Biometrics 60, 954–964
December 2004

Using Image and Curve Registration for Measuring the Goodness of Fit of Spatial and Temporal Predictions

Cavan Reilly,[1,*] Phillip Price,[2] Andrew Gelman,[3] and Scott A. Sandgathe[4]

[1]Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota 55454, U.S.A.
[2]Indoor Environment Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, U.S.A.
[3]Department of Statistics, Columbia University, New York, New York 10027, U.S.A.
[4]Applied Physics Laboratory, University of Washington, 1013 NE 40th Street, Seattle, Washington 98105, U.S.A.
[*]*email:* cavanr@biostat.umn.edu

Summary. Conventional measures of model fit for indexed data (e.g., time series or spatial data) summarize errors in y, for instance by integrating (or summing) the squared difference between predicted and measured values over a range of x. We propose an approach which recognizes that errors can occur in the x-direction as well. Instead of just measuring the difference between the predictions and observations at each site (or time), we first "deform" the predictions, stretching or compressing along the x-direction or directions, so as to improve the agreement between the observations and the deformed predictions. Error is then summarized by (a) the amount of deformation in x, and (b) the remaining difference in y between the data and the deformed predictions (i.e., the residual error in y after the deformation). A parameter, λ, controls the tradeoff between (a) and (b), so that as $\lambda \to \infty$ no deformation is allowed, whereas for $\lambda = 0$ the deformation minimizes the errors in y. In some applications, the deformation itself is of interest because it characterizes the (temporal or spatial) structure of the errors. The optimal deformation can be computed by solving a system of nonlinear partial differential equations, or, for a unidimensional index, by using a dynamic programming algorithm. We illustrate the procedure with examples from nonlinear time series and fluid dynamics.

Key words: Calculus of variations; Deformation; Dynamic programming; Errors-in-variables regression; Goodness of fit; Image registration; Morphometrics; Spatial distribution; Time series; Variance components.

1. Introduction

A fundamental problem in data analysis is quantifying the fit of predictions to data. To address this question in a spatial or temporal context, it is conventional to think of predictions (and observations) as descriptions of y as a function of one or more spatial or time indexes x, and to summarize errors by integrating some discrepancy function over x (e.g., the mean-squared error in y or mean absolute error in y).

In fact, though, errors—both in the sense of discrepancies between predictions and observations, and in the sense of variations of observed values about their true values—can occur in the y or x directions, not just in y. Measures of misfit should recognize that fact. From this perspective, predictions should be locally stretched or compressed ("deformed") in the x-direction (or directions) so as to improve the agreement between deformed predictions and the observations. Overall error can be summarized both in terms of the error in y that remains following the deformation, and the amount of deformation in x that was performed. This work differs from errors-in-variables regression (e.g., Madansky, 1959) in that we are modeling continuous deformation, not independent errors, in x.

Our work is conceptually related to "morphing" (derived from "metamorphosing"), a technique used in image registration, in which one image is continuously deformed into another. Morphing has been used in many applications and has a long history in mapping and biology. Our work goes beyond existing uses of morphing in that we use it as a tool to measure model fit. Moreover, most morphing methods (e.g., Goshtasby, 1988; Bookstein, 1991; Lee et al., 1996; Yan et al., 2000) require identification of specific features shared by both images, which may be easy for movie images (eyes, nose, mouth, etc.) or satellite photos (roads, buildings, etc.) but is not so simple for comparing model predictions to observations. For example, predictions and observations may have different numbers of local maxima and local minima. A recent overview of the field of image registration from a statistical perspective, including a novel implementation, is provided by Glasbey and Mardia (2001). While we develop a new variational approach to image registration here, the primary objective is to use these techniques to measure model misfit. This new method is necessary because we must perform the image registration many times (existing methods rely on high-dimensional nonlinear optimization and can be quite slow).

There have been several attempts to quantify error in location in the applied literature. For example, Martinson, Menke, and Stoffa (1982) discuss a one-dimensional problem similar to ours, in a signal processing context. In addition, there have been some attempts in the meteorological literature to define distortion measures for model misfit (e.g., Hoffman et al.,

1995; Marzban, 1998; Ebert and McBride, 2000). These methods are more restrictive than the methods developed here and typically require substantial user input. The one-dimensional problem is also treated by Ramsay and Silverman (1997, 2002), but their focus is on registration, not using registration to understand model fit.

2. Summarizing Errors in Terms of Scaling and Deformation

2.1 *Mathematical Formulation*

If observations and predictions are defined for all x in some set A, then the scaling of y and a deformation function $f(x)$ can be combined into a total error measure, as in Ramsay and Silverman (1997),

$$\mathcal{I}_\lambda(y, \hat{y}) \equiv \min_{f \in \mathcal{D}} \left\{ \int_A G(y(x), \hat{y}(f(x)))\, dx + \lambda \int_A F(x, f(x))\, dx \right\}. \quad (1)$$

Here, G is a discrepancy metric between y and $\hat{y}$, and F measures the amount of deformation, defined by $f(x)$, relative to the identity map $f(x) = x$. Because F measures the discrepancy between the deformation and the identity map, we require $F(x, x) = 0$ for all x in the region over which there are measurements. By deforming the predictions (rather than the observations), we favor sets of predictions that have more extreme values.

The parameter λ controls the tradeoff between prediction error in y and deformation in x. As $\lambda \to \infty$, even minor deformations become strongly penalized, thus $f(x) - x \to 0$, hence $\mathcal{I}_\lambda$ approaches the integrated squared error. On the other hand, if $\lambda = 0$, then any allowed deformation can be attained. In between, λ might be set based on external knowledge or preference; for example, a researcher evaluating hurricane prediction models might stipulate that an error of 20 km/hour in peak wind speed is equivalent to an error of 50 km in location. In other cases, it can be helpful to perform calculations over a range of λ so as to investigate the tradeoff between scaling and deformation, as we illustrate in examples below.

We characterize the set of allowable deformations as those that occur in elastic deformation. For predictions and observations that are defined over the same region, the two configurations must have the same boundaries, so we constrain our deformations to be one-to-one functions. If f is not onto, the deformation could discard some predicted values, and this could lead to misleading results because arbitrarily poor predictions could match the observations completely after applying such a deformation. A physical requirement of elasticity is that the deformation continuously moves each particle of the initial configuration to one and only one location in the final configuration. The analytical expression of this requirement is that the determinant of the Jacobian matrix of the deformation be strictly positive (see, e.g., Boresi and Chong, 2000). With a unidimensional index this condition is simply that the deformation be strictly increasing, as in Ramsay and Silverman (2002).

2.2 *Bayesian Interpretation*

As with many nonparametric methods, the deformed prediction surface can be interpreted as a Bayes posterior estimate under certain assumptions about the data-generation and measurement processes (e.g., Wahba, 1978). Exploring this connection can help us better understand when the procedure should do well and in what situations it would be less appropriate. For the methods in this article, the deformations correspond to a model on x and the errors correspond to a model on $y \mid x$. In addition, in a Bayesian framework, tuning parameters such as λ in (1) can be formulated as hyperparameters and estimated from hierarchically structured datasets. (We do not work with such data structures in this article, but we find it helpful to at least consider how to set up the problem.)

In conventional practice, predictions are related to sampled values in terms of a measurement error model, $y(x) = \hat{y}(x) + \epsilon(x)$, where ϵ is some error process, and the accuracy of the predictions is assessed via some functional of ϵ. Here we, instead, have the model $y(x) = \hat{y}(f(x)) + \epsilon(x)$, where f is the deformation. Assume that we have independent, constant-variance measurement/prediction error after we take account of the deformation of the index variable, and denote the variance of this error by σ^2. In this case, the f that minimizes $\mathcal{I}_\lambda$ in (1) is the posterior mode corresponding to some prior distribution on the space of deformations.

The expression $\mathcal{I}_\lambda$ in (1) can be interpreted as the negative of the log-posterior density of the functional parameter f given data y and predictions $\hat{y}$. The error metric G then represents the log likelihood of the data corresponding to some specified probability distribution on the errors ϵ (e.g., the sum of squared errors corresponds to independent Gaussian errors with equal variances).

