

Tutorial

Prior information, sampling distributions, and the curse of dimensionality

Andrew Curtis* and Anthony Lomax†

ABSTRACT

This tutorial addresses geometrical issues that concern the specification of high-dimensional sampling distributions in Bayesian inversion. We illustrate that simple, low-dimensional geometrical concepts that are sometimes used to construct such distributions may become completely distorted (and even untrue) in higher dimensional problems. This has important implications for Bayesian inversion: if a convenient sampling distribution is constructed using low dimensional geometrical concepts which cause it to differ from the distribution representing our prior information, these differences can become extremely expensive to correct in higher dimensions. Indeed, they may make a nonlinear inversion computationally intractable when this need not be the case. A crucial factor in Bayesian inversion is, therefore, whether one firmly believes in a particular prior distribution. If so, this distribution may constitute the most efficient sampling distribution, even in cases where it is not straightforward to draw samples from that prior distribution. The sampling artifacts described above then become irrelevant since they represent true prior beliefs.

INTRODUCTION

Bayesian inference or inversion to constrain model parameters using measured data consists of two main stages (Tarantola and Valette, 1982; Tarantola, 1987):

- 1) Specifying *prior* information, that is, all information available from sources independent of the current data, usually described by a *prior probability distribution function*.

- 2) Updating the prior information with additional information gained from current data, the resultant state of information usually being described by a *posterior probability distribution function*.

Ideally, the second stage requires that we construct a map of the misfit surface over the portion of model space consistent with the prior information, where the misfit measures the difference between the synthetic data predicted by any model and the observed data. In practical geophysical problems the number of parameters (dimensionality) is usually high, and the mapping between models and synthetic data (the forward function) may be nonlinear. Thus the misfit surface is high dimensional and is often complicated with many extrema (e.g., Lomax and Snieder, 1995; Sambridge, 1998).

Two approaches to exploring the misfit surface are deterministic or stochastic (pseudorandom) sampling of the portion of the model space consistent with the prior distribution, and linearized inversion techniques. Linearized inversion often appears to be computationally efficient because it uses local, quadratic approximations to the misfit surface to take steps from an initial model that (generally) improve the data fit. However, only a single point solution and associated local uncertainty estimate (based on the local quadratic approximation) are found. If the choice of initial model was based upon weak prior information, a local misfit minimum may be found in multimodal inverse problems. In such situations, no indication is given that a local rather than global minimum has been found, and the computational cost of checking this can be extremely high.

In principal, global sampling techniques do not have these problems since they allow all parts of model space that might give good data fits to be sampled and thus provide more complete information about the posterior uncertainty distribution (Kirkpatrick et al., 1983; Smith et al., 1992; Lomax and

Manuscript received by the Editor December 28, 1998; revised manuscript received April 17, 2000.

*Schlumberger Cambridge Research Ltd., High Cross, Madingley Road, Cambridge, CB3 0EL, United Kingdom. E-mail: curtis@cambridge.scr.slb.com.

†UMR Geosciences Azur, 250 Rue Albert Einstein (bat 4), 06560 Valbonne, France. E-mail: lomax@faillie.unice.fr.

© 2001 Society of Exploration Geophysicists. All rights reserved.

Snieder, 1995; Mosegaard and Tarantola, 1995; Vinther and Mosegaard, 1996; Sambridge, 1998). Sampling methods, however, are computationally expensive in problems with many model parameters and weak prior information because a large volume of model space must be explored. In such situations, the initial sampling distribution is sometimes also incrementally weighted by the data fit so that the samples gradually become distributed according to the posterior distribution (called importance sampling, see e.g., Lepage, 1978; Mosegaard and Tarantola, 1995; Sambridge, 1998). Although this sampling scheme is more efficient in principal, the same portion of model space must be explored initially (if only to check that parts of it don't fit the data).

In practise, we must specify a sampling distribution that respects the prior information, but also makes the inverse problem tractable. Often the distribution that represents prior information exactly is difficult to sample directly; a more convenient initial sampling distribution is sometimes chosen (e.g., uniform, Gaussian, etc.). This sampling distribution may then be weighted incrementally by data fit. To obtain the formal posterior probability distribution, one then corrects for any differences between the initial sampling and the prior distributions (Lepage, 1978; Sen and Stoffa, 1995).

