ADAPTIVE KERNELS AND OCCAM'S RAZOR IN INVERSION PROBLEMS

R. FISCHER, W. VON DER LINDEN AND V. DOSE
Max-Planck-Institut für Plasmaphysik, EURATOM Association
*POB 1533, D-85740 Garching, Germany**

Proceedings of the Maximum Entropy Conference 1996, South Africa

Abstract.
Following the adaptive-kernel methods in density estimation theory we extend the quantified maximum entropy concept by locally varying correlations. The smoothing property of the adaptive kernels are determined self-consistently in the framework of Bayesian probability theory with Occam's razor as driving force for the simplest model consistent with the data. The power of the adaptive-kernel approach, the suppression of ringing and noise fitting, is demonstrated on the density estimation problem of the Old Faithful geyser, on a mock data set representing common physical inversion problems and on a real-world problem.

Key words: Inverse Problem, Ringing, Over-Fitting, Adaptive Kernels, Occam's Razor

1. Introduction

Bayesian inference in ill-posed inversion problems with entropic prior information suffers from the occurrence of spurious sources due to over-fitting of the data [1, 2, 3, 4, 5, 6]. The "ringing" is intrinsic to the entropic prior which has no local smoothing [6, 7]. In view of the many successes of the entropic prior in the regularization of ill-posed inversion problems many attempts are made to reduce spurious noise fitting by the introduction of spatial correlations. Gull [1] supposed that the default model m of the entropic prior should be generated from a blurred image of hidden variables $\tilde{m}$: $m = B * \tilde{m}$. Skilling [3] introduced spatial smoothing into the quantified maximum entropy formalism (QME) via a hidden image which is blurred with a global correlation function. Due to the failure of the entropic prior in the continuum limit Sibisi and Skilling recently superposed QME in terms of Dirichlet priors but adheres to the global smoothing blur function [8]. An alternative approach to take smoothness into account has been proposed by Silver [4] what he called Quantum Statistical Inference (QSI). In addition to the features of classical entropy, like positivity and convex optimization, quantum entropy enforces correlations, such as local smoothing.

Local smoothness of the image f may be important prior information which is not included in standard maximum entropy (ME) methods. Common examples are the image reconstruction, density estimation and interpolation where continuous derivatives are expected. Puetter [5] incorporates local smoothness by the convolution of a local smoothing kernel with the distribution of a "pseudo" image. It is much like the "preblur on a hidden image" [3] but with local blurring width. Puetter calls the local blurring kernel "pixon" to point on the intimate relation of the discretization on a pixel grid with pixel widths small enough to fit the data and large enough to minimize the number of degrees of freedom (DOF). The results of the pixon approach are very well looking but Puetter fails in the translation of his Bayesian formulation into numerical codes. He uses hand-tuned non-unique pixon widths small enough to yield $\chi^2 = N$.

The local kernel approach is a standard technique in density estimation theory with adaptive (multivariate) kernels [9]. For the kernel functional and the determination of the window width plenty of sophisticated proposals are made [9, 10]. For a unique decision for the kernel functional and the determination of the window width we use hierarchical Bayes arguments.

Effectively, with the introduction of correlations the number of effective degrees of freedom (eDOF) in the reconstruction is reduced. The driving force to high correlations and low eDOF enters the Bayesian probability theory via Occam's Razor. The best hypothesis consistent with a data set is the simplest one [11, 12].

Section 2 introduces the adaptive-kernel approach with Bayesian probability theory. In Section 3 we present examples of density approximation and image reconstruction and compare the results with recently published other approaches. Section 4 concludes.

