BAYESIAN PREDICTIVE INFERENCE FOR SAMPLES FROM SMOOTH PROCESSES

by

J. V. Zidek
and
S. Weerahandi

Technical Report #106

Department of Statistics, University of British Columbia

March, 1991

WS/PR:ddp

BAYESIAN PREDICTIVE INFERENCE FOR SAMPLES FROM SMOOTH PROCESSES

by

J. V. Zidek and S. Weerahandi [1]
The University of British Columbia and Bellcore

ABSTRACT

This paper extends a method of the authors to obtain a Bayes linear procedure for predicting the value at a specified time of a future sample path. It is supposed that data are available from a number of sample paths which are deemed to be related to the future sample path. The method is local and relies on a simple linear model which derives from the Taylor expansions of the processes at the point at which the inference is required. It is argued that the resulting Bayes linear predictor is approximately the same as that from any one of a family of complex Bayes linear predictors which obtain under the assumption that the processes possess several derivatives. An illustrative application to growth curve analysis is used to bring out some of the strengths and weaknesses of the proposed method.

Key words and Phrases. Linear smoothing; Bayes linear predictors; nonparametric regression; multiple time series; splines; kernel regression methods; locally weighted regression; multiple regression; LOESS; growth curve analysis.

syntax error file title, between lines 38 and 38

Prepared in part under support to SIMS from the United States Environmental Protection Agency and in part from a grant provided by the Natural Science and Engineering Research Council of Canada.

1. INTRODUCTION

A problem of predictive inference for multiple time series is the subject of this paper. The proposed solution uses an extension of the method of Weerahandi and Zidek (1988, hereafter WZ), described by Cleveland and Devlin (1988) as a Bayesian version of locally weighted regression. A preliminary version of this paper is contained in Weerahandi and Zidek (1986).

The context of our work is that in which data are obtained from each of a set of sample paths, like growth curves for example. The observables are in a partitioned data column vector, $Y = (Y_1^T, \cdots, Y_m^T)^T$ with $Y_i : n_i \times 1 = (Y_i(t_{i1}), \cdots, Y_i(t_{in_i})$, for $i = 1, \cdots, m$. It is supposed that $Y = S + N$, where S and N are partitioned in conformation with Y. The uncorrelated coordinates of N represent noise; they have mean zero and a common variance, σ_N^2. And $S_i = (S_i(t_{i1}), \cdots, S_i(t_{in_i}))$ where $S_i(t)$ is $P+1$ times differentiable in quadratic mean.

An object of particular interest is $\beta_{m+1}^T = (S_{m+1}(t_{m+1}), S_{m+1}^{(1)}(t_{m+1}), \cdots, S_{m+1}^{(p)}(t_{m+1}))$ where S_{m+1} represents a possibly as yet unsampled process and t_{m+1}, a possibly as yet unused sample point. Bracketed superscripts denote L_2 derivatives of the process and $p \leq P$. More generally, the object of inferential interest may be $\beta = (\beta_1^T, \cdots, \beta_m^T, \beta_{m+1}^T)^T$, where β_i is defined for all i as in the case $i = m + 1$.

As is well known, the optimal linear, that is, Bayes linear procedure with respect to a quadratic loss function is

$$\beta_Y = E\beta + \alpha_{\beta Y}(Y - EY) \tag{1.1}$$

where in general, for any two random vectors U and V, $\alpha_{U|V} = \Gamma_{UV}\Gamma_{VV}^{-1}$, and $\Gamma_{UV} = E(U - EU)(V - EV)^T$. The reliability of β_Y is specified by $\Gamma_{\beta \cdot Y}$ where, in general, for any two random vectors, U and V, $U \cdot V = U - [EU + \alpha_{U|V}(V - EV)]$ and $\Gamma_{U \cdot V} = E(U \cdot V)(U \cdot V)^T$. It is also well known that a Bayes linear procedure is Bayes when Y and β have a joint Gaussian distribution. But even when the Gaussian distribution does not obtain, linear procedures are commonly used as they are here because of their simplicity.

In spite of their relative simplicity, the global modeling required to specify these procedures can be very demanding. A first step towards the development of such a model, which may simplify reasoning by enabling its hierarchical specification, begins with the linear model,

$$Y = EY + A(\beta - E\beta) + E\ , \tag{1.2}$$

where $A = \alpha_{Y\beta}$ and E is independent of β. By construction, E has mean zero and a covariance matrix given by $C = \Gamma_{Y \cdot \beta}$. Model (1.2) yields

$$\Gamma_{YY} = C + A\Gamma_{\beta\beta}A^T \tag{1.3}$$

and

$$\Gamma_{\beta Y}=\Gamma_{\beta\beta}A^T. \tag{1.4}$$

From equations (1.3) and (1.4), an expression for $\alpha_{\beta Y}$ is readily found. A well known alternative expression is:

$$\alpha_{\beta Y}=\Gamma_{\beta\cdot Y}A^T C^{-1}, \tag{1.5}$$

where

$$\Gamma_{\beta\cdot Y}=(\Gamma_{\beta\beta}^{-}+A^T C^{-1}A)^{-1}. \tag{1.6}$$

Even with the help of the model (1.2) the task of specifying the global prior distribution may be substantial. And it may not be necessary. If the sampling points around the point of interest, $t=\tau$, are of sufficiently high density and a semi-Markov-like property defined precisely in Section 2 is believed to hold, then the data outside a window located at $t=\tau$ would be relatively unimportant. Sometimes the data outside such a window are excluded (c.f. Muller, 1987), usually on heuristic grounds, we believe.

Here the implication of excluding the data outside an appropriate data window is that the elicitation of prior information can be restricted to that concerning just certain local rather than global parameters. These will be defined precisely in Section 2.

Excluding data as described above seems desirable where permissible not only because this simplifies the task of prior modeling, but also because this increases the procedure's robustness. It avoids the risk and possible consequences of misspecifying the global model, say by making a convenient choice from the time series catalogue, with its preponderance weakly stationary models. And it avoids the potential negative impact of errors in those data which contribute little to the optimal procedure anyway.

To construct a "local" model, suppose remote data have already been excluded so that the coordinates of Y are just the data for which $\Delta_{ij}=t_{ij}-\tau$ are small. Taylor's theorem implies that

$$Y = A^0\beta + E^0 \tag{1.7}$$

where A^0 is an approximation to A defined in the next section; E^0 includes the Taylor remainders, has expected value zero, is approximately uncorrelated with β, and has a covariance matrix, C^0 which would be expected to be small in magnitude. It will be argued on largely heuristic grounds that equations (1.2) and (1.7) are approximately equivalent under suitable regularity conditions. Zidek and Weerahandi (1990) go further in a special case and rigorously derive bounds on the approximation errors involved in substituting (1.7) for (1.2) when the range of the coordinates of Y is an arbitrary finite dimensional inner product space, $p=0$, and $P\geq 2$. In any case, the result is a Bayes linear procedure which may be viewed as an approximation to the "true" Bayes linear procedure, which would be obtained by specifying completely the a priori

model for the processes involved. Likewise, an approximation, $\Gamma^0_{\beta\cdot Y}$, is obtained for $\Gamma_{\beta\cdot Y}$. The latter is an important object in that it indicates the uncertainty in the Bayes linear rule.

