DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 1048
December 11, 2001

Penalized Log Likelihood Density Estimation, via Smoothing-Spline ANOVA and $ranGACV$ - Comments to Hansen and Kooperberg, 'Spline Adaptation in Extended Linear Models'

Grace Wahba, Yi Lin and Chenlei Leng

wahba,yilin,chenlei@stat.wisc.edu

http://www.stat.wisc.edu/~wahba, ~yilin, ~chenlei

This report has been submitted to Statistical Science as invited Comments to Mark H. Hansen and Charles Kooperberg, 'Spline Adaptation in Extended Linear Models', to appear. An abstract and two additional figures have been added. Research partly supported by NSF Grant DMS0072292 and NIH Grant EY09946.

Comments to "Spline Adaptation in Extended Linear Models", by Mark H. Hansen and Charles Kooperberg

Grace Wahba, Yi Lin and Chenlei Leng
wahba,yilin,chenlei@stat.wisc.edu
Department of Statistics
University of Wisconsin-Madison

Abstract

We thank Hansen and Kooperberg (HK) for an interesting paper discussing model selection methods in the context of Extended Linear Models. We comment on their univariate density estimation studies, which maximize the log likelihood in a low dimensional linear space. They consider spline bases for this space and consider greedy and Bayesian methods for choosing the knots. We describe a penalized log likelihood univariate density estimate, and compare the estimate to those studied by HK. Then we describe the multivariate version of our estimate, based on a Smoothing Spline ANOVA model. A randomized Generalized Approximate Cross Validation estimate for the smoothing parameters is obtained and an example is given. This represents work in progress.

1 Introduction and Thanks

The authors present greedy and Bayesian model selection frameworks for studying adaptation in the context of an extended linear model, with application to logspline density estimation and bivariate triogram regression models. We will confine our remarks to the density estimation case. The authors define the setup of their 'extended linear model' as finding $g \in G$ to maximize the log likelihood

$$l(g) = \sum_i l(g, W_i) \tag{1}$$

where G is a linear space, generally of much lower dimension than the sample size n. Generally the famous bias-variance tradeoff is controlled (most likely primarily) by the dimension of G, as well as other parameters involved in the choice of G or spline spaces in Hansen and Kooperberg (HK), the number of knots governs the dimension of the space, and the number and location of the knots are to be chosen according to several Bayesian methods and compared with a greedy method. Knot selection in the context of (1) is a difficult but not impossible task, as the authors clearly show. The authors are to be thanked for an interesting study of Bayesian knot selection methods and their comparison with a greedy knot selection method.

To contrast with the ELM approach in the paper, we will examine a penalized likelihood method for the same (log) density estimation problem. It is based on solving a variational problem in an infinite dimensional (Hilbert) space, where the problem has a Bayesian flavor, and where the solution to the variational problem is (essentially) known to lie in a particular n dimensional subspace. Then the smoothing parameter(s) are chosen by a predictive loss criteria. If the penalty functional is square integral second derivative, the n-dimensional subspace is spanned by a basis of cubic splines with knots at the observation points. At this point we can take one of several points of view. The three that are relevant to the discussion here are: (i) Solve the variational problem exactly, (ii) Find a good approximation to the solution of the variational problem, by using a representative or a random sample of the knots, instead of the complete set, when the sample size

is large and (iii) Instead of using the solution of the variational problem as the 'gold standard' as in (ii), use a greedy algorithm to choose a subset of the knots, actually, a subset of the *representers*, (Wahba (1990)), which reduce to the knots in the case of polynomial splines. This will have the effect of letting the 'wiggliness' of the solution vary where there are more observations, and/or more variable responses. Then the variational problem is solved in the greedily chosen subspace. This so-called hybrid approach was taken in Luo & Wahba (1997) in a Gaussian regression problem, using a relatively simple greedy algorithm, and, as was also found in Stone et al (1997) more knots are located near sharp features, as well as where there are more observations.

We will focus on a density estimation version of (ii) in the rest of this discussion. To carry out this program we need a criteria for the choice of the smoothing parameters appropriate for density estimation and we will use randomized $GACV$ for density estimation, (to be defined), which is a proxy for the comparative Kullback-Liebler distance of the 'truth' from the estimate. In this discussion we will first give some details for the univariate case and compare the results to Table 2 of HK. Loosely speaking, the results compare fairly favorably with all of the estimates whose MISE performance is given in Table 2 with the exception of the two largest sample sizes in the 'sharp peak' example. After commenting on these results, we will then describe some work in progress, in which the penalized likelihood estimate is extended to several dimensions via a smoothing spline ANOVA (SS-ANOVA) model. We briefly demonstrate a three dimensional result. The conceptual extension of the penalized likelihood method to higher dimensions is fairly straightforward, and the real thrust of the work is to be able to estimate densities in higher dimensions. One of the rationales behind the use of the SS-ANOVA model for density estimation in several dimensions is that the pattern of main effects and interactions has an interesting interpretation in terms of conditional dependencies, and can thus be used to fit graphical models (Darroch, Lauritzen & Speed (1980), Whittaker (1990), Jordan (1998)) nonparametrically.