2. Adaptive kernels

Density estimation, image reconstruction or interpolation in their most general form is an inverse problem, where the data D are related to the image f via the model A and distorted by noise N: $D(x) = A(f) + N(x)$. We impose smoothness on f through a convolution of a smoothing kernel functional B with a hidden ("latent") density h:

$$f(x) = \int dy B\left(\frac{x-y}{b(y)}\right) h(y) \quad . \tag{1}$$

Note the local blur width in the hidden space. The shape and width of the kernel B are to be determined by Bayesian methods. Knowing the normalization, the mean and the standard deviation of B, the kernel functional form has to be Gaussian according to the ME principle [13]:

$$B\left(\frac{x-y}{b(y)}\right) = \frac{1}{\sqrt{2\pi}b(y)} \exp\left[-\frac{1}{2}\left(\frac{x-y}{b(y)}\right)^2\right] . \tag{2}$$

We do not expect the shape of the kernel functional to matter greatly but in some applications the infinite range of a Gaussian may not be advantageous. For example, if we expect to be sharp edges in an image the prior information would favor a compact kernel like truncated paraboloids [5].

The width, $b(y)$, of the kernel, B, has to be determined via model selection. The reason for this is the strong correlation between the blur-width distribution b

and the complexity of our descriptive model. As will be described quantitatively below, the effective number of parameters (eDOF) describing the image is given by b. This can be easily seen in the limits $b \to \infty$ and $b \to 0$. For $b \to \infty$ the image is described solely by the mean of the density h with an eDOF of one. Whereas for $b \to 0$ the kernel is the δ-functional which results in $f(x) = h(x)$. We obtain the conventional result where the image is reconstructed pointwise with an eDOF of N, where N is the number of cells of h. Blur widths between the two limits result in local smoothing with locally varying eDOF. Gull [1] terms b as the "model-blur point spread function" indicating the model character of the b-parameters. We have to ask for the blur widths b which are enforced by the data, namely the posterior probability density $P(b|D)$. Two terms will compete: the likelihood probability density will favor a large eDOF, and thus small values of b, and we expect Occam's razor to penalize a large eDOF.

The posterior probability density of b given the data D is:

$$\begin{aligned} P(b|D) &= \int \mathcal{D}h P(b,h|D) \\ &= \int \mathcal{D}h \frac{P(D|b,h)P(h|b)P(b)}{P(D)} \quad . \end{aligned} \tag{3}$$

For the three examples presented below different likelihood probability densities $P(D|f(b,h))$ are required: For the density estimation problem of the Old Faithful geyser the likelihood consists of a product of densities, f, at the N_D eruption durations, D_i:

$$P(D|f(b,h)) = \prod_{i=1}^{N_D} f(D_i) \quad . \tag{4}$$

The maximum likelihood solution subject to the normalization constraint $\sum_{i=1}^N f_i = 1$ in terms of the digitized density, f_i, is

$$f_i^{ML} = g_i/N_D \quad , \tag{5}$$

where g_i is the number of eruptions in duration window i. In the case of normal distributed uncertainties or in the case where only the variance of the uncertainties can be estimated, the likelihood is given by the common misfit $\chi^2 = \sum_{i=1}^{N_D}((D_i - A_i(f))/\sigma_i)^2$, namely $P(D|b,h) \propto \exp(-0.5\chi^2)$. For the prior probability density $P(h|b)$ we use the entropic prior constraining positivity in f which is independent of b [3]:

$$P(h|b,\alpha) = P(h|\alpha) = \left(\prod_i (\frac{\alpha}{2\pi h_i})^{\frac{1}{2}} \right) \exp(\alpha S(h)) \quad , \tag{6}$$

with the entropy

$$S(h) = \sum_i^N (h_i - m_i - h_i \ln(\frac{h_i}{m_i})) \quad . \tag{7}$$

The prior $P(h)$ has to be determined by marginalization of the regularization parameter α [6]. For simplicity we use the historic ME approach which yields a criterion for α [2]. In our previous papers it is shown that historic ME provides quite the same regularization compared with the marginalization over α, whereas the evidence approximation tends to over-fit the data.

At any rate, we will see that the importance of the prior is strongly reduced due to the reduced eDOF. The range over which α has to be varied to observe significant changes in the image exceeds by far the differences of α of the various approaches.

