Journal of the Chinese Statistical Association
Vol. 58, (2020) 23–44

Smoothing Spline Estimation for Skew-Symmetric Density Functions

Sheng-Mao Chang[1,†], **Nan-Cheng Su**[2] **and Yunchan Chi**[1]

[1]Department of Statistics and Institute of Data Science, National Cheng Kung University

[2]Department of Statistics, National Taipei University

ABSTRACT

The location-scale skew-symmetric distribution, consisting of two parts: a symmetric density function and a skew function, is attracting increasing attention, especially when observed data are obviously asymmetric. In this article, given the location and the scale parameter, we propose a smoothing spline estimation of the skew function, and then apply a profile likelihood approach to estimate the location and the scale. Moreover, an approximate cross-validation is derived to estimate relative Kullback-Leibler distances to alleviate the computation burden in choosing an optimal smoothing parameter for smoothing spline modeling. The proposed skew function estimator is twice differentiable and hence a Newton approach can be applied to find the maximum profile likelihood estimation. The performance of the proposed approach was examined by simulation and real data examples.

Key words and phrases: Approximate cross validation, Profile likelihood, Skew-normal distribution, Skew-symmetric distribution, Smoothing spline.

JEL classification: C13, C14.

[†]Corresponding to: Sheng-Mao Chang
E-mail: smchang@mail.ncku.edu.tw

1. Introduction

Because the symmetry of observed data is more often the exception rather than the rule, the technique of constructing rival distributions against to the normal distribution is attracting increasing attention in disciplines ranging from biomedical, engineering, and financial science, among others. In some applications, the location and / or scale parameter are of major concern, e.g., biomedical measurements; whereas in other applications, probabilities of events are on demand, e.g., default probabilities in finance. We refer to [7] for a variety of such applications. Rao [19] explained this asymmetry as being due to the fact that the observed data are drawn from a symmetric distribution with certain distortion, not completely randomly sampled. Modeling the distortion together with a symmetric distribution has been proven a success in many applications. A large class of this kind of modeling has been proposed in the literature and most of these models are originated from the skew-normal distribution, Azzalini [3]. The skew-normal density function has the form $2\phi(x)\Phi(\alpha x)$ for some $\alpha \in \mathbb{R}$ where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and the cdf of standard normal distribution. The $\phi(\cdot)$ represents the part of random sampling, the $\Phi(\cdot)$ represents the distortion, and $\alpha = 0$ if there is no distortion.

To extend the skew-normal, Wang *et al.* [22] defined the class of skew-symmetric (SS) family. Specifically, a location-scale SS density has the form

$$\zeta(x; \mu, \sigma, \gamma, \pi) = \frac{2}{\sigma} f_\gamma \left(\frac{x-\mu}{\sigma} \right) \pi \left(\frac{x-\mu}{\sigma} \right)$$

where $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$, γ is a vector of parameter other than location and scale, $f_\gamma(\cdot)$ is a symmetric pdf defined in $\mathbb{R}$, and the skew function $\pi(x)$ satisfies $0 \leq \pi(x) \leq 1$ and $\pi(\mu + x) + \pi(\mu - x) = 1$, $\forall x \in \mathbb{R}$. As an example, Nadarajah and Kotz [17], Arellano-Valle *et al.* [1] and Arnold and Lin [2] investigated the skew-symmetric distribution with $f(z) = \phi(z)$ for various π where $z = (x - \mu)/\sigma$; similarly, Gupta *et al.* [9] and Nadarajah and Kotz [18] studied the skew-symmetric distribution with $\pi(z) = H(\alpha z)$ where $\alpha \in \mathbb{R}$ and H is a cdf satisfying H' is symmetric about 0, and f is one of the following pdf's: normal, Student's t, Cauchy, Laplace, logistic, and uniform. General

properties of skew-symmetric distribution have been studied in [4, 7].

Furthermore, Ma *et al.* [10] proposed a semiparametric approach to estimate (μ, σ). Their approach is insensitive to the posited skew function, so a reasonable guess of $\pi(\cdot)$ should suffice to obtain a consistent estimation of (μ, σ). An obvious consequence is that the overall density estimation of $\zeta(\cdot)$ may deviate a lot from the histogram of data which fails the estimation of tail probabilities. In order to have an accurate estimation of the SS density, Ma and Hart [11] further proposed a constrained local likelihood approach which applies local information to estimate the parameters (μ, σ) involved in the skew function. They suggested modeling the skew function by $\pi(x) = H(g(x))$, where $H : \mathbb{R} \to [0, 1]$ is the cdf of any continuous symmetric distribution and $g(x)$ is a specialized function which is not differentiable at observed points. Frederic [6] provided a B-spline modeling for $g(\cdot)$ with penalties. Essentially, B-spline is a linear combination of functions and these functions depend on a partition of a chosen interval in the real line. The breakup points are called interior knots. The choice of interior knots plays an important role of curve estimations and the optimal choice of knots, in general, increases the computation burden dramatically. Several optimal choices can be found in [14, 13], among others.

To completely waive the knot selection issue, smoothing spline, using all the data points as knots, is used for the skew function estimation. We refer to [21, 8] for a detailed construction for smoothing spline and its various applications. However, for skew function estimation, the smoothing spline theory does not apply directly because we are looking for an odd (and smooth) function rather than a smooth function from a general function space. Finally, the profile likelihood approach is applied to estimate parameters. To prevent ambiguity, hereafter, we call $\theta^{\mathrm{T}} = (\mu, \sigma, \gamma^T)$ the parameter and call $\pi(\cdot)$ the skew function, although the skew function can be treated as a (vector of) nuisance parameter. The additional parameters γ represents the parameter of the symmetric density function except the location μ and scale σ. For example, if the density function is the t-distribution with degrees of freedom ν then $\gamma = \nu$. The remainder of this article is arranged as follows. In Section 2, we define the proposed skew function estimator as well as the estimation procedure followed by a short comment on

its large sample property. In Section 3, we address issues on fitting the smoothing spline and choosing optimal tuning parameter for smoothing splines followed by simulation and real data analyses in Section 4. Finally, conclusions are drawn in Section 5.

2. Method

We begin with the skew function estimation via smoothing spline when θ is given. The estimation of θ is therefore done by the profile likelihood approach. For the skew function modeling, other than aforementioned constrains, we further assume that $\pi(\cdot)$ is twice differentiable and square integrable over a closed interval. These conditions are required to fit a smooth skew function.