2 Penalized Log Likelihood Density Estimation

Our density estimate is based on the penalized log likelihood estimate of Silverman (1982). When going to higher dimensions we will use the basic ANOVA decomposition idea in Gu (1993). Our density estimate will have compact support Ω, which will be scaled to the unit interval or the unit cube in E^d and then rescaled back after fitting. Let the density $p = e^g$ with g in some reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with square seminorm $J(g)$, where the null space of J contains the constant function and is low dimensional. Letting $x_i \in \Omega$, Silverman showed that the penalized log likelihood minimization problem: min $g \in \mathcal{H}$

$$-\frac{1}{n}\sum_{i=1}^{n} g(x_i) + \lambda J(g) \tag{2}$$

subject to the condition

$$\int_\Omega e^g = 1 \tag{3}$$

is the same as the minimizer of

$$\mathcal{I}_\lambda(g) = -\frac{1}{n}\sum_{i=1}^{n} g(x_i) + \int_\Omega e^g + \lambda J(g). \tag{4}$$

We will describe the estimate in general form so that its extension from the univariate to the multivariate case is clear. Let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\mathcal{H}_0$ is the null space of J, and let the reproducing kernel for $\mathcal{H}_1$ be $K(x, x')$. If the term $\int_\Omega e^g$ were not in (4), then (it is well known that) the minimizer of (4) would be in $\mathcal{H}^n \equiv \mathcal{H}_0 \oplus span\{\xi_i, i = 1, \cdots, n\}$, where $\xi_i(x) = K(x, x_i)$. (ξ_i is known as a representer.) We will therefore feel confident that the minimizer of (4) in $\mathcal{H}^n$ is a good approximation to the minimizer of (4) in $\mathcal{H}$. In fact, we will seek a minimizer in $\mathcal{H}^N = \mathcal{H}_0 \oplus span\{\xi_{i_r}, r = 1, \cdots N\}$ where the i_r is a representative subset chosen sufficiently large that the minimizer in $\mathcal{H}^N$ is a good approximation to the minimizer in $\mathcal{H}^n$.

In order to carry out penalized log likelihood estimation a method for choosing λ is required. We have obtained a randomized Generalized Approximate Cross Validation ($ranGACV$) estimate for λ, for density estimation. We briefly describe it here, details will be given elsewhere. Let f_λ be the estimate of the log density, and let $f_\lambda^{[-i]}(x_i)$ be the estimate with the ith observation left out. Define the ordinary leaving-out-one function as

$$V_0(\lambda) = OBS(\lambda) + D(\lambda) \tag{5}$$

where

$$OBS(\lambda) = -\frac{1}{n}\sum_{i=1}^{n} f_\lambda(x_i) \tag{6}$$

and

$$D(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[f_\lambda(x_i) - f_\lambda^{[-i]}(x_i)]. \tag{7}$$

Elsewhere (to appear) we show that $nD(\lambda)$ can be approximated by the trace of the inverse Hessian of $\mathcal{I}_\lambda$ with respect to $f_\lambda(x_i), i = 1, \cdots, n$ and that it can be estimated by a randomization technique as follows. Let $\mathcal{I}_\lambda(g, y)$ be

$$\mathcal{I}_\lambda(g, y) = -\frac{1}{n}\sum_{i=1}^{n} y_i g(x_i) + \int_\Omega e^g + \lambda J(g). \tag{8}$$

When $y = (1, \cdots, 1)'$ then (8) becomes (4). Letting f_λ^y be the minimizer of (8), $D(\lambda)$ is estimated as

$$\hat{D}(\lambda) = \frac{1}{n\sigma_\epsilon^2}\epsilon'(f_\lambda^{y+\epsilon} - f_\lambda^y) \tag{9}$$

where $y = (1, \cdots, 1)'$, ϵ is a random vector with mean 0 and covariance $\sigma_\epsilon^2 I$, and, with some abuse of notation $f_\lambda^z = (f_\lambda^z(x_1), \cdots, f_\lambda^z(x_n))'$. Several replicates in ϵ may be used for greater accuracy. Then

$$ranGACV(\lambda) = OBS(\lambda) + \hat{D}(\lambda). \tag{10}$$

Our numerical results (to appear) show that $ranGACV$ is a good proxy for the comparative Kullback Liebler distance between the density determined by f_λ and the true density.

