Partial Informative Normal and Spline Smoothing

Paul Speckman and Dongchu Sun

University of Missouri - Columbia

June 16, 2003
Outline

- Spline smoothing
- Bayesian interpretation
- Partial informative normal
- Fully Bayesian analysis and properties
- Additive models
- Adaptive spline smoothing
Spline Smoothing

Consider 1-d nonpara. regression with unknown $g$,

$$y_i = g(t_i) + \varepsilon_i, \; \varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \delta_0). \quad (1)$$

Estimate $g$ from solving the minimization problem,

$$\min_g \left[ \sum_{i=1}^{n} \{y_i - g(t_i)\}^2 + \eta \int_0^1 g^{(p)}(t)^2 dt \right]. \quad (2)$$

The general theory can be described in terms of a reproducing kernal Hilbert space $\mathcal{H}$ on a domain $\mathcal{X}$ with an inner product such that point evaluation is a bounded linear functional.

– See Wahba (1990) or Gu (2002).
**Fact:** There is a symmetric function $R(s, t)$, such that $R(s, \cdot) \in \mathcal{H}$, $\forall s \in \mathcal{X}$, and 

$$g(s) = \langle g, R(s, \cdot) \rangle, \quad \forall g \in \mathcal{H}.$$ 

Example of **polynomial splines**, let $R = R_0 + R_1$, $R_0$ ia a kernel associated with the $p$ dim. space of polynomials of degree less than $p$, and $R_1$ is a kernel associated with $J(g) = \int_0^1 [g(t)^{(p)}]^2 \, dt$.

*e.g.*, $R_1(s, t) = \int_0^1 \frac{(s - u)^{p-1}}{(p - 1)!} \frac{(t - 1)^{p-1}}{(p - 1)!} \, du$.

It can be shown that the solution to (2) has the form

$$g(t) = \sum_{k=1}^{p} c_i \frac{t^{k-1}}{(k - 1)!} + \sum_{i=1}^{n} d_i R_1(t_i, t).$$
Moreover, \( J(g) = \sum_{i=1}^{n} \sum_{j=1}^{n} d_i d_j R(t_i, t_j) \).

This set-up generalizes. In a number of important cases, there is a seminorm \( J(g) \) on a function space over a domain \( \mathcal{X} \) with null space spanned by functions \( \phi_j(t), j = 1, \ldots , p \), and an associated kernel \( R_1(s, t) \) such that

\[
g(t) = \sum_{k=1}^{p} c_i \phi_k(t) + \sum_{i=1}^{n} d_i R_1(t_i, t).
\] (3)
Now let \( T = (T_{ij})_{n \times p}, \ T_{ij} = \phi_j(t_i) \) and let \( \Sigma \) be the \( n \times n \) matrix with elements \( R(t_i, t_j) \). Then the solution to (2) is obtained by minimizing

\[
(y - Td - \Sigma c)'(y - Td - \Sigma c) + \eta c'\Sigma c.
\]

Differentiation with respect to \( c \) and \( d \) shows that the solution satisfies

\[
\begin{cases}
\Sigma\{(\Sigma + \eta I)c + Td - y\} = 0, \\
T'(\Sigma c + Td - y) = 0.
\end{cases}
\]

Since the kernel \( R(s, t) \) is positive semidefinite, \( \Sigma \) is a covariance matrix.
Consider the prior

\[ g = T\theta + x, \]

where \( x \sim N_n(0, \delta_1 \Sigma) \) and \( \theta \sim N_p(0, \delta_2 I_p) \).

Let \( a = \delta_2 / \delta_1 \), \( \eta = \delta_0 / \delta_1 \), and \( G_a = a TT' + \Sigma \).

The posterior of \( z \) is normal with

\[
E(g \mid y, a, \eta) = \hat{g}_a = G_a (G_a + \eta I)^{-1} y, \\
\text{var}(g \mid y, a, \eta) = \delta_1 \{ G_a - G_a (G_a + \eta I)^{-1} G_a \},
\]

If \( \delta_2 \to \infty \) \((a \to \infty)\), \( \lim_{a \to \infty} \hat{g}_a = \hat{g}_\infty \) is exactly the solution to system (4).
The limit of this posterior distribution has a convenient characterization, at least for theoretical purposes. In fact, the limiting distribution is

$$\mathcal{N}((I + \eta A)^{-1} y, \delta_0 (I + \eta A)^{-1}).$$  \hspace{1cm} (5)

where $A$ is define by

$$\begin{cases} 
\Sigma^{-1} - \Sigma^{-1} T (T' \Sigma^{-1} T)^{-1} T' \Sigma^{-1} & \text{if } \Sigma > 0, \\
G_1^{-1} - G_1^{-1} T (T' G_1^{-1} T)^{-1} T' G_1^{-1} & \text{if } G_1 > 0.
\end{cases}$$

This can be derived from Wahba (1990) or Gu (2002), at least for $\Sigma > 0$. A concise derivation is given in Speckman and Sun (2003).

