Journal of Mathematical Imaging and Vision 11, 27–43 (1999)

A Full Bayesian Approach to Curve and Surface Reconstruction

DANIEL KEREN
Department of Computer Science, The University of Haifa, Haifa 31905, Israel
dkeren@cs.haifa.ac.il

MICHAEL WERMAN
Institute of Computer Science, The Hebrew University, Jerusalem 91904, Israel
werman@cs.huji.ac.il

Abstract. When interpolating incomplete data, one can choose a parametric model, or opt for a more general approach and use a non-parametric model which allows a very large class of interpolants. A popular non-parametric model for interpolating various types of data is based on regularization, which looks for an interpolant that is both close to the data and also "smooth" in some sense. Formally, this interpolant is obtained by minimizing an error functional which is the weighted sum of a "fidelity term" and a "smoothness term".

The classical approach to regularization is: select "optimal" weights (also called hyperparameters) that should be assigned to these two terms, and minimize the resulting error functional.

However, using only the "optimal weights" does not guarantee that the chosen function will be optimal in some sense, such as the maximum likelihood criterion, or the minimal square error criterion. For that, we have to consider *all* possible weights.

The approach suggested here is to use the full probability distribution on the space of admissible functions, as opposed to the probability induced by using a single combination of weights. The reason is as follows: the weight actually determines the probability space in which we are working. For a given weight λ, the probability of a function f is proportional to $\exp(-\lambda \int f_{uu}^2\, du)$ (for the case of a function with one variable). For each different λ, there is a different solution to the restoration problem; denote it by f_λ. Now, if we had known λ, it would not be necessary to use all the weights; however, all we are given are some noisy measurements of f, and we do not know the correct λ. Therefore, the mathematically correct solution is to calculate, for every λ, the probability that f was sampled from a space whose probability is determined by λ, and average the different f_λ's weighted by these probabilities. The same argument holds for the noise variance, which is also unknown.

Three basic problems are addressed is this work:

- Computing the MAP estimate, that is, the function f maximizing $\Pr(f/D)$ when the data D is given. This problem is reduced to a one-dimensional optimization problem.
- Computing the MSE estimate. This function is defined at each point x as $\int f(x)\Pr(f/D)\mathcal{D}f$. This problem is reduced to computing a one-dimensional integral.
  In the general setting, the MAP estimate is not equal to the MSE estimate.
- Computing the pointwise uncertainty associated with the MSE solution. This problem is reduced to computing three one-dimensional integrals.

Keywords: Bayesian interpolation, regularization, hyperparameters

1. Introduction

In many areas of science and engineering, regularization [38, 40] is used to reconstruct objects from partial data. In computer vision, some references are [3, 13, 14, 35, 36]. Reconstruction of surfaces from partial data has been studied in many other fields, for example petroleum exploration [29], geology [4], electronics [2], estimation of the gravitational field of the earth [20], and medical imaging [21, 22]. In the field of maximum entropy, a similar idea is used for reconstruction of missing or corrupted data, as in image restoration [6, 9, 30–32].

The data D can be sparse, e.g., the height of a small number of points on a surface, or dense but incomplete, e.g., the case of optical flow and shape from shading [12] where data is available at many points but consists of the function's or its derivative's value in a certain direction only. The first difficulty in solving this problem stems from the multitude of possible solutions, each satisfying the partial data; which one should be chosen?

Regularization overcomes these difficulties by choosing among the possible objects one which approximates the given data and is also "smooth". This embodies an important assumption—that the "smoother" the object, the more probable it is. Formally, a *cost functional* $M(f)$ is defined for every object f by $M(f) = D(f) + \lambda S(f)$, where $D(f)$ measures the distance of f from the given data, $S(f)$ measures the smoothness of f, and $\lambda > 0$ is a parameter. The f chosen is the one minimizing $M(\ )$.

In the one-dimensional case, one can minimize $M(f) = \sum_{i=1}^{n} \frac{[f(x_i)-y_i]^2}{2\sigma^2} + \lambda \int_0^1 f_{uu}^2\, du$. In the two-dimensional case, one can minimize $M(f) = \sum_{i=1}^{n} \frac{[f(x_i,y_i)-z_i]^2}{2\sigma^2} + \lambda \int_0^1 \int_0^1 (f_{uu}^2 + 2f_{uv}^2 + f_{vv}^2)\, du\, dv$.

The Bayesian interpretation of this approach is: we are given the data D and want to find the function f which maximizes $\Pr(f/D) \propto \Pr(D/f)\Pr(f)$. Assuming a Gaussian noise model with variance σ^2, $\Pr(D/f) \propto \frac{1}{\sigma^n} \exp(-\frac{1}{2\sigma^2} \sum [f(x_i) - y_i]^2)$. Adopting a physical model, it is common to define $\Pr(f) \propto \exp(-\lambda \int f_{uu}^2\, du)$. Hence $\Pr(f/D) \propto \exp(-M(f))$, and the function minimizing $M(\ )$ maximizes the likelihood. Since the model is Gaussian, the MAP function is also the MSE function.

We will assume from now on that the functions are defined on the interval $[0, 1]$, or the unit square $[0, 1] \times [0, 1]$, and will usually omit those limits in the integrals. It is also possible to integrate the smoothness term $f_{uu}^2 + 2f_{uv}^2 + f_{vv}^2$ over all the plane, but we are interested in dealing with functions which are defined only on a bounded subset, so we choose to compute the integral on the unit interval/square. This is not a severe limitation for computer vision purposes.

The question is, how does one choose λ and σ? There are various methods for doing that, and some are mentioned in the following section. Most regularization schemes we are familiar with choose one combination of weights and use them alone to interpolate the function.

However, this approach finds the maximum likelihood (MAP) estimate for the interpolant f only for a given λ and σ. But the MAP estimate should maximize the following:

$$\int_w \Pr(f/D, w)\Pr(w/D)\, dw$$

where w varies over the set of all possible weights.

The reason for this is as follows: we *do not know* what the real weights are; all we know is the probability of each set of weights. Therefore, the probability of a certain f is the sum of its probability for every choice of weights, multiplied by the probability of the weights. Let us give an analogy: suppose an (uneven) dice is given, and each of its sides contains a list of probabilities for ten tasks. The dice is tossed, and then one proceeds to randomly choose a task, using the probability list on the side of the dice which turned up. In order to compute the probability that this experiment results in the undertaking of task number five, for instance, it is not correct to consider only the probability of task number five on the side of the dice which has the highest probability of turning up! Naturally, the probability that the toss will lead to undertaking task number five is the sum of the probabilities of each side, multiplied by the probability of task number five on that side.

For the restoration problem we wish to solve, the sides of the dice are analogous to the different weights, and the functions are analogous to the different tasks. Clearly, it's incorrect to use only the weights with the highest probability.

If $\Pr(w/D)$ has some nice properties—for instance, it is unimodal, symmetric, and concentrated around the pair of weights $w_{\max}$ which maximize $\Pr(w/D)$—it may be reasonable to approximate $\int_w \Pr(f/D, w)\Pr(w/D)\, dw$ by approximating the integrand with a rectangular function around $w_{\max}$. However, the distribution $\Pr(w/D)$ can be complicated and

this approximation will then fail [24, 33]; see also an example of such a data set and the corresponding probability distribution it induces on the weights, in this paper (Figs. 12, 13).

In this paper, it will be shown how to find the function f maximizing $\int_w \Pr(f/D, w)\Pr(w/D)\,dw$. Two other problems which are addressed are computing the expectation of a linear functional on the function space, such as the value of the function at a point. This expectation is the minimal square error (MSE) function. Another problem is to compute the pointwise uncertainty associated with the MSE estimate.

These three quantities—the MAP, the MSE, and the uncertainty—are perhaps the three most important estimators for a statistical entity, and it is therefore very important to rigorously compute them.

These problems are addressed in Sections 3–5. The technical details are described in the appendices. Some examples are given, of 1 and 2D data and its interpolants.

In Appendix 1, we note that if one wants to find the "optimal weights"—that is, the pair of weights $w = \{\lambda, \sigma\}$ which maximizes $\Pr(w/D)$—this can be reduced to a one-dimensional optimization problem, although the optimization is over a two-dimensional hyperparameter space.

In the future, we hope to find efficient numerical algorithms to speed up the algorithms described here.

2. Previous Work

A very popular method for determining the smoothing parameter λ is Generalized Cross Validation, GCV (bootstrapping) [5, 40]. The idea is to choose a λ such that the data points will predict one another. Formally, a function $V_0(\lambda)$ is defined as follows: for each sample point (x_k, y_k), $1 \le k \le l$, f_k is defined to be the spline minimizing

$$\sum_{i \neq k}^{l} [f(x_i) - y_i]^2 + \lambda \int f_{uu}^2\,du$$

i.e., the spline interpolating all the data points but the kth. $V_0(\lambda)$ is then defined as $\sum_{k=1}^{l} [f_k(x_k) - y_k]^2$, and the λ chosen is the one minimizing $V_0(\,)$. This algorithm is called Ordinary Cross-Validation (OCV).

An improvement of this method is the GCV algorithm [5] which proceeds as follows. Since the f interpolating the l data points is a linear combination of the set $\{y_1, y_2 \ldots y_l\}$, there is a matrix $A(\lambda)$ satisfying

$$\begin{pmatrix} f(x_1) \\ . \\ . \\ . \\ f(x_l) \end{pmatrix} = A(\lambda) \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_l \end{pmatrix} \tag{1}$$

$A(\lambda)$ is used to define a modified version of V_0:

$$V(\lambda) = \sum_{k=1}^{l} w_k(\lambda)[f_k(x_k) - y_k]^2,$$

$$w_k(\lambda) = \left[\frac{1 - A_{kk}(\lambda)}{\frac{1}{n} T_r(I - A(\lambda))}\right]^2$$

(T_r stands for Trace and I for the $l \times l$ identity matrix). The λ chosen is the one minimizing $V(\,)$. After this, λ is used to estimate σ. In [37], a few methods for choosing the smoothing parameter are analyzed.

Bayesian model selection is another approach for choosing an "optimal" smoothing parameter. To the best of our knowledge, it was first suggested to apply Bayesian model selection to regularization in the pioneering work of Szeliski [33]. There, the following question is posed: *given the data D, what is the most probable value of the smoothing parameter λ?* More recent work in this direction was done by MacKay in [24], and [23], which contains an extensive study on approximations to the ideal Bayesian approach, which, as the author correctly notes, is difficult to implement.

