Smoothing Spline Estimation of Variance Functions

Anna LIU, Tiejun TONG and Yuedong WANG *

September 19, 2006

Abstract

This article considers spline smoothing of variance functions. We focus on selection of the smoothing parameters and develop three direct data-driven methods: unbiased risk (UBR), generalized approximate cross validation (GACV) and generalized maximum likelihood (GML). In addition to guaranteed convergence, simulations show that these direct methods perform better than existing indirect UBR, generalized cross validation (GCV) and GML methods. The direct UBR and GML methods perform better than the GACV method. An application to array-based comparative genomic hybridization data illustrates the usefulness of the proposed methods.

KEY WORDS: array-based comparative genomic hybridization; generalized approximate cross validation; generalized maximum likelihood; heteroscedasticity; smoothing parameter; unbiased risk.

1. INTRODUCTION

Modeling local variability in terms of variance functions is an important problem with a wide range of applications. For example, variance function estimation is necessary in finance, quality control and immunoassay for measuring volatility or risk (Andersen and Lund 1997, Gallant and Tauchen 1997), experimental design (Box 1988), prediction (Carroll 1987, Yao and Tong 1994)

*Anna Liu is Assistant Professor, Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003 (email: anna@math.umass.edu); Tiejun Tong is Postdoctoral Associate, Department of Epidemiology and Public Health, Yale University, New Haven, CT 06520 (email: tiejun.tong@yale.edu); and Yuedong Wang is Professor, Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106 (email: yuedong@pstat.ucsb.edu). This research was supported by NIH Grant R01 GM58533. The authors thank the editor, the associate editor and two referees for their constructive comments and suggestions that have led to a substantial improvement in the article. Address for correspondence: Yuedong Wang, Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106.
and calibration (Raab 1981, Watters, Carroll and Spiegelman 1987). Variance estimation is especially important for detecting genes with differential expression across experimental conditions based on microarray data (Huang and Pan 2002, Wang and Guo 2004). With a small number of replicated array experiments, the standard estimates of variances are unreliable. Various methods have been proposed to improve estimation of the variances which usually lead to more powerful tests (Huang and Pan 2002, Wang and Guo 2004, Cui, Hwang, Qiu, Blades and Churchill 2005). We apply our variance function estimation methods to array-based comparative genomic hybridization (aCGH) data in Section 5. More applications of variance function estimation can be found in Carroll and Ruppert (1988). 

Research on nonparametric estimation of variance functions has attracted a great deal of attention (Carroll 1982, Silverman 1985, Hall and Carroll 1989, Ruppert, Wand, Holst and Hssjer 1997, Fan and Yao 1998, Yuan and Wahba 2004, Dai and Guo 2005). Most research concentrates on heteroscedastic regression. Both local polynomial smoothers and smoothing splines were used to model the variance function nonparametrically. Within the smoothing spline framework, Yuan and Wahba (2004) used the GACV method to select the smoothing parameter for estimating the variance function, while Dai and Guo (2005) treated squared pseudo-residuals (lag-one differences) as Gaussian data. The main goal of this paper is to develop and compare several data-driven smoothing parameter selection methods for the smoothing spline estimation of variance functions.

To simplify exposition, we focus on the situation when direct observations on a variance function are available. Specifically, we have independent observations \( \{(x_i, y_i), \ i = 1, \ldots, n\} \) where

\[
y_i = \exp(f(x_i))\chi^2_{i,k}/k,
\]

\( \chi^2_{i,k} \) are iid Chi-square random variables with \( k \) degrees of freedom. Our goal is to estimate the variance function \( f \) nonparametrically.

One typical situation leading to model (1) is the following heteroscedastic regression model with replicates

\[
z_{ij} = \mu_i + \exp(f(x_i)/2)\epsilon_{ij}, \quad i = 1, \ldots, n; \ j = 1, \ldots, k + 1,
\]

where \( \epsilon_{ij} \sim N(0,1) \). Then the sample variances \( y_i = \sum_{j=1}^{k+1}(z_{ij} - \sum_{j=1}^{k+1} z_{ij}/(k + 1))^2/k \) follow model (1). See Section 5 and Raab (1981) for real examples of model (2).

conditions, these squared residuals and pseudo-residuals follow model (1) asymptotically (Fan and Yao 1998, Dai and Guo 2005) with \( k = 1 \). Therefore, results in this article also shed light on the selection of smoothing parameters for these methods where \( k \) is small.

We introduce smoothing spline models and an estimation procedure in Section 2. In Section 3, we review some existing indirect methods and propose three direct methods for estimating the smoothing parameters. We report results of extensive simulations in Section 4 and apply our methods to a real aCGH data set in Section 5.

### 2. SMOOTHING SPLINE MODELS FOR VARIANCE FUNCTIONS

We assume model (1) in the remainder of this paper. Considering the Chi-square distribution as a special case of the Gamma distribution, we have \( y_i \sim \text{Gamma}(k/2, 2\exp(f(x_i))/k) \) with log-likelihood

\[
    l_i(f_i) = -ky_i\exp(-f_i)/2 - kf_i/2 + c(y_i),
\]

where \( f_i = f(x_i) \) and \( c(y_i) = \log(y_i^{k/2-1}(k/2)^{k/2}/\Gamma(k/2)) \), which is independent of \( f_i \). Spline smoothing methods for the exponential family in Wahba, Wang, Gu, Klein and Klein (1995) can be employed to estimate the function \( f \). Note, however, that instead of the canonical link (reciprocal) used in the previous publications, we use the logarithmic link in this paper in order to free the positive constraint on the variance function. We also propose new direct methods for selecting the smoothing parameters.