For the prior probability $P(b)$ positivity constraints and scale invariance would require Jeffreys' prior in N dimensions. But we know the scale of the b's: the upper limit of each kernel width b_i is given by the total image range (overall smoothing) and the lower limit is given by the value of b_i which gives vanishing contributions to neighboring image cells (no smoothing). In addition we favor a blur width distribution where blur widths from neighboring cells differ by no more than one cell width. This enforces smooth variations in b. Otherwise the data would be artificially discriminated into background and signals, although in some cases this may be advantageous. Thus we chose the Gaussian prior for the derivative of b, measured in cell units, with a standard deviation of one for $b \in [b_{min}, b_{max}]$:

$$P(b) \quad \propto \quad \begin{cases} \exp(-0.5 \sum_i (b_i - b_{i-1})^2) & : \quad b \in [b_{min}, b_{max}] \\ 0 & : \quad \text{elsewhere} \quad . \end{cases} \tag{8}$$

It appears that the prior on b is more important than the prior on h. For example, the constant prior for b results in images with greater details and larger uncertainties. Apparently, the additional smoothing constraint on b is necessary to regularize the ill-conditioned determination of the overlapping kernels.

The integration in Eqn. 3 is done in two ways: with steepest-descent approximation (SDA) [1] and with Monte-Carlo importance sampling (see for example [14]). With SDA we obtain the maximum-a-posteriori (MAP) image and the curvature of the posterior probability density at the MAP image and with MCI we get the mean values of b and the variance of the total probability density. The results obtained with both methods are compared to obtain information about weak data constraints and about the erroneous approximations of the steepest-descent method [6].

The factor which favors the simplest model is *Occam's razor*. Bayes theorem does recognize simplicity as one component of the inference. The quantity that governs Occam's razor is the volume of the prior probability covered by the high-likelihood region picked by the data [11]. Two cases have to be distinguished: the models differ in the number of parameters (DOF) or, as in the present case of the adaptive kernels, the models have the same number of DOF but differ in the values of additional model parameters which determine the eDOF. In the former case the prior probability density of the parameters depends critically on the manifold of the parameter space and its volume. The prior probability competes with the likelihood probability. In the latter case the prior probability of the parameters varies only via the explicit dependency on the model parameters. The volume of the hypothesis space is constant. Occam's razor is solely determined by the volume of the likelihood probability which depends implicitly on the model parameters. In the adaptive-kernel approach the DOF is given by the number of image cells, N, and the model parameters b define the complexity of the model (eDOF).

A measure for the eDOF of the model parameters, b, is given by the sum of the square roots of the eigenvalues of $B^T B$:

$$eDOF \quad = \quad \sum_i \sqrt{eigenvalue_i(B^T B)} \quad . \tag{9}$$

The largest eigenvalue is one due to normalization of B. For $b \to 0$ the eDOF is N and for $b \to \infty$ the eDOF is one. The large eigenvalues define eigenvectors which are essential for describing the data and the small eigenvalues define eigenvectors describing insignificant contributions due to noise.

The discrimination between significant data constraints and insignificant noise contributions is supported by the locality of the model-blur b. The common case is the occurrence of both image regions containing significant data information and image regions carrying only informations about the overall background. In the former case we need nearly pointwise reconstructions and in the latter case the information given by the data can be described only by a few eDOF.

For the best estimate of the kernel-width distribution, b^*, we can use the MAP, $max_b P(b|D)$, or the mean, $< b_i >= \int d^N b \; b_i P(b|D)$. If the probability density $P(b|D)$ is not sharply peaked at one particular kernel-width distribution, we have to do the full marginalization over b.

