
STABILIZED CROSS-VALIDATION OF SMOOTHNESS IN DENSITY DECONVOLUTION

David Kent

Cornell University

Abstract:

We consider density estimation under measurement error with the Smoothness-Penalized Deconvolution (SPeD) estimator. The estimator has a tuning parameter regulating the smoothness of the estimate, and proper choice of this parameter is critical for forming good estimates. We derive the cross-validation choice of tuning parameter for the SPeD estimator, but it performs very poorly. We introduce a stabilized cross-validation (SCV) criterion which unbiasedly estimates the mean integrated squared error (MISE) for a smaller sample size, and use asymptotic arguments to obtain an appropriate tuning parameter from this stabilized criterion. We show that the SCV is a strongly consistent estimator of the MISE, and that it is the minimum variance unbiased estimator of the MISE. In a simulation study, we show that the SCV approach outperforms the previously recommended choice of tuning parameter in nearly all settings, and in a majority of the settings, SPeD with the SCV outperforms the classic deconvoluting kernel estimator with its recommended choice of tuning parameter.

Key words and phrases: Penalty parameter, deconvolution, measurement error,
cross-validation, density estimation

Contents

1	Introduction	3
1.1	The Estimator	6
2	Cross-Validation of the Error	11
2.1	Cross-Validation in Density Deconvolution	12
2.2	Stabilized Cross-Validation	14
2.3	Relationship between the CV and SCV Criteria	18
3	Theory for $SCV_n(\alpha, m)$	20
4	Simulations	23
5	Appendix	29
5.1	Proofs of Results	29
5.2	Lemmas	38

1. Introduction

Suppose X has unknown pdf f , E has known pdf g , and X and E are independent. Then $Y = X + E$ has pdf $h(y) = g * f(y) = \int f(t)g(y - t) dt$. Suppose we observe a sample $Y_1, \dots, Y_n \stackrel{iid}{\sim} h$, and we wish to estimate the density f of X . This is the deconvolution problem. The recent Smoothness-Penalized Deconvolution (SPeD) estimator (Kent, Ruppert, 2024; Yang, Kent, et al., 2021; Yang, Apley, et al., 2020) depends critically on a penalty parameter α that regulates the smoothness of the estimate. If α is too large, the estimate is too smooth, and interesting features of f may be absent; if α is too small, the estimate is too wiggly, and contains many features not present in f .

In Yang, Apley, et al., 2020, the authors choose α by a *SURE* criterion, similar to Stein’s unbiased risk estimation. Specifically, they consider a new hypothetical sample Y_1^0, \dots, Y_n^0 identically distributed as and independent from the original sample, and denote by \hat{h} and \hat{h}^0 density estimates of h from the original and hypothetical samples, respectively. Letting \hat{f}_α denote the estimator fit to the original sample, they choose α by minimizing an estimate of the SURE criterion $\text{SURE}(\alpha) = \mathbb{E} \|\hat{h}^0 - g * \hat{f}_\alpha\|^2$, where the expectation is over both the original and hypothetical samples, which determine \hat{f}_α and \hat{h}^0 , respectively.

The authors note that this SURE approach occasionally chooses an α that is far too small, leading to a very rough estimate. In our simulations in Section 4—see, for example, Figure 4—we will see that these too-small choices of α can lead to catastrophically poor estimates, which poses difficulties in practice. This is not a huge surprise in retrospect: the deconvolution problem is ill-posed (Kent, Ruppert, 2024, Section 2) in the sense that for two estimates \hat{f} and \hat{f}' , the values $g * \hat{f}$ and $g * \hat{f}'$ can be very close even when \hat{f} and \hat{f}' are very different. But since $\text{SURE}(\alpha)$ measures the error in $g * \hat{f}_\alpha$, this criterion may be insensitive to choice of α even when the error in \hat{f}_α is quite sensitive. This observation aligns with the observation in Yang, Apley, et al., 2020 that the problem occurs when there is a “wide range of nearly-optimal $[\alpha]$ values,” indicating that the problem occurs when the error criterion is insensitive to the α values.

Cross-validation has been suggested for choosing the bandwidth in the deconvolution kernel density estimator (Stefanski, Carroll, 1990; Hesse, 1999; Youndjé, Wells, 2008), though the intuition behind it is less transparent in deconvolution than in classical density estimation. In Section 2.1 we translate the intuitive justification for cross-validation into the deconvolution problem, and derive the cross-validation criterion for the SPeD estimator. We find that it suffers from extremely poor performance, regularly

choosing α too small. In fact, many realizations of $CV(\alpha)$ are monotonically increasing on any reasonable interval we might optimize over (see an example plotted in Figure 1); in these cases, the minimizer of $CV(\alpha)$ is determined by the largely-arbitrary lower bound of the optimization interval.

In classical density estimation, the cross-validation criterion consists of a non-random term that closely approximates the integrated variance, and a random term that estimates the integrated squared bias (IBS); the same is true in deconvolution. In Hall, Marron, Park, 1992, explaining the large variability of cross-validated bandwidths in the classical case, the authors note that “estimates of integrated squared bias tend to suffer from high variance when bandwidths of the order $h^{-1/5}$ are considered”—but of course this is the order of the optimal bandwidth, so it is exactly the order of the bandwidths we should be searching in cross-validation. We see a similar phenomenon in practice: on the scale of the penalty parameter minimizing the MISE, the component of $CV(\alpha)$ estimating the bias has high variance.

This insight makes us wish we were considering α of a larger order. In Section 2.2 we do exactly that, introducing a procedure we call *stabilized cross-validation* (SCV). In SCV, we adjust the CV criterion to estimate the error for a sample of size $m \ll n$; this adjusted criterion consists of a non-random term closely approximating the integrated variance (for a sample

of size m), plus *the same random term* (to a very close approximation) that occurs in the CV criterion approximating the IBS. For a given α , the variance of the IBS is essentially unchanged, but when we deal with the SCV criterion we are considering α of a larger order than with the CV criterion due to the increased integrated variance term. For $m \approx n^{1/2}$, the SCV criterion is much better behaved, and its minimum has much lower variance. After minimizing the SCV criterion to estimate α'_m suitable for a sample of size m , we use an idea from Hall, 1990 to parlay this into an estimate $\hat{\alpha}_n$ suitable for a sample of size n . Namely, we use *a priori* knowledge of the asymptotic rate of the optimal α_n , say $\alpha_n \sim \kappa b_n$ for a known sequence b_n and unknown constant κ , and form $\hat{\alpha}_n = (b_n/b_m)\alpha'_m$.

1.1 The Estimator

The estimator we address is a slight modification of the SPeD addressed in Kent, Ruppert, 2024, which will simplify its analysis. In their definition of the estimator, they begin with a preliminary density estimate h_n of h ; here, we replace h_n by the empirical distribution, which amounts to replacing $\int[g * u]h_n$ by $\frac{1}{n} \sum_{j=1}^n g * u(Y_j)$. For a similar approach in a different problem, see Wager, 2014. Here we present three equivalent representations of this estimator (cf. Theorem 2 of Kent, Ruppert, 2024). We will denote

the Fourier transform by an overset twiddle, so that $\tilde{f}(t) = \mathbb{E}[e^{-itX}] = \int e^{-itx} f(x) dx$. Let $P_n = \frac{1}{n} \sum_{j=1}^n \delta_{Y_j}$ denote the empirical distribution, and $\tilde{P}_n(t) = \frac{1}{n} \sum_{j=1}^n e^{-itY_j}$ its Fourier transform. We assume throughout that \tilde{g} is Lebesgue-almost-everywhere non-zero and $\int f(x)^2 dx < \infty$.

Definition 1. The *smoothness penalized deconvolution* or *SPeD* estimator is defined by the following equivalent characterizations.

- (i) $f_n^\alpha = \operatorname{argmin}_{u \in H_{2,m}(\mathbb{R})} \|g * u\|^2 - \frac{2}{n} \sum_{j=1}^n g * u(Y_j) + \alpha \|u^{(\nu)}\|^2$,
- (ii) $f_n^\alpha(x) = \frac{1}{n} \sum_{i=1}^n \varphi_\alpha(x - Y_i)$, and
- (iii) $\tilde{f}_n^\alpha(t) = \tilde{\varphi}_\alpha(t) \tilde{P}_n(t)$,

where $\tilde{\varphi}_\alpha(t) = \frac{\overline{\tilde{g}(t)}}{|\tilde{g}(t)|^2 + \alpha |t|^{2\nu}}$ and $\varphi_\alpha(x) = \lim_{r \rightarrow \infty} \frac{1}{2\pi} \int_{-r}^r e^{itx} \tilde{\varphi}_\alpha(t) dt$ is its inverse Fourier transform. The Sobolev space of functions with square-integrable m th derivative is denoted by $H_{2,m}(\mathbb{R})$. The proof that these are equivalent representations is essentially identical to the proof of Theorem 2(i)-(iii) in Kent, Ruppert, 2024.

Proposition 1. *The following facts about the estimator will be useful later.*

- (i) $\int \operatorname{Var}(f_n^\alpha(x)) dx = \frac{1}{n} \|\varphi_\alpha\|^2 - \frac{1}{n} \|\varphi_\alpha * g * f\|^2$
- (ii) $\|f_n^\alpha\|^2 = \frac{1}{n} \|\varphi_\alpha\|^2 + \frac{1}{n^2} \sum_{j=1}^n \sum_{k \neq j} \varphi_\alpha * \varphi_\alpha(x + Y_j - Y_k)$

(iii) $\mathbb{E}\|f_n^\alpha\|^2 = \frac{1}{n}\|\varphi_\alpha\|^2 + \frac{n-1}{n}C_{\alpha,f}$. $C_{\alpha,f}$ does not depend on n .

For the following two facts, assume $\int |f'(x)|^2 dx < \infty$.

(iv) (Normal errors) Assume $g(x) = (2\pi)^{-1}e^{-x^2/2}$. If $\alpha_n = \kappa \frac{\log n}{n}$, then

$$\text{MISE}(\alpha_n) = O([\log n]^{-1}).$$

(v) (Cauchy errors) Assume $g(x) = \frac{1}{\pi(1+x^2)}$. If $\alpha_n = \kappa \frac{(\log n)^2}{n}$, then

$$\text{MISE}(\alpha_n) = O([\log n]^{-2}).$$

Proof of Proposition 1. Part (i): First, since $f_n^\alpha = \frac{1}{n} \sum_{j=1}^n \varphi_\alpha(x - Y_j)$ is a sum of independent summands, $\int \text{Var}(f_n^\alpha(x)) dx = \frac{1}{n} \int \text{Var}(\varphi_\alpha(x - Y_1)) dx$.

Writing the variance in terms of the moments,

$$\int \text{Var}(\varphi_\alpha(x - Y_1)) dx = \int \mathbb{E}[\varphi_\alpha(x - Y_1)^2] dx - \int \mathbb{E}[\varphi_\alpha(x - Y_1)]^2 dx$$

For the first,

$$\begin{aligned} \int \mathbb{E}[\varphi_\alpha(x - Y_1)^2] dx &= \int \int \varphi_\alpha(x - y)^2 h(y) dy dx \\ &= \int \int \varphi_\alpha(x - y)^2 dx h(y) dy \\ &= \int \|\varphi_\alpha\|^2 h(y) dy \\ &= \|\varphi_\alpha\|^2. \end{aligned}$$

For the second, note that $\mathbb{E}[\varphi_\alpha(x - Y_1)] = \varphi_\alpha * h(x) = \varphi_\alpha * g * f(x)$, so

$\int \mathbb{E}[\varphi_\alpha(x - Y_1)]^2 dx = \|\varphi_\alpha * g * f\|^2$. Combining these gives the desired result.

