

Nonparametric Density Deconvolution By Weighted Kernel Estimators

Berwin A Turlach
statba@nus.edu.sg

Department of Statistics and Applied Probability
National University of Singapore

joint work with Martin L Hazelton

JSM, Denver, 4 August 2008 – 1 / 23

- Introduction
- Our approach
- Practical Implementation
- Numerical results
- Concluding remarks

- Introduction
- Our approach
- Practical Implementation
- Numerical results
- Concluding remarks

JSM, Denver, 4 August 2008 – 2 / 23

The setting

We observe a univariate random sample Y_1, \dots, Y_n from a density g , where

$$Y_i = X_i + Z_i \quad (i = 1, \dots, n).$$

Here

- X_1, \dots, X_n are independent and identically distributed with unknown continuous density f , and
- the measurement errors Z_1, \dots, Z_n form a random sample from the continuous density η which we assume to be known.

Our goal is to obtain a nonparametric estimate of f from the observed sample.

JSM, Denver, 4 August 2008 – 3 / 23

Classical approach (I)

The densities f , g and η are related by the convolution equation

$$g(y) = f * \eta(y) = \int f(x)\eta(y - x) dx$$

and so estimation of f is a deconvolution problem.

The standard kernel density estimator constructed from Y_1, \dots, Y_n is

$$\check{g}(y) = \check{g}(y; h) = \frac{1}{n} \sum_{i=1}^n K_h(y - Y_i)$$

where $K_h(y) = K(y/h)/h$, and K is a kernel function.

The estimator \check{g} targets g rather than the desired density f , but it can be adapted by modifying the kernel function.

JSM, Denver, 4 August 2008 – 4 / 23

- Introduction
- The setting
- Classical approach (I)
- Classical approach (II)
- Our approach
- Practical Implementation
- Numerical results
- Concluding remarks

- Introduction
- The setting
- Classical approach (I)
- Classical approach (II)
- Our approach
- Practical Implementation
- Numerical results
- Concluding remarks

- Introduction
- The setting
- Classical approach (I)
- Classical approach (II)
- Our approach
- Practical Implementation
- Numerical results
- Concluding remarks

The required deconvoluting estimator \check{f} is related to \check{g} and the error density η by $\check{g} = \check{f} * \eta$, so that $\psi_{\check{f}} = \psi_{\check{g}}/\psi_{\eta}$ where ψ_a denotes the characteristic function of a density a .

It follows that

$$\check{f}(x) = \frac{1}{n} \sum_{i=1}^n K_h^Z(x - Y_i)$$

where $K_h^Z(u) = h^{-1}K^Z(u/h; h)$ and

$$K^Z(u; h) = \frac{1}{2\pi} \int e^{-itu} \frac{\psi_K(t)}{\psi_{\eta}(t/h)} dt.$$

- Introduction
- Our approach
- Proposal (I)
- Motivation (I)
- Motivation (II)
- Proposal (II)
- Proposal (III)
- Practical Implementation
- Numerical results
- Concluding remarks

Consider weighted kernel estimators of the form

$$\hat{f}_{\mathbf{w}}(x) = \hat{f}_{\mathbf{w}}(x; h) = \frac{1}{n} \sum_{i=1}^n w_i K_h(x - Y_i),$$

where $\mathbf{w} = (w_1, \dots, w_n)^T$ is a vector of non-negative weights.

Usually, we constrain \mathbf{w} by $\bar{w} = \frac{1}{n} \sum_{i=1}^n w_i = 1$ so as to ensure that $\int \hat{f}_{\mathbf{w}}(x) dx = 1$.

K is a kernel function satisfying $\int K(y) dy = 1$, $K(y) \geq 0$, $K(y) = K(-y)$ and $\mu_2 = \mu_2(K) = \int K(y)y^2 dy < \infty$.

