DATA-DRIVEN PRIORS FOR HYPERPARAMETERS IN REGULARIZATION

DANIEL KEREN
Department of Mathematics, The University of Haifa
Haifa 31905, Israel [‡]
AND
MICHAEL WERMAN
Institute of Computer Science, The Hebrew University
Jerusalem 91904, Israel [§]

Abstract. 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 is to select weights that should be assigned to these two terms, and minimize the resulting error functional. However, using only these "optimal weights" does not guarantee that the chosen function will be optimal in some sense. 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.

Key words: Bayesian estimation, regularization, Gaussian measures on function spaces.

1. Introduction

In many areas of science and engineering, regularization [1,2] is used to reconstruct functions from partial data. In the field of maximum entropy, a similar idea is used for reconstruction of missing or corrupted data [3–7].

Regularization chooses among the possible functions one which approximates the given data and is also "smooth". A *cost functional* $M(f)$ is defined for every

[‡] Email: dkeren@mathcs2.haifa.ac.il

[§] Email: werman@cs.huji.ac.il

K. M. Hanson and R. N. Silver (eds.), Maximum Entropy and Bayesian Methods, 77–84.

function 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 $M(f) = \sum_{i=1}^{n} \frac{[f(x_i) - y_i]^2}{2\sigma^2} + \lambda \int_0^1 f_{uu}^2 du$. Due to lack of space only the one-dimensiona case will be presented here, however this work was extended and applied to functions of two variables as well.

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 Caussian noise model with variance σ^2, $Pr(f/D) \propto \frac{1}{\sigma^n} \exp(-\frac{1}{2\sigma^2} \sum [f(x_i) - y_i]^2)$. 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.

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. However, all regularization schemes we are familiar with choose one combination of weights and use them alone to interpolate the function; but, this approach fails to find the maximum likelihood (MAP) estimate for the interpolant f, as it uses only one set of weights λ and σ to construct f. However, 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.

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 this integral 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 [8,9]; see also an example of such a data set and the corresponding probability distribution it induces on the weights, in this paper (Figure 4).

In this paper, it will be shown how to find the function f maximizing $\int_w Pr(f/D, w) Pr(w/D) dw$.

We also address the questions of computing the MSE function, and the pointwise uncertainty associated with it.

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.

2. Previous Work

A very popular method for determining the smoothing parameter λ is Generalized Cross Validation, GCV (bootstrapping) [10,2]. In [11], a few methods for choosing the smoothing parameter are analyzed.

A different approach, which also chooses an "optimal" smoothing parameter and uses it, is that of Bayesian model selection which, to the best of our knowledge, was first suggested in the pioneering work of Szeliski [9]. 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 [8]. Another method for choosing the smoothing parameter is presented in [12]. In [13], 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 Figure 1). 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. Nontheless, 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 [3,4]. 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 calculation of the MSE estimate given here.

Finally, recent work reported in [14,15] 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 for them is assumed.

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

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

where the $\sqrt{\lambda}$ in the first integral normalizes the probability distribution on the function space [16].

The expression above 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 $F_1(\alpha) = \int_0^\infty \sqrt{\lambda}\exp(-\lambda\alpha)Prior(\lambda)d\lambda$ and $F_2(\beta) = \int_0^\infty \frac{1}{\sigma^n}\exp(-\frac{\beta}{2\sigma^2})Prior(\sigma)d\sigma$.

Note that, obviously, $F_1()$ and $F_2()$ are monotonically decreasing.

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.

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

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

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

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

so, we have to find for which λ this expression equals α. Diagonalizing A by an orthonormal U, $UAU^t = D$, and denoting $Z = UY$, 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 $\sum[f(x_i)-y_i]^2 + \lambda \int f_{uu}^2 du$. Without going into all the technical details, let us just state that this equals $\beta = ||AU^t(D + \lambda I)^{-1}Z - Y||^2$, another expression which 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 $F_1(\alpha)F_2(\beta)$

and we have to maximize $C(\alpha)$ for $0 \le \alpha \le \int (f_{interpolate})^2_{uu} du$, where $f_{interpolate}$ is the interpolant which passes through the data points. This range covers all the relevant functions, because $f_{interpolate}$ is the interpolant of the type we're studying which maximizes $\int f^2_{uu} du$ (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 $Pr(f) \propto \exp(-\lambda \int f^2_{uu} du)$. Under this probability, which is Gaussian, the MSE function, denoted $(f_{opt})_{\lambda,\sigma}$, is equal to the MAP function and there is a closed-form expression for it (given in the previous section). It can be proved that

$$E_x = \int f(x) Pr(f/D)\mathcal{D}f = \int_\lambda \int_\sigma (f_{opt})_{\lambda,\sigma}(x) Pr(M_{\lambda,\sigma}/D) d\lambda d\sigma$$