Part (ii):

$$\begin{aligned}
\|f_n^\alpha\|^2 &= \int \left(\frac{1}{n} \sum_{j=1}^n \varphi_\alpha(x - Y_j) \right)^2 dx \\
&= \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n \int \varphi_\alpha(x - Y_j) \varphi_\alpha(x - Y_k) dx \\
&= \frac{1}{n^2} \sum_{j=1}^n \int \varphi_\alpha(x - Y_j)^2 dx + \frac{1}{n^2} \sum_{j=1}^n \sum_{k \neq j} \int \varphi_\alpha(x - Y_j) \varphi_\alpha(x - Y_k) dx \\
&= \frac{1}{n^2} \sum_{j=1}^n \int \varphi_\alpha(x)^2 dx + \frac{1}{n^2} \sum_{j=1}^n \sum_{k \neq j} \int \varphi_\alpha(x) \varphi_\alpha(x - (Y_k - Y_j)) dx \\
&= \frac{1}{n} \|\varphi_\alpha\|^2 + \frac{1}{n^2} \sum_{j=1}^n \sum_{k \neq j} \varphi_\alpha * \varphi_\alpha(Y_k - Y_j)
\end{aligned}$$

Part (iii): Here, we just take the expectation of the expression in Part (ii).

The only random term is in the summand and its expectation $C_{\alpha,f}$ is identical for all j and k , so the value of the double sum is $n(n-1)C_{\alpha,f}$.

Parts (iv)- (v): Define $f^\alpha(x) = \varphi_\alpha * h(x)$ as in Kent, Ruppert, 2024; this is exactly the expected value of the estimator in Definition 1. We have $\text{MISE}(\alpha) = \int (f^\alpha - f)^2 + \int \text{Var}(f_n^\alpha)$. By part (i), $\int \text{Var}(f_n^\alpha) \leq \frac{1}{n} \|\varphi_\alpha\|^2$, which can be bounded in the following way for $\alpha < M$. By dropping $\alpha|t|^{2\nu}$ in the denominator, it can be seen that $|\tilde{\varphi}_\alpha(t)|^2 \leq |\tilde{g}(t)|^{-2}$. Applying the inequality $a + b \geq 2\sqrt{ab}$ to the denominator yields $|\tilde{\varphi}_\alpha(t)|^2 \leq \frac{1}{2} \alpha^{-\frac{1}{2}} |t|^{-\nu}$. Then $\int |\tilde{\varphi}_\alpha(t)|^2 dt \leq \frac{1}{\alpha} \int \min\{M|\tilde{g}(t)|^{-2}, \frac{1}{4}|t|^{-2\nu}\} dt$. The integral is finite because $|\tilde{g}(t)|^{-2}$ is integrable near zero, while $|t|^{-2m}$ is integrable away from

zero. Thus $\int \text{Var}(f_n^\alpha) \leq \frac{C}{n\alpha}$, with C depending on g , ν , and M but not on f .

For normal errors, the integrated squared bias is bounded in Kent, Ruppert, 2024 Lemma 5(i), so that $\int (f^\alpha - f)^2 \leq C_1[\nu \log(\nu^{-1}\alpha^{-\frac{1}{\nu}})]^{-1}$. Combining this with the bound for the variance and $\alpha_n = \kappa \frac{\log n}{n}$ we have

$$\begin{aligned} \text{MISE}(\alpha_n) &\leq C_1[\nu \log(\nu^{-1}\alpha^{-\frac{1}{\nu}})]^{-1} + C_2(n\alpha_n)^{-1} \\ &= C_1\nu^{-1}[-\log \nu - \nu^{-1}(\log \kappa + \log \log n - \log n)]^{-1} + C_3(\log n)^{-1} \\ &\sim C_4 \log n \end{aligned}$$

For Cauchy errors, by Kent, Ruppert, 2024 Lemma 5(ii), $\int (f^\alpha - f)^2 \leq C_1[\nu \log(\nu^{-1}\alpha^{-\frac{1}{2\nu}})]^{-2}$. Combining this with the bound for the variance and $\alpha_n = \kappa \frac{(\log n)^2}{n}$ we have

$$\begin{aligned} \text{MISE}(\alpha_n) &\leq C_1\nu^{-2}[\log(\nu^{-1}\alpha_n^{-\frac{1}{2\nu}})]^{-2} + C_2(n\alpha_n)^{-1} \\ &= C_1\nu^{-2}[\log \nu + (2\nu)^{-1}(\log \kappa + 2 \log \log n - \log n)]^{-2} + C_3(\log n)^{-2} \\ &\sim C_4(\log n)^{-2} \end{aligned}$$

□

Note that $0 \leq \|\varphi_\alpha * g * f\|^2 \leq \|f\|^2$ for all α , so Proposition 1 (i) implies that $\int \text{Var}(f_n^\alpha(x)) dx = \frac{1}{n}\|\varphi_\alpha\|^2 + O(\frac{1}{n})$.

2. Cross-Validation of the Error

Cross-validation, proposed in Rudemo, 1982, is a data-based method for estimating a constant shift of the integrated squared error $\|\hat{f}_n - f\|^2$ in an estimate \hat{f}_n . The cross-validation criterion has the same expectation as $\|\hat{f}_n - f\|^2 - \|f\|^2 = \|\hat{f}_n\|^2 - 2 \int \hat{f}_n f$. Since the constant shift $\|f\|^2$ does not depend on the estimate \hat{f}_n , this criterion can be used in place of $\|\hat{f}_n - f\|^2$ for choosing among possible \hat{f}_n , typically by setting a tuning parameter. In the following section we derive the CV criterion for SPeD.

Following that, we introduce the related SCV criterion, which is estimates the error in the estimator fit to a smaller sample size $m \ll n$; in this context, the particular integrated error $\|\hat{f}_m - f\|^2$ makes less sense, so we will focus on the mean integrated squared error (MISE). Both the CV and SCV criteria can be regarded as an unbiased estimator of a vertical shift of the MISE.

Definition 2. For the SPeD estimator applied to a sample of size n , define

$$\begin{aligned} R(\alpha, n) &= \mathbb{E}\|f_n^\alpha - f\|^2 - \|f\|^2 \\ &= \mathbb{E}\|f_n^\alpha\|^2 - 2 \int \mathbb{E}f_n^\alpha f \end{aligned}$$

2.1 Cross-Validation in Density Deconvolution

Least-squares cross-validation for classical (error-free) density estimation can be motivated by noticing that in the integrated squared error of an estimate \hat{f} , i.e. $\|\hat{f} - f\|^2 = \|\hat{f}\|^2 - 2 \int \hat{f}f + \|f\|^2$, the first term is known and the last term is constant, so we need only estimate the second term. In classical density estimation, the data X_i have pdf f , and contemplating the integral $\int \hat{f}f$ we might attempt to approximate this by a sample average $\frac{1}{n} \sum \hat{f}(X_i)$. This turns out to differ in expectation from $\int \hat{f}f$ because \hat{f} depends on X_i ; this is corrected by using $\frac{1}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i)$, where \hat{f}_{-i} refers to the estimator fit on the data excluding observation i . The resulting expression $\|\hat{f}\|^2 - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i)$ is the (unbiased) cross-validation estimate of the error (Scott, Terrell, 1987).

If we are estimating a density f , but observe data with a different density h (as we are in the deconvolution problem), this line of thinking grinds to a halt when we think of approximating $\int \hat{f}f$ by a sample average – we have no samples with density f that we can use to that end. But we can extend the idea in the following way. Assume the data Y_i have density $h = \mathcal{C}f$ for some linear operator \mathcal{C} with adjoint \mathcal{D} (satisfying $\int u[\mathcal{C}v] = \int [\mathcal{D}u]v$); assume also that \mathcal{D} is injective and \hat{f} is in the range of \mathcal{D} . Then we manipulate that integral like so: $\int \hat{f}f = \int [\mathcal{D}\mathcal{D}^{-1}\hat{f}]f = \int [\mathcal{D}^{-1}\hat{f}][\mathcal{C}f] =$

$\int[\mathcal{D}^{-1}\hat{f}]h$. Now we can move forward and approximate $\int[\mathcal{D}^{-1}\hat{f}]h$ by the sample average $\frac{1}{n} \sum_{i=1}^n \mathcal{D}^{-1}\hat{f}_{-i}(Y_i)$.

In deconvolution, we have $h(y) = \mathcal{C}f(y) = \int f(t)g(y-t)dt$, in which case \mathcal{D} is defined by $\mathcal{D}u(y) = \int g(t-x)u(t)dt$. The Fourier transform of $\mathcal{D}u$ is $\tilde{g}\tilde{u}$, so our assumption that \tilde{g} is a.e. non-zero guarantees that \mathcal{D} is injective. For the SPeD estimator, we can see that f_n^α is in the range of \mathcal{D} by observing that $\tilde{f}_n^\alpha = \tilde{g}\tilde{\varphi}_\alpha\tilde{P}_n/\tilde{g}$. Now $\tilde{u} = \tilde{\varphi}_\alpha\tilde{P}_n/\tilde{g}$ is square-integrable, so its Fourier inverse u exists; thus $f_n^\alpha = \mathcal{D}u$. The operator \mathcal{D}^{-1} can be defined explicitly through Fourier transforms.

Definition 3. $\mathcal{D}^{-1}u(x) = \frac{1}{2\pi} \int e^{itx}\tilde{u}(t)/\overline{\tilde{g}(t)}dt$ for u in the range of \mathcal{D} .

Finally, we can state the cross-validation estimate of the error. It may be useful to note that the same form is obtained if one follows the approach of Youndjé, Wells, 2008; there, the authors substitute an unbiased estimator of $|\tilde{h}(t)|^2$ in $R(\alpha, n)$ after taking Fourier transforms.

Definition 4. The *cross-validation criterion* is

$$\text{CV}_n(\alpha) = \|f_n^\alpha\|^2 - \frac{2}{n} \sum_{j=1}^n \mathcal{D}^{-1}f_{n,-j}^\alpha(Y_j) \quad (2.1)$$

2.2 Stabilized Cross-Validation

The cross-validation criterion performs poorly in practice. The value of $\text{CV}_n(\alpha)$ is very unstable near $\alpha_n = \operatorname{argmin} R(\alpha, n)$, and the minimizer of $\text{CV}_n(\alpha)$ is often much too small; furthermore, realizations of $\text{CV}_n(\alpha)$ are often monotonically increasing in an extremely large interval around α_0 , and in these cases the minimum is determined in practice by the arbitrary left endpoint. To remedy this, we introduce a related criterion $\text{SCV}_n(\alpha, m)$ that estimates $R(\alpha, m)$ with m possibly different from the sample size n ; when $m \ll n$, $\text{SCV}_n(\alpha, m)$ does not suffer from the deficiencies of $\text{CV}_n(\alpha)$. Notice that $\text{SCV}_n(\alpha, n) = \text{CV}_n(\alpha)$.

Definition 5. The *stabilized cross-validation criterion* is

$$\text{SCV}_n(\alpha, m) = \frac{(m-1)n}{m(n-1)} \|f_n^\alpha\|^2 + \frac{n-m}{m(n-1)} \|\varphi_\alpha\|^2 - \frac{2}{n} \sum_{j=1}^n \mathcal{D}^{-1} f_{n,-j}^\alpha(Y_j) \quad (2.2)$$

Just as $\text{CV}_n(\alpha)$ is an unbiased estimate of $R(\alpha, n)$, this stabilized criterion $\text{SCV}_n(\alpha, m)$ is unbiased estimate of the error for a sample of size m .