Another method for choosing the smoothing parameter is presented in [10]. In [28], the behavior of the smoothing spline over a range of smoothing parameters is studied, and is then used to construct a confidence interval for the smoothing parameter.

The problem with methods that use a single set of weights is that the choice of the values of λ and σ is sometimes very sensitive to the data. Since these values are crucial to the shape of the fitted curve or surface, it turns out that sometimes a small change in the data drastically changes the shape of the fitted curve or surface (see Figs. 1, 2 for curve fitting and 4, 5 for surface fitting). Another problem is that, although it can be proved that GCV has some nice asymptotic properties, the choice of the "optimal" values of λ and σ is heuristic in nature. Nonetheless, the algorithm performs well in general and is widely used; there are very

sophisticated numerical methods for implementing the GCV algorithm.

Work which proceeds in a direction somewhat similar to the one given here is presented in [6, 32]. However, this work is in the realm of entropy and therefore the mathematical framework is rather different from ours; for instance, there is no analog to the concept of the MSE estimate, neither to the concept of the uncertainty of the solution. Moreover, these works assume an uninformative prior on the hyperparameter, while we determine it according to the data. While assuming an uninformative prior on the hyperparameter and integrating it out using this prior may well have a beneficial effect—for instance, in stabilizing the restoration procedure—it is inherently different from the Bayesian approach presented here.

Finally, recent work reported in [25, 26] concerns the problem of computing the MAP solution, in a Bayesian framework, by integrating over the space of smoothing parameters and noise. For "integrating out" these two parameters, a uniform prior on them is assumed, which, as noted, is very different from the prior used in this work.

3. Computing the MAP Estimate

In order to compute the MAP estimate, we have to maximize $\Pr(f/D)$ over all functions f. Using Bayes' rule, $\Pr(f/D) \propto \Pr(D/f)\Pr(f)$. In order to compute this, one needs to integrate over all values of λ, σ, resulting in

$$\int_0^\infty \sqrt{\lambda}\exp\left(-\lambda\int f_{uu}^2\,du\right)\mathrm{Prior}(\lambda)\,d\lambda \cdot \int_0^\infty \frac{1}{\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum[f(x_i)-y_i]^2\right)\mathrm{Prior}^n(\sigma)\,d\sigma$$

In this expression, the priors on λ, σ are general, and not unique to our problem. These are not the priors determined by the data (which are computed in Section 10). In this work, we have used either log-normal or flat priors; the choice of the prior had hardly any effect on the results.

Note that a given λ corresponds to the probability $\exp(-\lambda\int f_{uu}^2\,du)$ on the function space. The entire probability, over the whole space, therefore equals

$$\int \Pr(f)\mathcal{D}f = \int \exp\left(-\lambda\int f_{uu}^2\,du\right)\mathcal{D}f$$

and, using the change of variables $g = \sqrt{\lambda}f$, it is easy to see that this equals

$$\frac{1}{\sqrt{\lambda}}\int \exp\left(-\int f_{uu}^2\,du\right)\mathcal{D}f$$

it is therefore necessary to multiply the probability by $\sqrt{\lambda}$. This normalization ensures that the structure of the probability, for every λ, does indeed define a probability space. This is the explanation for the $\sqrt{\lambda}$ in the first integral in the equation for $\Pr(f/D)$.

The expression for $\Pr(f/D)$ has to be maximized over the space of admissible functions. Let us write it more compactly as $F_1(\int f_{uu}^2\,du)F_2(\sum[f(x_i)-y_i]^2)$, where $F_1(\alpha) = \int_0^\infty \sqrt{\lambda}\exp(-\lambda\alpha)\mathrm{Prior}(\lambda)\,d\lambda$ and $F_2(\beta) = \int_0^\infty \frac{1}{\sigma^n}\exp(-\frac{\beta}{2\sigma^2})\mathrm{Prior}^n(\sigma)\,d\sigma$.

Note that, obviously, $F_1(\,)$ and $F_2(\,)$ are monotonically decreasing in α and β respectively.

It is possible to turn this optimization problem to a one-dimensional optimization by setting $\int f_{uu}^2\,du$ to a constant α, and then minimizing $\sum[f(x_i)-y_i]^2$ over all functions f such that $\int f_{uu}^2\,du = \alpha$.

Using Lagrange multipliers, this problem transforms into one resembling "standard" regularization: find a λ such that the function f minimizing $\sum[f(x_i)-y_i]^2 + \lambda\int f_{uu}^2\,du$ satisfies $\int f_{uu}^2\,du = \alpha$, where λ is the Lagrange multiplier.

In Section 10.1, it is proved that the f minimizing $\sum[f(x_i)-y_i]^2 + \lambda\int f_{uu}^2\,du$ is given by $f(x) = (H_{x_1}(x)\ldots H_{x_n}(x))(A+\lambda I)^{-1}(y_1\ldots,y_n)^t$, where

$$H_x(\xi) = \begin{cases} 0 \le \xi \le x: & \dfrac{(x-1)\xi(x^2-2x+\xi^2)}{6} \\ x \le \xi \le 1: & \dfrac{x(\xi-1)(x^2+\xi^2-2\xi)}{6} \end{cases}$$

and $A_{i,j} = H_{x_i}(x_j)$. Let us denote the data vector $(y_1,\ldots,y_n)$ by Y. After some manipulations,

$$\int f_{uu}^2\,du = Y(A+\lambda I)^{-1}A(A+\lambda I)^{-1}Y^t$$

so, we have to find for which λ this expression equals α. Diagonalizing A by an orthonormal U, $UAU^t = D$, and denoting $Z^t = UY^t$, the expression for $\int f_{uu}^2\,du$ reduces to

$$\sum \frac{d_i Z_i^2}{(d_i+\lambda)^2}$$

where d_i are the diagonal elements of D. Finding a λ for which this equals α is fast, as this function is

monotonically decreasing in λ, and we can solve the problem by binary search.

After finding λ, we have to compute $\sum[f(x_i) - y_i]^2$, where f minimizes $\sum[f(x_i) - y_i]^2 + \lambda \int f_{uu}^2\, du$. As noted above (see also Section 10.1), this f equals $(H_{x_1}(x) \dots H_{x_n}(x))(A + \lambda I)^{-1} Y^t$. Therefore

$$\begin{aligned} f(x_i) &= \big(H_{x_1}(x_i) \dots H_{x_n}(x_i)\big)(A + \lambda I)^{-1} Y \\ &= (A_{i,1} \dots A_{i,n})(A + \lambda I)^{-1} Y^t \end{aligned}$$

and so

$$\begin{aligned} \sum[f(x_i) - y_i]^2 &= \|(f(x_1) \dots f(x_n)) - Y\|^2 \\ &= \|A(A + \lambda I)^{-1} Y^t - Y^t\|^2 \\ &= \|AU^t(D + \lambda I)^{-1} U Y^t - Y^t\|^2 \\ &= \|AU^t(D + \lambda I)^{-1} Z^t - Y^t\|^2 \end{aligned}$$

this expression can be computed fast since it involves inverting a diagonal matrix, and since AU^t needs to be computed only once.

Now, all that's left is to compute $F_1(\alpha)F_2(\beta)$. $F_1(\,)$ and $F_2(\,)$ are one-dimensional integrals with rather simple integrands, and can be computed fast (or perhaps stored in a table).

What remains is to maximize $F_1(\alpha)F_2(\beta)$ over α (recall that β is not a free parameter, as it is determined by α).

The algorithm therefore tries to maximize a function $C(\alpha)$ which is defined as follows:

1. compute $F_1(\alpha)$
2. compute the (single) λ_α which satisfies $\sum \frac{d_i Z_i^2}{(d_i + \lambda_\alpha)^2} = \alpha$. This is fast because, as noted, $\sum \frac{d_i Z_i^2}{(d_i + \lambda)^2}$ is monotonically decreasing in λ (A is positive definite, so $d_i > 0$).
3. define $\beta = \|AU^t(D + \lambda_\alpha I)^{-1} Z - Y\|^2$
4. compute $F_2(\beta)$
5. return $F_1(\alpha)F_2(\beta) = C(\alpha)$

and we have to maximize $C(\alpha)$ for $0 \le \alpha \le \int (f_{\text{interpolate}})_{uu}^2\, du$, where $f_{\text{interpolate}}$ is the interpolant which passes through the data points. This range covers all the relevant functions, because $f_{\text{interpolate}}$ is the interpolant of the type we're studying which maximizes $\int f_{uu}^2\, du$ (it corresponds to $\lambda = 0$).

This is a one-dimensional optimization problem, which we solve numerically. The solution is reasonably fast, taking a few seconds on a workstation.

4. Computing the MSE Estimate

An estimator which for some purposes is more useful than the MAP estimate is the MSE estimate. Its value at x is defined by $E_x = \int f(x) \Pr(f/D) \mathcal{D} f$.

In order to compute this integral, the following approach is taken. Let us define a probability structure $M_{\lambda,\sigma}$ on the space of admissible functions. In this space, we assume the measurement noise is σ, and the prior distribution of the function f is $\Pr(f) \propto \exp(-\lambda \int f_{uu}^2\, du)$. Under this probability, which is Gaussian, the MSE function, denoted $(f_{\text{opt}})_{\lambda,\sigma}$, is equal to the MAP function and there is a closed-form expression for it (given in the previous section). In Section 10.1, we show that

$$\begin{aligned} E_x &= \int f(x) \Pr(f/D) \mathcal{D} f \\ &= \int_\lambda \int_\sigma (f_{\text{opt}})_{\lambda,\sigma}(x) \Pr(\lambda, \sigma/D)\, d\lambda\, d\sigma \end{aligned}$$

The idea is to decompose the complicated probability structure over the function space to a weighed sum of simple (Gaussian) probability structures, over each of which we can easily calculate the desired integral.

In Section 10.1, $\Pr(M_{\lambda,\sigma}/D)$ is computed, and the following expression for E_x is derived:

$$\frac{\int \frac{1}{\sqrt{v}} |A + vI|^{-(1/2)} \big(H_{x_1}(x) \dots H_{x_n}(x)\big)(A + vI)^{-1} Y^t [Y(A + vI)^{-1} Y^t]^{(4-n)/2}\, dv}{\int \frac{1}{\sqrt{v}} |A + vI|^{-(1/2)} [Y(A + vI)^{-1} Y^t]^{(4-n)/2}\, dv}$$

where A and H_{x_i} are the same as in Section 3.