Suppose that we have n random samples, say $x_1, \ldots, x_n$, from a skew-symmetric distribution. The negative log-likelihood of the SS distribution can be written as

$$l(\theta, \pi) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\log\left(\frac{2}{\sigma}f_\gamma\left(\frac{x_i-\mu}{\sigma}\right)\right) + \log\pi\left(\frac{x_i-\mu}{\sigma}\right)\right\}. \tag{1}$$

We will use these two expressions interchangeably. When θ is known, without loss of generality, assume that the population has $\mu = 0$ and $\sigma^2 = 1$. To relax the first constraint on π, $\pi : \mathbb{R} \to [0, 1]$, define function $g : \mathbb{R} \to \mathbb{R}$ and $\pi(x) = L(g(x)) = e^{g(x)}/1+e^{g(x)}$. With the restriction $\pi(x)+\pi(-x) = 1$ and after rearrangement, we have $g(x) + g(-x) = 0$ for all x. This implies that g must be an odd function. Note that, in Introduction, we mentioned that Ma and Hart [11] suggested using $\pi(x) = H(g(x))$ for arbitrary cdf $H(\cdot)$ whose probability density function is symmetric. Here, we choose $\pi(x) = L(g(x))$, where $L(\cdot)$ is the cdf of logistic distribution, to ensure the weak convergence. Other choices of $H(\cdot)$ may need additional manipulation to prove the convergence. For more details, please see Appendix A and B.

For a general situation, let $\pi_\theta = \pi((x-\mu)/\sigma)$ and $g_\theta = g((x-\mu)/\sigma)$. In this article, we search for the best g_θ from the mth order Soblov space $W_m(S) = \{g : g^{(i)}, i = 0, \cdots, m-1,$ are absolutely continuous, and $\int_S [g^{(m)}(x)]^2 dx < \infty\}$ and $S = [-0.5, 0.5]$. However, g_θ must be an odd function and thus we define the subspace $W_{mO}(S)$ as the collection of all odd functions in $W_m(S)$. Consequently, when θ is given, the smoothing

spline estimator of g_θ is

$$g_{\theta\lambda}^{SS} = \underset{g_\theta \in W_{mO}(S)}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^{n} \{-g_\theta(x_i) + \log\left(1 + \exp(g_\theta(x_i))\right)\} + \lambda^2 I_m^2(g_\theta) \right] \tag{2}$$

where $\lambda > 0$ and $I_m^2(g_\theta) = \int_S \left[g_\theta^{(m)}(x)\right]^2 dx$, the so called roughness penalty. The explicit form of $I_m^2(g_\theta)$ is shown in Appendix C. The calculation of $g_{\theta\lambda}^{SS}$ is deferred to Section 3.1 and Appendix C. In short, if we observe n independent variables, say $x_1, \cdots, x_n$, then, according to the chosen kernel, we can find a design matrix $\mathbf{Z} \in \mathbb{R}^{n \times n_1}$, $n < n_1$, and a vector $b \in R^{n_1}$ such that $(g_{\theta\lambda}^{SS}(x_1), \cdots, g_{\theta\lambda}^{SS}(x_n))^T = \mathbf{Z}b$. Note that m is determined by the kernel. This linear form is helpful in choosing the tuning parameter λ in terms of deriving the approximated cross validation, as shown in Section 3.2.

With $\pi_{\theta\lambda}^{SS} = L(g_{\theta\lambda}^{SS})$, the profile likelihood estimator of θ is simply the minimizer of the following function

$$l_p(\theta; \lambda) = -\frac{1}{n} \sum_{i=1}^{n} \left\{ \log\left(\frac{2}{\sigma} f_\gamma\left(\frac{x_i - \mu}{\sigma}\right)\right) + \log\left(\pi_{\theta\lambda}^{SS}\left(\frac{x_i - \mu}{\sigma}\right)\right) \right\} \tag{3}$$

where the subscript p stands for profile. In this sequel, the suggested estimation procedure is as follows:

1. Design a sequence of smoothing parameter $\{\lambda_j\}$, $j = 1, \cdots, q$, $q \geq 2$. Set initial values of μ and σ^2 by sample mean and sample variance, respectively.

2. For each λ_j, estimate θ_j by maximizing (3) and then estimate $\pi_{\theta_j\lambda_j}^{SS}$ by minimizing (2).

3. Choose the best estimation θ_j and $\pi_{\theta_j\lambda_j}^{SS}$ by the approximated cross validation approach defined in Section 3.2.

Last, we address the large sample property of the profile estimator of θ. According to the semiparametic approach proposed by [10], the efficient score estimator of θ is also consistent when plugging-in the smoothing spline estimator $\pi_{\theta\lambda}^{SS}$ in the sense that this estimator is a submodel of the true model. On the other hand, the consistency and the normality of the maximum profile likelihood estimator $\hat{\theta}$ both hold well if the

condition $||\pi^{SS}_{\hat{\theta}\lambda} - \pi^*_{\theta^*}|| = O_p(||\hat{\theta} - \theta^*||) + o_p(n^{-1/4})$ can be proven [15], where θ^* and π^* denote the true parameter and the true skew function, respectively. In general, the convergence rate of the smoothing spline is $o_p(n^{-m/(2m+1)})$, [8]. However, we estimate π and penalize the log profile likelihood with respect to g (but not to π) so modification is needed. The result is shown in Theorem 1 and the convergence is defined with respect to Hellinger's distance $h(g, g^*) = [0.5 \int (\sqrt{g(x)} - \sqrt{g^*(x)})^2 dx]^{1/2}$. The proof is deferred to Appendix A. Finally, with $m = 1$, the convergence rate of the smoothing spline estimator is of order $o_p(n^{-1/3})$, which is faster than $o_p(n^{-1/4})$, so the asymptotic normality of the proposed profile likelihood estimate θ holds whenever $m \geq 1$.

Theorem 1. *Let g^* be the true underlying function and $g^{SS}_{\theta\lambda}$ be the smoothing spline estimator defined in (2). Then*

$$h\left(L\left(g^{SS}_{\theta^*\lambda}\right), L(g^*)\right) = O_P(\lambda)(1 + I(g^*))$$

provided $\lambda^{-1} = O_P\left(n^{m/2m+1}\right)(1 + I(g^*))^{1/2}$.

3. Computation Issues

3.1 Fitting Smoothing Splines

Wahba [21] and Gu [8] demonstrated several ways to search for functions from W_m by smoothing splines. They were applied to various applications. However, in our case, the usual reproducing kernels are not applicable because we are looking for an odd function from W_m. To this end, we modify our optimization problem and show that the original problem (2) and the new problem (4), defined below, result in the same minimizer. Note that, in terms of profile likelihood, the following functions g and η depend on θ but we omit the subscript θ for shorthand.