3 The Univariate Estimate

The procedure is to start with N representers. In the one-dimensional case we choose roughly equally spaced order statistics. Fix λ large. Use a Newton Raphson iteration to estimate the coefficients of f_λ in the basis functions spanning $\mathcal{H}^N$. Evaluate $ranGACV(\lambda)$. Decrease λ and repeat, until the minimizer over λ is found. Double N and repeat. Compare the resulting estimates with N and $2N$, if they agree within a specified tolerance, stop, otherwise double N again. We tried this penalized log likelihood estimate on the examples in HK, using $\mathcal{H} = W_2^2 \equiv \{g : g, g' abs.cont., g'' \in \mathcal{L}_2\}$ and $J(g) = \int_0^1 (g''(x))^2$. In this case $\mathcal{H}_0$ is spanned by linear functions and $K(x, x') = k_2(x)k_2(x') - k_4([x - x'])$, $x \in [0, 1]$ where $[\tau]$ is the fractional part of τ and $k_m(x) = B_m(x)/x!$ where B_m is the mth Bernoulli polynomial. The estimate is a cubic spline (Wahba (1990)) with knots at the x_{i_r}. In the one dimensional case this is not the most efficient way to compute this estimate, since a B-spline basis is available given the knots, and that will lead to a sparse linear system, whereas the present representation does not. However, this representation generalizes easily to higher dimensional estimates. In our experiment the maximum allowed N was 48. We made 100 replicates of each case in Table 2 of HK, and computed the MISE in the same way as HK did, by averaging the squared difference over 5001 equally spaced quadrature points in the three intervals (for the normal, slight bimodal and sharp peak cases) of $[-5, 5], [-7, 7]$ and $[0, 12]$.

dist	sample size	MISE (pen.log.lik)	HK (Table 2(i))	Ratio (pen.log.lik/HK)
normal	50	0.01859	0.02790	0.666
	200	0.00435	0.01069	0.407
	1000	0.00071	0.00209	0.340
	10000	0.00014	0.00020	0.700
bimodal	50	0.01358	0.02502	0.543
	200	0.00372	0.00770	0.483
	1000	0.00079	0.00164	0.482
	10000	0.00011	0.00020	0.550
peak	50	0.10011	0.15226	0.657
	200	0.03045	0.03704	0.822
	1000	0.02152	0.00973	2.212
	10000	0.01624	0.00150	10.83

We note that the ratio column suggests that this estimate is among the better estimates in HK's Table 2 with the exception of the $n = 1000$ and $n = 10000$ cases for the peak example.

4 Multivariate Smoothing Spline ANOVA Density Estimation

The univariate penalized log likelihood density estimation procedure we have described can be generalized to the multivariate case in various ways. Here we describe the smoothing spline ANOVA (SS-ANOVA) model. The use of SS-ANOVA in a density estimate was suggested by Gu (1993), who also gave a method for choosing the smoothing parameter(s). It can be shown that (for the same smoothing parameters) the estimates of Gu and Silverman are mathematically equivalent,

however we found the variational problem in Silverman easier to compute. The problem in d dimensions is transformed to the d-dimensional unit cube, and $x_i = (x_{i1}, \cdots, x_{id})$. $\mathcal{H}$ will be an RKHS on the d dimensional cube which is formed as the direct sum of subspaces of the tensor product of d one dimensional RKHS's. Details of SS-ANOVA models may be found in Wahba (1990), Wahba, Wang, Gu, Klein & Klein (1995) Lin, Wahba, Xiang, Gao, Klein & Klein (2000). Letting $u = (u_1, \cdots, u_d) \in [0,1]^d$, we have

$$g(u) = \mu + \sum_{\alpha=1}^{d} g_\alpha(u_\alpha) + \sum_{\alpha \neq \beta} g_{\alpha\beta}(u_\alpha, u_\beta) + ... \tag{11}$$

where the terms satisfy averaging conditions analogous to those in ordinary ANOVA that insure identifiability, and the series may be truncated somewhere. The interesting feature of this representation of a log density is the fact that the presence or absence of interaction terms determines the conditional dependencies, that is, a graphical model see Whittaker (1990)). For example the main effects model represents independent component random variables, and if, for example $d = 3$ and the g_{23} and g_{123} terms are missing then the second and third component random variables are conditionally independent, given the first.

Let $\tilde{\mathcal{H}}$ be the d-fold tensor product of W_2^2 and let $\mathcal{H}$ be the subspace of $\tilde{\mathcal{H}}$ consisting of the direct sum of subspaces containing the terms retained in the expansion. (They are orthogonal in $\tilde{\mathcal{H}}$) We have $\int_0^1 g_\alpha(u_\alpha) du_\alpha = 0$, and so forth. The penalty functional $J(g)$ of (4) becomes $J_\theta(g)$ where the θ represents a vector of (relative) weights on separate penalty terms for each of the components of (11). As before $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\mathcal{H}_0$ is the (low dimensional) null space of J_θ. Let $K_\theta(x, x'), x, x' \in [0,1]^d$ be the reproducing kernel for $\mathcal{H}_1$ where θ has been incorporated into the norm on $\mathcal{H}_1$. (See Wahba (1990) Chapter 10.) Let $\xi_i(x) = \xi_{i\theta}(x) = K_\theta(x, x_i)$. The same arguments hold as in the one dimensional case, and we seek a minimizer of (4) (with $J = J_\theta$) in $\mathcal{H}^N = \mathcal{H}_0 \oplus span\{\xi_{i_r\theta}, r = 1, \cdots, N\}$, and λ and θ are chosen using the $ranGACV$ of (10).