This is a closed-form expression, but it involves a one-dimensional integral whose computation is non-trivial due to the complicated form of the integrand. Currently, we are investigating ways to speed-up the computation of this integral, which computationally is the bottleneck of the algorithm suggested here.

5. Computing the Uncertainty Associated With the Interpolant

In [5, 15, 19, 24, 27, 33, 34, 39, 40], the problem of assigning a measure of uncertainty to the regularizing

interpolant is addressed. This is very important, because usually one wants not only to know the curve (surface) which is optimal in some sense, but also to know how reliable this curve (surface) is. We chose to extend the method suggested in [15], defining the uncertainty of the interpolant at the point x as

$$\int [f(x) - E_x]^2 \Pr(f/D) \mathcal{D} f$$

the details are given in Appendix 2. As was the case with E_x, we obtain a closed-form solution, but its computation is non-trivial.

Just as the MSE is dependent on the hyperparameters, so is the uncertainty. This is demonstrated in Fig. 10. Two nearly identical data sets result in the GCV algorithm choosing very different values of the hyperparameter λ, and this results not only in a very different MSE estimate, but also in very different uncertainty intervals (more details in Section 7).

6. The 2D Case

All the results in the previous sections have been extended to the 2D case (surface reconstruction). There is one technical difficulty to overcome: the computation of the two-dimensional functions which are the equivalent of the functions $H_x(\xi)$. That is, it is necessary to find functions $G_{x,y}(u, v)$ (also called *reproducing kernels*) which satisfy, for every two-dimensional function f, which satisfies some boundary conditions, the equality $(f, G_{x,y})_{2D} = f(x, y)$, where

$$(f, g)_{2D} = \int\int (f_{uu} g_{uu} + 2 f_{uv} g_{uv} + f_{vv} g_{vv})\, du\, dv.$$

As opposed to the one-dimensional reproducing kernels, which have a simple form (cubic splines), there is no known closed form expression for the 2D reproducing kernels. In [15, 16] this problem is addressed, and it is shown how to quickly compute the functions $G_{x,y}(u, v)$ to any desired accuracy using an approximation on a finite subspace. Due to the fact that the inner product $(f, g)_{2D}$ is nearly orthogonal on subspaces spanned by trigonometric functions, convergence is very fast, and a subspace of reasonably low dimension is good enough to compute the functions to a very high accuracy. We have implemented this approximation, and used it to restore 2D functions; see Figs. 4–7.

7. Examples

A simple pattern—one cycle of a sinusoidal function—was contaminated with Gaussian noise, with a variance equal to five percent of the amplitude, and then the resulting data was interpolated using the GCV algorithm and the methods suggested in the previous sections. The instability of the GCV is demonstrated by noting that changing the value of the data at a single point radically changes the shape of the fitted curve (Figs. 1 and 2). The MSE estimates for these two data sets are presented in Fig. 3. They were calculated using Eq. (8) of Appendix 1.

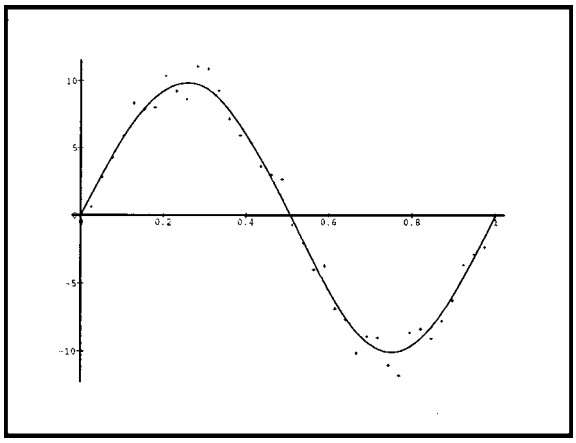

Figure 1. GCV chooses a "standard" value of λ, to interpolate sinusoidal data contaminated by Gaussian noise.

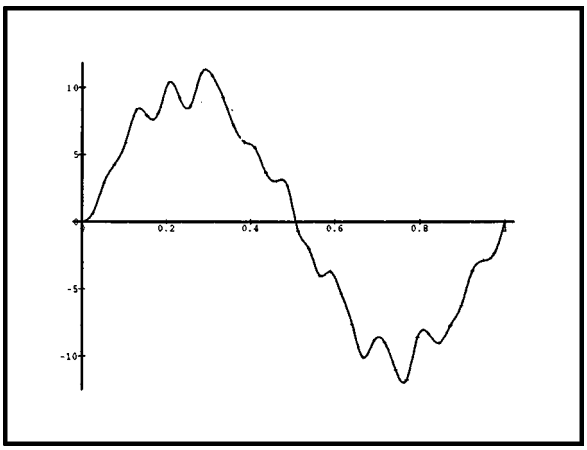

Figure 2. For a data set differing from that of Fig. 1 in only one point, GCV chooses a very small value of λ, resulting in a completely different fit.

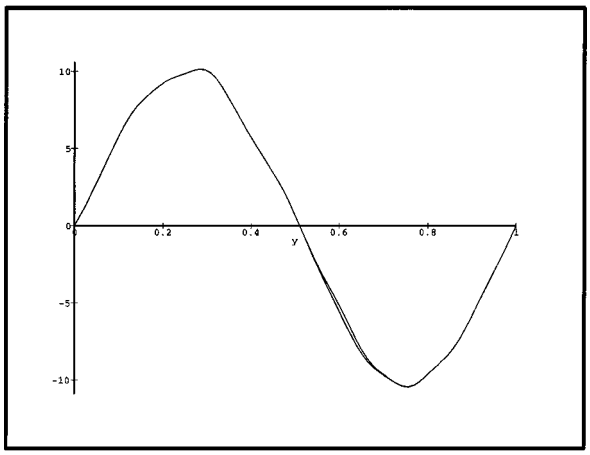

Figure 3. The MSE estimate for the data sets of Figs. 1 and 2, obtained using Eq. (8). Fits are almost identical.

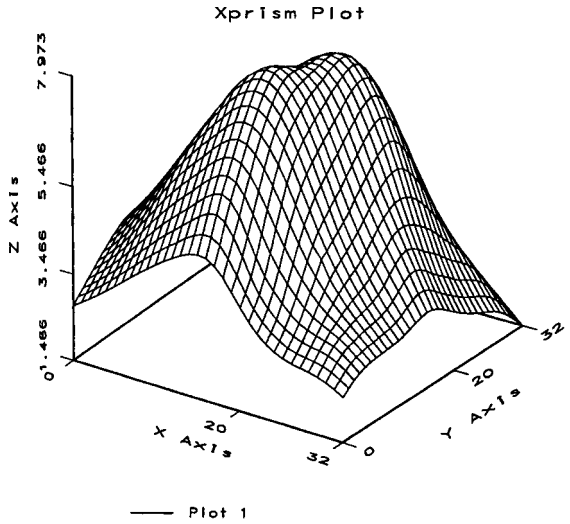

Figure 4. GCV reconstruction for the function $125x(1-x)y(1-y)$, contaminated by Gaussian noise. This is a "typical" reconstruction for such data.

We have run some tests on 2D data which was created by adding Gaussian noise with a variance of 0.1 to the function $125x(1-x)y(1-y)$. Figures 4 and 5 demonstrate how the GCV returns radically different results for two data sets which differ only in one point—this is because, just like in the case of the data in Figs. 1 and 2, this slight change caused GCV to choose a very different value of λ. Figure 6 shows the MSE estimate for the data of Fig. 4 (also using Eq. (8)). Applying Eq. (8) to the data of Fig. 5 results in a surface which is almost identical to that of Fig. 6.

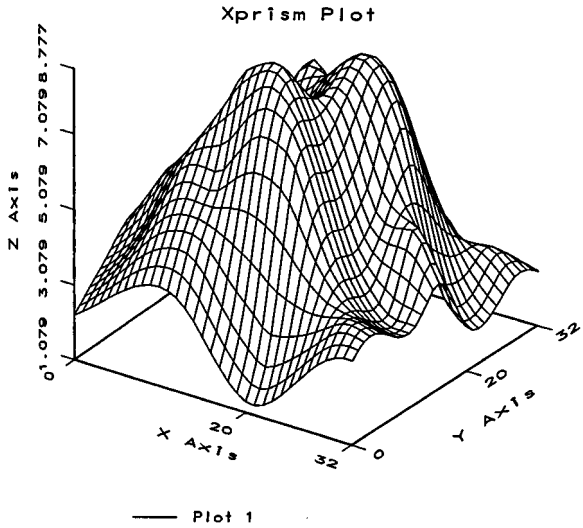

Figure 5. For a data set differing from that of Fig. 4 in a single point, GCV finds a radically different interpolation.

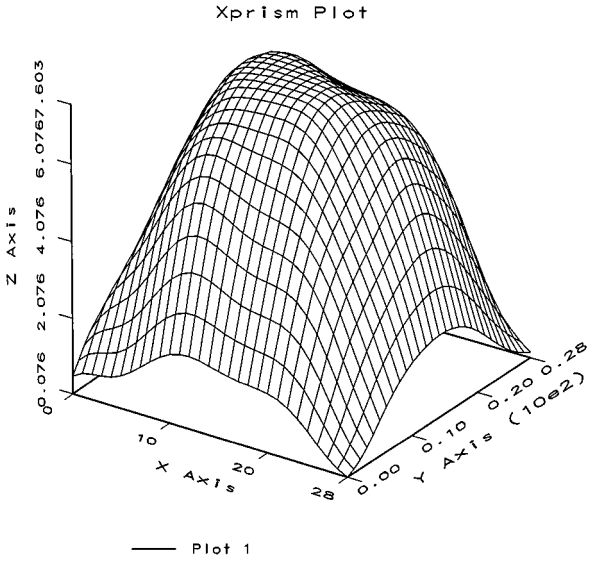

Figure 6. The MSE estimate for the data set of Fig. 4, obtained using Eq. (8).

In Fig. 7, the result of restoring real data is given; the points were sampled from a depth image of a human face, and Gaussian noise with a variance of 1 was added to them.

In Fig. 8, the two MAP reconstructions for the data sets of Figs.1 and 2 are given, with the data.

In Fig. 9, the MSE estimate and confidence intervals for a data set are given. The data is a sample of the x-coordinates of a hand-written word.