Define $g(x) = [\eta(x) - \eta(-x)]/2$ for any $\eta \in W_m$. Then, g is an odd function, *i.e.* $g \in W_{mO}$. The smoothing spline estimator of $\eta(\cdot)$ is therefore

$$\underset{\eta \in W_m(S)}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ -\left[\eta(x_i) - \eta(-x_i)\right]/2 + \log\left(1 + e^{[\eta(x_i) - \eta(-x_i)]/2}\right) \right\} + \lambda^2 I_m^2(\eta) \right]. \quad (4)$$

We define the new estimator of g as $\hat{g}(x) = [\hat{\eta}(x) - \hat{\eta}(-x)]/2$. On the other hand, for any $g \in W_m$, one can always find infinity many η such that $[\eta(x) - \eta(-x)]/2 = g(x)$, e.g., for every $c \in \mathbb{R}$, $\eta_c(x) = \eta(x) + c$ results in the same $g(x)$. This means that if there exists a true g then there exist more than one η resulting in the same g. Therefore, solving (2) is preferred rather than solving (4).

Furthermore, we notice that the likelihood parts in two problems are identical but the penalty functions are not. The idea to link these two optimization problems is factorizing the function $\eta \in W_m$ into an odd function $\eta_O \in W_{mO}$ and an even function $\eta_E \in W_{mE}$, *i.e.* $\eta = \eta_O + \eta_E$. This turns out that the penalty function in (4) can be written as

$$I_m^2(\eta) = \int_S [\eta_O^{(m)}(x)]^2 dx + 2\int_S \eta_O^{(m)}(x)\eta_E^{(m)}(x)dx + \int_S [\eta_E^{(m)}(x)]^2 dx$$

and the second integrand of the right-hand-sided is an even function and hence equals zero, *i.e.* $I_m^2(\eta) = I_m^2(\eta_O) + I_m^2(\eta_E)$. Also, since W_{mE} and W_{mO} are subspaces and orthogonal complements of W_m the uniqueness of the factorization is based on the completeness of W_m and W_m is in fact complete. With this knowledge, Theorem 2 below shows the equivalence of two estimates of g. The proof is deferred to Appendix B. The final smoothing spline fitting is identical to the one in [8] and, to make this article self contained, we show it in Appendix C.

Theorem 2. *Given $\lambda \in \mathbb{R}^+$, denote the minimizer of (4) as $\hat{\eta}_\lambda(x)$ and $\hat{g}_\lambda^{(1)}(x) = [\hat{\eta}_\lambda(x) - \hat{\eta}_\lambda(-x)]/2$; the minimizer of (2) as $\hat{g}_\lambda^{(2)}(x)$. Then, $\hat{g}_\lambda^{(1)}(x) = \hat{g}_\lambda^{(2)}(x)$.*

3.2 Choosing Smoothing Parameter

The smoothing parameter λ can be obtained by minimizing the Kullback-Leibler (KL) distance between the true function and the estimated function. The KL distance is defined as $\mathrm{KL}(l, l_\lambda) = \mathrm{E}_l[\log(l/l_\lambda)]$ where l denotes the true distribution and l_λ denotes the estimated distribution with tuning parameter λ. In our case, assuming θ is known, we are interested in choosing a λ such that $\mathrm{KL}(\pi^*, \pi_{\theta\lambda}^{SS})$ is minimized where π^* is the true skew function. Note that π is in fact a function of g, so hereafter we represent π by

g and shorthand the smoothing spline estimator $g_{\theta\lambda}^{SS}$ by g_λ for convenience. Ignoring terms in (3) that λ does not involve, we have the relative KL (RKL) distance,

$$\mathrm{RKL}(\lambda) = -\mathrm{E}_{\pi^*}\left[g_\lambda(X) - \log\left(1 + e^{g_\lambda(X)}\right)\right]$$

The cross-validation (CV) method suggests an unbiased estimator of RKL as

$$\widehat{\mathrm{RKL}}_{CV}(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\left\{g_\lambda^{[i]}(x_i) - \log\left(1 + e^{g_\lambda^{[i]}(x_i)}\right)\right\}$$

where $g_\lambda^{[i]}$ is the solution of the delete-one-observation version of (2). By doing so, n smoothing spline curve fits are required.

To save the computational burden, we derived an approximate CV (ACV) by mimicking the one mentioned in [8]. Before moving forward, we define that, for a function or parameter ψ, $\tilde{\psi}$ means full data estimation and $\psi^{[i]}$ means leave-one-out estimation. The rationale of ACV derivation is as follows. First, assuming θ and λ are known, the $\mathrm{RKL}(\lambda)$ can be approximated by a quadratic form (applying Taylor expansion with respect to g at $\tilde{g}_\lambda$). The smoothing spline estimator takes the form $\mathbf{Z}b$ and thus the approximated $\mathrm{RKL}(\lambda)$ is a quadratic form of b. In this sequel, g can be approximated in one Newton step. Second, with the assumption that dropping one observation does alter estimation a little, the leave-one-out estimate $g_\lambda^{[i]}$ is approximated by the minimizer of the approximated $\mathrm{RKL}(\lambda)$ by ignoring all the information from the i^{th} individual. The detailed derivation is deferred to Appendix D. Finally, an approximation of RKL_{CV} is

$$\widehat{\mathrm{RKL}}_{ACV}(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\left\{g_{a\lambda}^{[i]}(x_i) - \log\left(1 + e^{g_{a\lambda}^{[i]}(x_i)}\right)\right\}$$

where $g_{a\lambda}^{[i]}$ is provided in Appendix C. The smoothing parameter λ can be chosen so that $\widehat{\mathrm{RKL}}_{ACV}$ is minimized. Note that the leave-one-out estimate $g_\lambda^{[i]}$ requires several Newton iteration steps, whereas the approximate version $g_{a\lambda}^{[i]}$ needs only one Newton step, and hence the savings from using the latter can be substantial.

There are several other computationally efficient ways to approximate CV. One simple but effective method is the K-fold CV, which divides data into K pieces. Each

time, $K-1$ pieces are used to model fitting and the remaining one is used for model validation or performance evaluation. Usually, five-fold CV (FFCV) is competitive with other approximated versions. We will compare FFCV and proposed ACV with a specific skew function in the next section.

4. Simulation and Case Study

Note that, when performing a simulation we know the underlying distribution, so we can evaluate exact RKL by Monte Carlo integration. We call this method the *optimal method* and the resulting RKL as the *optimal RKL*.