Proposition 2. $\mathbb{E} \text{SCV}_n(\alpha, m) = R(\alpha, m)$

Proof of Proposition 2. First, note that $\mathbb{E} \int f_n^\alpha f = \int [\varphi_\alpha * h] f$. Now, consider the sum: $\mathbb{E} \frac{1}{n} \sum_j \mathcal{D}^{-1} f_{n,-j}^\alpha(Y_j) = \mathbb{E} \mathcal{D}^{-1} \varphi_\alpha(Y_2 - Y_1)$. Recalling that

$h = \mathcal{C}f$, we write $\mathbb{E}\mathcal{D}^{-1}\varphi_\alpha(Y_2 - Y_1) = \int h(t) \int [\mathcal{D}^{-1}\varphi_\alpha](x - t)\mathcal{C}f(x) dx dt = \int h(t) \int \varphi_\alpha(x - t)f(x) dx dt = \int [\varphi_\alpha * h]f$, so that $\mathbb{E}_n \sum_j \mathcal{D}^{-1}f_{n,-j}^\alpha(Y_j) = \mathbb{E} \int f_n^\alpha f = \mathbb{E} \int f_m^\alpha f$ (since this quantity does not depend on n), bringing us to $\mathbb{E}\hat{R}(\alpha, m) = \frac{(m-1)n}{m(n-1)}\mathbb{E}\|f_n^\alpha\|^2 + \frac{n-m}{m(n-1)}\|\varphi_\alpha\|^2 - 2\mathbb{E} \int f_m^\alpha f$. Finally, use Proposition 1 (iii) twice (once with n and once with m) to find that $\frac{(m-1)n}{m(n-1)}\mathbb{E}\|f_n^\alpha\|^2 + \frac{n-m}{m(n-1)}\|\varphi_\alpha\|^2 = \frac{1}{m}\|\varphi_\alpha\|^2 + \frac{m-1}{m}C(\alpha) = \mathbb{E}\|f_m^\alpha\|^2$, where $C(\alpha)$ is the same constant as in Proposition 1 (iii). Bringing this all together, we have $\mathbb{E}\hat{R}(\alpha, m) = \mathbb{E}\|f_m^\alpha\|^2 - 2\mathbb{E} \int f_m^\alpha f = R(\alpha, m)$, as needed. \square

The obvious objection here is: “OK, maybe you’ve given me a nice way to choose α when my sample size is m . But my sample size is n !” To make the leap from the minimum of $R(\alpha, m)$ to the minimum of $R(\alpha, n)$, we take an asymptotic point of view inspired by Hall, 1990. Suppose that the sequence $\alpha_n^* = \operatorname{argmin}_{\alpha \in \mathcal{A}_n} R(\alpha, n)$ satisfies $\alpha_n^* \sim \kappa^* b_n$ for some known sequence b_n . Then we have that $\kappa^* \sim \alpha_m^*/b_m$, so $\alpha_n^* \sim (b_n/b_m)\alpha_m^*$. Now we estimate α_m^* by choosing the α'_m that minimizes $\operatorname{SCV}_n(\alpha, m)$, and approximate α_n^* by $\hat{\alpha}_n = (b_n/b_m)\alpha'_m$. For example, when the measurement errors are Gaussian and f has a square-integrable first derivative, $b_n = n^{-1} \log n$ (Proposition 1 (iv)); in that case we obtain α'_m by minimizing $\operatorname{SCV}_n(\alpha, m)$, and use $\hat{\alpha}_n = \frac{m \log n}{n \log m} \alpha'_m$. The intuition behind SCV is pictured in Figure 1. The procedure can be summarized as follows.

Definition 6. Suppose that $\alpha_n^* = \operatorname{argmin}_{\alpha \in \mathcal{A}_n} R(\alpha, n)$ satisfies $\alpha_n^* \sim \kappa^* b_n$ for known sequence b_n , and choose $m \ll n$. The *stabilized cross-validation procedure* (SCV procedure) is defined by

1. Compute $\alpha'_m = \operatorname{argmin}_{\alpha \in \mathcal{A}_n} \operatorname{SCV}_n(\alpha, m)$.
2. Set $\hat{\alpha}_n = (b_n/b_m)\alpha'_m$.
3. Use $f_n^{\hat{\alpha}_n}$ to estimate f .

The sequence b_n depends on smoothness assumptions for the target density f ; for instance with normal errors, $b_n = n^{-1} \log n$ attains optimal rates for f with one square-integrable derivative. But if we restrict to densities with k square-integrable derivatives, we should use $b_n = n^{-1}(\log n)^k$, letting the smoothness parameter decay more slowly (Kent, Ruppert, 2024). In Definition 6, we simply made a minimal smoothness assumption. We also propose an adaptive procedure for normal errors, in which we choose k from the data, motivated by the fact that $\alpha_n^* \approx \kappa n^{-1}(\log n)^k$, so $\log \alpha_n^* \approx \log \kappa + k \log \log n - \log n$.

Definition 7. Choose $m_i \ll n$, $i = 1, \dots, n_m$ and $m \ll n$. The *adaptive stabilized cross-validation procedure* (ASCV procedure) for normal errors is defined by

1. Compute $\alpha'_{m_i} = \operatorname{argmin}_{\alpha \in \mathcal{A}_n} \operatorname{SCV}_n(\alpha, m_i)$ for $i = 1, \dots, n_m$.
2. Fit the regression model $\log \alpha'_{m_i} = \beta_0 + \beta_1 \log \log m_i - \log m_i + \varepsilon_i$, and set $\hat{b}_n = n^{-1}(\log n)^{\hat{\beta}_1}$.
3. Set $\hat{\alpha}_n = (\hat{b}_n/\hat{b}_m)\alpha'_m$.
4. Use $f_n^{\hat{\alpha}_n}$ to estimate f .

Remark 1 (Choosing m). Definitions 6 and 7 both require choice of $m \ll n$.

The SCV procedures both involve a statistical estimate and an asymptotic approximation: the value α'_m is a statistical estimate of the optimal α_m^* , and we then use the asymptotic approximation $\alpha_n^* \approx (b_n/b_m)\alpha_m^*$. The smaller m is with respect to n , the more well-behaved are the estimates α'_m ; on the other hand, the asymptotic approximation $\alpha_n^* \approx (b_n/b_m)\alpha_m^*$ relies on both n and m being large. So the trade-off is that when m is very small, the estimate $\alpha'_m \approx \alpha_m^*$ is quite good, but the asymptotic approximation $\alpha_n^* \approx (b_n/b_m)\alpha_m^*$ may be poor, while when m is closer to n , the situation is reversed, and the α'_m is poorly behaved (cf. Figure 1) while the approximation is quite good.

In Theorem 1, we prove uniform consistency for the $\operatorname{SCV}_n(\alpha, m)$ criterion estimating $R(\alpha, m)$ under Assumption (A2) relating m to the rate of convergence $r_n = R(\alpha_n^*, n)$ of the optimal $\alpha_n^* \sim \kappa b_n$. For normal errors,

2.3 Relationship between the CV and SCV Criteria

we show in Example 1 that the assumption is equivalent to $m = n^{\frac{1-\delta}{2}}$ with $0 < \delta < 1$. We recommend choosing m large, but still satisfying this condition; in all the simulations here, we use $m = n^{\frac{1}{2}-10^{-3}}$. In practice, $m = \sqrt{n}$ seems to work just as well, but we did notice degradation of performance for larger powers like $m = n^{\frac{3}{5}}$ and above.

When choosing m_i in the ASCV procedure, we recommend using equally-spaced values close to, but not exceeding m ; we found that $m_1 = n^{\frac{1}{2}-\frac{1}{20}}$ through $m_{n_m} = n^{\frac{1}{2}-10^{-3}}$ worked well.

2.3 Relationship between the CV and SCV Criteria

In view of Proposition 1 (i), define

$$\overline{\text{IV}}(\alpha, n) = \frac{1}{n} \|\varphi_\alpha\|^2.$$

This approximation to the integrated variance satisfies $\overline{\text{IV}}(\alpha, n) = \int \text{Var}(f_n^\alpha(x)) dx + O(\frac{1}{n})$. For reasons we will explain momentarily, define

$$\widehat{\text{IBS}}(\alpha) = \frac{1}{n^2} \sum_{j=1}^n \sum_{k \neq j} \varphi_\alpha * \varphi_\alpha(Y_j - Y_k) - \frac{2}{n} \sum_{j=1}^n \mathcal{D}^{-1} f_{n,-j}^\alpha(Y_j).$$

By Proposition 1 (ii), we may write

$$\text{CV}_n(\alpha) = \overline{\text{IV}}(\alpha, n) + \widehat{\text{IBS}}(\alpha).$$

By Proposition 2 with $m = n$,

$$\begin{aligned}\mathbb{E} \text{CV}_n(\alpha) &= \|f_n^\alpha - f\|^2 - \|f\|^2 \\ &= \int \text{Var}(f_n^\alpha(x)) dx + \int (\mathbb{E} f_n^\alpha(x) - f(x))^2 dx - \|f\|^2,\end{aligned}$$

so

$$\mathbb{E} \widehat{\text{IBS}}(\alpha) = \int (\mathbb{E} f_n^\alpha(x) - f(x))^2 dx - \|f\|^2 + O\left(\frac{1}{n}\right).$$

Thus $\widehat{\text{IBS}}(\alpha)$ estimates a vertical shift of the integrated squared bias and furthermore is the only random term in $\text{CV}_n(\alpha)$. Exactly this observation is made for cross-validation in classical density estimation in Hall, Marron, Park, 1992.

Using Proposition 1 (ii) again, we can write

$$\begin{aligned}\text{SCV}_n(\alpha, m) &= \overline{\text{IV}}(\alpha, m) + \frac{m-1}{mn(n-1)} \sum_{j=1}^n \sum_{k \neq j} \varphi_\alpha * \varphi_\alpha(Y_j - Y_k) \\ &\quad - \frac{2}{n} \sum_{j=1}^n \mathcal{D}^{-1} f_{n,-j}^\alpha(Y_j) \\ &\approx \frac{n}{m} \overline{\text{IV}}(\alpha, n) + \widehat{\text{IBS}}(\alpha),\end{aligned}$$

showing that to a good approximation, the $\text{SCV}_n(\alpha, m)$ and $\text{CV}_n(\alpha)$ criteria contain the same estimates of the bias and variance, with the variance rescaled for the smaller sample size. This shows why the $\text{SCV}_n(\alpha, m)$ is better behaved: it is because $\text{Var}(\widehat{\text{IBS}}(\alpha))$ decreases as α increases, and when $m \ll n$, the minimum of $R(\alpha, m)$ (and $\text{SCV}_n(\alpha, m)$) occurs at larger α .

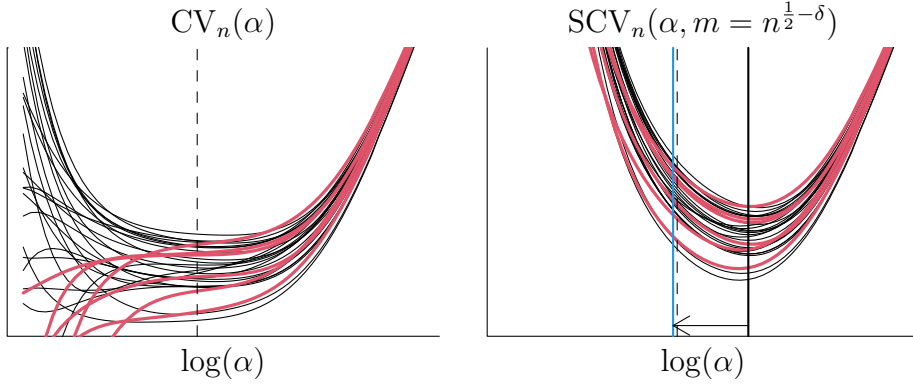


Figure 1: Thirty realizations of the CV and SCV functions under Density 2 (see Section 4) with Gaussian errors, $n = 1000$, and 10% measurement error; SCV is computed with $m = n^{\frac{1}{2}-\delta}$ with $\delta = 10^{-3}$. Realizations plotted in red have $CV(\alpha)$ monotonically increasing over the whole domain; the $SCV_n(\alpha, m)$ functions computed from those same realizations are plotted in red. The dashed and vertical lines are $\alpha_n^* = \operatorname{argmin} R(\alpha, n)$ and α_m^* , respectively. The blue vertical line is $(b_n/b_m)\alpha_m^*$.