In the Bayesian formulation, the penalty function F represents the logarithm of a prior distribution on the space of functions f. Different specifications of F in terms of squared differences between $f(x)$ and x, or their differentials, correspond to various Gaussian process priors for f (quadratic log-density functions), with the added restriction that the determinant of the Jacobian matrix of f be strictly positive.

2.3 *Choosing the Functions F and G*

The error measure, G, may be any function that summarizes the error in y, such as the squared error, relative error, or any other measure of fit. Because specification of such functions has been well explored in the statistical literature, for convenience we simply use squared error here:

$$G(y(x), \hat{y}(f(x))) \equiv [y(x) - \hat{y}(f(x))]^2. \quad (2)$$

In some applications, other choices may be more sensible. Recently, Ramsay and Silverman (2002) proposed

$$G_A(y(x), \hat{y}(f(x))) \equiv [Ay(x) - \hat{y}(f(x))]^2,$$

for some scalar A. This error measure will be useful when there is multiplicative error, but there are two problems with this formulation: if $\hat{y} \approx 0$ then $A \approx 0$ (thus the estimate of f is largely meaningless), and there are identifiability problems when we try to estimate the pair (A, f), as we discuss in Section 4.1. If multiplicative error is suspected, one should

first log the data prior to summarizing the error with a sum of square type measure.

2.3.1 *Choosing F for unidimensional index.* The function F is a measure of the amount of deformation represented by the deformation f. To the extent that the function f differs from the identity map, it represents a stretching and compression of the x dimension. Natural choices for the function that quantifies the deformation include $F(x, f(x)) \equiv [f(x) - x]^2$, $F(x, f(x)) \equiv [\log f'(x)]^2$, and $F(x, f(x)) \equiv [f'(x) - 1]^2$. Just as there are many possible choices for measuring error in y, there are many possible choices for quantifying the amount of deformation in x. For this article, we use

$$F(x, f(x)) \equiv \left[\frac{df(x)}{dx} - 1\right]^2. \tag{3}$$

Somewhat surprisingly, the choice of the deformation penalty function has implications for the smoothness of the resulting deformation. As a simple example, suppose $A = [0, 1]$, $y(x) = 0$, and $\hat{y}(x) = x$. For this example, we can find an explicit expression for the deformation using several different deformation penalty functions. If $F(x, f(x)) = [f'(x) - 1]^2$ then the optimal deformation is $\hat{f}(x) = \sinh(\lambda^{-1/2}x)/\sinh(\lambda^{-1/2})$, whereas if $F(x, f(x)) = [f(x) - x]^2$, then the optimal deformation is $\hat{f}(0) = 0, \hat{f}(1) = 1$, and $\hat{f}(x) = (\lambda x)/(1 + \lambda)$ for $x \in (0, 1)$. Below, we present methods which allow us to find the first of these deformations; the second is found by completing the square under the integral in the definition of $\mathcal{I}_\lambda$. The second of these deformations is not continuous for any value of λ, while the first is continuous for all $\lambda > 0$. Because we seek continuous deformations, we actually want the continuous function that is as close as possible to the discontinuous deformation for the second penalty function. Below, we demonstrate how we can exploit the smoothness of a deformation to obtain an algorithm, which finds the optimal deformation rapidly, and for this reason we find the derivative penalties useful.

2.3.2 *Choosing F for multidimensional index.* The multidimensional case presents substantial computational challenges. We focus here on the two-dimensional case; generalizing to higher dimensions is then straightforward. In two dimensions, the deformation is a vector-valued function $f(x_1, x_2) = (f_1(x_1, x_2), f_2(x_1, x_2))$. For this article, as a direct generalization of the unidimensional approach, to measure the extent of deformation we simply compare each partial derivative of f to 1 using squared error loss. For a thorough discussion regarding the choice of deformation penalty and the many possibilities, the reader is referred to Glasbey and Mardia (2001). These other deformation measures could also be used with the approach proposed here because these measures are functionals of the derivatives of the deformation.

3. Computation

To find the function f that minimizes the integral in the definition of our statistic, we use the calculus of variations. This technique allows us to find f by solving a certain boundary value problem (the Euler–Lagrange equations). While it can be difficult to solve these problems numerically in more than one dimension, we have solved them successfully for several applied problems, some of which are presented below. The advantage of the use of the calculus of variations is that we can find deformations very rapidly (in less than a minute for the two-dimensional example presented below). In contrast, if one finds f by nonlinear optimization, as in Glasbey and Mardia (2001), then according to the authors it takes roughly 30 minutes for some problems. Moreover, the solution one obtains using nonlinear optimization algorithms could just be a local optimum.

Throughout this article, we interpolate between values at observed sites when necessary in order to reconstruct a continuous set of observations and predictions from a set of observations and predictions at a finite collection of sites. The need for interpolation is clearest in the unidimensional context because sites must map into and onto the sites with a map that is strictly monotone, and the only strictly monotone bijective map from one finite ordered set into another with the same cardinality is the identity. We parameterize f by its values at the sites where we have observations and linearly interpolate to obtain the value of f at locations between the observed sites. In the univariate case, values of f are needed only at the x values of the observations and at the x values onto which the predictions are mapped. One could use a different interpolation approach to find a smooth, continuous deformation f, but this is not necessary in the univariate case, and the details of the result would depend on the interpolation method. In the two-dimensional implementation we use numerical methods that define f at locations other than the observed sites.

3.1 *Using Dynamic Programming to Solve the One-Dimensional Problem*

In the unidimensional case, we can use dynamic programming (see Bellman, 1957) to obtain an approximation to the solution to the Euler–Lagrange equation. The resulting algorithm is extremely stable numerically and will always find the global optimum. In contrast, existing methods rely on high-dimensional optimization and are subject to the usual problems we associate with such methods, such as convergence to local optima and long computing times. Dynamic programming is a method for solving sequential decision problems where the goal is to optimize some function of the entire sequence of decisions. The unidimensional problem can be construed as a sequential decision problem in which we find $f(x)$ as we allow x to traverse the interval A. This formulation also indicates why there is no direct generalization of this procedure to higher-dimensional problems. Our dynamic programming implementation only allows determination of an approximate solution because the values of the range of f must be a discrete set, but because this discrete set can approach the set of real numbers the solution can reach any desired level of accuracy.

The basic idea of the application of dynamic programming to our problem is to relate the minimized value of the approximation to $\mathcal{I}_\lambda$ (which restricts the range of f) when there are n observations, $\mathcal{F}_n$, to the value of this discrete approximation when there are $n - 1$ observations, $\mathcal{F}_{n-1}$.

Suppose the n observations are equally spaced in x, and label the sites 1 to n. Let $\mathcal{C}_m$ be the set of piecewise linear, strictly increasing functions with nodes at $\{1, \ldots, n\}$ such that the values at the nodes are of the form $(m + j)/m$ for

$j = 0, 1, \ldots, m(n-1)$ with $m > 1$. Let $\mathcal{C}_{i,j} = \{f \in \mathcal{C}_m : f(i) = (m+j)/m\}$, let

$$\mathcal{F}_i(j) = \min_{f \in \mathcal{C}_{i,j}} \int_1^i [y(x) - \hat{y}(f(x))]^2 + \lambda[f'(x) - 1]^2 \, dx.$$

As m increases, the gap between possible values of $f(i)$ decreases, and this gap can be made arbitrarily small by choosing a sufficiently large value of m. For f to be bijective we require $f(1) = 1$ and $f(n) = n$.