When Y and β have a joint Gausssian distribution, our approximations yield an approximate posterior distribution for β and hence β_{m+1}. This may in turn be used to find attributes of interest for β such as credibility regions.

The problem of specifying the local prior distribution for β is discussed in Section 2, as is that of specifying the covariance hyperparameters.

Our interest in the problem addressed in this paper derived from discussions with Dr Ned Glick in 1978 when a very preliminary version of the approximation given in WZ was formulated. The practical problem addressed in those discussions was predictive inference for a future child in a certain growth curve study where there was just one datum from each of a random sample of children. The problem is different from that of classical time series analysis where a number of often equally spaced values are obtained from a single sample path, although the latter is a special case of the general situation considered here.

The example presented in Section 4 also derives from a growth curve study but here there are several values from each child's growth curve and these are taken at the same equally spaced values for all the children. Our results are in close agreement with those obtained by Fearn (1975) by a different Bayesian analysis.

2. APPROXIMATE BAYES LINEAR PROCEDURES

In this section, the model in equation (1.7) will be derived in an explicit form from (1.2). In this derivation, $U = O_p(\delta)$ will mean $|U/\delta|$ has bounded expectation for any $\delta>0$ when U any random vector . An analogous meaning is assigned to $u = O(\delta)$ when u is not a random object.

Let us adopt the following notation:

$$
\begin{aligned}
a_{ij} &= \Delta_j^r/r! \quad , r=0,1,\ldots,P \ , \\
a_{ij} &= (a_{ij0}, \ldots, a_{ijp}) \ , \quad \text{and} \\
\beta_{ir} &= S_i^{(r)}(\tau) \ , \quad r=0,1,\ldots,P \ , \\
\beta_i &= (\beta_{i0}, \ldots, \beta_{ip})^T \ , \quad i=1,2\ldots,m+1 \ , \\
\beta &= (\beta_1^T, \ldots, \beta_{(m+1)}^T)^T \ .
\end{aligned}
\tag{2.1}
$$

By the assumptions of Section 1 and Taylor's theorem

$$S_{ij} = \sum_{r=0}^{p} a_{ijr}\beta_r + O_p(|\Delta|_{ij}^{p+1}) \tag{2.2}$$

Thus

$$ES_{ij} = \sum_{r=0}^{p} a_{ijr}E\beta_r + O_p(|\Delta|_{ij}^{p+1})$$

and hence

$$\begin{aligned} \bar{S}_{ij} &\overset{\Delta}{=} S_{ij} - ES_{ij} \\ &= \sum_{r=0}^{p} a_{ijr}\bar{\beta}_r + O_p(|\Delta|_{ij}^{p+1}) \end{aligned}$$

where $\bar{\beta}_r \overset{\Delta}{=} \beta_r - E\beta_r$ from which it follows that $\bar{\beta}_r = \bar{S}_i^{(r)}(\tau)$.

A straight forward calculation now gives

$$\alpha_{\bar{S}_{ij}\beta}\bar{\beta} = \sum_{r=0}^{p} a_{ijr}\bar{\beta}_r + \sum_{r=p+1}^{p} a_{ijr}\alpha_{\beta_r\beta}\bar{\beta} + O_p(|\Delta|_{ij}^{p+1}) \tag{2.3}$$

and

$$E(S_{ij}\cdot\beta) = \sum_{r=p+1}^{p} a_{ijr}E(\beta_r\cdot\beta) + O_p(|\Delta|_{ij}^{p+1}) . \tag{2.4}$$

Combining equations (2.3) and (2.4) yields

$$S_{ij}(\beta) = \sum_{r=0}^{p} a_{ijr}\beta_r + \sum_{r=p+1}^{p} a_{ijr}\beta_r(\beta) + O_p(|\Delta|_{ij}^{p+1}) . \tag{2.5}$$

where, in general, $U(V) = E(U) + \alpha_{UV}(V - E(V))$. Reinvoking equation (2.2) gives

$$S_{ij}\cdot\beta = \sum_{r=p+1}^{p} a_{ijr}\beta_r\cdot\beta + O_p(|\Delta|_{ij}^{p+1}) . \tag{2.6}$$

Since by definition, $S_{ij} = S_{ij}(\beta) + S_{ij}\cdot\beta$, a fundamental decomposition is obtained:

$$S_{ij} = \sum_{r=0}^{p} a_{ijr}\beta_r + \sum_{r=p+1}^{p} a_{ijr}\beta_r(\beta) + \sum_{r=p+1}^{p} a_{ijr}(\beta_r\cdot\beta) + O_p(|\Delta|_{ij}^{p+1}) . \tag{2.7}$$

By combining the second and third terms in equation (2.7), equation (2.2) is obtained; the latter is the basis on which Weerahandi and Zidek (1986) build their Bayes linear inferential procedure.

In this paper we share with Zidek and Weerahandi (1990) the goal of finding a single model which approximates those described by equation (1.2) in the sense that it yields a local approximation to each member of the class of Bayes linear procedures which are implied by (1.2). We therefore use the decomposition in (2.7) to suggest an

approximation to the model in equation (1.2). In particular to order p, the approximation A_i^0 to A_i has $i-j$ th row defined by

$$A_{ij}^0\beta = a_{ij}\beta_i \quad .$$

Equation (2.7) yields an estimate of the error in this approximation:

$$(A_{ij} - A_{ij}^0)\beta = \sum_{r=p+1}^{P} a_{ijr}\alpha(\beta_r,\beta)\beta \quad .$$

An analogous approximation, C^0 for C, is obtained below along with an estimate of the error in the approximation. Zidek and Weerahandi (1990) use these estimates of the errors in the approximations to A and C to determine bounds on the approximation errors induced in β_Y and $\Gamma_{\beta,Y}$ but we will not seek such bounds here.

From equation (2.6) we obtain,

$$\Gamma[(S_{ij}-\beta)(S_{kl}-\beta)] = \sum_{r=p+1}^{P}\sum_{s=p+1}^{P} a_{ijr}a_{kls}\Gamma_{rs}^{(r,s)} + O(|\Delta_{ij}|^{P+1}+|\Delta_{kl}|^{P+1}) \quad , \qquad (2.8)$$

for all i,j,k,l, where $\Gamma_{rs}^{(r,s)} = \Gamma_{\beta_r,\beta,\beta_s,\beta}$ is the residual covariance between $S_i^{(r)}(\tau)$ and $S_k^{(s)}(\tau)$ when the linear effect of β have been factored out. Assume like Weerahandi and Zidek (1988,1990) in the special cases they treat, $P \geq 2p+2$. Then to the order of the linear model for S based on β, the square root of the absolute value of the quantity in (2.8) is zero; locally the Taylor expansion has removed all variation and covariation in the S-processes. This was the heuristic basis for the model proposed in the special cases treated by Weerahandi and Zidek(1986, 1988).

