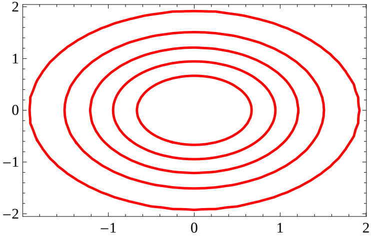
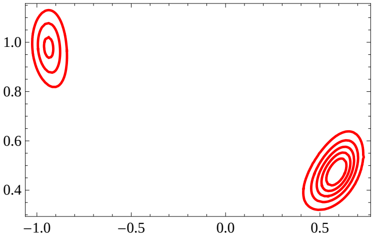
Minimization for conditional simulation: relationship to optimal transport

Dean Oliver, Uni CIPR

28 May 2013

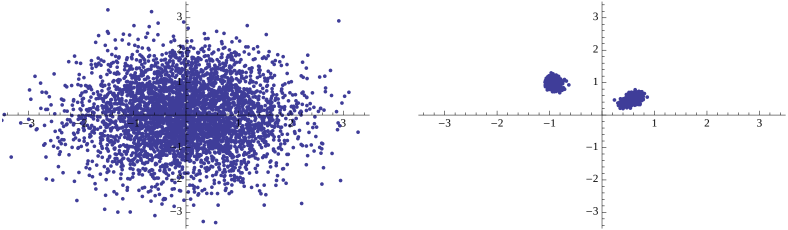
Initial uncertainty in reservoir model parameters	$\xrightarrow{\text{data}}$	Final uncertainty in reservoir model parameters





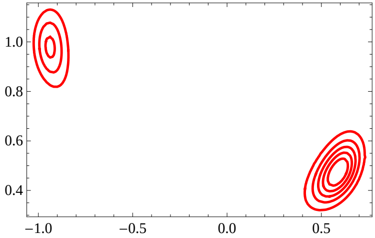
Bayes' rule relates prior pdf to posterior pdf. For high model dimensions, generally need Monte Carlo methods for integration.

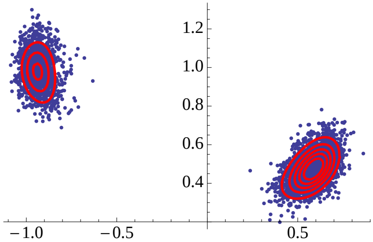
Initial ensemble of reservoir model parameters $\xrightarrow{\text{data}}$ Updated ensemble of reservoir model parameters



Ensembles of particles represent prior pdf and posterior pdf.
Compute expectations by summing over samples.

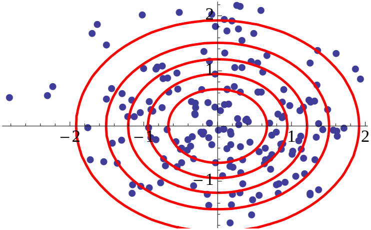
Sampling from the conditional pdf





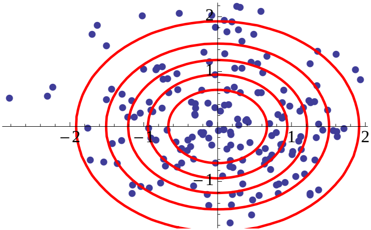
Samples need to be distributed correctly. MCMC? Rejection/acceptance? Scale poorly in high model/data dimensions.

Update the ensemble of samples

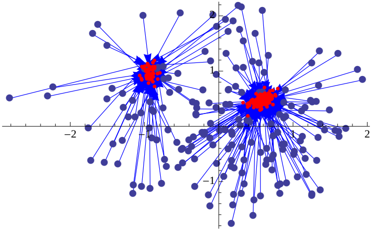


prior

Update the ensemble of samples



prior



posterior

Each sample from prior is *updated*, not *resampled* from posterior.
How to update?