For simplicity, we assume that the domain of the function \( f \) is \([0, 1]\) and that \( f \) belongs to the reproducing kernel Hilbert space (RKHS)

\[
    W_m([0, 1]) = \left\{ f : f, f', \cdots, f^{(m-1)} \text{absolutely continuous, } \int_0^1 (f^{(m)})^2 dx < \infty \right\}. \tag{4}
\]

Our methods apply to general smoothing spline and smoothing spline ANOVA models (Wahba et al. 1995). The smoothing spline estimate of \( f \), \( f_\lambda \), is the minimizer of the penalized likelihood

\[
    -\sum_{i=1}^{n} l_i(f_i) + \frac{n\lambda}{2} \int_0^1 (f^{(m)})^2 dx, \tag{5}
\]

where \( \lambda \) is a smoothing parameter controlling the trade-off between the goodness-of-fit and the smoothness of the function. Let \( B_r \) be Bernoulli polynomials, \( \phi_r(x) = B_{r-1}(x)/(r-1)! \), \( r = 1, \cdots, m \), and \( R_1(s, t) = \phi_{m+1}(s)\phi_{m+1}(t) + (-1)^{m-1}\phi_{2m+1}(s-t) \). Then the solution to (5) is (Wahba et al. 1995, Gu 2002)

\[
    f_\lambda(x) = \sum_{i=1}^{m} d_i \phi_i(x) + \sum_{i=1}^{n} c_i R_1(x_i, x). \tag{6}
\]
For any fixed $\lambda$, coefficients $c = (c_1, \cdots, c_n)^T$ and $d = (d_1, \cdots, d_m)^T$ can be solved by the Newton procedure (Wahba et al. 1995). Let $u_i = -dl_i(f_i)/df_i = -ky_i\exp(-f_i)/2 + k/2$ and $w_i = -d^2l_i(f_i)/df_i^2 = ky_i\exp(-f_i)/2$. The Newton procedure updates $c$ and $d$ by iteratively reformulating the minimization problem (5) as

$$\sum_{i=1}^{n} w_{i-}(\tilde{y}_i - f_i)^2 + n\lambda \int_{0}^{1} (f^{(m)})^2 dx,$$

where $\tilde{y}_i = f_i - u_i/w_{i-}$ is the pseudo-data; the subscript $\text{minus}$ denotes its corresponding quantities evaluated at $f$ from the previous Newton iteration (Gu 1992, Wahba et al. 1995).

### 3. METHODS FOR SELECTING SMOOTHING PARAMETERS

A good choice of the smoothing parameter $\lambda$ is crucial to the performance of the smoothing spline estimate $f_\lambda$. One measure of the discrepancy between the spline estimate $f_\lambda$ and the true function $f$ is the Kullback-Leibler distance (Wahba et al. 1995)

$$KL(f, f_\lambda) = \frac{1}{n} \sum_{i=1}^{n} E_{f_i}(l_i(f_i) - l_i(f_\lambda)) = \frac{k}{2n} \sum_{i=1}^{n} (\exp(f_i - f_\lambda) + f_\lambda) - \frac{k}{2n} \sum_{i=1}^{n} (1 + f_i).$$

Ignoring the last term in (8), which is independent of $\lambda$, and also the multiplying constant $k/2$, we consider the comparative Kullback-Leibler criterion

$$CKL(f, f_\lambda) = \frac{1}{n} \sum_{i=1}^{n} (\exp(f_i - f_\lambda) + f_\lambda).$$

One approach is to find $\lambda$ such that $CKL(f, f_\lambda)$ is minimized. However, similar to other loss functions, $CKL(f, f_\lambda)$ cannot be minimized directly since $f$ is unknown. Therefore, we need to find estimates or proxies of the CKL criterion. We review some existing indirect methods in Section 3.1 and propose three direct data-driven methods in Sections 3.2-3.4.

#### 3.1 Indirect Methods

Note that (7) is the penalized weighted least squares for the working variables $\tilde{y}_i$ and working weights $w_{i-}$. An indirect (or iterative) method chooses a smoothing parameter at each iteration for the reformulated problem (7) and hopes it will converge (Gu 1992). To estimate $\lambda$ at each iteration, we may use the UBR, GCV or GML methods which respectively minimize the following UBR, GCV and GML scores

$$U_\ast(\lambda) = \frac{1}{n} \| (I - A(\lambda)) W_{-1}^{1/2} \tilde{y} \|^2 + 2\frac{\hat{\sigma}^2}{n} tr A(\lambda),$$

$$V_\ast(\lambda) = \frac{1}{n} \| (I - A(\lambda)) W_{1}^{1/2} \tilde{y} \|^2,$$

$$M_\ast(\lambda) = \frac{\tilde{y}^T W_{-1}^{1/2} (I - A(\lambda)) W_{-1}^{1/2} \tilde{y}}{[\det^+(W_{-1}^{1/2} (I - A(\lambda)) W_{-1}^{1/2})]^{1/(n-m)}}.$$
where $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T$, $W_\ast = \text{diag}(w_1, \ldots, w_n)$, $f_{\lambda i} = f_\lambda(x_i)$, $A(\lambda)$ satisfies $(w_{1/2}^{\lambda 1}, \ldots, w_{1/2}^{\lambda n})^T = A(\lambda)(w_{1/2}^{1/2}, \ldots, w_{1/2}^{1/2})^T$, $\hat{\sigma}^2 = \sum_{i=1}^n u_i^2/\mu_i$, and det$^+$ is the product of the nonzero eigenvalues (Wahba et al. 1995, Gu 2002).