Let $\hat{f}$ be the optimal deformation for $f \in \mathcal{C}_m$. If $\hat{f}(i) = (m+j)/m$, then the requirement that the deformation does not tear the predictions implies that $\hat{f}(l) \in (1, (m+j)/m)$ for $l = 2, \ldots, i-1$. The dynamic programming principle demands that for any optimal deformation $\hat{f}$, if $\hat{f}(i-1) = (m+k)/m$, then $\hat{f}(l)$ for $l > i-1$ must be the optimal deformation for the initial condition, which specifies $\hat{f}(i-1) = (m+k)/m$. Hence

$$\mathcal{F}_i(j) = \min_{k \in [1,j]} \Bigg[\mathcal{F}_{i-1}(k) + \int_{i-1}^i ([y(x) - \hat{y}(f(x))]^2 + \lambda[f'(x) - 1]^2) \, dx \Bigg], \tag{4}$$

for $i = 2, \ldots, n-1$ and $j = \lceil 1 + (i-1)/m \rceil, \ldots, \lfloor n - (n-i)/m \rfloor$. The f in this expression is just a function of i, j, k because the expression entails $f(i-1) = (m+k)/m$ and $f(i) = (m+j)/m$, and other values of f are obtained via linear interpolation. Explicitly incorporating linear interpolation in the integral in equation (4) leads to

$$\begin{aligned} &\int_{i-1}^i ([y(x) - \hat{y}(f(x))]^2 + \lambda[f'(x) - 1]^2) \, dx \\ &= \frac{1}{3} \sum_{l=\lfloor \frac{m+k}{m} \rfloor}^{\lceil \frac{j}{m} \rceil} \frac{1}{\beta_{i-1} - \hat{\beta}_l \frac{j-k}{m}} \\ &\quad \times \left\{ \left[\alpha_{i-1} + \left(i - 1 + \frac{lm-k}{j-k} \right) \beta_{i-1} - \hat{\alpha}_l - (l+1)\hat{\beta}_l \right]^3 \right. \\ &\quad \left. - \left[\alpha_{i-1} + \left(i - 1 + \frac{lm-mk}{j-k} \right) \beta_{i-1} - \hat{\alpha}_l - l\hat{\beta}_l \right]^3 \right\} \\ &\quad + \lambda \left(\frac{j-k}{m} - 1 \right)^2, \end{aligned} \tag{5}$$

where $\alpha_k = y_k - k(y_{k+1} - y_k)$, $\beta_k = y_{k+1} - y_k$, and $\hat{\alpha}_k$ and $\hat{\beta}_k$ are defined analogously, but with hats on the y's. Let $\mathcal{G}(i,j,k)$ be the term in solid brackets in equation (4), so that $\mathcal{F}_i(j) \equiv \min_{k \in [1,j]} \mathcal{G}(i,j,k)$. Once we compute $\mathcal{F}_i(j)$ by recording $\arg\min_{k \in [1,j]} \mathcal{G}(i,j,k)$ for all i, j, it is simple to find the optimal deformation because $\hat{f}(n) = n$ and $\hat{f}(l) = (m + \arg\min_{k \in [1, f(l+1)m - m]} \mathcal{G}(l, \hat{f}(l+1)m - m, k))/m$ for $l < n$. Software that implements the procedure is available at `www.biostat.umn.edu/~cavanr`.

3.2 *Calculus of Variations for the One-Dimensional Problem*

Another way to find the optimal deformation is to solve the Euler–Lagrange equations using general techniques for the solution of boundary value problems. The differential equation using the deformation penalty from (3) is

$$\lambda \frac{d^2 f}{dx^2} = \hat{y}'(f(x))[\hat{y}(f(x)) - y(x)],$$

and if the observations and predictions are on the unit interval, the boundary conditions are $f(0) = 0$ and $f(1) = 1$. Unfortunately, this differential equation is a singular perturbation problem (for $\lambda \to 0$), and the singularly perturbed problem does not necessarily have a unique solution; see, e.g., De Jager and Furu (1996) or O'Malley (1991). For example, if $A = [0, 1]$, $y(x) = x$, and $\hat{y}(x) = 2 + x$, then there is no solution to the equation with $\lambda = 0$. In practice this means that we cannot always find a numerical approximation to the solution of the two-point boundary value problem for small λ, due to numerical instabilities. This same example illustrates that methods that rely on nonlinear optimization can encounter problems because the optimal solution can be on the boundary of the parameter space.

Another problem with the calculus of variations approach is that we cannot guarantee that the resulting deformation is strictly monotone. One is tempted to use a penalty function to rule out nonmonotone solutions (i.e., add a term to the definition of $\mathcal{I}_\lambda$, which penalizes nonmonotone solutions to the extent that the numerical techniques will not produce nonmonotone solutions), but such a technique will not work: it is easy to show that if there is no monotone solution to our two-point boundary value problem without a nonmonotonicity penalty function, then adding such a penalty function to rule out nonmonotonic solutions generates a boundary value problem with no solutions at all. (The proof rests on the uniqueness of solutions to ordinary differential equations.) This same problem is encountered by other methods; for example, Glasbey and Mardia (2001) give a method for checking whether the warping function is bijective but do not use these constraints when they find the warp function. We take the same approach here. In practice, solutions are monotone provided λ exceeds some problem-specific value.

We think the dynamic programming approach is the method of choice for unidimensionally indexed data (because it is numerically stable and finds the global optimum), so we use this method exclusively for such data. However, the calculus of variations approach generalizes more directly to higher-dimensioned indexes, as we discuss here.

3.3 *Calculus of Variations in Multiple Dimensions*

We measure the deviation of a deformation from the identity map with a two-dimensional (and, by implication, higher-dimensional) extension of (3), by comparing the diagonal of the Jacobian of the deformation to the identity matrix. The measure of total misfit is

$$\mathcal{I}_\lambda(y, \hat{y}) = \min_{f \in \mathcal{D}} \left\{ \iint_A \left([y(x_1, x_2) - \hat{y}(f_1(x_1, x_2), f_2(x_1, x_2))]^2 + \lambda \left[\left(\frac{\partial f_1}{\partial x_1} - 1 \right)^2 + \left(\frac{\partial f_2}{\partial x_2} - 1 \right)^2 \right] \right) dx_1 \, dx_2 \right\}.$$

The calculus of variations gives the following nonlinear elliptic system of partial differential equations (PDEs) for the

deformation f that minimizes the previous integral,

$$\lambda \frac{\partial^2 f_i}{\partial x_i^2} = \frac{\partial \hat{y}}{\partial f_i}(f_1(x_1, x_2), f_2(x_1, x_2)) \times [\hat{y}(f_1(x_1, x_2), f_2(x_1, x_2)) - y(x_1, x_2)],$$

for $i = 1, 2$; see, e.g., Renardy and Rogers (1993). We assume measurements and predictions are on the unit square. The deformation must be one to one, hence the boundaries must map to themselves, so the boundary conditions are $f_1(0, x_2) = f_2(x_1, 0) = 0$, $f_1(1, x_2) = f_2(x_1, 1) = 1$, $f_1(x_1, 0) = f_1(x_1, 1) = x_1$, and $f_2(0, x_2) = f_2(1, x_2) = x_2$ for all (x_1, x_2) in the unit square.

We then solve the system of PDEs (approximately) using numerical methods. We have found the use of Brandt's full approximation storage algorithm helpful in numerically solving these systems; see, e.g., Hackbusch (1985) and Press et al. (1992). This algorithm uses a multigrid, finite difference approach to numerically solve the system of equations. Because the system is only weakly coupled (due to the choice of penalty function), finding the solution is relatively easy. With other choices for the deformation penalty function, one obtains different sets of equations. Due to the singular perturbation at $\lambda = 0$, one should find optimal deformations with a decreasing sequence of λ's because the algorithm can become unstable when λ gets small. Software for the solution of these systems is available at `www.biostat.umn.edu/~cavanr`.

