A Bayesian Framework for Regularization *

July 6, 1994

Daniel Keren and Michael Werman
Institute for Computer Science
The Hebrew University of Jerusalem
Jerusalem 91904, Israel
dkeren@cs.huji.ac.il, werman@cs.huji.ac.il

Abstract

Regularization is a popular method for interpolating sparse data, as well as smoothing data obtained from noisy measurements. Simply put, regularization looks for an interpolating or approximating function which is both close to the data and also "smooth" in some sense. Formally, this function is obtained by minimizing an error functional which is the sum of two terms, one measuring the distance from the data, the other measuring the smoothness of the function.

The classical approach to regularization is: decide the relative weights that should be given to these two terms, and minimize the resulting error functional. This approach, however, suffers from two serious flaws: there is no rigorous way to compute these weights and it does not find the function which is the MAP estimate for the interpolation problem. This may result in grave instability in the reconstruction process; two data sets which are arbitrarily close to each other can be assigned two radically different interpolants.

Our solution to this problem is through the Bayesian approach: instead of using only the *optimal* estimates for the weights, construct a probability distribution over all possible weights. Then, the probability of each function is equal to the integral of its probability over all possible weights, scaled by the probability of the weights. The MAP estimate is the function which maximizes this probability.

*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.

The main contribution of this work is the construction of the probability distribution for the weights. This also allows to rigorously solve the smaller problem of finding the MAP estimate for the weights.

The optimal weights are still important – for instance, for classification problems, and for separation of signal from noise. We show that finding the optimal weights can be reduced from a two-dimensional optimization problem to a one-dimensional one. We also study the reliability of the estimate for the optimal smoothing parameter.

Keywords: Regularization, smoothing parameter, function spaces

1 Introduction

In computer vision, regularization [28] is used to reconstruct objects from partial data [26, 27, 10, 3]. Reconstruction of surfaces from partial data has been studied in many other fields, for example petroleum exploration [22], geology [4], electronics [2] and medical imaging. The data 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 [9] 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? Also, data instances which are not compatible with others can cause singularities in the solution.

The regularization approach 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 function* $M(f)$ 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 $M()$.

The one-dimensional case, which will be addressed in this work, is to minimize $M(f) = \sum_{i=1}^{n} \frac{[f(x_i) - y_i]^2}{\sigma^2} + \lambda \int_0^1 [f''(v)]^2 dv$.

Without loss of generality, we will assume from now on that the functions are defined on the interval $[0, 1]$, and will omit those limits in the integrals.

However, this approach fails to find the MAP estimate for the interpolant f, as it uses only the optimal weights to construct f. But, what happens if that function has a relatively small probability for a wide range of other weights? The MAP estimate should maximize the following:

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

where w varies over the set of all possible weights. The main contribution of this work (Section 3) is the computation of the integrand, $Pr(f/w)Pr(w)$; space limitations don't allow us to elaborate on how the integral is computed.

A word of caution: it seems that one doesn't really need both λ and σ, but only the quantity $\lambda\sigma^2$, as it completely determines the interpolant in the "classical" regularization approach. However, for the approach presented here, one has to compute the joint distribution. Moreover, even if one wishes to use only the optimal value of $\lambda\sigma^2$ for the reconstruction, the values of λ and σ are important for classification and in order to determine how much noise was present in the measurement process.

A method which is extensively used is the *Generalized Cross Validation* (GCV) algorithm. In Section 4, it is demonstrated that the GCV algorithm returns completely different interpolants for two arbitrarily close sets of data; this is because it relies only on the optimal weights, and does not bring into consideration the "global probability" of the interpolant.

In Section 5 an estimator fo the optimal λ is presented and analized. In Section 6, a realistic example (hand-written data) is analized. Section 7 summarizes the work and suggests directions for further research.

2 Previous Work

The most popular method for determining the smoothing parameter λ is that of the generalized cross validation (GCV) [5], which is described in more detail in Section 4. Intuitively, GCV tries to choose a λ such that the data points predict each other; specifically, it tries to minimize the sum of differences between y_i and $S_i(x_i)$, where S_i is the spline that reconstructs the data set with the point (x_i, y_i) removed. As will be demonstrated, GCV can be very unstable, in the sense that it gives two radically different values of λ for two data sets which are arbitrarily close to each other. Also, there does not seem to be a theoretical justification for using GCV for arbitrary data, although it has some nice asymptotic properties.

A different approach, which is closer to ours, is that of Bayesian model selection which, to the best of our knowledge, was first suggested in the pioneering work of Szeliski [24]. There, the following question is posed: *given the data D, what is the most probable value of* λ*?* More recent work in this direction was done by MacKay [18]. This article suggests a different approach, namely, computing the probability distribution by directly integrating over the (infinite-dimensional) space of all possible interpolants.

3 Computing the Joint Probability for λ, σ

Suppose we have a data set D and we want to describe or fit it with a member of some model M. The Bayes solution is to find f which satisfies

$$\max_{f \in M} Pr(f/D) = \max_{f \in M} \frac{Pr(D/f)Pr(f/M)}{Pr(D)} \propto \max_{f \in M} Pr(D/f)Pr(f/M)$$

Regularization, for instance, can be formalized in this way because

$$Pr(D/f) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp(-\sum_{i=1}^{n}\frac{[f(x_i)-y_i]^2}{2\sigma^2})$$

(assuming uncorrelated Gaussian noise of constant variance) and the prior distribution is

$$Pr(f) \propto \exp(-\lambda\int [f''(v)]^2 dv)$$

which resembles the Boltzmann distribution [24, 6, 11, 12, 23, 21, 19, 17, 25, 20]. Multiplying, we get that the f chosen from M should maximize $\exp(-M(f))$, or minimize $M(f)$. This simple analysis shows how regularization is consistent with Bayes rule for choosing the MAP estimate, given λ.

Now, what if a few models are possible? As explained before, the first step is to compute the probability of each model. In our case, the models are indexed by two continuous parameters, λ and σ. Thus, a specific choice of the two parameters is equivalent to the claim "this function was sampled from a function space with a specific prior, and the measurement was corrupted by a specific noise".

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

$$Pr(M_{\lambda,\sigma}/D) = \frac{Pr(D/M_{\lambda,\sigma})Pr(M_{\lambda,\sigma})}{Pr(D)} \propto Pr(D/M_{\lambda,\sigma}) = \frac{\int_{M_{\lambda,\sigma}} Pr(D/f)Pr(f/M_{\lambda,\sigma})df}{\int_{M_{\lambda,\sigma}} Pr(f/M_{\lambda,\sigma})df} \quad (1)$$

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.

Here we have assumed an improper uniform prior on the family of models. Currently, we are studying different priors. Using different priors will be straightforward, as it appears only as a multiplier to the final probability we have computed (Equation 6).

The main goal of this work is to compute this integral; it turns out that although the space $M_{\lambda,\sigma}$ is infinite dimensional, it is possible to reduce the integral 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 Equation 6, so that the optimal λ and σ are the ones that maximizes Equation 6. Then, it is shown how this is equivalent to a simpler, one-dimensional, optimization problem.

The question of how to compute such integrals as those appearing in Equation 1 – which are defined over domains that are infinite dimensional – has been solved for some types of integrals in the realm of pure mathematics [8, 14, 16, 29, 15, 7]. It was applied to the types of spaces used in regularization in [12, 13]. The space $M_{\lambda,\sigma}$ is a "Hilbert space" [30]. Let us recall that if U is a subspace of a Hilbert space H, its *orthogonal subspace*, $U^{\perp}$, is defined as following –

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

It is well known that for every $h \in H$, there are $u_1 \in U$ and $u_2 \in U^{\perp}$ so that $u_1 + u_2 = h$. Also, u_1 and u_2 are unique. 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 $\int_0^1 [f''(v)]^2 dv$ 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 $\{f \in L_2^2 \mid f(0) = f(1) = 0\}$. The reason is that otherwise the denominator in Equation 1 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.

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

$$\frac{1}{(2\pi)^{\frac{n}{2}}(\frac{\rho}{\sqrt{2}})^n} \frac{\int_f \exp(-(\lambda \int [f''(v)]^2 dv + \sum_{i=1}^{n} \frac{[f(x_i) - y_i]^2}{\rho^2}))df}{\int_f \exp(-\lambda \int [f''(v)]^2 dv)df} =$$

$$\frac{\exp(-\frac{\|Y\|^2}{\rho^2})}{\pi^{\frac{n}{2}}\rho^n} \frac{\int_f \exp(-(\lambda \int [f''(v)]^2 dv + \frac{1}{\rho^2}\sum_{i=1}^{n} f^2(x_i) - \frac{2}{\rho^2}\sum_{i=1}^{n} y_i f(x_i)))df}{\int_f \exp(-\lambda \int [f''(v)]^2 dv)df}$$

where Y is the data vector, $\{y_1, y_2 .... y_n\}$.