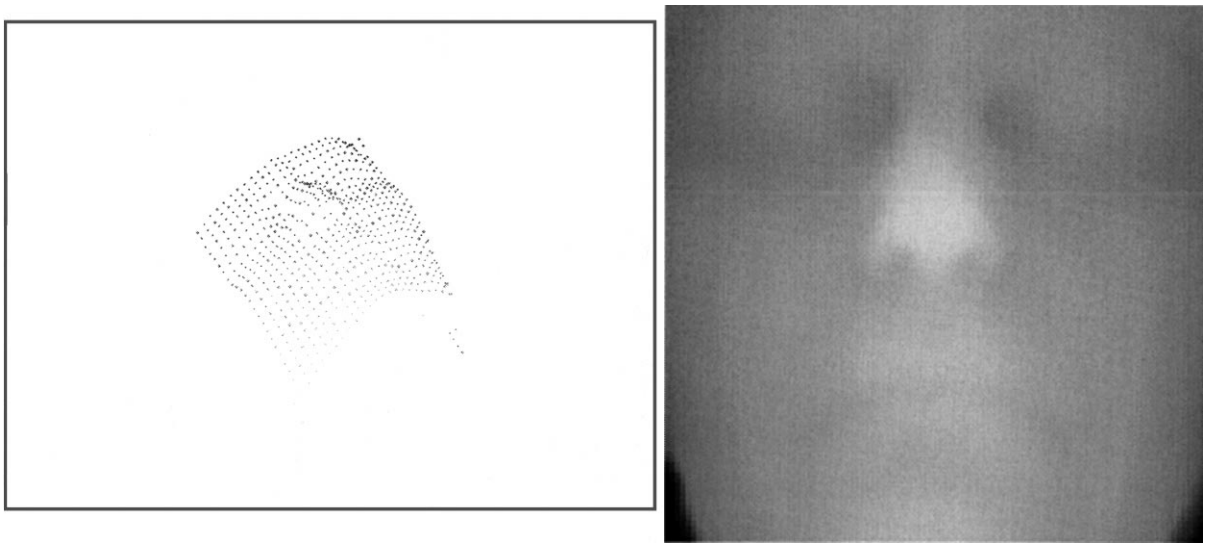

Figure 7. Data sampled from depth image of human face (left), and its reconstruction (right), obtained using Eq. (8) (a 25×25 sample from the 100×100 data is shown).

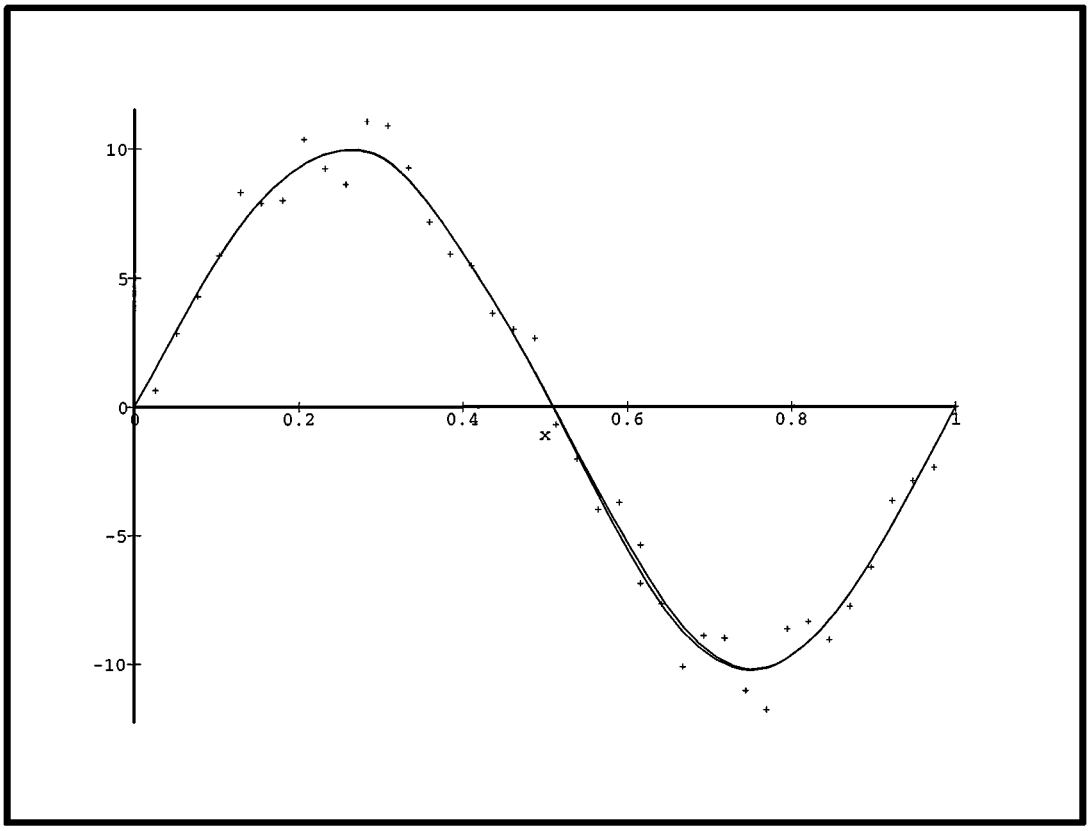

Figure 8. The suggested method for computing the MAP estimate, for the data sets of Figs. 1 and 2.

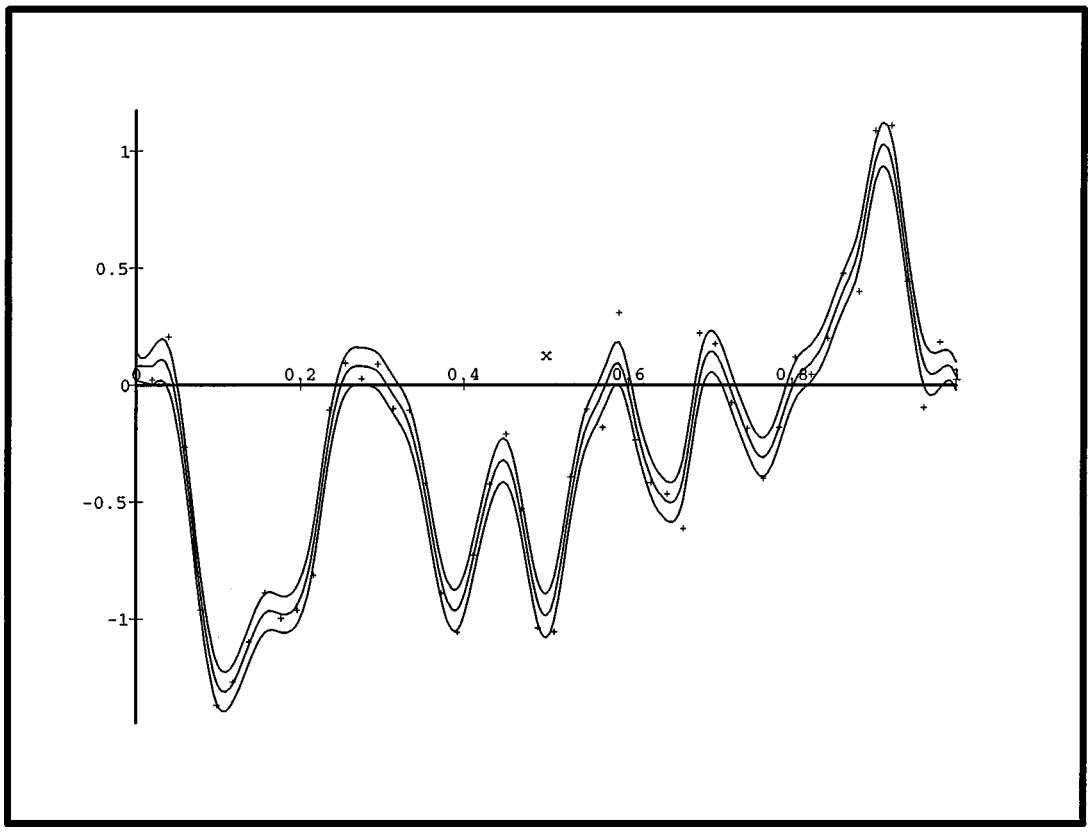

Figure 9. MSE function and confidence intervals for an evenly sampled data set. The little "+" signs are the data points, the middle curve is the MSE interpolant, and the upper and lower curves consist of the upper and lower confidence intervals, with a width of (pointwise) one standard deviation. The pointwise variance was computed using Eq. (9).

In Fig. 10, the importance of integrating over the different weights when computing the uncertainty intervals is demonstrated. For the data sets of Figs. 1 and 2, the GCV algorithm chose two very different values of λ, although the data sets are nearly identical (they slightly differ in one point only). The lower graph shows the value of the uncertainty, for the λ corresponding to the restoration in Fig. 1; the upper graph shows the corresponding quantity for the λ corresponding to the restoration in Fig. 2. Note the instability in the values of the uncertainty, which are caused by using only one hyperparameter.

In Fig. 11, the interpolant and confidence intervals are given for data unevenly sampled from a sinusoid with noise added to it. One can see that the uncertainty is larger in areas which are far from the sample points. The uncertainty at the endpoints is zero, because we constrain our functions to be zero at the endpoints (see Appendix 1).

Finally, we give an example which explains why one has to integrate over all the weights. In Fig. 12, two data sets are shown, superimposed. As one can see, they are almost identical. In Fig. 13, the (scaled) probability distribution for the weights λ, σ of one of the data sets is plotted. It has two distinct peaks, which are rather far apart; the location of the peaks correspond to the location of the most probable weights for the two data sets of Fig. 12. Therefore, the interpolants for the data sets of Fig. 12 which use only the most probable weights are drastically different, although the data sets are almost identical.