This tutorial concerns geometrical concepts that might be used to construct practical sampling distributions when the number of model parameters M (or model space dimensionality) is low ($M = 1 \leq 3$). When applied to high-dimensional model spaces, these concepts do not always give results that might be expected, and may make correction for differences between the sampling and prior distributions very inefficient. The mathematics included herein is not new; our main objective is to ensure that the implications of the concepts presented are more widely appreciated within the inversion community.

PRIOR INFORMATION

In many practical inverse problems, prior information is vague. It tends to be of the form, "I know very roughly what values my parameters will take, and I do not expect them to go outside of certain known bounds." Consider a case where we wish to constrain a velocity model of the earth by Bayesian inversion of seismic data. Say the P velocity model $\mathbf{m} = [m_1, \dots, m_M]^T$ is parameterized by constant velocities m_i in M individual depth ranges that span the whole depth range of interest. Prior information on \mathbf{m} may be that we expect each velocity to be somewhere around 5 km/s from studies conducted previously. Also, we know that previous, reliable studies in the region have never estimated a velocity outside of the range [4, 6] km/s within the relevant depth range, and thus we believe that 4 and 6 km/s are truly velocity bounds.

Such information about the bounds provides what are called *independent linear constraints* on each parameter [i.e., constraints of the form $L^j(\mathbf{m}) < b^j$, for some linear functions L^j of each m_i independently, where b^j are constants]. This should be distinguished from geophysical problems involving quadratic (or higher-order) constraints (e.g., energy constraints in potential field problems). The latter implicitly introduce parameter correlations and insurmountable problems in higher dimensions (see Backus, 1988, and Scales, 1996, for discussions of the "curse" of dimensionality). Here, we concentrate on independent linear constraints since these are often encountered in seismological inverse problems.

To solve the inverse problem stochastically, we must build this prior information into a sampling distribution. Notice that in this case there is no explicit prior probability distribution; unless very carefully chosen, any particular choice of sampling distribution is likely to impose different constraints on the sample set than are implied by the prior information given above.

We focus on sampling distributions that can be applied to each parameter independently since these are generally easier to implement (and hence are more often used) than those with interparameter dependency. We will consider two sampling goals that turn out to be difficult to attain in high-dimensional problems:

- 1) The sampling should occur across the full parameter ranges, but should be relatively dense close to the central model (5 km/s).
- 2) The sampling should not be excessive close to, or outside of, the parameter bounds (4 and 6 km/s).

UNIFORM SAMPLING DISTRIBUTIONS

Unless we have specific information that defines the form of the prior probability of velocities within the range [4, 6] km/s, it is often the case that we choose a uniform sampling distribution across this velocity range. This is especially common when Bayesian inversion will be performed using stochastic sampling methods because pseudorandom, uniform sampling is easy to perform. If dimension $M = 2$, this representation defines a uniform probability over the square in model space defined by $4 \leq m_1 \leq 6$, and $4 \leq m_2 \leq 6$; if dimension $M = 3$, then this defines a uniform probability over the corresponding cube (Figure 1). If the number of dimensions $M > 3$, then the uniform probability is defined over a hypercube. If individual model parameters have different velocity ranges, then the hypercube would become stretched to form a hypercuboid. In such cases, we assume that parameters can be scaled appropriately to have a common range, after which the discussion and figures presented here are directly applicable. With this representation, we do not place any emphasis on the central point (5 km/s) on each model parameter axis.

The discussion that follows concerns the nature of this uniform distribution as the number of model space dimensions increases. Let the length of each edge of the square (cube, henceforth hypercube) be a . Call this the a -hypercube with volume $V_M^a = a^M$ in M dimensions. Consider embedding a smaller hypercube of side length fa where $f < 1$ (the fa -hypercube, Figure 1). The ratio of the volume of the fa -hypercube to the volume of the a -hypercube is

$$\frac{V_M^{fa}}{V_M^a} = \frac{(fa)^M}{a^M} = f^M. \quad (1)$$

Since $f < 1$, the ratio in equation 1 tends to zero as $M \rightarrow \infty$.

