathbf{d} && (96) \\ &\propto p_M(\mathbf{m}) L(\mathbf{m}) && (97) \end{aligned}$$

where $L(\mathbf{m})$ is the likelihood function. Consider a few special cases. If the forward model is deterministic, we have $\pi(\mathbf{m}) \propto p_D(\mathbf{G}(\mathbf{m})) p_M(\mathbf{m})$, which is the usual expression used in Bayesian approaches to inverse problems [72]. If the forward model remains stochastic but the measurements are perfect, we have $\pi(\mathbf{m}) \propto \varphi(\mathbf{d}_{\text{obs}}|\mathbf{m}) p_M(\mathbf{m})$, which is, for example, the posterior expression used to estimate parameters of a disease model in [79] and to identify polynomial chaos coefficients in [41].

**Compared to equation (1.89) in [89], we have assumed the homogeneous probability density on the data space, $\mu_D(\mathbf{d})$, to be constant since $\mathcal{D} = \mathbb{R}^d$.

Note that Tarantola's expression (96) may also be derived in a more typically Bayesian way. First, recognize that $\pi(\mathbf{m})$ in (96) is really $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$; similarly, $p_D(\mathbf{d})$ is really $p(\mathbf{d}|\mathbf{d}_{\text{obs}})$. Then

$$
\begin{aligned}
p(\mathbf{m}|\mathbf{d}_{\text{obs}}) &= \int p(\mathbf{m},\mathbf{d}|\mathbf{d}_{\text{obs}})\mathrm{d}\mathbf{d} && (98)\\
&= \int p(\mathbf{m}|\mathbf{d})p(\mathbf{d}|\mathbf{d}_{\text{obs}})\mathrm{d}\mathbf{d} && (99)\\
&\propto \int \varphi(\mathbf{d}|\mathbf{m})p(\mathbf{m})p(\mathbf{d}|\mathbf{d}_{\text{obs}})\mathrm{d}\mathbf{d} && (100)\\
&\propto p_M(\mathbf{m})\int \varphi(\mathbf{d}|\mathbf{m})p_D(\mathbf{d})\mathrm{d}\mathbf{d} && (101)
\end{aligned}
$$

where in (100) we have applied Bayes' rule and in (101) we have just removed the prior from the integral and reverted to the original notation.

3.2 Markov chain Monte Carlo

The parameter inference problem now reduces to a computational question: how to efficiently sample from (96) when the forward model $\varphi(\mathbf{d}|\mathbf{m})$ cannot be expressed analytically?

Consider a Markov chain on the product space $\mathcal{M}\times\mathcal{D}$. We would like to construct a chain whose stationary distribution is given by $\pi(\mathbf{m},\mathbf{d})\propto p_M(\mathbf{m})\varphi(\mathbf{d}|\mathbf{m})p_D(\mathbf{d})$. A marginal distribution of this chain will be $\pi(\mathbf{m}) = p(\mathbf{m}|\mathbf{d}_{\text{obs}})$. We employ a standard Metropolis-Hastings construction. (See [42] for an excellent review of MCMC methods.) A Metropolis-Hastings algorithm involves the specification of a proposal distribution $q(\mathbf{m}',\mathbf{d}'|\mathbf{m},\mathbf{d})$. The proposed "move" from $(\mathbf{m},\mathbf{d})$ to $(\mathbf{m}',\mathbf{d}')$ is then accepted with probability

$$
\begin{aligned}
\alpha &= \min\left(1,\frac{\pi(\mathbf{m}',\mathbf{d}')q(\mathbf{m},\mathbf{d}|\mathbf{m}',\mathbf{d}')}{\pi(\mathbf{m},\mathbf{d})q(\mathbf{m}',\mathbf{d}'|\mathbf{m},\mathbf{d})}\right)\\
&= \min\left(1,\frac{\varphi(\mathbf{d}'|\mathbf{m}')p_D(\mathbf{d}')p_M(\mathbf{m}')q(\mathbf{m},\mathbf{d}|\mathbf{m}',\mathbf{d}')}{\varphi(\mathbf{d}|\mathbf{m})p_D(\mathbf{d})p_M(\mathbf{m})q(\mathbf{m}',\mathbf{d}'|\mathbf{m},\mathbf{d})}\right). \qquad (102)
\end{aligned}
$$