3. Theory for $SCV_n(\alpha, m)$

The criterion $SCV_n(\alpha, m)$ can be expressed as a constant (depending on α and m) plus a U-statistic. From this, it can be shown that $SCV_n(\alpha, m)$ is the minimum variance unbiased estimator of $R(\alpha, m)$.

Proposition 3.

$$SCV_n(\alpha, m) = \frac{1}{m} \|\varphi_\alpha\|^2 + \frac{1}{\binom{n}{2}} \sum_{1 \leq j < k \leq n} \theta_\alpha(Y_j - Y_k), \quad (3.3)$$

where $\theta_\alpha(x) = \frac{m-1}{m} \varphi_\alpha * \varphi_\alpha(x) - 2\mathcal{D}^{-1}\varphi_\alpha(x)$. Furthermore, for $k < 2\nu - 1$

the kernel $\theta_\alpha(x)$ has the bounds

$$\sup_x |\theta_\alpha^{(k)}(x)| \leq \frac{C_{M,k}}{\alpha} \quad (3.4)$$

for all $\alpha < M$, where the constant $C_{M,k}$ depends on M and k .

Proposition 4. *If $\hat{R}_n(\alpha, m) = \hat{R}(\alpha, m; Y_1, \dots, Y_n)$ is an unbiased estimator of $R(\alpha, m)$, then $\text{Var}(\hat{R}_n(\alpha, m)) \geq \text{Var}(\text{SCV}_n(\alpha, m))$*

Now we proceed to some consistency results for $\text{SCV}_n(\alpha, m)$ in estimating $R(\alpha, m)$.

Assumptions. Fix g, f , and $0 < \lambda < \mu$.

(A1) There are sequences b_n and r_n such that $\inf_{\lambda b_n \leq \alpha \leq \mu b_n} R(\alpha, n) = r_n$.

(A2) The sequence m_n is chosen such that $(r_{m_n} b_{m_n})^2 n = n^\delta$ with $\delta > 0$.

Under Assumptions (A1) and (A2), we can show that with probability 1, the stabilized cross-validation criterion converges uniformly to the shifted mean integrated squared error that it is estimating.

Theorem 1. *Assume (A1) and (A2). Then with probability 1,*

$$\sup_{\lambda b_m \leq \alpha \leq \mu b_m} |\text{SCV}_n(\alpha, m) - R(\alpha, m)| = o(r_m).$$

A simple corollary says that the minimizer of $\text{SCV}_n(\alpha, m)$ asymptotically minimizes the true risk function $R(\alpha, m)$. Note that the function $R(\cdot, m)$ here still refers to the risk of the estimator using a fixed penalty parameter.

Corollary 1. *Assume (A1) and (A2). Then with probability 1,*

$$R(\alpha'_m, m) = R(\alpha_m^*, m) + o(r_m),$$

so

$$\frac{R(\alpha'_m, m)}{R(\alpha_m^*, m)} \sim 1.$$

As a bonus, an observation of Chacón, Tenreiro, 2024 leads directly from Corollary 1 to a strongly consistent estimator of $\|f\|^2$.

Corollary 2. *Assume (A1) and (A2). Then with probability 1,*

$$-\text{SCV}_n(\alpha'_m, m) = \|f\|^2 + O(r_m).$$

Since $r_m \rightarrow 0$, it follows that $-\text{SCV}_n(\alpha'_m, m) \rightarrow \|f\|^2$ with probability 1.

Example 1 (Normal errors). If g is a normal density and we assume f has one square-integrable derivative, then by Proposition 1 (iv) $\alpha_n = \kappa b_n = \kappa \frac{\log n}{n}$ is optimal, and yields rate of convergence $r_n = (\log n)^{-1}$. Then $(b_m r_m)^2 = \kappa^2 m^{-2}$, so Theorem 1 holds if we use $m = n^{\frac{1-\delta}{2}}$ for $0 < \delta < 1$.

Example 2 (Cauchy errors). If g is a Cauchy density and we assume f has one square-integrable derivative, then by Proposition 1 (v) $\alpha_n = \kappa b_n = \kappa \frac{(\log n)^2}{n}$ is optimal, and yields rate of convergence $r_n = (\log n)^{-2}$. Then $(b_m r_m)^2 = \kappa^2 m^{-2}$, and again if we use $m = n^{\frac{1-\delta}{2}}$ for $0 < \delta < 1$ then Theorem 1 holds.

4. Simulations

To assess the practical performance of the SCV approach, we test it on a battery of densities with a variety of features. In Marron, Wand, 1992, the authors propose fifteen densities which exhibit a variety of challenging features for curve estimation, all represented as mixtures of normal densities. We use here the first eight, which are the unimodal and bimodal examples, all pictured in Figure 2. Density 1 is the standard normal; densities 2 & 3 are slightly and strongly skewed, respectively; density 4 has $\frac{1}{3}$ of its mass in a sharp peak at the mean; density 5 has $\frac{9}{10}$ of its mass in a sharp peak at its mean (or from another point of view, $\frac{1}{10}$ of its mass spread far from the mean); densities 6 and 7 are bimodal with slightly and very separated modes, respectively; density 8 is bimodal with modes of different width.

We aim to estimate each of these target example densities under measurement error. We simulate $Y_i = X_i + E_i$, $i = 1, \dots, n$ with the X_i drawn independently from the target density f and $E_i \sim N(0, \sigma_E^2)$ with $\sigma_E^2 = \frac{1}{9} \text{Var}(X_1)$, which results in $\frac{\text{Var}(X_i)}{\text{Var}(Y_i)} = \frac{1}{10}$, so that 10% of the variance in our data is from the measurement error. Simulations are performed with $n = 100, 500$, and 1000. For each f and n , we compute $\alpha_n^* = \text{argmin}_\alpha \mathbb{E} \|f_n^\alpha - f\|^2$, the MISE-optimal smoothing parameter; we will call this the “optimal” choice since among fixed choices of α it has the

smallest MISE. For each iteration, we choose $\hat{\alpha}_{\text{SCV}}$ by the adaptive SCV (ASCV) procedure described in Definition 7 with $m = n^{\frac{1}{2}-10^{-3}}$ and five equally spaced m_i between $m_1 = n^{\frac{1}{2}-\frac{1}{20}}$ and $m_5 = m$ (see Remark 1 for suggestions about choosing m); we also choose $\hat{\alpha}_{\text{SURE}}$ by the SURE procedure of Yang, Apley, et al., 2020. We fit the estimator f_n^α and compute the ISE $\|f_n^\alpha - f\|^2$ for each of $\alpha = \alpha_n^*$, $\hat{\alpha}_{\text{ASCV}}$, and $\hat{\alpha}_{\text{SURE}}$; we also fit the classical deconvoluting kernel estimator (DKDE) using the data-based bandwidth choice recommended by Delaigle, Gijbels, 2004 and compute its ISE. For each density f and sample size n we estimate the MISE from $n_{\text{sim}} = 10^4$ iterations.

To compare the ASCV, SURE, and DKDE procedures, we compare the proportional increase in MISE over the optimal α_n^* , for example we look at $100 \times (\text{MISE}_{\text{ASCV}} / \text{MISE}_{\text{optimal}} - 1)$ for the ASCV procedure. These values are displayed in Table 1. In 22 of the 24 simulation settings, our proposed ASCV procedure outperforms the SURE procedure. The SPeD estimator with our ASCV procedure outperformed the DKDE with plug-in bandwidth in 16 of the 24 settings. Another interesting observation is that the ASCV procedure's worst case yields a MISE 81% larger than the optimal choice of parameter; SURE and DKDE have worst cases with MISE 670% and 254% larger than optimal, respectively.

In assessing the SURE procedure, the authors saw that it occasionally “selects too small a [penalty parameter],” and in their Figure 3 they show an example of a catastrophically small penalty parameter, where the estimate has many narrow modes, though the target density is unimodal (Yang, Apley, et al., 2020). These simulations bear this out—in Figure 5, the small penalty parameters are apparent in the long left tails of the SURE histograms, and in Figure 6, we can see that when the SURE estimate has extremely large error, this is typically when it has selected a very small penalty parameter. The ASCV procedure does not have this same catastrophic failure mode. It is therefore worth quantifying: how extreme is each procedure’s “poor” estimates? We’ll consider e.g. $\text{ISE}_{\text{ASCV}} / \text{ISE}_{\text{Optimal}}$, so that an ASCV estimate is poor if its error is much larger than the error of the estimate using α_n^* . We then consider the 99th percentile of this quantity, plotted for three target densities in Figure 4; density 2 was the worst case for both SURE and ASCV, density 1 was the worst case for DKDE, and density 5 was the best case for both ASCV and SURE. The results corroborate the comments of Yang, Apley, et al., 2020: the SURE procedure’s poor estimates may be *much* worse than optimal, even 200 times larger. On the other hand, the ASCV and DKDE procedure’s poor estimates are more pedestrian, around 8-16 times larger.

In Table 1, the negative value for ASCV in target density 5, $n = 1000$ may arouse suspicion; the fact that this number is negative means that the ASCV choice of tuning parameter resulted in a *lower* MISE than the what we have called the optimal choice of tuning parameter. First, recall that our notion of optimality involved a fixed, non-random tuning parameter; α_n^* is, for a given f and n , a fixed value that minimizes the MISE. There are certainly data-dependent choices of α which could give lower MISE; for example $\hat{\alpha}_n^* = \operatorname{argmin}_{\alpha} \|f_n^{\alpha} - f\|^2$ minimizes the error for each realization, and has MISE smaller than α_n^* and all other choices of α . Since ASCV is data-dependent, it is plausible that it can result in MISE lower than α_n^* . Testing the hypothesis that $\text{MISE}_{\text{ASCV}} < \text{MISE}_{\text{Optimal}}$ in our simulations results in a t -statistic of $t = 28.78$, so we can conclude that in this one simulation setting, the ASCV choice is indeed better than the fixed value α_n^* that minimizes the MISE.

Acknowledgements

The work on this paper would not have been possible without the guidance, support, and encouragement of my PhD advisor David Ruppert.

