From Estimation of Traffic Flows to Deconvolution of Densities: Some Statistical Linear Inverse Problems

Martin Hazelton
m.hazelton@massey.ac.nz

Massey University

11 October 2007
As a D.Phil student, I was interested in nonparametric smoothing.

I later worked as a research assistant on mathematical and statistical problems in transportation science.

These research areas remained very important to me every since.

Until recently I regarded them as pretty much separate.
Fifteen Years Later
Some Recent Research Interests

**Nonparametric Smoothing**
- Kernel binary regression
- Estimation of geographical relative risk surfaces
- Adaptive smoothing methods
- Boundary correction methods
- Density deconvolution

**Transportation Science**
- Traffic assignment: modelling and inference
- Speed estimation
- Estimation of origin-destination (trip) matrices
Density Deconvolution

Model: \( Y = X + Z \)
- \( Y \) is ‘contaminated’ observation; density \( g \).
- \( X \) is uncontaminated latent variable; density \( f \).
- \( Z \) is measurement error; known density \( \pi \).
- Densities related by convolution formula

\[
g(y) = \int f(y - z)\pi(z) \, dz
\]

Data: Observe random sample \( Y_1, \ldots, Y_n \).

Aim: Estimation of \( f \).
Example: Blood Pressure Measurements

- Data on $n = 285$ male subjects aged 56 years and over (from Framingham health study).
- Any subject’s blood pressure measurement will vary due to
  - Error in measurement;
  - Physiological variations.
- Model $Y = X + Z$ with
  - $Y$ is measured systolic blood pressure (in mmHg);
  - $X$ is ‘long term mean’ blood pressure;
  - $Z$ is ‘measurement error’; model $Z \sim N(0, 9.0^2)$. 
Systolic BP: Convoluted Density Estimate

Presented at the University of Otago
11 October 2007
Trip Matrix Estimation: An Example Road Network

- Road network below from English city of Leicester.
- Roads represented by (directed) network links (arcs).
- Junctions and origins/destinations of flow represented by nodes.
**Trip Matrix Estimation: The Model**

**Data:** Observe traffic counts \( g_1, \ldots, g_N \) on (subset of) network links

**Model:**

\[
g_j = \sum_{i=1}^{M} f_i \pi_{ij}
\]

- \( f_i \) is realized flow for origin-destination (OD) pair \( i \)
- \( \pi_{ij} \) proportion of flow for OD pair \( i \) passing through link \( i \)
- Typically \( M \gg N \).

**Aim:** Reconstruct OD flows \( f_1, \ldots, f_M \) and/or estimate parameters of underlying distribution.
Positive Linear Inverse Problems

- Recall fundamental equations:

  \[ g(y) = \int f(y - z)\pi(z)\,dz \quad \text{(density deconvolution)} \]

  \[ g_j = \sum_{i=1}^{M} f_i\pi_{ij} \quad \text{(trip matrix estimation)} \]

- In both cases want to solve for \( f > 0 \).
- Both are instances of positive linear inverse problems.
- Unified for measures \( F, G, \Pi \) in equation

  \[ G(\cdot) = \int \Pi(z, \cdot)F(dz) \]
Regularization

- Common problem is that number of observations (much) less than number of unknowns.
- In case of trip matrix estimation, this means multiple patterns of OD flow consistent with observed link counts.
- Unique solution to such an ‘ill posed’ problem requires regularization.
- In traffic flow this can be achieved by introducing additional information (e.g. prior trip matrix, assumption of maximum entropy etc.)
Density Deconvolution

The remainder of this talk will focus on density deconvolution, and some joint work with Berwin Turlach (National University of Singapore).

Will cover

1. The difficulty of density deconvolution
2. Classical kernel deconvolution
3. Weighted kernel deconvolution
4. Nonparametric estimation of weights
5. Semiparametric estimation of weights
6. Numerical results
Reminder of the Deconvolution Problem

**Model:** \[ Y = X + Z \]
- \( Y \) is ‘contaminated’ observation; density \( g \).
- \( X \) is uncontaminated latent variable; density \( f \).
- \( Z \) is measurement error; known density \( \pi \).
- \( g(y) = \int f(y - z)\pi(z)\,dz = f \ast \pi(y) \)

**Data:** Observe random sample \( Y_1, \ldots, Y_n \).

**Aim:** Estimation of \( f \).
For nonparametric density deconvolution:

- Best achievable convergence rate is $O(1/\log(n)^\alpha)$ when $\pi$ is normal (where $\alpha$ depends on smoothness of $f$). Carroll & Hall (1988).