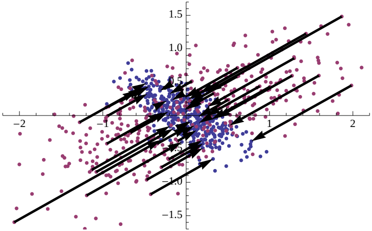
How to update samples?

"The Ensemble Kalman Filter (EnKF) [is] a continuous implementation of the Bayesian update rule" (Li and Caers, 2011).

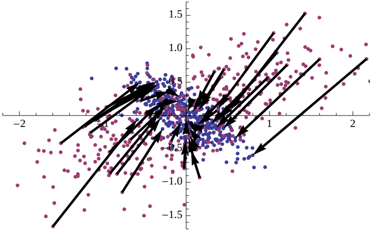
"These methods use . . . 'samples' that are drawn independently from the given initial distribution and assigned equal weights. . . . When observations become available, Bayes' rule is applied either to individual samples . . . " (Park and Xu, 2009)

Note: Bayes rule explains how to update probabilities, but not how to update samples.
(Particle filters do update the probability of each sample using Bayes rule.)

Two transformations from prior to posterior pdf



$$y = \mu_y + L_y[L_y \Sigma_x L_y]^{-1/2} L_y (x - \mu_x)$$



$$y = \mu_y + L_y L_x^{-1} (x - \mu_x)$$

Two equally valid transformations from prior to posterior for linear-gaussian problem. Bayes rule is not sufficient to specify the transformation.

Why it matters

1. If prior is Gaussian and posterior is Gaussian, then any linear transformation of variables that obtains the correct posterior mean and covariance is OK. (Any version of EnKF or EnSRF.)
2. What transformation to use when the observation operator is nonlinear?
   2.1 Randomized maximum likelihood (Kitanidis, 1995; Oliver et al., 1996)
   2.2 Optimal map (El Moselhy and Marzouk, 2012)
   2.3 Implicit filters (Chorin et al., 2010; Morzfeld et al., 2012)
   2.4 Continuous data assimilation or multiple data assimilation (Reich, 2011; Emerick and Reynolds, 2013)
   2.5 Ensemble-based iterative filters/smoothers

Monge's transport problem

Solve for the optimal transformation $s^*(\cdot)$ that minimizes

$$s^* = \arg\min_s \int \|x - s\,(x)\|^2 \; p(x)\,dx \qquad \text{such that } q_s(x) = q(x)$$

Reminder:

- $x \sim p(x)$
- $s(x) = y \sim q_s(y)$, if $x \sim p(x)$
- $q(y)$ is the target pdf of transformed variables

Explicitly (with Gaussian prior)

Let X be a multivariate normal random variable with probability density p,

$$p(x) = c_p \exp\left(-\frac{1}{2}(x-\mu)^T C_x^{-1}(x-\mu)\right),$$

and let Y be a random vector defined by $s(X) = Y$. The probability density of Y is

$$\begin{aligned} q_s(y) &= p\left(s^{-1}(y)\right) \det S^{-1}(y) \\ &= c_p \exp\left(-\frac{1}{2}(s^{-1}(y)-\mu)^T C_x^{-1}(s^{-1}(y)-\mu)\right) \det S^{-1}(y) \end{aligned}$$

where

$$S^{-1}(y) = \nabla s^{-T}.$$

The posterior pdf, for a variable Y with prior distribution $p(y)$ and observation relation $d^o = h(y) + \epsilon_d$ for $\epsilon_d \sim N(0, C_d)$ is

$$q(y) = c_q \exp\left(-\frac{1}{2}(y-\mu)^T C_x^{-1}(y-\mu) \right.$$
$$\left. -\frac{1}{2}(h(y)-d^o)^T C_d^{-1}(h(y)-d^o)\right).$$

If we wish to sample from the posterior, we must seek transformations such that $q_s(y) = q(y)$.

