VU VRIJE UNIVERSITEIT AMSTERDAM

# VU Research Portal

## Asymptotics in Deconvolution Models

Donauer, S.

2009

***document version***
Publisher's PDF, also known as Version of record

**Link to publication in VU Research Portal**

***citation for published version (APA)***
Donauer, S. (2009). *Asymptotics in Deconvolution Models: Approximating Perfect Knowledge*.

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

**Take down policy**
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

**E-mail address:**
**vuresearchportal.ub@vu.nl**

Download date: 22. Jan. 2022

# Deconvolution

# Two

# The Estimator

In this chapter we cover basic properties of the Maximum Likelihood Estimator (MLE) for a class of deconvolution models with decreasing noise density. The estimator is formally defined in Section 2.1. Existence and uniqueness are addressed, as well as a characterization in terms of Fenchel optimality conditions (Section 2.2). We propose a Newton procedure with iterative minimization in order to compute the MLE. Details and an illustration of this algorithm end this chapter (Section 2.3).

## 2.1 Definition of the MLE

The deconvolution model has been studied from the perspective of a density estimation problem (i.e. estimating $f_0$) as well as a problem of estimating $F_0$. One difficulty that is encountered in all these approaches is that inverting (1.1.2) in order to express $F_0$ in terms of $h_0$ is ill-posed in the sense that there exists no stable solution for $F_0$, i.e. small changes in $h_0$ can cause large fluctuations in $F_0$.

Various methods are used in the literature to mostly estimate $f_0$. One popular approach (see Devroye (1989), Liu and Taylor (1989) and Fan (1991b)) is based on the relation $\Psi_{f_0} = \Psi_{h_0}/\Psi_g$, where $\Psi_k$ denotes the Fourier transform of a density $k$, which is an immediate consequence of (1.1.2). A kernel density estimator for $h_0$ is used to estimate $\Psi_{h_0}$ leading to an estimator of $\Psi_{f_0}$ since $g$ and therefore $\Psi_g$ is known. In a second step, an inverse Fourier transformation is used to give an estimator for $f_0$. However, some regularization is needed to guarantee the existence of the inverse Fourier transform. Truncated integration (Diggle and Hall, 1993) as well as so-called damping factors are just two ideas to deal with this technical difficulty.
Another route is to approximate $f_0$ by an element of a family of densities depending on a finite dimensional parameter. Mendelsohn and Rice (1982) suggest a linear combination of certain splines with a fixed set of knots where the coefficients are the unknown. Hermite polynomials are one alternative, used by Masry and Rice (1992). Plugging these approximations into (1.1.2) allow us to estimate the parameter using the available data which yields an estimator for $f_0$. The latter reference also describes so-called derivative deconvolution in a Gaussian setting, i.e. $Y \sim N(0, \sigma^2)$ with $\sigma^2$ known. An approximation of the Fourier transform of $g$ by a power series implies that $f_0$ can be expressed by a power series based on the derivatives $h_0^{2k}, k = 0, 1, \ldots$, which then can be estimated from the data.
Several proposed methods are based on constructing certain wavelets. See for example Starck and Bijaoui (1994), Fan and Koo (2002) or Starck and Murtagh (2002). Using either of these methods to estimate $f_0$, an estimator for the distribution function $F_0$ itself can be obtained by integrating the density estimator. One example can be found in Zhang (1990, Section 3).

Alternatively, estimators for $F_0$ can be defined directly. Cordy and Thomas (1997), for example, approximate $F_0$ by a finite mixture with known component functions and use a maximum likelihood approach to estimate the unknown coefficients of the mixture representation. In the case that $g$ is decreasing, van Es et al. (1998) define an isotonic inverse estimator (IIE) which uses results from isotonic regression leading to an explicit representation of the estimator (see also Section 1.2).

The nonparametric *Maximum Likelihood Estimator* (MLE) also estimates $F_0$ directly. This very natural and well-defined estimator, denoted in the sequel by $\hat{F}_n$, was first introduced in Groeneboom and Wellner (1992) for the case that $F_0$ lives on the positive half line. Contrary to the IIE no explicit form of $\hat{F}_n$ is available which makes it quite difficult to study this estimator.

