Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal. (2011), Vol. 4, Issue 1, 106 – 110
e-ISSN 2070-5948, DOI 10.1285/i20705948v4n1p106

EXCESSIVE EFFICIENCY IN STATISTICAL INTERPOLATION DUE TO TOO MANY CONSTRAINTS

Adam C. Hillier and Steven C. Gustafson*

Air Force Institute of Technology, AFIT/ENG, Wright-Patterson AFB, Ohio 45433, United States of America

Received 15 June 2010; Accepted 15 October 2010
Available online 26 April 2011

***Abstract**: A mean function and a variance function are specified which interpolate given points in that the mean function intersects the points and the variance function is zero at the points. The specification places simple and reasonable constraints on the form of these functions and on the form of their extrapolation. Using an elementary three-point problem, it is shown that the resulting variance function is unreasonably small over the point domain. This excessive efficiency illustrates that too many constraints can yield unwarranted and overly-optimistic results in statistical interpolation.*

***Keywords**: Interpolation, statistical efficiency*

1. Introduction

Statistical interpolation, as considered here, specifies a mean function that intersects given points and a variance function that is zero at the points. Such interpolation has an extensive literature; see, for example, [1]. The work reported here uses an elementary three-point problem to demonstrate a pitfall. In particular, the specification places simple and reasonable constraints on the form of the mean and variance functions and on the form of their extrapolation. A consequence is that the resulting variance function is unreasonably small over the point domain. This excessive efficiency illustrates that too many constraints can yield unwarranted and overly-optimistic results.

As motivation, consider the following practical problem. A deep sea oil rig wellhead is leaking oil. To stop the flow of oil, it is crucial to know how the temperature at the well head has changed with time. Three temperature measurements were made ten minutes apart, and but additional measurements cannot be made for safety reasons. Before a promising technique to stop the flow can be tried, temperature values must be estimated every minute for the 20 minutes

* E-mail: gustafson.steven@gmail.com

between the first and last measurement. How can the 20 temperature values be estimated, and what is the accuracy of the estimates?

2. Specification

A mean function and a variance function are specified for the motivational problem as follows:

1. Normalize the temperature versus time values by translating, scaling, and rotating so that the resulting points have zero mean and unit variance horizontally and vertically and so that the x axis is their least squares line. Thus the points are (x_i, y_i), $i = 1, 2, \dots, n$, $x_1 < x_2 < \dots < x_n$, where $\sum x_i = 0$, $\sum y_i = 0$, $\sum x_i^2 = n$, $\sum y_i^2 = n$, $\sum x_i y_i = 0$, and here $n = 3$.

2. Let the mean function intersect the points, where this function is a weighted sum of Gaussian kernels, each centered on a point and of the same width such that roughness is minimized. Thus the mean function is

$$\mu(x) = \sum A_i \exp[-(x - x_i)^2/(2s^2)], \qquad (1)$$

where the A_i and s are such that $\int_{-\infty}^{\infty} \mu(x)''^2 dx$ is minimized subject to $\mu(x_i) = y_i$.

3. Let the deviation function be zero at and only at the points, where this function is the positive square root of the quadratic variance function[1] for the least squares line plus a weighted sum of Gaussian kernels, each centered on a point and of the same width such that the function squared has a maximum in the point domain. Thus the deviation function is:

$$\sigma(x) = (1 + 1/n + x^2/n)^{1/2} + \sum B_i \exp[-(x - x_i)^2/(2t^2)], \qquad (2)$$

where the B_i and t are such that $\sigma^2(x)$ is maximized for $x_1 < x < x_n$ subject to $\sigma(x) = 0$ if and only if $x = x_i$.

4. Let the mean of y at any x be $\mu(x)$, let its variance be $\sigma^2(x)$, and invert the normalization to obtain an estimated temperature and its variance for any time.

[1] This well-known variance function (see, for example, [2]) is obtained for normalized points drawn from a Gaussian probability density with mean a + bx and unit variance, where the prior density of parameters a and b is Gaussian with large variance. Thus the density of y given parameters a and b is proportional to $\exp[-(y - a - bx)^2/2]$, and the density of the points given the parameters is proportional to $\exp[-\sum(y_i - a - \sum bx_i)^2/2]$. By Bayes' rule, the density of y given the points is proportional to $\int_{-\infty}^{\infty} \exp[-(y - a - bx)^2/2] \exp[-\sum(y_i - a - bx_i)^2/2] da\, db$. Analytical evaluation of this integral yields a Gaussian density with a zero mean function and a variance function $1 + 1/n + x^2/n$.