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

$$(f,g)_1 = \frac{1}{\rho^2}\sum_{i=1}^{n} f(x_i)g(x_i) + \lambda \int f''(v)g''(v)dv$$

$$(f,g)_2 = \lambda \int f''(v)g''(v)dv$$

For every x_i, let us denote by H_{x_i} the function which satisfies, for every $f \in M_{\lambda,\sigma}$, $\int_0^1 H''_{x_i}(v)f''(v)dv = f(x_i)$. We can explicitly calculate this function, following the same method as in [12]:

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

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 in Equation 1 reduces to

$$\frac{\exp(-\frac{\|Y\|^2}{\rho^2})}{\pi^{\frac{n}{2}}\rho^n} \frac{\int_f \exp(-[(f,f)_1 + (f,f_0)_2])df}{\int_f \exp(-(f,f)_2)df} \tag{2}$$

This integral can be computed using the fact that the function space can be decomposed into a direct sum, where the 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

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

when restricted to W, $(,)_1$ and $(,)_2$ define the same inner product. Moreover, if $h \in M_{\lambda,\sigma}$ 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 [12], we have from dimension arguments the following important result

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

next, let us write the expression in the exponent of the integrand in the numerator of Equation 2 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 Equation 2 is $(f_W, f_W)_2 + (f_{W^{\perp}}, f_{W^{\perp}})_2$. Writing the appropriate exponents as products, e.g.

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

we see that the integrals over W cancel out, and the expression in Equation 2 is equal to

$$\frac{\exp(-\frac{\|Y\|^2}{\rho^2})}{\pi^{\frac{n}{2}}\rho^n} \frac{\int_{W^{\perp}} \exp(-[(f,f)_1 + (f,f_0)_2])df}{\int_{W^{\perp}} \exp(-(f,f)_2)df} \tag{3}$$

This expression is computed by identifying $W^{\perp}$ with $\mathcal{R}^n$. The n-dimensional vector $(u_1, u_2....u_n)$ 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:

$$\frac{\exp(-\frac{\|Y\|^2}{\rho^2})}{\pi^{\frac{n}{2}}\rho^n} \quad \frac{\int_{\mathcal{R}^n} \exp(-[u\Lambda_1 u^T + (u, u_0)])du}{\int_{\mathcal{R}^n} \exp(-(u\Lambda_2 u^T))du} \tag{4}$$

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

$$(\Lambda_2)_{i,j} = (h_{x_i}, h_{x_j})_2 = h_{x_i}(x_j)$$

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

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

$$\Lambda_2 = \frac{A}{\lambda}$$

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

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

and so the expression of Equation 4 equals

$$\frac{\exp(-\frac{\|Y\|^2}{\rho^2})}{\pi^{\frac{n}{2}}\rho^n} \quad |\Lambda_2|^{\frac{1}{2}}|\Lambda_1|^{-\frac{1}{2}}\exp(\frac{1}{4}u_0\Lambda_1^{-1}u_0^t) \tag{5}$$

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 $|\Lambda_1|^{-\frac{1}{2}} = \lambda^n\rho^n|A|^{-\frac{1}{2}}|\lambda\rho^2 I+A|^{-\frac{1}{2}}$.

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

$$\frac{1}{\pi^{\frac{n}{2}}\rho^n} \quad |\Lambda_2|^{\frac{1}{2}}|\Lambda_1|^{-\frac{1}{2}} = \frac{\lambda^{\frac{n}{2}}}{\pi^{\frac{n}{2}}}|\lambda\rho^2 I + A|^{-\frac{1}{2}}$$

Next, we turn to calculate the exponent in Equation 5: 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

$$\frac{1}{4}u_0\Lambda_1^{-1}u_0^t = \frac{1}{4}(\frac{2}{\lambda\rho^2})^2\lambda^2\rho^2 YA(\lambda\rho^2 I + A)^{-1}Y^t = \frac{YA(\lambda\rho^2 I + A)^{-1}Y^t}{\rho^2}$$

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

$$\frac{\lambda^{\frac{n}{2}}}{\pi^{\frac{n}{2}}}|A+\lambda\rho^2 I|^{-\frac{1}{2}}\exp(-\lambda Y(A+\lambda\rho^2 I)^{-1}Y^t) \tag{6}$$

A further simplification is possible if one uses the fact that every positive definite symmetric matrix can be diagonalized by an orthonormal matrix. So, let U be an orthonormal matrix satisfying $UAU^t = D$ and $U^tDU = A$. Then

$$|A+\lambda\rho^2 I| = |U^t(D+\lambda\rho^2 I)U| = |U^t||D+\lambda\rho^2 I||U| = |D+\lambda\rho^2 I| = (d_{11}+\lambda\rho^2)(d_{22}+\lambda\rho^2)...(d_{nn}+\lambda\rho^2)$$

and also

$$Y(A+\lambda\rho^2 I)^{-1}Y^t = Y(U^tDU+\lambda\rho^2 I)^{-1}Y^t = Y[U^t(D+\lambda\rho^2 I)U]^{-1}Y^t =$$

$$YU^{-1}(D+\lambda\rho^2 I)^{-1}(U^t)^{-1}Y^t = YU^t(D+\lambda\rho^2 I)^{-1}UY^t = (YU^t)(D+\lambda\rho^2 I)^{-1}(YU^t)^t$$

and, denoting $Z = YU^t$, this is equal to

$$\frac{Z_1^2}{d_{11}+\lambda\rho^2}+\frac{Z_2^2}{d_{22}+\lambda\rho^2}+...+\frac{Z_n^2}{d_{nn}+\lambda\rho^2}$$

so, the expression for the probability can be written as

$$\frac{1}{\pi^{\frac{n}{2}}}\frac{\lambda^{\frac{n}{2}}\exp(-\lambda(\frac{Z_1^2}{d_{11}+\lambda\rho^2}+\frac{Z_2^2}{d_{22}+\lambda\rho^2}+...+\frac{Z_n^2}{d_{nn}+\lambda\rho^2}))}{\sqrt{(d_{11}+\lambda\rho^2)(d_{22}+\lambda\rho^2)...(d_{nn}+\lambda\rho^2)}} \tag{7}$$

This expression is easier to compute than Equation 6, as its computation for any range of λ and σ requires the inversion of matrix A only once. the next goal is to maximize this expression. This can be turned into an easier, one-dimensional, optimization problem as follows. Let us substitute u for λ and v for $\lambda\rho^2$. Then, we have

$$\frac{1}{\pi^{\frac{n}{2}}}\frac{u^{\frac{n}{2}}\exp(-u(\frac{Z_1^2}{d_{11}+v}+\frac{Z_2^2}{d_{22}+v}+...+\frac{Z_n^2}{d_{nn}+v}))}{\sqrt{(d_{11}+v)(d_{22}+v)...(d_{nn}+v)}} = \frac{1}{\pi^{\frac{n}{2}}}\frac{u^{\frac{n}{2}}}{K_2(v)}exp(-uK_1(v))$$

where the definition of $K_1()$ and $K_2()$ is the obvious one. keeping v constant, we can maximize over u (discarding for the moment quantities which depend only on v):

$$\frac{\partial}{\partial u}(u^{\frac{n}{2}}exp(-uK_1(v))) = \frac{n}{2}u^{\frac{n}{2}-1}exp(-uK_1(v)) - K_1(v)u^{\frac{n}{2}}exp(-uK_1(v))$$

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

$$\frac{1}{\pi^{\frac{n}{2}}}\frac{[\frac{n}{2K_1(v)}]^{\frac{n}{2}}}{K_2(v)}exp(-\frac{n}{2K_1(v)}K_1(v)) = \frac{(\frac{n}{2\pi e})^{\frac{n}{2}}}{K_2(v)[K_1(v)]^{\frac{n}{2}}}$$

maximizing this is equivalent to minimizing $K_2^2(v)[K_1(v)]^n$, or

$$(v+d_{11})(v+d_{22})...(v+d_{nn})\left[\frac{Z_1^2}{v+d_{11}}+\frac{Z_2^2}{v+d_{22}}+...+\frac{Z_n^2}{v+d_{nn}}\right]^n \tag{8}$$

after the optimal v is found, one can easily extract the optimal λ, which is equal to $\frac{n}{2K_1(v)}$, and ρ, which is $\sqrt{\frac{2vK_1(v)}{n}}$, hence the optimal σ is $\sqrt{\frac{vK_1(v)}{n}}$

Next, we touch on some asymptotic properties of the optimal λ and ρ. The first one is trivial, but offers some insight into the meaning of the prior defined on the function space:

Lemma 1 $\lim_{\lambda \to 0} Pr(M_{\lambda,\sigma}) = 0$, *no matter what the data is.*

Proof: this is obvious from equation 6.