We will give a three dimensional example, essentially to demonstrate that the calculations are possible and the $ranGACV$ reasonable in higher dimensions. The SS-ANOVA model for this example contained only the main effects and two factor interactions, and we had altogether 6 smoothing parameters, parameterized in a convenient manner (details to appear elsewhere). For fixed smoothing parameters λ, θ the coefficients in the expansion in $\mathcal{H}^N$ are obtained via a Newton-Raphson iteration. In this case integrations over $[0,1]^3$ are required, and we used quadrature formulae based on the hyperbolic cross points, see (Novak & Ritter (1996), Wahba (1978)). These quadrature formulae seem particularly appropriate for SS-ANOVA models and make high dimensional quadrature feasible. Then the $ranGACV$ was minimized over smoothing parameters via a 6-dimensional downhill simplex calculation.

The underlying true density used in the example is $p(x) = 0.5N(\mu_1, \Sigma) + 0.5N(\mu_2, \Sigma)$, where $\mu_1 = (0.25, 0.25, 0.25)$, $\mu_2 = (0.75, 0.75, 0.75)$,

$$\Sigma = \begin{pmatrix} 10 & 0 & 10 \\ 0 & 20 & 30 \\ 10 & 30 & 80 \end{pmatrix}^{-1} = \begin{pmatrix} 0.14 & 0.06 & -0.04 \\ 0.06 & 0.14 & -0.06 \\ -0.04 & -0.06 & 0.04 \end{pmatrix}.$$

(This density has a non zero three factor interaction which is not in our two factor model.) In this example the sample size was $n = 1000$. $N = 40$ and the 40 representers were randomly chosen from among the n possibilities. The $N = 80$ estimate was essentially indistinguishable from the $N = 40$ case. (Note that the smoothing parameters will not generally be the same in the two cases.) Figure 1 gives cross sections of the true density, and Figure 2 gives the SS-ANOVA penalized log likelihood

estimate. Figure 3 compares the $ranGACV$ and the CKL ($CKL(\lambda) = -\int_\Omega f_{\lambda,\theta}(u)p(u)du$) as a function of iteration number in a downhill simplex minimization of the $ranGACV$. For this report we have added Figures 4 and 5 which are the same as Figures 1 and 2 but from a different viewing angle.