Example[1]

- Prior pdf:
$$p(x) = N(\mu_x, \Sigma_x)$$
- Target pdf:
$$q(x) = N(\mu_y, \Sigma_y)$$

$$y = s(x) := \mu_y + L_y[L_y \Sigma_x L_y]^{-1/2} L_y (x - \mu_x)$$

where $L_x = \Sigma_x^{1/2}$ and $L_y = \Sigma_y^{1/2}$.

Minimizes the expected value of the squared distance $\|X - Y\|^2$

[1] Olkin and Pukelsheim (1982); Knott and Smith (1984); Rüschendorf (1990)

Speculation — why would the minimum distance solution be desirable?

Obtaining a transformation with optimal transport properties may only be important insofar as it simplifies the sampling problem except that the 'natural' optimal transport solution might be more robust to deviations from ideality.

Examples include those in which samples from the prior are complex geological models, in which case making as small of a change as possible might be beneficial.

Transformations between spaces of unequal dimensions

Consider the mapping from $X \in \mathbb{R}^N$ to $Y = s(X) \in \mathbb{R}^M$ where $M < N$ such that

$$\int \|x - As(x)\|^2_{\Sigma_x} \, p(x) \, dx$$

is minimized and the pdfs for x and $y = s(x)$ are

$$x \sim N(\mu_x, \Sigma_x)$$
$$y \sim N(\mu_y, \Sigma_y).$$

Transformations between spaces of unequal dimensions (continued)

The transformation

$$y = \arg\min_y (x - Ay)^T \Sigma_x^{-1} (x - Ay)$$
$$= \left(A^T \Sigma_x^{-1} A\right)^{-1} A^T \Sigma_x^{-1} x.$$

is a solution for the special case

$$\Sigma_y = \left(A^T \Sigma_x^{-1} A\right)^{-1}$$

and

$$\mu_y = \left(A^T \Sigma_x^{-1} A\right)^{-1} A^T \Sigma_x^{-1} \mu_x.$$

Transformations between spaces of unequal dimensions

Note that if

$$A = \begin{bmatrix} I \\ H \end{bmatrix} \quad \text{and} \quad \Sigma_{xd} = \begin{bmatrix} C_x & 0 \\ 0 & C_d \end{bmatrix}$$

then

$$\begin{aligned} y &= \left(A^T \Sigma_{xd}^{-1} A\right)^{-1} A^T \Sigma_{xd}^{-1} \begin{bmatrix} x \\ d \end{bmatrix} \\ &= x + C_x H^T \left(H C_x H^T + C_d\right)^{-1} (d - Hx) \end{aligned}$$

is the optimal transport transformation of $[x \quad d]$ to y for a particular cost function.

(Note that this is the perturbed observation form of the EnKF, or RML for linear inverse problem.)

Nonlinear Transformations

Consider now the problem of determining an approximation of the optimal transformation s^* that satisfies

$$s^* = \arg\min_s \int \left\| \begin{bmatrix} x \\ d \end{bmatrix} - \begin{bmatrix} s(x,d) \\ h(s(x,d)) \end{bmatrix} \right\|^2_{\Sigma_{xd}} p(x,d)\, dx\, dd$$

subject to $y = s(x,d)$ is distributed as $q(y)$ for

$$\begin{aligned} p(x,d) = c_p \exp \Bigg(-\frac{1}{2}(x-\mu)^T C_x^{-1} (x-\mu) \\ - \frac{1}{2}(d-d^o)^T C_d^{-1}(d-d^o) \Bigg), \end{aligned}$$

The prior is Gaussian but the data relationship is nonlinear.