8. Conclusions and Further Research

This work suggests a straightforward and mathematically rigorous approach for solving three basic problems in curve and surface reconstruction, which are very common in many areas: finding the MAP

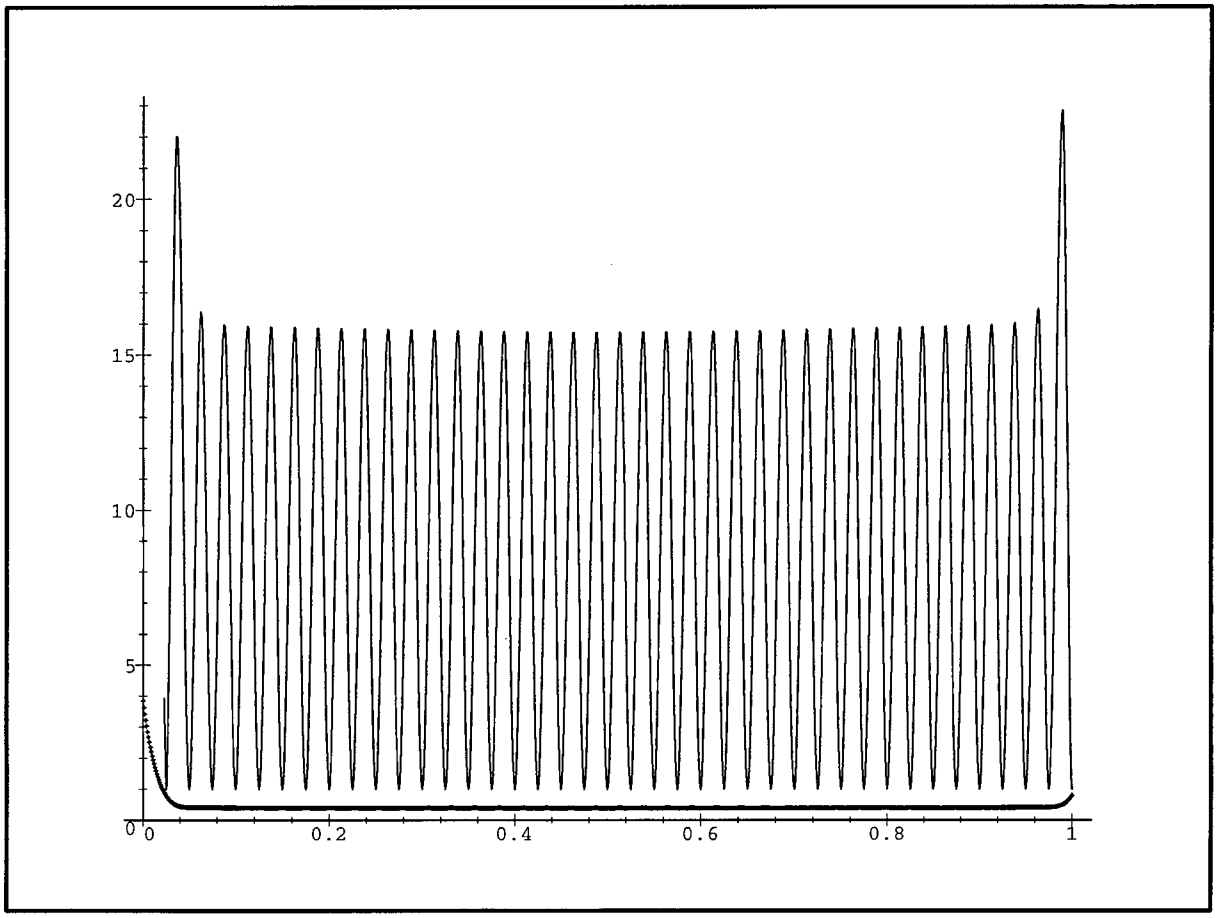

Figure 10. The height of the confidence intervals for the two different values of λ chosen by the GCV algorithm for the data sets of Fig. 1 (lower thick line) and Fig. 2 (sinusoidal).

interpolant, finding the MSE interpolant, and computing the uncertainty associated with the interpolant.

In the future, we hope to present algorithms for speeding up the computation of these three entities, as well as to expand the model to handle discontinuities. The problem of detecting and handling discontinuities in the data is especially important in the area of computer vision [7, 21, 22, 36]. For that, we plan to extend the Sobolev space to include functions with discontinuities.

Appendix 1: Computing $\Pr(\lambda, \sigma/D)$ and E_x

Call the model that assumes λ as a smoothing parameter and σ as the measurement noise $M_{\lambda,\sigma}$. In this model, $\Pr(f) \propto \exp(-\lambda \int f_{uu}^2\, du)$. Given a data set D, we compute $\Pr(\lambda, \sigma/D)$. Using Bayes rule:

$$\begin{aligned}\Pr(\lambda, \sigma/D) &= \frac{\Pr(D/\lambda, \sigma)\text{Prior}(\lambda, \sigma)}{\Pr(D)} \\ &\propto \Pr(D/\lambda, \sigma)\text{Prior}(\lambda, \sigma) \\ &= \frac{\int \Pr(D/f)\Pr(f/\lambda, \sigma)\mathcal{D}f}{\int \Pr(f/\lambda, \sigma)\mathcal{D}f}\text{Prior}(\lambda, \sigma)\end{aligned} \tag{2}$$

where the denominator is introduced to turn the distribution on the functions f into a probability, by normalizing its integral on the whole space to 1.

Since the data is given, it is the same for all models and can be eliminated from consideration. We have used the prior $\text{Prior}(\lambda, \sigma) = \lambda^{-\frac{5}{2}}$, for the following reason. Intuitively, spaces with a prior

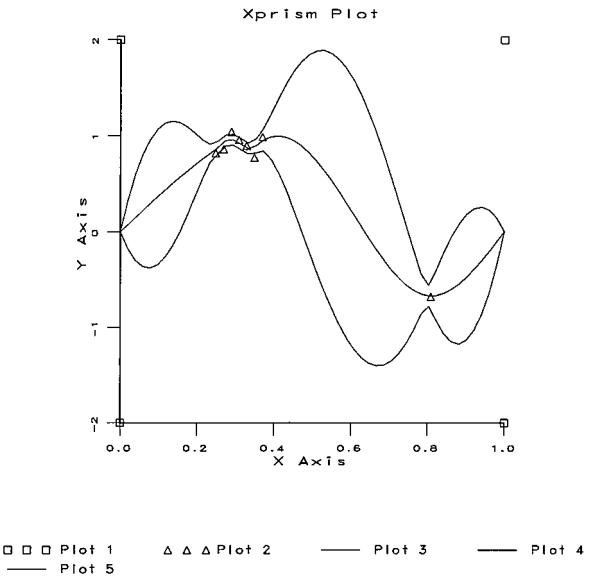

Figure 11. MSE function and confidence intervals for an unevenly sampled data set. The little triangles are the data points, the middle curve is the MSE interpolant, and the upper (lower) curves consist of the upper (lower) confidence intervals, with a width of (pointwise) one standard deviation. The pointwise variance was computed using Eq. (9).

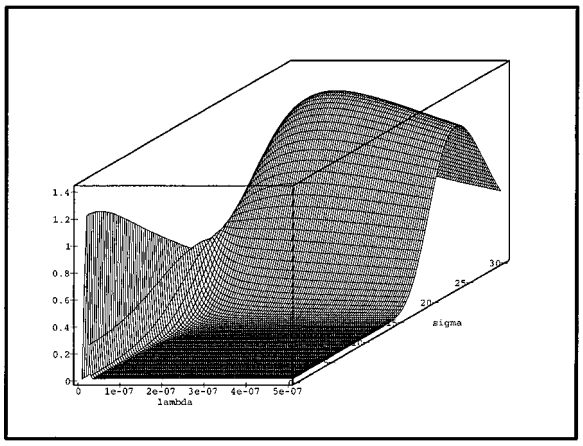

Figure 13. The (scaled) probability distribution for one of the data sets of Fig. 12.

distribution determined by a large λ are "very much alike", in the sense that random samples from these spaces are very similar [16]. It makes sense therefore to use the average smoothness of the functions to determine the prior. This average smoothness is $\int (\int f_{uu}^2\, du) \exp(-\lambda \int f_{uu}^2\, du) \mathcal{D}f$. The one-dimensional equivalent is $\int x^2 \exp(-\lambda x^2)\, dx$, which equals $\lambda^{-\frac{3}{2}}$.

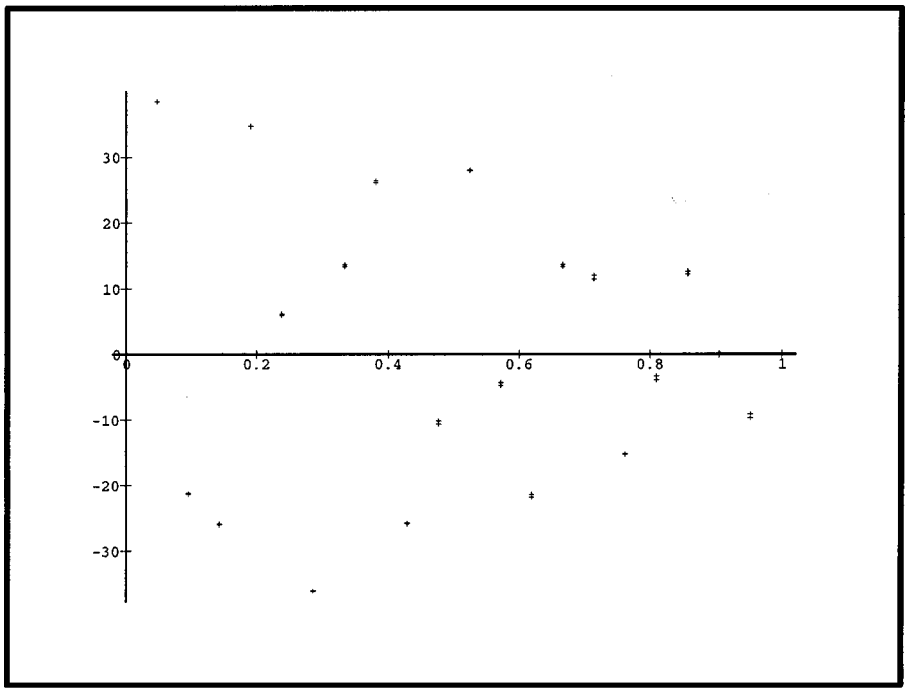

Figure 12. Two nearly identical data sets superimposed.

The prior, Prior(λ), should therefore satisfy

$$b^{-\frac{3}{2}} - a^{-\frac{3}{2}} = d^{-\frac{3}{2}} - c^{-\frac{3}{2}} \to \int_a^b \text{Prior}(\lambda)\, d\lambda = \int_c^d \text{Prior}(\lambda)\, d\lambda$$

so, Prior(λ) $= \lambda^{-\frac{5}{2}}$. This agrees with intuition—since the spaces for small values of λ differ much more than those with large values of λ, the prior is much greater when λ is small.