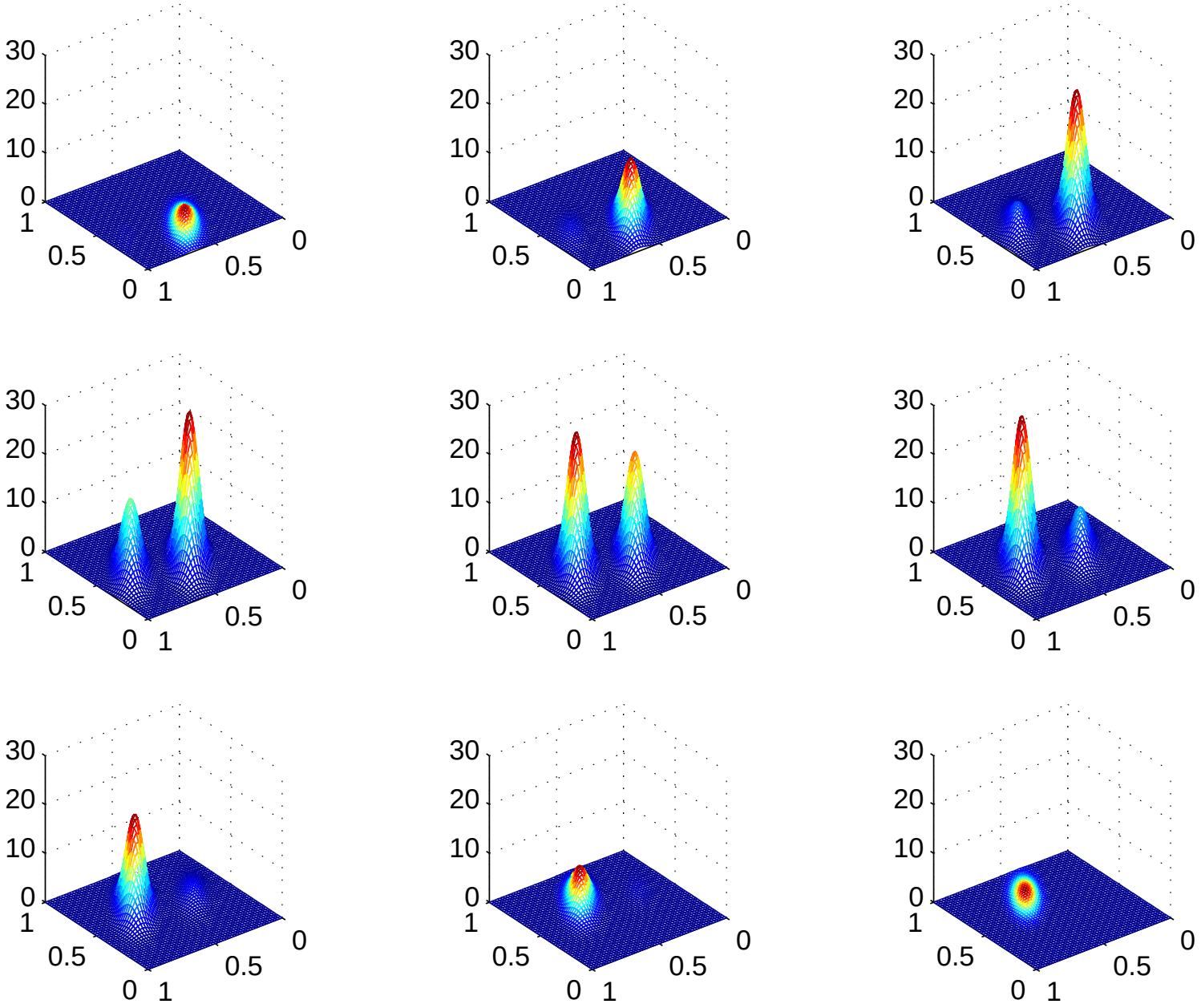

Figure 1: The true density. $x_1 = .1, ..., .9.$ is fixed in the plots, left to right, then top to bottom.

5 Closing Remarks

We have compared a penalized likelihood density estimate with *ranGACV* to choose the smoothing parameter(s) to the greedy density estimate and the Bayesian estimates in ELM models considered by HK. Fairly favorable results were obtained except in the highest n peak cases. We have shown that these penalized likelihood estimates can be extended to the multivariate case (work in progress). It remains to develop tests to allow the construction of graphical models from the SS-ANOVA estimates in higher dimensions.

We would be interested in knowing to what extent the Bayesian model selection methods can be incorporated in ELM estimates for the multivariate case.

Splines of various flavors have been widely adopted in many statistical problems. It is interesting to compare the various flavors and we are pleased to compliment the authors and contribute to the discussion.

Acknowledgment

Research supported in part by NIH Grant RO1 EYO9946 and NSF Grant DMS0072292.

References

Darroch, J., Lauritzen, S. & Speed, T. (1980), 'Markov and log linear interaction models for contingency tables', *Ann. Statist.* **8**, 522–539.

Gu, C. (1993), 'Smoothing spline density estimation: A dimensionless automatic algorithm', *Journal of the American Statistical Association* **88**, 495–504.

Jordan, M. (1998), *Learning in Graphical Models*, Kluwer.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. & Klein, B. (2000), 'Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV', *Ann. Statist.* **28**, 1570–1600.

Luo, Z. & Wahba, G. (1997), 'Hybrid adaptive splines', *J. Amer. Statist. Assoc.* **92**, 107–114.

Novak, E. & Ritter, K. (1996), 'High dimensional integration of smooth functions over cubes', *Numer. Math.* **75**, 79–97.

Silverman, B. (1982), 'On the estimation of a probability density function by the maximum penalized likelihood method', *Ann. Statist.* **10**, 795–810.

Wahba, G. (1978), Interpolating surfaces: High order convergence rates and their associated designs, with applications to x-ray image reconstruction, Technical Report 523, Dept. of Statistics, University of Wisconsin, Madison, WI.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1995), 'Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy', *Ann. Statist.* **23**, 1865–1895. Neyman Lecture.

Whittaker, J. (1990), *Graphical Models in Applied Mathematical Multivariate Statistics*, Wiley.