Intuitively, this is true because as λ tends to zero, the prior distribution on $\{f(x_1), f(x_2)...f(x_n)\}$ (which always has zero mean) tends to have a covariance matrix with infinite entries (recall that this covariance matrix is $\frac{A}{2\lambda}$). Since the sampled values are finite, the probability of them having been sampled from such a distribution tends to zero.

Lemma 2 *As the optimal choice of $v = \lambda\rho^2$ approaches infinity, the optimal noise variance approaches the average of the squared norm of the data, $\frac{\|Y\|^2}{n}$.*

Proof: it was shown before that the optimal choice of ρ^2 is $\frac{2vK_1(v)}{n}$. Substituting the expression for $K_1(v)$ yields

$$\frac{2v[\frac{Z_1^2}{v+d_{11}} + \frac{Z_2^2}{v+d_{22}} + ... + \frac{Z_n^2}{v+d_{nn}}]}{n} = \frac{2[\frac{Z_1^2}{1+\frac{d_{11}}{v}} + \frac{Z_2^2}{1+\frac{d_{22}}{v}} + ... + \frac{Z_n^2}{1+\frac{d_{nn}}{v}}]}{n}$$

which tends to $2\frac{\|Z\|^2}{n}$ as v tends to infinity. Since Z has the same norm as Y and $\rho^2 = 2\sigma^2$, the result follows.

The intuition behind this is also straight-forward. Recall that the reconstructed spline minimizes $M(f) = M(f) = \sum_{i=1}^{n} \frac{[f(x_i) - y_i]^2}{\sigma_i^2} + \lambda \int_0^1 [f^{''}(v)]^2 dv = \frac{1}{\sigma^2}(\sum_{i=1}^{n}[f(x_i) - y_i]^2 + \lambda\sigma^2 \int_0^1 [f^{''}(v)]^2 dv)$. So, as $\lambda\sigma^2$ tends to infinity, the spline is forced to be a linear function. However, we are working in the space of functions which assume zero at the end points of the interval; so, the reconstructed spline will be zero, which really means the data is just noise; so, the method described here chooses the magnitude of the noise to be equal to the magnitude of the data.

Lemma 3 *For any choice of data Y there is a λ_0 such that is $\lambda < \lambda_0$ the probability of the model $M_{\lambda,\sigma}$ is monotonically decreasing in σ.*

Proof: let us hold λ constant and look for the optimal ρ. Then $\lambda^{\frac{n}{2}}$ can be discarded, being a constant, and the expression for the probability reduces to $\frac{\exp(-\lambda K_1(\lambda\rho^2))}{K_2(\lambda\rho^2)}$. To simplify, let us square this expression (makes no difference to the property of monotonicity) and also replace ρ^2 by μ (also makes no difference). Then the expression assumes the form $\frac{\exp(-\lambda K_1(\lambda\mu))}{K_2^2(\lambda\mu)}$, or

$$\frac{\exp(-2\lambda[\frac{Z_1^2}{d_{11}+\lambda\mu} + \frac{Z_2^2}{d_{22}+\lambda\mu} + ... + \frac{Z_n^2}{d_{nn}+\lambda\mu}])}{(d_{11}+\lambda\mu)(d_{22}+\lambda\mu)...(d_{nn}+\lambda\mu)}$$

differentiating this with respect to μ results in a cumbersome expression; however, its sign is equal to the sign of

$$\frac{(2\lambda Z_1^2 - d_{11}) - \lambda\mu}{(\lambda\mu + d_{11})^2} + \frac{(2\lambda Z_2^2 - d_{22}) - \lambda\mu}{(\lambda\mu + d_{22})^2} + \ldots + \frac{(2\lambda Z_n^2 - d_{nn}) - \lambda\mu}{(\lambda\mu + d_{nn})^2} \tag{9}$$

and if $2\lambda Z_i^2 < d_{ii}$ for every i, the derivative is always negative, and the probability therefore decreases with ρ, or σ. As an auxiliary result, we get that if $\mu > 2\max_i(Z_i^2)$ (or, equivalently, $\sigma^2 > \max_i(Z_i^2)$) the probability is decreasing; so, for a given λ, we have presented a lower and upper bound on the optimal choice of σ. One cannot hope to prove a similar result for λ, as its optimal value can be arbitrarily large or small.

These two results can be explained as follows. As λ approaches zero, the prior on the function space becomes more and more "relaxed", allowing any data set – even one that oscillates a lot – to have a relatively large probability. At some stage, the prior becomes "too relaxed for the data". This happens when the covariance matrix of the prior, $\frac{A}{2\lambda}$, assumes values which are large relative to the data. In that case, adding noise only increases the covariances, and the probability decreases.

The second observation is easier to explain; intuitively, it means that there is no reason to assume that the measurement noise is larger than the data.

4 The GCV Algorithm and How it Compares With the Bayesian Approach

A common method for determining λ is *cross-validation* [5]. 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

$$\sum_{i \neq k}^{l} [f(x_i) - y_i]^2 + \lambda \int [f''(\xi)]^2 d\xi$$

e.g. the spline interpolating all the data points but the k-th. $V_0(\lambda)$ is then defined as $\sum_{k=1}^{l} [f_k(x_k) - y_k]^2$, and the λ chosen is the one minimizing $V_0()$.

An extension of this method is the *Generalized cross-validation* (GCV) [5] which proceeds as follows. First, since the spline 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

$$\begin{pmatrix} f(x_1) \\ . \\ . \\ . \\ f(x_l) \end{pmatrix} = A(\lambda) \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_l \end{pmatrix} \tag{10}$$

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

$$V(\lambda) = \sum_{k=1}^{l} w_k(\lambda)[f_k(x_k) - y_k]^2, \quad w_k(\lambda) = [\frac{1 - A_{kk}(\lambda)}{\frac{1}{n}T_r(I - A(\lambda))}]^2$$

(T_r stands for Trace and I for the $l \times l$ identity matrix). The λ chosen is the one minimizing $V()$. In this work, we have used a version of the cross validation algorithm written by M.F. Hutchinson, and described in the ACM. Trans. Math. Software Vol 12, No. 2, June 1986, p. 150. The algorithm uses the GCV algorithm described in [5], and it gives identical results to the GCV version we implemented. Experiments were run on a Sparc work station. One can see that when the data strongly oscillates, the GCV algorithm can run into grave problems. Before giving some actual results, let us explain why this happens.

Look at Figure 1 (all figures appear at the end of the paper). The data to be interpolated consists of a sinusoidal pattern, sampled at the peaks. When choosing the smoothing parameter λ, the GCV consecutively discards each sample points, reconstructs the spline from the remaining ten points, and measures how much it deviates from the discarded point. These measures are weighted differently, but this is not a crucial matter, as in this case (equally spaced sample points) the weights are almost similar. The goal is to minimize the sum of the deviation over all the points. Figure 1 demonstrates what happens when the eighth point $(0.7, -10)$ is discarded. If λ is small, the deviation between the value of the spline reconstructed using the other ten points and between the value of the original measurement (-10) is large, because the spline obtains a large positive value at the point $x = 0.7$. On the other hand, when λ is large, this deviation is smaller, because the reconstructed spline tends to be a straight line and it doesn't miss $(0.7, -10)$ by such a large amount. This is true of the other sample points. Consequently, GCV chooses a very large λ in this case, and the resulting interpolant is a straight line.

A similar phenomena occurred when the GCV program mentioned before was tested. This program returns the λ it chose, as well as its estimate to the variance of the noise present at the measurements and the resulting spline. In order to compare the output to that of the method presented in this paper, which assumes the functions are equal to 0 at the ends of the interval, the variance of the noise at the interval's ends was defined to be very small compared to the (uniform) noise on the rest of the interval. In Figure 2 are the results of running the GCV algorithm with an input resembling that of Figure 1. "Plot2" (the little triangles) show the data points, and "Plot 1" (the straight line) shows the interpolant found by GCV algorithm.