Motivation (I)

Motivation (II)

- Introduction
- Our approach
- Proposal (I)
- Motivation (I)
- Motivation (II)
- Proposal (II)
- Proposal (III)
- Practical Implementation
- Numerical results
- Concluding remarks

Assume we have an oracle that tells us what f/g is. Consider the weight vector $\mathbf{w}_0 = (w_{01}, \dots, w_{0n})^T$ defined by

$$w_{0i} = \frac{f(Y_i)}{g(Y_i)} \quad (i = 1, \dots, n).$$

With this choice of weights we have

$$\begin{aligned} E[\hat{f}_{\mathbf{w}_0}(x)] &= E \left[\frac{f(Y)K_h(x - Y)}{g(Y)} \right] \\ &= \int \frac{f(y)}{g(y)} K_h(x - y)g(y) dy \\ &= f * K_h(x). \end{aligned}$$

This is precisely the same mean as for a standard kernel estimator constructed from uncontaminated data.

- Introduction
- Our approach
- Proposal (I)
- Motivation (I)
- Motivation (II)
- Proposal (II)
- Proposal (III)
- Practical Implementation
- Numerical results
- Concluding remarks

Standard asymptotic expansions show that

- bias $\{\hat{f}_{\mathbf{w}_0}(x)\} \approx h^2 \mu_2 f''(x)/2$, whence
- the integrated squared bias may be approximated by $h^4 \mu_2^2 R(f'')/4$ where $R(a) = \int a(y)^2 dy$ for any squared integrable function a .
-

$$\text{var}\{\hat{f}_{\mathbf{w}_0}(x)\} = \frac{1}{n} \left[\int \frac{f^2(y)}{g(y)} K_h^2(x - y) dy - \{f * K_h(x)\}^2 \right]$$

assuming that the support of f is a subset of the support of g , whence

- the integrated variance may be approximated by

$$\text{IV}\{\hat{f}_{\mathbf{w}_0}\} \approx \frac{1}{nh} R(K) \int \frac{f(y)^2}{g(y)} dy.$$

- Introduction
- Our approach
- Proposal (I)
- Motivation (I)
- Motivation (II)
- Proposal (II)
- Proposal (III)
- Practical Implementation
- Numerical results
- Concluding remarks

If the estimator $\hat{f}_{\hat{w}}$ is good then one would expect $\hat{f}_{\hat{w}} * \eta \approx g$.

While g is unknown, we do have a natural estimator \check{g} .

This suggests that we might search for a vector of positive weights solving the linear system

$$\hat{f}_{\hat{w}} * \eta(y) = \check{g}(y),$$

possibly subject to the constraint $\bar{w} = 1$.

Our preferred approach is to minimize a global measure of the discrepancy between $\hat{f}_{\hat{w}} * \eta$ and \check{g} . Specifically, we look for weights to minimize the criterion

$$Q = Q(\mathbf{w}) = \int \{\hat{f}_{\hat{w}} * \eta(y) - \check{g}(y)\}^2 dy.$$

- Introduction
- Our approach
- Proposal (I)
- Motivation (I)
- Motivation (II)
- Proposal (II)
- Proposal (III)
- Practical Implementation
- Numerical results
- Concluding remarks

When $\eta = \phi_\sigma$ (a normal density with zero mean and variance σ^2) and the kernel is also normal, $K_h = \phi_h$, then Q has the simple expression

$$Q(\mathbf{w}) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n [w_i w_j \phi_{\sqrt{2}\lambda}(Y_i - Y_j) + \phi_{\sqrt{2}h}(Y_i - Y_j) - 2w_i \phi_\omega(Y_i - Y_j)]$$

where $\lambda^2 = h^2 + \sigma^2$ and $\omega^2 = 2h^2 + \sigma^2 = \lambda^2 + h^2$.

Theorem:

If \hat{w} denotes the minimizer of Q , then, under mild regularity conditions, the estimator $\hat{f}_{\hat{w}}$ is consistent for f .

Bandwidth selection

Optimization and Regularization (I)

- Introduction
- Our approach
- Practical Implementation
- Bandwidth selection
- Optimization and Regularization (I)
- Optimization and Regularization (II)
- Selecting the Ridge Penalty Constraint
- Numerical results
- Concluding remarks

The choice of bandwidth is typically critical in terms of performance when implementing kernel smoothers.

We found over a range of numerical experiments that our deconvolution estimator operates well using standard methods of bandwidth selection for density estimation from uncontaminated data.