After computing$Pr(M_{\lambda,\sigma}/D)$, the following expression for E_x can be derived:

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

5. Computing the Uncertainty Associated With the Interpolant

In [17,2,16,8,18,9], 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 [16], defining the uncertainty of the interpolant at the point x as $\int[f(x)-E_x]^2 Pr(f/D)\mathcal{D}f$ As was the case with E_x, we obtain a closed-form solution, but its computation is non-trivial.

6. Examples

A simple pattern – one cycle of a sinusoidal function – is contaminated with Gaussian noise, and then the resulting data is 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 (Figure 1). In Figure 2, The MSE (left) and MAP (right) estimates for these two data sets are presented. In Figure 3, the MSE estimate and confidence intervals for two data sets are given. On the left, the data is a sample of the x-coordinates of a hand-written word. On the right, 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.

Finally, we give an example which explains why one has to integrate over all possible weights. In Figure 4, two data sets are shown, superimposed. As one can see, they are almost identical. Also, the (scaled) joint probability distribution of the weights λ, σ for 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. Therefore, the interpolants for the data sets which use only the most probable weights are drastically different, although the data sets are almost identical.

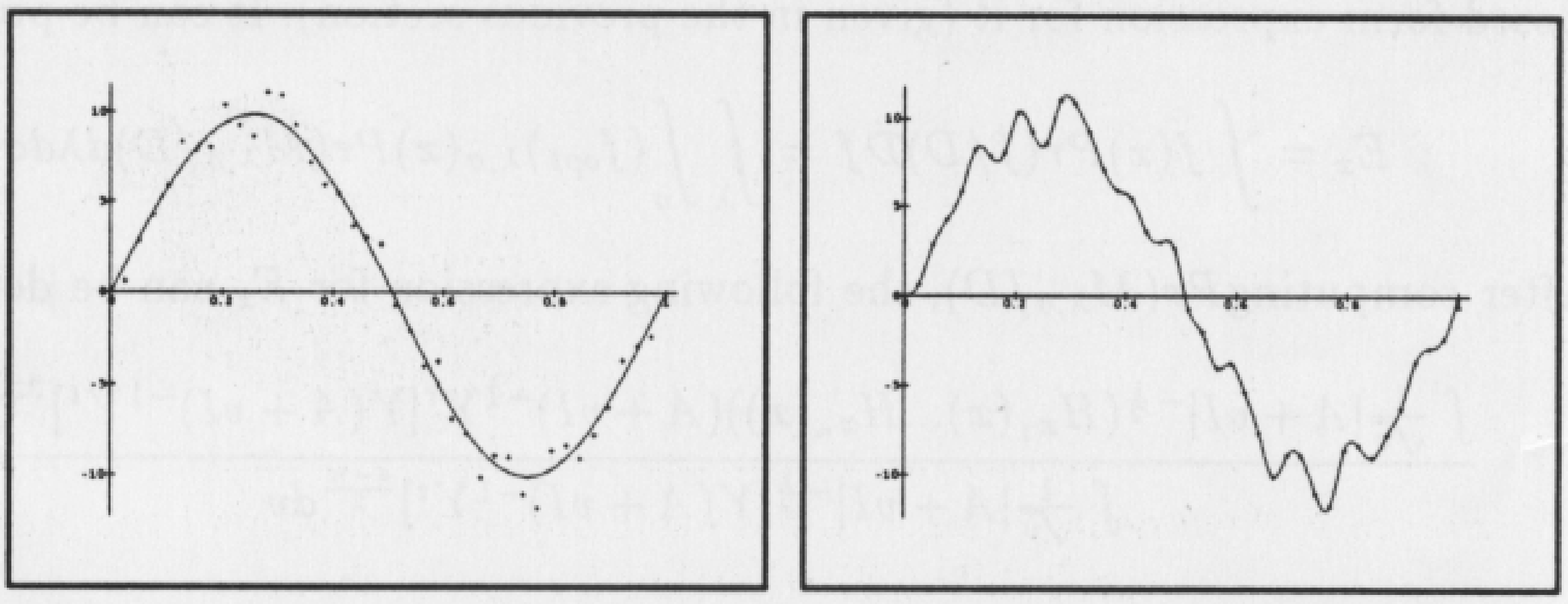

Figure 1: Instability of GCV: for two data sets differing in one point, GCV gives two very different interpolants.