As can be seen, the GCV chose a very large value of λ; actually, it does not return the value of λ but that of $\frac{\lambda}{1+\lambda}$, and that was equal to 1.0, which corresponds to $\lambda = \infty$. The variance of the noise returned by the program is 25. In other words, the GCV decides that the sinusoidal pattern resulted from a flat signal with noise added to it.

It is even more striking to compare this result with what happens when the sinusoidal pattern above is replaced by a set of measurements which results when each sample point of the original pattern is replaced by two points which are very close to each other. The corresponding data and reconstruction obtained when applying the GCV algorithm are displayed in Figure 3. Again, 'Plot2" (the little triangles) show the data points, and "Plot 1" shows the interpolant found by GCV algorithm.

In this case, the reconstruction agrees with the data, and the variance of the data is estimated as zero, while λ is also chosen to be very small! Hence, GCV makes two very different decisions with regard to two sets of data which look very much alike.

Why does this happen? The reason is that, for the data set in Figure 1, GCV chooses the optimal λ as infinity (as does our method), and then finds the function f that maximizes $Pr(f/M_\infty)$. This f is, of course, the zero function, because every non-zero function has a zero probability in M_∞. However, there are many other functions that have a non-zero probability for other priors, for which the Bayesian probability is hardly smaller than that of M_∞! To illustrate this point, let us look at a graph of the probability of $M_{\lambda,\sigma}$, given the data of Figure 1, for a large range of values (Figure 4): It is clear that the probability is rather "flat" in the λ direction, indicating that functions other than the zero function have a relatively large probability, when integrating over all values of λ.

In order to demonstrate how the probability for λ and σ look for arbitrary data, here is its plot for such data, which is given in Figure 5, together with the reconstruction for the optimal values of λ (0.000277) and σ (0.497). Figure 6 shows a portion of the graph for optimizing the value of $\lambda\rho^2$ (Equation 8). Recall that here, the optimal value is the minima of the function. In Figure 7, the joint probability for λ and σ is plotted.

The expression to be optimized for v (Equation 8) is

$$(0.112+v)(0.007+v)(0.001+v)(0.0004+v)(0.00018+v)(0.00009+v)(0.00005+v)(0.000030+v)(0.000021+v)(0.000017+v)\cdot$$
$$\left(\frac{201.98}{0.113+v}+\frac{0.460}{0.007+v}+\frac{6.266}{0.0014+v}+\frac{1.655}{0.00044+v}+\frac{0.225}{0.00018+v}+\frac{0.037}{0.00009+v}+\frac{0.026}{0.00005+v}+\frac{0.79}{0.00003+v}+\frac{0.50}{0.00002+v}+\frac{0.049}{0.000017+v}\right)^{10}$$

and the expression to be minimized for λ and σ simultaneously is

$$\frac{\lambda^5 e^{-\lambda\left(\frac{201.9}{0.11+2\lambda\sigma^2}+\frac{0.46}{0.007+2\lambda\sigma^2}+\frac{6.26}{0.0013+2\lambda\sigma^2}+\frac{1.65}{0.00044+2\lambda\sigma^2}+\frac{0.22}{0.00018+2\lambda\sigma^2}+\frac{0.037}{0.00009+2\lambda\sigma^2}+\frac{0.025}{0.000049+2\lambda\sigma^2}+\frac{0.79}{0.00003+2\lambda\sigma^2}+\frac{0.50}{0.000021+2\lambda\sigma^2}+\frac{0.049}{0.000016+2\lambda\sigma^2}\right)}}{\sqrt{(0.11+2\lambda\sigma^2)(0.007+2\lambda\sigma^2)(0.0013+2\lambda\sigma^2)(0.00044+2\lambda\sigma^2)(0.00018+2\lambda\sigma^2)(0.00008+2\lambda\sigma^2)(0.000049+2\lambda\sigma^2)(0.00003+2\lambda\sigma^2)(0.000021+2\lambda\sigma^2)(0.000017+2\lambda\sigma^2)}}$$

which, as was shown before, is really equivalent to an easier one-dimensional optimization.

5 Numerically Estimating λ

This section deals with the following practical problem: given that a certain function (or functions) were sampled from a prior distribution satisfying $Pr(f) \propto \exp(-\lambda \int [f''(v)]^2 dv)$, estimate λ. As will be shown, the accuracy of the estimate presented here depends on the amount of measurement noise present, as well as on the number and location of the sample points.

In the simplest case, when the measurement noise is known to be zero, it is possible to explicitly write down the MAP estimate for λ; it is the value which maximizes the probability given in Equation 6, or $\lambda_{MAP} = \frac{n}{2YA^{-1}Y^t}$; from now on, we will use the standard notation $\hat{\lambda}$. The amount of trust one assigns to this estimate is equal to $\int_{\mathcal{R}^n}(\hat{\lambda}-\frac{n}{2YA^{-1}Y^t})^2 Pr(Y/M_\lambda)dY$. However, computing this intergal is difficult due to convergence problems; therefore, we estimate $\frac{1}{\lambda}$, for which the estimate above translates to $\frac{2YA^{-1}Y^t}{n}$. We will look for a slightly more general estimate, $\hat{(\frac{1}{\lambda})} = \beta YA^{-1}Y^t$. The quality of this estimate in the presence of noise with standard deviation σ is

$$\int_{\mathcal{R}^n}[\frac{1}{\lambda} - \beta YA^{-1}Y^t]^2 Pr(Y/M_\lambda, \sigma)dY$$

according to Equation 6 this is equal to (recall that $\rho = \sqrt{2}\sigma$)

$$\frac{\lambda^{\frac{n}{2}}}{\pi^{\frac{n}{2}}|A+\lambda\rho^2 I|^{\frac{1}{2}}}\int_{\mathcal{R}^n}[\frac{1}{\lambda} - \beta YA^{-1}Y^t]^2 \exp(-\lambda(Y(A+\lambda\rho^2 I)^{-1}Y^t))dY$$

after some manipulations, this integral turns out to be

$$\frac{1}{\lambda^2} - \frac{\beta}{\lambda}\sum_i \frac{1}{\lambda_i} + \beta^2(\sum_{i<j}\frac{1}{2\lambda_i\lambda_j} + \frac{3}{4}\sum_i \frac{1}{\lambda_i^2})$$

where $\lambda_i = \frac{\lambda}{1+\lambda\rho^2 d_i}$, and $\{d_i\}_{i=1}^n$ are the eigenvalues of A^{-1}.

If no noise is present ($\rho = 0$) this reduces to

$$\frac{1}{\lambda^2}(1 - \beta n + \frac{n(n+2)}{4}\beta^2)$$

since we are trying to estimate the variance of the error for $\frac{1}{\lambda}$, it makes sense to study only the expression $(1 - \beta n + \frac{n(n+2)}{4}\beta^2)$. It achieves a minimum value of $\frac{2}{n+2}$, for $\beta = \frac{2}{n+2}$. So, when the number of sample points grows, the variance of the estimator decreases in a rate which, asymptomatically, is optimal (due to the Cramer-Rao bound). It is interesting to note that in the noiseless case, the goodness of the estimate is not affected by the location of the sample points. The situation is rather different for the case where noise is present. Currently, we are still studying the general problem, and here we shall restrict ourselves to a special case, where there are three sample points.

The variance of the error of the estimator $\beta YA^{-1}Y^t$ is

$$1 - 3\beta - \beta\lambda\rho^2 d_1 - \beta\lambda\rho^2 d_2 - \beta\lambda\rho^2 d_3 + \frac{15\beta^2}{4} + \frac{5\beta^2\lambda\rho^2 d_2}{2} + \frac{5\beta^2\lambda\rho^2 d_1}{2} + \frac{\beta^2\lambda^2\rho^4 d_1 d_2}{2}$$
$$+\frac{5\beta^2\lambda\rho^2 d_3}{2} + \frac{\beta^2\lambda^2\rho^4 d_1 d_3}{2} + \frac{\beta^2\lambda^2\rho^4 d_2 d_3}{2} + \frac{3\beta^2\lambda^2\rho^4 {d_1}^2}{4} + \frac{3\beta^2\lambda^2\rho^4 {d_2}^2}{4} + \frac{3\beta^2\lambda^2\rho^4 {d_3}^2}{4}$$

if we want to choose a β minimizing this, we have to "integrate out" λ and ρ. This leaves us with a function of β, and also of d_1, d_2, d_3 (which we assume are known, as they can be evaluated from the sample points). We can minimize this expression for β; the result is a complicated expression which we shall not present here. However, it is instructive to compare the behavior of the estimator for various values of λ and ρ for two different sets of sample points. If we sample at $\{0.25, 0.5, 0.75\}$, we have a better estimate of λ than if we sample at $\{0.1, 0.11, 0.12)\}$; this is because when the sample points are very close to each other, the difference between the values of the sampled function at them is small relative to the noise, and it hard to tell if the differences are a result of the noise or of the inherent function itself. This also explains why the error of the estimator grows when the noise grows.