The posterior probability density for f given the "best" estimate for the blurring kernel widths, b^*, is

$$\begin{aligned} P(f|D,b^*) \quad &\propto \quad \int \mathcal{D}h P(D|b^*,f,h) P(f|h,b^*) P(h) \\ &\propto \quad \int \mathcal{D}h P(D|\hat{f}(h,b^*)) \delta(f - \hat{f}(h,b^*)) \exp(\alpha S(h)) \quad . \end{aligned} \tag{10}$$

The Bayesian analysis yields the entire probability distribution which is however somewhat complicated due to the presence of the delta function. This is of minor importance since our interest usually focuses on expectation values of some functionals of f. We represent the posterior probability distribution by its moments:

$$\begin{aligned} < f_l^\nu > \quad &= \quad \int \mathcal{D}f f_l^\nu P(f|D,b^*) \\ &\propto \quad \int \mathcal{D}h \hat{f}_l^\nu P(D|\hat{f}(h,b^*)) \exp(\alpha S(h)) \quad . \end{aligned} \tag{11}$$

We are mainly interested in the posterior expectation, $< f_l >$, and point estimates of the variance of the image, $var(f_l) =< f_l^2 > - < f_l >^2$. The variance (uncertainty) of the image f is correlated with the kernel width. To illustrate this we consider an example where only the smoothing properties of the kernels are addressed. The mock data should consist of N samples of a constant, distorted by a constant noise level σ. In the limit $b \to 0$ with no smoothing the variance of the uncorrelated image points is σ^2. In the limit $b \to \infty$ with overall smoothing, the data are described by their mean value and the variance is σ^2/N, the variance of the mean. Between the to limiting cases, the variance is approximately given by σ^2/b. Thus, unstructured data regions which can be described with large kernel widths result in reconstructions with a low error level.

3. Examples

We illustrate the performance of the adaptive-kernel approach using the textbook density-estimation example of the eruptions of the Old Faithful geyser. Figure 1 shows the raw data for the duration of 107 eruptions displayed as a histogram using 100 bins. The best estimate (the mean of the posterior) for the density obtained with adaptive kernels is shown as solid line. The error bars correspond to one standard deviation. The second moment of the posterior probability density is especially for durations with only a few eruptions not sufficient for describing the posterior since it is significantly skew.

To get an impression of the quality of the estimate we have to compare it with recently published results. Silver [15] compares the results for the Old Faithful geyser obtained with QSI with the classical ME estimate. The classical ME estimate consists of spikes at the positions of the data because it assumes no correlations between neighboring bins in f. In QSI the smoothness properties are determined by a global Lagrange multiplier β and the degree of the differential smoothing operator. The optimal QSI estimate looks very pleasant and is very close to the estimate obtained with adaptive kernels.

Like QSI, the kernel approach provides point estimates of errors on f which are about the same as for the QSI result. The QSI approach is fundamentally new and interesting, especially the important "dual-variable" feature [15], but the adaptive-kernel approach is more comprehensible and easy to implement in existing ME-codes. Furthermore, the present form of the QSI approach does not consider the adaptive resolution concept.

The best density estimate for the eruption duration obtained with the Dirichlet prior in conjunction with global preblur is more structured [8]. In addition to the strong tendency of the Dirichlet prior towards zero the more structured image may be a fact of the globality of the blur width. To fit the data the maximal blur width is determined by the sharpest significant structure. The smoothness property and thus the suppression of ringing is limited to this maximal blur width. Our observation is that the results depend only slightly upon the prior probability once the adaptive resolution concept is used.

In this proceedings you find also the Old-Faithful density-estimation problem solved with an adaptive-spline approach. The result looks quit similar to the one obtained with the adaptive-kernel approach. The adaptive splines may be identified with non-overlapping polynomial kernels with widths, positions and polynomial parameters determined by the data in the same way as the adaptive