Approximate solution

Consider the transformation defined by the solution of

$$y^* = \arg\min_y \left(\begin{bmatrix} x^* \\ d^* \end{bmatrix} - \begin{bmatrix} y \\ h(y) \end{bmatrix} \right)^T \begin{bmatrix} C_x & 0 \\ 0 & C_d \end{bmatrix}^{-1} \left(\begin{bmatrix} x^* \\ d^* \end{bmatrix} - \begin{bmatrix} y \\ h(y) \end{bmatrix} \right).$$

If h is differentiable, then the variables x^*, d^*, and y^* are related as

$$x^* = y^* + C_x H_*^T C_d^{-1}[h(y^*) - d^*]$$

Approximate solution

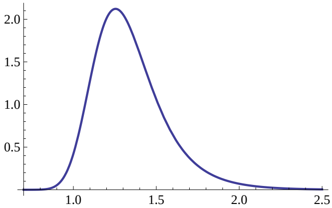
For small $y - y^*$, the pdf of transformed variables

$$\log(q_s(y)) = \frac{1}{2}(y-\mu)^T C_x^{-1}(y-\mu) + \frac{1}{2}(h(y)-d_o)^T C_d^{-1}(h(y)-d_o) + u(\hat{\delta}, d^*, y^*) + \log|J|$$

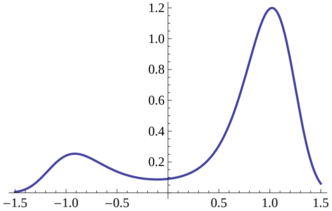
is approximately equal to the target pdf.
$u(\hat{\delta}, d^*, y^*)$ comprises terms that do not depend on y.
The Jacobian determinant of the transformation is

$$J = \left|\frac{\partial(x,d)}{\partial(y,\delta)}\right| = \begin{vmatrix} I & -C_x H_*^T C_d^{-1} \\ H_* & I \end{vmatrix}$$

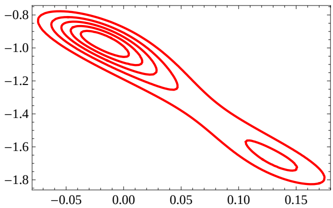
Examples



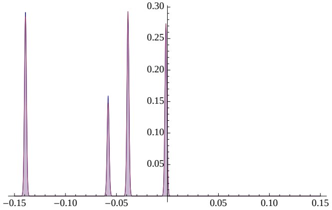
1. univariate: posterior pdf is unimodal but skewed



2. univariate: posterior pdf is bimodal

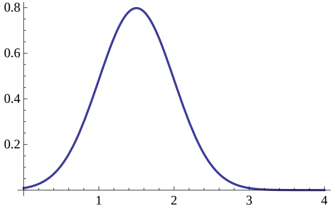


3. bivariate: posterior pdf is bimodal

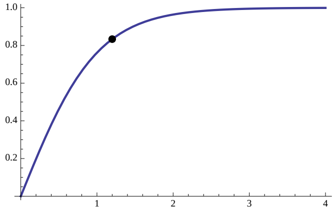


4. 1000 D: posterior pdf has 4 modes

Example 1: unimodal but skewed



prior pdf for model var



$d = h(x) = \tanh(x)$

10,000 independent samples sampled from the prior distribution and mapped to the posterior.

Example 1: unimodal but skewed

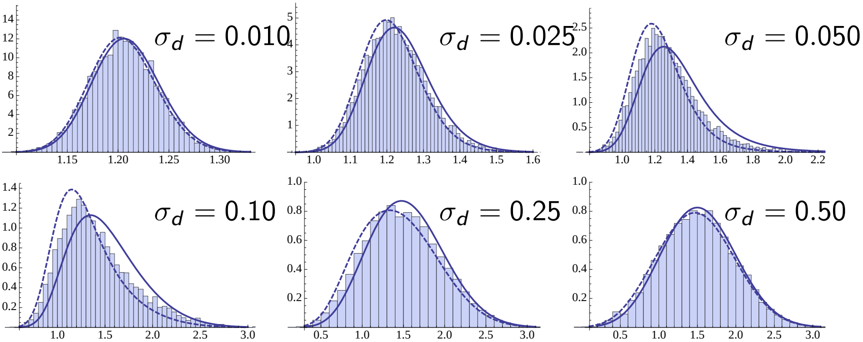
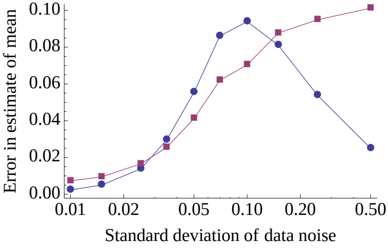
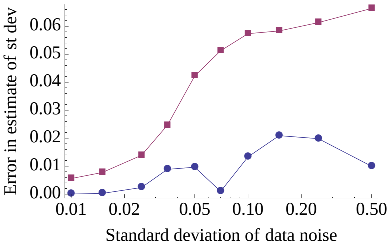