4. One-Dimensional (Time-Series) Examples

4.1 *An Example with an Explicit Solution*

Before getting into real-data applications of the method, we explore some of the mathematical challenges that arise with deformation, even in one dimension, in the context of a simple example with an explicit solution. Suppose $y(x) = 0$ and $\hat{y}(x) = x$ for $x \in [0, \frac{1}{2}]$ and $\hat{y}(x) = 1 - x$ for $x \in [\frac{1}{2}, 1]$. The associated two-point boundary value problem is then

$$\lambda \frac{d^2 f}{dx^2} = \begin{cases} f & \text{for } 0 \le f(x) < \frac{1}{2}, \\ f - 1 & \text{for } \frac{1}{2} \le f(x) \le 1, \end{cases}$$

with boundary conditions $f(0) = 0$ and $f(1) = 1$. There are infinitely many (weak) solutions when $\lambda = 0$ (namely $f_a(x) = 1_{\{x>a\}}$ for all $a \in (0, 1)$) but for any $\lambda > 0$ the optimal deformation is

$$f_\lambda(x) = \begin{cases} \dfrac{\sinh(\lambda^{-1/2} x)}{2\sinh(\lambda^{-1/2}/2)} & \text{if } 0 \le x \le \frac{1}{2}, \\ \dfrac{\sinh\left(\lambda^{-1/2}(x-1)\right)}{2\sinh(\lambda^{-1/2}/2)} + 1 & \text{if } \frac{1}{2} \le x \le 1. \end{cases}$$

This solution is monotone for all λ. Figure 1 shows graphs of the optimal deformation and the deformed version of the predictions for various values of λ. As we allow λ to approach zero, the solution approaches the discontinuous function $f(x) = 0$ for $x \le \frac{1}{2}$ and $f(x) = 1$ for $x > \frac{1}{2}$.

This example illustrates that the optimal deformation will try to wipe out peaks in $\hat{y}$ that are not in y by squeezing them into oblivion. In the limit, the discontinuous function we obtain completely erases the peak. This example illustrates a fundamental problem with methods for registration: deformation of a curve can appear to reduce prediction error even when we think the curves do not differ in terms of phase. Ramsay and Li (1998) note a similar phenomenon and see this as a weakness, but we do not. For the above example, the predictions are very close to the observed values for values of x near the boundaries. We can think of some of these as misplaced predictions; hence, we should move these predictions toward 0.5, and this is what the optimal deformation does in this example. In the limit, there is no error in y after

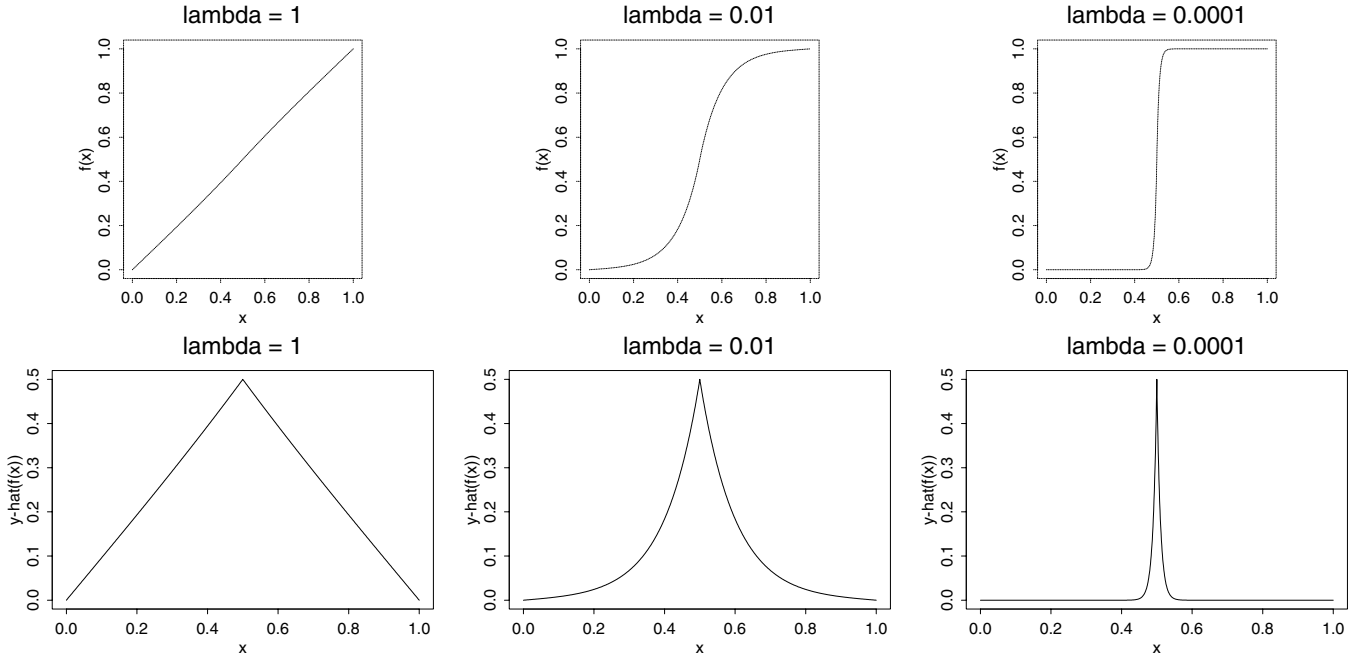

Figure 1. The optimal deformation $f(x)$ for each of the three values of λ for an example in which the solutions can be calculated analytically (top row). Deformed predictions for each of the three models (bottom row).

accounting for deformation. Note that if the observations had been all -1 we would find the same optimal deformation as above, but after deformation with the limiting deformation, the predictions would still differ from the observations by 1. This illustrates that the idea of error in x and error in y are indeed distinct notions.

To see the identifiability problems associated with the parameters (A, f) in the approach of Ramsay and Li (1998) noted in Section 2.3, suppose $\hat{y}(x)$ is as above but now $y(x) = \hat{y}(x)/2$. Then there are two solutions that result in no error after applying the transformation corresponding to the pair: namely $(2, x)$ and $(1, f^*(x))$ where $f^*(x) = x/2$ for $x \in [0, 1/2]$ and $f^*(x) = x/2 + 1/2$ for $x \in (1/2, 1]$. Hence, for small λ there will be two nearly equivalent modes.

4.1.1 *Practical implications of the singular perturbation.* Discontinuities (or extremely sharp deformations) will lead to great difficulties with most numerical implementations. For example, we could try to parameterize f as a piecewise linear function on a mesh, but this would lead to numerical underflow/overflow problems as we take λ smaller if we parameterize the solution so that all first differences of the deformation are positive. If we, instead, were to solve the two-point boundary value problem by discretizing and then numerically solving the system of equations, then, as $\lambda \to 0$, we would have to use some sort of adaptive grid that gets much denser at the high end of the interval, or use a very fine mesh in order to obtain a reasonably accurate solution. Our finite difference solution to the set of PDEs associated with the two-dimensional problem has difficulty in solving the discretized set of equations on the coarsest grid when λ becomes too small.

The problem with the calculus of variations approach can be remedied by finding the deformation over a class of better-behaved functions. Dynamic programming avoids the singular perturbation problems because the restriction on the range imposes implicit constraints on the minimum and maximum derivative of the deformation. Restricting the size of the derivative of the deformation when implementing the calculus of variations method can serve the same purpose (although it is not obvious how we would implement this restriction). The explicit example illustrates how a restriction on the maximum derivative puts an upper bound on how much a peak can get squeezed. Similarly, the lower bound on the derivative of the deformation ($1/m$ in the dynamic programming approach) constrains how much stretching is allowed.