Let the proposal distribution q be specified as follows:

$$
\begin{aligned}
q(\mathbf{m}',\mathbf{d}'|\mathbf{m},\mathbf{d}) &= q_1(\mathbf{d}'|\mathbf{m}',\mathbf{m},\mathbf{d})q_2(\mathbf{m}'|\mathbf{m},\mathbf{d})\\
&= q_D(\mathbf{d}'|\mathbf{m}')q_M(\mathbf{m}'|\mathbf{m})\\
&= \varphi(\mathbf{d}'|\mathbf{m}')q_M(\mathbf{m}'|\mathbf{m}) \qquad (103)
\end{aligned}
$$

In other words, we have factored the joint proposal into a conditional q_1 and a marginal q_2, then chosen particular forms for these factors. The proposal for the observable, q_D,

is simply the stochastic forward model. The proposal for the model parameters, q_M, can be chosen quite flexibly. For instance, one could choose a random walk proposal: $\mathbf{m}' \sim N(\mathbf{m}, \sigma)$. Alternately, choosing $q_M(\mathbf{m}'|\mathbf{m}) = q_M(\mathbf{m}')$ yields the independence sampler. Assume that we are at least using a symmetric proposal: $q_M(\mathbf{m}'|\mathbf{m}) = q_M(\mathbf{m}|\mathbf{m}')$. (This assumption is not particularly restrictive, and is easily removed from the expression below.) Plugging the proposal distribution (103) into (102) gives the acceptance probability:

$$\alpha = \min\left(1, \frac{p_D(\mathbf{d}')p_M(\mathbf{m}')}{p_D(\mathbf{d})p_M(\mathbf{m})}\right) \tag{104}$$

The Metropolis-Hastings scheme based on proposal distribution (103) and acceptance probability (104) yields a correlated sequence of samples from $\pi(\mathbf{m}, \mathbf{d})$ (95). The density of the forward model $\varphi(\mathbf{d}|\mathbf{m})$ is never explicitly computed; we simply need to draw a sample from this distribution at each step of the MCMC chain. The steps comprising the scheme are summarized as follows:

1. If at $(\mathbf{m}^{(t)}, \mathbf{d}^{(t)})$, propose a move to $\mathbf{m}'$ according to q_M.
2. Sample the observable $\mathbf{d}'$ from the stochastic forward model $\varphi(\mathbf{d}'|\mathbf{m}')$.
3. Calculate $\alpha(\mathbf{m}', \mathbf{d}', \mathbf{m}^{(t)}, \mathbf{d}^{(t)})$ according to (104).
4. Accept $(\mathbf{m}^{(t+1)}, \mathbf{d}^{(t+1)}) = (\mathbf{m}', \mathbf{d}')$ with probability α; otherwise put $(\mathbf{m}^{(t+1)}, \mathbf{d}^{(t+1)}) = (\mathbf{m}^{(t)}, \mathbf{d}^{(t)})$. Return to step #1.

3.3 Discussion and comparison to other schemes

A similar MCMC scheme, relying only on the ability to draw samples from the stochastic model, was proposed by Marjoram *et al.* [71]. The algorithms discussed therein (and the associated posterior distributions) do not account for measurement noise or uncertainty; in effect, they assume that $p_D(\mathbf{d}) = \delta(\mathbf{d} - \mathbf{d}_{\text{obs}})$. These MCMC schemes thus define Markov chains only on $\mathcal{M}$, not on $\mathcal{M} \times \mathcal{D}$. However, [71] provides a simple proof that these methods satisfy the detailed balance condition, and this proof appears adaptable to the MCMC scheme given here.

