CARNEGIE MELLON UNIVERSITY

NONSTATIONARY GAUSSIAN PROCESSES
FOR REGRESSION AND SPATIAL MODELLING

A DISSERTATION
SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
for the degree
DOCTOR OF PHILOSOPHY
in
STATISTICS

by
Christopher Joseph Paciorek

Department of Statistics
Carnegie Mellon University
Pittsburgh Pennsylvania 15213

May 2003

Abstract

Recent work in the areas of nonparametric regression and spatial smoothing has focused on modelling functions of inhomogeneous smoothness. In the regression literature, important progress has been made in fitting free-knot spline models in a Bayesian context, with knots automatically being placed more densely in regions of the covariate space in which the function varies more quickly. In the spatial statistics literature, attention has focused on using nonstationary covariance structures to account for inhomogeneity of the spatial field.

In this dissertation, I use nonstationary covariance functions in Gaussian process (GP) prior distributions over functions to perform both nonparametric regression and spatial smoothing in a Bayesian fashion. I extend the kernel convolution method of Higdon et al. (1999) to create a class of nonstationary covariance functions. I prove that the nonstationary covariance functions retain the differentiability properties of the stationary correlation functions on which they are based, provided there is sufficient smoothness in the underlying kernel structure used to generate the nonstationarity. The stationary Matérn covariance function has desirable differentiability properties; the generalized kernel convolution method developed here provides a Matérn-based nonstationary covariance function.

I develop a generalized nonparametric regression model and assess difficulties in identifiability and in fitting of the model using Markov Chain Monte Carlo (MCMC) algorithms. Of particular note, I show how to improve MCMC performance for non-Gaussian data based on an approximate conditional posterior mean. The modelling approach produces a flexible response surface that responds to inhomogeneity while naturally controlling overfitting. For Gaussian errors, on test datasets in one dimension, the GP model performs well, but not as well as the free-knot spline method. However, in two and three dimensions, the nonstationary GP model seems to outperform

both free-knot spline models and a stationary GP model. Unfortunately, as implemented the method is not feasible for datasets with more than a few hundred observations because of the computational difficulties involved in fitting the model.

The nonstationary covariance model can also be embedded in a spatial model. In particular, I analyze spatiotemporal climate data, using a nonstationary covariance matrix to model the spatial structure of the residuals. I demonstrate that the nonstationary model fits the covariance structure of the data better than a stationary model, but any improvement in point predictions relative to a stationary model or to the maximum likelihood estimates is minimal, presumably because the data are very smooth to begin with. My comparison of various correlation models for the residuals highlights the difficulty in fitting high-dimensional covariance structures.

Acknowledgements

I would like to acknowledge the advice, suggestions, support, and friendship of a number of people who helped me during the writing of this thesis and the rest of my time as a graduate student in the Department of Statistics. First, I would like to thank my advisor, Mark Schervish, for his ongoing involvement in this work. In his understated way, Mark offered hands-on advice without hands-on management, giving suggestions that often guided me in a better direction. He has always been willing to delve into the details of the work and talk through issues, sometimes long past a reasonable meeting length. I would also like to thank the other members of my committee. James Risbey got me started on the Advanced Data Analysis project that led to this thesis. I thank him for his collaboration, advice, and friendship. I thank Doug Nychka for an enjoyable and productive visit to NCAR as well as for his various suggestions and his assistance with the wavelet model. I thank Valérie Ventura for her support and for her collaboration on two related projects. Larry Wasserman offered an open door, of which I did not take as much advantage as I would have liked toward the close of my graduate work. The statistics department staff and the rest of the faculty in the department have helped me in various ways, and I thank them for making the department a great place to do graduate work. I'd also like to thank Chris Holmes for the Matlab code on his website.

I want to thank the graduate students in the department. In particular, Fang Chen, my longtime officemate, has tolerated with good humor my often foul moods during the writing of this thesis.

Seven years is a long time to be in graduate school; I'm glad Susan Davidson was there for many of them. I'd also like to thank old friends who do not need to be named to know their importance to me. Little did we know in grade school and high school that I would be 31 when I finally finished school. John, how many seconds have I been in school?

Finally, but most importantly, I thank my parents for setting me on the road with what I would need to get to this point.

Contents

List of Tables

List of Figures