- Essentially, very difficult to detect small wiggles (in tails) even for huge sample sizes.

- For modest sample sizes, detection of even gross features like bimodality can be challenging.
The Difficulty of Density Deconvolution

Spot the Difference # 1

Two convoluted densities ($g$)
The Difficulty of Density Deconvolution

Spot the Difference # 1

Two convoluted densities \( g \)
The Difficulty of Density Deconvolution

Spot the Difference # 2

Corresponding deconvoluted densities ($f$)

- $f(x)$
- $f(x)$
The Difficulty of Density Deconvolution

Spot the Difference # 2

Corresponding deconvoluted densities ($f$)
Classical Kernel Deconvolution

Kernel Density Estimation

- Popular approach in nonparametric density estimation.
- Standard kernel density estimate constructed from data \(X_1, \ldots, X_n\) is

\[
\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)
\]

- \(K(\cdot)\) is kernel function; assumed to be a p.d.f. with symmetry \(K(x) = K(-x)\);
- \(K_h(x) = h^{-1}K(x/h)\) is scaled kernel;
- \(h\) is the bandwidth, which controls smoothness of estimate.
Estimate $\hat{f}$ is aggregate of ‘bumps’ centred at data points.
In deconvolution must modify density estimate constructed from contaminated data $Y_1, \ldots, Y_n$ to account for measurement error.

Classical approach adapts kernel function. Viz:

$$
\hat{g}(y) = \hat{f} \ast \pi(y) \Rightarrow \phi_{\hat{f}}(t) = \phi_{\hat{g}}(t)/\phi_{\pi}(t)
$$

where $\phi_f$ is characteristic function of $f$ etc. Hence

$$
\hat{g}(y) = n^{-1} \sum_{i=1}^{n} K_h(y - Y_i) \Rightarrow \hat{f}(y) = n^{-1} \sum_{i=1}^{n} K_{h}^{Z}(y - Y_i; h)
$$

where $K_{h}^{Z}(\cdot; h)$ is kernel with characteristic function $\phi_K/\phi_{\pi}$. 
Deconvolution kernel $K^Z_h(\cdot; h)$ will take negative values, and so generally will $\hat{f}$.

Evaluation of deconvolution kernel is computationally expensive for multivariate data with complex measurement error structure.

Performance of classical kernel deconvolution not terribly impressive:

- Small to medium $n$: simulation and other numerical results indicates that kernel deconvolution does very poorly for $\pi = N(0, \sigma^2)$ unless $\sigma$ is small. E.g. Fan (1991).
- Slow asymptotic convergence applies.
Weighted kernel density estimator:

\[ \hat{f}_w(y) \equiv \hat{f}_w(y; h) = n^{-1} \sum_{i=1}^{n} w_i K_h(y - Y_i) \]

- \( w = (w_1, \ldots, w_n)^T \) vector of non-negative weights;
- to make \( \int \hat{f}_w(x) \, dx = 1 \), impose constraint \( \bar{w} = n^{-1} \sum_{i=1}^{n} w_i = 1 \).
- Helpful to think in terms of weighting function \( w(\cdot) \), with \( w_i = w(Y_i) \).
Can think of deconvolution in terms of biased data.

Observed $Y_1, \ldots, Y_n$ from $g$, but want to sample from $f$.

Biased sampling equivalent to candidate sampling from $f$, then accepting data with probability proportional to $g/f$.