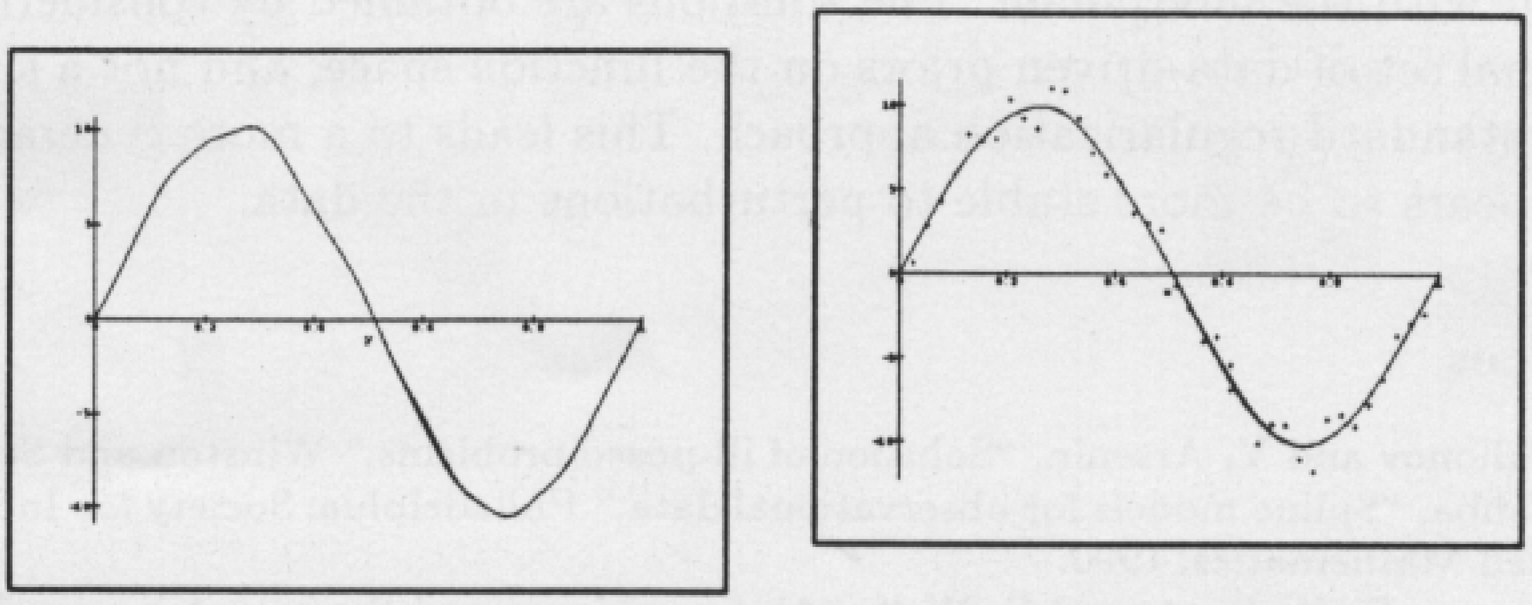

Figure 2: The MSE (left) and MAP (right) estimates for the data sets of Figure 1.