Unless some information on the noise is given, no prior is assumed on σ. However, we have experimented with different priors—such as log-normal—and the results seem to hardly depend on the specific prior.

Although the space $M_{\lambda,\sigma}$ is infinite dimensional, it is possible to reduce Eq. (2) to a quotient of two integrals defined on a finite dimensional space. The rest of this section is dedicated to this reduction, culminating in the expression of Eq. (7).

The problem of computing such integrals as those appearing in Eq. (2)—which are defined over infinite dimensional domains—has been solved for some types of integrals in the realm of pure mathematics [8, 11, 17–19, 41]. It was applied to the types of spaces used in regularization in [15, 16]. The space $M_{\lambda,\sigma}$ is a "Hilbert space" [42]. We will need to use the notion of an *orthogonal subspace*; let us recall that if U is a subspace of a Hilbert space H, its orthogonal subspace, $U^{\perp}$, is defined as—

$$U^{\perp} = \{h \in H / u \in U \Longrightarrow (u, h) = 0\}$$

It is well known that for every $h \in H$, there are unique $u_1 \in U$ and $u_2 \in U^{\perp}$ such that $u_1 + u_2 = h$. They are called the *projections* of u on U and $U^{\perp}$, and are denoted h_U and $h_{U^{\perp}}$.

The Hilbert space used in this work is the space of all functions which can serve as interpolants in the framework of regularization. Since we have to use functions f for which the smoothness term $\int f_{uu}^2\, du$ is defined, the natural space is the *Sobolev space* L_2^2, which consists of the functions having a second derivative which is square integrable (for an extensive treatment of Sobolev spaces, see [1]).

For technical reasons, we restrict ourselves to the subspace of L_2^2 which is defined by $\{f \in L_2^2 / f(0) = f(1) = 0\}$. The reason is that otherwise the denominator in Eq. (2) is not defined. We will keep denoting the model/function space by $M_{\lambda,\sigma}$. Note that this is not really a restriction—any two numbers B_0 and B_1 can be used for boundary conditions at 0 and 1, simply by subtracting the linear function which obtains the values B_0 and B_1 at 0 and 1. One can use different constraints, such as fixing the function's and its derivative's value at some point.

It turns out that the calculations are a little simpler if we make a change of variable, $\rho = \sqrt{2}\sigma$. The expression for the probability of $M_{\lambda,\sigma}$ given D is then

$$\frac{1}{(2\pi)^{n/2}\left(\frac{\rho}{\sqrt{2}}\right)^n} \frac{\int \exp\left(-\left(\lambda \int f_{uu}^2\, du + \sum_{i=1}^n \frac{[f(x_i)-y_i]^2}{\rho^2}\right)\right)\mathcal{D}f}{\int \exp\left(-\lambda \int f_{uu}^2\, du\right)\mathcal{D}f}$$

$$= \frac{\exp\left(-\frac{\|Y\|^2}{\rho^2}\right)}{\pi^{n/2}\rho^n} \frac{\int \exp\left(-\left(\lambda \int f_{uu}^2\, du + \frac{1}{\rho^2}\sum_{i=1}^n f^2(x_i) - \frac{2}{\rho^2}\sum_{i=1}^n y_i f(x_i)\right)\right)\mathcal{D}f}{\int \exp\left(-\lambda \int f_{uu}^2\, du\right)\mathcal{D}f}$$

where Y is the data vector, $(y_1, y_2 \ldots y_n)$.

Now, let us simplify the last expression by defining two inner products on $M_{\lambda,\sigma}$:

$$(f, g)_1 = \frac{1}{\rho^2}\sum_{i=1}^n f(x_i)g(x_i) + \lambda \int f_{uu} g_{uu}\, du$$

$$(f, g)_2 = \lambda \int f_{uu} g_{uu}\, du$$

For every x_i, let us denote by H_{x_i} the function which satisfies, for every $f \in M_{\lambda,\sigma}$, $\int (H_{x_i})_{uu} f_{uu}\, du = f(x_i)$. We can explicitly calculate this function, following the same method as in [15]:

$$H_x(\xi) = \begin{cases} 0 \le \xi \le x: & \dfrac{(x-1)\xi(x^2 - 2x + \xi^2)}{6} \\ x \le \xi \le 1: & \dfrac{x(\xi-1)(x^2 + \xi^2 - 2\xi)}{6} \end{cases}$$

note that this expression depends only on the location of the sample points x_i, and not the value of the samples y_i. As it turns out, this saves a lot of computation.

Finally, let us define for each i the function $h_{x_i} = \frac{H_{x_i}}{\lambda}$. Obviously, $(f, h_{x_i})_2 = f(x_i)$ for every $f \in M_{\lambda,\sigma}$, and so, if we define $f_0 = -\frac{2}{\rho^2}\sum_{i=1}^n y_i h_{x_i} = -\frac{2}{\lambda\rho^2}\sum_{i=1}^n y_i H_{x_i}$, then $(f, f_0)_2 = -\frac{2}{\rho^2}\sum_{i=1}^n y_i f(x_i)$.

After these definitions, the expression for the probability reduces to

$$\frac{\exp\left(-\frac{\|Y\|^2}{\rho^2}\right)}{\pi^{n/2}\rho^n}\frac{\int \exp(-[(f, f)_1 + (f, f_0)_2])\mathcal{D}f}{\int \exp(-(f, f)_2)\mathcal{D}f} \quad (3)$$

This integral can be computed using the fact that the function space can be decomposed into a direct sum, where the two inner products $(,)_1$ and $(,)_2$ differ from each other only on one of the summands, which is finite dimensional. Specifically, let us define a subspace W of $M_{\lambda,\sigma}$ by

$$W = \{f \in M_{\lambda,\sigma} \mid f(x_1) = f(x_2) = \ldots = f(x_n) = 0\}$$

when restricted to W, $(,)_1$ and $(,)_2$ define the same inner product. Moreover, if $f \in M_{\lambda,\sigma}$ and $g \in W$, then $(f, g)_1 = (f, g)_2$.

Now, if $f \in W$, then for every $1 \le i \le 1$, $(f, h_{x_i})_1 = f(x_i) = 0$, hence $h_{x_i} \in W^\perp$. Since the h_{x_i}'s are linearly independent [15], we have from dimension arguments the following important result

$$W^\perp = \text{span}\{h_{x_1}, h_{x_2} \ldots h_{x_n}\}$$

next, let us write the expression in the exponent of the integrand in the numerator of Eq. (3) using the decomposition into W and $W^\perp$:

$$(f, f)_1 + (f, f_0)_2 = (f_W, f_W)_2 + (f_{W^\perp}, f_{W^\perp})_1 + (f_{W^\perp}, f_0)_2$$

here, we have used the fact that $f_0 \in W^\perp$ (obvious, since it is a linear combination of the h_{x_i}'s), and also the fact that, restricted to W, the two inner products are the same.

Similarly, the expression in the exponent of the integrand in the denominator of Eq. (3) is $(f_W, f_W)_2 + (f_{W^\perp}, f_{W^\perp})_2$. Writing the appropriate exponents as products, e.g.,

$$\exp(-[(f, f)_1 + (f, f_0)_2]) = \exp(-(f_W, f_W)_2) \times \exp(-(f_{W^\perp}, f_{W^\perp}))_1 \exp(-(f_{W^\perp}, f_0))_2$$

we see that the integrals over W cancel out, and the expression in Eq. (3) is equal to

$$\frac{\exp\left(-\frac{\|Y\|^2}{\rho^2}\right)}{\pi^{n/2}\rho^n}\frac{\int_{W^\perp} \exp(-[(f, f)_1 + (f, f_0)_2])\mathcal{D}f}{\int_{W^\perp} \exp(-(f, f)_2)\mathcal{D}f} \quad (4)$$

This expression is computed by identifying $W^\perp$ with $\mathcal{R}^n$. The n-dimensional vector $(u_1, u_2 \ldots u_n)$ is identified with $\sum_{i=1}^n u_i h_{x_i}$. There is no need to worry about the Jacobian of this transformation, as it appears both in the numerator and denominator and hence cancels out. We are left with the following:

$$\frac{\exp\left(-\frac{\|Y\|^2}{\rho^2}\right)}{\pi^{n/2}\rho^n}\frac{\int_{\mathcal{R}^n} \exp\left(-[u\Lambda_1 u^T + (u, u_0)]\right) du}{\int_{\mathcal{R}^n} \exp(-(u\Lambda_2 u^T))\, du} \quad (5)$$

where $(,)$ denotes the usual scalar product on $\mathcal{R}^n$, and

$$(\Lambda_2)_{i,j} = \left(h_{x_i}, h_{x_j}\right)_2 = h_{x_i}(x_j)$$
$$(\Lambda_1)_{i,j} = \left(h_{x_i}, h_{x_j}\right)_1 = (h_{x_i}, h_{x_j})_2 + \frac{1}{\rho^2}\sum_{k=1}^n h_{x_i}(x_k) h_{x_j}(x_k)$$

and

$$[u_0]_i = -\frac{2}{\rho^2}\sum_{k=1}^n y_k h_{x_k}(x_i)$$

defining an $n \times n$ matrix A by $A_{i,j} = H_{x_i}(x_j)$, we have

$$\Lambda_2 = \frac{A}{\lambda}$$
$$\Lambda_1 = \frac{A}{\lambda} + \frac{A^2}{\lambda^2\rho^2}$$
$$u_0 = -\frac{2}{\lambda\rho^2} YA$$

and so the expression of Eq. (5) equals

$$\frac{\exp\left(-\frac{\|Y\|^2}{\rho^2}\right)}{\pi^{n/2}\rho^n} |\Lambda_2|^{1/2} |\Lambda_1|^{-(1/2)} \exp\left(\frac{1}{4} u_0 \Lambda_1^{-1} u_0^t\right) \quad (6)$$

now, $\Lambda_1 = \frac{A}{\lambda} + \frac{A^2}{\lambda^2\rho^2} = \frac{1}{\lambda^2\rho^2}(\lambda\rho^2 A + A^2) = \frac{A}{\lambda^2\rho^2}(\lambda\rho^2 I + A)$, and so $|\Lambda_1|^{-(1/2)} = \lambda^n \rho^n |A|^{-(1/2)} |\lambda\rho^2 I + A|^{-(1/2)}$.