4.1.2 *Bayesian interpretation.* We can view the singular perturbation problem under a Bayesian lens, applying the ideas of Section 2.2 to this specific model. As above, assume the deformation is piecewise linear with nodes at the sampled locations, but now suppose $f(i) = f(i-1) + \eta(i)$ where $\eta(i)$ is a sequence of i.i.d. positive random variables with mean 1 and variance σ_f^2. If we replace the integrals in the definition of our statistic by summations, then we have the statistic

$$\mathcal{S}_\lambda(y, \hat{y}) = \min_{f \in \mathcal{D}} \sum_{i=1}^{I} [y(i) - \hat{y}(f(i))]^2 + \lambda \sum_{i=1}^{I} [f(i) - f(i-1) - 1]^2,$$

and this makes sense if we define $\hat{y}(i)$ by interpolation for non-integral $f(i)$. The deformation for a given λ is the same deformation we obtain from minimizing $\sum_i [y(i) - \hat{y}(f(i))]^2/\sigma^2 + \sum_i [f(i) - f(i-1) - 1]^2/\sigma_f^2$, and so $\hat{f}$ has an interpretation as a linear Bayes estimate (or as a posterior mode for a Gaussian model) under our nonlinear deformation and additive error model. Within this context, λ can be interpreted as a variance ratio. Decreasing λ is equivalent to specifying a larger value of σ_f relative to σ, and in the limit $\lambda \to 0$, $\sigma_f \to \infty$, corresponding to a noninformative prior distribution on the distribution of the increments.

4.2 *An Example Using the Canadian Lynx Series*

We illustrate the one-dimensional method with the often-analyzed series of Canadian lynx trapped in the Mackenzie River area from 1821 to 1934 (Elton and Nicholson, 1942). Although these data have been analyzed dozens of times (see Tong, 1990, for a review), our intention is not to compare all methods of prediction for this dataset but rather to use this example to illustrate the ability of our method to compare the fits of different sets of predictions. We compare three simple models, two of which are biologically motivated. For illustration, we fit each model to the first 80 years of data in order to predict the final 34 years (only these final years are shown in the plots).

4.2.1 *Three models fit to the lynx data.* For our first set of predictions (which we refer to as model A), we model the natural logarithm of the number of lynx trapped as a stochastic process and use the best-fitting autoregression (in terms of Akaike's information criterion) in order to predict the number of lynx trapped. The chosen autoregression has four terms, and the predictions are shown in Figure 2A, along with the actual data from the last 34 years of the series.

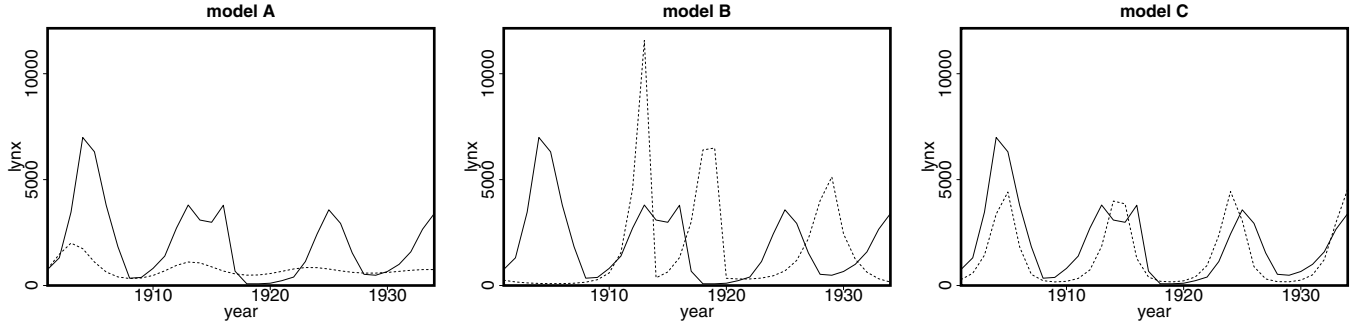

Figure 2. Observed number of trapped Canadian lynx (solid lines) and three sets of predictions (dashed lines). The models predict the time series in different ways and have different patterns of errors.

Our other predictions are based on an explicit biological model, in particular, the simplest predator–prey model of species interaction, with the predators and prey being lynx and snowshoe hare; see Reilly and Zeringue (2004) for more detail. If $u_1(x)$ is the number of lynx at time x, and $u_2(x)$ is the number of snowshoe hare at time x, then the simplest deterministic predator–prey model assumes these quantities are related via the nonlinear system,

$$\frac{du_1}{dx} = -\alpha_1 u_1 + \beta_1 u_1 u_2,$$
$$\frac{du_2}{dx} = \alpha_2 u_2 - \beta_2 u_1 u_2,$$

where α_j, β_j for $j = 1, 2$ are positive parameters.

To obtain predictions, we need to add an equation relating the measured lynx trappings, $y(x)$, for $x = 1, \ldots, n$, to the size of the lynx population. If α_0 is the proportion of the lynx population trapped (assumed constant over time), then we assume

$$y(x) = \alpha_0 u_1(x) + \epsilon(x),$$

where $\epsilon(x)$, $x = 1, \ldots, n$ is an independent Gaussian measurement error process. If we use a forward difference representation of the dynamics, allow for Gaussian noise in the system equations, and set the measurement variance to zero, we can use the (linear) Kalman filter to obtain predictions once prior distributions are specified for all the parameters (this is model B). The prior information (parameterized as independent normal distributions) to which we have access is not of very high quality, but the predictions (using the posterior mode) do match the approximate magnitude and temporal spacing of peaks and valleys in the numbers of lynx trapped (see Figure 2B).

A third set of predictions was obtained by letting $\theta_1(x) = \log(\beta_2 u_1(x))$ and $\theta_2(x) = \log(\beta_1 u_2(x))$ to obtain the system,

$$\log(y(x)) = \alpha + \theta_1(x) + \delta(x),$$
$$\frac{d\theta_1}{dx} = \mathrm{e}^{\theta_2} - \alpha_1,$$
$$\frac{d\theta_2}{dx} = \alpha_2 - \mathrm{e}^{\theta_1},$$

where $\delta(x)$ for $x = 1, \ldots, n$ is a sequence of independent Gaussian measurement errors (and we use the natural logarithm). We then conduct Bayesian inference for the six parameters in this model (two initial conditions, α, α_1, α_2, and the standard deviation of the measurement noise) using the same prior means but with much larger variances than in model B to obtain model C. The predictions produced by this model (using the posterior mode) are shown in Figure 2C.

Figure 3 displays the optimal deformation and the deformed predictions for all three models using $\lambda = 0$. Setting $\lambda = 0$ is of interest because the resulting deformation is independent of the metric F chosen used to quantify the departure of the deformation from the identity.

4.2.2 *Summarizing the prediction errors of the fitted models.* We compare the models by examining their summary measures of misfit; that is, $G(y(x), \hat{y}(f(x)))$ from (2) and $F(x, f(x))$ from (3), for a range of λ. The left plot in Figure 4 displays the root mean integrated squared error (RMISE), $\left(\int G(y(x), \hat{y}(x))\, dx / \int dx\right)^{1/2}$, plotted by the standardized deformation penalty, $\left(\int F(x, f(x))\, dx / \int dx\right)^{1/2}$, for all three models using the dynamic programming algorithm. To construct