Even though the posterior $\pi(\mathbf{m}, \mathbf{d})$ (95) may be the stationary and limiting distribution of the Markov chain defined by our scheme, important practical considerations remain. In particular, the chain should *mix rapidly* within the support of the posterior, and thus the acceptance probabilities α should not be too low. (Nor should they be too high; see [42]!) This issue is discussed by Marjoram *et al.* [71]. In the basic MCMC method of that paper, moves are accepted only if $\mathbf{d}' = \mathbf{d}_{\text{obs}}$, and in a high-dimensional or continuous state space this will occur very rarely. One solution then proposed is to accept any $\mathbf{d}'$ for which $\rho(\mathbf{d}', \mathbf{d}_{\text{obs}}) < \varepsilon$, where ρ is some suitably defined metric and ε is a small constant; this effectively samples the posterior $p(\mathbf{m} \,|\, \rho(\mathbf{d}, \mathbf{d}_{\text{obs}}) < \varepsilon)$. A further alternative proposed

in [71] is to condition not on the data, but on summary statistics of the data that are *close* to being a sufficient statistic for $\mathbf{m}$. (Recall that a sufficient statistic $\mathcal{S}$ for $\mathbf{m}$ is one for which $p(\mathbf{d}|\mathbf{m},\mathcal{S}) = p(\mathbf{d}|\mathcal{S})$. In practice $\mathcal{S}$ may be difficult to identify.) Such methods fall under the headline of "approximate Bayesian computation."

In the present MCMC scheme, the introduction of p_D and a Markov process on the data space ensures that our acceptance probability is not so stringent; in effect, we may pit the variability of the model against the precision of the measurements. So it is unclear whether low acceptance probabilities will be an issue here, and if so, under what circumstances. Numerical explorations will be useful.

3.4 An alternative robust scheme

If acceptance rates are too small under the MCMC scheme of §3.2, a more robust—though less efficient—alternative may be useful. As before, we would like to explore the posterior $p(\mathbf{m},\mathbf{d}|\mathbf{d}_{\text{obs}})$. With a generic proposal distribution $q(\mathbf{m}',\mathbf{d}'|\mathbf{m},\mathbf{d})$, the resulting acceptance probability is given in (102). Now let

$$q(\mathbf{m}',\mathbf{d}'|\mathbf{m},\mathbf{d}) = q_M(\mathbf{m}'|\mathbf{m})p_D(\mathbf{d}') = q_M(\mathbf{m}'|\mathbf{m})p(\mathbf{d}'|\mathbf{d}_{\text{obs}}) \tag{105}$$

In other words, we propose values for the model outputs from the observational density p_D, independent of the current state of $\mathbf{d}$. Assuming a symmetric $q_M(\cdot|\cdot)$, the acceptance probability then reduces to:

$$\alpha = \min\left(1, \frac{\varphi(\mathbf{d}'|\mathbf{m}')p_M(\mathbf{m}')}{\varphi(\mathbf{d}|\mathbf{m})p_M(\mathbf{m})}\right) \tag{106}$$

The "likelihood" term $\varphi(\mathbf{d}'|\mathbf{m}')$ can be evaluated via kernel density estimation [44, 48], which entails simulating from the forward model to obtain a large set of realizations for any proposed $\mathbf{m}'$. As noted in the introduction, this approach is most suitable for stochastic forward models that are inexpensive to simulate. The Metropolis-Hasting scheme based on (105) and (106) should be understood as an extension of [41] to include observational errors.

This MCMC scheme could also modified in order to employ single-component or block updates of $\mathbf{m}$, for better mixing. Many variations are possible, one of which is as follows:

1. At the start of MCMC iteration t, propose a new set of model outputs $\mathbf{d}'$ from $p_D(\cdot)$
2. Accept this move with probability $\alpha = \min\left(1, \frac{\varphi(\mathbf{d}'|\mathbf{m}^{(t)})}{\varphi(\mathbf{d}^{(t)}|\mathbf{m}^{(t)})}\right)$. Upon acceptance or rejection, put $\mathbf{d}^{(t+1)} = \mathbf{d}'$ or $\mathbf{d}^{(t+1)} = \mathbf{d}^{(t)}$.
3. Beginning with $i = 1$, propose a move for component or block m_i of $\mathbf{m}$, from a symmetric proposal $q_M^i(m_i'|m_i^{(t)})$.