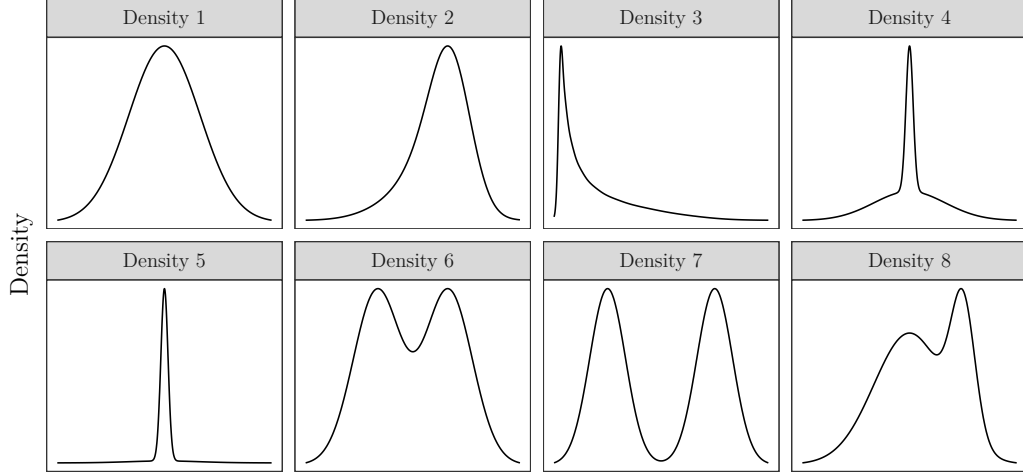


Figure 2: *The first eight target densities of Marron, Wand, 1992 which we test against in simulations.*

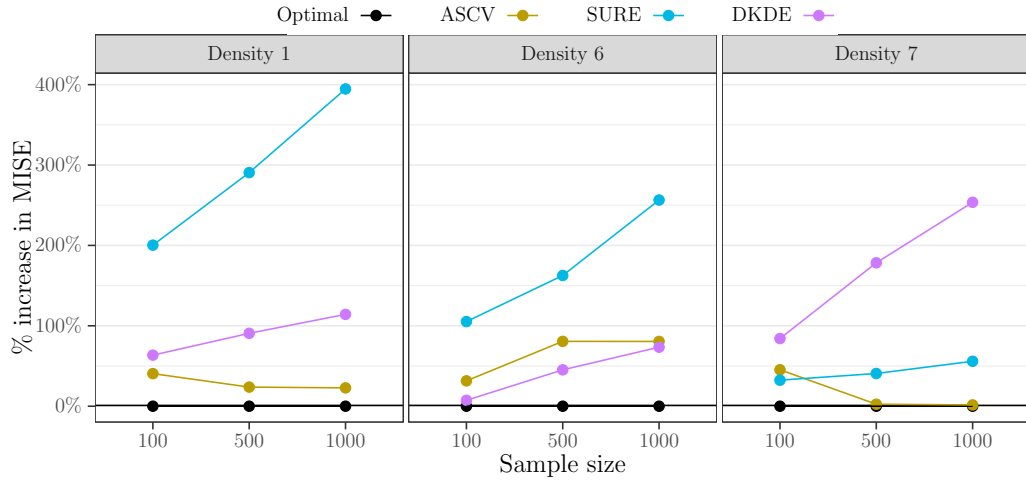


Figure 3: *Based on $n_{sim} = 10^4$ simulations, the percent increase in MISE over the MISE-optimal SPeD. For example, the ASCV values are $(MISE_{ASCV}/MISE_{Optimal} - 1) \times 100\%$.*

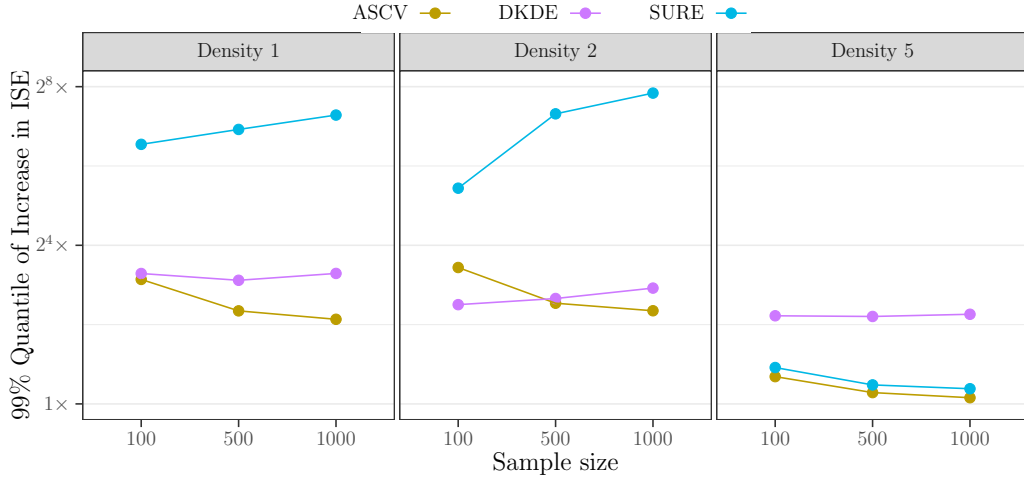


Figure 4: Based on $n_{sim} = 10^4$ simulations. The 0.99-quantile of increase in ISE over the MISE-optimal penalty. For example, the ASCV values are the 0.99-quantile of $ISE_{ASCV}/ISE_{Optimal}$. In other words: benchmarked against the MISE-optimal parameter, if we have done much worse than usual, how much worse have we done? Note the logarithmic y-axis.

		Target density							
Method		1	2	3	4	5	6	7	8
$n = 100$	ASCV	41	49	13	13	15	32	45	17
	SURE	200	145	23	8*	24	105	32*	80
	DKDE	64	43	16	12	106	7	84	3
$n = 500$	ASCV	24	23	9	5	1	81	2	48
	SURE	291	396	11	9	9	163	41	127
	DKDE	91	74	27	23	164	45	178	23
$n = 1000$	ASCV	23	23	8	1	-2	81	2	57
	SURE	395	670	11	3	6	256	56	181
	DKDE	114	100	34	20	186	73	254	39

Table 1: For the target densities shown in Figure 2, the percent increase in MISE over the MISE-optimal SPeD. For each target density/sample size, the best method (incurring the smallest increase in MISE) is marked in bold. In the two settings where the SURE procedure outperforms the ASCV procedure, the value is marked with an asterisk.

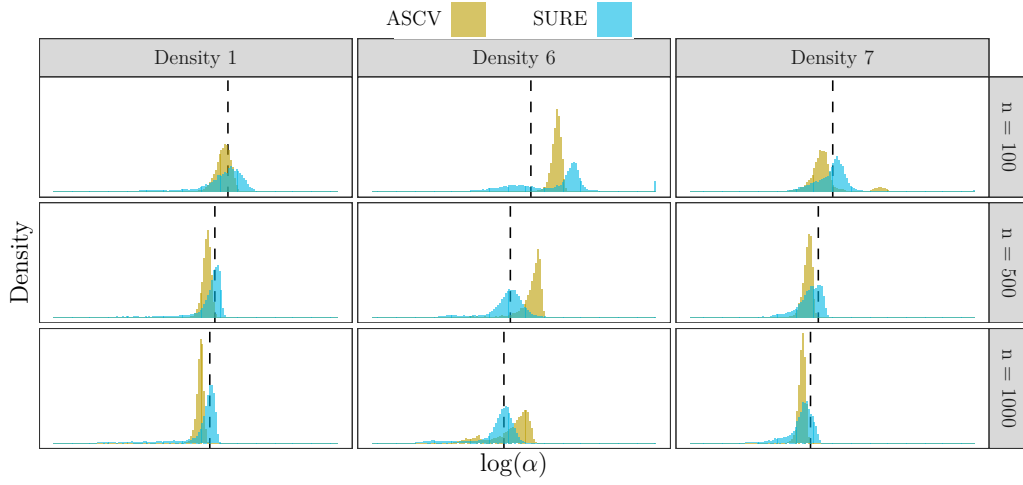


Figure 5: Histograms of α_{ASCV} and α_{SURE} in the $n_{\text{sim}} = 10^4$ simulations. Vertical dashed line marks the MISE-optimal value α_n^* in each setting. The only panel here in which SURE has lower MISE than ASCV is Density 7, $n = 100$.

5. Appendix

5.1 Proofs of Results

Proof of Proposition 3. Using Proposition 1 (ii) and rearranging,

$$\begin{aligned} \text{SCV}_n(\alpha, m) &= \frac{1}{m} \|\varphi_\alpha\|^2 \\ &\quad + \frac{2}{\binom{n}{2}} \sum_{1 \leq j < k \leq n} \left[\frac{m-1}{m} \varphi_\alpha * \varphi_\alpha - 2\mathcal{D}^{-1} \varphi_\alpha \right] (Y_j - Y_k). \end{aligned}$$

Now, the Fourier transform of $\theta_\alpha(x)$ is

$$\begin{aligned} \tilde{\theta}_\alpha(t) &= \frac{m-1}{m} \frac{|\tilde{g}(t)|^2}{(|\tilde{g}(t)|^2 + \alpha|t|^{2\nu})^2} - \frac{2}{|\tilde{g}(t)|^2 + \alpha|t|^{2\nu}} \\ &= \frac{m-1}{m} \frac{|\tilde{g}(t)|^2}{(|\tilde{g}(t)|^2 + \alpha|t|^{2\nu})^2} - \frac{2|\tilde{g}(t)|^2 + 2\alpha|t|^{2\nu}}{(|\tilde{g}(t)|^2 + \alpha|t|^{2\nu})^2} \\ &= -\frac{\frac{m+1}{m} |\tilde{g}(t)|^2 + 2\alpha|t|^{2\nu}}{(|\tilde{g}(t)|^2 + \alpha|t|^{2\nu})^2}, \end{aligned}$$

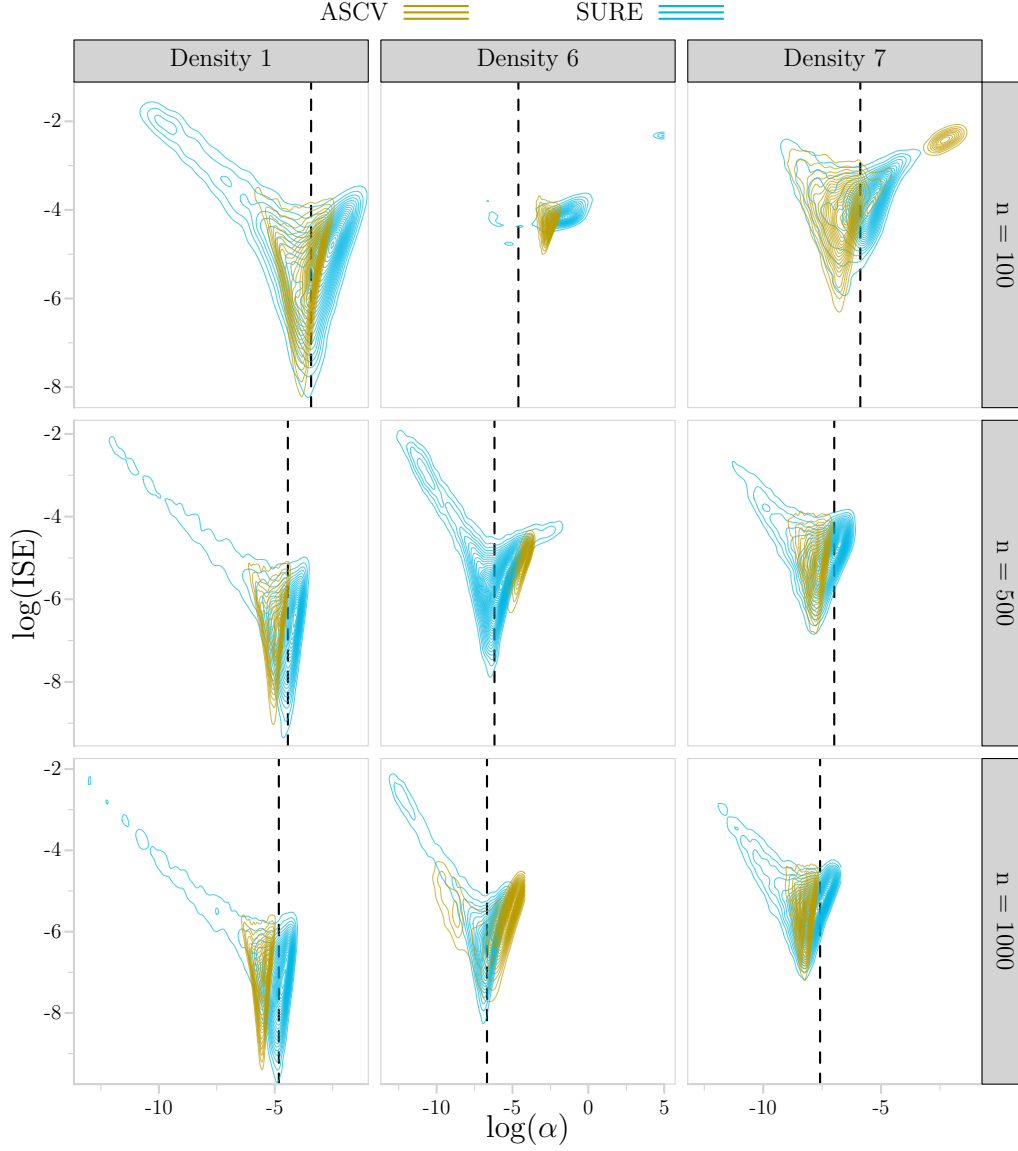


Figure 6: Contour plots of the probability density of $(\log \alpha, \log \text{ISE})$ for ASCV and SURE in the $n_{\text{sim}} = 10^4$ simulations. Vertical dashed line marks the MISE-optimal value α_n^* in each setting.