This is intuitive, since the weights are chosen with respect to an error criterion Q defined in terms of estimates of g rather than of f . We wish $\hat{f}_{\hat{w}} * \eta$ to be as close as possible to g , and hence implement \check{g} using a bandwidth calibrated for estimation of g . This bandwidth is then inherited by $\hat{f}_{\hat{w}}$.

For our numeric work, we used the Sheather–Jones plug-in bandwidth selector.

- Introduction
- Our approach
- Practical Implementation
- Bandwidth selection
- Optimization and Regularization (I)
- Optimization and Regularization (II)
- Selecting the Ridge Penalty Constraint
- Numerical results
- Concluding remarks

Optimizing $Q(\mathbf{w})$ under the constraints that the weights are non-negative and sum to n leads to the following quadratic program:

$$\begin{aligned} & \underset{\mathbf{w}}{\text{minimise}} && \frac{1}{2} \mathbf{w}^T \mathbf{Q} \mathbf{w} - \mathbf{b}^T \mathbf{w} \\ & \text{subject to} && \sum_{i=1}^n w_i = n \\ & && 0 \leq w_i, \quad i = 1, \dots, n \end{aligned}$$

Theoretically the matrix \mathbf{Q} in the objective function is positive definite, in finite precision arithmetic \mathbf{Q} is typically singular.

This problem can be solved using various algorithm:

- Interior point algorithm,
- Homotopy algorithm (calculate the complete solution path);
- Gradient projection methods.

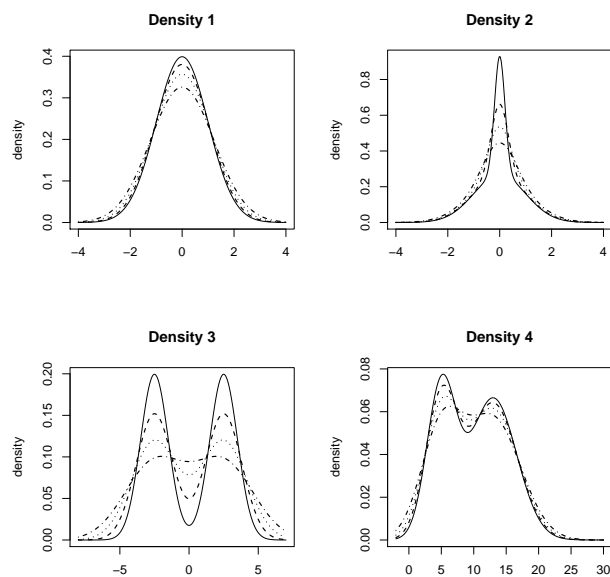
Optimization and Regularization (II)

Due to the numerical singularity of \mathbf{Q} , the solution to this problem is too variable to be of practical use and some regularisation is needed.

This leads to the final version of our proposed method:

$$\begin{aligned} & \underset{\mathbf{w}}{\text{minimise}} && \frac{1}{2} \mathbf{w}^T (\mathbf{Q} + \frac{\gamma}{n} \mathbf{I}) \mathbf{w} - \mathbf{b}^T \mathbf{w} \\ & \text{subject to} && \sum_{i=1}^n w_i = n \\ & && 0 \leq w_i, \quad i = 1, \dots, n \end{aligned}$$

Densities used



Selecting the Ridge Penalty Constraint

Select γ using maximization of a five-fold likelihood cross-validation criterion:

- Randomly partition the data into five blocks of roughly equal size.
- Compute a log-likelihood for the elements of each block using a weighted density estimate constructed from all the other data and aggregate the results. This produces cross-validation criterion

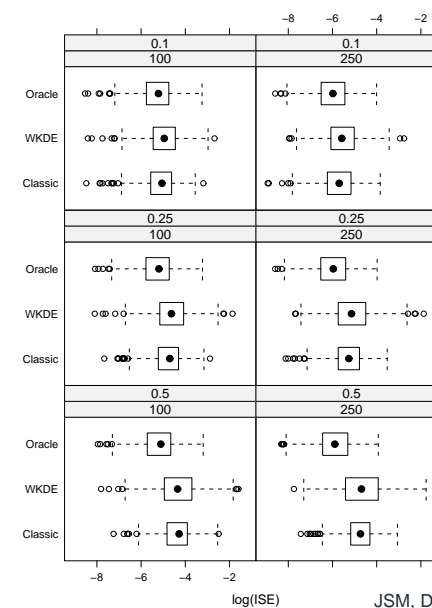
$$CV(\gamma) = CV(\hat{\mathbf{w}}(\gamma)) = \sum_{i=1}^n \log\{\hat{f}_{\hat{\mathbf{w}}}^{\setminus i} * \eta(Y_i)\}.$$

- We select γ by maximizing $CV(\gamma)$ for $\gamma > 0$.