4.1 The Performance of ACV

In our first simulation study, we generated 200 datasets each with 100 random samples from the density $2\phi(x)\Phi(-3x^3+x)$. Three methods, ACV, FFCV and Optimal method, were applied to choose smoothing parameter and hence to evaluate RKL's. The optimal choices of ACV, FFCV, and optimal RKL are shown in the upper panel of Figure 1. On average, it appears that λ's chosen by Optimal RKL are smaller than by ACV, and λ's chosen by ACV are smaller than by FFCV. However, λ's derived by ACV deviate from λ's derived by Optimal RKL in a higher rate then λ's derived by FFCV. From the lower panel of Figure 1, FFCV performs better than ACV in terms of curve variation. However, ACV performs well enough to compensate for the loss of model fitting and the consumption of computing time. In the lower panel of Figure 1, 200 fits (gray lines) are overlapped. The true curves (black line) are almost in the middle of gray shadow except at the two ends. The optimal method, calculating RKL using Monte Carlo integration with 1000 samples, has smaller variation than the two others though it is not very obvious. Again, this shows that ACV and FFCV approximate RKL very well and are reliable for choosing smoothing parameters.

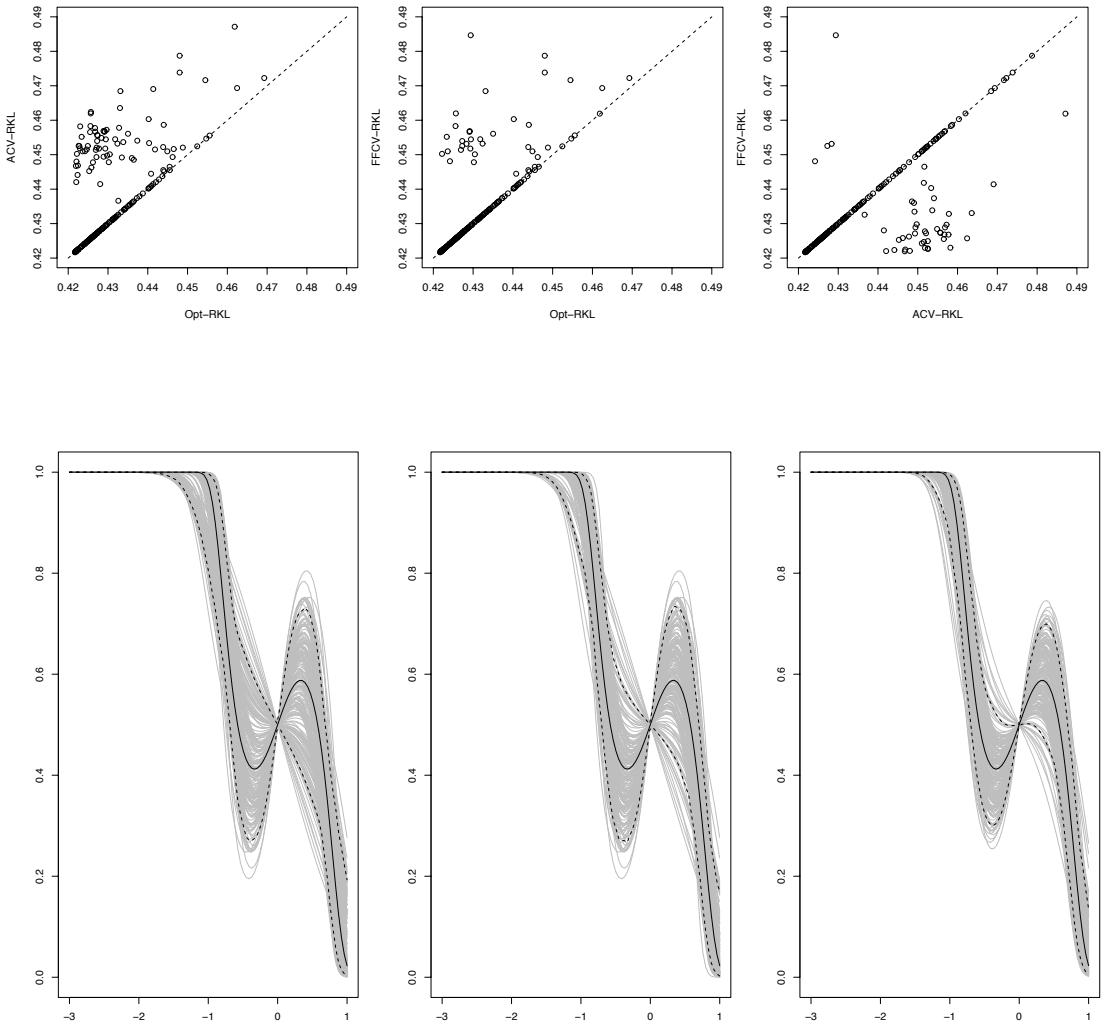

Figure 1: The upper panel shows optimal λ's derived by ACV, FFCV, and Optimal RKL whereas the lower panel shows skew-function fits via ACV, FFCV, and optimal RKL, from left to right. Each of them contains 200 fits (gray lines), one true curve (black line) and pointwise 90% C.I. (dotted black lines.) Figures in the lower panel have x-axis x and y-axis $\pi^{SS}_{\theta^*_\lambda}(x)$.

4.2 Simulation Study

According to the previous simulation, we observe that ACV performs well and saves computation greatly. So, we apply ACV approach to choose the tuning parameter in the following simulations to examine the performance of the proposed smoothing spline method. Two skew functions are considered: $\pi_1(x) = \Phi(\sin(-3x))$ and $\pi_2(x) = \Phi(-3x^3 + x)$. These skew functions were also adopted in [10, 11]. We generated 1000 datasets, each consisting of 100 or 200 samples from the density function $2\phi(x; \mu, \sigma^2)\pi_{\theta,j}(x)$, $j = 1, 2$ with $\mu = 3$, $\sigma^2 = 1$ or 2 and $\pi_{\theta,j}(x) = \pi_j((x - \mu)/\sigma)$. For the skew function, we evaluated the sample mean square errors

$$\sum_{i=1}^{n} \left[\pi_{\hat{\theta},j}(x_i) - \hat{\pi}_{\hat{\theta},j}(x_i)\right]^2 /n,$$

which is the empirical version of the integrated mean square error (IMSE) in nonparametric curve fitting literature. Results are summarized in Table 1 which shows that the parameter estimation is numerically unbiased because all 95% confidence intervals cover their own underlying parameter values. In addition, the estimated standard errors were very close to the sample standard errors. The standard error estimates were carried out by the inverse of the approximate Hessian, evaluated by numerical differences. However, in contrast to small variance ($\sigma = 1$), large variance ($\sigma = 2$) tended to result in underestimation of both μ and σ. Fortunately, when the sample size increased from 100 to 200, the underestimation was alleviated. Last, larger variance σ^2 resulted in larger IMSE and $\pi_{\theta,1}$ yields larger IMSE than $\pi_{\theta,2}$. Overall, IMSE decreased when sample size increased.

4.3 Australian Institute of Sport Data

Next, we analyzed the Australian Institute of Sport data [5] which consists of 11 biological measurements made on 102 male and 100 female Australian athletes. We only analyzed one of these biological measurements, the body mass index (BMI) which is continuous and appeals to follow normal distribution in general population. It is

Table 1: Simulation Results.