The target criteria of these indirect methods change throughout the iterations. Therefore, one conceptual problem of the indirect methods is that their overall target criteria are not explicitly known. Asymptotically, $U_\ast(\lambda)$ provides a proxy of $KL(f, f_\lambda)$ (Wang, Wahba, Chappell and Gu 1995). No justifications have been provided for the indirect GCV and GML methods. One practical problem with the indirect methods is that they do not guarantee convergence. For Binomial and Poisson data, extensive simulations indicate that convergence is achieved for almost all situations (Wang et al. 1995). The performance of these indirect methods for Gamma data has not been studied. For Binomial and Poisson data, some direct methods have been developed and have been found to work better than the indirect methods (Xiang and Wahba 1996, Gu and Xiang 2001, Yuan 2005). However, direct methods for Gamma data have not been developed.

### 3.2 Unbiased Risk Method

In this subsection we derive an unbiased estimate of $E(CKL(f, f_\lambda))$. Let $h_\lambda(i, z, \cdot)$ be the minimizer of (5) when the $i$th observation, $y_i$, is replaced by $z$. Let $g_{\lambda i}(z) = h_\lambda(i, z, x_i)$ and $v(t) = \int_0^t \exp(-g_{\lambda i}(z))z^{k/2-1}dz$. Then for any fixed $y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)^T$, when $k \geq 3$, we have

$$E(\exp(f_i - f_{\lambda i})) = \frac{\exp(f_i)}{\Gamma(k/2)} \left( \frac{2 \exp(f_i)}{k} \right)^{-\frac{k}{2}} \int_0^\infty \exp(-g_{\lambda i}(t))t^{\frac{k}{2}-1} \exp \left( \frac{-kt}{2 \exp(f_i)} \right) dt$$

$$= \frac{\exp(f_i)}{\Gamma(k/2)} \left( \frac{2 \exp(f_i)}{k} \right)^{-\frac{k}{2}} \left\{ v(t) \exp \left( \frac{-kt}{2 \exp(f_i)} \right) \right\}^\infty_0 \int_0^\infty v(t) \exp \left( \frac{-kt}{2 \exp(f_i)} \right) dt$$

$$= \frac{k}{2 \Gamma(k/2)} \left( \frac{2 \exp(f_i)}{k} \right)^{-\frac{k}{2}} \int_0^\infty v(t) \exp \left( \frac{-kt}{2 \exp(f_i)} \right) dt$$

$$= \frac{k}{2} E \left( v(y_i)y_i^{-(k/2-1)} \right),$$

where we used the facts that $E(\exp(-f_{\lambda i}))$ exists and $v(t) \exp(-kt/2 \exp(f_i))|_0^\infty = 0$ (Appendix A). The derivation is similar to those in Berger (1980) and Wong (2006), where the above facts were assumed as conditions. Then, when $k \geq 3$, an unbiased estimator of $E(CKL(f, f_\lambda))$ is

$$UBR(\lambda) = \frac{1}{n} \sum_{i=1}^n \left( \frac{k}{2} v(y_i)y_i^{-(k/2-1)} + f_{\lambda i} \right).$$

The direct UBR estimate of $\lambda$ is the minimizer of $UBR(\lambda)$. Gaussian quadrature may be used to approximate $v(y_i) = \int_0^{y_i} \exp(-g_{\lambda i}(z))z^{k/2-1}dz$. However, it requires calculating
h_\lambda(i, z, \cdot) for several different values of z, which can be computationally intensive when the sample size is large. Simulations indicate that \exp(g_\lambda(z)) is approximately linear in z: \exp(g_\lambda(z)) \approx \beta_0 + \beta_1 z. We have tried several methods to compute the coefficients \beta_0 and \beta_1, and found that the following two work well: (i) compute \lambda_i(y_i) and \lambda_i(y_i/2), and then compute \beta_0 and \beta_1 as the intercept and slope of the line joining two points (y_i/2, \exp(g_\lambda(y_i/2))) and (y_i, \exp(g_\lambda(y_i))); (ii) from Taylor expansion \exp(g_\lambda(z)) \approx \exp(g_\lambda(y_i)) + \exp(g_\lambda(y_i)) (\partial g_\lambda/\partial z)|_{z=y_i} (z - y_i) \approx \exp(f_{\lambda_i}) + \exp(f_{\lambda_i}) d_{ii} (z - y_i), where (\partial g_\lambda/\partial z)|_{z=y_i} \approx d_{ii}; this is based on a similar argument as that in Appendix B, and \lambda_i is defined later in Section 3.3. Thus \beta_0 \approx \exp(f_{\lambda_i})(1 - d_{ii} y_i) and \beta_1 \approx \exp(f_{\lambda_i}) d_{ii}. Based on the linear approximation, \nu(y_i) can be calculated through recursive formulas obtained by lengthy algebra (not shown). Extensive simulations indicate that the above linear approximations lead to similar UBR estimates of the smoothing parameters \beta_0 and \beta_1 also lead to similar UBR estimates. We use the linear approximation based on the Taylor expansion in our simulations since it requires the least amount of computation.