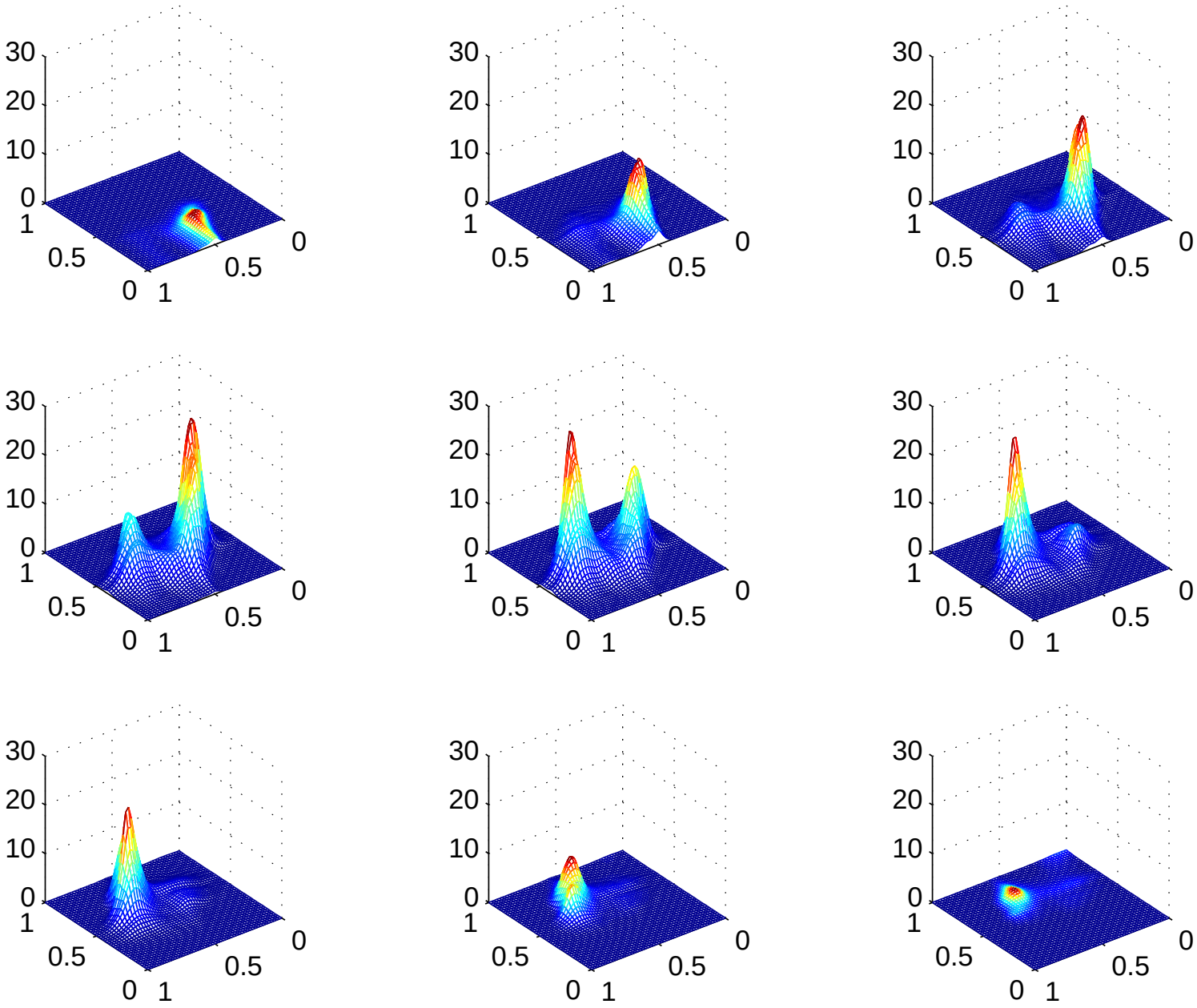

Figure 2: The estimated density. $x_1 = .1, ..., .9$ is fixed in the plots, left to right, then top to bottom.