		$\pi_{\theta,1}$			$\pi_{\theta,2}$		
		ave.	ave. s.e. est.	emp. s.e.	ave.	ave. s.e. est.	emp. s.e.
$n = 100$	$\mu = 3$	3.0015	0.0558	0.0655	2.9893	0.0716	0.0686
	$\sigma = 1$	0.9975	0.0736	0.0752	0.9953	0.0852	0.0782
	IMSE	0.2123	–	–	0.0085	–	–
$n = 100$	$\mu = 3$	2.9975	0.1105	0.1281	2.8578	0.1384	0.2548
	$\sigma = 2$	1.9958	0.1465	0.1488	1.8836	0.1540	0.2455
	IMSE	0.2126	–	–	0.0716	–	–
$n = 200$	$\mu = 3$	3.0006	0.0409	0.0447	2.9979	0.0538	0.0600
	$\sigma = 1$	1.0001	0.0514	0.0488	0.9990	0.0600	0.0594
	IMSE	0.2083	–	–	0.0049	–	–
$n = 200$	$\mu = 3$	3.0056	0.0818	0.0895	2.9153	0.1066	0.2025
	$\sigma = 2$	1.9913	0.1014	0.1070	1.9458	0.1140	0.1532
	IMSE	0.2060	–	–	0.0095	–	–

reasonable to believe that athlete group may be the consequence of "bias sampling" or of "selection" and to assume that BMI's in athletic population follows a skew-normal distribution (or other generalization.)

Figure 2 shows contour plots for the log-profile-likelihood of 102 male (left) and 100 female (right) athletes' BMI. Each summit marked by a cross locates the MLE of location μ and scale σ parameter. The right-most plot shows the resulting skew functions of male (dotted line) and female (solid line). Small bars on the top represent the normalized BMI's of males while bars on the bottom are for females. Since the likelihood may not be strictly convex, we plotted the contours of likelihood in Figure 2. It seems to be smooth and unimodal but using grid search is recommended to find good initial values. The skew function is shown in the very right-hand-panel of Figure 2. Results are shown in Table 2. In general, the proposed method yields estimates which are closer to the normal fit than the results of [10].

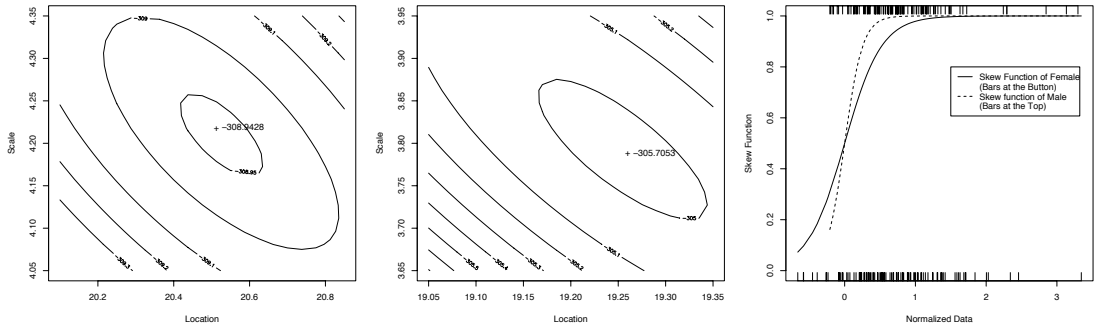

Figure 2: From right to left: contour of the log likelihood (1) for male data, contour for female data, and skew function estimates for both male and female.

Table 2: Australian Institute of Sport Data.

Method	Male				Female			
	$\hat{\mu}$	$\widehat{se(\hat{\mu})}$	$\hat{\sigma}$	$\widehat{se(\hat{\sigma})}$	$\hat{\mu}$	$\widehat{se(\hat{\mu})}$	$\hat{\sigma}$	$\widehat{se(\hat{\sigma})}$
Normal	23.90	0.27	2.77	0.20	21.99	0.26	2.65	0.19
π^{SS}	20.79	0.66	2.93	0.29	19.45	0.42	3.65	0.34

The row named Normal presents the normal fit of the data whereas the π^{SS} row presents the proposed fit.

4.4 Peruvian Indian Pulse Data

This dataset consists of 39 pulse rates of Peruvian indians [20]. The pulse data seems to have a heavier right-tail and there are no other covariates to explain the skewness. Skew-symmetric modeling is therefore suggested. Again, the contour plot of log likelihood against two parameters is shown in Figure 3. With a relevant fine-search, parameter estimates are summarized in Table 3. Since the dataset consists of only 39 individuals, we are less confident to claim that it really follows a skew-symmetric distribution. However, our density fit (the very left panel of Figure 3) captures the major peak and a protuberance at the right end.

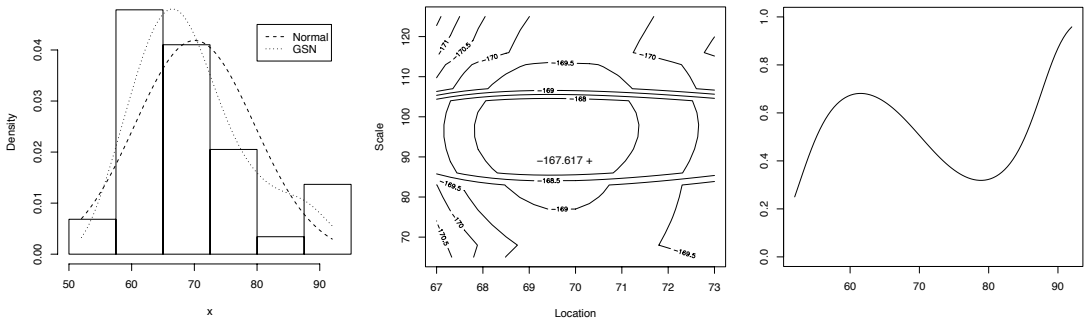

Figure 3: The left panel is the histogram of Peruvian Indian Pulse data. The dashed line represents the proposed fit. The middle panel is the log profile likelihood contour plot. The right panel is the skew function solved by proposed method where the x-axis is pulse and the y-axis is the estimated skew-function.

Table 3: Peruvian Indian Pulse Data.

Method	$\hat{\mu}$	$\widehat{se(\hat{\mu})}$	$\hat{\sigma}$	$\widehat{se(\hat{\sigma})}$
Normal	70.051	1.526	9.594	1.108
π^{SS}	70.193	2.933	9.409	0.918

The row named Normal presents the normal fit of the data whereas the π^{SS} row presents the proposed fit.