4. Accept or reject this move with probability

$$\alpha = \min\left(1, \frac{\varphi(\mathbf{d}^{(t+1)}|m_i', \mathbf{m}_{-i})\ p_M(m_i', \mathbf{m}_{-i})}{\varphi(\mathbf{d}^{(t+1)}|m_i^{(t)}, \mathbf{m}_{-i})\ p_M(m_i^{(t)}, \mathbf{m}_{-i})}\right) \tag{107}$$

   where $\mathbf{m}_{-i} \equiv (m_1^{(t+1)}, \ldots, m_{i-1}^{(t+1)}, m_{i+1}^{(t)}, \ldots)$. Since we are really interested in the ratio of full conditionals of m_i, the above expression for the acceptance probability could be simplified, depending on the form of the prior and the forward model. Put $m_i^{(t+1)}$ equal to m_i' or $m_i^{(t)}$, upon acceptance or rejection.

5. Repeat the previous two steps for successive i until every component of $\mathbf{m}$ has been updated.

6. Return to step #1.

In closing, we note several further ideas that may improve the state of the art beyond the parameter estimation schemes considered here:

- How to construct approximations of $\varphi(\mathbf{d}|\mathbf{m})$, or of ratios such as $\varphi(\mathbf{d}'|\mathbf{m})/\varphi(\mathbf{d}|\mathbf{m})$ or $\varphi(\mathbf{d}|\mathbf{m}')/\varphi(\mathbf{d}|\mathbf{m})$? In particular, can we seek approximations based on the actual details/physics/structure of the forward model?

- How best to use these approximations, for instance, in two-stage MCMC schemes [81] that still occasionally appeal to the exact forward model?

- What about single-component MCMC schemes that use only small portions of the data at each iteration, thus rendering acceptances based on p_D less restrictive?

- In a similar vein, could greater efficiency be obtained via sequential Monte Carlo schemes (e.g., Bayesian filtering) [87, 22], that recursively condition on increasing amounts of data?

References

[1] P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center, Oslo, Norway, 1997.

[2] R. J. Adler and J. E. Taylor. *Random Fields and Geometry*. Springer, 2007.

[3] V. Akcelik, G. Biros, O. Ghattas, K. Long, and B. van Bloemen Waanders. A variational finite element method for source inversion for convective-diffusive transport. *Finite Elements in Analysis and Design*, 39:683–705, 2003.

[4] O. M. Alifanov. *Inverse Heat Transfer Problems*. Springer, 1994.

[5] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

[6] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[7] R. Aster, B. Borchers, and C. Thurber. *Parameter Estimation and Inverse Problems*. Academic Press, 2004.

[8] S. Balakrishnan, A. Roy, M. G. Ierapetritou, G. P. Flach, and P. G. Georgopoulos. Uncertainty reduction and characterization for complex environmental fate and transport models: an empirical Bayesian framework incorporating the stochastic response surface method. *Water Resources Research*, 39(12):1350, 2003.

[9] J. V. Beck, B. Blackwell, and C. R. S. Clair. *Inverse Heat Conduction: Ill-Posed Problems*. Wiley, 1985.

[10] G. Berkooz, P. Holmes, and J. L. Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25:539–575, 1993.

[11] J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–41, 1995.

[12] D. Calvetti and E. Somersalo. A Gaussian hypermodel to recover blocky objects. *Inverse Problems*, 23:733–754, 2007.

[13] R. Cameron and W. Martin. The orthogonal development of nonlinear functionals in series of Fourier-Hermite functionals. *Annals of Mathematics*, 48:385–392, 1947.

[14] A. Chorin. Hermite expansions in Monte Carlo computation. *Journal of Computational Physics*, 8:472–482, 1971.

[15] A. Chorin. Gaussian fields and random flow. *J. Fluid Mech.*, 63:21–32, 1974.

[16] J. A. Christen and C. Fox. MCMC using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810, 2005.

[17] R. Cools and P. Dellaportas. The role of embedded integration rules in Bayesian statistics. *Statistics and Computing*, 6:245–250, 1996.

[18] R. Courant and D. Hilbert. *Methods of Mathematical Physics*, volume 1. Wiley-Interscience, 1953.

[19] N. A. C. Cressie. *Statistics for Spatial Data*. Wiley, 1993.

[20] B. Debusschere, H. Najm, A. Matta, O. Knio, R. Ghanem, and O. Le Maître. Protein labeling reactions in electrochemical microchannel flow: Numerical simulation and uncertainty propagation. *Physics of Fluids*, 15(8):2238–2250, 2003.