Recall that $\mathcal{F}_{[0,\infty)}$ denotes the set of all distribution functions on $[0,\infty)$ and that $\mathcal{H} = \{h_F : F \in \mathcal{F}_{[0,\infty)}\}$ where $h_F$ was defined in (1.1.2) by

$$h_F(z) = \int_{[0,z]} g(z-x)\,dF(x),\ z \geq 0, \tag{2.1.1}$$

for a fixed known noise density $g$ with $h_0 = h_{F_0}$. The empirical distribution function of $h_0$ and its empirical counterpart are denoted by $H_0$ and $H_n$, respectively. Integration by parts leads for every $F \in \mathcal{F}_{[0,\infty)}$ to the useful representation

$$h_F(z) = g(0)F(z) + \int_{[0,z]} g'(z-y)F(y)\,dy,\ z \geq 0. \tag{2.1.2}$$

We consider models with noise densities $g : [0,\infty) \to [0,\infty)$ that are bounded, decreasing and absolutely continuous with

$$g(y) = -\int_y^\infty g'(w)\,dw = g(0) + \int_0^y g'(w)\,dw,\ y \geq 0. \tag{2.1.3}$$

The upper support point of $F \in \mathcal{F}_{[0,\infty)}$ is denoted by $S_F = \inf\{x \geq 0 : F(x) = 1\}$ with the abbreviation $S_0 = S_{F_0}$.

Definition 2.1.1 (MLE $\hat{F}_n$).
*The MLE is a distribution function $\hat{F}_n \in \mathcal{F}_{[0,\infty)}$ that maximizes the log likelihood (divided by $n$) $\Psi_n$ over the class $\mathcal{F}_{[0,\infty)}$, where*

$$\begin{aligned}\Psi_n(F) &= \frac{1}{n}\sum_{i=1}^n \log\big(h_F(Z_i)\big) \\ &= \int_{[0,\infty)} \log\Big(\int_{[0,z]} g(z-x)\,dF(x)\Big)\,dH_n(z).\end{aligned} \tag{2.1.4}$$

This definition leads to a meaningful estimator:

Lemma 2.1.2 (Existence and Uniqueness).
*A maximizer of $\Psi_n$ exists and is contained in*

$$\begin{aligned}\mathcal{F}_1 = \big\{F \in \mathcal{F}_{[0,\infty)} : F \text{ is constant on intervals } [z_{i-1}, z_i), 1 \leq i \leq n,\\ F(0) = 0, F(z_1) > 0, F(z_n) = 1\big\}\end{aligned}$$

*where $z_0 = 0 \leq z_1 \leq \ldots \leq z_n$ are the ordered observations. Within the class $\mathcal{F}_1$ the MLE is unique. If $g$ is strictly decreasing, the MLE is the unique maximizer of $\Psi_n$ over the whole set $\mathcal{F}_{[0,\infty)}$.*

Proof.
First note that $|\Psi_n| \not\equiv \infty$ on $\mathcal{F}_1$. The empirical distribution function $H_n$ is indeed an element of $\mathcal{F}_1$ with $|\Psi_n(H_n)| < \infty$ since $g(0) \geq h_{H_n}(z_i) \geq g(0)/n$ for all $z_i$. Note also that whenever $F(z_1) = 0$ we have $\Psi_n(F) = -\infty$ since then $\log(h_F(z_1)) = \log(0)$. Therefore one can assume that every maximizer of $\Psi_n$ attains a positive value at $z_1$.

Now let $F \in \mathcal{F}_{[\,0,\infty)}$ be an arbitrary distribution function with $F(0) = 0$ and $F(z_1) > 0$. We show that we can construct a piecewise constant perturbation $\tilde{F} \in \mathcal{F}_1$ of $F$ with $\Psi_n(\tilde{F}) \geq \Psi_n(F)$. Define

$$\tilde{F}(x) \;=\; \sum_{i=1}^{n-1} F(z_i)\mathbb{1}_{[\,z_i, z_{i+1})}(x) + \mathbb{1}_{[\,z_n,\infty)}(x),\; x \geq 0.$$