This result is important in high-dimensional inverse problems where the sampling distribution is specified as uniform within a hypercube, as above. It states, in the velocity problem, for instance, that as the number of dimensions increases, the hypervolume of our distribution becomes almost completely contained within a thin band close to the hypercubic faces (since the result is true even for $f = 0.99$). This clearly contravenes goal 2.

This result is in direct contrast with our perceptions from one-, two-, and three-dimensional model spaces that we might "visualize" in our minds. In one dimension, for instance, if we

collect a large set of random Monte Carlo samples according to a uniform distribution, we perceive that we are likely to obtain good coverage over the whole range of models within the bounds, including those models near the point centrally located between the bounds. Thus, even if we expect that the velocity model might be roughly centrally located between the bounds, we may choose to use a uniform sampling distribution to avoid the possibility of any incorrect bias towards centrally located models. This choice may seem reasonable if our expectations are not sufficiently precise to be represented by any particular distribution [see Scales and Snieder (1997) for a discussion of

how “reasonable” this is!]. However, if we apply the same philosophy to high-dimensional model spaces, the results above show that samples from a uniform distribution in a hypercube will almost never lie close to the central model. Thus, a uniform sampling distribution in high-dimensional model spaces does not even approximately satisfy either of goals 1 or 2.

The character of typical velocity models sampled from a uniform distribution within a 10-dimensional hypercube and its fa -hypercube where $f = 1/2$ are shown in Figure 2 (top left and top right, respectively). Models sampled from either hypercube typically include dramatic variations from parameter

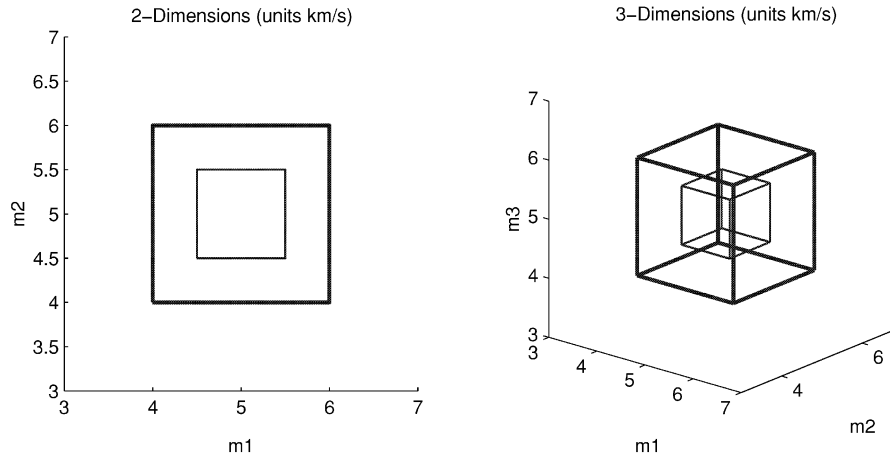


FIG. 1. Two- and three-dimensional model space where each model parameter lies within the range [4, 6] km/s. The inner square and cube have half the edge length of the outer ones [$f = 1/2$ in Equation (1)].