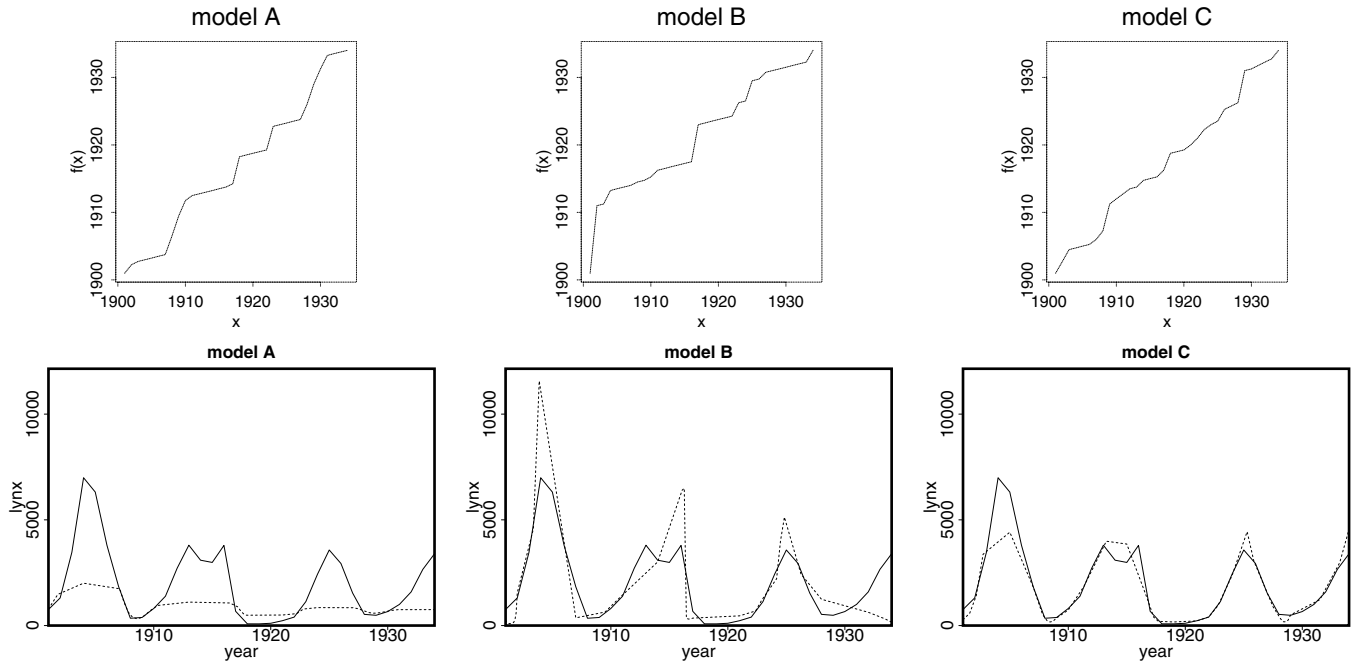

Figure 3. The optimal deformation $f(x)$ for each of the three models in Figure 2 using $\lambda = 0$—that is, prediction error in y is minimized with no concern for deformation error (top row). The data (solid lines) and deformed predictions (dashed lines) for each of the three models (bottom row).

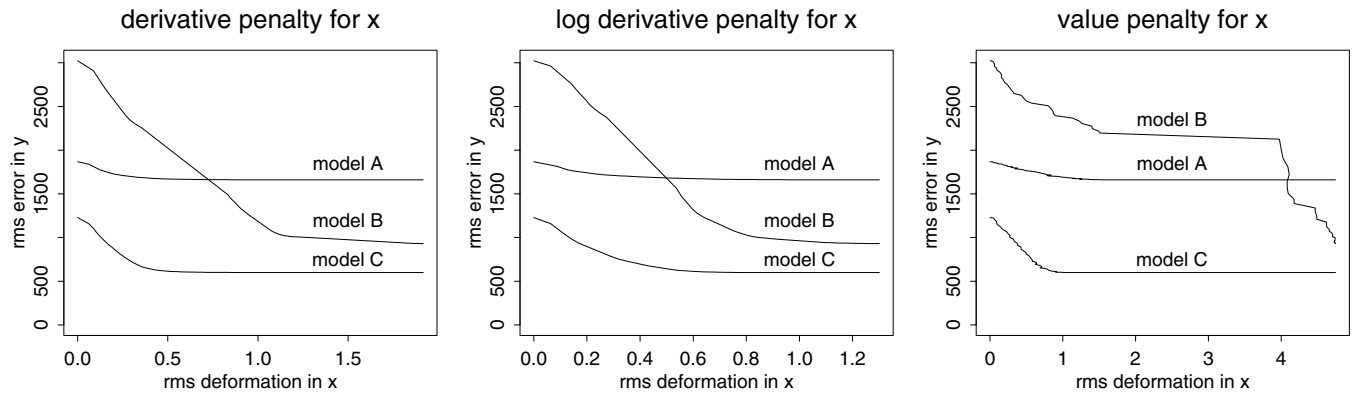

Figure 4. The tradeoff between prediction error in y and deformation error in x for the three models displayed in Figure 2. These three plots show deformation error as measured in three different ways.

these curves, we found the optimal deformation for hundreds of values of λ, then computed the components of error using these deformations.

Under these fitting criteria, model C dominates the other two models: for any given mean squared error (measured by G), the prediction from model C requires the least amount of deformation (as measured by F). Conversely, for any level of deformation, model C fits with the lowest mean squared error in y. Models A and B are very poor by comparison to model C. As Figure 3 illustrates, model A underpredicts the peak numbers of lynx trapped by a factor of 5–10, though it does correctly predict the approximate year and duration of at least the first couple of peaks. Model B, on the other hand, predicts the magnitudes and durations of the peaks within a factor of 2 or so, but misplaces the peaks badly in time.

The left plot in Figure 4 shows that most of the error in model A is error in the units of number of trappings: no amount of deformation substantially reduces the RMISE in the number of lynx trapped. In contrast, a large portion of the error in model B is error in placement of predictions over time: if moderate deformations are allowed (where the average difference between the deformation's derivative and 1 is about 1), the RMISE in lynx trapped is greatly reduced. This implies that if it is more important to predict the approximate magnitude of large fluctuations over extended periods than to predict the exact timing of these fluctuations (e.g., suppose we are trying to determine an inventory schedule over an extended period), then model B is a better guide than model A.

4.2.3 *Considering other measures of deformation, F.* The measure of deformation we have used thus far seems reasonable for many applications. However, it is relatively easy to alter the dynamic programming procedure to allow for different measures of deformation. We illustrate with the lynx predictions: the plot in the center of Figure 4 shows, for each of the three fitted models, the RMISE versus the standardized version of another deformation penalty, $F(x,\ f(x)) = [\log(f'(x))]^2$. The primary difference between this deformation penalty and the previous is this treats stretching and compression of the x-axis in a symmetric fashion (whereas the previous penalty favors compression of the x-axis). The rightmost panel in Figure 4 displays the errors using the penalty function $F(x,\ f(x)) = [f(x) - x]^2$. The λ values for this case are not directly comparable to those with the other measures of deformation; indeed, the units are not even the same: here, λ has units of lynx^2/ year^2 (once we standardize G and F by dividing by the length of time over which we integrate). The curves for model A and model B cross at $\lambda = 6900$, so that these models are "equally good" if we are indifferent between an error of 1 year or an error of 83 lynx ($6900^{1/2} \approx 83$).

5. A Two-Dimensional (Image) Example

5.1 *An Example Using Meteorological Data*

Our two-dimensional example uses a dataset from the University of Washington Department of Atmospheric Sciences Short-Range Ensemble Forecast System (SREF). The SREF is an ensemble of the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5), a community research numerical weather prediction model. Each member of the ensemble uses a different global analysis from numerical weather centers around the globe (see Table 1) as the initial condition for its solution. One member is also initialized from the "centroid" analysis, an analysis created by finding the mean for each parameter of the other six global analyses. Due to uncertainty in the initial conditions and slight differences in model physical approximations, each analysis is slightly different and results in a different numerical solution. The SREF is run daily for the northeast Pacific Ocean and the northwestern United States and Canada. The charts shown extend from the dateline (180) to 90E and from 30N to 60N. Numerical calculations are on a Lambert conformal grid with a grid spacing of approximately 36 km or 126 by 150 grid points.

The charts are the 48-hour predictions for the 500 mb pressure surface height of the seven ensemble member numerical forecasts initialized on 00 GMT October 30, 2002. The verifying analysis is the centroid analysis, the mean of the seven global analyses, for 00 GMT November 1, 2002. We treat the verifying analysis as the observations because these are the best estimates of what occurred on November 1 since they use observations from October 30 and the next 2 days to estimate the true pressure distribution on November 1. Pressure cannot be measured directly everywhere on the globe, so it

Table 1
Summary of numerical weather prediction methods used

Model name	Description	Production center
AVN	Global Forecast System	U.S. National Center for Environmental Prediction
CMCG	Global Environmental Multi-Scale Model	Canadian Meteorological Center
ETA	ETA Limited-Area Mesoscale Model	U.S. National Center for Environmental Prediction
GASP	Global Analysis and Prediction Model	Australian Bureau of Meteorology
NGPS	Navy Operational Global Atmospheric Prediction System	Fleet Numerical Meteorological and Oceanographic Center
TCWB	Global Forecast System	Taiwan Central Weather Bureau
CENT	Multianalysis Centroid (mean of the six models listed above)	University of Washington Atmospheric Sciences Department

must be reconstructed from a number of different sources. The weather pattern is a large blocking ridge extending up from the NE Pacific Ocean into British Columbia and two low-pressure systems starting to cut under the main ridge. Figure 5 shows the seven sets of predictions and the verifying analysis.