Following that using of weights $w_i = f(Y_i)/g(Y_i)$ compensates for bias.
Using optimal weight function \( w(x) = f(x)/g(x) \) we get

\[
\mathbb{E}[\hat{f}_{w}(y)] = \mathbb{E}[w_{n}K_{h}(y - Y_{n})]
\]

\[
= \int w(x)K_{h}(y - x)g(x) \, dx
\]

\[
= \int K_{h}(y - x)f(x) \, dx
\]

\[
= \mathbb{E}[K_{h}(y - X_{n})]
\]

Use of weights compensates for measurement error at expense of modest increase in estimator’s variance.
Nonparametric Estimation of Weights

- Need to estimate $w_1, \ldots, w_n$.
- Can do so by comparing two different estimates of $g$:
  - A direct kernel density estimate of $g$;
  - An implied weighted kernel density estimate of $g$.
- A direct kernel density estimate (unweighted) is
  \[
  \hat{g}(y) = n^{-1} \sum_{i=1}^{n} K_h(y - Y_i)
  \]
- Deconvolution estimate $\hat{f}_w(x)$ implies weighted kernel density
  \[
  \bar{g}(y) = \pi * \hat{f}_w(x) = n^{-1} \sum_{i=1}^{n} w_i \pi * K_h(y - Y_i)
  \]
Intuitively, select weights $w_1, \ldots, w_n$ to minimize discrepancy between $\tilde{g}(y)$ and $\hat{g}(y)$.

E.g. integrated squared difference (ISD):

$$\hat{w} = \arg\min_{w: \sum w = 1} \int \left\{ \tilde{g}(y; h) - \hat{g}(y; h) \right\}^2 dy$$
ISD objective function is a quadratic function of $w_1, \ldots, w_n$:

$$\int \left\{ \tilde{g}(y; h) - \bar{g}(y; h) \right\}^2 dy = \frac{1}{2} w^T Q w - b^T w$$

However, $Q$ is typically almost singular (often completely so in finite precision arithmetic).

This is an example of a lack of regularity in this inverse problem.

Might solve by adding ridge-regression type regularization (but that’s another story).

Other problem – how to select bandwidth?
Make the deconvolution problem easier (regularize it).

One method is to reduce degrees of freedom of weight vector $w$.

Can do by modelling the density ratio $f/g$ parametrically.

The result is a semiparametric specification of $f$. 
Semiparametric Estimation of Weights

Semiparametric Model

- Let $g_0(x)$ denote true sampling density, and define semiparametric family for $f$:

$$f(x|\theta) = w(x|\theta)g_0(x)$$

- This implies parametric family for $g$

$$g(x|\theta) = f(\cdot|\theta) \ast \pi(x)$$

- True parameter $\theta_0$ satisfies

$$g(x|\theta_0) = f(\cdot|\theta_0) \ast \pi(x) = f_0 \ast \pi(x) = g_0(x).$$

- Then $w_0(x) = w(x|\theta_0) = f_0(x)/g_0(x)$ is true density ratio (i.e. optimal weighting function).
Semiparametric Estimation of Weights
Approach to Inference

- Density ratio (weight function) $w(\cdot | \theta)$ is parametric.
- Need to estimate $\theta$.
- Can do so by constructing a semiparametric likelihood.
- If model is mis-specified (i.e. true weight function lies outside parametric class), will end up estimating closest approximation to target density (in sense of Kullback-Leibler divergence).
Can write

\[ g(y|\theta) = \int \pi(y - x) \, dF(x|\theta) \]

where \( F(\cdot|\theta) \) is distribution function corresponding to density \( f(\cdot|\theta) \).

Plug-in estimation: replace \( F(\cdot|\theta) \) with \( \hat{F}_w \), the weighted empirical measure placing probability mass \( w_i = w_i(\theta) \) at \( Y_i \) for \( i = 1, \ldots, n \).