Figure 8 shows the relative error of the estimator as a function for λ and σ when the sample points are $\{0.25, 0.5, 0.75\}$, and Figure 9 for samples taken at $\{0.1, 0.11, 0.12)\}$.

6 Computing the Prior for λ

use the fact that the integral of the second derivative squared bounds the function to demonstrate that for large lambdas, all functions which are not contained in a small strip around y=0 have a very small probability.

the criterion to differentiate between spaces is the average smoothness which turned out to be $\frac{1}{\lambda^{3/2}}$. So, the prior is taken to be $\frac{1}{\lambda^{5/2}}$.

There is no prior for σ (if no information is given).

7 Computing the Expectation of the value at x

$$E[f(x)/D] = \int_\sigma \int_\lambda E[f(x)/D, M_{\lambda,\sigma}] Pr(M_{\lambda,\sigma}) d\lambda d\sigma =$$

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

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

$$\frac{\int_v \int_u \frac{1}{\sqrt{v}} (H_{x_1}(x)...H_{x_n}(x))(A+vI)^{-1} Y^t \frac{u^{\frac{n-6}{2}}}{\pi^{\frac{n}{2}}} |A+vI|^{-\frac{1}{2}} \exp(-uY(A+vI)^{-1} Y^t) du dv}{\int_v \int_u \frac{1}{\sqrt{v}} \frac{u^{\frac{n-6}{2}}}{\pi^{\frac{n}{2}}} |A+vI|^{-\frac{1}{2}} \exp(-uY(A+vI)^{-1} Y^t) du dv}$$

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

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

8 Examples

Next, we demonstrate how the proposed approach results in stable fits. A simple pattern – one cycle of a sinusoidal function - is contaminated with noise, and then the resulting data is smoothed using the GCV algorithm and using the method suggested in the previous section. 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 (Figures 10 and 11). The results of the suggested method are given in Figure 12.

9 Conclusions and Further Research

A novel approach for looking at interpolation is suggested, which evaluates the MAP function given the data, by maximizing the integral of each possible interpolant over all possible models, scaled by the probability of each model. It is demonstrated why "classical" regularization fails to find the MAP estimate. The probability of the models is calculated, allowing one to rigorously determine the optimal parameters for regularization. Estimates for one of the optimal weights are studied.

In the future, we plan to demonstrate how the integral over all the weights is computed, and also study other priors/models, notably those that belong to the realm of "robust statistics". These pose a fascinating mathematical challenge because, usually, the probability they assign to models can not be expressed in terms of inner products on the function spaces involved.