3. Results

Since the measurement times are equally spaced, the three normalized points are (-p, -q), (0, 2q), (p, -q), where $p = (3/2)^{1/2}$ and $q = (1/2)^{1/2}$. The mean function for these points has been reported previously [3]. Figure 1 shows the variance function, which is the square of Equation (2). Figure 2 shows the maximum variance achieved in the point domain versus the parameter t in Equation (2), where the t values that satisfy the iff constraint are indicated. Note the following:

1. The variance function is zero at the points, extrapolates to the quadratic for the least squares line, and between the points has the largest variance (i.e., the smallest efficiency) that its form allows.

2. The maximum variance over the point domain is nearer the end points than the center point.

3. Values of t that satisfy the iff constraint are irregularly located.

4. The maximum variance is 0.0075, so the maximum standard deviation is 0.087. Since the data points have unit variance, the precision (defined as the ratio of the estimated standard deviation to the standard deviation of the data) is better than 8.7 % in the point domain. Also, for x magnitudes less than 0.1 the precision is much better than 1%.

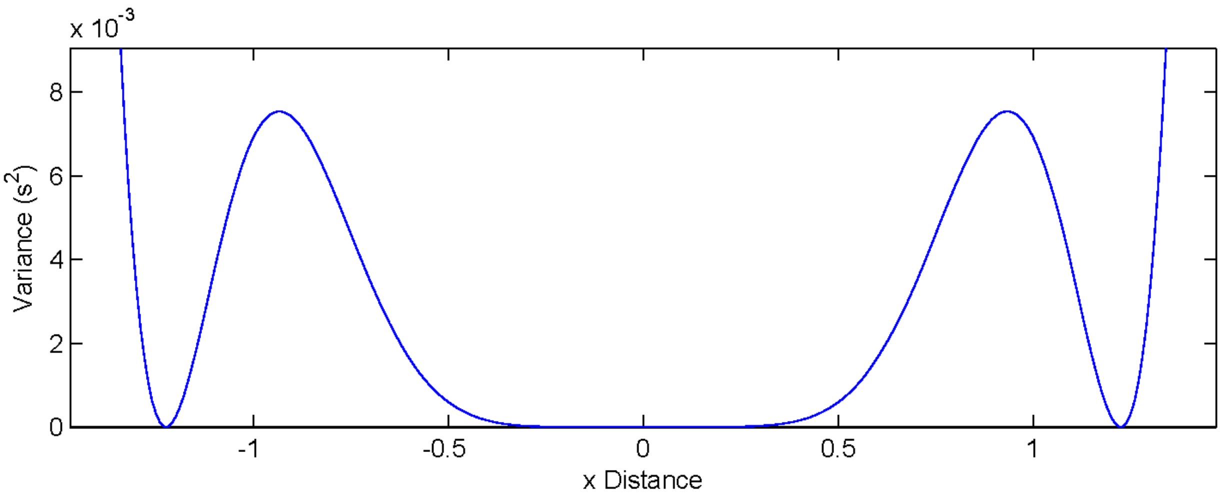

Figure 1. The variance function $\sigma^2(x)$ for three normalized points, where $\sigma(x)$ is given by Equation (2). This function is zero at the points, extrapolates to the quadratic for the least squares line, and between the points has the largest variance (i.e., the smallest efficiency) that its form allows. The maximum variance over the point domain is nearer the end points than the center point, as is reasonable. The maximum variance is 0.0075, so the maximum standard deviation is 0.087. Since the data points have unit variance, the precision (defined as the ratio of the estimated standard deviation to the standard deviation of the data) is better than 8.7 % in the point domain. Also, for x magnitudes less than 0.1 the precision is much better than 1%. As discussed in Section 4, these percentages are unreasonably small.