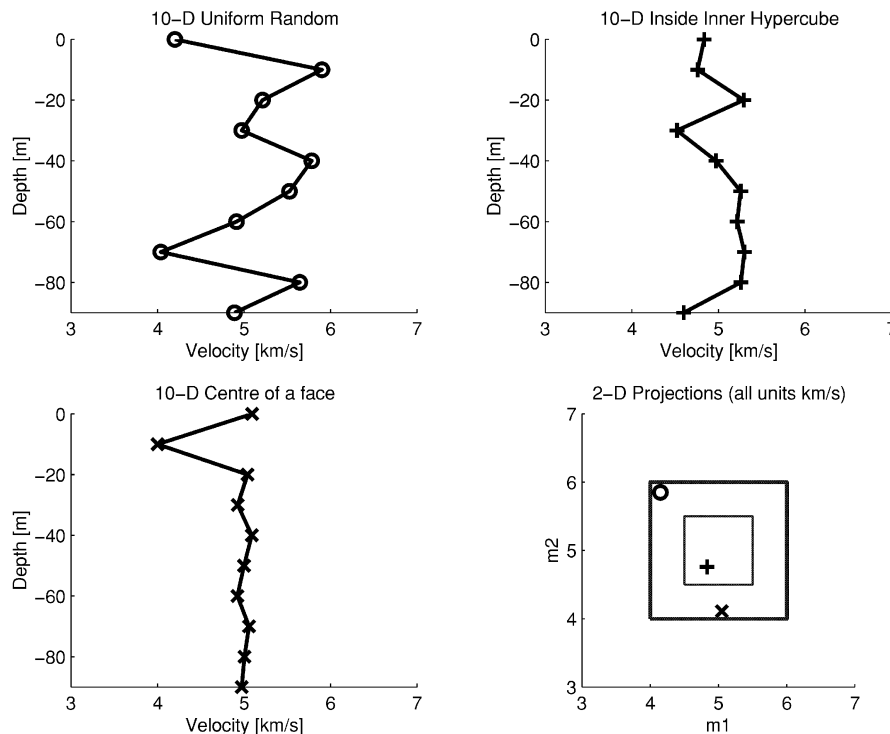


FIG. 2. Example 10-dimensional random models sampled uniformly from the velocity a -hypercube ($a = 2$, top left), the inner fa -hypercube ($f = 1/2$) (top right), close to the center of one of the a -hypercubic faces (bottom left), and their projections onto the first two dimensions (bottom right).

to parameter, but those from the fa -hypercube are very much more localized around the central model (see also the lower right plot). The mathematics above states that there are very many more models “like” that shown top left than “like” those shown top right.

This is illustrated explicitly in Figure 3 in which uniform probability density functions inside a -hypercubes of increasing dimensionality were sampled and the proportion of samples that lay within the fa -hypercube ($f = 1/2$) is plotted. This decreases dramatically as dimensionality increases; the reduction is significant even in only a few dimensions. Practical inverse problems may include hundreds of dimensions. Hence, if we are interested in satisfying goal 1, for example, then very many samples may be required to obtain even a single sample from the central portion of the model space in which we are most interested. Correction of the sampling distribution to reflect the prior information would therefore be fairly inefficient.

The reason that so many models lie outside of the fa -hypercube is that even a single model parameter lying further than fa from the mean value causes this to be true (e.g., Figure 2, lower left and lower right). One way to sample models that are more likely to lie in the fa -hypercube is therefore to sample smooth models. To illustrate this, we use the first-order measure

$$s = \frac{1}{M} \sum_{i=2}^M |m_i - m_{i-1}| \quad (2)$$

(normalised by the number of dimensions M) to measure model smoothness. In Figure 3, we also plot the proportion of “smooth” models ($s < 0.25$) that lie within the fa -hypercube. Smoothing certainly increases the proportion of models within

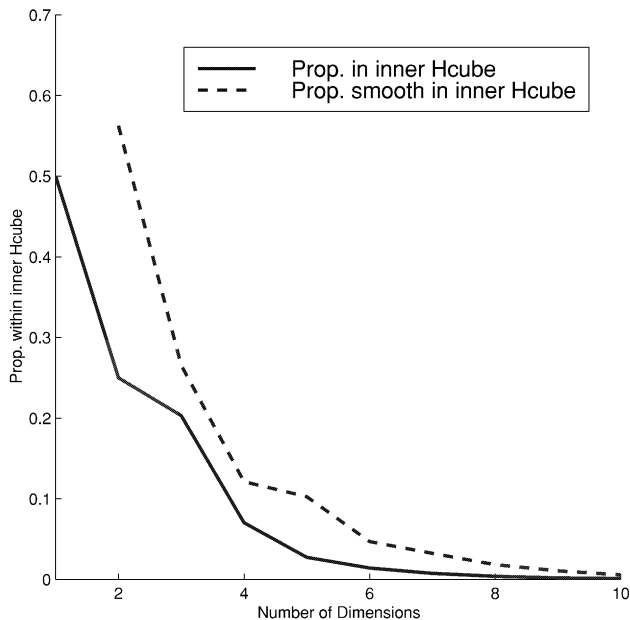


FIG. 3. Proportion of models randomly and uniformly sampled from a -hypercubic model spaces similar to those in Figure 1 that lie within the fa -hypercube ($f = 1/2$, solid), plotted as a function of dimension. The dashed line shows the proportion of smooth models that lie within the fa -hypercube, where a smoothing threshold of $s < 0.25$ was used in equation (2).

this region, but this criterion is clearly insufficient to stop the vast majority of models lying outside of it as the number of dimensions increases.