[21] B. Debusschere, H. Najm, P. Pébay, O. Knio, R. Ghanem, and O. Le Maître. Numerical challenges in the use of polynomial chaos representations for stochastic processes. *SIAM Journal of Scientific Computing*, 26(2):698–719, 2004.

[22] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo methods in Practice*. Springer, 2001.

[23] J. Dunkley, M. Bucher, P. G. Ferreira, and K. Moodley. Fast and reliable Markov chain Monte Carlo technique for cosmological parameter estimation. *Monthly Notices of the Royal Astronomical Society*, 356(3):925–936, 2005.

[24] Y. Efendiev, T. Y. Hou, and W. Luo. Preconditioning Markov chain Monte Carlo simulations using coarse-scale models. *SIAM Journal on Scientific Computing*, 28:776–803, 2006.

[25] M. Evans and T. Swartz. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, 10(3):254–272, 1995.

[26] S. N. Evans and P. B. Stark. Inverse problems as statistics. *Inverse Problems*, 18:R55–R97, 2002.

[27] G. S. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer, 1996.

[28] P. Frauenfelder, C. Schwab, and R. A. Todor. Finite elements for elliptic problems with stochastic coefficients. *Computer Methods in Applied Mechanics and Engineering*, 194:205–228, 2005.

[29] D. Gamerman. *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Chapman & Hall, 1997.

[30] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[31] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534, 2006.

[32] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall CRC, 2 edition, 2003.

[33] M. G. Genton. Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2001.

[34] C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In E. M. Keramidas, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, volume 23, pages 156–163. Interface Foundation of North America, 1991.

[35] C. J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.

[36] R. Ghanem. Probabilistic characterization of transport in heterogeneous media. *Comput. Methods Appl. Mech. Engrg.*, 158:199–220, 1998.

[37] R. Ghanem. Ingredients for a general purpose stochastic finite element formulation. *Comput. Methods Appl. Mech. Engrg.*, 168:19–34, 1999.

[38] R. Ghanem. Stochastic finite elements for heterogeneous media with multiple random non-gaussian properties. *ASCE J. Eng. Mech.*, 125:26–40, 1999.

[39] R. Ghanem, J. Red-Horse, and A. Sarkar. Modal properties of a space-frame with localized system uncertainties. In *8th ASCE Specialty Conference of Probabilistic Mechanics and Structural Reliability*, number PMC200-269. ASCE, 2000.

[40] R. Ghanem and P. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer Verlag, New York, 1991.

[41] R. G. Ghanem and A. Doostan. On the construction and analysis of stochastic models: Characterization and propagation of the errors associated with limited data. *Journal of Computational Physics*, 217:63–81, 2006.

[42] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[43] W. P. Gouveia and J. A. Scales. Resolution of seismic waveform inversion: Bayes versus Occam. *Inverse Problems*, 13:323–349, 1997.

[44] A. G. Gray and A. W. Moore. Nonparametric density estimation: Toward computational tractability. In *SIAM International Conference on Data Mining*. SIAM, 2003.

[45] M. Grigoriu. *Stochastic Calculus*. Birkhäuser, 2002.

[46] D. Higdon, H. Lee, and C. Holloman. Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems. *Bayesian Statistics*, 7:181–197, 2003.

[47] P. G. Hoel, S. C. Port, and C. J. Stone. *Introduction to stochastic processes*. Waveland Press, 1987.

[48] A. T. Ihler. Kernel density estimation toolbox for MATLAB. http://ttic.uchicago.edu/ ihler/code/kde.php.

[49] C. Jackson, M. K. Sen, and P. L. Stoffa. An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions. *Journal of Climate*, 17:2828–2841, 2004.

[50] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Springer, 2005.

[51] J. Kaipio and E. Somersalo. Statistical inverse problems: Discretization, model reduction, and inverse crimes. *Journal of Computational and Applied Mathematics*, 198(2):493–504, 2007.

[52] J. P. Kaipio, V. Kolehmainen, M. Vauhkonen, and E. Somersalo. Inverse problems with structural prior information. *Inverse Problems*, 15:713–729, 1999.

[53] S. Kakutani. Spectral analysis of stationary Gaussian processes. In J. Neyman, editor, *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 239–247. University of California Press, 1961.