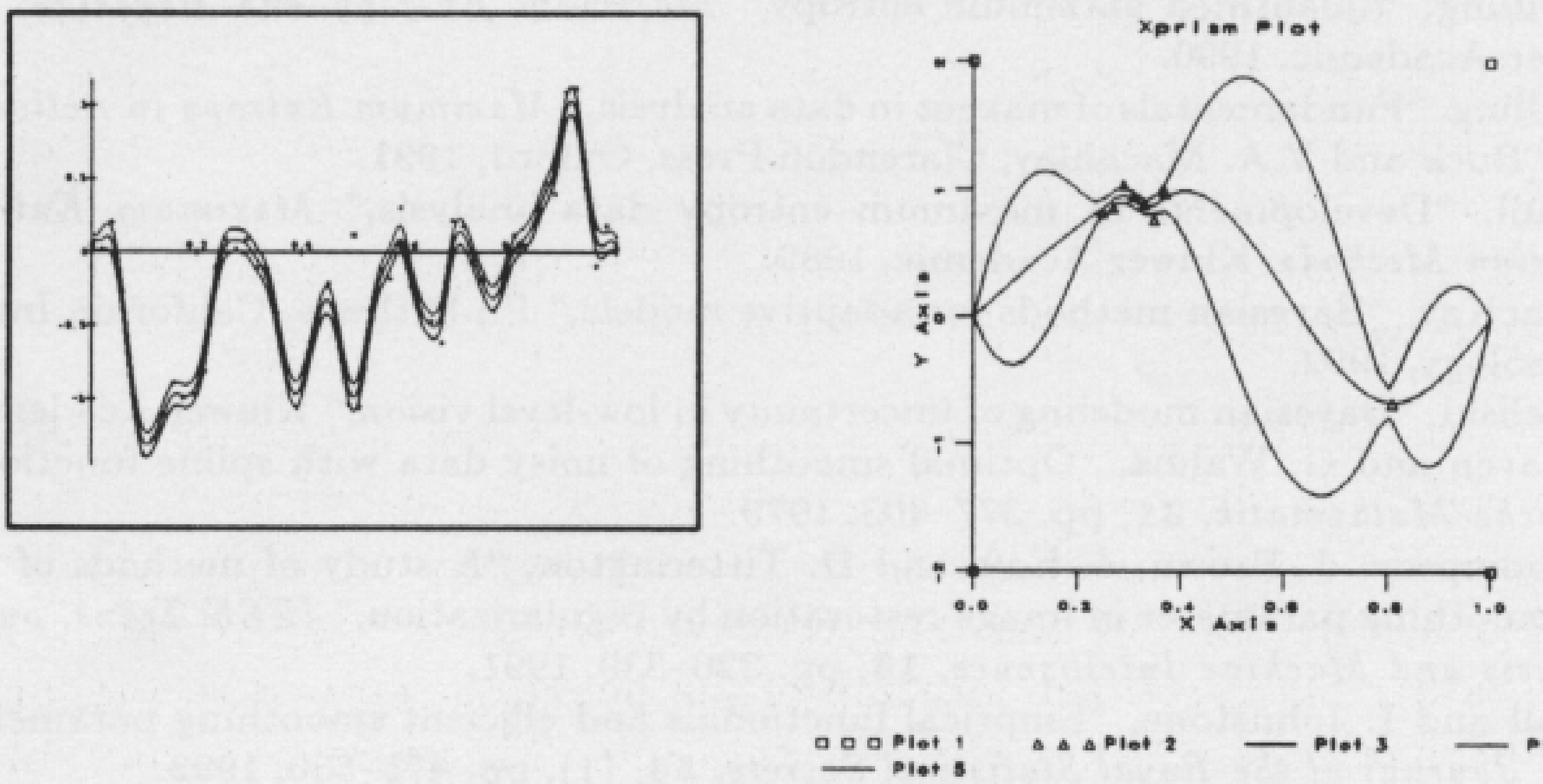

Figure 3: MSE function and confidence intervals for an evenly and unevenly sampled data set.