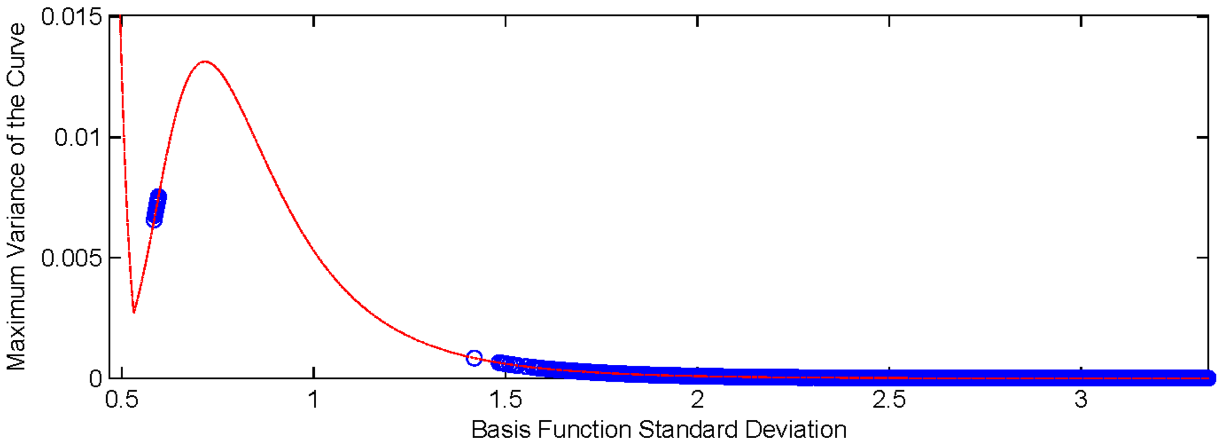

Figure 2. The maximum variance achieved in the point domain versus the parameter t in Equation (2), where the t values that satisfy the if-and-only-if constraint are indicated by circles. As discussed in Section 3, these values are irregularly located.

4. Discussion

Result (1) is required by the specification. Result (2) is reasonable since the variance should be larger at x values not surrounded by points. Result (3) may seem strange because only simple quadratic and sum-of-Gaussian forms are employed in the solution. However, the technical literature contains many examples of chaotic behavior arising from simple forms; see, for example, [4].

Result (4) indicates precision percentages that are unreasonably small, i.e., it is unreasonable to suppose, based data which consists of only three measurement points, that additional data can be estimated between the points with a precision of better (and in some regions much better) than 8.7 %. Note that this unwarranted and overly-optimistic precision is obtained even though the variance function is maximized over the point domain, and thus the percentages are as large or non-optimistic as the variance function specification permits.

An explanation for the unreasonable precision is that the variance function, although it uses only simple quadratic and sum-of-Gaussian forms, has two constraints which force the variance function to be small. The first constraint is that the variance at the points is zero so that the measurements at the points are without error. This constraint is reasonable if the measurements are much more precise than the estimates (e.g., analog measurements with precisions of 0.01% or better are common). The second constraint is that the variance function extrapolates to the quadratic function associated with the least squares line. This constraint is reasonable if the measurements are known to extrapolate to a linear trend.

A conclusion is that an excessive number of constraints on the form of the mean and variance functions and on the form of their extrapolation can lead to a variance function which is unreasonably small over the point domain. This conclusion applies even if each constraint is simple and reasonable. The resulting excessive efficiency illustrates that too many constraints can yield estimates with unwarranted and overly-optimistic precision.

Acknowledgement

Helpful discussions with Dr. R. K. Martin are gratefully acknowledged. The following statement is required by the U. S. Government: The views expressed in this paper are those of the authors and do not reflect the official policies or positions of the United States Air Force, the US Department of Defense, or the US Government.

References

[1]. MacKay, D.J.C. (1992). Bayesian Interpolation. *Neural Computation*, 4, 415-447.
[2]. Sivia, D.S. (1996). *Data Analysis: A Bayesian Tutorial*. Oxford: Oxford Univ. Press.
[3]. Guild, E.M., Like, E.C., and Gustafson, S.C. (2009). Fast Cardinal Interpolation. *The Open Cybernetics and Systemics Journal*, 3, 85-89.
[4]. Sprott, J.C. (2003). *Chaos and Time-Series Analysis*. Oxford: Oxford Univ. Press.