3.3 Generalized Approximate Cross Validation Method

Let \tilde{f}_\lambda^{(-i)} be the minimizer of (5) without the ith observation and \tilde{f}_\lambda^{(-i)} = \tilde{f}_\lambda^{(-i)}(x_i). Replacing \exp(f_i - f_{\lambda_i}) by \exp(-f_{\lambda_i}^{(-i)}) and ignoring the multiplying constant 1/n, we obtain a cross-validation estimate of \text{C KL}(f, f_\lambda)

\[ CV(\lambda) = \sum_{i=1}^{n} \left( y_i \exp(-f_{\lambda_i}^{(-i)}) + f_{\lambda_i} \right). \] (14)

It is usually expensive to compute \text{CV}(\lambda) for large n. We now introduce an approximation of \text{CV}(\lambda). Let \mathbf{f = (f_1, \ldots, f_n)}^T, \mathbf{T}_{n \times m} = \{\phi_i(x_i)\}_{i=1}^{n} \{\phi_j(x_j)\}_{j=1}^{m} and \mathbf{\Sigma} = \{R_i(x_i, x_j)\}_{i,j=1}^{n}. Let \mathbf{T = (Q_1, Q_2)(R^T 0^T)^T} be the QR decomposition of \mathbf{T} and \mathbf{\Omega = Q_2Q_2^T}\Sigma Q_2^{\dagger}Q_2^T where ^\dagger is the Moore-Penrose generalized inverse. Then \mathbf{f = Td + \Sigma c}, \int_0^1 (f^{(m)})^2 dx = \mathbf{f^T \Omega f}, and the penalized likelihood (5) can be rewritten as (Xiang and Wahba 1996)

\[ J = -\sum_{i=1}^{n} l_i(f_i) + \frac{n\lambda}{2} \mathbf{f^T \Omega f}. \] (15)

Let \mathbf{W_\lambda = \text{diag}(k y_1 \exp(-f_{\lambda_1})/2, \ldots, k y_n \exp(-f_{\lambda_n})/2)}, \mathbf{V = \text{diag}(k \exp(-f_{\lambda_1})/2, \ldots, k \exp(-f_{\lambda_n})/2)}, and \mathbf{D = (W_\lambda + n\lambda \Omega)^{-1} V}. An approximation of \text{CV}(\lambda) is (Appendix B)

\[ ACV(\lambda) = L(\lambda) + \sum_{i=1}^{n} \frac{d_{ii} \exp(-f_{\lambda_i}) y_i (y_i - \exp(f_{\lambda_i}))}{1 - d_{ii} \exp(f_{\lambda_i})}, \]

where \mathbf{L(\lambda) = \sum_{i=1}^{n} (y_i \exp(-f_{\lambda_i}) + f_{\lambda_i}) and d_{ii} is the ith diagonal element of D}. Replacing \exp(f_{\lambda_i}) d_{ii} by tr(\mathbf{W_0^{1/2} D W_0^{1/2}})/n where \mathbf{W_0 = \text{diag}(\exp(f_{\lambda_1}), \ldots, \exp(f_{\lambda_n}))}, we obtain a gen-
Gaussian data, therefore, its performance is unknown. An approximation has not been explored as a tool for selecting the smoothing parameter for non-Gaussian data, and then fitting the transformed data using a penalized least square smoothing spline.

3.4 Generalized Maximum Likelihood Method

Let \( F(x) = \sum_{i=1}^{M} \theta_i \phi_i(x) + b^2 Z(x) \) be the prior for the function \( f \), where \( \theta_i \overset{iid}{\sim} N(0, a) \), \( b = 2/kn\lambda \), \( Z(x) \) is a Gaussian process independent of \( \theta_i \)'s with \( E(Z(x)) = 0 \) and \( E(Z(s)Z(t)) = R_1(s, t) \). We assume that observations \( y = (y_1, \ldots, y_n)^T \) are generated according to model (1) conditional on \( f = F \). As \( a \to \infty \), Gu (1992) showed that the posterior mean \( E(F(x)|y) \) approximately equals the spline estimate \( f_\lambda(x) \).

Let \( u_{ic} \) and \( w_{ic} \) be \( u_i \) and \( w_i \), respectively evaluated at the convergence point of the Newton procedure. Let \( u_c = (u_{1c}, \ldots, u_{nc})^T, W_c = \text{diag}(w_{1c}, \ldots, w_{nc}), f_\lambda = (f_{\lambda 1}, \ldots, f_{\lambda n})^T, y_c = f_\lambda - W_c^{-1}u_c, \Sigma_c = W_c^{1/2} \Sigma W_c^{1/2} \) and \( T_c = W_c^{1/2} T \). Let \( Q_{1c} Q_{2c} (R_c^T 0^T)^T \) be the QR decomposition of \( T_c \), and \( U \Lambda U^T \) be the spectral decomposition of \( Q_{2c}^T \Sigma_c Q_{2c} \) where \( \Lambda = \text{diag}(\lambda_{vn} \nu = 1, \ldots, n-m) \). Let \( z = (z_1, \ldots, z_{n-m})^T = U^T Q_{2c}^T W_c^{1/2} y_c \). Then, using equation (16) in Liu, Meiring and Wang (2005) and ignoring a constant, we have an approximation of the negative log marginal likelihood of \( y \),

\[
GML(\lambda) = -\sum_{i=1}^{n} l_i(f_{\lambda i}) - \frac{k}{4} u_c^T W_c^{-1} u_c + \frac{k}{2} \sum_{\nu=1}^{n-m} \left( \ln(\lambda_{vn}/n\lambda + 1) + \frac{k z_{\nu}^2/2}{\lambda_{vn}/n\lambda + 1} \right) + \ln |R_c|.
\]

The direct GML estimate of \( \lambda \) is the minimizer of \( GML(\lambda) \). Liu et al. (2005) used the approximated marginal likelihood to construct GML tests for generalized linear models. This approximation has not been explored as a tool for selecting the smoothing parameter for non-Gaussian data, therefore, its performance is unknown.