For $i \in \{1, \ldots, n-1\}$ we have

$$\begin{aligned} h_{\tilde{F}}(z_i) \;&=\; \int_{[\,0,z_i\,]} g(z_i - x)\, d\tilde{F}(x) \;=\; \sum_{1\leq j\leq i} g(z_i - z_j)\,(F(z_j) - F(z_{j-1})) \\ &=\; \sum_{1\leq j\leq i} g(z_i - z_j) \int_{(z_{j-1},z_j\,]} dF(x) \;=\; \sum_{1\leq j\leq i} \int_{(z_{j-1},z_j\,]} g(z_i - z_j)\, dF(x) \\ &\geq\; \sum_{1\leq j\leq i} \int_{(z_{j-1},z_j\,]} g(z_i - x)\, dF(x) \;=\; \int_{[\,0,z_i\,]} g(z_i - x)\, dF(x) \;=\; h_F(z_i) \quad (2.1.5) \end{aligned}$$

since $g$ is decreasing. A similar computation yields for $z_n$

$$\begin{aligned} h_{\tilde{F}}(z_n) \;&\geq\; \int_{[\,0,z_{n-1}\,]} g(z_n - x)\, dF(x) + g(0)(1 - F(z_{n-1})) \\ &\geq\; \int_{[\,0,z_{n-1}\,]} g(z_n - x)\, dF(x) + g(0)(F(z_n) - F(z_{n-1})) \\ &\geq\; \int_{[\,0,z_n\,]} g(z_n - x)\, dF(x) \;=\; h_F(z_n). \end{aligned}$$

Hence

$$\Psi_n(\tilde{F}) - \Psi_n(F) \;=\; \frac{1}{n}\sum_{i=1}^{n} \Big[\log\big(h_{\tilde{F}}(z_i)\big) - \log\big(h_F(z_i)\big)\Big] \;\geq\; 0$$

which allows the conclusion that if a maximizer of $\Psi_n(F)$ over $\mathcal{F}_{[\,0,\infty)}$ exists, there is also one contained in $\mathcal{F}_1$.

The existence of a maximizer of $\Psi_n$ over $\mathcal{F}_1$ follows from a compactness argument in $\mathbb{R}^n$. Each $F \in \mathcal{F}_1$ is completely determined by the vector $(F(z_1), \ldots, F(z_n))$ contained in $\{u \in \mathbb{R}^n : 0 \leq u_1 \leq \cdots \leq u_n \leq 1\}$. Hence, maximizing $\Psi_n$ over $\mathcal{F}_1$ is a finite

dimensional optimization problem over a subset of the compact set $[\,0,1\,]^n$. Define

$$\mathcal{F}_2 \;=\; \{F \in \mathcal{F}_1 : \Psi_n(F) \geq \Psi_n(H_n)\}$$

and note that $\operatorname{argmax}_{\mathcal{F}_1} \Psi_n(F) = \operatorname{argmax}_{\mathcal{F}_2} \Psi_n(F)$ since $H_n \in \mathcal{F}_1$, so that we only need to show existence with respect to $\mathcal{F}_2$. Note that $\Psi_n$ is continuous on $\mathcal{F}_2$ (by identifying also $\mathcal{F}_2$ with a subset of $[\,0,1\,]^n$) and bounded since

$$\Psi_n(F) \;=\; \frac{1}{n}\sum_{i=1}^{n} \log h_F(Z_i) \;\leq\; \log g(0).$$

Moreover, the set $\mathcal{F}_2$ (still viewed as a subset of $[\,0,1\,]^n$) is compact in $\mathcal{F}_1$ as the inverse image of the closed interval $[\,\Psi_n(H_n), \log g(0)\,]$. Compactness and continuity imply the existence of a maximizer of $\Psi_n$ in $\mathcal{F}_2$ and hence in $\mathcal{F}_1$.

Uniqueness in $\mathcal{F}_1$ follows from the strict concavity of the logarithm by the following argument. Assume that $F^{(1)} \in \mathcal{F}_1$ and $F^{(2)} \in \mathcal{F}_1$ are both maximizers of $\Psi_n$ over $\mathcal{F}_1$. Then $(F^{(1)} + F^{(2)})/2 \in \mathcal{F}_1$, satisfying

$$\begin{aligned}\Psi_n\left((F^{(1)} + F^{(2)})/2\right) \;&=\; \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{2}h_{F^{(1)}}(Z_i) + \frac{1}{2}h_{F^{(2)}}(Z_i)\right)\\ >\; \frac{1}{2}\Psi_n\left(F^{(1)}\right) + \frac{1}{2}\Psi_n\left(F^{(2)}\right) \;&\geq\; \Psi_n\left(F^{(1)}\right) \qquad (2.1.6)\end{aligned}$$

where the last inequality is due to $\Psi_n\left(F^{(1)}\right) \geq \Psi_n(F)$ for all $F \in \mathcal{F}_1$. But then (2.1.6) is a contradiction to exactly this property of $F^{(1)}$.