Figure 1: Samples from the Gaussian prior distribution (mean 1.5) were mapped using the minimization transformation. The blue solid curve shows the true pdf. The dashed curve shows the product of the true pdf and the Jacobian determinant of the transformation.

Example 1: unimodal but skewed



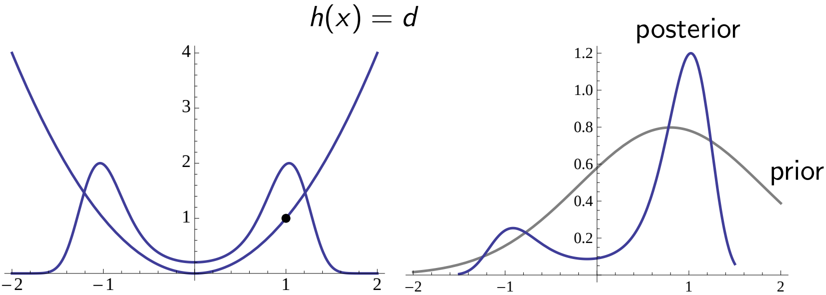
(a) Absolute error in estimate of mean.



(b) Absolute error in estimate of standard deviation.

Figure 2: Comparison of empirical moments from 10,000 independent samples using minimization-based sampling with true moments. The magenta dots are 2 standard deviations in the error of samples of 100 realizations from the true distribution.

Example 2: single variable, bimodal

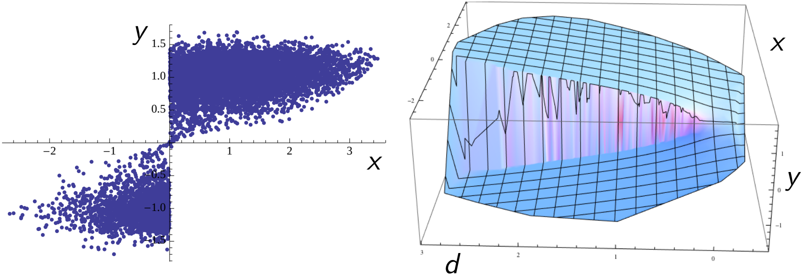


(a) Likelihood (b) prior and posterior

Figure 3: The observation tells that $x \approx \pm 1$. Prior says that $x \approx 0.8$.

Will sample from prior, transform to posterior.

Example 2: single variable, bimodal

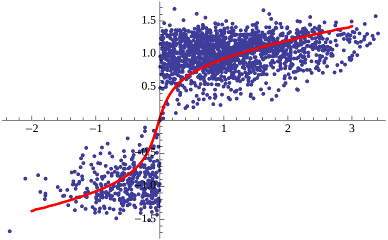


(a) y vs x (b) y vs (x, d)

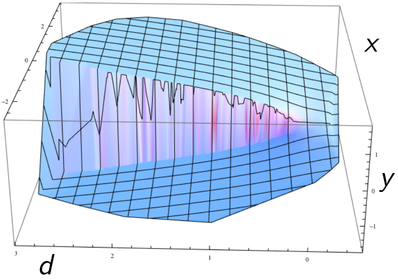
Figure 4: Transformed samples from approximate posterior vs samples from prior.