But a global approximant to the quantity in (2.8) is required which will yield a locally weighted predictive procedure and at the same time preserve its local character. In Weerahandi and Zidek (1986,1988), the approximant was chosen to be a diagonal matrix for simplicity. However in Zidek and Weerahandi (1990) where $p=0$, the locally dominant term on the right hand side of equation (2.4) is retained to capture the residual covariance structure (and enable bounds on approximation errors to be found). This term will also be retained here.

Let the approximation to C be given by

$$C_{(ij)(kl)}^0 = \Sigma_{(ij)(kl)} + \eta_{(ij)(kl)} + \zeta_{(ij)(kl)} \quad . \qquad (2.9)$$

where $\Sigma_{(ij)(kl)}=0$ unless $i=k$ and $j=l$ when $\Sigma_{(ij)(ij)}$ the variance of the noise in $Y_i(t_{ij})$ is positive. Furthermore

$$\eta_{(ij)(kl)} = \Delta_{(ij)(kl)}\Gamma_{\alpha}^{(p+1,p+1)}$$

where

$$\Delta_{(ij)(kl)} = \frac{\Delta_{ij}^{(p+1)}\Delta_{kl}^{(p+1)}}{[(p+1)!]^2}\ e^{-(|\Delta_{ij}|+|\Delta_{kl}|)/\Delta_0} \quad . \qquad (2.10)$$

Finally

$$\zeta_{(ij)(kl)} = |\Delta_{ij}|^{(p+1)}|\Delta_{kl}|^{(p+1)}(|\Delta_{ij}|+|\Delta_{kl}|)\sigma_{(ij)(kl)} \quad , \tag{2.11}$$

where the $\sigma_{(ij)(kl)}$ must be selected to assure positive definiteness of the resulting approximant; $(\sigma_{(ij)(kl)})$ could, in particular, be a diagonal matrix.

The approximation in equation (2.9) is meant to replace the higher order terms in (2.8) on the one hand and at the same time insure by making $\sigma_{(ij)(kl)}$ sufficiently large, that values of $Y_i(t_{ij})$ for which $|\Delta_{ij}|$ is large are "windowed out". This leads to the approximate model in equation (1.7), subject to the selection of the covariance hyperparameters in equation (2.9).

The more rigorous approach of Zidek and Weerahandi (1990) would entail partitioning Y as $(P^T,R^T)^T$ where P is the vector of data values in a window at $t=\tau$. The general results of Section 2 of Zidek and Weerahandi (1990) could now be applied. The error, $\beta_Y-\beta_P$, is (ibid, Theorem 1)

$$\Gamma_{\beta R\cdot P}\Gamma_{R\cdot P}^{-1}R \tag{2.12}$$

where $\Gamma_{\beta R\cdot P} = \Gamma_{\beta R} - \Gamma_{\beta P}\Gamma_{P\beta}^{-1}\Gamma_{R\beta}$ and $\Gamma_{R\cdot P}$ is defined in the Introduction. It can be shown that the R to P correlation structure can be recovered from $\Gamma_{S\cdot\beta}$; the result is

$$\Gamma_{\beta R\cdot P}\Gamma_{R\cdot P}^{-1} = D\alpha^*[\Gamma_{R\cdot P\beta} + \alpha^* D\alpha^{*T}]^{-1} \tag{2.13}$$

where

$$\alpha^* = \alpha_{R P\cdot\beta}\alpha_{P\beta} - \alpha_{R\beta} \quad ,$$

$$\alpha_{R P\cdot\beta} = \Gamma_{R P\cdot\beta}\Gamma_{P\beta}^{-1} \quad ,$$

$$\Gamma_{R\cdot P\beta} = \Gamma_{R\cdot\beta} - \Gamma_{R P\cdot\beta}\Gamma_{P\beta}^{-1}\Gamma_{P R\cdot\beta} \quad ,$$

and

$$D = [\Gamma_{\beta\beta}^{-1} + \alpha_{\beta P}^T\Gamma_{P\cdot\beta}\alpha_{P\beta}]^{-1} \quad .$$

The expression in (2.13) is useful in obtaining error bounds for the approximation to β_Y given by equation (1.7).

In any case, reduction to P from Y can only be justified if it is believed that

$$||\beta_Y-\beta_P||/||\beta_Y|| \tag{2.14}$$

is small. This is a generalized semi-Markov property. For an $AR(1)$ process (c.f. Zidek and Weerahandi (1990)), the error will be zero when P consists of just data points on either side of $t=\tau$. Of course, the error in (2.14) need not be evaluated as long as it is believed to be small and this is the heuristic, presumably, which underlies all methods which use data windows. Of course, the validity of the approximation of β_P by β^{0_P} derived from equation (1.7) still obtains under the regularity conditions given here. But unless the generalized semi Markov condition holds, valuable information in the

data points remote from the point at which inference in being made, $t = \tau$, may be lost.

To rigorously justify the model in equation (1.7) entails showing the error in (2.14) is small when β_Y and β_P are replaced by this corresponding approximants, β^{0Y} and β^{0Y}. Zidek and Weerahandi (1990) address this issue (when $p=0$). It is plausible that this error will be small in the present context when $\sigma_{(ij)(kl)}$ in equation (2.11) is sufficiently large.

In summary, the analysis of this section has lead us to the linear model on equation (1.7), where the ij th row of A^0 is given by

$$A^0_{ij}\beta = a_{ij}\beta_i$$

for all β, where E^0 is uncorrelated with β, and where $C^0 = \Gamma^{0Y+}$ is determined by equation (2.9). This in turn leads to an approximate Bayes linear procedure. It should be emphasized that in this section we are assuming the covariance hyperparameters have been specified. We address the problem of specifying them in the next section where we give a particular implementation of our proposed approximation.

3. EXCHANGEABLE GAUSSIAN PROCESS

In this section a special case of the model in equation (1.7) will be investigated and a further approximation to the residual covariance matrix introduced. For simplicity, the subscript, "0", imposed in Sections 1 and 2 on A and C to denote their approximations, will be suppressed.