4. SIMULATIONS

One simple and commonly used approach is to transform \( y_i \) in model (1) using the logarithm scale and then fitting the transformed data using a penalized least square smoothing spline.
We refer to this simple approach as the transformation method. In the following simulation, we use the GCV method to select the smoothing parameter for the transformation approach.

We generate data from model \( (1) \) with \( x_i = i/n \). We use a factorial design with seven functions

\[
\begin{align*}
  f_1(x) &= 2 \sin(2\pi j x) + 3, \quad j = 1, 2, 3, \\
  f_4(x) &= 3 - (5x - 2.5)^2, \\
  f_5(x) &= \logit((-1.6x + .9) I\{x \leq .5\} + (1.6x - .7) I\{x > .5\}), \\
  f_6(x) &= \logit((3.5x/3) I\{x \leq .6\} + .7 I\{x > .6\}), \\
  f_7(x) &= 0.218 - 4.312x,
\end{align*}
\]

four different sample sizes \( n = 100, 200, 300, 400 \), and four different degrees of freedom \( k = 1, 2, 3, 4 \). Functions \( f_1 - f_7 \) are similar to those in Gu and Xiang (2001). We fit model \( (1) \) using eight methods: direct UBR, direct GML, GACV1, GACV2, indirect UBR, indirect GCV, indirect GML and transformation. For each simulation setting, we repeat the simulation 100 times.

Because UBR, GACV1, GACV2 and GML are all continuous functions of the smoothing parameter, we used the quasi-Newton method to optimize them. Specifically, the R function “optim” with the option “L-BFGS-B” is used. This option performs the quasi-Newton optimization and allows the user to specify lower and upper bounds for the smoothing parameter. We used -10 and 3 as the bounds for \( \log_{10}(n\lambda) \). The R code is available from the authors.

Contrary to our experience with the Binomial and Poisson data, the indirect methods sometimes fail to converge. Table 1 lists the number of replications in which the indirect UBR method failed to converge. The non-convergence problem of this method is quite severe for small \( k \). When \( k = 1 \), the non-convergence problem remains even when the sample size is as large as \( n = 400 \). The indirect GML and GCV failed to converge occasionally with frequency ranging from 1% to 4%. Convergence is achieved by all other methods.

Figure 1 contains typical curves of CKL, UBR, GACV2 and GML for \( f(x) = f_5(x) \), \( n = 100 \), and \( k = 1 \) and \( k = 2 \). Minimum points are marked by crosses. The same plots for other simulation settings are similar. This figure shows that the UBR function is a good estimate of the CKL criterion, and that the GACV1, GACV2 and GML functions are good proxies. Minima of these functions are usually quite close.

Figures 2 and 3 show boxplots of the CKL values for \( n = 200 \). The CKL values of the GACV1 and the transformation methods are not shown as they are both generally larger, and much larger for the transformation method, than those obtained from other methods. The
boxplots when $k = 4$ are not shown since they are similar to those when $k = 3$. The same plots for other sample sizes are similar. In general, the direct GML and UBR methods perform the best. The indirect UBR method suffers from the non-convergence problem, and when it does converge, it performs as well as the GML and UBR methods. Although the indirect GML and GCV converge most of the time, they tend to over-smooth which leads to larger CKL values. The GACV₁ and GACV₂ methods also tend to over-smooth occasionally. Even though the unbiasedness property is established for $k \geq 3$ only, we nevertheless applied the direct UBR method to the cases with $k = 1$ and $k = 2$. Performance of the direct UBR method for these cases is similar to that for larger $k$.

To take a closer look at the performance of various methods, we plot CKL’s of the indirect UBR, indirect GML and UBR vs CKL’s of the GML for $f(x) = f_6(x)$, $n = 100$, and $k = 1$ and $k = 3$ in Figure 4. The same plots for other simulation settings are similar. In general, the GML tends to have smaller CKL’s than the indirect GML; the UBR has similar CKL’s as the GML except for a few bad cases; and the indirect UBR has similar CKL’s as the GML when it converges.

Figure 5 shows the 5th, 50th and 95th best estimates ordered by CKL for the first four functions, with $n = 100$ and $\lambda$ selected by the GML method. The same plots for other simulation settings and with the UBR method are similar. We conclude that the smoothing spline estimates of the variance functions with smoothing parameters selected by the GML and UBR methods perform well. The GML method is new for non-Gaussian data. We postulate that it can be extended to Binomial and Poisson data, and can perform better than some existing methods.