For each model we solved for the optimal deformation several dozen times using a decreasing sequence of values for λ. As in the lynx example, we can compare the models by examining the tradeoff between the deformation penalty and the RMISE (see Figure 6). The model with the greatest RMISE before deformation (TCWB) has the greatest amount of error in spatial placement, as witnessed by how much the RMISE can be lowered by spatial deformation. Based on examination of Figure 5, and given the RMISE without any deformation it is not clear that AVN and CENT are really better predictions than the others: perhaps the others simply slightly misplaced an area of high or low pressure. Figure 6 demonstrates that this is not the case because allowing for deformation does not largely alter the ordering of the quality of the predictions. Such conclusions would not be possible without a method that considers different components of error.

Finally, Figure 7 shows the optimal deformation for two sets of predictions. The misfit of the TCWB predictions largely consists of a shift to the upper left compared to the observations, while the misfit of the AVN predictions involves some twisting in the middle of the field. If such large-scale shifts are witnessed for a given model in many instances (as we see for TCWB here), then this suggests a simple way to improve the model.

6. Discussion

Predictions for indexed data are usually summarized in terms of error in y, but this is often inappropriate because modeling misspecification or misfit is often better summarized in terms of error in y and in x.

In this approach, the parameter λ controls the tradeoff between deformation in x and error in y. If $\lambda = 0$, the

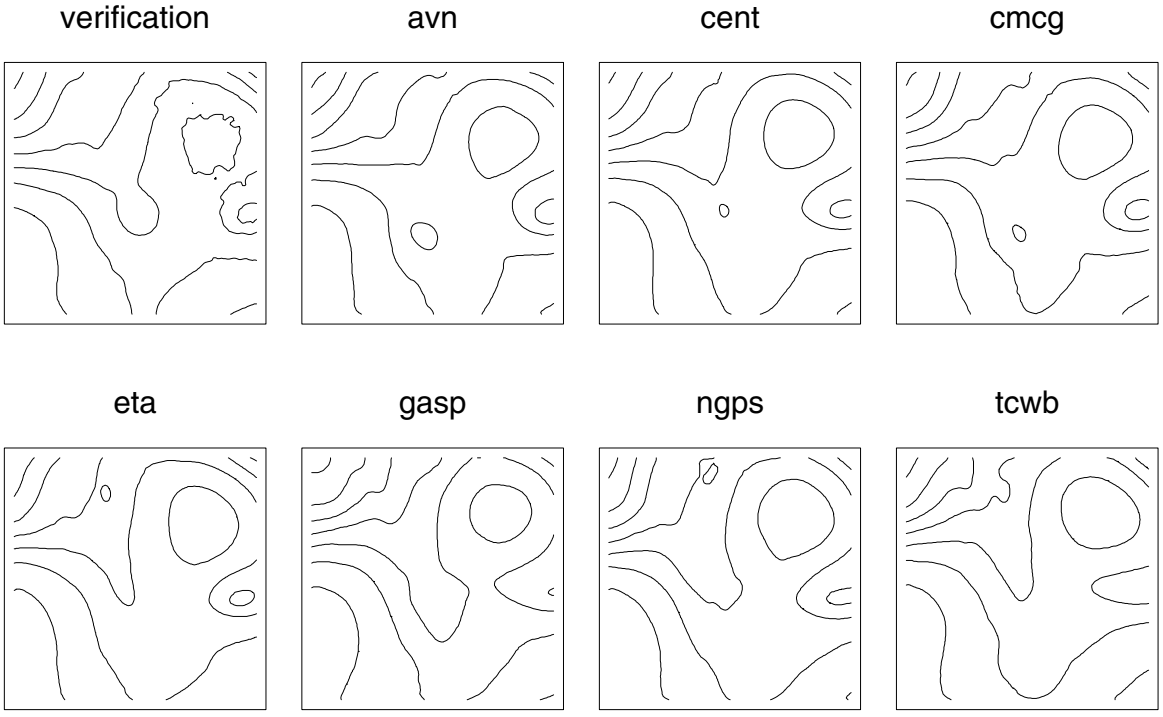

Figure 5. The verification analysis and seven sets of predictions of pressure.

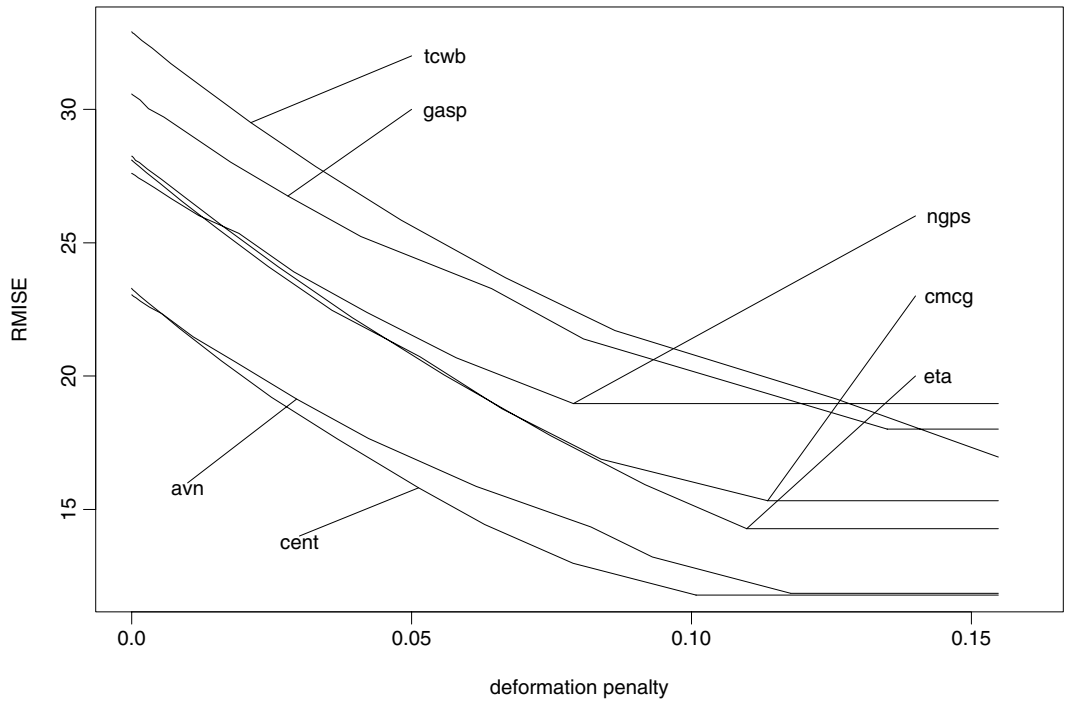

Figure 6. The tradeoff in deformation error and RMISE for the seven sets of predictions in Figure 5.

deformation minimizes the integrated error in y (subject to the elasticity constraint). As $\lambda \to \infty$, no deformation is allowed, and all error is assumed to occur in y. The tradeoff parameter λ may be selected by the researcher, or may in some cases be estimated from data. Alternatively, results can be calculated for a selection of values of λ as a way of investigating the properties of the model misfit.