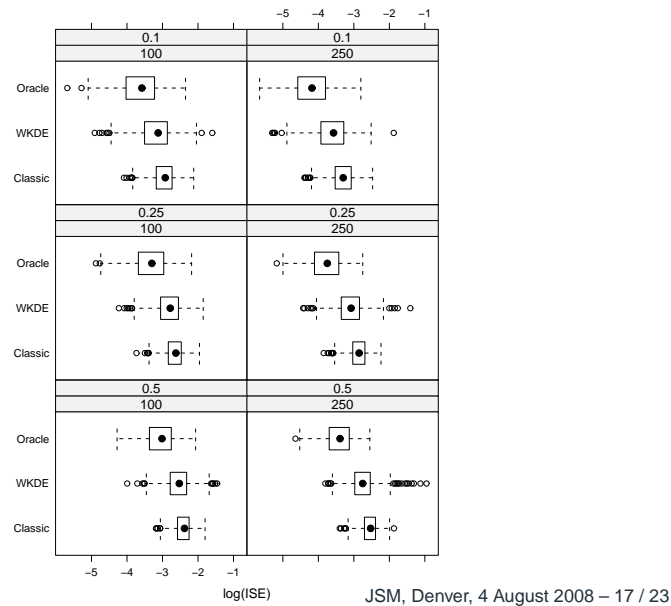
$\hat{f}_{\hat{\mathbf{w}}}^{\setminus i}$ denotes the weighted kernel estimator constructed only using those data which are in a block different to that of data point i .

For each of the five density estimates $\hat{f}_{\hat{\mathbf{w}}}^{\setminus i}$ corresponding to the five blocks, separate weights are computed based on the given value of γ .