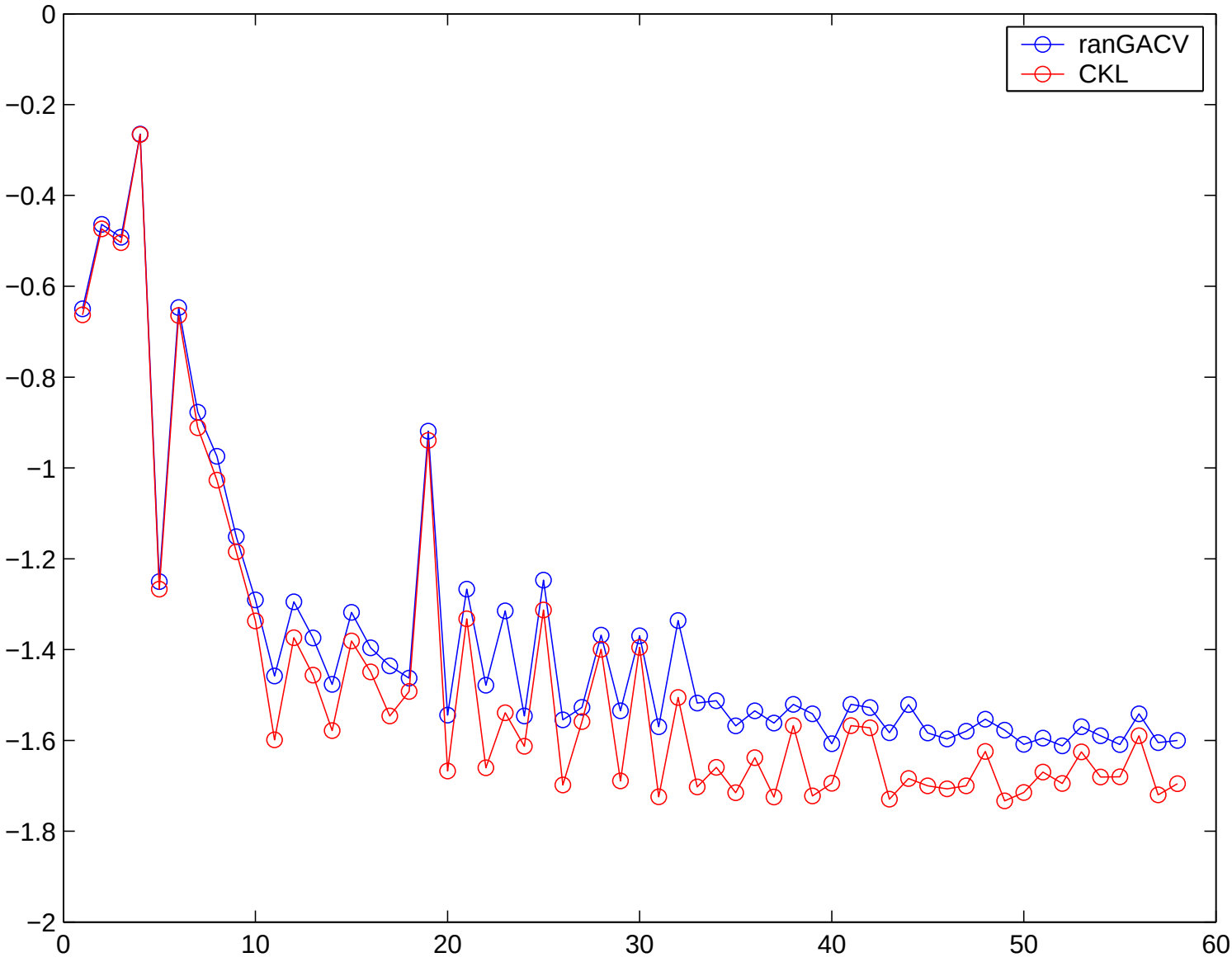

Figure 3: The $ranGACV$ and the CKL compared. The horizontal axis is iteration number, using the downhill simplex method. The $ranGACV$ is minimized and the $ranGACV$ and CKL are computed at the minimizer at each step.