5. Concluding Remarks

For finding the MLE of the profile likelihood, an inevitable difficulty is that $l_p(\theta)$ may not be concave in terms of θ, so a good initial value is important to the convergence of the algorithm. Following [10], initial values are taken in a reasonable range. If these initial values result in different θ or $\pi(\cdot)$ then the MLE is chosen such that that the corresponding likelihood (1) is maximized.

We advocate the use of a smoothing spline to estimate the skew function because the proposed estimator has two major merits: first, unlike the local likelihood approach [11], the skew function estimator is twice differentiable everywhere and, second, unlike the B-spline approach [6], there is no additional consideration in choosing interior knots. Finally, together with the proposed ACV, the estimates of θ and π behave no worse than the one using original CV and our approach reduces computation burden greatly.

Appendix A. Proof of Theorem 1

The proof of Theorem 1 is almost identical to Theorem 10.6 in [16] but with different arguments in Hellinger metric. Before moving forward, we define useful subsets: $\mathcal{G}_m(M) = \{g : g \in W_m, I_m(g) \leq M\}$ for some $1 \leq M < \infty$; $\mathcal{G}_{mO}(M) = \mathcal{G}_m(M) \cap W_{mO}$; and $\mathcal{G}_m^F(\varpi, M) = \{\varpi(g) : g \in \mathcal{G}_{mO}(M)\}$ for some function ϖ. These subsets will be adopted in various situations below in order to show convergence. Besides, analogous to Lemma 10.5 of [16], we have basic inequality

$$h^2(\nu(\hat{g}), \nu(g^*)) + 4\lambda^2 I_m^2(\hat{g}) \leq 16 \int \rho \circ \nu(\hat{g}) d(P_n - P) + 4\lambda^2 I_m^2(g^*)$$

where $\rho \circ \nu(g) \equiv \rho(\nu(g))$, $\rho(w) = 0.5 \log[(w + w^*)/2w^*]$, $\nu(g) = -g + \log(1 + e^g)$ and $\log \pi = \nu(g)$. The function ρ is well defined since $\nu(w) > 0 \ \forall w \in \mathbb{R}$.

Equipping with above basic inequality, the proof of convergence is done by showing that the regularity conditions in Theorem 10.6 of van de Geer (2000) hold. These conditions are: 1) for any $M > 1$ and $g \in G_1 \equiv \mathcal{G}_m^F(\rho \circ \nu, M) = \{\rho \circ \nu(g) : g \in W_{mO}, I_m(g) + I_m(g_0) \leq M\}$, and $||g - g_0|| \leq h(g, g_0)$; 2) $\sup_{g \in G_1} h(g, g_0) \leq c_0 M$, for some $c_0 > 0$; 3) $H_B(\delta, G_1, P) \leq A \left(\frac{M}{\delta}\right)^{1/m}$, for all $\delta > 0$; and 4) $\sup_{g \in G_1} |g - g_0|_\infty \leq c_0 M$. Here, $H_B(\cdot)$ denotes the bracketing entropy of the corresponding class of function under square-norm metric. Condition 1 and 2 are inherited by Hellinger metric so we omit them. Condition 3 has been shown in Lemma 1 below together with the fact that G_1 has finite entropy [16]. Finally, condition 4 is satisfied because both ν and ρ are Lipschitz and monotone decreasing which can be proved by taking derivatives, see Lemma 2.

Lemma 1. *Let φ be Lipschitz and monotone. For any $\delta > 0$ and $m \in \mathcal{N}$, we have*

$$H_B(c_1\delta, \mathcal{G}_m^F(\varphi, M), P) \leq H_B(\delta, \mathcal{G}_m(M), P) \leq A \left(\frac{M}{\delta}\right)^{\frac{1}{m}},$$

for some positive constant c_1.

Proof. The second inequality is proven by [16]. Therefore, we only show the first one. By the definition of bracketing, for all $g \in \mathcal{G}_m$, there exists a pair $\{g_L, g_U\}$ such that

$g_L \le g \le g_U$ and $||g_U - g_L|| \le \delta$, $\delta > 0$. The smallest number of such pairs is associated with the entropy. Let $\varphi(\cdot)$ be a continuous monotone and Lipschitz function. Due to the monotonicity and Lipschitz property, $\varphi(g_L) \le \varphi(g) \le \varphi(g_U)$ for monotone increasing and $\varphi(g_U) \le \varphi(g) \le \varphi(g_L)$ for monotone decreasing. Consequently, $||\varphi(g_U) - \varphi(g_L)|| \le c_1||g_U - g_L|| \le c_1\delta$. □

Lemma 2. *For any $g \in \mathcal{G}_{mO}(M)$, we have 1) $\nu(g)$ is Lipschitz and monotone decreasing in g; and 2) $\rho \circ \nu(g)$ is Lipschitz and monotone decreasing in g.*

Proof. With the fact that if the first derivative of a function is bounded then this function is Lipschitz, two results can be shown by simply taking derivative. For any $g_1, g_2 \in \mathcal{G}_{mO}(M)$,

1. Since $-1 < \frac{d}{dg}\nu(g) = -1 + \frac{e^g}{1+e^g} < 0$, by mean value theorem, we have $|\nu(g_1) - \nu(g_2)| = |\frac{d}{dg}\nu(g)|_{g=g_m}|g_1 - g_2|$ where g_m is an intermediate value between g_1 and g_2. Therefore, $|\nu(g_1) - \nu(g_2)| \le c_1|g_1 - g_2|$ with $c_1 = 1$.

2. Again,
$$0 < \frac{d}{d\nu}\rho(\nu) = \frac{1}{2}\frac{1}{\nu + \nu^*} < \frac{1}{2\nu^*} \equiv c_2$$
where $\nu = \nu(g)$ and $\nu^* = \nu(g^*)$. This implies
$$|\rho \circ \nu(g_1) - \rho \circ \nu(g_2)| \le c_2|\nu(g_1) - \nu(g_2)| \le c_2c_1|g_1 - g_2|,$$
i.e. $\rho \circ \nu(g)$ is Lipschitz. Furthermore, since $\frac{d}{dg}\nu < 0$ and $\frac{d}{d\nu}\rho > 0$ we conclude that $\rho \circ \nu(g)$ is monotone decreasing in g.