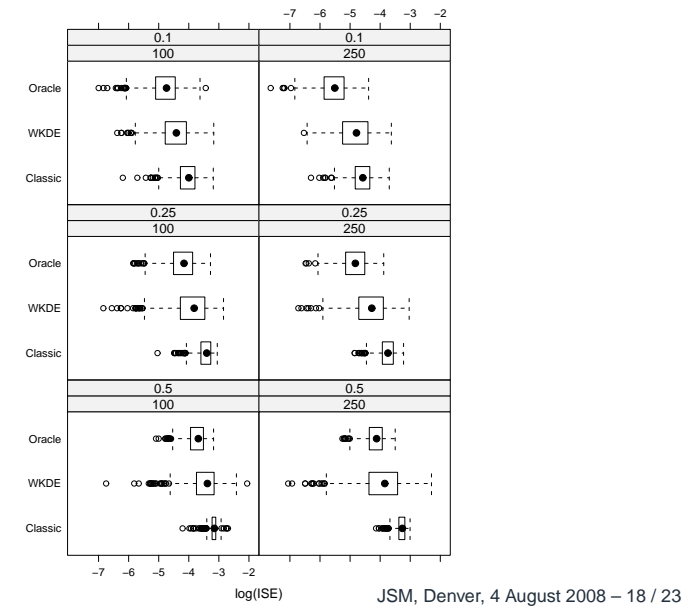
Results for density 1



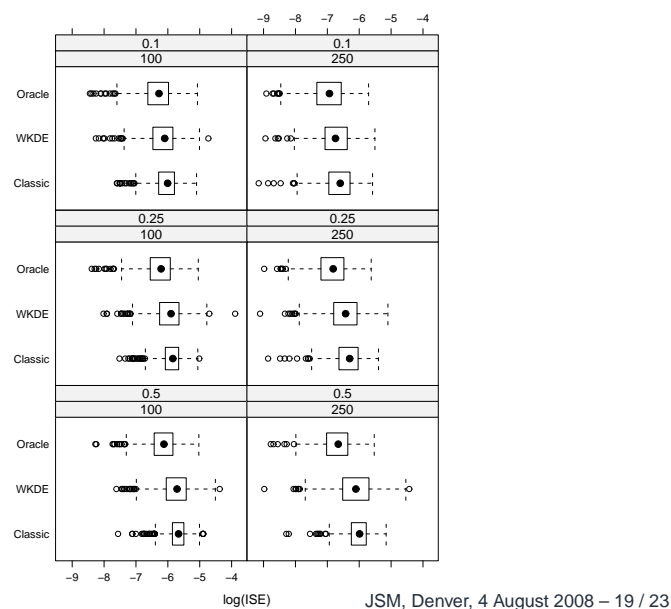
Results for density 2



Results for density 3



Results for density 4

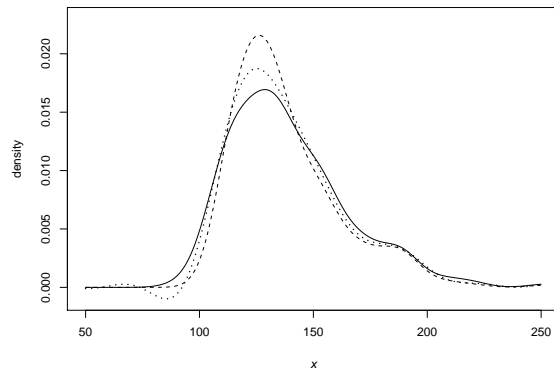


Comparing classical and WKDE method

| $\text{var}(Z)/\text{var}(X)$ | n | Density 1 | Density 2 | Density 3 | Density 4 |
|-------------------------------|-----|-----------|-----------|-----------|-----------|
| 0.1 | 100 | 0.498 | 0.890 | 0.912 | 0.778 |
| 0.1 | 250 | 0.460 | 0.934 | 0.698 | 0.758 |
| 0.25 | 100 | 0.526 | 0.862 | 0.868 | 0.690 |
| 0.25 | 250 | 0.456 | 0.894 | 0.888 | 0.694 |
| 0.5 | 100 | 0.552 | 0.826 | 0.818 | 0.600 |
| 0.5 | 250 | 0.508 | 0.866 | 0.920 | 0.636 |

Pairwise comparison of the classical and WKDE methods. The tabulated values show the proportion of simulated data sets for which WKDE returned a lower integrated squared error than the classical method.

- Introduction
- Our approach
- Practical Implementation
- Numerical results
 - Densities used
 - Results for density 1
 - Results for density 2
 - Results for density 3
 - Results for density 4
 - Comparing classical and WKDE method
- Real data example**
- Concluding remarks



Density estimates for systolic blood pressure data for $n = 285$ men from the Framingham study. The solid line is the density estimate ignoring measurement error. The dashed line is the weighted kernel deconvolution estimate. The dotted line is the classical deconvolution estimate.

- Introduction
- Our approach
- Practical Implementation
- Numerical results
- Concluding remarks
- Conclusion**
- Further work

We have proposed a new approach to density deconvolution, based on the use of weighted kernel estimators. Our methodology has a number of advantages over the classical kernel technique for deconvolution.

- our weighted kernel estimator avoids the spurious wiggles and regions of negative density that arise from the shape of the effective kernels employed in the classical method,
 - results from our simulation study indicate that weighted kernel estimation can provide tangible improvements in performance over the classical estimators for moderate sample sizes; and
 - our approach generalizes very simply to multivariate setting, even when the measurement error is correlated across variables.
- Implementation of the classical method in such circumstances can be challenging because of the complexity of the integrals required to compute the effective deconvolution kernels.

Further work

- Introduction
- Our approach
- Practical Implementation
- Numerical results
- Concluding remarks
- Conclusion
- Further work**

It is natural to compute the weights for our kernel estimator so as to minimize some measure of discrepancy between the standard kernel estimate from the contaminated data on the one hand, and the convolution of the weighted deconvolution estimate with the measurement error density on the other hand.

Our preference is to use an integrated squared difference between these densities but there are many alternatives that seem reasonable, at first sight at least. For example, one could seek to minimize the integrated absolute difference between the densities, or seek a perfect match between the densities on some finite set of values.

There remains scope for further research in this matter.