The above discussion shows that, as the dimensionality of the model space increases, the choice of sampling distribution becomes more and more important. Ultimately the cost of evaluating the data fit offered by any particular model always limits the number of stochastic samples that may be taken. Hence, in high-dimensional model spaces where some information is available about the “most likely” region of model space, the specification of very narrow or nonuniform sampling distributions that sample this region efficiently becomes crucial to the computational tractability of Bayesian inference and inversion.

GAUSSIAN SAMPLING DISTRIBUTIONS

One way to accomplish the requirement of centrally-concentrated sampling is to sample according to a Gaussian probability distribution in each model parameter (defined by a mean model and variance or uncertainty around the mean). The resulting joint distribution of all model parameters is also a multidimensional Gaussian distribution. This approach is especially common in linearized Bayesian inversions since formal theory exists to update a Gaussian prior with information from data with Gaussian uncertainties to obtain a Gaussian probability distribution as the linearized inverse problem solution (e.g., Tarantola and Valette, 1982). Clearly, in one-dimensional problems, the prior information then implies that the bulk of the prior probability mass (the area, volume, hypervolume beneath the distribution function) is located around the mean or central model. Hence, if a random Monte Carlo sample was taken from the Gaussian distribution, most of the samples would end up in some neighborhood of the mean.

Consider now a Gaussian distribution in higher dimensional model spaces. If the probability of any model \mathbf{m} under consideration is described by a multivariate Gaussian distribution, then it depends on the difference $\mathbf{r} = \mathbf{m} - \bar{\mathbf{m}}$ between the current model and the mean (or central) model $\bar{\mathbf{m}}$. The probability density function in M model dimensions can be written as,

$$G(\mathbf{m}; \mathbf{C}) = \frac{1}{[(2\pi)^M |\mathbf{C}|]^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \mathbf{r} \mathbf{C}^{-1} \mathbf{r}^T\right), \quad (3)$$

where \mathbf{C} is a covariance matrix. Depending on the form of \mathbf{C} , this prior distribution can introduce a combination of *damping* (models are more probable closer to the central model), *smoothing* (smooth models are more probable than rough models) or *roughening* (opposite of smoothing) to an inverse problem (Figure 4). Below we consider mainly damping, but similar discussions pertain to other cases.

If the covariance matrix has no nonzero off-diagonal elements (\mathbf{C} is a diagonal matrix), it forms a damping distribution. The magnitudes of the diagonal elements govern the strength with which each model parameter is biased towards the central model. If each model parameter is normalized in proportion to the corresponding diagonal element of \mathbf{C} , all diagonal elements become equal with value σ^2 . The Gaussian density function $G(\mathbf{m}; \mathbf{C})$ in M dimensions is then spherically symmetric around the central model and becomes a function only of

the radius $r = \|\mathbf{m} - \bar{\mathbf{m}}\|$ from model $\bar{\mathbf{m}}$ and of the variance σ^2 (Figure 4, left and center):

$$G(\mathbf{m}; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right). \quad (4)$$

Despite the fact that a one-dimensional Gaussian distribution was applied to each coordinate axis independently as in the uniform case, notice that the resulting multidimensional distribution is not “hypercubic” in nature (Figure 4, center), and probability is not concentrated close to the hypercubic faces as it was in the uniform case.

However, notice what happens as the number of dimensions becomes high: for any fixed $\sigma > 0$ (where σ is determined from prior information), almost every sample will contain one or more parameters with values outside of the range [4, 6] km/s. This is true because we select each parameter m_i independently from identical one-dimensional Gaussians $G(m_i; \sigma^2)$; as the number of selections (parameters) tends towards infinity, one or more must lie in the region outside of any fixed, finite parameter range because the probability is nonzero in this region. Thus, by using Gaussian sampling distributions we contravene goal 2.