5. APPLICATION

Comparative genomic hybridization (CGH) is a technique designed for detecting segmental genomic alterations. Recent advances in array-based CGH (aCGH) technology have enabled examination of chromosomal regions in unprecedented detail, revolutionizing our understanding of chromosome alterations associated with tumorigenesis and many developmental abnormalities (Albertson and Pinkel 2003, Mantripragada, Buckley, de Stahl and Dumanski 2004). The aCGH technology uses two differentially labeled test and reference DNAs which are cohybridized to cloned genomic fragments immobilized on glass slides. The hybridized DNAs are then detected in two different fluorochromes, and the significant deviation from unity in the ratios of the digitized intensity values is indicative of copy-number differences between the test
and reference genomes (Wang and Guo 2004). Accurate identification of amplified or deleted regions requires estimates of variances (Moore, Pallavicini, Cher and Gray 1997, Wang and Guo 2004). The number of replicated arrays is typically small, owing to the cost. Therefore, the standard sample variance estimates are unreliable and some simple improvements have been proposed (Moore et al. 1997, Tusher, Tibshirani and Chu 2001, Huang and Pan 2002, Wang and Guo 2004). Wang and Guo (2004) considered the variance as a function of the physical locations in the genome. They applied the lowess smoother to the logarithm of the sample variances with a fixed smoothing parameter. They have shown that using the smoothed variances in the t-test can lead to large gains in power. As we have shown in Section 4, the simple approach to smoothing the logarithm of the sample variances is less efficient when the number of replications is small. Also, the selection of the smoothing parameter in Wang and Guo (2004) is somewhat arbitrary.

To illustrate our new methods, we downloaded a well-known BAC array data set from the website http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html. The data result from an experiment aimed at measuring copy number changes for the cell strains (test samples) against normal male reference DNAs (references), which were co-hybridized on CGH arrays containing 2,460 BAC and P1 clones in triplicate (7,380 spots) (Snijders, Nowak, Segraves, Brown, Conroy, Hamilton, Hindle, Huey, Kimura, Law, Myambo, Palmer, Ylstra, Yue, Gray, Jain, Pinkel and Albertson 2001).

For each chromosome of a cell strain, we assume model (2) where $z_{ij}$ is the $j$th replicate of log$_2$ fluorescent ratio of clone $i$, $\mu_i$ is the mean fluorescent ratio, $x_i$ is the physical position along the chromosome, and $k = 2$. A non-zero $\mu_i$ corresponds to an alteration. Wang and Guo (2004) introduced a smoothed t-statistic with variances smoothed along the genome. They have shown that the smoothed t-statistic always improves the performance over the standard t-statistic. We focus on estimation of the variance function. For simplicity we transform the genome position $x_i$ into the interval [0, 1].

We fitted variance functions to all chromosomes of all cell strains. Figure 6 shows observations, fits and confidence intervals for chromosome 10 of the cell strain GM13330010220, and chromosomes 11 and 12 of the cell strain GM03134001218. For the lowess fits, as in Wang and Guo (2004), we used 30% of the data for smoothing at each position. For the smoothing spline fit, we used the direct GML method to select the smoothing parameter. The approximate Bayesian confidence intervals were computed using the methods in Wahba et al. (1995). Note that the lowess fits are based on the logarithm of the sample variances. Therefore, visually, they are in the center of the data in Figure 6 since variances are plotted on the logarithm scale. However, it is known that naive estimates based on the logarithm of the sample variances are biased downward (Cui et al. 2005, Tong and Wang n.d.). A bias correction factor which is greater than one should have been applied before fitting. The lowess estimates are more rough
since the smoothing span has been set at 30%, which is the somewhat arbitrary choice used in Wang and Guo (2004).

Some variances are relatively large and some variances are very small (Figure 6). Replacing variances by their smoothed estimates also reduces the effect of outliers and prevents clones with very small variances from dominating the results. Other approaches have been proposed to deal with very small variances (Tusher et al. 2001).

6. CONCLUSIONS

Nonparametric estimation of variance functions has been well-studied in the literature, while the related problem of selection of the smoothing parameters has received less attention. Within the smoothing spline framework, the existing indirect UBR, GCV and GML methods for selecting the smoothing parameter do not guarantee convergence. In this article we have proposed the direct UBR, GACV and GML methods and found that the direct UBR and GML methods have the best performance in general.

APPENDIX A

We show that when \( k \geq 3 \), \( (a) \ v(t) \exp(-kt/2 \exp(f_i)) \rceil_0^\infty = 0 \), and \( (b) \ E(\exp(-f_{\lambda i})) < \infty \).

For \( (a) \), we show that \( \lim_{t \to 0} v(t) = 0 \) and \( \lim_{t \to \infty} (v(t) \exp(-kt/2 \exp(f_i))) = 0 \). We use the following fact that for any \( a > 0 \) and \( x \),

\[
a \exp(-x) + x \geq 1 + \ln a.
\]  

(17)

The above inequality can be verified by finding the minimum of the left hand side as a function of \( x \).

Note that \( y_{-i} \) is fixed and the \( i \)th observation is replaced by \( z \). The penalized likelihood (5) is equivalent to

\[
I(f) = z \exp(-f_i) + f_i + \sum_{j \neq i} (y_j \exp(-f_j) + f_j) + \frac{n\lambda}{k} \int_0^1 (f^{(m)})^2 dx.
\]  

(18)

Note that \( h_{\lambda}(i, z, \cdot) \) is the minimizer of (18) and \( g_{\lambda i}(z) = h_{\lambda}(i, z, x_i) \).