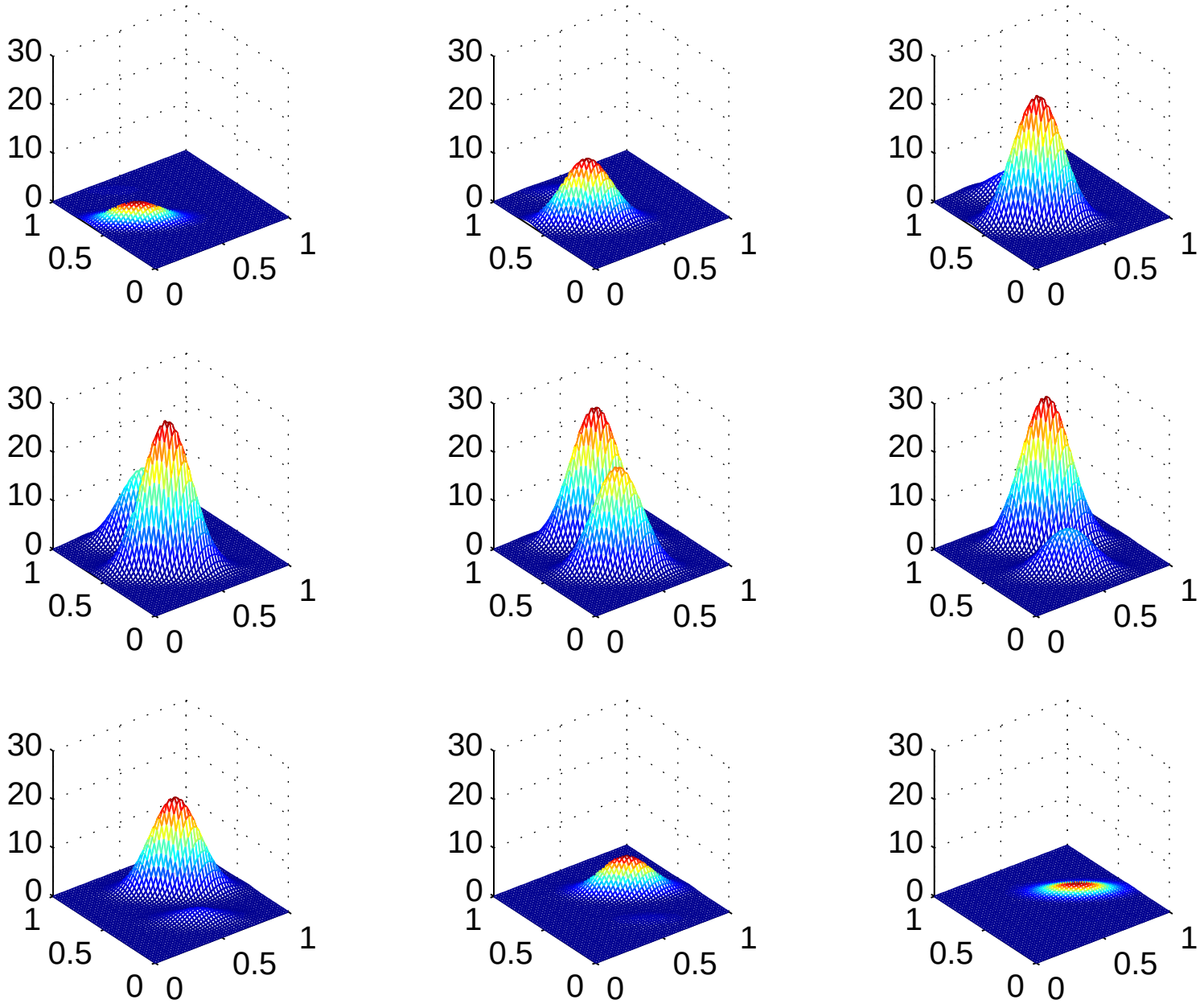

Figure 4: The true density. Same as Figure 1 but from a different viewing angle.

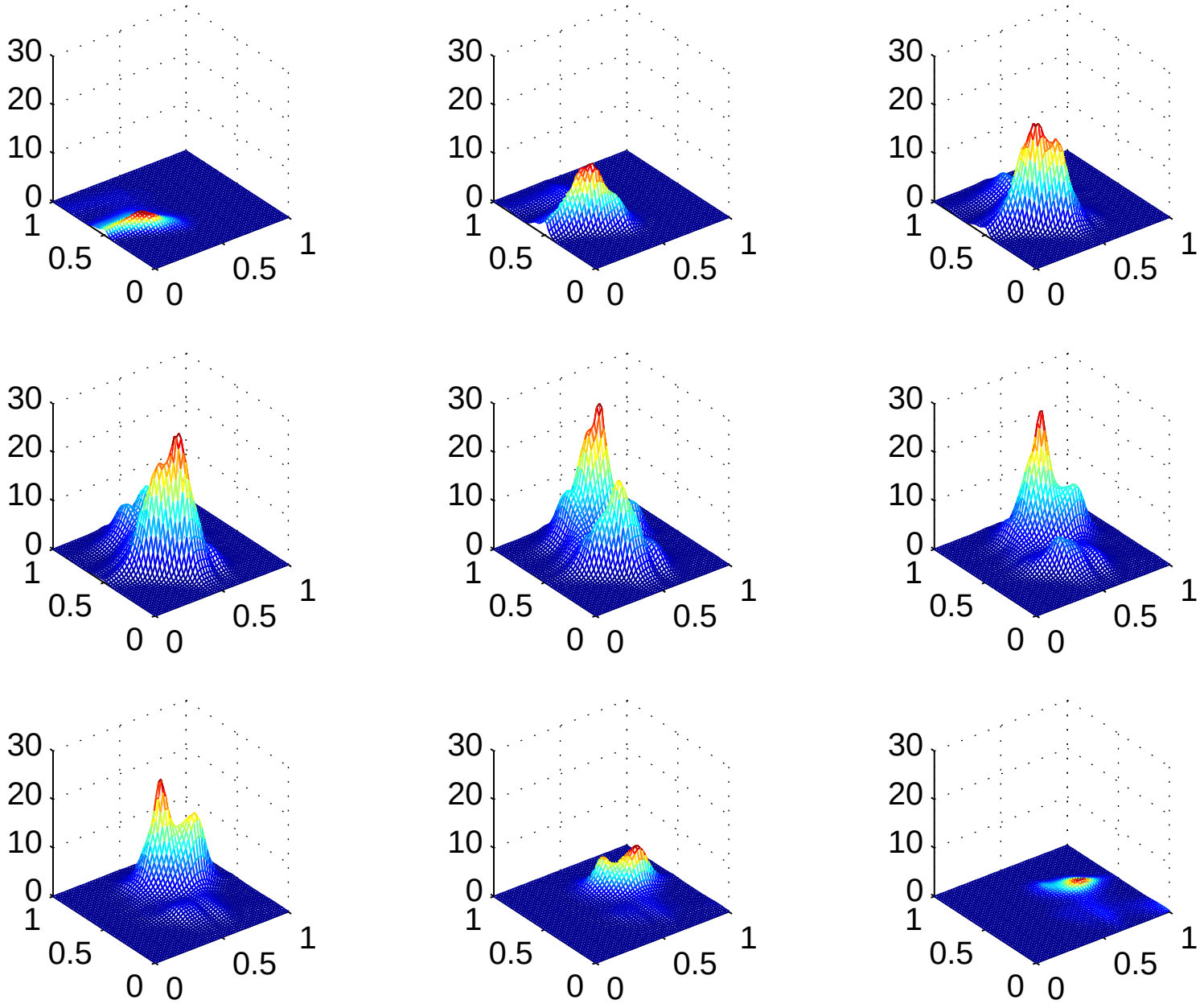

Figure 5: The estimated density. Same as Figure 2 but from the same viewing angle as Figure 4