A method sometimes used to avoid such unwanted parameter values is to truncate the Gaussian distribution. That is, either parameters that lie outside of the desired range are reselected until a value within the range is obtained, or the Gaussian “tails”

outside of the range are simply removed and the remaining part is renormalised to obtain a sampling distribution such as that shown in the left plot of Figure 4.

However, notice that there is now a discontinuity in the sampling density at parameter values 4 and 6 km/s. As the number of dimensions becomes large, almost every model will have a few parameters with values just inside of this range, whereas none will have values just outside of it. This may give undue importance to these values. After all, if no prior studies have found velocities outside of the range [4, 6] km/s, then neither have they found velocities outside of [3.9, 6.1] km/s. Is the former range really so significant compared to the latter? If this sampling phenomenon is perceived to be a problem, then again we contravene goal 2.

TAPERED SAMPLING DISTRIBUTIONS

Another possible way to focus sampling on a specific region of the model space is to apply prior information in the form of a centrally uniform probability distribution that tapers to zero at each end of the range of interest of each parameter axis (henceforth, *tapered-uniform* distributions, Figure 5, left). In this way one might hope to reduce the relatively large number of samples that lie close to the hypercube edges, while still ensuring that model parameters lie within the specified bounds.

Using this strategy, the joint distribution of all model parameters consists of a uniform distribution within an inner

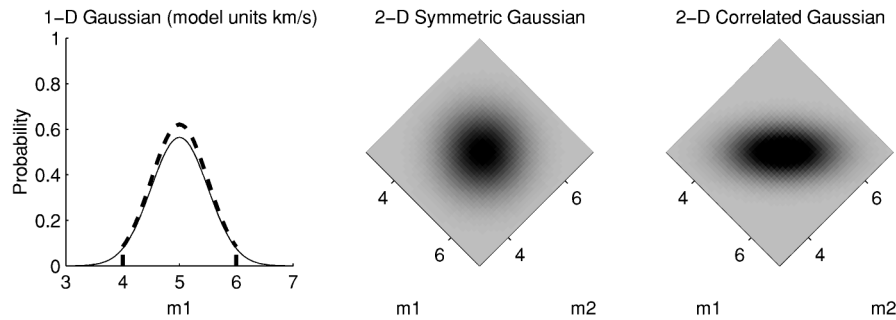


FIG. 4. Uncorrelated Gaussian distributions in one (left) and two (center) dimensions, and a correlated Gaussian in two dimensions (right). Also shown is a truncated, renormalized Gaussian distribution (left, dashed line).

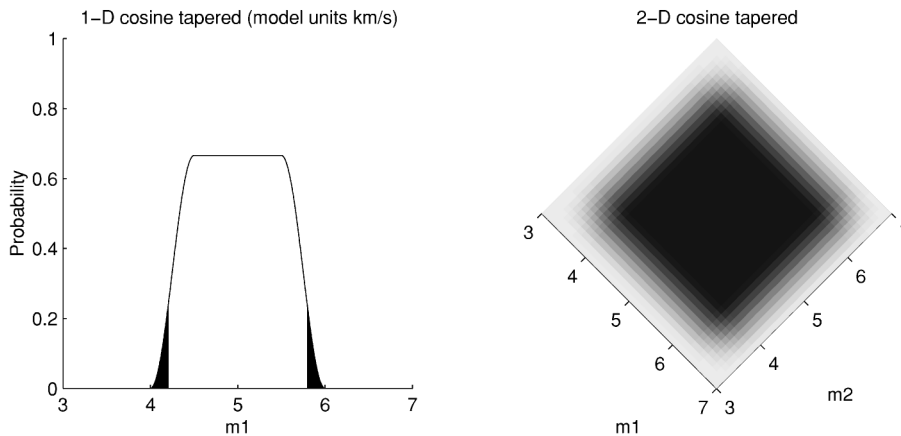


FIG. 5. Prior distributions in one and two dimensions formed by applying a cosine taper to the edges of the uniform hypercubic distribution. The gray scale on the right-hand plot represents the probability (black being most probable). The shaded area in the left plot represents the area A within ϵ of the bounds referred to in the text.