The problem of specifying θ, the vector of covariance hyperparameters, will be addressed below, but suppose for now it has been specified. Assume the S_i 's and the noise processes are Gaussian, that the noise is homoscedastic and that the S_i 's are exchangeable. To be precise, we suppose that in equation (2.9) $\Sigma_{(ij)(kl)} = \sigma^2$ or 0 according as $i=k$ and $j=l$ or not, and $\sigma_{(ij)(kl)} = \sigma^2_R$ or 0 according as $i=k$ and $j=l$ or not. Furthermore $\Gamma^{(p+1,p+1\beta)}_{ik}$ has a common value, say γ_B, for all pairs, $i \neq k$ and likewise $\Gamma^{(p+1,p+1\beta)}_{ii} = \gamma_W$ for all i. Assume $\gamma_B = \gamma_W = 0$, an assumption which is justified under reasonable conditions indicated in the Appendix. Finally suppose that conditional on γ,

$$\beta_i \mid \gamma \overset{ind}{\sim} N_{(p+1)}(\gamma, \Lambda) \ ,$$

in the spirit of Lindley and Smith (1972, hereafter LS). Indeed, if γ is supposed to have a uniform (improper) prior distribution, the results of LS show that

$$\beta \mid Y, \theta \sim N_{(m+1)(p+1)}(D_\theta d_\theta, D_\theta) \qquad (3.1)$$

where in the notation of LS

$$D_0^{-1} = A_1^T C_1^{-1} A_1 + C_2^{-1} - C_2^{-1} A_2 (A_2^T C_2^{-1} A_2)^{-1} A_2^T C_2^{-1} ,$$

and

$$d_0 = A_1^T C_1^{-1} Y .$$

To translate these results in the present context, set $C_1 = C$, $A_1 = A$, $C_2 = Diag\{\Lambda, \cdots, \Lambda\}$, $A_2 = (I, \cdots, I)^T$, an $(m+1)(p+1)\times(p+1)$ matrix with I denoting the $(p+1)\times(p+1)$ identity matrix.

It is straightforward to show that

$$A_2^T C_2^{-1} A_2 = (m+1)\Lambda^{-1} .$$

Thus

$$C_2^{-1} A_2 (A_2^T C_2^{-1} A_2)^{-1} A_2^T C_2^{-1} = (m+1)^{-1} A_2 \Lambda^{-1} A_2^T .$$

The assumptions made above in this section entail

$$C = \sigma^2 I + \sigma_g^2 D$$

where

$$D = Diag\{D_1, \cdots, D_m\}$$

and

$$D_i = Diag\{|\Delta_{i1}|^{2p+3}, \cdots, |\Delta_{in_i}|^{2p+3}\} .$$

Now

$$D_0^{-1} = e - (m+1)^{-1} A_2 \Lambda^{-1} A_2^T ,$$

where $e = Diag\{e_1, \cdots, e_{(m+1)}\}$, $e_i = v_i + \Lambda^{-1}$, $v_i = a_i^T c_i^{-1} a_i$ or 0 according as $i \le m$ or $i = m+1$, $c_i = \sigma^2 I_{n_i} + \sigma_g^2 D_{i,}$ $i = 1, \ldots, m$, and $a_{i,}$ $i = 1, \cdots, m$ is the $n_i \times (p+1)$ matrix whose j-th row is a_{ij} , $j = 1, \cdots, n_i$, $i = 1, \cdots, m$, that is,

$$a_{ij} = (a_{ij0}, \cdots, a_{ijp}) ,$$

with $a_{ijr} = \Delta_{ij}^r / r!$ for $r = 0, \ldots, p$. In fact, $a_{i,}$ $i = 1, \cdots, m$ is a submatrix of A (A^0 in the last section) defined just below equation (2.7) with $A = (Diag\{a_1, \cdots, a_m\}, 0)$.

We note in passing that when, as in the next section, the n_i are identical, and the observations for each process Y_i are taken at the same time points, then the a_i are identical as are the c_i's. So the v_i, $i = 1, \cdots, m$ are identical in this case.

From the well known matrix equation (see LS), $(u + v^T v)^{-1} = u^{-1} - u^{-1} v^T (t^{-1} + v u^{-1} v^T)^{-1} v u^{-1}$,

$$D_0 = e^{-1} + e^{-1} A_2 F^{-1} A_2^T e^{-1} ,$$

where $F = \Lambda \sum_{i=1}^{m} F_i^{-1} \Lambda$ and $F_i = v_i^{-1} + \Lambda$.

Of particular interest is the marginal posterior distribution of $\beta_{(m+1)}$ which is easily deduced from equation (3.1). It is

$$\beta_{(m+1)} | Y, \theta \sim N_{(p+1)}(\beta^*_{(m+1)}, \Sigma_{(m+1)}) , \tag{3.2}$$

where

$$\Sigma_{(m+1)} = \Lambda + [\sum_{i=1}^{m} F_i^{-1}]^{-1} ,$$

$$\beta^*_{(m+1)} = [\sum F_i^{-1}]^{-1} \sum F_i^{-1} \hat{\beta}_i ,$$

and $\hat{\beta}_i = (a_i^T c_i^{-1} a_i)^{-1} (a_i^T c_i^{-1} Y_i)$. In other words, $\beta^*_{(m+1)}$ is a weighted average of the local least squares estimators of the β_i. When the F_i's are equal as they are in the next section, the weights are identically equal to $1/m$. The distribution in (3.2) serves as a basis for inference about β_{m+1} if the covariance hyperparameters have been specified. However, in practice these parameters will not usually be specified at this first stage of hierarchical modeling so we turn briefly to the issue underlying these considerations.

The variance of the noise process (in the homoscedastic case) is a global covariance hyperparameter unlike the other parameters in our approximate model which are local. Our theory is deficient in that it does not deal adequately with global model parameters, inference about which should be based on all the data. Instead for all approximately linear models (in the sense of Sacks and Ylvisaker, 1979), inference in both repeated sampling contexts like that of Sacks and Ylvisaker and Bayesian contexts (see Zidek and Weerahandi, 1990) will rely primarily on the data for which the contribution from the local model is small. Here this means, the data in a window at τ. To deal with the complex issue of estimating the global components of approximately linear models, entails simultaneous inference across all τ. A model like that underlying the theory of splines for example, (c.f. Wahba, 1982) would be needed. It should be emphasized however, that the priors for splines are highly specialized and we believe they are insufficiently flexible as to adequately represent a reasonable spectrum of prior opinion.

The seemingly natural approach to dealing with the remaining covariance hyperparameters (see equation (3.1)) entails putting a distribution on σ^2_δ and Λ. And provided that the conditional expectations of the S processes given these parameters does not depend on them, then this may be done. That is, it does not matter whether the approximation step in going from equation (1.2) to (1.7) is carried out before or after "marginalizing out" these covariance hyperparameters. However, it is not clear how well the marginal distribution from (1.2) would be approximated by that from its simpler relative, (1.7). Clearly the Gaussian model of Section 3 will be lost either way. For practical reasons, it is tempting to adopt the simple model in (1.7) as we do here. But this matter deserves further investigation.

The determination of the marginal posterior distribution of θ is straightforward. The result is

$$\pi(\theta|Y) = K|C|^{-1/2}|A|^{-m/2}|D_\theta|^{1/2}\exp[-1/2\ Y^T G^{-1} Y]\ \pi(\theta) \qquad (3.3)$$

where K is the normalization constant whose exact value is unnecessary for our purposes, $G = C^{-1} - C^{-1}AD_\theta A^T C^{-1}$ and $\pi(\theta)$ is the prior density of θ.