[54] V. A. Karkhin, V. V. Plochikhine, and H. W. Bergmann. Solution of inverse heat conduction problem for determining heat input, weld shape, and grain structure during laser welding. *Science and Technology of Welding and Joining*, 7(4):224–231, 2002.

[55] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, 63(3):425–464, 2001.

[56] G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

[57] C. H. Lai, C. S. Ierotheou, C. J. Palansuriya, K. A. Pericleous, M. S. Espedal, and X. C. Tai. Accuracy of a domain decomposition method for the recovering of discontinuous heat sources in metal sheet cutting. *Computing and Visualization in Science*, 2:149–152, 1999.

[58] O. Le Maître, R. Ghanem, O. Knio, and H. Najm. Uncertainty propagation using Wiener-Haar expansions. *Journal of Computational Physics*, 197(1):28–57, 2004.

[59] O. Le Maître, O. Knio, H. Najm, and R. Ghanem. A stochastic projection method for fluid flow I. Basic formulation. *Journal of Computational Physics*, 173:481–511, 2001.

[60] O. Le Maître, H. Najm, R. Ghanem, and O. Knio. Multi-resolution analysis of Wiener-type uncertainty propagation schemes. *Journal of Computational Physics*, 197:502–531, 2004.

[61] O. Le Maître, M. Reagan, H. Najm, R. Ghanem, and O. Knio. A stochastic projection method for fluid flow II. Random process. *Journal of Computational Physics*, 181:9–44, 2002.

[62] H. Lee, D. Higdon, Z. Bi, M. Ferreira, and M. West. Markov random field models for high-dimensional parameters in simulations of fluid flow in porous media. *Technometrics*, 44(3):230–241, 2002.

[63] H. K. H. Lee, C. H. Holloman, C. A. Calder, and D. M. Higdon. Flexible Gaussian processes via convolution. Discussion paper 2002-09, Duke University, Department of Statistical Science, 2002.

[64] W. Li and O. A. Cirpka. Efficient geostatistical inverse methods for structured and unstructured gris. *Water Resources Research*, 42:W06402, 2006.

[65] M. Loève. *Probability Theory II*. Springer, 4 edition, 1978.

[66] T. Loredo. Computational technology for Bayesian inference. In D. M. R. Plante and D. Roberts, editors, *ASP Conf. Ser., Astronomical Data Analysis Software and Systems VIII*, volume 172. 1999.

[67] W. Luo. *Wiener chaos expansion and numerical solutions of stochastic partial differential equations*. PhD thesis, California Institute of Technology, May 2006.

[68] D. J. C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11(5):1035–1068, 1999.

[69] A. Malinverno. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, 151:675–688, 2002.

[70] F. Maltz and D. Hitzl. Variance reduction in Monte Carlo computations using multidimensional Hermite polynomials. *Journal of Computational Physics*, 32:345–376, 1979.

[71] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci USA*, 100(26):15324–15328, 2003.

[72] Y. M. Marzouk, H. N. Najm, and L. A. Rahn. Stochastic spectral methods for efficient Bayesian solution of inverse probelms. *Journal of Computational Physics*, 224(2):560–586, 2007.

[73] W. Meecham and D. Jeng. Use of the Wiener-Hermite expansion for nearly normal turbulence. *J. Fluid Mech.*, 32:225–249, 1968.

[74] A. Mohammad-Djafari. Bayesian inference for inverse problems. In *Bayesian inference and Maximum Entropy Methods in Science and Engineering*, volume 21, pages 477–496, 2002.

[75] K. Mosegaard and M. Sambridge. Monte Carlo analysis of inverse problems. *Inverse Problems*, 18:R29–R54, 2002.

[76] D. S. Oliver, L. Cunha, and A. C. Reynolds. Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. *Mathematical Geology*, 29(1):61–91, 1997.

[77] W. Press, S. A. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK, 2 edition, 1992.

[78] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[79] J. Ray, Y. M. Marzouk, H. N. Najm, M. Kraus, and P. Fast. A Bayesian approach for estimating bioterror attacks from patient data. Submitted to Statistics in Medicine, 2007.

[80] M. Reagan, H. Najm, R. Ghanem, and O. Knio. Uncertainty quantification in reacting flow simulations through non-intrusive spectral projection. *Combustion and Flame*, 132:545–555, 2003.

[81] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2 edition, 2004.