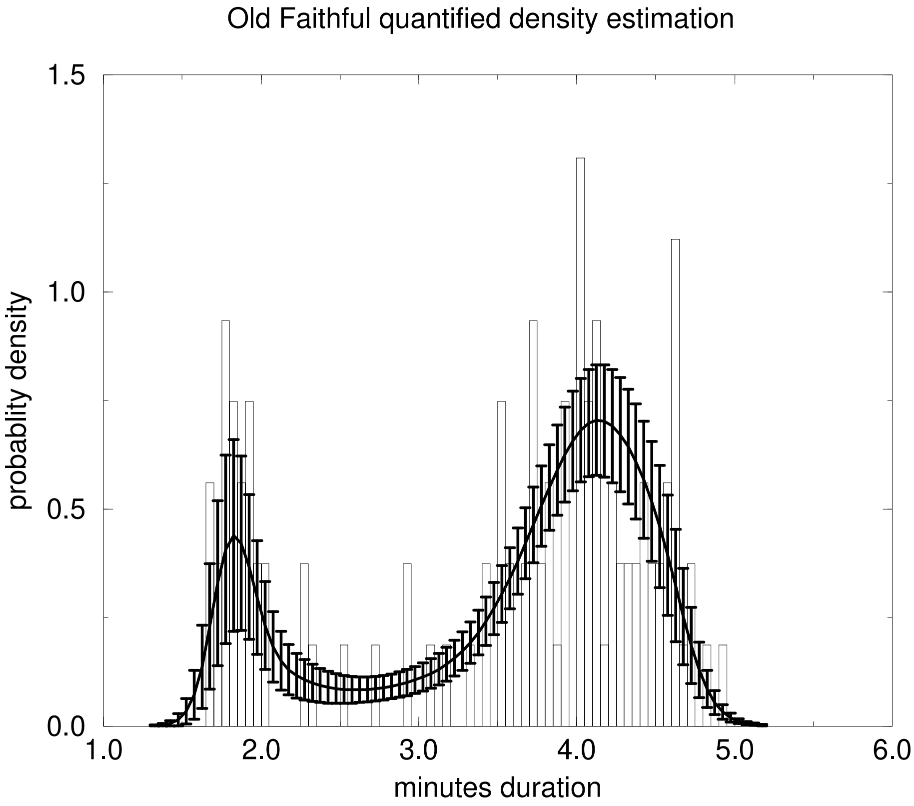

Figure 1: Old Faithful Eruptions - 107 measurements of the duration of geyser eruptions are displayed as a histogram with 100 bins. The solid line is the mean of the posterior density probability. The error bars indicating the $\pm\sigma$ uncertainty and the mean are obtained by MC integration.

Gaussian kernels. The main difference is that the spline approach deals explicitly with different numbers of parameters (DOF) where the adaptive-kernel approach has a constant parameter space. Furthermore the adaptive-spline approach is more difficult to be applied to problems with more than one dimension, like image reconstruction.

A density estimation problem, with a maximum likelihood solution consisting of δ-functionals, is a particular challenge for ME. The typical point spread functions involved in inversion problems are already locally smooth, so that additional local smoothing might be less important [15]. In our next example with a mock data set we show that even in the case of locally smooth point spread functions additional local smoothing is of great importance.