By combining (3.2) and (3.3), the joint distribution of $\beta_{(m+1)}$ and θ is obtained and from this in turn the marginal distribution of $\beta_{(m+1)}$. We will not discuss in general, the problem of determining the prior distribution of θ. But in the next section we will make a particular choice for the example considered there.

4. APPLICATION TO GROWTH CURVE ANALYSIS.

In this section the approximate model developed in Sections 2 and 3 is applied to data used by Grizzle and Allen (1969) in their frequency theory analysis and Fearn (1975) in his Bayesian analysis of growth curves. These data consist of the Ramus heights (in mm), of 20 boys at 8, 8 1/2, 9 and 9 1/2 years of age. The data were collected to establish a normal growth curve for the use of orthodontists.

The example is a challenging one chosen to bring out strengths and weaknesses of our proposed method. As we will argue below, its use seems inappropriate here given the assumptions we have made in its development. And Fearn's analysis suggests the regression curve of Ramus height on age is very well approximated by a parametric (in fact, linear) model so our nonparametric approach seems unrealistic from the outset. Nevertheless, surprisingly good agreement with Fearn's results will be obtained.

Figure 1 depicts the data along with the results of our analysis. From a superficial examination, the inter-sampling points seem large while our theory is designed for data which are clustered around the point of inference. But sampling intensity is a relative concept and must be measured against the inherent variability of the process. So our immediate reaction to Figure 1 may be premature. Indeed, we do not yet know how to assess the adequacy of sampling intensity. There does not seem to be a generally satisfactory way of addressing this issue even though it underlies discrete time series analysis which always concerns itself, implicitly at least, with samples from underlying continuous time processes.

The number of derivatives to include in the model of equation (1.7) is somewhat arbitrary, although in some situations there may be physical models which dictate a useful upper bound. For example, if the underlying S_i's were Brownian paths, then of necessity p would be zero. However, in this case the condition in the Section 2 that $P \geq 2p+2$ would be violated; the justification given in Section 2 for using our method would be lost, although the method could still be applied and may even be justified on

other grounds.

Derivatives must be included in equation (1.7) when inference about them is required. First derivatives, for example, might well be of interest as indicators of local trends. Now if the values of the process over a neighborhood at $t=t_{m+1}$ were simultaneously estimated these derivatives could simply be calculated. But our method does not permit simultaneous inference. Moreover, the joint distribution of the derivatives and values of the process, if required, would not be given by this approach.

As we have formulated the growth curve problem, estimates of the process derivatives are not required, and this leads us to choose $p=0$. We believe the underlying growth processes to be quite smooth, at least thrice differential, so that we can view our inferential procedures as approximating the Bayes linear procedures which would obtain from imposing a global stochastic model on the underlying processes.

In general, there is a second potential reason for including derivatives in the model of equation (1.2). This is the need to bring in available prior knowledge about these derivatives which would otherwise be lost in the approximation of the residual covariance. In the present example, for instance, the growth functions must have non-negative derivatives and we should have included this knowledge through an appropriate prior distribution. Since we have chosen $p=0$, partly for simplicity and partly in keeping with our desire to challenge the proposed method under less than ideal conditions of implementation, our analysis can undoubtedly be improved upon. A more realistic approach to growth curve analysis is the subject of current work.

Let us make the realistic choice in this situation of $\sigma = 0$, even though this assumption of no noise causes us to violate the assumption which helps to justify the approximation of Section 3. It is reassuring that nonetheless our analysis leads to results which are in good agreement with those of Fearn (1976).

To specify the joint posterior density function given in (3.3) we need to specify a prior distribution for θ. It is convenient to let $\omega = \sigma_\beta^{-2}\Lambda$. Our lack of knowledge about θ suggests adopting a vague prior distribution to describe our uncertainty about it. Because of the complicated structure of the covariance, we choose the Jeffrey's prior computed by Weerahandi and Zidek (1986) to give an operational interpretation of the notion of "vague". But we recognize there is some arbitrariness about this choice, which has the improper density function, $\sigma_\beta^{-4}\pi_1(\omega)$, with respect to $d\sigma_\beta^2 d\omega$, where $\pi_1(\omega) = (1 + \omega v)^{-1}$ and v denotes the common value of $v_i = a_i^T c_i^{-1} a_i$ in this special case (see Weerahandi and Zidek, ibid).

We may factor σ_β^2 out of a number of objects like D_0 and leave behind a factor which depends on θ only through ω; it will be useful to designate such resulting factors with a single asterisk. Thus, $D_0 = \sigma_\beta^2 D_0^*$ for example, where $D_0^* = (e^*)^{-1} + (e^*)^{-1}A_2(F^*)^{-1}A_2^T(e^*)^{-1}$, $e^* = Diag\{e_1^*,...,e_{(m+1)}^*\}$, $e_i^* = v_i^* + \omega^{-1}$ and $v_i^* = a_i^T D_i^{-1} a_i$.

With the help of this new notation, we have, after incorporating the σ_β^2 from the change of variables, $d\Lambda = \sigma_\beta^2 d\omega$,

$$\pi(\beta_{(m+1)},\omega,\sigma_\beta^2|Y) = KF(\omega)\ (\sigma_\beta^2)^{-(n+2)/2}\exp[-(\sigma_\beta^{-2}T^*)/2]\ , \tag{4.1}$$

where $F(\omega) = \pi_1(\omega)(\Sigma^*_{(m+1)})^{-1/2}\omega^{-m/2}(D_0^*)^{1/2}$, $\Sigma^*_{(m+1)} = (m+1)\omega/m + (nv^*)^{-1}$, when v^* is used to represent the common value of the v_i^*'s, $T^* = (\beta_{(m+1)} - \beta^*_{(m+1)})^2\ (\Sigma^*_{(m+1)})^{-1} + Y^T D_2^* Y$, and it will be recalled that here $\beta^*_{(m+1)} = m^{-1}\sum\hat{\beta}_i$ while $D_2^* = D^{-1} - D^{-1}A^T(D_0^*)AD^{-1}$. By integrating out σ_β^2 from equation (4.1) we deduce that

$$\frac{\beta_{(m+1)} - \beta^*_{(m+1)}}{\sigma_{(m+1)}} \mid Y \sim t_{n-1} \tag{4.2}$$

and is distributed independently of ω with density

$$\pi(\omega|Y) = \hat{K}F(\omega)(Y^T D_2^* Y)^{-n/2}\ , \tag{4.3}$$

where $\hat{K}$ is the normalization constant, $n = \sum n_i$, t_{n-1} is the Student's t distribution with $n-1$ degrees of freedom and $\sigma_{(m+1)} = (\ \Sigma^*_{(m+1)}Y^T D_2^* Y/(n-1)\ )^{1/2}$. The inferences on $\beta_{(m+1)}$ can be based on (4.2) and (4.3) as illustrated below.