□

Appendix B: Proof of Theorem 2

Proof. In order to preventing ambiguity, denote the minimizer of (2) with tuning parameter λ by $\hat{g}_\lambda^{(1)}$ and the minimizer of of (4) with λ by $\hat{\eta}_\lambda^{(2)}$ and hence, $\hat{g}_\lambda^{(2)}$. Let $\eta = \eta_O + \eta_E$ where $\eta \in W_m$, $\eta_O \in W_{mO}$ and $\eta_E \in W_{mE}$. Given η, η_O and η_E are

unique. Thus,

$$\begin{aligned}\hat{\eta}_\lambda^{(2)} &= \underset{\eta\in W_m}{\arg\min}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{-\left[\eta(x_i)-\eta(-x_i)\right]/2+\log\left(1+e^{[\eta(x_i)-\eta(-x_i)]/2}\right)\right\}+\lambda^2 I_m^2(\eta)\right]\\ &= \underset{\eta_O\in W_{mO},\eta_E\in W_{mE}}{\arg\min}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{-\eta_O(x_i)+\log\left(1+e^{\eta_O(x_i)}\right)\right\}+\lambda^2\left\{I_m^2(\eta_O)+I_m^2(\eta_E)\right\}\right]\\ &= \underset{\eta_O\in W_{mO}}{\arg\min}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{-\eta_O(x_i)+\log(1+e^{\eta_O(x_i)})\right\}+\lambda^2 I_m^2(\eta_O)\right]+\underset{\eta\in W_{mE}}{\arg\min}\,\lambda^2 I_m^2(\eta_E)\\ &= \hat{\eta}_{O,\lambda}^{(2)}+\eta_{E,\lambda}^{(2)}\end{aligned}$$

where $\eta_{E,\lambda}^{(2)}$ can be arbitrary function in W_{mE} such that $I_m(\eta_{E,\lambda}^{(2)})=0$. The first minimization of the last second equation is exactly problem (4) but replacing g by η_O, i.e. $\hat{\eta}_{O,\lambda}^{(2)}=\hat{g}_\lambda^{(1)}$. Besides, since $\hat{g}_\lambda^{(2)}=\left[\hat{\eta}_\lambda^{(2)}(x)-\hat{\eta}_\lambda^{(2)}(-x)\right]/2=\hat{\eta}_{O,\lambda}^{(2)}$ we conclude $\hat{g}_\lambda^{(1)}=\hat{g}_\lambda^{(2)}$. □

Appendix C: Smoothing Spline Estimation

In [8], the polynomial smoothing spline corresponds to a reproducing kernel R which can be orthogonally partitioned as $R=R_0+R_1$. One particular kernel for W_2 is $R_0(x,y)=1+k_1(x)k_1(y)$ and $R_1(x,y)=k_2(x)k_2(y)-k_4(|x-y|)$ where $k_1(z)=z-0.5$, $k_2(z)=k_1^2(z)/2-1/24$ and $k_4(z)=(k_1^4(z)-k_1^2(z)/2+7/240)/24$. By representation theory, $\eta(x)=d_1+d_2x+\sum_{j=1}^n c_iR_1(x_j,x)$ for some unknown constant $c=(c_1,\cdots,c_n)^{\mathrm{T}}$ and $d=(d_1,d_2)^{\mathrm{T}}$. Recall that $g(x)=[\eta(x)-\eta(-x)]/2$ and hence, d_1 can be arbitrary. We set $d_1=0$. Write $\mathbf{Q}c+\mathbf{S}d$ where the (i,j)th element of $\mathbf{Q}$ are $\mathbf{Q}_{ij}=\mathbf{Q}_{x_i}(x_j)=R_1(x_i,x_j)$ and $\mathbf{S}_{ij}=\mathbf{S}_{x_i}(x_j)=x_i$. Consequently, solving (4) is equivalent to find the solution of $\tau=(c^{\mathrm{T}},d)^{\mathrm{T}}$. Thus, we have

$$pl(g)=-\frac{1}{n}\sum_{i=1}^{n}W_i^{\mathrm{T}}\tau+\frac{1}{n}\sum_{i=1}^{n}\log(1+\exp\{W_i{}^{\mathrm{T}}\tau\})+\lambda^2c^{\mathrm{T}}\mathbf{Q}c \tag{5}$$

where $W_{ij}=[\mathbf{Q}_{x_i},\mathbf{S}_{x_i}](x_j)-[\mathbf{Q}_{x_i},\mathbf{S}_{x_i}](-x_j)$ and $\mathbf{W}_i$ represents the ith row of matrix $\underline{\mathrm{W}}$. The gradient and Hessian of (5) are

$$\frac{1}{n}\mathbf{W}^{\mathrm{T}}(p_\tau-1_n)+2\lambda^2\mathbf{Q}'\tau \quad\text{and}\quad \frac{1}{n}\mathbf{W}^{\mathrm{T}}V_\tau\mathbf{W}+2\lambda^2\mathbf{Q}',$$

respectively, where we define

$$\mathbf{Q}' = \begin{bmatrix} \mathbf{Q} & 0 \\ 0^{\mathrm{T}} & 0 \end{bmatrix}, \quad p_{\tau,i} = \frac{\exp\{\mathbf{W}_i^{\mathrm{T}}\tau\}}{1+\exp\{\mathbf{W}_i^{\mathrm{T}}\tau\}},$$

$\mathbf{V}_\tau$ as a diagonal matrix with diagonal elements $\mathbf{V}_{\tau,ii} = p_{\tau,i}(1-p_{\tau,i}),\ i = 1, ..., n$ and 1_n as a 1-vector with length n. Applying Newton-Raphson algorithm, the updating formula is

$$\tau^{new} = \tau - \left[\frac{1}{n}\mathbf{W}^{\mathrm{T}}\mathbf{V}_\tau\mathbf{W} + 2\lambda^2\mathbf{Q}'\right]^{-}\left[\frac{1}{n}\mathbf{W}^{\mathrm{T}}(p_\tau - 1_n) + 2\lambda^2\mathbf{Q}'\tau\right]$$

where "$^-$" represents the generalized inverse.

Note that the aforementioned reproducing kernel is defined on $[0, 1]$ whereas our problem is on $S = [-0.5, 0.5]$. This does not fail the result because both $k_2(\cdot)$ and $k_4(\cdot)$ are functions of $k_1(x) = x - 0.5$. So, in practice, we can scale the data into $[0, 1]$ and use the regular polynomial kernel or scale the data into $[-0.5, 0.5]$ and use the modified kernel with $k_1(x) = x$.