We expect the main benefit of this approach to be decomposition of errors into separate components, perhaps for a range of λ as in the lynx example, but there are cases in which the statistic $\mathcal{I}_\lambda$ may be important in its own right. For example, if the deformation measure (and tradeoff parameter) are chosen correctly, then $\mathcal{I}_\lambda$ can be directly used for model selection. This application requires more study; for example, we have not investigated the sampling behavior of the statistic $\mathcal{I}_\lambda$, without which differences in $\mathcal{I}_\lambda$ for any given value of λ cannot be deemed due to systematic factors rather than noise. In a Bayesian predictive context this is quite simple because uncertainty in the predictions leads to uncertainty in $\mathcal{I}_\lambda$. If the predictions are a function of a vector of parameters,

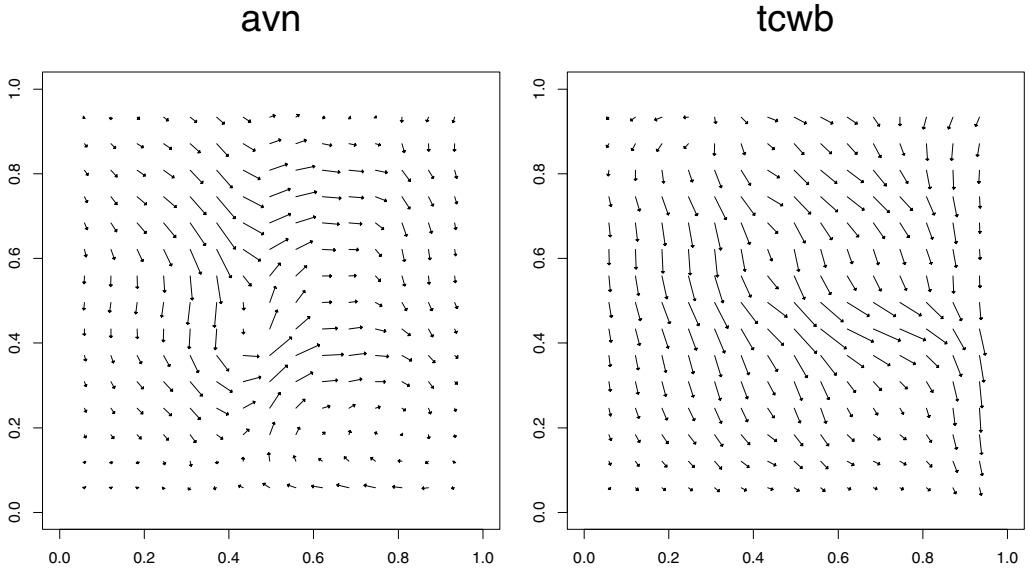

Figure 7. The limiting allowable deformation for two of the sets of pressure predictions from Figure 6 represented as vector fields.

then $\mathcal{I}_\lambda$ is a function of these same parameters, so if we have posterior samples of the parameter vector, then we use these samples to numerically integrate the parameter vector out of $\mathcal{I}_\lambda$.

Acknowledgements

We thank Upmanu Lall for helpful comments; the National Science Foundation for grants SBR-9708424, DMS-9796129, and SES-0084368; the Office of Naval Research grant N00014-01-10745; and the U.S. Department of Energy for support under contract no. DE-AC03-76SF00098.

Résumé

Les mesures conventionnelles d'ajustement à un modèle pour des données indexées (par exemple des séries temporelles ou des données spatiales) résument les erreurs en y, par exemple en intégrant (ou en sommant) le carré de la différence entre les valeurs prédites et les valeurs mesurées dans un large intervalle de valeurs de x. Nous proposons une approche qui envisage que les erreurs peuvent aussi se produire en x. Plutôt que de calculer simplement la différence entre prédiction et observation à chaque site (ou date), nous ≪déformons≫ d'abord les prévisions, en étirant ou en comprimant dans la (ou les) direction(s) x, pour améliorer l'accord entre les observations et les prévisions ≪déformées≫. L'erreur est alors résumée par a) la quantité de ≪déformation≫ en x, et b) la différence restant en y entre les données et les prévisions ≪déformées≫ (autrement dit l'erreur résiduelle en y après ≪déformation≫). Un paramètre, λ, contrôle le compromis entre a) et b), de sorte que si $\lambda \to \infty$ aucune déformation n'est autorisée, alors que si $\lambda \to 0$ la déformation minimise les erreurs en y. Dans certaines applications la déformation elle-même est intéressante acr elle caractérise la structure (temporelle ou spatiale) des erreurs. La déformation optimale peut être obtenue en résolvant un système d'équations aux dérivées partielles non-linéaires, ou pour un indice unidimensionnel en utilisant un algorithme de programmation dynamique. Nous illustrons la procédure par des exemples de séries temporelles non-linéaires et de dynmique des fluides.

References

Ames, W. (1992). *Numerical Methods for Partial Differential Equations*. New York: Academic Press.

Bellman, R. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Bookstein, F. L. (1991). *Morphometric Tools for Landmark Data*. Cambridge, U.K.: Cambridge University Press.

Boresi, A. and Chong, K. (2000). *Elasticity in Engineering and Mechanics*. New York: Wiley.

Campbell, M. and Walker, A. (1977). A survey of statistical work on the Mackenzie River Series of Annual Canadian lynx trappings for the years 1821–1934 and a new analysis. *Journal of the Royal Statistical Society A* **140,** 411–431.

De Jager, E. and Furu, J. (1996). *The Theory of Singular Perturbations*. Amsterdam: Elsevier.

Ebert, E. and McBride, J. (2000). Verification of precipitation in weather systems: Determination of systematic errors. *Journal of Hydrology* **239,** 179–202.

Elton, C. and Nicholson, M. (1942). The ten-year cycle in the numbers of lynx in Canada. *Journal of Animal Ecology* **11,** 215–244.

Frank, L. (1997). *Singular Perturbations in Elasticity Theory*. Amsterdam: IOS Press.

Glasbey, C. and Mardia, K. (2001). A penalized likelihood approach to image warping. *Journal of the Royal Statistical Society B* **63,** 465–514.

Goshtasby, A. (1988). Registration of images with geometric distortion. *IEEE Transactions on Geoscience and Remote Sensing* **26,** 60–64.

Hackbusch, W. (1985). *Multi-Grid Methods and Applications*. New York: Springer-Verlag.

Hoffman, R. N., Liu, Z., Louis, J.-F., and Grassotti, C. (1995). Distortion representation of forecast errors. *Monthly Weather Review* **123,** 2758–2770.

Lee, S., Wolberg, G., Chwa, K., and Shin, S. Y. (1996). Image metamorphosis with scattered feature constraints. *IEEE Transactions on Visualization and Computer Graphics* **2,** 337–354.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association* **54,** 173–205.

Martinson, D. G., Menke, W., and Stoffa, P. (1982). An inverse approach to signal correlation. *Journal of Geophysical Research* **87**(B6), 4807–4818.

Marzban, C. (1998). Scalar measures of performance in rare-event situations. *Weather and Forecasting* **13,** 753–763.

Mass, C. F., Ovens, D., and Westrick, K. (2002). Does increasing horizontal resolution produce more skillful forecasts? *Bulletin of the American Meteorological Society* **83,** 407–430.

Ogden, R. (1984). *Non-Linear Elastic Deformations*. New York: Wiley.

O'Malley, R. (1991). *Singular Perturbation Methods for Ordinary Differential Equations*. New York: Springer-Verlag.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C*. Cambridge, U.K.: Cambridge University Press.

Ramsay, J. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society, Series B* **60,** 351–363.

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. New York: Springer-Verlag.

Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis*. New York: Springer-Verlag.

Reilly, C. and Zeringue, A. (2004). Improved predictions of lynx trappings using a biological model. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, A. Gelman and X. Meng (eds), 297–308. New York: Wiley.

Renardy, M. and Rogers, R. (1993). *An Introduction to Partial Differential Equations*. New York: Springer-Verlag.

Tong, H. (1990). *Non-Linear Time Series*. Oxford: Oxford University Press.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society B* **40,** 364–372.

Yan, R., Tokuda, N., Miyamichi, J., and Ni, Y. (2000). Image morphing by spatial thin-plate spline transformation. *Information Processing Society of Japan* **37**.

Received June 2003. *Revised January* 2004.
Accepted March 2004.