• Note that $\text{rank}(A) = n - p$. 
PIN (partially informative normal) and AR priors

A random vector $X \in \mathbb{R}^n$ has a **partially informative normal** distribution with parameters $(\mu, A)$, denoted by $PIN(\mu, A)$, if its density is

$$f(x) \propto |A|_{+}^{1/2} \exp \left\{ -\frac{1}{2}(x - \mu)'A(x - \mu) \right\},$$

where $\mu \in \mathbb{R}^n$ is location parameter, $A$ is a nonnegative definite matrix, $|A|_{+} = \text{the product of the positive eigenvalues of } A$ (included for use subsequent in hierarchical models.)

– If $A > 0$, it is multivariate normal with mean $\mu$ and precision matrix $A$. 

The interesting cases occur when $A$ is singular and the distribution is improper. It can be interpreted as the convolution of a constant prior on the null space of $A$ and a proper normal on the range of $A$. In practice, $A$ is often only known up to an unknown parameter $\delta_1$. When $\mu = 0$ and $A$ has rank $n - p$, the prior has improper density

$$f(z) \propto \delta_1^{-(n-p)/2}|A|^{1/2} \exp\left(-\frac{1}{2\delta_1}z'Az\right).$$

- The limit of priors pertaining to spline smoothing has exactly this form. In fact, we can show that as $a \to \infty$, the limit of the priors specified for $g$ is exactly PIN, with $A$ defined by (5).
Another way to look at PIN: CAR and IAR

It follows that

\[
    f(z_i \mid z_j, j \neq i) = \left( \frac{a_{ii}}{2\pi\delta_1} \right)^{\frac{1}{2}} \exp \left\{ -\frac{a_{ii}}{2\delta_1} \left( z_i - \sum_{j=1}^{n} c_{ij} z_j \right)^2 \right\},
\]

where \( c_{ii} = 0 \) and \( c_{ij} = -\frac{a_{ij}}{a_{ii}} \) for \( i \neq j \).

- This is the conditional autoregressive (CAR) process (Besag, 1974).
- When \( |A| = 0 \), it is an intrinsic autoregressive (IAR) prior.
Examples of PINs

Fahrmeir & Wagenpfeil (1996): 2nd order difference

\[ z_k = 2z_{k-1} - z_{k-2} + \xi_k, \quad \xi_k \overset{iid}{\sim} N(0, \delta_1), \text{ i.e.,} \]

\[ f(z) \propto \delta_1^{-\frac{n-2}{2}} \prod_{k=3}^{n} \phi \left( \frac{(z_k - 2z_{k-1} + z_{k-2})}{\sqrt{\delta_1}} \right), \]

where \( \phi \) is the density of \( N(0, 1) \). Fahrmeir & Lang (2001) generalized the prior to approximate divided differences for non-equally spaced points. For the equally-spaced case, the second order form above easily generalizes to a \( p \)-th order form,

\[ f(z) \propto \delta_1^{-\frac{n-p}{2}} \prod_{k=p+1}^{n} \phi \left( \sum_{j=0}^{p} \binom{p}{j} z_{k-j} / \sqrt{\delta_1} \right). \]
Fully Bayesian Hierarchical Models

- Need priors on variance components $\delta_0$ and $\delta_1$. Carter and Kohn (1994) chose a small $b_0 > 0$,

$$\pi(\delta_1, \delta_0) \propto \frac{1}{\delta_0} \exp\{ -b_0/\delta_0 \},$$


$$\pi_r(\delta_1, \delta_0) \approx \frac{1}{\delta_1^{1/2} \delta_0}.$$

- constant priors for $(\delta_0, \delta_1)$ or $(\sqrt{\delta_0}, \sqrt{\delta_1})$. 
• In general, we assume that $\delta_0$ and $\delta_1$ are \textit{apriori} independent inverse gamma type, with density

$$\pi_i(\delta) \propto \frac{1}{\delta_i^{a_i+1}} \exp\left(-\frac{b_i}{\delta_i}\right)$$

for some real constants $(a_i, b_i)$.

• Inverse gamma $(\epsilon, \epsilon)$ has been used.

• Computationally simple with MCMC.

• Properiety of joint posterior of $(g, \delta_0, \delta_1)$.

• Posterior is improper if $a_1 = b_1 = 0$. 
We consider asymptotic properties for a variety of estimates:

- Generalized Cross Validation
- Generalized Maximum Likelihood
- Type II ML estimates
- marginal posterior mode

Under the usual Sobolov space frequentist assumptions on the function to be estimated, consistency and asymptotic normality of the estimated smoothing parameter are proved.
Our results show that the asymptotic distribution of the marginal posterior mode of the smoothing parameter does not depend on the prior distributions for a general class priors. The relative rates of convergence of the smoothing parameter estimates agree with those previously obtained for cross-validated kernel density estimates.

— converges .... but slowly ......

Bayesian posterior mode of smoothing parameter $\eta = \delta_0 / \delta_1$ performs better than GCV, with a smaller asymptotic variance.
To estimate functions of several variables, additive models have long been a popular technique for dimension reduction. For example, consider

\[ y_i = g(t_{1i}, t_{2i}) + \varepsilon_i, \quad i = 1, \ldots, n. \]

A common assumption is that

\[ g(t_1, t_2) = \mu + g_1(t_1) + g_2(t_2), \]

where for identifiability \( g_j \) is orthogonal to the space of constants for all \( j \).