Appendix D: Derivation of the Approximate Cross Validation

The ACV is derived by a second order Taylor expansion. Denote g as the true underlying function and $\tilde{g}$ as the estimate using (5). By expanding g at $\tilde{g}$, we have

$$g - \log(1+e^g) \approx K(\tilde{g}) - \frac{1}{2}\left[(g-\tilde{g})\frac{e^{\tilde{g}/2}}{1+e^{\tilde{g}}} - \frac{1}{e^{\tilde{g}/2}}\right]^2$$

where $K(g) = g - \log(1+e^g) + e^{-g}/2$. Notice that $\tilde{g}$ is a known function and hence the first summand is a fixed value, say q. Thus, the leave-one-out version of the approximate log-likelihood has the form

$$Apl_{[i]} = q_{[i]} + \frac{1}{2(n-1)}\left\{\tau^{\mathrm{T}}\mathbf{A}_i\tau - 2\tau^{\mathrm{T}}\mathbf{A}_i\tilde{\tau} - 2\tau^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\Lambda^{[i]}p_2 + q_2\right\} + \lambda^2\tau^{\mathrm{T}}\mathbf{Q}'\tau$$

where $\mathbf{A}_i = \mathbf{W}^{\mathrm{T}}\Gamma^{[i]}\mathbf{W}$, $\Gamma = \Lambda^2$, Λ is a diagonal matrix with (i,i) element $e^{\tilde{g}(x_i)/2}/(1+e^{\tilde{g}(x_i)})$; p_2 is a vector with i^{th} element $e^{-\tilde{g}(x_i)/2}$; for matrix $\mathbf{B}$, $\mathbf{B}^{[i]}_{ij} = \mathbf{B}_{ij}$, $\forall j \neq i$, and $\mathbf{B}^{[i]}_{ii} = 0$; and q_2 is another function which involves no τ. Consequently the minimizer of $Apl_{[i]}$ is

$$\tau_{[i]} = \left[\mathbf{A}_i + 2\lambda^2(n-1)\mathbf{Q}'\right]^{-}\left[\mathbf{A}_i\tilde{\tau} + \mathbf{W}^{\mathrm{T}}\Lambda^{[i]}p_2\right].$$

Let $g^a_{[i]} = \mathbf{W}\tau_{[i]}$, $i = 1, 2, ..., n$ and the approximation of $\widehat{\mathrm{RKL}}_{CV}$ is

$$\widehat{\mathrm{RKL}}_{ACV} = -\frac{1}{n}\sum_{i=1}^{n}\left\{g^a_{[i]}(x_i) - \log\left(1 + e^{g^a_{[i]}(x_i)}\right)\right\}.$$

The smoothing parameter λ is chosen so that $\widehat{\mathrm{RKL}}_{ACV}$ is minimized.

References

[1] Arellano-Valle, R.B., Gómez, H.W. and Quintana, F.A. (2004). A new class of skew-normal distributions. *Commun. Statist. Theory Meth.*, 33: 1975-1991.

[2] Arnold, B.C. and Lin, G.D. (2004). Characterizations of the skew-normal and generalized chi distributions, *Sankhyā*, 66: 593-606.

[3] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Stat*, 12: 171-178.

[4] Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *J. Roy. Soc. Ser. B*, 65: 367-389.

[5] Cook, R.D. and Weisberg, S. (2005). An Introduction to Regression Graphics. Wiley, NY .

[6] Fredric, P. (2011). Modeling skew-symmetric distributions using B-spline and penalties. *J Stat. Plan. Infer*, 141: 2878-2890.

[7] Genton, M.G. (2004). Skew-Elliptical Distributions and Their Applications: A

Journey Beyond Normality. Edited Volume, Chapman & Hall/CRC, Boca Raton, FL.

[8] Gu, C. (2002) Smoothing Spline ANOVA Models. Springer, NY.

[9] Gupta, A.K., Chang, F.C. and Huang, W.J. (2002). Some skew-symmetric models. *Random Oper. and Stoch. Equ.*, 10: 133-140.

[10] Ma, Y., Genton, M.G. and Tsiatis, A.A. (2005). Locally efficient semiparametric estimators for generalized skew-elliptical distributions. *J. Am. Stat. Assoc.*, 100: 980-989.

[11] Ma, Y. and Hart, J.D. (2007). Constrained local likelihood estimators for semiparametric skew-normal distributions. *Biometrika*, 94: 119-134.

[12] Ma, Y. and Genton, M.G. (2004). A flexible class of skew-symmetric distributions. *Scand. J. Statist.*, 31: 459-468.

[13] Miyata, S. and Shen, X. (2005). Free-knot splines and adaptive knot selection. *Journal of Japanese Statistical Society*, 35: 303-324.

[14] Molinari, N., Durand, J. and Sabatier, R. (2004). Bounded optimal knots for regression splines. *Comput. Stat. Data An.*, 45: 159-178.

[15] Murphy, S.A. and van der Vaar, A.W. (2000). On profile likelihood. *J. Am. Stat. Assoc.*, 95: 449-465.

[16] Van de Geer, S.A. (2000). Empirical Processes in M-Estimation. Cambridge University Press, U.K..

[17] Nadarajah, S. and Kotz, S. (2003). Skewed distributions generated by the normal kernel. *Statist. Prob. Lett.*, 65: 269-277.

[18] Nadarajah, S. and Kotz, S. (2006). Skew distributions generated from different families. *Acta App. Math.*, 91: 1-37.

[19] Rao, C.R. (1985). Weighted Distributions Arising Out of Methods of Ascertainment: What Populations Does a Sample Represent? In a Celebration of Statistics: The ISI Centenary Volume, A.G. Atkinson and S.E. Fienberg, eds, New York: Springer-Verlag, pp. 543-569.

[20] Ryan, T.A. Jr., Joiner, B.L. and Ryan, B.F. (1985). The Minitab Student Handbook, Boston: Duxbery Press.

[21] Wahba, G. (1990). Spline Models for Observational Data. Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM.

[22] Wang, J., Boyer, J. and Genton, M.G. (2004). A skew-symmetric representation of multivariate distributions. *Stat. Sinica*, 14: 1259-1270.

[Received February 2019; accepted September 2019.]

偏斜化對稱分佈函數之曲線平滑估計

張升懋[1,†]　蘇南誠[2]　嵇允嬋[1]

[1]國立成功大學統計學系暨數據科學研究所

[2]國立臺北大學統計學系

摘　要

當資料呈現非對稱分佈時, 考慮有關位置尺度族之偏斜化對稱分佈是一可行的方法。位置尺度族偏斜化對稱分佈函數由二部份組成, 其一是一屬於尺度族的對稱分佈函數, 其二則是一偏斜函數。本文採用最大化剖面概似函數方法估計位置及尺度參數, 用無母數方法中之曲線平滑方法估計偏斜函數。此外, 我們亦推導出相對 Kullback-Leibler 距離, 用以簡化使用資料交叉驗證的計算量。對於曲線平滑方法, 本文使用的核函數是二次可微, 所以文獻上的大樣本性質及計算估計量的演算法得以套用。我們最後以電腦模擬及二個實證分析來展示本文所提出方法之可用性。

關鍵詞: 近似交叉驗證、剖面概似函數、偏斜化對稱分佈函數、曲線平滑估計。

JEL classification: C13, C14.

[†]通訊作者: 張升懋
E-mail: smchang@mail.ncku.edu.tw