$$y^* = \arg\min_y \left(\begin{bmatrix} x^* \\ d^* \end{bmatrix} - \begin{bmatrix} y \\ h(y) \end{bmatrix} \right)^T \begin{bmatrix} C_x & 0 \\ 0 & C_d \end{bmatrix}^{-1} \left(\begin{bmatrix} x^* \\ d^* \end{bmatrix} - \begin{bmatrix} y \\ h(y) \end{bmatrix} \right).$$

Example 2: single variable, bimodal



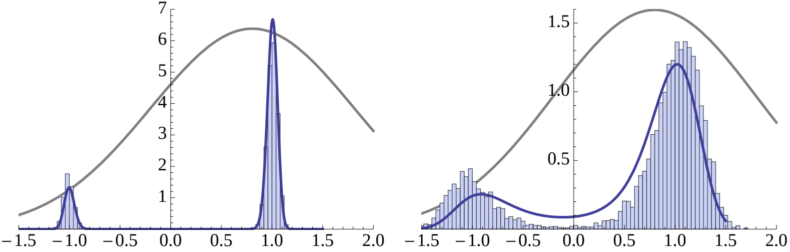
(a) y vs x



(b) y vs (x, d)

Figure 5: Transformed samples from approximate posterior vs samples from prior. Red curve shows the (correct) optimal map defined for minimizing expected distance between x and y.

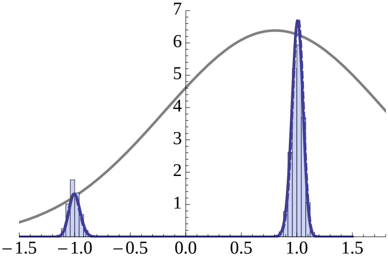
Example 2: single variable, bimodal



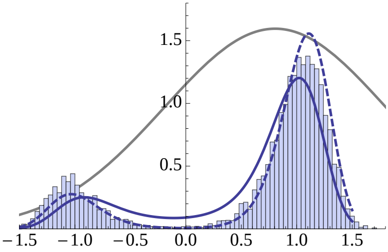
(a) $\sigma_d = 0.1$ (b) $\sigma_d = 0.5$

Figure 6: Distributions of transformed samples for two values of noise in the observation. Samples from the prior distribution (gray curve) were mapped using the minimization transformation. The blue solid curve shows the true pdf. Clearly under-sampled in region between modes.

Example 2: single variable, bimodal



(a) $\sigma_d = 0.1$



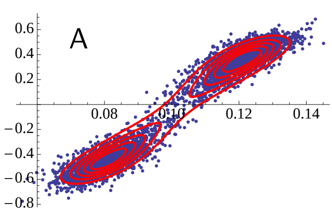
(b) $\sigma_d = 0.5$

Figure 7: Same as previous slide, but the dashed curve shows the product of the true pdf and the Jacobian determinant of the transformation.

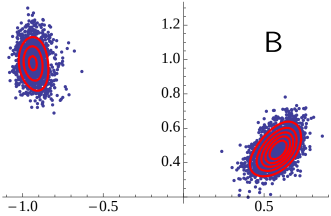
Example 3: two variables, bimodal

1. The locations of the modes in 20 experiments were randomly sampled.
2. Bimodality is established through nonlinearity of one observation of $x_1^2 + x_2^2$.
3. Second observation is made of a linear combination of the two variables
4. Approximately 4000 samples for each experiment

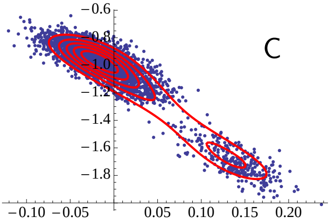
Example 3: two variables, bimodal



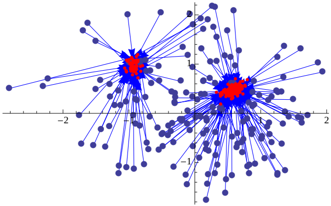
(a) Example A: modes relatively close, good sampling.