[82] G. O. Roberts. Markov chain concepts related to sampling algorithms. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 45–58. Chapman and Hall, 1996.

[83] W. Schoutens. *Stochastic Processes and Orthogonal Polynomials*. Springer, 2000.

[84] M. Seeger. Gaussian processes for machine learning. *International Journal of Neural Systems*, 14(2):1–38, 2004.

[85] U. Shavit and A. Furman. The location of deep salinity sources in the Israeli coastal aquifer. *J. Hydrology*, 250:63–77, 2001.

[86] L. Sirovich. Turbulence and the dynamics of coherent structures. Part 1: Coherent structures. *Quarterly of Applied Mathematics*, 45(3):561–571, 1987.

[87] S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods. *Proc Natl Acad Sci USA*, 104(6):1760–1765, 2007.

[88] B. Sommeijer, L. Shampine, and J. Verwer. RKC: An explicit solver for parabolic PDEs. *J. Comput. Appl. Math.*, 88:315–326, 1997.

[89] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, 2005.

[90] L. Tenorio. Statistical regularization of inverse problems. *SIAM Review*, 43(2):347–366, 2001.

[91] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.

[92] L. Tierney. Introduction to general state-space Markov chain theory. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 59–74. Chapman and Hall, 1996.

[93] J. G. Verwer, B. P. Sommeijer, and W. Hundsdorfer. RKC time-stepping for advection-diffusion-reaction problems. CWI Report MAS-E0405, CWI, Amsterdam, The Netherlands, 2004.

[94] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[95] X. Wan and G. E. Karniadakis. An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. *Journal of Computational Physics*, 209:617–642, 2005.

[96] J. Wang and N. Zabaras. A Bayesian inference approach to the inverse heat conduction problem. *International Journal of Heat and Mass Transfer*, 47:3927–3941, 2004.

[97] J. Wang and N. Zabaras. Hierarchical Bayesian models for inverse problems in heat conduction. *Inverse Problems*, 21:183–206, 2005.

[98] J. Wang and N. Zabaras. Using Bayesian statistics in the estimation of heat source in radiation. *International Journal of Heat and Mass Transfer*, 48:15–29, 2005.

[99] J. Wang and N. Zabaras. A Markov random field model of contamination source identification in porous media flow. *International Journal of Heat and Mass Transfer*, 49:939–950, 2006.

[100] N. Wiener. The homogenous chaos. *American Journal of Mathematics*, 60:897–936, 1938.

[101] C. K. I. Williams and C. E. Rasmussen. Gaussian processes for regression. In D. S. Teuretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1996.

[102] D. Xiu and G. Karniadakis. Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Computer Methods in Applied Mechanics and Engineering*, 191:4927–4948, 2002.

[103] D. Xiu and G. Karniadakis. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal of Scientific Computing*, 24(2):619–644, 2002.

[104] D. Xiu, D. Lucor, C.-H. Su, and G. Karniadakis. Stochastic modeling of flow-structure interactions using generalized polynomial chaos. *ASME J. Fluids Engineering*, 124:51–59, 2002.

Appendix: Reproducing kernel Hilbert space

Here we briefly review properties of reproducing kernel Hilbert spaces (RKHS). Useful connections among RKHS, Gaussian process priors, regularization, and the Karhunen-Loève expansion are discussed in §2.2.6. Much more detailed expositions of RKHS may be found in [6, 94, 84, 2].

A RKHS is a Hilbert space H of functions for which pointwise evaluation is a bounded linear functional. Let $L_t : H \to \mathbb{R}$ denote such a functional, where elements of H are real-valued functions defined on some domain D; for $f \in H$, $L_t(f) = f(t)$. By the Riesz representation theorem, there exists a unique $K_t \in H$ called the representer of evaluation at t, such that

$$L_t f = \langle f, K_t \rangle_H = f(t) \tag{108}$$

for all $f \in H$, where $\langle \cdot, \cdot \rangle_H$ is the inner product in H. The function $K(t,s) \equiv K_t(s)$ is called the *reproducing kernel*, since the reproducing property

$$\left\langle K(t,\cdot), K(t',\cdot) \right\rangle_H = K(t,t') \tag{109}$$

follows directly from (108). Now consider two elements f and g of H, with $f \equiv \sum_i \alpha_i K(t_i, \cdot)$ and $g \equiv \sum_j \beta_j K(t_j, \cdot)$. Their inner product $\langle f, g \rangle$ is

$$\left\langle \sum_i \alpha_i K_{t_i}, \sum_j \beta_j K_{t_j} \right\rangle = \sum_{i,j} \alpha_i \beta_j \left\langle K_{t_i}, K_{t_j} \right\rangle = \sum_{i,j} \alpha_i \beta_j K(t_i, t_j) \tag{110}$$