Since $|\Lambda_2|^{1/2} = \lambda^{-(n/2)}|A|^{1/2}$, we have that

$$\frac{1}{\pi^{n/2}\rho^n}\,|\Lambda_2|^{1/2}|\Lambda_1|^{-(1/2)} = \frac{\lambda^{n/2}}{\pi^{n/2}}|\lambda\rho^2 I + A|^{-(1/2)}.$$

Next, we turn to calculate the exponent in Eq. (6): it is equal to $\frac{1}{4}u_0\Lambda_1^{-1}u_0^t - \frac{\|Y\|^2}{\rho^2}$. Using the fact that $u_0 = -\frac{2}{\lambda\rho^2}YA$, we get

$$\begin{aligned}\frac{1}{4}u_0\Lambda_1^{-1}u_0^t &= \frac{1}{4}\left(\frac{2}{\lambda\rho^2}\right)^2\lambda^2\rho^2 YA(\lambda\rho^2 I + A)^{-1}Y^t \\ &= \frac{YA(\lambda\rho^2 I + A)^{-1}Y^t}{\rho^2}\end{aligned}$$

to get the total exponent we have to subtract $\frac{\|Y\|^2}{\rho^2}$ from this, which results in

$$\frac{YA(\lambda\rho^2 I + A)^{-1}Y^t - \|Y\|^2}{\rho^2} = -\lambda Y(A + \lambda\rho^2 I)^{-1}Y^t$$

and, all in all, the probability is

$$\frac{\lambda^{n/2}}{\pi^{n/2}}|A + \lambda\rho^2 I|^{-(1/2)}\exp(-\lambda Y(A + \lambda\rho^2 I)^{-1}Y^t) \quad (7)$$

If one wishes to find the MAP weights—that is, the λ and σ maximizing Eq. (7)—this can be reduced to a one-dimensional optimization problem as follows. Let us substitute u for λ and v for $\lambda\rho^2$. Then,

$$\begin{aligned}&\frac{\lambda^{n/2}}{\pi^{n/2}}|A + \lambda\rho^2 I|^{-(1/2)}\exp(-\lambda Y(A + \lambda\rho^2 I)^{-1}Y^t) \\ &\quad = \frac{1}{\pi^{n/2}}\,\frac{u^{n/2}}{K_2(v)}\exp(-uK_1(v))\end{aligned}$$

where the definitions of $K_1(\,)$ and $K_2(\,)$ are the obvious ones. Keeping v constant, we can maximize over u (discarding for the moment quantities which depend only on v):

$$\begin{aligned}&\frac{\partial}{\partial u}(u^{n/2}\exp(-uK_1(v))) \\ &\quad = \frac{n}{2}u^{n/(2-1)}\exp(-uK_1(v)) \\ &\qquad - K_1(v)u^{n/2}\exp(-uK_1(v))\end{aligned}$$

which is zero when $u = \frac{n}{2K_1(v)}$. Substituting this back into the expression for the probability yields

$$\begin{aligned}&\frac{1}{\pi^{n/2}}\,\frac{\left[\frac{n}{2K_1(v)}\right]^{n/2}}{K_2(v)}\exp\left(-\frac{n}{2K_1(v)}K_1(v)\right) \\ &\quad = \frac{\left(\frac{n}{2\pi e}\right)^{n/2}}{K_2(v)[K_1(v)]^{n/2}}\end{aligned}$$

If $v_{\max}$ maximizes this expression, we can easily extract the optimal λ, which is equal to $\frac{n}{2K_1(v_{\max})}$, and ρ, which is $\sqrt{2vK_1(v_{\max})/n}$, hence the optimal σ is $\sqrt{v_{\max}K_1(v_{\max})/n}$.

Again, it is important to emphasize that these λ and σ are "optimal" only in the sense that they maximize $\Pr(M_{\lambda,\sigma})$, and that the MAP and MSE estimates can be rather different from the estimate $(f_{\text{opt}})_{\lambda_{\max},\sigma_{\max}}$ obtained using only these "optimal" λ and σ.

As noted before, $(f_{\text{opt}})_{\lambda_{\max},\sigma_{\max}}$ may be a good approximation to the MSE estimate if $\Pr(M_{\lambda,\sigma})$ is unimodal, symmetric, and concentrated around $M_{\lambda_{\max},\sigma_{\max}}$. In that case, it will be very useful to find $\{\lambda_{\max}, \sigma_{\max}\}$, because computing $(f_{\text{opt}})_{\lambda_{\max},\sigma_{\max}}$ is faster than computing the MSE estimate using Eq. (8). The simple observation above shows that one can reduce the search for $\{\lambda_{\max}, \sigma_{\max}\}$ from a two-dimensional minimization problem to a one-dimensional one.

A1.1 Computing the Expectation of the Value at x

If L is a functional on f, its expectation given D is

$$E[L(f)/D] = \int L(f)\Pr(f/D)\mathcal{D}f$$

if we want to compute the value of a function at a point x, then $L(f)$ is simply the evaluation at x, and the expectation can be computed, according to Fubini's theorem, by first evaluating it for each $\{\lambda, \sigma\}$ pair, and then integrating over all such pairs, weighing each one by its probability conditioned by the data D:

$$\begin{aligned}E_x &= E[f(x)/D] \\ &= \int\int E[f(x)/D, \lambda, \sigma]\Pr(\lambda, \sigma/D)\,d\lambda\,d\sigma\end{aligned}$$

First, we have to compute $E[f(x)/D, \lambda, \sigma]$. However, the probability distribution on $M_{\lambda,\sigma}$ is Gaussian, so it is enough to find the MAP estimate. As shown in Eq. (5), the probability of u in $M_{\lambda,\sigma}$ is $\exp(-[u\Lambda_1 u^T + (u, u_0)])$ (using the same notations as those in this appendix). Therefore, the u maximizing the probability has to minimize $u\Lambda_1 u^T + (u, u_0)$, so $u = \frac{1}{2}\Lambda_1^{-1}u_0$. Substituting the expressions previously derived in the appendix,

$$[u_0]_i = -\frac{2}{\rho^2}\sum_{k=1}^{n} y_k h_{x_k}(x_i)$$

and

$$\Lambda_1 = \frac{A}{\lambda} + \frac{A^2}{\lambda^2\rho^2}$$

$$u_0 = -\frac{2}{\lambda\rho^2} YA$$

it is easy to verify that

$$E[f(x)/D,\lambda,\sigma] = \big(H_{x_1}(x)\dots H_{x_n}(x)\big)(A+\lambda\rho^2 I)^{-1}Y^t$$

And so $E[f(x)/D]$ equals

$$\frac{\int\int \big(H_{x_1}(x)\dots H_{x_n}(x)\big)(A+\lambda\rho^2 I)^{-1}Y^t \frac{\lambda^{(n-5)/2}}{\pi^{n/2}}|A+\lambda\rho^2 I|^{-(1/2)}\exp(-\lambda Y(A+\lambda\rho^2 I)^{-1}Y^t)\,d\lambda\,d\sigma}{\int\int \frac{\lambda^{(n-5)/2}}{\pi^{n/2}}|A+\lambda\rho^2 I|^{-(1/2)}\exp(-\lambda Y(A+\lambda\rho^2 I)^{-1}Y^t)\,d\lambda\,d\sigma}$$

using the change of variables $u=\lambda$, $v=\lambda\rho^2$, the integral transforms to

$$\frac{\int\int \frac{1}{\sqrt{v}}\big(H_{x_1}(x)\dots H_{x_n}(x)\big)(A+vI)^{-1}Y^t \frac{u^{(n-6)/2}}{\pi^{n/2}}|A+vI|^{-(1/2)}\exp(-uY(A+vI)^{-1}Y^t)\,du\,dv}{\int\int \frac{1}{\sqrt{v}}\frac{u^{(n-6)/2}}{\pi^{n/2}}|A+vI|^{-(1/2)}\exp(-uY(A+vI)^{-1}Y^t)\,du\,dv}$$

the inner integral is a Gamma function, hence the last expression reduces to

$$\frac{\int \frac{1}{\sqrt{v}}|A+vI|^{-(1/2)}\big(H_{x_1}(x)\dots H_{x_n}(x)\big)(A+vI)^{-1}Y^t[Y(A+vI)^{-1}Y^t]^{(4-n)/2}\,dv}{\int \frac{1}{\sqrt{v}}|A+vI|^{-(1/2)}[Y(A+vI)^{-1}Y^t]^{(4-n)/2}\,dv} \tag{8}$$

This integral is computed numerically. As in Section 3, A is diagonalized to save time when computing the integrand.

Appendix 2: Computing the Pointwise Uncertainty

The computation of the uncertainty at a point resembles the one carried out in [15], but is somewhat more complicated, because in that work it was assumed that the probability of a function depended on a single pair of weights, $\{\lambda,\sigma\}$. Let us proceed with the computation:

$$\int [f(x)-E_x]^2 \Pr(f/D)\mathcal{D}f = \int [f(x)-E_x]^2 \times \left[\int\int \Pr(f/D,\lambda,\sigma)\Pr(\lambda,\sigma/D)\,d\lambda\,d\sigma\right]\mathcal{D}f =$$

(due to Fubini's theorem)

$$\int\int\left[\int [f(x)-E_x]^2 \Pr(f/D,\lambda,\sigma)\mathcal{D}f\right] \times \Pr(\lambda,\sigma/D)\,d\lambda\,d\sigma =$$

$$\int\int\left[\int [f^2(x)-2E_x f(x)+E_x^2] \times \Pr(f/D,\lambda,\sigma)\mathcal{D}f\right]\Pr(\lambda,\sigma/D)\,d\lambda\,d\sigma =$$

(recalling the definition of E_x)

$$\int\int\left[\int f^2(x)\Pr(f/D,\lambda,\sigma)\mathcal{D}f\right] \times \Pr(\lambda,\sigma/D)\,d\lambda\,d\sigma - E_x^2$$

Proceeding as with the computation in the previous appendix, the inner integral is equal to

$$\int f^2(x)\Pr(f/D,\lambda,\sigma)\mathcal{D}f = \frac{\int f^2(x)\exp(-[(f,f)-2(f,f_0)])\mathcal{D}f}{\int \exp(-[(f,f)-2(f,f_0)])\mathcal{D}f}$$

where $(f,g)=\frac{1}{\rho^2}\sum f(x_i)g(x_i)+\lambda\int f_{uu}\,g_{uu}\,du$, and $f_0=\sum y_i h_{x_i}$, where h_{x_i} are the reproducing kernels satisfying $(f,h_{x_i})=f(x_i)$ (note that these are different than the H_{x_i}).