and then $|\tilde{\theta}_\alpha(t)| \leq \frac{2}{|\tilde{g}(t)|^2 + \alpha|t|^{2\nu}}$, so that

$$\begin{aligned}\theta_\alpha^{(k)}(x) &\leq \frac{1}{2\pi} \int \frac{|t|^k}{|\tilde{g}(t)|^2 + \alpha|t|^{2\nu}} dt \\ &= \frac{1}{\pi} \int_0^\infty \frac{t^k}{|\tilde{g}(t)|^2 + \alpha t^{2\nu}} dt.\end{aligned}$$

Now we integrate this bound near and away from zero. Choose $\varepsilon > 0$ s.t.

$\tilde{g}(t) \geq c > 0$ for all $0 \leq t \leq \varepsilon$, and bound as

$$\begin{aligned}\theta_\alpha^{(k)}(x) &\leq \frac{1}{\pi} \int_0^\varepsilon \frac{t^k}{|\tilde{g}(t)|^2 + \alpha t^{2\nu}} dt + \frac{1}{\pi} \int_\varepsilon^\infty \frac{t^k}{|\tilde{g}(t)|^2 + \alpha t^{2\nu}} dt \\ &\leq \frac{1}{\pi} \int_0^\varepsilon \frac{t^k}{|\tilde{g}(t)|^2 + \alpha t^{2\nu}} dt + \frac{1}{\alpha\pi} \int_\varepsilon^\infty t^{k-2\nu} dt \\ &\leq C_k + C'_k \alpha^{-1} \\ &= C_{k,M} \alpha^{-1}\end{aligned}$$

□

Proof of Proposition 4. The basic idea that follows is similar to Theorem 3 of Lee, 1990. The estimand can be written $R(\alpha, m) = \int \int \theta_\alpha(y_1 - y_2) h(y_1) h(y_2) dy_1 dy_2$, with kernel $\theta_x(x) = \frac{1}{m} \|\varphi_\alpha\|^2 + \frac{m-1}{m} \varphi_\alpha * \varphi_\alpha(x) - 2\mathcal{D}^{-1} \varphi_\alpha(x)$, so $R(\alpha, m)$ is a functional of degree 2 defined for densities in $\mathcal{H} = \{h = g * f | f \in \mathcal{F}\}$, with $\mathcal{F} = \{f \in L_2 | \int f = 1, f \geq 0\}$; the U-statistic in Proposition 3 is a symmetric, unbiased estimate formed with this same kernel $\theta_\alpha(x)$.

To prove the minimum variance assertion, suppose we knew that the symmetric unbiased estimator was unique; i.e. if $\hat{U} = U(Y_1, \dots, Y_n)$ is sym-

metric and $\mathbb{E}\hat{U} = R(\alpha, m)$, then $U(y_1, \dots, y_n) = \frac{1}{\binom{n}{2}} \sum_{1 \leq j < k \leq n} \theta_\alpha(y_j - y_k)$ (we will prove this momentarily). Let $\hat{R} = \hat{R}(y_1, \dots, y_n)$ be any unbiased estimator of $R(\alpha, m)$; form a symmetric estimator by averaging $\hat{R}(y_1, \dots, y_n)$ over all permutations of the arguments: $\hat{R}^{[n]}(y_1, \dots, y_n) = \sum \frac{1}{n!} \hat{R}(y_{i_1}, y_{i_2}, \dots, y_{i_n})$. This is also unbiased. Since the symmetric unbiased estimator is unique, $\hat{R}^{[n]}(y_1, \dots, y_n) = \text{SCV}_n(\alpha, m)$. Now apply Cauchy-Schwarz:

$$\begin{aligned} (\text{SCV}_n(\alpha, m))^2 &= \left(\hat{R}^{[n]}(y_1, \dots, y_n) \right)^2 \\ &= \left(\sum \frac{1}{n!} \hat{R}(y_{i_1}, y_{i_2}, \dots, y_{i_n}) \right)^2 \\ &\leq \sum \left(\frac{1}{n!} \right)^2 \sum \left(\hat{R}(y_{i_1}, y_{i_2}, \dots, y_{i_n}) \right)^2 \\ &= \frac{1}{n!} \sum \left(\hat{R}(y_{i_1}, y_{i_2}, \dots, y_{i_n}) \right)^2 \end{aligned}$$

Taking expectations,

$$\begin{aligned} \mathbb{E}[(\text{SCV}_n(\alpha, m))^2] &\leq \frac{1}{n!} \sum \mathbb{E} \left[\left(\hat{R}(y_{i_1}, y_{i_2}, \dots, y_{i_n}) \right)^2 \right] \\ &= \mathbb{E} \left[\left(\hat{R}(y_{i_1}, y_{i_2}, \dots, y_{i_n}) \right)^2 \right] \end{aligned}$$

Since $\mathbb{E} \text{SCV}_n(\alpha, m) = \mathbb{E} \hat{R}$, $\text{Var}(\text{SCV}_n(\alpha, m)) \leq \text{Var}(\hat{R})$.

Now we show that the symmetric unbiased estimator is unique. It suffices to show that for symmetric kernels $\psi(y_1, y_2, \dots, y_n)$, if

$$\int \cdots \int \psi(y_1, \dots, y_n) h(y_1) \cdots h(y_n) dy_1 \cdots dy_n = 0$$

for all $h \in \mathcal{H}$, then $\psi(y_1, \dots, y_n) = 0$ for almost all $y_1, \dots, y_n \in \mathbb{R}$. Recall that if $h \in \mathcal{H}$, then $h = g * f$ for $f \in \mathcal{F}$. So we may define $K(x_1, \dots, x_n) = \int \cdots \int \psi(x_1 + t_1, \dots, x_n + t_n) g(t_1) \cdots g(t_n) dt_1 \cdots dt_n$ and write instead

$$\int \cdots \int K(x_1, \dots, x_n) f(x_1) \cdots f(x_n) dx_1 \cdots dx_n = 0$$

for all $f \in \mathcal{F}$. Note that K is symmetric due to the symmetry of ψ . We will be finished if we show that $K(x_1, \dots, x_n) = 0$ for almost all x_1, \dots, x_n for the following reason. A function vanishes a.e. if and only if its Fourier transform vanishes a.e.; so if we show that $K(x_1, \dots, x_n)$ vanishes a.e., then so too does its Fourier transform \tilde{K} . Now, the Fourier transform of K is $\tilde{K}(s_1, \dots, s_n) = \tilde{\psi}(s_1, \dots, s_n) \overline{\tilde{g}(s_1) \cdots \tilde{g}(s_n)}$. Since this is zero a.e., and $\tilde{g}(\cdot)$ is non-zero a.e., we conclude that $\tilde{\psi}(s_1, \dots, s_n) = 0$ a.e., which ultimately implies that $\psi(x_1, \dots, x_n) = 0$ for almost all x_1, \dots, x_n , which is what we seek.

Consider the symmetric multilinear form

$$M(u_1, \dots, u_n) = \int \cdots \int K(x_1, \dots, x_n) u_1(x_1) \cdots u_n(x_n) dx_1 \cdots dx_n.$$

By Lemma 2 and the hypothesis that $M(f, f, \dots, f) = 0$ for all $f \in \mathcal{F}$, we know that $M(u_1, u_2, \dots, u_n) = 0$ whenever $u_1, \dots, u_n \in \text{span}(\mathcal{F})$. For any Borel set B , the indicator $\mathbf{1}_B \in \text{span}(\mathcal{F})$, so the indicator of any product of n Borel sets can be written as $u_1(x_1) \cdots u_n(x_n)$ with $u_1, \dots, u_n \in \text{span}(\mathcal{F})$.

Thus $\int_B K(x_1, \dots, x_n) dx_1 \dots dx_n = 0$ for any product B of Borel sets, so $K(x_1, \dots, x_n) = 0$ almost everywhere. \square

Proof of Theorem 1. Let $\mathcal{A}_n = \{\alpha_i : \alpha_i = \lambda b_m + \frac{i-1}{k_n-1}(\mu - \lambda)b_m\}$, with $k_n = \lceil n^p \rceil$, $p > 1$. Then Lemma 3 holds for $\alpha_i \in \mathcal{A}_n$, and in fact the probabilities are summable, so it holds almost surely, so with probability 1,

$$\sup_{\alpha \in \mathcal{A}_n} |\text{SCV}_n(\alpha, m) - R(\alpha, m)| = o(r_m).$$

Fix an arbitrary $\alpha' \in [\lambda b_m, \mu b_m]$. We want to show that

$$|\text{SCV}_n(\alpha', m) - R(\alpha', m)| \leq C_n r_m$$

for some sequence $C_n \rightarrow 0$ which does not depend on α' . Let $\alpha_j \in \mathcal{A}_n$ be the element nearest to α' , and upper bound the difference as

$$\begin{aligned} & |\text{SCV}_n(\alpha', m) - R(\alpha', m)| \\ &= |\text{SCV}_n(\alpha', m) - \text{SCV}_n(\alpha_j, m) \\ &\quad - R(\alpha', m) + R(\alpha_j, m) + \text{SCV}_n(\alpha_j, m) - R(\alpha_j, m)| \end{aligned} \tag{5.5}$$

The following can be seen by applying expanding $\text{SCV}_n(\alpha, m)$ using Proposition 1 (ii) and noticing that the $\frac{1}{m} \|\varphi_\alpha\|^2$ terms cancel.

$$\text{SCV}_n(\alpha, m) - R(\alpha, m) = T_1(\alpha) - T_2(\alpha),$$

with

$$T_1(\alpha) = \frac{m-1}{nm(n-1)} \sum_{j=1} \sum_{k \neq j} \varphi_\alpha * \varphi_\alpha(\alpha)(Y_j - Y_k) - \frac{2}{n} \sum_{j=1}^n \mathcal{D}^{-1} f_{n,-j}^\alpha(Y_j)$$

and

$$T_2(\alpha) = \frac{n+1}{n} \|\varphi_\alpha * g * f\|^2 - 2 \int [\varphi_\alpha * g * f(x)] f(x) dx.$$

Applying this to the paired terms in Equation 5.5 and then applying the triangle inequality, we find the upper bound

$$\begin{aligned} |\text{SCV}_n(\alpha', m) - R(\alpha', m)| &\leq |\text{SCV}_n(\alpha_j, m) - R(\alpha_j, m)| \\ &\quad + |T_1(\alpha') - T_1(\alpha_j)| \\ &\quad + |T_2(\alpha') - T_2(\alpha_j)|. \end{aligned} \tag{5.6}$$

Since $\alpha_j \in \mathcal{A}_n$, by Lemma 3 there is a sequence $C_{0,n} \rightarrow 0$ s.t. $|\text{SCV}_n(\alpha_j, m) - R(\alpha_j, m)| \leq C_{0,n} r_m$, so now we only focus on the final two terms.