From the properties of the inner product, it is straightforward to show that $K(t,s) : D \times D \to \mathbb{R}$ is symmetric and positive definite. Conversely, it can be shown that for every positive definite function K on $D \times D$ there exists a unique RKHS $H(K)$ with K as its reproducing kernel [6]. In fact, $H(K)$ is the completion of the space spanned by $\{f(\cdot) = K(x, \cdot)\}_{x \in D}$ with respect to the inner product (110).

An intuitive appreciation of RKHS may come by contrasting (108) with the situation in $L^2(D)$. In $L^2(D)$, the representer of evaluation is a delta function:

$$\int_D f(s)\delta(s-t)ds = f(t) \tag{111}$$

but the delta function is not in $L^2(D)$, and thus L^2 is not a RKHS. (Indeed, $L^2(D)$ has no bounded linear evaluation functional.) In a RKHS, the existence of the representer $K_t \in H(K)$ implies a relationship between a function value at t and its value at *other* points $t' \in D$ [84]. The Gaussian kernel, for instance, captures the notion that "close is relevant." Other kernels may capture periodicity in f, polynomial trends, and other features.

If the reproducing kernel K is also a Mercer kernel, we can alternatively construct the RKHS $H(K) \subset L^2(D)$ from eigenfunctions of the integral operator defined by the kernel. The eigenfunctions ψ_i and eigenvalues λ_i are given by

$$\int_D K(s,t)\psi_i(t)dt = \lambda_i \psi_i(s) \tag{112}$$

The kernel K can then be expanded in terms of the orthonormal sequence of eigenfunctions $\{\psi_i\}$:

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_i \psi_i(s) \psi_i(t) \tag{113}$$

Now consider two functions $f, g : D \to \mathbb{R}$, with projections

$$f_i = \int_D f(s) \psi_i(s) ds \tag{114}$$

and similarly for g. We define their inner product as:

$$\langle f, g \rangle_K = \left\langle \sum_i f_i \psi_i, \sum_j g_j \psi_j \right\rangle \equiv \sum_i \frac{f_i g_i}{\lambda_i} \tag{115}$$

which induces the norm

$$\|f\|_K^2 = \left\langle \sum_i f_i \psi_i, \sum_j f_i \psi_i \right\rangle \equiv \sum_i \frac{f_i^2}{\lambda_i} \tag{116}$$

It can be shown [94] that $f \in H(K)$ if and only if $\|f\|_K^2 < \infty$, that is, iff:

$$\sum_i \frac{f_i^2}{\lambda} < \infty \tag{117}$$

Requiring the RKHS norm to be finite in effect enforces some "smoothness" on members of $H(K)$, in contrast to $L^2(D)$; since we divide by λ_i in (116), the Fourier coefficients f_i must decay quickly enough for the sum to be finite. The representer property (108) is, of course, preserved under the definition of inner product in (115). In fact the inner products (110) and (115) are equivalent [94].

DISTRIBUTION:

4	MS	9051	Youssef Marzouk, 8351
1	MS	9051	Habib Najm, 8351
1	MS	9051	Bert Debusschere, 8351
1	MS	9051	Philippe Pébay, 8351
1	MS	9051	Dawn Manley, 8351
1	MS	9055	Tom Settersten, 8353
1	MS	9054	Andy McIlroy, 8350
1	MS	9054	Larry Rahn, 8350
1	MS	9159	Paul Boggs, 8961
1	MS	9159	Jaideep Ray, 8964
1	MS	1318	Bart van Bloemen Waanders, 1414
2	MS	9018	Central Technical Files, 8944
2	MS	0899	Technical Library, 4536
1	MS	0123	D. Chavez, LDRD Office, 1011