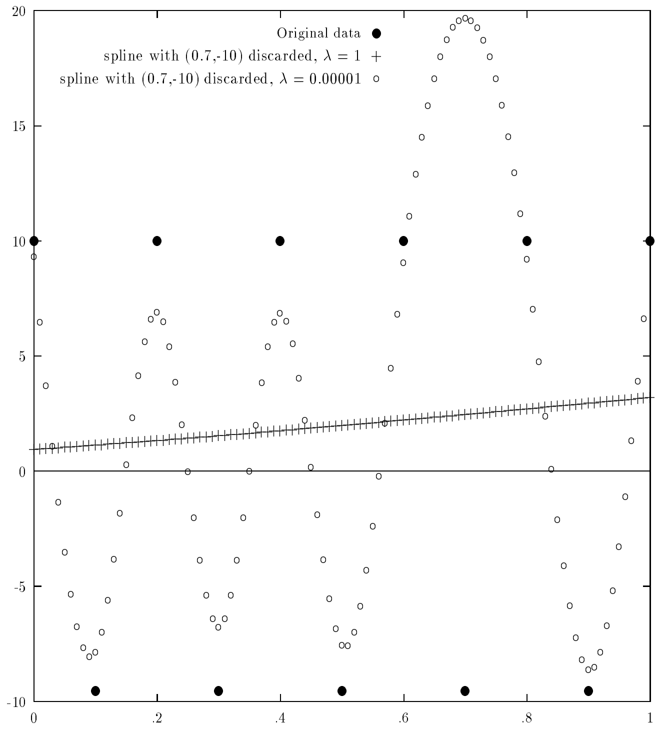

Figure 1: Why GCV chooses large λ for certain data

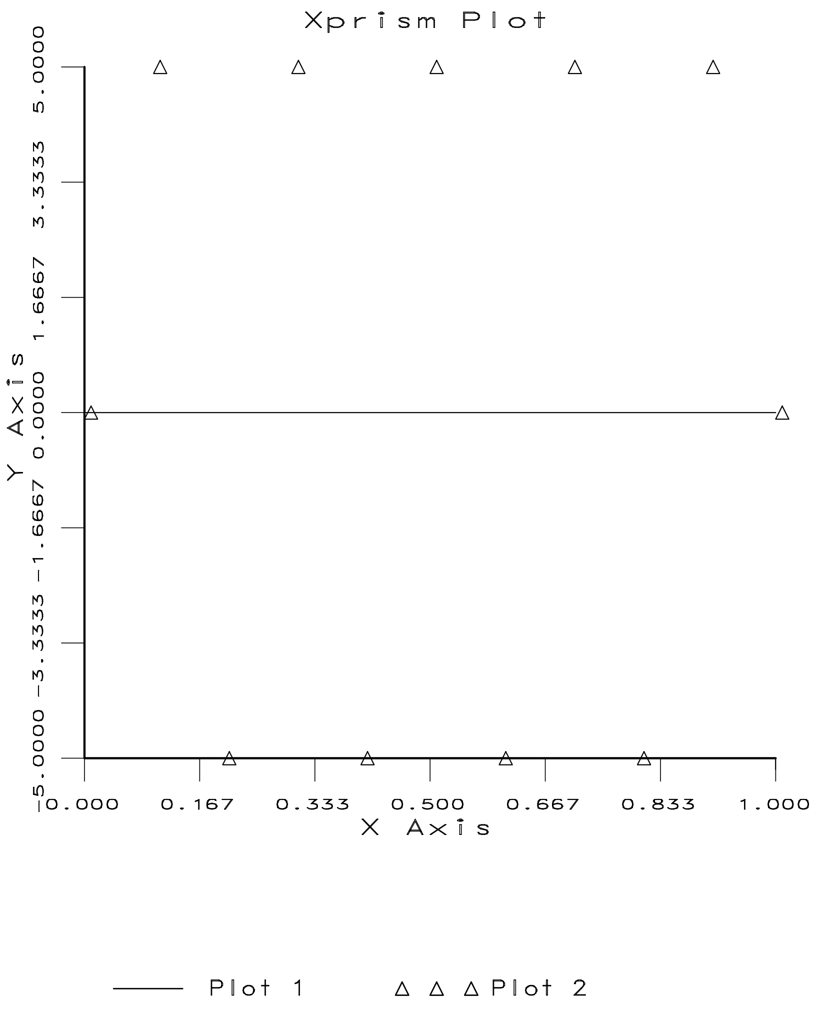

Figure 2: GCV reconstruction

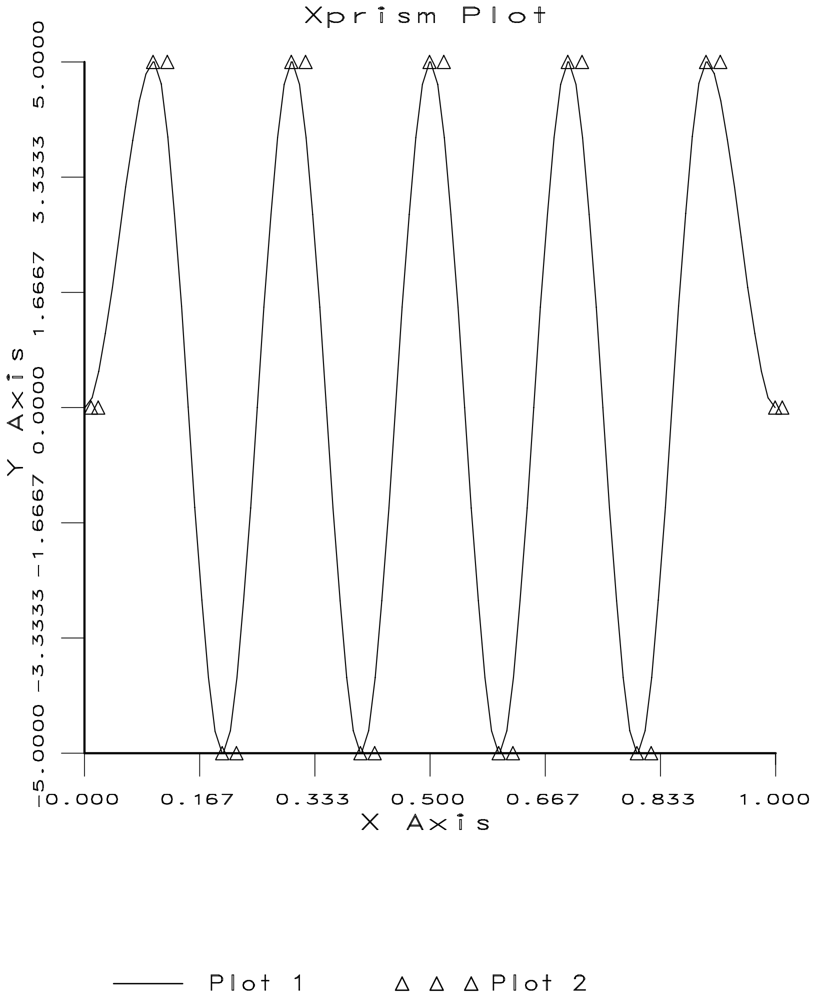

Figure 3: GCV reconstruction for data which is very similar to data of Fig. 2

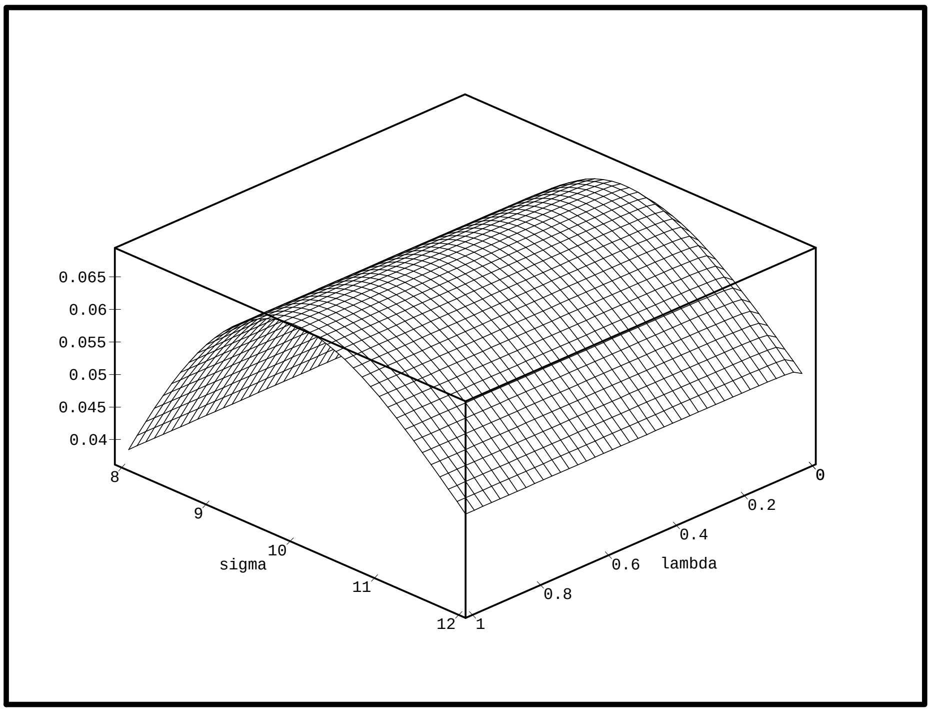

Figure 4: Probability distribution for λ and σ, given data of Fig. 2

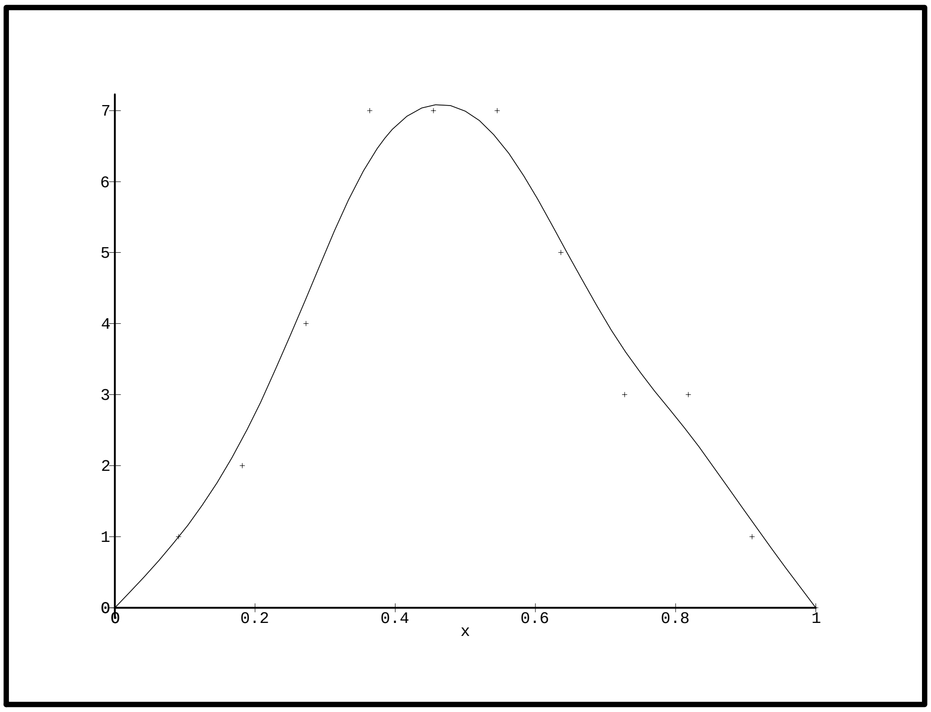

Figure 5: Data and reconstruction with optimal λ and σ

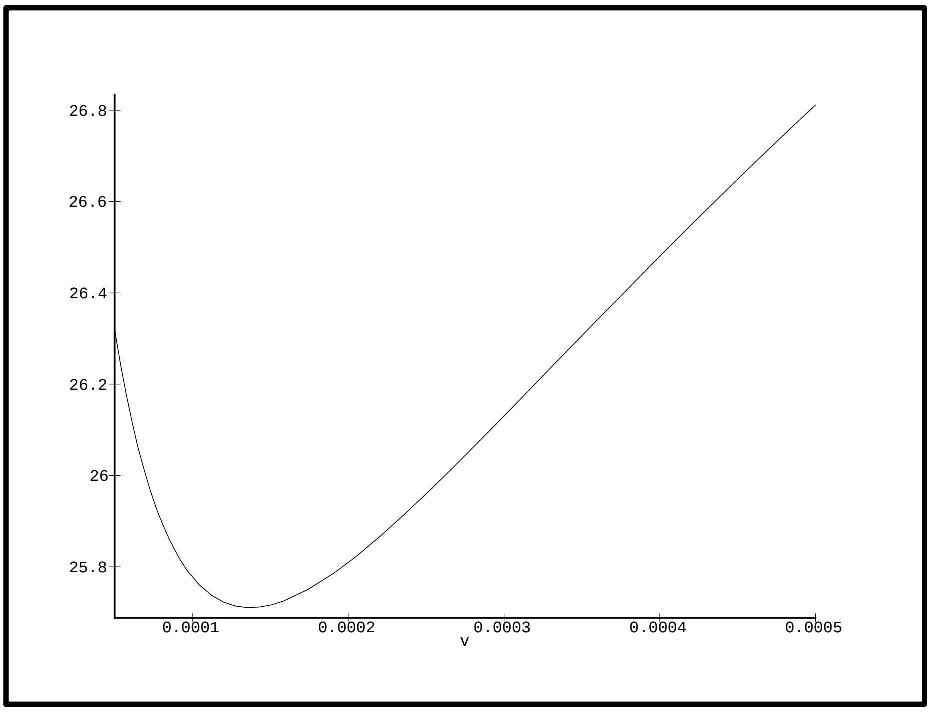

Figure 6: Graph of expression in Eq. 8, for optimizing $v = \lambda\rho^2$ (data of Fig. 5)