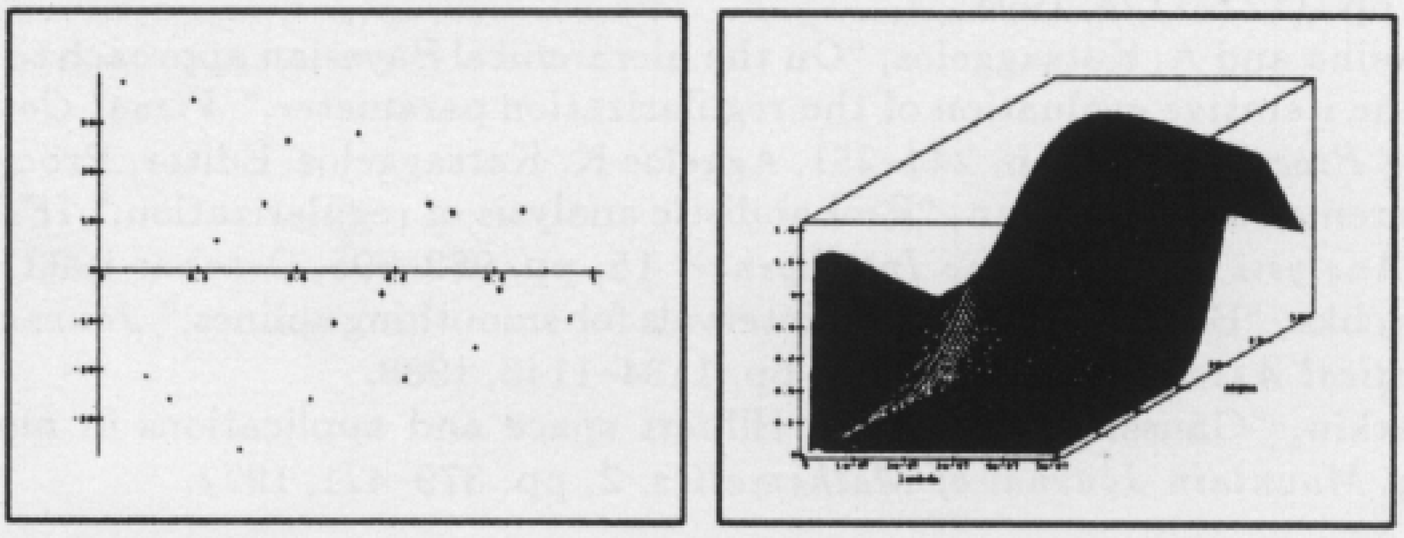

Figure 4: Two nearly identical data sets superimposed, and the (scaled) probability distribution for one of them.

7. Conclusions and Further Research

This work suggests a straightforward approach for solving three basic problems in curve and surface reconstruction, which are very common in many areas: finding the MAP interpolant, finding the MSE interpolant, and computing the uncertainty associated with the interpolant. The solutions are obtained by considering a two-dimensional set of data-driven priors on the function space, and not a single prior as in the standard regularization approach. This leads to a more general solution, which appears to be more stable to perturbations in the data.

References

1. A. Tikhonov and V. Arsenin, "Solution of ill-posed problems," Winston and Sons, 1977.
2. G. Wahba, "Spline models for observational data," Philadelphia: Society for Industrial and Applied Mathematics, 1990.
3. C. Strauss, D. Wolpert, and E. Wolf, "Alpha evidence and the entropic prior," *Maximum Entropy and Bayesian Methods*, Kluwer Academic, 1995.
4. R. Fischer, W. Von Der Linden, and V. Dose, "On the importance of α marginalization in mamimum entropy," *Maximum Entropy and Bayesian Methods*, Kluwer Academic, To Appear.
5. J. Skilling, "Quantified maximum entropy," *Maximum Entropy and Bayesian Methods*, Kluwer Academic, 1990.
6. J. Skilling, "Fundamentals of maxent in data analysis," *Maximum Entropy in Action*, Edited by B. Buck and V.A. Macaulay, Clarendon Press, Oxford, 1991.
7. S. Gull, "Developments in maximum entropy data analysis," *Maximum Entropy and Bayesian Methods*, Kluwer Academic, 1989.
8. D. MacKay, "Bayesian methods for adaptive models," Ph.D thesis, California Institute of Technology, 1992.
9. R. Szeliski, "Bayesian modeling of uncertainty in low-level vision," Kluwer Academic, 1989.
10. P. Craven and G. Wahba, "Optimal smoothing of noisy data with spline functions," *Numerische Mathematik*, **31**, pp. 377–403, 1979.
11. A. Thompson, J. Brown, J. Kay, and D. Titterington, "A study of methods of choosing the smoothing parameter in image restoration by regularization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **13**, pp. 326–339, 1991.
12. P. Hall and I. Johnstone, "Emprical functionals and efficient smoothing parameter selection," *Journal of the Royal Statistical Society*, **54**, (1), pp. 475–530, 1992.
13. D. Nychka, "Choosing a range for the amount of smoothing in nonparametric regression," *Journal of the American Statistical Association*, **86**, (415), pp. 653–664, 1991.
14. R. Molina, "On the hierarchical Bayesian approach to image restoration: Applications to astronomical images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **16**, (11), pp. 1122–1128, 1994.
15. R. Molina and A. Katsaggelos, "On the hierarchical Bayesian approach to image restoration and the iterative evaluation of the regularization parameter," *Visual Communications and Image Processing'94*, pp. 244–251, Aggelos K. Katsaggelos, Editor, Proc. SPIE 2308, 1994.
16. D. Keren and M. Werman, "Probabilistic analysis of regularization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **15**, pp. 982–995, October 1993.
17. D. Nychka, "Bayesian confidence intervals for smoothing splines," *Journal of the American Statistical Association*, **83**, (404), pp. 1134–1143, 1988.
18. F. Larkin, "Gaussian measure in Hilbert space and applications in numerical analysis," *Rocky Mountain Journal of Mathematics*, **2**, pp. 379–421, 1972.