Numerical Illustration

Returning to the specific application on Ramus heights of male children described in the first paragraph of this section, we now illustrate how the foregoing theory can be used to carry out inferences on the Ramus height, say $H=\beta_{m+1}$, of a representative child who may or may not have been sampled yet. In Figure 1, the Ramus height data of each boy are plotted against his ages at the times the heights were measured; the data points of the 20 boys are labeled A, B,... and ,T.

Shown in the same figure are the estimated values of the mean growth curve, $\hat{H}(t)=\sum\hat{\beta}_i(t)/20$ and the 95% point-wise estimated Bayesian credibility band of $H(t)$. In estimating the mean growth curve, $\hat{\beta}_i(t_{m+1})$ was computed for a range of values of $t=t_{(m+1)}$ using the formula given in equation (3.2). The 95% point-wise credibility band for H was also obtained by varying $t=t_{(m+1)}$ using (4.2) in conjunction with the posterior distribution of ω given by (4.3) and the vague prior for ω described above. The analysis entails solving, for each t with the help of numerical integration, the appropriate equation to determine the 97.5 th percentile of the Student's t distribution with 79 degrees of freedom, but this is straightforward. The details may be found in Weerahandi and Zidek (1986).

It is of interest that, although we have not assumed a parametric model, the estimated normal growth curve of Ramus heights was found to be slightly concave but almost linear over the sampled range of ages. Hence the model assumed by Fearn

(1975) seems reasonable; the inherent technical difficulties of his approach make the construction of Bayesian credibility intervals exceedingly difficult. An advantage of our nonparametric approach lies in its capability to handle even highly nonlinear growth curves such as those treated by Berkey (1982) without needing to identify appropriate parametric models.

5. CONCLUDING REMARKS.

Bayesians have not contributed to the theory of nonparametric regression and smoothing for continuous parameter processes to anything like the same degree as frequentists. This remark ignores the extensive literature on the classical theory of time stationary time series and Kriging both of which, it could well be argued, only have meaning in a Bayesian framework. However, neither is usually considered a Bayesian theory. A paper much more in the character of this one is that of O'Hagen (1979); but unlike O'Hagen, we are not concerned with adaptive modeling and our focus is concerned with the implications of local smoothness for inference at a fixed point. The recent manuscript of O'Hagen (1989) on numerical quadrature as well as that of Angers and Delampady (1990) on smoothing entail Bayesian approaches to continuous parameter processes, although the approaches taken are quite different to that of this paper. Lack of space prevents us from giving a detailed survey; instead we refer the interested reader to the recent review of Sacks, Welch, Mitchell, and Wynn (1989).

There is an immense and rapidly growing list of repeated sampling school contributions to the literature of smoothing and nonparametric regression. Recently some work has been published on the topic of this paper, smoothing for multiple time series. References may be found in the forthcoming paper of Fraiman and Iribarren (1991) who mix a nonrandom population mean function with random sample paths for individuals, $i = 1,...,m$ which deviate from the population mean function by a zero mean, autocorrelated stochastic processes. Inference is simultaneously about the population mean function and its first derivative, as both the number of individuals, m and the number of data per individual , $n_i, i = 1,..., m$ approach infinity. The n_i are supposed to be identical as are the equispaced sampling points for all individuals. The latter become increasingly dense as the amount of data increases; interest focuses on consistency and the asymptotic distributions for a wide class of linear inferential procedures all of which are locally weighted means whose weights become increasingly concentrated at the point, t_* of inferential interest. The same locally weightings are used for each individual and the linear inferential procedure averages across individuals with equal weights, as is justified by the method of Section 3 for balanced sampling at equi-spaced sampling points (but not otherwise).

We have general concerns about the validity of repeated sampling theory. And we have specific concerns about the value of the consistency criterion which cannot rule out obviously inefficient procedures. For example, if it were known that the population mean function were similar at several widely separated parameter points, the local averages from data windows at these points should obviously be combined in some way. But single window procedures would nevertheless be consistent.

The very elegant and general theory of Fraiman and Iribarren (1991) embraces a large family of potential linear procedures, as the local weightings are required to satisfy only certain rather weak conditions. The virtue of generality can be a shortcoming when it is confronted with a finite sample size in that the theory does not point to a good choice of weights. However, point estimation in this context seems rather well understood so this criticism may be misplaced. A more significant problem is that of providing reasonably good indicators such as residual covariances and reliability (confidence or credibility) intervals, of the accuracy of point estimances. This topic has not been much addressed in the theory of nonparametric regression and smoothing.

We have briefly reviewed the paper of Fraiman and Iribarren (1991) to give the flavor of the repeated sampling school theory for this problem. The issues we have identified, among others, have lead us to the approach of the present paper.

Bayes linear procedures (see equation (1.1)), the objects of interest in this paper, are often derived from linear models like that in equation (1.2) to enable hierarchical modeling. That is the case here where inference is about the value at a particular time of a future sample path. Observations with noise from each of several related sample paths are deemed to be available. The model in equation (1.7) is convenient for incorporating the information that the processes are at least p times differentiable. We have justified this model as an approximation to that in (1.2) for every member of a class, B, of models that express the knowledge that the processes are p times differentiable. It is argued that the approximation will be good if these processes are actually $P \geq 2p+2$ times differentiable and the data are sampled with sufficient intensity near the point of interest. "Good" means that the resulting Bayes linear procedure derived from the model in (1.7) will be in good agreement with that derived from (1.2) for every member of the class B. It is also argued that the residual covariance matrices obtained from (1.2) and (1.7) will likewise be in close agreement under these circumstances. This justification is analogous to consistency in the frequency setting.

Our proposed method has a number of potential competitors including kernel, locally weighted regression (LWR) and spline methods. In fact Cleveland (1988) cites our method as a Bayesian version of the LWR method. And it shares the simplicity of the latter; the model in equation (1.7) is just the conventional linear model of regression analysis and so susceptible to analysis by the host of methods and computational software which have been developed for such models.

But as this paper attempts to show, the potential domain of application of the method is extremely broad and so it may enjoy some advantages over its competitors. The theory in Section 2 addresses data from a fairly complex sampling plan in a seemingly natural way. As Zidek and Weerahandi (1990) show, vector valued processes can readily be accommodated. And in current work with Dr M. Delampady and Ms. Irene Yee, likelihood function estimators for time series are developed from the same starting point. The work of Joe, Ma and Zidek (1986) based on the proposed Bayesian approach suggests a computationally cheap alternative to cross validation. In fact, we view this paper as illustrating a general Bayesian framework for addressing problems of interpolating and extrapolating locally smooth functions, rather than as merely providing an additional method for tackling them.