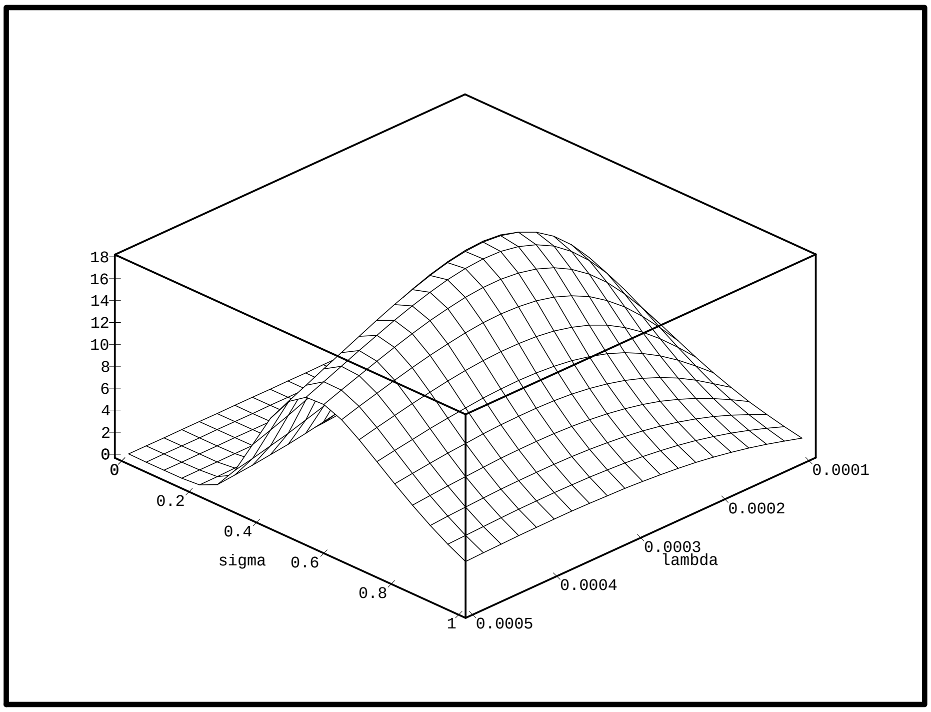

Figure 7: Graph of joint Probability for λ and σ (data of Fig. 5)

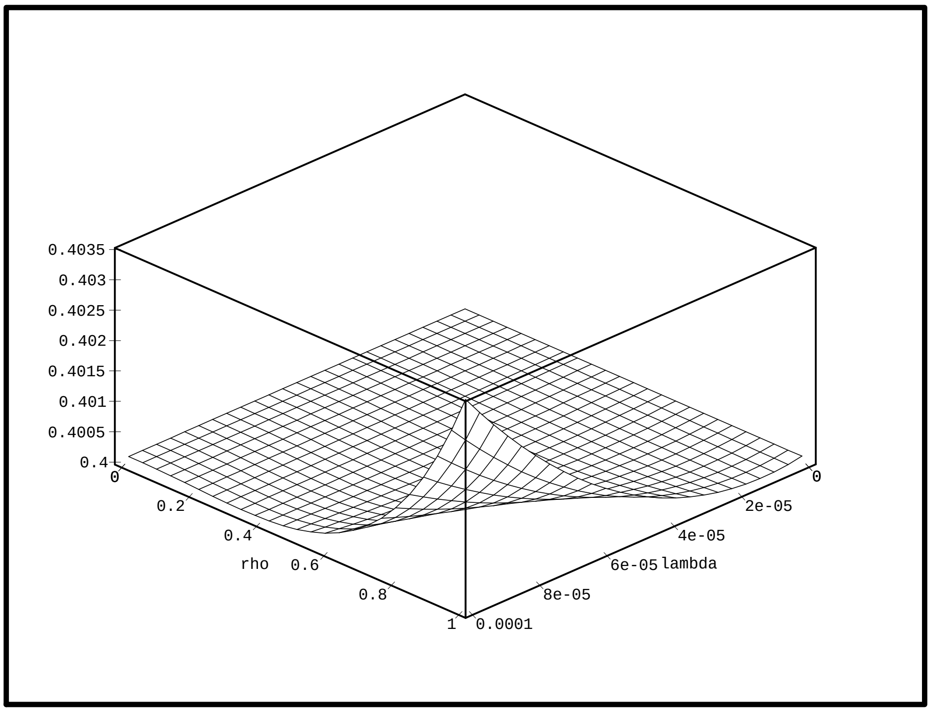

Figure 8: Error of estimator for $\frac{1}{\lambda}$, sample points=$\{0.25, 0.5, 0.75\}$.

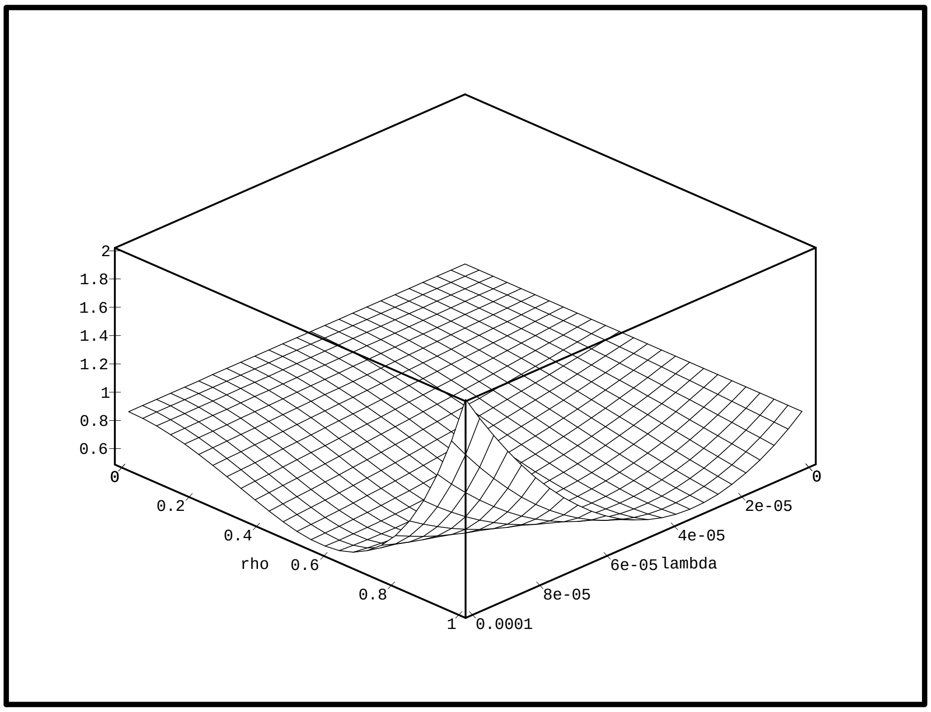

Figure 9: Error of estimator for $\frac{1}{\lambda}$, sample points=$\{0.1.0.11, 0.12\}$.

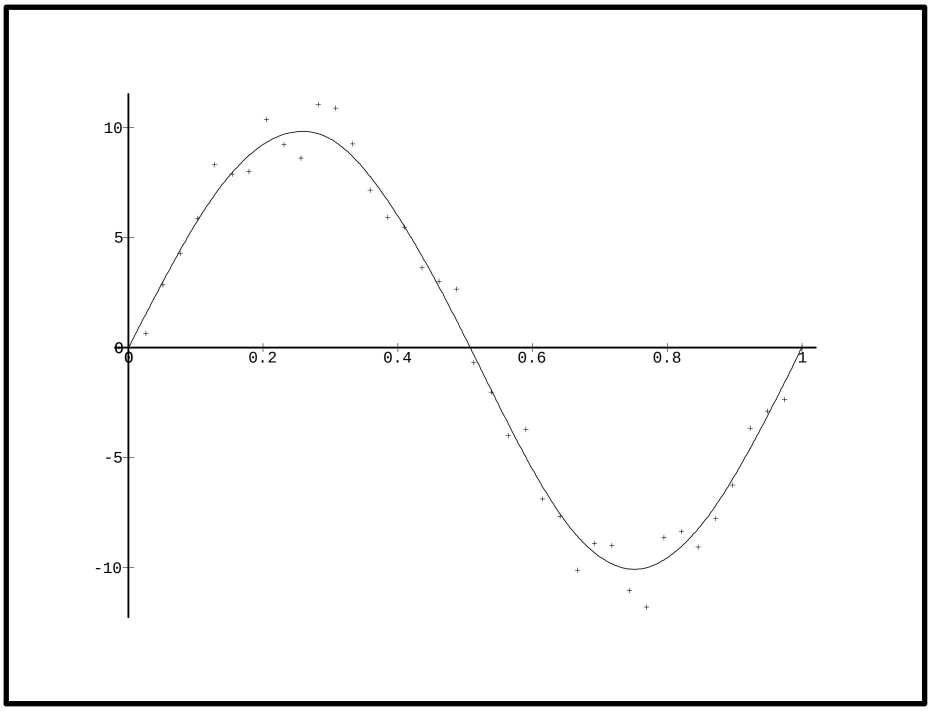

Figure 10: GCV chooses a “standard” value of λ.