If $g$ is strictly decreasing, $\hat{F}_n$ maximizes $\Psi_n$ uniquely over $\mathcal{F}_{[\,0,\infty)}$ since any other assignment of mass than at the observation points would strictly decrease the likelihood since the inequality in (2.1.5) would then be strict. □

Note that in general a maximizer of $\Psi_n$ does not have to be unique over $\mathcal{F}_{[\,0,\infty)}$. Let $g$ be constant on $[\,0,\alpha\,]$ for some $\alpha > 0$. Choose $n = 1$ and assume that $0 < z_1 < \alpha$. The unique maximizer $\hat{F}_1 \in \mathcal{F}_1$ of $\Psi_1$ must be $\hat{F}_1(x) = \mathbf{1}_{[\,z_1,\infty)}(x)$, $x \geq 0$, according to the previous lemma. Hence $\Psi_1(\hat{F}_1) = \log g(0)$. Now, let $F^*(x) = \mathbf{1}_{[\,0,\infty)}(x)$, $x \geq 0$. Then $\Psi_1(F^*) = \log g(z_1) = \log g(0) = \Psi_1(\hat{F}_1)$. In what follows we always take $\hat{F}_n$ as the unique piecewise constant maximizer of $\Psi_n$ over $\mathcal{F}_1$ in the sense of Lemma 2.1.2.

Definition 2.1.3 (Set of Points of Jump).
*The set* $\mathcal{T}_n := \{x : \hat{F}_n(x) - \hat{F}_n(x-) > 0\} = \{\tau_{n,1}, \ldots, \tau_{n,m}\} \subset \{z_1, \ldots, z_n\}$ *with* $m \leq n$ *denotes the collection of points of jump of* $\hat{F}_n$.

In the sequel we usually omit the dependence of the points of jump on $n$ in the notation.

Definition 2.1.4 (MLE $\hat{h}_n$).
*The maximum likelihood estimator* $\hat{h}_n \in \mathcal{H} = \{h_F : F \in \mathcal{F}_{[0,\infty)}\}$ *of the observation density* $h_0$ *is defined by* $\hat{h}_n = h_{\hat{F}_n}$ *where* $\hat{F}_n$ *is as in Definition 2.1.1.*

Note that if $g$ has compact support $[0, S_g]$ with $0 < S_g < \infty$, the support of $\hat{h}_n$ is $[0, \tau_m + S_g] \subset [0, S_0 + 2S_g]$. In the sequel we always take the support to be $[0, S_0 + S_g]$. The following example illustrates $\hat{F}_n$ and $\hat{h}_n$.

Example (Exponential Deconvolution).
Figures 2.1 and 2.2 show the MLE $\hat{F}_n$ and the corresponding $\hat{h}_n$, respectively, in the exponential deconvolution model ($g(y) = \exp(-y)\mathbf{1}_{[0,\infty)}(y)$). The estimators are computed using the algorithm presented in the following section and are based on 100 observations using $F_0 = U[0, 5]$.
Note the piecewise decreasing structure of $h_{\hat{F}_n}$ with upward jumps at points contained in $\mathcal{T}_n$. This behavior is clear is from (2.1.1), using that $\hat{F}_n$ is piecewise constant and $g$ decreasing.

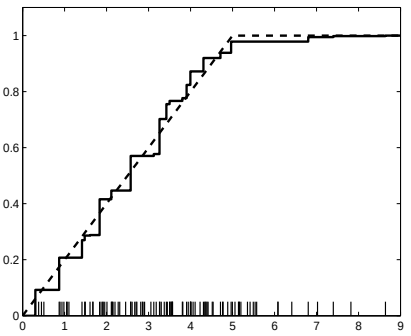


Figure 2.1: $\hat{F}_{100}$ (solid), $F_0$ (dashed).

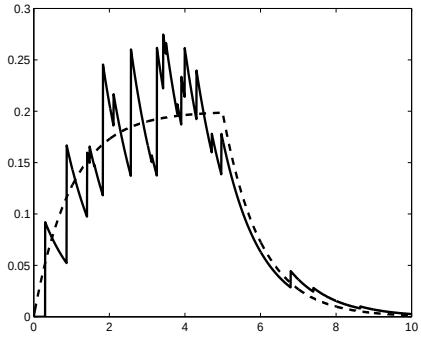


Figure 2.2: $\hat{h}_{100}$ (solid), $h_0$ (dashed).