When the processes which generate the data have P derivatives with $p \leq P \leq 2p+2$, the justification of this paper for the proposed method is lost. An appropriate approximation is unclear. Suppose (i) $m=1$ and inference is about values of the single S process; (ii) $P=p$ (iii) the process is weakly stationary; and (iv) the residuals to the left and right of $t=t_{m+1}$ from fitting the model in (1.7), are uncorrelated. Then it can be shown that the process is AR(p) and the resulting residual correlations may be taken to be approximately zero. This argument lead Weerahandi and Zidek (1988) to the approximation they chose, one which corresponds to that of Section 3. Clearly as $P \geq p$ increases and more of the process derivatives are buried in the model's residual term, so intuitively it seems that the residual correlation must increase. Therefore it would seem to be desirable to choose p as large as possible. But the price is increased modeling complexity and given the increased likelihood of prior model misspecification, it is not clear the tradeoff is worthwhile. Clearly this is a matter for further study.

The approach underlying the analysis of this paper can be applied without the justification described above provided the model expresses well the investigator's prior views. In particular it can be applied even when the data are not dense around the point of inference. In such situations values inferred from automatic procedures like spline fitting, become mathematical artifacts of their definitions. Indeed, such procedures can have other unforeseen properties. A simultaneous frequency property called "intriguing" (cf. Nychka (1988)) is observed by Wahba (1983) for point-wise intervals generated by a Bayesian approach. Nychka (1988) seeks to "remove some of the mystery" by interchanging the randomness in the function (here the S's) assumed by Wahba (1983) with the determinism of the observation times (here the t_{ij} 's) so that the latter are now endowed with stochastic uncertainty.

We believe the properties of the method proposed here may be more predictable. In particular, the investigator can ensure that values inferred from the proposed method are in accord with prior experience. However, accomplishing this would require that the investigator use an informative prior rather than that adopted in Section 3 and in

the illustrative example of Section 4.

We will now summarize some of the concerns and open questions that remain. Given its fundamental importance, a better understanding of the extended semi-Markov property of Section 2 is required. This is needed to justify all inferential procedures, including those of this paper, which make local inferences either by local weighting or by relying only on data windows.

We need a better understanding of the role of the components of local curvature, like $\Gamma^{(p+1)(p+1)\beta}_{\alpha}$. These may be important for incorporating data at moderate distance from the point at which inference is being made, even when the extended semi Markov property holds. For simplicity this quantity was ignored in our analysis, albeit with some justification as provided in the Appendix.

The hierarchical development of priors through linear models like that in equation (1.2) is important. But it is not clear how well the resulting marginal distribution is approximated through the adoption of the simpler model in (1.7) instead. Presumably the answer depends on the sampling intensity around the point of inference.

REFERENCES

Angers, J.F., and Delampady, M. (1990). Hierarchical Bayesian estimation of a function. Sims Tech. Rep. No. 144, Department of Statistics, University of British Columbia.

Berkey, C.S. (1982). Bayesian Approach for a nonlinear growth model. *Biometrics*, 38, 953-961.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

Cleveland, W.S. and Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596-610.

Fraiman, R. and Iribarren, G.P. (1991). Nonparametric regression estimation in models with weak error dependence. *Journal of Multivariate Analysis*, To appear.

Heckman, N.E. (1988). Minimax estimates in a semiparametric model. *Journal of the American Statistical Association*, 83, 1090-1096.

Joe, H., Ma, Wilson, and Zidek, J.V. (1986). A Bayesian nonparametric univariate smoothing method with applications to acid rain data analysis. SIMS Technical Report No. 47, Department of Statistics, University of British Columbia.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes Estimates for the linear model(with Discussion). *Journal of the Royal Statistical Society*, B, 34, 1-41.

Muller, Hans-Georg (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association*, 82, 231-238.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association*, 83, 1134-1143.

O'Hagen, A. (1978). Curve fitting and optimal design for prediction (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 40, 1-42.

O'Hagen, A. (1989) Bayesian quadrature. Tech. Rep., Department of Statistics, University of Warwick

Sacks, J. and Ylvisaker, D. (1978). Linear estimation for approximately linear models. *Annals of Statistics*, 6, 1122-1137.

Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, 4, 409-435.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross validated smoothing spline. *Journal of the Royal Statistical Society*, B, 45, 133-150.

Weerahandi, S. and Zidek, J.V. (1986). Analyses of multiple time series by Bayesian and empirical Bayesian nonparametric methods. SIMS Tech. Rep. No. 96, Department of Statistics, University of British Columbia.

Weerahandi, S. and Zidek, J.V. (1988). Bayesian nonparametric smoothers for regular processes. *The Canadian Journal of Statistics*, 16, 1, 61-73.

Zidek, J.V., Weerahandi, S. (1990). Approximate Bayes linear smoothers for continuous processes. Submitted.

APPENDIX

An argument in support of the approximate residual covariance matrix adopted in Section 3 will now be given.

In the notation of Section 2, suppose $\gamma_B = \Gamma_{kk}^{(p+1,p+1\,\beta)}$ for all $i \neq k$ while $\gamma_W = \Gamma_{ii}^{(p+1,p+1\,\beta)}$ for all i. Let $\gamma_d = \gamma_W - \gamma_B$. Then the approximate residual covariance matrix in equation (2.9) becomes

$$C = C_1 + C_2 \tag{A.1}$$

where $C_1 = Diag\ \{c_{11}, \ldots, c_{1m}\}$ and $C_2 = \gamma_B \delta\delta^T$ and for all $i=1,\ldots,m$, $j=1,\ldots,n_i$,

$$c_{1i} = d_i + \gamma_d \delta_i \delta_i^T \ ,$$

$$\delta^T = (\delta_1^T, \ldots, \delta_m^T) \ ,$$

$$d_i = \Sigma_i + \zeta_i \ ,$$

$$\Sigma_i = Diag\ \{\Sigma_{(1)(1)}, \ldots, \Sigma_{(n_i)(n_i)}\} \ \delta_i = (\delta_i, \ldots, \delta_{in_i}) \ ,$$

$$\delta_{ij} = \Delta_{ij} \exp\{-|\Delta_{ij}|/\Delta_0\}/(p+1)! \ ,$$

$$\zeta_i = 2Diag\ \{\zeta_{i1}, \ldots, \zeta_{in_i}\}$$

and

$$\zeta_{ij} = 2|\Delta_{ij}|^{2p+3}\sigma_{ij} \ .$$

From the well-known matrix identity,

$$(x+y^T z y)^{-1} = x^{-1} - x^{-1}y^T(z^{-1} + yx^{-1}y^T)yx^{-1} \ , \tag{A.2}$$

$$C^{-1} = C_1^{-1} - K_B C_1^{-1}\delta\delta^T C_1 \ , \tag{A.3}$$

where $K_B = (\gamma_B^{-1} + \delta^T C_1^{-1}\delta)^{-1}$. Now let $d_i = Diag\ \{d_{i1}, \ldots, d_{in_i}\}$. Then

$$C_1^{-1} = Diag\ \{c_{11}^{-1}, \ldots, c_{1m}^{-1}\} \tag{A.4}$$

where from equation (A.2),

$$c_{1i}^{-1} = d_i^{-1} - K_i d_i^{-1}\delta_i\delta_i^T d_i^{-1} \tag{A.5}$$

with $K_i = \gamma_d + \delta_i d_i \delta_i$.