hypercube with probability density tapering to zero along all boundaries of the outer hypercube (Figure 5, right). If the range of each parameter is $2a$ and the taper is applied within the distance $a/2$ of both bounds on each coordinate axis, then the inner hypercube has edges of length a . Applying the arguments in the section above on uniform sampling to the inner hypercube, we infer immediately that as the number of dimensions increases, samples lying within the inner hypercube will tend to lie close to the inner hypercubic faces. However, since in this case the inner hypercube can be made arbitrarily small, this fact need not contravene goal 1.

As the edge length of the inner hypercube is reduced, the proportion of samples with parameter values lying within parameter ranges spanned by the tapers increases. This begs the question, how concentrated does the sampling close to outer hypercubic faces become? In fact, as long as there is a nonzero probability that each parameter lies within distance ϵ of the bounds for $\epsilon > 0$, and such tapers are applied to each parameter independently, then as the number of dimensions increases, the probability that at least n parameters of any model sample lie within ϵ of the bounds tends towards one for any given n . In some situations, this may be perceived to contravene goal 2.

To overcome this, the form of the taper needs to depend on the number of dimensions. Say, for example, that for each model sample we wish to place a limit m on the expected number of parameter values that lie within ϵ of the bounds. If the area under the one-dimensional sampling distribution for each parameter within ϵ of the bounds is A (shaded area in Figure 5, left), then the probability that n parameter values of any particular model sample fall within ϵ of the bounds follows a binomial probability distribution with expected value $\bar{n} = A \times M$ where the model has M dimensions (e.g., Wetherill, 1982). If we wish to place a limit m on the expected number of samples that have such extreme values, we must choose a taper such that area $A \leq m/M$. Hence as M increases, A must be reduced, and thus the form of the taper must depend on M . Example cosine tapers that could be used for model spaces with one, two, and three dimensions are shown in Figure 6.

To summarise, tapered uniform distributions applied to each parameter independently may be constructed such that they satisfy goals 1 and 2 respectively if (1) the inner hypercube is sufficiently small that the faces lie “close enough” to the central model, and (2) the taper shape is itself a function of the number of model space dimensions such that the area beneath the taper used is reduced in line with the condition $A \leq m/M$ given above.

DISCUSSION

The above examples illustrate how geometrical concepts that often guide the specification of sampling distributions in low-dimensional problems may produce undesirable distributions in high-dimensional problems. In complex, nonlinear inverse problems with large model spaces, carefully selected sampling distributions may be the only means of creating computationally tractable problems from otherwise insoluble ones. We have demonstrated that certain rules of thumb, often applied in low-dimensional problems can lead to massive increases in computational workload in the inverse problem solution (Figure 3).

For instance, if uniform sampling distributions are used when a priori we might expect models close to some particular model

to be more likely than those further away, this increase in computational workload tends to infinity as the number of dimensions increases. Prior information is almost never purely uniform in reality, so it may be better to use other, more centrally-concentrated distributions.

Gaussian distributions can be used to increase sampling density close to any given mean model estimate, but produce artifacts at parameter bounds that can not be removed by simple truncation. Tapered-uniform distributions allow sampling to be centrally concentrated without producing unwanted increases in sampling density close to the parameter bounds, but only if the form of the tapers used is chosen as a function of the number of model parameters.

This tutorial follows on naturally from that of Scales and Snieder (1997) in which different formalisms for representing prior information were discussed. Here, we have shown that whatever formalism is employed, the choice of sampling distribution used to explore possible inverse problem solutions also becomes very important in high-dimensional problems. A poor choice may produce unwanted sampling artifacts and even render problems computationally intractable when this need not be the case. When constructing such sampling distributions, it is important to examine mathematically the behavior of the sampling density as the number of dimensions grows. It is not good to rely on intuition which is derived from the three-dimensional world in which we live; most of us have no intuition about 10-, 20-, or 100-dimensional spaces at all.