(b) Example B: modes relatively distant, good sampling.

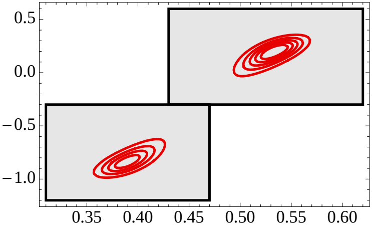


(c) Example C: modes relatively close, poor sampling.

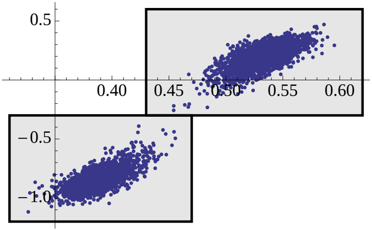


(d) Example B: mapping of individual samples from prior to posterior.

Example 3: Quantitative comparison



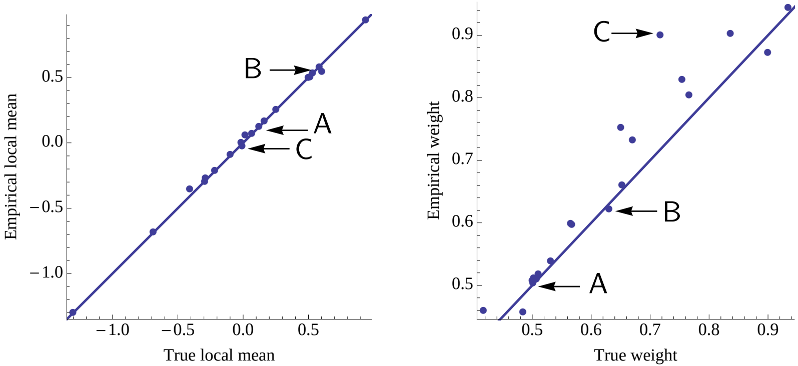
true pdf



sampling

For each example, we compute the local mean for each of the modes of the distribution and the total probability for each mode.

Example 3: Summary results



(a) Comparison of sampled estimate of local mean for modes with true local mean.

(b) Comparison of sampled estimate of weight on one of the modes with true weight.

Figure 9: Approximately 4000 optimization-based samples are used to compute the approximate weights and means for both modes of the sampled distributions. Labeled points refer to experiments shown on a previous slide.

Example 4: 4 modes, high dimension

The prior probability density is Gaussian with independence of variables and unit range:

$$x \sim A \exp[-0.5 x^T x]$$

The likelihood is the sum of delta functions of random weights:

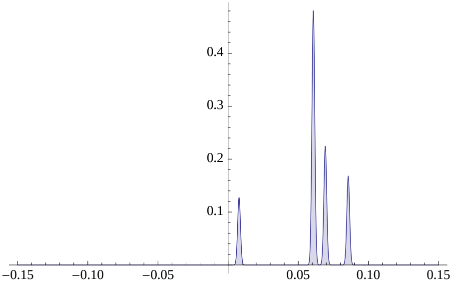
$$L[m|d] = \sum_{i=1}^{4} b_i \delta(x - \alpha_i)$$

so that the posteriori pdf that we wish to sample is

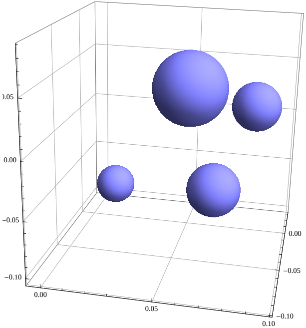
$$p[m = \alpha_i | d] \propto b_i \exp[-0.5\, \alpha_i^T \alpha_i]$$

The α_i were normally distributed with mean 0 but small variance.