In Fig. 2 a mock data set is shown consisting of a Lorentzian and a step

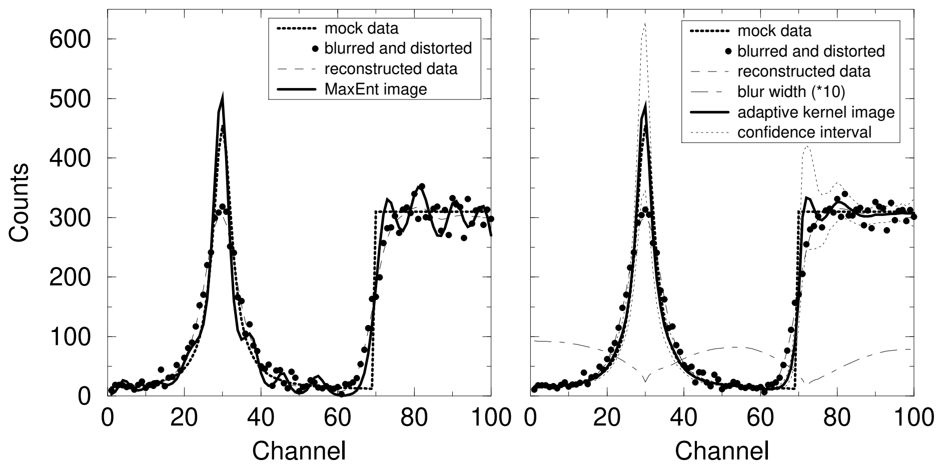

Figure 2: Mock data (dotted line) convolved with a Gaussian and distorted by Poisson noise (circles). Left panel: The ME deconvolution with α-marginalization (solid line) shows a frightening ringing level. Right panel: The adaptive-kernel approach results in a much smoother reconstruction. The ringing disappeared in the Lorentz region and is greatly reduced in the step region. The point estimates of the error on the reconstruction is shown as a confidence interval.

blurred by a Gaussian. The data were distorted by Poisson noise. This is a typical data set for various spectroscopy experiments: a spectral density consisting of a quasi-particle peak and the onset of continuum excitations superposed by a constant background. The step is introduced to study the worst conceivable case. The left panel depicts the reconstruction obtained by QME using the full Bayesian approach without any approximations. The regularization parameter α is marginalized as proposed by Strauss et al. [16] and elaborated by us [6]. The reconstruction looks frightening: the step is not reproduced properly and the image is superposed by considerable ringing. This is due to the fact that the image-grid used contains more DOF than the information content of the data can assign.

The right panel shows the result obtained with the adaptive-kernel approach with the respective point estimates of the errors (confidence interval). The reconstruction hardly deviates from the mock data. The ringing disappeared completely on the slopes of the Lorentzian and is greatly reduced at the step. The remaining Gibbs oscillation on top of the step is damped. This worst case is not encountered in real-spectrum restoration where spectra don't exhibit such strong discontinuities. In Fig. 2 the kernel blur width b, scaled by a factor of 10, is also shown. Whenever the data reveal significant structures the blur width is reduced. Note the correlation between the blur width and the confidence interval. Large blur widths result in small confidence intervals due to averaging over many data points improving the statistics considerably. The eDOF calculated with Eqn. 9 is merely 7.5 as compared to 100 used before.

A complex real-world problem is the following: In order to deconvolute a spectral density distribution measured by X-ray photoemission spectroscopy the apparatus function is needed. Generally, it is not known but it can be determined if we first measure a spectrum of known density distribution which is depicted in the left panel of Fig. 3. The QME approach without spatial correlations results in an apparatus function with a considerable ringing level shown in the right panel of Fig. 3. For further deconvolutions this apparatus function is nearly worthless due to huge uncertainties concomitant with the ringing. The adaptive-kernel approach results in a much smoother apparatus function with considerably lower uncertainties. With this technique it has been possible, for the first time, to deconvolute point-spread-functions which are themselves deconvoluted form-free from noisy data.

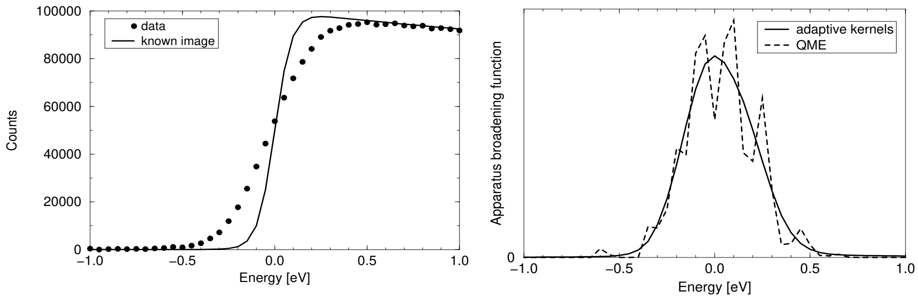

Figure 3: Deconvolution of the apparatus function (right panel) from a measured spectrum of known density (left panel). The QME approach without correlations results in an apparatus function with a considerable ringing level. The adaptive-kernel approach results in a much smoother apparatus function.