In conclusion, if we firmly believe that our prior probability distribution represents our prior information, then this distribution may form the most efficient sampling distribution, even in cases where sampling directly from that distribution may be nontrivial and computationally expensive. Any remaining sampling artifacts such as those described herein, while

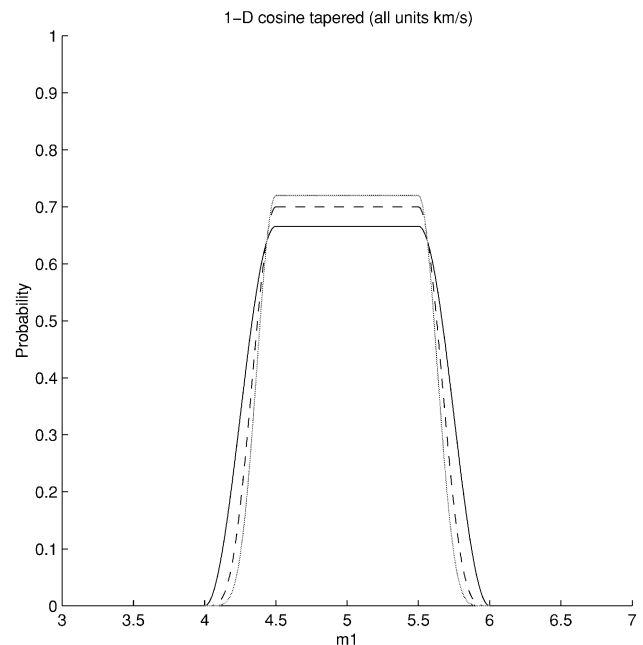


FIG. 6. Example symmetrically cosine tapered-uniform distributions where the taper varies with increasing dimensionality (black, one dimension; dashed, two dimensions; gray, three dimensions).

interesting, then become irrelevant since they represent true prior beliefs.

ACKNOWLEDGMENTS

We were very grateful for constructive reviews from John Scales, Malcolm Sambridge, Albert Tarantola, and Sven Treitel, which inspired many significant changes in the manuscript.

REFERENCES

- Backus, G. E., 1988, Hard and soft prior bounds in geophysical inverse problems: *Geophys. J.*, **94**, 249–261.
- Kirkpatrick, S., Gelatt, C. D., Jr., and Vecchi, M. P., 1983, Optimization by simulated annealing: *Science*, **220**, 671–680.
- Lepage, G. P., 1978, A new algorithm for adaptive multidimensional integration: *J. Comp.*, **27**, 192–203.
- Lomax, A., and Snieder, R., 1995, Identifying sets of acceptable solutions to non-linear, geophysical inverse problems which have complicated misfit functions: *Nonlin. Proc. in Geophys.*, **2**, 222–227.
- Mosegaard, K., and Tarantola, A., 1995, Monte Carlo sampling of solutions to inverse problems: *J. Geophys. Res.*, **100**, No. B7, 12431–12447.
- Sambridge, M., 1998, Exploring multidimensional landscapes without a map: *Inverse Prob.*, **14**, 427–440.
- Scales, J. A., 1996, Uncertainties in seismic inverse calculations, *in* Jacobson, B. H., Mosegaard, K., and Sibani, P., Eds., *Inverse methods*: Springer-Verlag, 79–97.
- Scales, J. A., and Snieder, R., 1997, To Bayes or not to Bayes?: *Geophysics*, **62**, 1045–1046.
- Sen, M., and Stoffa, P. L., 1995, *Global optimization methods in geophysical inversion*: Elsevier Science Publishers.
- Smith, M. L., Scales, J. A., and Fischer, T. L., 1992, Global search and genetic algorithms: *The Leading Edge*, **11**, 22–26.
- Tarantola, A., 1987, *Inverse problem theory*: Elsevier Science Publishers.
- Tarantola, A., and Valette, B., 1982, Inverse problems = quest for information: *J. Geophys.*, **50**, 159–170.
- Vinther, R., and Mosegaard, K., 1996, Seismic inversion through tabu search: *Geophys. Prosp.*, **44**, 555–570.
- Wetherill, G. B., 1982, *Elementary statistical methods* (third edition): Chapman and Hall.