We assume that the constant function, \( f_0(x) = 0 \), belongs to the model space. This is true when the model space is \( W_m([0, 1]) \). \( I(f_0) = z + \sum_{j \neq i} y_j \). Since \( h_{\lambda}(i, t, \cdot) \) is the minimizer of
$I(f)$, we have

$$z \exp(-g_{\lambda_i}(z)) + g_{\lambda_i}(z) \leq I(h_\lambda(i, z, \cdot)) - \sum_{j \neq i} (y_j \exp(-h_\lambda(i, z, x_j)) + h_\lambda(i, z, x_j))$$

$$\leq I(f_0) - \sum_{j \neq i} (y_j \exp(-h_\lambda(i, z, x_j)) + h_\lambda(i, z, x_j))$$

$$\leq z + \sum_{j \neq i} y_j - \sum_{j \neq i} (1 + \ln y_j), \quad (19)$$

where the last inequality is based on (17). Therefore, as $z \to 0$, $z \exp(-g_{\lambda_i}(z)) + g_{\lambda_i}(z)$ is bounded above.

We claim that for any power $p > 1$,

$$\lim_{z \to 0} z^p \exp(-g_{\lambda_i}(z)) = 0. \quad (20)$$

Otherwise, there exists some $p_1 > 1$ such that $\limsup_{z \to 0} z^{p_1} \exp(-g_{\lambda_i}(z)) > 0$. Let $p_2 = (p_1 + 1)/2$. Then $\limsup_{z \to 0} z^{p_2} \exp(-g_{\lambda_i}(z)) = \infty$. There exists a sequence $z_m \to 0$ such that $-g_{\lambda_i}(z_m) + p_2 \ln z_m \to \infty$. Also, $-g_{\lambda_i}(z_m) + \ln z_m \geq -g_{\lambda_i}(z_m) + p_2 \ln z_m \to \infty$. Therefore, for large enough $m$, we have $(-g_{\lambda_i}(z_m) + \ln z_m)/2 \geq 1/(p_2 - 1) > 0$. Thus, for large enough $m$,

$$z_m \exp(-g_{\lambda_i}(z_m)) + g_{\lambda_i}(z_m) = \exp(-g_{\lambda_i}(z_m) + \ln z_m) + g_{\lambda_i}(z_m)$$

$$\geq 1 - g_{\lambda_i}(z_m) + \ln z_m + (-g_{\lambda_i}(z_m) + \ln z_m)^2/2 + g_{\lambda_i}(z_m)$$

$$\geq \ln z_m + (-g_{\lambda_i}(z_m) + \ln z_m)/(p_2 - 1)$$

$$= (-g_{\lambda_i}(z_m) + p_2 \ln z_m)/(p_2 - 1).$$

The left hand side is bounded above, while the right hand side approaches $\infty$, a contradiction.

Taking $p = 1.1$ in (20), when $k \geq 3$, we have

$$\limsup_{t \to 0} v(t) = \limsup_{t \to 0} \int_0^t (z^{1.1} \exp(-g_{\lambda_i}(z))) z^{\frac{k}{2} - 1} dz = \limsup_{t \to 0} \int_0^t z^{\frac{k}{2} - 2.1} dz = 0. \quad (21)$$

Therefore, $\limsup v(t) = 0$ since $v(t) \geq 0$.

As $z \to \infty$, from (19), we have $z \exp(-g_{\lambda_i}(z)) + g_{\lambda_i}(z) \leq 2z$. On the other hand, from (17) with $a = z/2$, $-g_{\lambda_i}(z) \leq z \exp(-g_{\lambda_i}(z))/2$ when $z \geq 2$. Therefore, $z \exp(-g_{\lambda_i}(z)) \leq 2z - g_{\lambda_i}(z) \leq 2z + z \exp(-g_{\lambda_i}(z))/2$. There exists some large $T$ such that

$$\exp(-g_{\lambda_i}(z)) \leq 4, \quad z \geq T. \quad (22)$$

Now,

$$v(t) = \int_0^T \exp(-g_{\lambda_i}(z)) z^{\frac{k}{2} - 1} dz + \int_T^t \exp(-g_{\lambda_i}(z)) z^{\frac{k}{2} - 1} dz \leq v(T) + \frac{8}{k} t^{\frac{k}{2}}.$$
Then
\[
\limsup_{t \to \infty} \left( v(t) \exp\left( -\frac{kt}{2 \exp(f_i)} \right) \right) \leq \limsup_{t \to \infty} \left( v(T) + \frac{8}{k} t^{\frac{3}{2}} \right) \exp\left( -\frac{kt}{2 \exp(f_i)} \right) = 0.
\]

Therefore, \( \lim_{t \to \infty} (v(t) \exp(-kt/2 \exp(f_i))) = 0. \)

Now we prove (b) \( E(\exp(-f_{\lambda i})) < \infty. \) From (20), when \( k \geq 3, \) there exists a small enough \( \epsilon_1 \) such that \( E(\exp(-f_{\lambda i})I_{\{y_i < \epsilon_1\}}) \leq E(y_i^{-1}I_{\{y_i < \epsilon_1\}}) < \infty. \) Similarly, from (22), there exists a large enough \( \epsilon_2 \) such that \( E(\exp(-f_{\lambda i})I_{\{y_i > \epsilon_2\}}) < \infty. \) Then \( E(\exp(-f_{\lambda i})) = E(\exp(-f_{\lambda i})I_{\{y_i < \epsilon_1\}}) + E(\exp(-f_{\lambda i})I_{\{y_i > \epsilon_2\}}) + E(\exp(-f_{\lambda i})I_{\{\epsilon_1 \leq y_i \leq \epsilon_2\}}) < \infty. \)