Gives

\[ \bar{g}(y|\theta) = \int \pi(y - x) \, d\hat{F}_w(x) = \frac{1}{n} \sum_{i=1}^{n} w_i(\theta) \pi(y - Y_i) \]

for any given \( \theta \in \Theta \). Notice no bandwidth selection required.
Semiparametric Estimation of Weights
The Semiparametric Likelihood

- Use $\tilde{g}(y|\theta)$ to construct semiparametric log-likelihood.
- Gives

$$L(\theta) = \sum_{i=1}^{n} \log \{\tilde{g}(Y_i|\theta)\}$$

- Want to maximize $L(\theta)$ under constraint

$$n^{-1} \sum_{i=1}^{n} w(Y_i|\theta) = 1$$

- Denote constrained maximizer by $\hat{w}$. 
Semiparametric Estimation of Weights
(Un)constrained Maximization of the Likelihood

Can compute maximizer of $L(\theta)$ subject to weight constraint by unconstrained maximization of

$$Q(\theta) = L(\theta) - n^{-1} \sum_{i=1}^{n} w(Y_i|\theta)$$

Rationale: suppose $n^{-1} \sum_{i=1}^{n} w_i = t$. Then

$$Q(w) = \sum_{i=1}^{n} \log\{\bar{g}(Y_i|w(\theta))\} - t$$

$$= \sum_{i=1}^{n} \log\{t\bar{g}(Y_i|t^{-1}w(\theta))\} - t$$

$$= \log(t) + (1 - t) + Q(w/t) \leq Q(w/t)$$
Model $w$ as cubic spline on log-scale:

$$w_i = w(Y_i|\theta) = e^{s(Y_i|\theta)}$$

where $s(\cdot|\theta)$ is a cubic spline.

- Flexible approach which gives semiparametric method a nonparametric feel.
- This spline model includes as a special case the ‘default parametric model’ where $X$ and $Z$ are normal.
Semiparametric Estimation of Weights
Some Theory

Theorem

Assume:

- Model correctly specified;
- Some regularity conditions.

Then

\[ \mathbb{E}\{|\hat{w} - w_0|^2\} = o(n^{-4/5}) \]

- Interestingly, regularity conditions require that tails of \( g \) are not too thick.
- Intuitively, semiparametric log-likelihood too variable when \( g \) does have very heavy tails.
Using $\hat{w}$, compute deconvoluted density estimator

$$\hat{f}_w(y) = n^{-1} \sum_{i=1}^{n} \hat{w}_i K_h(y - Y_i)$$

Then

$$\mathbb{E}[\{\hat{f}_w(x) - f(x)\}^2] = O(n^{-4/5})$$

This is the standard optimal rate for kernel density estimation from uncontaminated data.

Only possibly because semiparametric model reduces difficulty of nonparametric deconvolution problem.
Semiparametric Estimation of Weights
Bandwidth Selection

- Practical performance of \( \hat{f}_w(x) \) dependent on bandwidth \( h \) (controls amount of smoothing).
- Can use leave-one-out (cross-validation) approach.
- Choose \( h \) to minimize

\[
\Delta(h) = \int \left\{ f_w(x; h) \right\}^2 dx - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \hat{w}_i \hat{w}_j K_h(Y_i - Y_j).
\]

- Considering weights as fixed, \( \Delta(h) \) is unbiased for mean integrated squared error of \( f_w(x; h) \) modulo a constant.
Numerical Results
Simulation Study

- Five target densities
- Three levels of measurement error (low, medium and high values for standard deviation of $Z$)
- Three sample sizes: $n = 100$, $n = 400$, $n = 900$.
- Two estimation methods:
  - Classical kernel deconvolution, CLAS;
  - Weighted kernel deconvolution, WKDE.
- For each combination, 400 data sets generated and integrated squared error (ISE) computed from each density estimate.
Numerical Results

Test Densities

Density 1

Density 2

Density 3

Density 4

Density 5
Numerical Results

Summary of Results

- WKDE has lower median ISE than CLAS in $39$ of the $5 \times 3 \times 3 = 45$ scenarios.
- Magnitude of the improvement of WKDE over CLAS can be large – in $20\%$ of scenarios the WKDE approach reduces the median ISE of CLAS by more than half.
- Advantage of WKDE most marked for the higher levels of measurement error.
- Overall, semiparametric outperforms nonparametric method even when semiparametric model is clearly mis-specified.

Remark

It is sometimes better to solve an approximate problem well than to solve an exact problem poorly.
Numerical Results
Deconvolution for Blood Pressure Data

Presented at the University of Otago 11 October 2007
References


http://www-ist.massey.ac.nz/mhazelton/