The extension of the adaptive-kernel approach on the deconvolution of an point-spread-function from a two-dimensional image is straightforward though the computational requirements might be huge.

4. Summary

We have shown that the intrinsic ringing level of ill-posed inversion problems can be reduced drastically by the introduction of adaptive resolution into Bayesian analysis. The concept of adaptive kernels provides local smoothness which eliminates the over-fitting and makes the result robust against noise corruption. The locality of the kernel widths takes account of the locality of the information content of the data. The complexity of a model is penalized self-consistently within the Bayesian framework. Occam's razor favors the smallest number of effective degrees of freedom supported by the data constraints. The adaptive-kernel approach provides reasonable point estimates of errors on the reconstruction by averaging over many data points depending on the local kernel width. Examples from the field of density estimation and spectral deconvolution show the advantageous of the adaptive resolution concept. In order to reconstruct the point-spread-function of an apparatus the adaptive-kernel approach is essential for further deconvolutions. We can now honestly use the statement that the Bayesian approach allowing for adaptive resolution yields the most uncommittal reconstruction compatible with the data.

References

1. S. Gull, "Developments in maximum entropy data analysis," in *Maximum Entropy and Bayesian Methods*, J. Skilling, ed., p. 53, Kluwer Academic Publishers, Dordrecht, 1989.
2. J. Skilling, "Classic maximum entropy," in *Maximum Entropy and Bayesian Methods*, J. Skilling, ed., p. 45, Kluwer Academic Publishers, Dordrecht, 1989.
3. J. Skilling, "Fundamentals of maxent in data analysis," in *Maximum Entropy in Action*, B. Buck and V. Macaulay, eds., p. 19, Clarendon Press, Oxford, 1991.
4. R. Silver, "Quantum statistical inference," in *Maximum Entropy and Bayesian Methods*, G. Heidbreder, ed., Kluwer Academic Publishers, Dordrecht, 1996. to be published.
5. R. Puetter, "Pixon-based multiresolution image reconstruction and the quantification of picture information content," *Int. J. Image Sys. & Tech.*, **6 (4)**, p. 314, 1995.
6. R. Fischer, W. von der Linden, and V. Dose, "On the importance of α marginalization in maximum entropy," in *Maximum Entropy and Bayesian Methods*, R. Silver and K. Hanson, eds., Kluwer Academic Publishers, Dordrecht, 1996. to be published.
7. W. von der Linden, R. Fischer, and V. Dose, "Evidence integrals," in *Maximum Entropy and Bayesian Methods*, R. Silver and K. Hanson, eds., Kluwer Academic Publishers, Dordrecht, 1996. to be published.
8. S. Sibisi and J. Skilling, "Prior distributions on measure space," *J. R. Statist. Soc. B*, **59**, to appear.
9. B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.
10. A. Izenman, "Recent developments in nonparametric density estimation," *J. Amer. Statist. Ass.*, **86**, p. 205, 1991.
11. E. Jaynes, *Probability Theory: The Logic of Science*. unpublished.
12. A. Garrett, "Ockham's razor," *Physics World*, p. 39, 1991.
13. L. Mead and N. Papanicolaou, "Maximum entropy in the problem of moments," *J. Math. Phys.*, **25 (8)**, p. 2404, 1984.
14. J. Ruanaidh and W. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*, Springer, New York, 1996.
15. R. Silver and H. Martz, "Applications of quantum entropy to statistics," *Journal of the American Statistical Association, 1994 Proceedings Toronto*, pp. 61–70, 1994.
16. C. Strauss, D. Wolpert, and D. Wolf, "Alpha, evidence, and the entropic prior," in *Maximum Entropy and Bayesian Methods*, G. Heidbreder, ed., Kluwer Academic Publishers, Dordrecht, 1996. to be published.