From equation (A.3) it follows that, with A defined in equation (1.7) and its superscript, "0", deleted,

$$A^T C^{-1} A = A^T C_1^{-1} A - K_B A^T C_1^{-1}\delta\delta^T C^{-1} A \ . \tag{A.6}$$

From (A.4)

$$A^T C_1^{-1} A = Diag\ \{a_1^T c_{11}^{-1} a_1, \ldots, a_m^T c_{1m}^{-1} a_m, 0\}$$

where $a_i = (a_{i1}^T, \ldots, a_{in_i}^T)^T : n_i \times (p+1)$, $i=1,\ldots,m$. And

$$A^T C_1^{-1} \delta = [(a_1^T c_{11}^{-1} \delta_1)^T, \ldots, (a_m^T c_{1m}^{-1} \delta_m)^T, 0]^T .$$

Now from (A.5) we obtain

$$a_i^T c_{1i}^{-1} a_i = a_i^T d_i^{-1} a_i - K_i a_i^T d_i^{-1} \delta_i \delta_i^T d_i^{-1} a_i . \tag{A.7}$$

and

$$a_i^T c_{1i}^{-1} \delta_i = a_i^T d_i^{-1} \delta_i - K a_i^T d_i^{-1} \delta_i \delta_i^T d_i^{-1} \delta_i . \tag{A.8}$$

Observe that

$$a_i^T d_i^{-1} \delta_i = \sum_j a_{ij}^T \delta_{ij} / (\Sigma_{ij} + \zeta_{ij}) . \tag{A.9}$$

But for large and small $|\Delta_{ij}|$, the summand in equation (A.9) is approximately zero because $\zeta_{ij} \propto |\Delta_{ij}|^{2p+3}$, and $\delta_{ij} \propto \Delta_{ij}^{p+1} \exp[-|\Delta_{ij}|/\Delta_0]$, respectively, become dominant. At most, the summands corresponding to moderate values of $|\Delta_{ij}|$ will contribute to equation (A.9). To simplify the model and the problem of specifying the prior parameters, it seems reasonable therefore to drop the second term in equation (A.7) at least if the $\Sigma_{ij} > 0$ are not unduly small.

Returning to (A.6), observe that

$$[A^T C_1^{-1} \delta]^T = [(a_1^T c_{11}^{-1} \delta_1)^T, \ldots, (a_m^T c_{1m}^{-1} \delta_m)^T, 0] .$$

which by the reasoning of the last paragraph might reasonably be approximated by a matrix of 0 's. In summary,

$$A^T C^{-1} A = A^T (\Sigma + \zeta)^{-1} A$$

is a plausible approximation. This corresponds to taking $\gamma_B = \gamma_W = 0$, that is, to ignoring the middle term in the approximation.

Analogous reasoning for $A^T C^{-1} Y$ yields a similar conclusion. This leads to the approximation in Section 3.

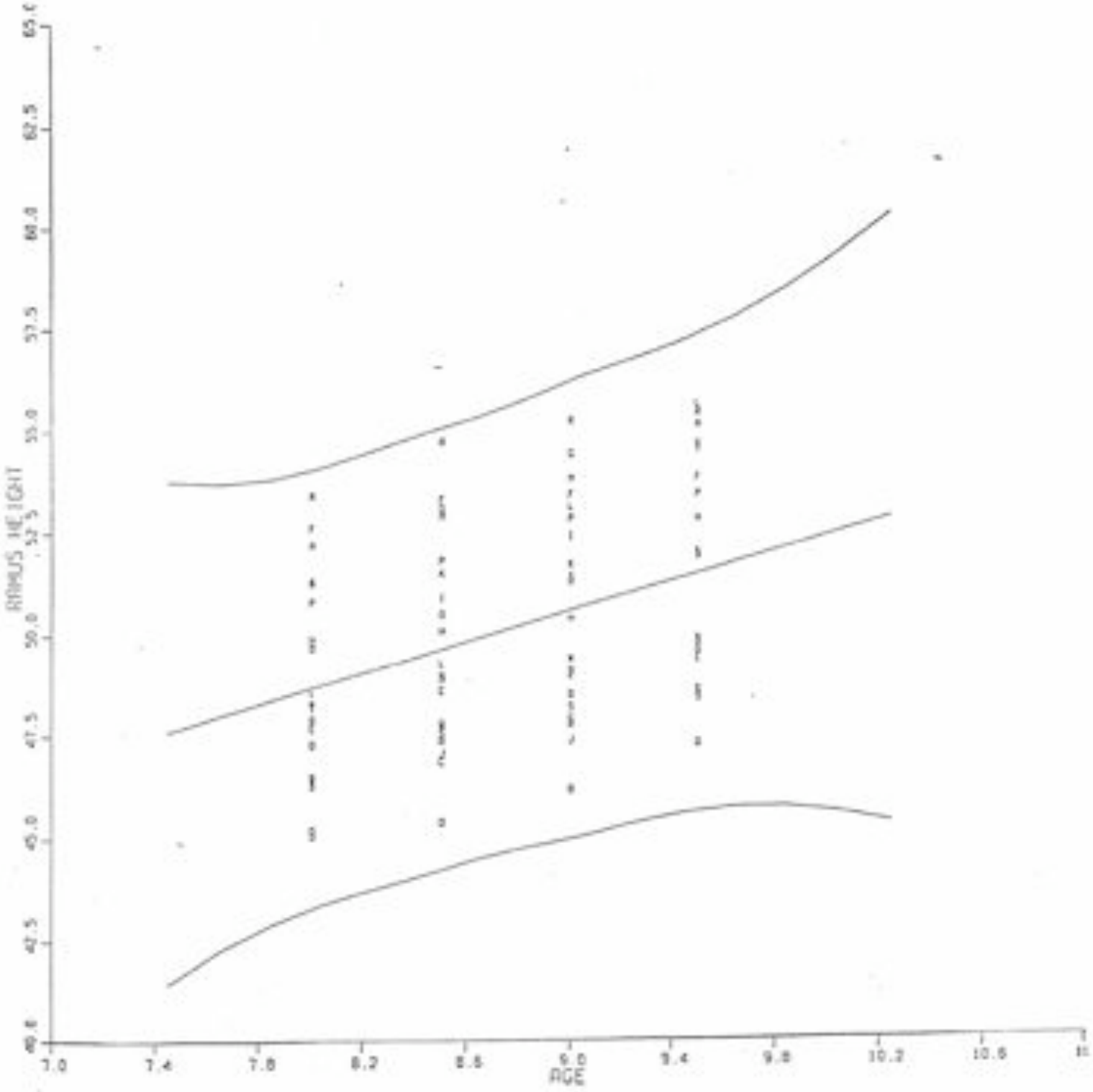

Figure 3.1. Normal growth curve of ramus heights and confidence band.