**APPENDIX B**

**Lemma 1.** (Leave-one-out lemma) \( h_{\lambda i}(i, \exp(f_{\lambda i}^{(-i)}), x) = f_{\lambda i}^{(-i)}(x). \)

The proof is omitted since it is similar to that in Xiang and Wahba (1996). Lemma 1 states that the estimated function based on \( y^{-i} = (y_1, \cdots, y_{i-1}, \exp(f_{\lambda i}^{(-i)}), y_{i+1}, \cdots, y_n)^T \) is the same as \( f_{\lambda i}^{(-i)}. \) It is easy to check that \( \partial^2 J/\partial f \partial f^T = W + n\lambda\Omega \) and \( \partial^2 J/\partial y \partial f^T = -V. \) Using Taylor expansions as in Xiang and Wahba (1996), we have \( (f_{\lambda i} - f_{\lambda i}^{(-i)})/(y_i - \exp(f_{\lambda i}^{(-i)})) \approx d_{ii}. \) Thus by (14),

\[
CV(\lambda) = L(\lambda) + \sum_{i=1}^{n} y_i \left( \exp(-f_{\lambda i}^{(-i)}) - \exp(-f_{\lambda i}) \right)
\approx L(\lambda) + \sum_{i=1}^{n} y_i \exp(-f_{\lambda i}) \left( f_{\lambda i} - f_{\lambda i}^{(-i)} \right) = L(\lambda) + \sum_{i=1}^{n} y_i \exp(-f_{\lambda i}) \left( f_{\lambda i} - f_{\lambda i}^{(-i)} \right) \frac{y_i - \exp(f_{\lambda i})}{y_i - \exp(f_{\lambda i}^{(-i)})} \frac{y_i - \exp(f_{\lambda i})}{y_i - \exp(f_{\lambda i}^{(-i)})} 
\approx L(\lambda) + \sum_{i=1}^{n} y_i \exp(-f_{\lambda i}) \left( f_{\lambda i} - f_{\lambda i}^{(-i)} \right) \frac{y_i - \exp(f_{\lambda i})}{y_i - \exp(f_{\lambda i}^{(-i)})} 
= L(\lambda) + \sum_{i=1}^{n} y_i \exp(-f_{\lambda i}) \frac{y_i - \exp(f_{\lambda i})}{y_i - \exp(f_{\lambda i}^{(-i)})} 
\approx L(\lambda) + \sum_{i=1}^{n} d_{ii} y_i \exp(-f_{\lambda i})(y_i - \exp(f_{\lambda i})) 
\approx L(\lambda) + \sum_{i=1}^{n} d_{ii} y_i \frac{y_i - \exp(f_{\lambda i})}{1 - d_{ii} \exp(f_{\lambda i})}
\]

**References**


Table 1: Number of replications out of 100 that failed to converge for the indirect UBR method.

<table>
<thead>
<tr>
<th></th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 3$</th>
<th>$k = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$f_1$</td>
<td>$f_{2,3}$</td>
<td>$f_4$</td>
<td>$f_5$</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>98</td>
<td>100</td>
<td>97</td>
<td>95</td>
</tr>
<tr>
<td>$n = 200$</td>
<td>97</td>
<td>100</td>
<td>94</td>
<td>97</td>
</tr>
<tr>
<td>$n = 300$</td>
<td>92</td>
<td>100</td>
<td>87</td>
<td>93</td>
</tr>
<tr>
<td>$n = 400$</td>
<td>88</td>
<td>100</td>
<td>88</td>
<td>97</td>
</tr>
</tbody>
</table>
Figure 1: Plots of CKL (solid lines), UBR (long-dashed lines), GACV1 (short-dashed lines), GACV2 (dotted lines) and GML (mixed dotted-dashed lines) criteria as functions of $\log_{10}(\lambda)$ for a typical simulated data set with $f(x) = f_5(x)$, $n = 100$, and $k = 1$ and $k = 2$. Minima of these curves are marked by crosses.
Figure 2: Boxplots of log CKL for $f_1$-$f_3$ with $n = 200$. Boxplots for indirect methods are based on converged replications.
Figure 3: Boxplots of log CKL for $f_4$-$f_7$ with $n = 200$. Boxplots for indirect methods are based on converged replications.
Figure 4: Plots of CKL’s of the indirect UBR (+), indirect GML (△) and UBR (×) vs CKL’s of GML for $f(x) = f_{\theta}(x)$, $n = 100$, and $k = 1$ and $k = 3$. 
Figure 5: Plots of the true functions (solid lines), the 5th (short-dashed lines), 50th (mixed dotted-dashed lines), and 95th (long-dashed lines) best estimates ordered by CKL. $n = 100$ and $\lambda$’s are selected by the GML method. To make the y-axis common for all panels, the curves in the last row are shifted up by 3.
Figure 6: Plots of observations and fits for chromosome 10 of the cell strain GM13330010220 (left), and chromosomes 11 (middle) and 12 (right) of the cell strain GM03134001218. Points are sample variances on the logarithm scale. Solid and mixed dotted-dashed lines are smoothing spline fits and lowess fits, respectively. Two dashed lines in each plot are 95% Bayesian confidence intervals of the smoothing spline fit.