Example 4: random test pdf



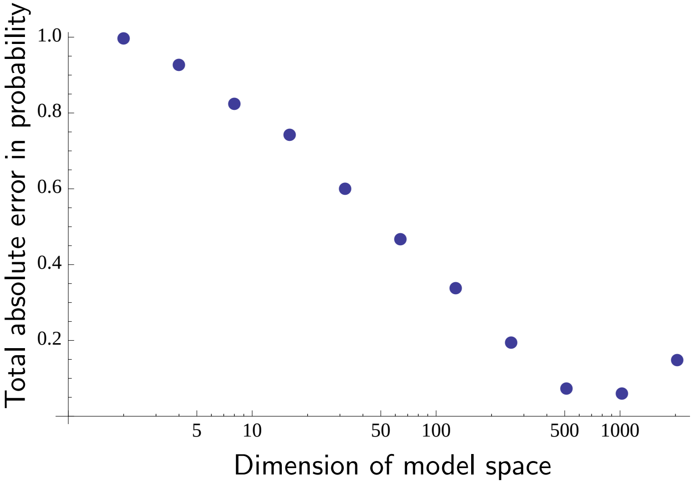
projection 1D



projection 3D

Computations in dimensions as high as 2000.

Example 3: Summary results



1000 random samples of α_i and β_i. 10,000 samples used to compute error for each set of α_i and β_i.

Key points

1. Bayes' rule does not specify how to update individual samples. Need some other criterion (such as 'minimum expected distance').
2. For some types of problems, can simplify the optimal transport problem (for probability density) by careful choice of a cost function. Then use unconstrained minimization.
3. Samples from RML should probably be weighted according to the Jacobian determinant of the transformation.

Remaining questions

1. Defining a cost function was key to getting a good approximation of optimal transport solution easily – how to generalize for nongaussian prior?
2. How to efficiently compute the Jacobian of transformation for weighting of updated samples?
3. Relationship of EnRML to RML?

A. J. Chorin, M. Morzfeld, and X. Tu. Implicit particle filters for data assimilation. *Communications in Applied Mathematics and Computational Science*, 5(2):221–240, 2010. doi: 10.2140/camcos.2010.5.221.

T. A. El Moselhy and Y. M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23): 7815–7850, 2012. ISSN 0021-9991. doi: 10.1016/j.jcp.2012.07.022.

A. A. Emerick and A. C. Reynolds. Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, 55:3–15, 2013. ISSN 0098-3004. doi: 10.1016/j.cageo.2012.03.011.

P. K. Kitanidis. Quasi-linear geostatistical theory for inversing. *Water Resour. Res.*, 31(10):2411–2419, 1995.

M. Knott and C. S. Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984. doi: 10.1007/BF00934745.

H. Li and J. Caers. Geological modelling and history matching of multi-scale flow barriers in channelized reservoirs: methodology and application. *Petroleum Geoscience*, 17:17–34, 2011.

M. Morzfeld, X. Tu, E. Atkins, and A. J. Chorin. A random map implementation of implicit filters. *Journal of Computational Physics*, 231(4):2049–2066, 2012. ISSN 0021-9991. doi: 10.1016/j.jcp.2011.11.022.

D. S. Oliver, N. He, and A. C. Reynolds. Conditioning permeability fields to pressure data. In *European Conference for the Mathematics of Oil Recovery, V*, pages 1–11, 1996.

I. Olkin and F. Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48(0):257–263, 1982. ISSN 0024-3795. doi: 10.1016/0024-3795(82)90112-4.

S. Reich. A dynamical systems framework for intermittent data assimilation. *BIT Numer Math*, 51(1):235–249, 2011. ISSN 0006-3835. doi: 10.1007/s10543-010-0302-4.

L. Rüschendorf. Corrigendum. *J. Multivar. Anal.*, 34(1):156, 1990. ISSN 0047-259X.