By a change of variables this turns out to be

$$\frac{\int [g(x)+f_0(x)]^2\exp(-(g,g))\mathcal{D}g}{\int g^2(x)\exp(-(g,g))\mathcal{D}g} = \int \big[f_0^2(x)+g^2(x)\big]\mu(Dg)$$

where $\mu(Dg)$ is the Gaussian measure induced by the inner product ( ) [8, 15, 19]. We need to compute $\int g^2(x)\mu(Dg)$. In [15] it is shown that this integral equals

$$\rho^2 V_x \left(B_1 + \lambda\rho^2 B_1^2\right)^{-1} V_x^T$$

where, if we denote $x = x_{n+1}$, $V_x = (H_{x_1}(x) \dots H_{x_{n+1}}(x))$, and B_1 is the $(n+1) \times (n+1)$ matrix defined by $(B_1)_{i,j} = H_i(x_j)$ (so, B_1 contains A as a sub-matrix).

Now we have to incorporate the exterior integrals, over λ and σ. We use the same substitution we used before to reduce this integral to a one-dimensional integral. Adding the other summands—that is, $-E_x^2$ and the integral of the squared expectation (the term corresponding to the integral of $f_0^2(x)$)—we finally get the following expression for the uncertainty, or variance, at x:

$$\frac{2\int \sqrt{v}|A+vI|^{-(1/2)}\left(H_{x_1}(x)\dots H_{x_n}(x)\right), H_x(x))\left(B_1+vB_1^2\right)^{-1}\left(H_{x_1}(x)\dots H_{x_n}(x)\right), H_x(x))^t[Y(A+vI)^{-1}Y^t]^{(6-n)/2}\,dv}{(n-6)\int \frac{1}{\sqrt{v}}|A+vI|^{-(1/2)}[Y(A+vI)^{-1}Y^t]^{(4-n)/2}\,dv}$$

$$+\frac{\int \frac{1}{\sqrt{v}}|A+vI|^{-(1/2)}\left[\left(H_{x_1}(x)\dots H_{x_n}(x)\right)(A+vI)^{-1}Y^t\right]^2[Y(A+vI)^{-1}Y^t]^{(4-n)/2}\,dv}{\int \frac{1}{\sqrt{v}}|A+vI|^{-(1/2)}[Y(A+vI)^{-1}Y^t]^{(4-n)/2}\,dv}$$

$$-\left[\frac{\int \frac{1}{\sqrt{v}}|A+vI|^{-(1/2)}\left(H_{x_1}(x)\dots H_{x_n}(x)\right)(A+vI)^{-1}Y^t[Y(A+vI)^{-1}Y^t]^{(4-n)/2}\,dv}{\int \frac{1}{\sqrt{v}}|A+vI|^{-(1/2)}[Y(A+vI)^{-1}Y^t]^{(4-n)/2}\,dv}\right]^2$$

Acknowledgments

We thank Prof. David Donoho of Stanford University, and Dr. David Steinberg from the University of Tel-Aviv, for their insightful comments on this work. We also thank Alan Williams from the University of Queensland, who showed us how to implement the 2D Generalized-Cross-Validation package, and Yoram Singer from the Hebrew University, who supplied us with the data used in one of the experiments. This paper was improved following suggestions from three anonymous reviewers; we are grateful for their efforts and constructive remarks.

This research has been sponsored by the U.S. Office of Naval Research under Grant N00014-93-1-1202, R&T Project Code 4424341—01.

References

1. R.A. Adams, *Sobolev Spaces*, Academic Press, 1975.
2. H. Akima, "Bivariate interpolation and smooth surface fitting based on local procedures," *Comm. ACM*, Vol. 17, pp. 26–31, 1974.
3. M. Bertero, T.A Poggio, and V. Torre, "Ill-posed problems in early vision," in *Proceedings of the IEEE*, Vol. 8, pp. 869–889, 1988.
4. R.J. Chorley, *Spatial Analysis in Geomorphology*, Methuen and Co., 1972.
5. P. Craven and G. Whaba, "Optimal smoothing of noisy data with spline functions," *Numerische Mathematik*, Vol. 31, pp. 377–403, 1979.
6. R. Fischer, W. Von Der Linden, and V. Dose, "On the Importance of α Marginalization in Maxium Entropy," *Maximum Entropy and Bayesian Methods (book chapter)*, Kluwer Academic, 1996, 229–236.
7. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 6, pp. 721–741, 1984.
8. L. Gross, "Integration and non-linear transformations in Hilbert space," *Transactions of the American Mathematical Society*, Vol. 94, pp. 404–440, 1960.
9. S.F. Gull, "Developments in maximum entropy data analysis," in *Maximum Entropy and Bayesian Methods*, J. Skilling (Ed.) Kluwer Academic, 1989.
10. P. Hall and I. Johnstone, "Emprical functionals and efficient smoothing parameter selection," *Journal of the Royal Statistical Society*, Vol. 54, No. 1, pp. 475–530, 1992.
11. E. Hille, "Introduction to the general theory of reproducing kernels," *Rocky Mountain Journal of Mathematics*, Vol. 2, pp. 321–368, 1972.
12. B. Horn, *Robot Vision*, MIT Press, 1986.
13. B.K.P Horn and B.G. Schunck, "Determining optical flow," *Artificial Intelligence*, Vol. 17, pp. 185–203, 1981.
14. D. Keren and M. Werman, "A Bayesian framework for regularization," in *12th International Conference on Pattern Recognition*, Vol. C, Jerusalem, 1994, pp. 72–76.
15. D. Keren and M. Werman, "Probabilistic analysis of regularization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 15, pp. 982–995, 1993.
16. Daniel Keren, "Probabilistic analyses of interpolation in computer vision," Ph.D. thesis, Hebrew University of Jerusalem, 1990.
17. J. Kuelbs, F.M. Larkin, and J.A. Williamson, "Weak probability disributions on reproducing kernel Hilbert spaces," *Rocky Mountain Journal of Mathematics*, Vol. 2, pp. 369–378, 1972.
18. H.H. Kuo, *Guassian Measures in Banach Spaces*, Springer-Verlag, 1975.

19. F.M. Larkin, "Gaussian measure in Hilbert space and applications in numerical analysis," *Rocky Mountain Journal of Mathematics*, Vol. 2, pp. 379–421, 1972.
20. S. Lauritzen, "Random orthogonal set functions and stochastic models for the gravity potential of the earth," *Stochastic Processes and Applications*, Vol. 3, pp. 65–72, 1975.
21. M. Lee, A. Rangarajan, I.G. Zubal, and G. Gindi, "A continuation method for emission tomography," *IEEE Transaction on Nuclear Science*, Vol. 40, No. 6, pp. 2049–2058, 1993.
22. S.J. Lee, A. Rangarajan, and G. Gindi, "Bayesian image reconstruction in spect using higher-order mechanical models as priors," *IEEE Transaction on Medical Imaging*, Vol. 4, pp. 669–680, December 1995.
23. D.J.C. MacKay, "Comparison of approximate methods in handling hyperparameters," *Neural Computation*, to appear.
24. D.J.C. MacKay, "Bayesian methods for adaptive models," Ph.D. thesis, California Institute of Technology, 1992.
25. R. Molina, "On the hierarchical Bayesian approach to image restoration: Applications to astronomical images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 11, pp. 1122–1128, 1994.
26. R. Molina and A.K. Katsaggelos, "On the hierarchical Bayesian approach to image restoration and the iterative evaluation of the regularization parameter," in *Visual Communications and Image Processing'94, Proc. SPIE 2308*, Aggelos K. Katsaggelos (Ed.), 1994, pp. 244–251.
27. D. Nychka, "Bayesian confidence intervals for smoothing splines," *Journal of the American Statistical Association*, Vol. 83, No. 404, pp. 1134–1143, 1988.
28. D. Nychka, "Choosing a range for the amount of smoothing in nonparametric regression," *Journal of the American Statistical Association*, Vol. 86, No. 415, pp. 653–664, 1991.
29. J.E. Robinson, H.A.K. Charlesworth, and M.J. Ellis, "Structural analysis using spatial filtering in interior plans of south-central alberta," *Amer. Assoc. Petrol. Geol. Bull.*, Vol. 53, pp. 2341–2367, 1969.
30. J. Skilling, *Quantified Maximum Entropy. Maximum Entropy and Bayesian Methods*, Kluwer Academic, 1990.
31. J. Skilling, "Fundamentals of maxent in data analysis," in *Maximum Entropy in Action*, B. Buck and V.A. Macaulay (Eds.). Clarendon Press: Oxford, 1991.
32. C.E.M. Strauss, D.H. Wolpert, and E.D. Wolf, "Alpha Evidence and the Entropic Prior," *Maximum Entropy and Bayesian Methods*, Kluwer Academic, 1995.
33. R. Szeliski, *Bayesian Modeling of Uncertainty in Low-Level Vision*, Kluwer, 1989.
34. S. Szeliski and D. Terzopoulos, "From splines to fractals," *SIGGRAPH*, 1989, pp. 51–60.
35. D. Terzopoulos, "Multi-level surface reconstruction," in *Multiresolution Image Processing and Analysis*, A. Rosenfeld (Ed.), Springer-Verlag, 1984.
36. D. Terzopoulos, "Regularization of visual problems involving discontinuities," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 8, pp. 413–424, 1986.
37. A.M. Thompson, J.C. Brown, J.W. Kay, and D.M. Titterington, "A study of methods of choosing the smoothing parameter in image restoration by regularization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 13, pp. 326–339, 1991.
38. A.N. Tikhonov and V.Y. Arsenin, *Solution of Ill-Posed Problems*, Winston and Sons, 1977.
39. G. Wahba, "Bayesian 'confidence intervals' for the cross-validated smoothing spline," *Journal of the Royal Statistical Society, Ser. B*, Vol. 45, pp. 133–150, 1983.
40. G. Wahba, *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics: Philadelphia, 1990.
41. G.W. Wasilkowski, "Optimal algorithms for linear problems with Gaussian measures," *Rocky Mountain Journal of Mathematics*, Vol. 16, pp. 727–749, 1986.
42. N. Young, *An Introduction to Hilbert Space*, Cambridge Mathematical Textbooks, 1988.