It is also possible (not on grid)

\[ g(t_1, t_2) = \mu + g_1(t_1, t_2) + g_2(t_1, t_2). \]
The problem of estimating $g$ can be approached in terms of penalized log likelihood again, and $g$ can be estimated by solving the minimization problem

$$
\min_{g_1,g_2} \left[ \sum_{i=1}^{n} \{y_i - g_1(t_{i1}) - g_2(t_{i2})\}^2 + \sum_{j=1}^{2} \eta_j \int \{g_j^{(p)}\}^2 \right]
$$

for suitable smoothing parameters $\eta_j$. Assuming the explanatory variables are not collinear, the method of backfitting described in Hastie & Tibshirani (1990) converges to a unique solution.
Recently, Bayesian formulations have appeared including Shively, Kohn, & Wood (1999), Hastie & Tibshirani (2000) and Wood, Kohn, Shively, and Jiang (2002).

**Question:** if $g_k \sim PIN(\delta_k^{-1} A_k)$, for fixed variance components $\delta_k$, what is the marginal prior for $g_1 + g_2$?
Theorem. Let $Z_i$ be indep. $PIN_n(A_i)$, $\text{rank}(A_i) = n - p_i$. Let

$$A_i = (\Gamma_{i1}, \Gamma_{i2}) \text{diag}(0_{p_1}, \Lambda_i)(\Gamma_{i1}, \Gamma_{i2})',$$

$$\Sigma_i = (\Gamma_{i1}, \Gamma_{i2}) \text{diag}(I_{p_i}, \Lambda_i^{-1})(\Gamma_{i1}, \Gamma_{i2})',$$

$T = \text{the base of column space of } (\Gamma_{11}, \Gamma_{21}),$

$\Sigma = \Sigma_1 + \Sigma_2,$

$A = \Sigma^{-1} - \Sigma^{-1}T(T'\Sigma^{-1}T)^{-1}T'\Sigma^{-1}.$

Then $Z_1 + Z_2 \sim PIN_n(0, A).$
The motivation is to adaptively model the precision parameter in the discretized prior for smoothing. For example, suppose \( z_k = 2z_{k-1} - z_{k-2} + \xi_k \), where \( \xi_k \overset{\text{iid}}{\sim} N(0, \delta_k) \). Heuristically,

- for flat regions of \( g \), need small innovations \( \xi_k \), so the variance \( \delta_k \) should be small.

- if \( g \) has comparatively sharp features, \( \delta_k \) should be large.

- Adaptively modeling \( \delta_k \)’s \( \iff \) variable bandwidth selection in density and function estimation.
Frequentist literature on adaptive estimation

- Pintore, Speckman, and Holmes (2003) considered spatially adaptive smoothing splines and have successfully applied piecewise constant smoothing parameters.

Bayesian autoregressive conditional heteroscedasticity (ARCH), GARCH and stochastic volatility models.

- Vrontos et al. (2000)
- Nakatsuma (2000)
Bayesian Adaptive Smoothing

The priors considered here are designed to provide a Bayesian alternative to frequentist methods.

**Stage 1:** \( f(z_1, \ldots, z_n \mid \tau_{p+1}, \ldots, \tau_n) \)

\[
\propto \prod_{k=p+1}^{n} \delta_k^{-\frac{1}{2}} \phi \left( \sum_{j=0}^{p} \binom{p}{j} z_{k-j} / \sqrt{\delta_k} \right).
\]

**Stage 2:** \( q^{th} \) difference priors on \( \gamma_k = \log(\delta_k) \),

\[
f(\gamma_{p+1}, \ldots, \gamma_n \mid \xi)
= \xi^{-\frac{n-p-q}{2}} \prod_{k=p+q+1}^{n} \phi \left( \sum_{j=0}^{q} \binom{q}{j} \gamma_{k-j} / \sqrt{\xi} \right).
\]
A Fully Bayesian Adaptive Smoothing

- data: $y_i = g(t_i) + \varepsilon_i, \varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \delta_0)$.
- $z_i = g(t_i)$ follows the adaptive prior.
- prior for hyperparameter $\xi \sim$ inverse gamma.
- numerical examples: $p = 2, q = 1$. 
• Strong connection between spline smoothing and Bayesian analysis.

• IAR (CAR) models with \( p > 1 \) are important in estimating functional relationships. Models with \( p = 1 \) such as those in image analysis and disease mapping are not appropriate for function estimation.

• Objective Bayesian analysis is important, especially in choosing priors for variance components.

• Proper posterior is necessary for MCMC.
- Bayesian posterior mode of smoothing parameter $\eta = \delta_0/\delta_1$ performs better than GCV, with a smaller asymptotic variance.

- MCMC in additive model is similar to backfitting in nonparametric regression. Bayes solution analysis gives convergence with data-driven choices for $\eta_1$ and $\eta_2$.

- Bayesian adaptive smoothing seems to be promising.