Definition 2.1.5 (MLE $\hat{S}_n$).
*The estimator* $\hat{S}_n$ *of* $S_0 = F_0^{-1}(1)$ *is defined by the plug-in estimator*

$$\hat{S}_n \quad = \quad \inf\{x \in \mathbb{R} : \hat{F}_n(x) = 1\}.$$

Due to the uniqueness of $\hat{F}_n$ also $\hat{S}_n$ is unique. Furthermore, since $\hat{S}_n$ is the last point of jump of $\hat{F}_n$ and therefore in particular an observation point, we have $\hat{S}_n \leq Z_{(n)} \leq S_0 + S_g$.

## 2.2 Characterization

As the solution of a well-defined, convex optimization problem, $\hat{F}_n$ can be characterized by necessary and sufficient optimality conditions (Fenchel conditions). For doing so, define for $F \in \mathcal{F}_{[0,\infty)}$ the process $C_F : [0,\infty) \to [0,\infty)$ by

$$C_F(x) \;:=\; \int_{[x,\infty)} \frac{g(z-x)}{h_F(z)}\, dH_n(z).$$

We will use the usual abbreviation $\hat{C}_n = C_{\hat{F}_n}$. Note that $\hat{C}_n$ is a well-defined function since $\hat{h}_n(z_i) > 0$ for $i = 1, \ldots, n$.

Theorem 2.2.1 (Characterization of $\hat{F}_n$).
*The piecewise constant function $\hat{F}_n \in \mathcal{F}_1$ maximizes $\Psi_n$ over $\mathcal{F}_{[0,\infty)}$ if and only if*

$$\hat{C}_n(x) \;\leq\; 1 \text{ for all } x \geq 0. \tag{2.2.1}$$

*Moreover, $\hat{C}_n(x) = 1$ for all $x \in \mathcal{T}_n$.*

Example.
Figures 2.3 to 2.6 illustrate the process $\hat{C}_n$. The underlying $F_0$ is the uniform distribution on $[0,5]$. Its MLE $\hat{F}_n$ is again computed along the lines discussed in the following section and the pictures identify the MLE as the optimal solution of the optimization problem (2.1.4). For that note that the crosses at $y = 1.1$ correspond to the locations of the points of jump of $\hat{F}_n$. The first two pictures show the behavior for $g_2(y) = 2(1-y)\mathbb{1}_{[0,1]}(y)$ for $n = 15$ and $n = 1500$. The third and fourth pictures correspond to the deconvolution model with the standard exponential kernel $g_{\exp}(y) = \exp(-y)\mathbb{1}_{[0,\infty)}(y)$ for the same sample sizes.

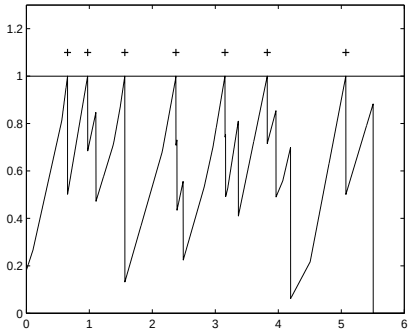


Figure 2.3: $\hat{C}_{15}$ for $g_2$.

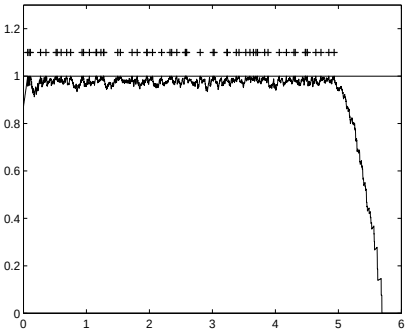


Figure 2.4: $\hat{C}_{1500}$ for $g_2$.

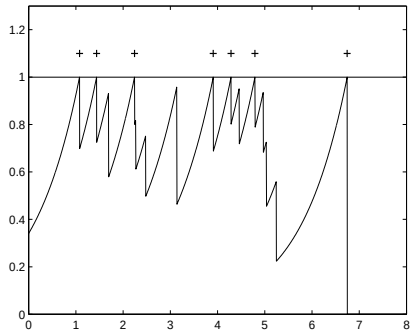


FIGURE 2.5: $\hat{C}_{15}$ for $g_{\exp}$.