Now, since α_j is the nearest element of \mathcal{A}_j to α' , we have $|\alpha_j - \alpha'| \leq \frac{(\mu - \lambda)b_m}{k_n - 1}$. Using Lemma 4 to bound $\frac{\partial}{\partial \alpha} T_2(\alpha)$,

$$\begin{aligned} |T_2(\alpha') - T_2(\alpha_j)| &= \left| \int_{\alpha'}^{\alpha_j} \frac{\partial}{\partial \alpha} T_2(\alpha) d\alpha \right| \\ &\leq \int_{\alpha'}^{\alpha_j} \left| \frac{\partial}{\partial \alpha} T_2(\alpha) \right| d\alpha \\ &\leq |\alpha_j - \alpha'| \sup_{\lambda b_m \leq \alpha \leq \mu b_m} \left| \frac{\partial}{\partial \alpha} T_2(\alpha) \right| \\ &\leq \frac{(\mu - \lambda)b_m}{k_n - 1} \times \frac{C_2}{\alpha} \\ &\leq \frac{(\mu - \lambda)b_m}{k_n - 1} \times \frac{C_2}{\lambda b_m} \\ &\leq \frac{C'_2}{k_n - 1}. \end{aligned}$$

Proceeding similarly for the $T_1(\alpha)$ term,

$$\begin{aligned} |T_1(\alpha') - T_1(\alpha_j)| &\leq \frac{(\mu - \lambda)b_m}{k_n - 1} \times \frac{C_1}{\alpha^2} \\ &\leq \frac{(\mu - \lambda)b_m}{k_n - 1} \times \frac{C_1}{\lambda^2 b_m^2} \\ &\leq \frac{C'_1}{b_m(k_n - 1)}. \end{aligned}$$

At this point we have

$$|\text{SCV}_n(\alpha', m) - R(\alpha', m)| \leq \left[C_{0,n} + \frac{C_1}{r_m b_m (k_n - 1)} + \frac{C_2}{r_m (k_n - 1)} \right] r_m,$$

and $C_{0,n} \rightarrow 0$, so if the other two terms in brackets vanish as $n \rightarrow \infty$, then we are finished. By Assumption (A2), $n(r_m b_m)^2 = n^\delta$ with $\delta > 0$. This implies that $\frac{C_1}{n^p (r_m b_m)^2} \rightarrow 0$ for any $p \geq 1$; thus $\frac{C_1}{n^{\frac{1}{2}} r_m b_m} \rightarrow 0$. Since we have

chosen $k_n > n$, $\frac{C_1}{r_m b_m(k_n-1)} \rightarrow 0$, as needed. Finally, $r_m > r_n$ since $m < n$, and since r_n is a rate of convergence in deconvolution, we have $n^{-1} = o(r_n)$. Thus $\frac{C_2}{r_m(k_n-1)} \leq \frac{C_2}{r_n(k_n-1)} \leq \frac{C_2}{r_n n} \rightarrow 0$.

□

Proof of Corollary 1. By Theorem 1, $\sup_{\alpha \in \mathcal{A}_n} |\text{SCV}_n(\alpha) - R(\alpha, m)| \leq C_n r_m$ with $C_n \rightarrow 0$. Write

$$\begin{aligned} R(\alpha'_m, m) - R(\alpha_m^*, m) &= R(\alpha'_m, m) - \text{SCV}_n(\alpha'_m) \\ &\quad + \text{SCV}_n(\alpha'_m) - \text{SCV}_n(\alpha_m^*) \\ &\quad + \text{SCV}_n(\alpha_m^*) - R(\alpha_m^*, m). \end{aligned}$$

The first and third pairs are addressed by Theorem 1, while the second pair is negative because α'_m minimizes $\text{SCV}_n(\cdot)$, so

$$R(\alpha'_m, m) - R(\alpha_m^*, m) \leq C_n r_m + 0 + C_n r_m = 2C_n r_m,$$

i.e. $R(\alpha'_m, m) - R(\alpha_m^*, m) = o(r_m)$.

□

Proof of Corollary 2. By Definition 2, $R(\alpha_m^*, m) = \mathbb{E}\|f_n^{\alpha_m^*} - f\|^2 - \|f\|^2$, i.e. $-R(\alpha_m^*, m) = \|f\|^2 - \mathbb{E}\|f_m^{\alpha_m^*} - f\|^2$. Since α_m^* minimizes $\mathbb{E}\|f_m^\alpha - f\|^2$ in α , $\mathbb{E}\|f_m^{\alpha_m^*} - f\|^2 = O(r_m)$. By Corollary 1, $-\text{SCV}_n(\alpha'_m, m) = -R(\alpha_m^*, m) + o(r_m) = \|f\|^2 + O(r_m)$ almost surely, as needed.

□

5.2 Lemmas

Lemma 1. *Let $\tilde{Q}_n(t) = \frac{1}{n(n-1)} \sum_{j=1}^n \sum_{k \neq j} e^{-it(Y_j - Y_k)}$. Then $\tilde{Q}_n(t)$ is an unbiased estimator of $|\tilde{h}(t)|^2$, i.e. $\mathbb{E}\tilde{Q}_n(t) = |\tilde{h}(t)|^2$.*

Proof of Lemma 1. Since $j \neq k$ in every summand and the Y_i are independent,

$$\begin{aligned} \mathbb{E}\tilde{Q}_n(t) &= \frac{1}{n(n-1)} \sum_{j=1}^n \sum_{k \neq j} \mathbb{E} e^{-it(Y_j - Y_k)} \\ &= \frac{1}{n(n-1)} \sum_{j=1}^n \sum_{k \neq j} \mathbb{E}[e^{-itY_j}] \mathbb{E}[e^{itY_k}] \\ &= \frac{1}{n(n-1)} \sum_{j=1}^n \sum_{k \neq j} \tilde{h}(t) \overline{\tilde{h}(t)} \\ &= |\tilde{h}(t)|^2 \end{aligned}$$

□

Lemma 2. *Let V be a vector space over the reals, and $\mathcal{F} \subset V$ be convex with $\text{span}(\mathcal{F}) \supset V$. Suppose $M : V^n \rightarrow \mathbb{R}$ is multilinear and symmetric in its arguments. If $M(u, u, \dots, u) = 0$ for all $u \in \mathcal{F}$, then $M(u_1, u_2, \dots, u_n) = 0$ for all $u_1, \dots, u_n \in V$.*

Proof of Lemma 2. Since $\text{span}(\mathcal{F}) \supset V$, it suffices to show that if $u_1, \dots, u_n \in \mathcal{F}$, then $M(u_1, \dots, u_n) = 0$. For brevity, for any symmetric multilinear form $L : V^p \rightarrow \mathbb{R}$ we will write $L_k(u, v) = L(u, \dots, v, \dots)$ with u appearing exactly k times and v appearing exactly $p - k$ times. We will also write

$\ell(u) = L(u, u, \dots, u)$. We prove this by induction on the number j of unique arguments u_1, \dots, u_n . The base case of $j = 1$ is exactly the hypothesis of the lemma.

For the inductive step, assume that if there are no more than j unique arguments, then $M(u_1, \dots, u_n) = 0$. Fix $j - 1$ elements $v_1, \dots, v_{j-1} \in \mathcal{F}$ and define a new symmetric, multilinear form by $M^j(u_1, \dots, u_{n-j+1}) = M(v_1, \dots, v_{j-1}, u_1, \dots, u_{n-j+1})$ (if $j = 1$, then we fix zero of the arguments, so $M^j = M$). Note that by the inductive hypothesis, $m^j(u) = 0$ for any $u \in \mathcal{F}$, since $m^j(u)$ is $M(u_1, \dots, u_n)$ evaluated with j or fewer unique arguments – specifically the arguments v_1, \dots, v_{j-1}, u . Fix $u, v \in \mathcal{F}$, and set $w_t = tu + (1 - t)v$. Since \mathcal{F} is convex, $w_t \in \mathcal{F}$, so $m^j(w_t) = 0$. By the linearity and symmetry of M^j , we can expand $m^j(w_t)$ to find

$$\begin{aligned} 0 &= m^j(w_t) \\ &= t^{n-j+1}m^j(u) + (1-t)^{n-j+1}m^j(v) + \sum_{k=1}^{n-1} \binom{n}{k} t^k (1-t)^{n-j+1-k} M_k^j(u, v) \\ &= \sum_{k=1}^{n-1} \binom{n}{k} t^k (1-t)^{n-j+1-k} M_k^j(u, v), \end{aligned}$$

where we used that $m^j(u) = m^j(v) = 0$. The functions $t^k(1-t)^{n-j+1-k}$ are linearly independent and the expression vanishes for all $t \in [0, 1]$, so the coefficients $\binom{n}{k} M_k^j(u, v)$ all vanish; in particular $M_k^j(u, v) = 0$ for all

$u, v \in \mathcal{F}$, and since v_1, \dots, v_{j-1} were also arbitrary, $M(u_1, \dots, u_n)$ vanishes whenever there are no more than $j + 1$ unique arguments, as needed.

□

Lemma 3. *For $\lambda b_m \leq \alpha_i \leq \mu b_m$, we have the bound*

$$\mathbb{P}\left(\sup_{1 \leq i \leq k_n} r_m^{-1} |\text{SCV}_n(\alpha_i, m) - R(\alpha_i, m)| > \varepsilon\right) \leq 2k_n \exp\{-C\varepsilon^2 n(r_m b_m)^2\}. \quad (5.7)$$

Thus if $(r_m b_m)^{-2} = o(n)$ and $k_n = o(e^{n(r_m b_m)^2})$,

$$\sup_{1 \leq i \leq k_n} |\text{SCV}_n(\alpha_i, m) - R(\alpha_i, m)| = o_p(r_m). \quad (5.8)$$

Proof of Lemma 3. We use McDiarmid's inequality. By the bound in Proposition 3, we have that $|\theta_\alpha(x) - \theta_\alpha(y)| \leq C/\alpha$. Use the U-statistic representation of $\text{SCV}_n(\alpha, m)$ to find, noticing that all terms not involving Y'_i vanish,

$$\begin{aligned} & |\text{SCV}_n(\alpha, m; Y_1, \dots, Y_i, \dots, Y_n) - \text{SCV}_n(\alpha, m; Y_1, \dots, Y'_i, \dots, Y_n)| \\ & \leq \left| \frac{2}{n(n-1)} \sum_{j \neq i} \theta_\alpha(Y_j - Y_i) - \theta_\alpha(Y_j - Y'_i) \right| \\ & \leq \frac{2}{n(n-1)} \sum_{j \neq i} |\theta_\alpha(Y_j - Y_i) - \theta_\alpha(Y_j - Y'_i)| \\ & \leq \frac{2}{n(n-1)} \cdot (n-1) \cdot \frac{C}{\alpha} \\ & = \frac{C}{n\alpha}. \end{aligned} \quad (5.9)$$

Thus we have the coordinatewise bounds $c_i = \frac{C}{n\alpha}$ so that $\sum_{i=1}^n c_i^2 = \frac{C^2}{n\alpha^2}$.