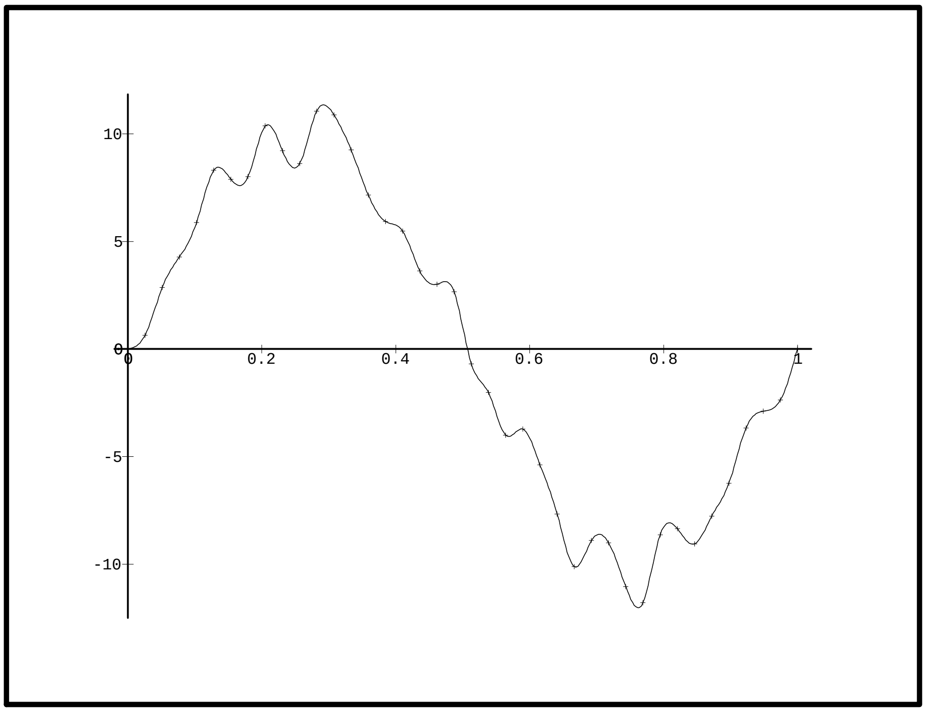

Figure 11: For a data set differing from that of Figure 10 in only one point, GCV chooses a very small value of λ, resulting in a completely different fit.

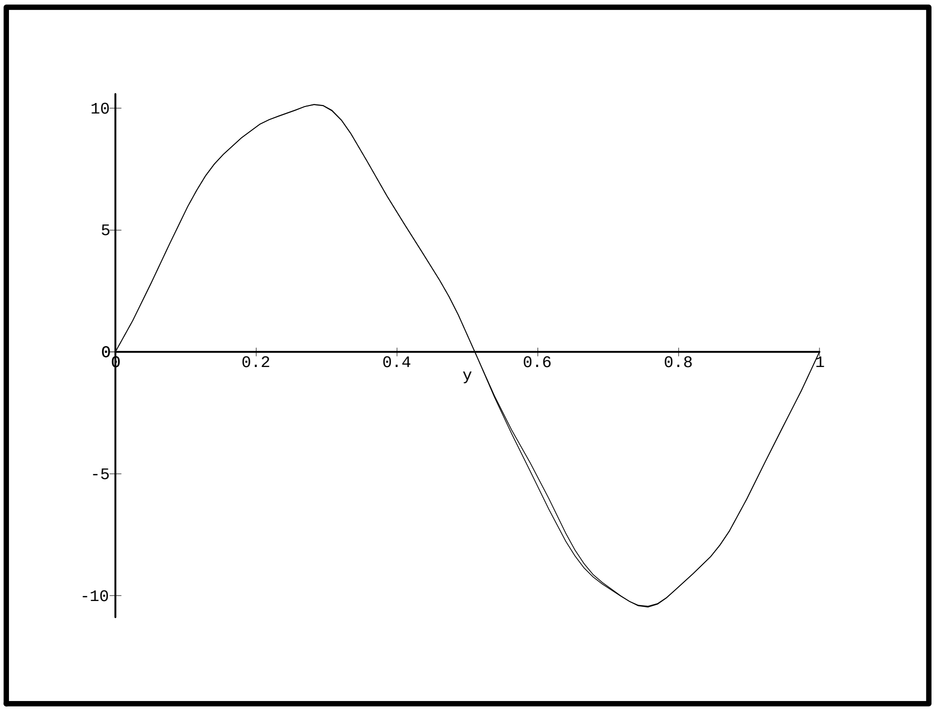

Figure 12: The suggested method for fitting, used on the data sets of Figures 10 and 11. Fits are almost identical.

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*, 17:26–31, 1974.

[3] M. Bertero, T.A Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 8:869–889, 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*, 31:377–403, 1979.

[6] S. Geman and D.Geman. Stochastic relaxation, gibbs distribution, and the bayesian restorat ion of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–741, June 1984.

[7] L. Gross. Integration and non-linear transformations in hilbert space. *Transactions of the American Mathematical Society*, 94:404–440, 1960.

[8] E. Hille. Introduction to the general theory of reproducing kernels. *Rocky Mountain Journal of Mathematics*, 2:321–368, 1972.

[9] B. Horn. *Robot Vision.* MIT Press, 1986.

[10] B.K.P Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[11] D. Keren and M. Werman. Variations on regularization. In *10'th International Conference on Pattern Recognition*, Atlantic City, 1990.

[12] D. Keren and M. Werman. Probabilistic analysis of regularization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15:982–995, October 1993.

[13] Daniel Keren. *Probabilistic Analyses of Interpolation in Computer Vision.* PhD thesis, Hebrew University of Jerusalem, 1990.

[14] J. Kuelbs, F.M. Larkin, and J.A. Williamson. Weak probability disributions on reproducing kernel hilbert spaces. *Rocky Mountain Journal of Mathematics*, 2:369–378, 1972.

[15] H.H Kuo. *Guassian Measures in Banach Spaces.* Springer-Verlag, 1975.

[16] F.M. Larkin. Gaussian measure in hilbert space and applications in numerical analysis. *Rocky Mountain Journal of Mathematics*, 2:379–421, 1972.

[17] Y.G. Leclerc. Image and boundary segmentation via minimal-length encoding on the co nnection machine. In *Image Understanding Workshop*, pages 1056–1069, 1989.

[18] David J.C. MacKay. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, 1992.

[19] J.L Marroquin. Deterministic bayesian estimation of markovian random fields with app lications to computational vision. In *International Conference on Computer Vision*, pages 597–601, London, May 1987.

[20] L. Matthies, R. Szeliski, and T. Kanade. Incremental estimation of dense depth maps from image sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 366–374, Ann Arbor, June 1988.

[21] D. Mumford and J. Shah. Boundary detection by minimizing fuctionals. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 22–26, San Francisco, June 1985.

[22] 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.*, 53:2341–2367, 1969.

[23] R. Szeliski. Regularization uses fractal priors. In *National Conference on Artificial Intelligence*, pages 749–754, 1987.

[24] R. Szeliski. *Bayesian Modeling of Uncertainty in Low-Level Vision*. Kluwer, 1989.

[25] S. Szeliski and D. Terzopoulos. From splines to fractals. In *SIGGRAPH*, pages 51–60, 1989.

[26] D. Terzopoulos. Multi-level surface reconstruction. In A. Rosenfeld, editor, *Multiresolution Image Processing and Analysis*. Springer-Verlag, 1984.

[27] D. Terzopoulos. Regularization of visual problems involving discontinuities. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8:413–424, August 1986.

[28] A.N Tikhonov and V.Y Arsenin. *Solution of Ill-Posed Problems*. Winston and Sons, 1977.

[29] G.W. Wasilkowski. Optimal algorithms for linear problems with gaussian measures. *Rocky Mountain Journal of Mathematics*, 16:727–749, 1986.

[30] N. Young. *An Introduction to Hilbert Space*. Cambridge Mathematical Textbooks, 1988.