Applying McDiarmid's inequality gives

$$\begin{aligned} \mathbb{P}(r_m^{-1} |\text{SCV}_n(\alpha, m) - R(\alpha, m)| > \varepsilon) &\leq 2 \exp\{-2C_1 \varepsilon^2 r_m^2 n \alpha^2\} \\ &\leq 2 \exp\{-2C_2 \varepsilon^2 (r_m b_m)^2 n\} \end{aligned}$$

A union bound gives the stated result. \square

Lemma 4. *With $T_1(\alpha)$ and $T_2(\alpha)$ defined as in the proof of Theorem 1, we have, for $\alpha < M$,*

$$\frac{\partial}{\partial \alpha} T_1(\alpha) \leq \frac{C_1}{\alpha^2} \quad \frac{\partial}{\partial \alpha} T_2(\alpha) \leq \frac{C_2}{\alpha}$$

The constant C_1 depends on M , but the constant C_2 does not.

Proof of Lemma 4. Taking Fourier transforms we have

$$\begin{aligned} T_1(\alpha) &= \frac{1}{2\pi} \int \frac{m-1}{m} |\tilde{\varphi}_\alpha(t)|^2 \tilde{Q}_n(t) dt - \frac{1}{\pi} \int \frac{1}{|\tilde{g}(t)|^2 + \alpha |t|^{2\nu}} \tilde{Q}_n(t) dt \\ &= \frac{1}{2\pi} \int \frac{\frac{m-1}{m} |\tilde{g}(t)|^2}{(|\tilde{g}(t)|^2 + \alpha |t|^{2\nu})^2} \tilde{Q}_n(t) dt - \frac{1}{\pi} \int \frac{1}{|\tilde{g}(t)|^2 + \alpha |t|^{2\nu}} \tilde{Q}_n(t) dt \end{aligned}$$

Differentiating under the integral sign,

$$\begin{aligned} \frac{\partial}{\partial \alpha} T_1(\alpha) &= -\frac{1}{\pi} \int \frac{\frac{m-1}{m} |t|^{2\nu} |\tilde{g}(t)|^2}{(|\tilde{g}(t)|^2 + \alpha |t|^{2\nu})^3} \tilde{Q}_n(t) dt + \frac{1}{\pi} \int \frac{|t|^{2\nu}}{(|\tilde{g}(t)|^2 + \alpha |t|^{2\nu})^2} \tilde{Q}_n(t) dt \\ &= \frac{1}{\pi} \int \frac{\alpha |t|^{4\nu} + \frac{1}{m} |t|^{2\nu} |\tilde{g}(t)|^2}{(|\tilde{g}(t)|^2 + \alpha |t|^{2\nu})^3} \tilde{Q}_n(t) dt. \end{aligned}$$

Since $|\tilde{Q}_n(t)| \leq 1$, we can bound this by

$$\begin{aligned} \frac{\partial}{\partial \alpha} T_1(\alpha) &\leq \frac{1}{\pi} \int \frac{\alpha |t|^{4\nu} + \frac{1}{m} |t|^{2\nu} |\tilde{g}(t)|^2}{(|\tilde{g}(t)|^2 + \alpha |t|^{2\nu})^3} dt \\ &= \frac{1}{\pi} \int \frac{\alpha |t|^{4\nu}}{(|\tilde{g}(t)|^2 + \alpha |t|^{2\nu})^3} dt + \frac{1}{m\pi} \int \frac{|t|^{2\nu} |\tilde{g}(t)|^2}{(|\tilde{g}(t)|^2 + \alpha |t|^{2\nu})^3} dt \end{aligned} \tag{5.10}$$

We'll show that $\frac{\partial}{\partial \alpha} T_1(\alpha) \leq \frac{C_1}{\alpha^2}$.

For the first integral in the upper bound of Equation 5.10, dropping the first and second term in the denominator of the integrand in turn yields upper bounds $\alpha^{-2}|t|^{-2\nu}$ and $\alpha|t|^{4\nu}|\tilde{g}(t)|^{-6}$ respectively. Now choose $\varepsilon > 0$ small enough that $|\tilde{g}(t)| \geq \delta > 0$ for all $0 \leq t \leq \varepsilon$. Then integrate the first bound away from zero and the second bound near zero:

$$\begin{aligned} \frac{1}{\pi} \int \frac{\alpha|t|^{4\nu}}{(|\tilde{g}(t)|^2 + \alpha|t|^{2\nu})^3} dt &\leq \frac{1}{\pi} \left[\alpha \int_{-\varepsilon}^{\varepsilon} t^{4\nu} |\tilde{g}(t)|^{-6} dt + \alpha^{-2} \int_{|t|>\varepsilon} |t|^{-2\nu} dt \right] \\ &\leq \frac{1}{\pi} \left[2\alpha\varepsilon^{4\nu+1}\delta^{-6} + \frac{2}{\alpha^2(2\nu-1)}\varepsilon^{1-2\nu} \right] \\ &\leq \frac{C'_1}{\alpha^2}. \end{aligned}$$

For the second integral in the upper bound of Equation 5.10, we take a similar approach. The integrand is upper-bounded by $\frac{4}{27\alpha^2}|t|^{-2\nu}$, which can be seen by applying the AM-GM inequality to $|\tilde{g}(t)|^2 + \frac{\alpha}{2}|t|^{2\nu} + \frac{\alpha}{2}|t|^{2\nu}$ in the denominator. On the other hand, dropping the $\alpha|t|^{2\nu}$ in the denominator yields a different upper bound $t^{2\nu}|\tilde{g}(t)|^{-4}$. Now integrate these bounds away from and near zero, respectively:

$$\begin{aligned} \frac{1}{m\pi} \int \frac{|t|^{2\nu}|\tilde{g}(t)|^2}{(|\tilde{g}(t)|^2 + \alpha|t|^{2\nu})^3} dt &\leq \frac{1}{m\pi} \left[\int_{-\varepsilon}^{\varepsilon} t^{2\nu} |\tilde{g}(t)|^{-4} dt + \frac{4}{27\alpha^2} \int_{|t|>\varepsilon} |t|^{-2\nu} dt \right] \\ &\leq \frac{1}{m\pi} \left[2\varepsilon^{2\nu+1}\delta^{-4} + \frac{8}{27\alpha^2(2\nu-1)}\varepsilon^{1-2\nu} \right] \\ &\leq \frac{C''_1}{\alpha^2}. \end{aligned}$$

Combining these two, we have that $\frac{\partial}{\partial \alpha} T_1(\alpha) \leq \frac{C_1}{\alpha^2}$.

Now, for $T_2(\alpha)$, taking Fourier transforms,

$$T_2(\alpha) = \frac{1}{2\pi} \left[\int \frac{\frac{n+1}{n} |\tilde{g}(t)|^4}{(|\tilde{g}(t)|^2 + \alpha |t|^{2\nu})^2} |\tilde{f}(t)|^2 dt - 2 \int \frac{|\tilde{g}(t)|^2}{|\tilde{g}(t)|^2 + \alpha |t|^{2\nu}} |\tilde{f}(t)|^2 dt \right]$$

Differentiating under the integral,

$$\begin{aligned} \frac{\partial}{\partial \alpha} T_2(\alpha) &= \frac{1}{\pi} \int \frac{\alpha |t|^{4\nu} |\tilde{g}(t)|^2 - \frac{1}{n} |t|^{2\nu} |\tilde{g}(t)|^4}{(|\tilde{g}(t)|^2 + \alpha |t|^{2\nu})^3} |\tilde{f}(t)|^2 dt \\ &\leq \frac{1}{\pi} \int \frac{\alpha |t|^{4\nu} |\tilde{g}(t)|^2}{(|\tilde{g}(t)|^2 + \alpha |t|^{2\nu})^3} |\tilde{f}(t)|^2 dt \end{aligned}$$

It can be seen that $\frac{\alpha |t|^{4\nu} |\tilde{g}(t)|^2}{(|\tilde{g}(t)|^2 + \alpha |t|^{2\nu})^3} \leq \frac{4}{27\pi\alpha}$ (apply the AM-GM inequality to

$|\tilde{g}(t)|^2 + \frac{\alpha}{2} |t|^{2\nu} + \frac{\alpha}{2} |t|^{2\nu}$), so

$$\left[\frac{\partial}{\partial \alpha} T_2(\alpha) \right] \leq \frac{8\|f\|^2}{27\pi} \alpha^{-1}. \quad (5.11)$$

□

David Kent, Cornell University

E-mail: dk657@cornell.edu

References

- Chacón, J. E., Tenreiro, C. (2024). “Estimation of density functionals via cross-validation”. en. *Statistica Neerlandica* n/a.n/a. ISSN: 1467-9574.
DOI: 10.1111/stan.12359.

REFERENCES

- Delaigle, A., Gijbels, I. (Mar. 2004). “Practical bandwidth selection in deconvolution kernel density estimation”. *Computational Statistics & Data Analysis* 45.2, pp. 249–267. DOI: 10.1016/S0167-9473(02)00329-8.
- Hall, P. (Feb. 1990). “Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems”. *Journal of Multivariate Analysis* 32.2, pp. 177–203. DOI: 10.1016/0047-259X(90)90080-2.
- Hall, P., Marron, J. S., Park, B. U. (Mar. 1992). “Smoothed cross-validation”. en. *Probability Theory and Related Fields* 92.1, pp. 1–20. ISSN: 1432-2064. DOI: 10.1007/BF01205233.
- Hesse, C. H. (Jan. 1999). “Data-driven deconvolution”. *Journal of Nonparametric Statistics* 10.4, pp. 343–373. ISSN: 1048-5252. DOI: 10.1080/10485259908832766.
- Kent, D., Ruppert, D. (2024). “Smoothness-Penalized Deconvolution (SPeD) of a Density Estimate”. *Journal of the American Statistical Association* 119.547, pp. 2407–2417. DOI: 10.1080/01621459.2023.2259028.

REFERENCES

- Lee, A. J. (1990). *U-statistics: theory and practice*. English. New York: M. Dekker. ISBN: 978-0-8247-8253-5.
- Marron, J. S., Wand, M. P. (June 1992). “Exact Mean Integrated Squared Error”. *The Annals of Statistics* 20.2, pp. 712–736. DOI: 10.1214/aos/1176348653.
- Rudemo, M. (1982). “Empirical Choice of Histograms and Kernel Density Estimators”. *Scandinavian Journal of Statistics* 9.2, pp. 65–78. ISSN: 0303-6898.
- Scott, D. W., Terrell, G. R. (Dec. 1987). “Biased and Unbiased Cross-Validation in Density Estimation”. *Journal of the American Statistical Association* 82.400, pp. 1131–1146. ISSN: 0162-1459. DOI: 10.1080/01621459.1987.10478550.
- Stefanski, L., Carroll, R. J. (Jan. 1990). “Deconvoluting Kernel Density Estimators”. *Statistics* 21.2, pp. 169–184. ISSN: 0233-1888. DOI: 10.1080/02331889008802238.
- Wager, S. (2014). “A Geometric Approach to Density Estimation with Additive Noise”. *Statistica Sinica* 24.2, pp. 533–554. ISSN: 1017-0405.

REFERENCES

- Yang, R., Apley, D. W., et al. (July 2020). “Density Deconvolution With Additive Measurement Errors Using Quadratic Programming”. *Journal of Computational and Graphical Statistics* 29.3, pp. 580–591. DOI: 10.1080/10618600.2019.1704294.
- Yang, R., Kent, D., et al. (July 2021). “Bias-corrected Estimation of the Density of a Conditional Expectation in Nested Simulation Problems”. *ACM Trans. Model. Comput. Simul.* 31.4, 22:1–22:36. ISSN: 1049-3301. DOI: 10.1145/3462201.
- Youndjé, É., Wells, M. T. (May 2008). “Optimal bandwidth selection for multivariate kernel deconvolution density estimation”. *TEST* 17.1, pp. 138–162. DOI: 10.1007/s11749-006-0027-5.