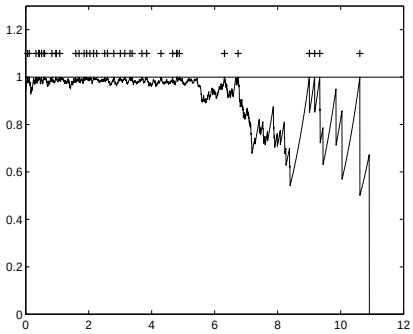


FIGURE 2.6: $\hat{C}_{1500}$ for $g_{\exp}$.

PROOF OF THEOREM 2.2.1.
First suppose that $\hat{F}_n$ maximizes $\Psi_n$ over $\mathcal{F}_{[0,\infty)}$. Define for a fixed $x \geq 0$ the function $F_x(z) := \mathbb{1}_{[x,\infty)}(z)$, $z \geq 0$, with $h_x(z) = h_{F_x}(z) = g(z-x)\mathbb{1}_{[x,\infty)}(z)$. Then for any $\varepsilon \in (0, 1]$ and $x \geq 0$, we have

$$\varepsilon^{-1}\left(\Psi_n(\hat{F}_n + \varepsilon(F_x - \hat{F}_n)) - \Psi_n(\hat{F}_n)\right) \quad \leq \quad 0$$

and hence, also using that $\Psi_n$ is concave,

$$\begin{aligned}
0 \;\geq\; & \lim_{\varepsilon \downarrow 0} \frac{\Psi_n(\hat{F}_n + \varepsilon(F_x - \hat{F}_n)) - \Psi_n(\hat{F}_n)}{\varepsilon} \\
= \;& \lim_{\varepsilon \downarrow 0} \varepsilon^{-1} \int_{[0,\infty)} \log\left(\frac{\hat{h}_n(z) + \varepsilon(h_x(z) - \hat{h}_n(z))}{\hat{h}_n(z)}\right) dH_n(z) \\
= \;& \lim_{\varepsilon \downarrow 0} \varepsilon^{-1} \int_{[0,\infty)} \log\left(1 + \varepsilon\left(\frac{h_x(z) - \hat{h}_n(z)}{\hat{h}_n(z)}\right)\right) dH_n(z) \\
= \;& \int_{[0,\infty)} \frac{h_x(z) - \hat{h}_n(z)}{\hat{h}_n(z)}\, dH_n(z) \;=\; \int_{[0,\infty)} \frac{h_x(z)}{\hat{h}_n(z)}\, dH_n(z) - 1
\end{aligned}$$

which implies (2.2.1).
Conversely, suppose that (2.2.1) is satisfied for $\hat{F}_n$. Choose $F \in \mathcal{F}_{[0,\infty)}$. Then

$$\begin{aligned}
\Psi_n(F) - \Psi_n(\hat{F}_n) \;=\; & \int_{[0,\infty)} \log\left(\frac{h_F(z)}{\hat{h}_n(z)}\right) dH_n(z) \\
\leq \;& \int_{[0,\infty)} \left(\frac{h_F(z)}{\hat{h}_n(z)} - 1\right) dH_n(z) \;=\; \int_{[0,\infty)} \frac{h_F(z)}{\hat{h}_n(z)}\, dH_n(z) - 1 \;\leq\; 0
\end{aligned}$$

since

$$\begin{aligned}\int_{[0,\infty)} \frac{h_F(z)}{\hat{h}_n(z)}\, dH_n(z) &= \int_{z\in[0,\infty)} \int_{x\in[0,z]} \frac{g(z-x)}{\hat{h}_n(z)}\, dF(x)\, dH_n(z) \\ &= \int_{z\in[0,\infty)} \int_{x\in[x,\infty)} \frac{g(z-x)}{\hat{h}_n(z)}\, dH_n(z)\, dF(x) = \int_{[0,\infty)} \hat{C}_n(x)\, dF(x) \\ &\leq \int_{[0,\infty)} dF(x) = 1.\end{aligned}$$

Hence, $\hat{F}_n$ maximizes $\Psi_n$ over $\mathcal{F}_{[0,\infty)}$.

To verify the last statement of the theorem, choose $x \in \mathcal{T}_n$. Then $x$ is a point of jump of $\hat{F}_n$ which implies that we get $(1-\varepsilon)\hat{F}_n + \varepsilon F_x \in \mathcal{F}_{[0,\infty)}$ for negative $\varepsilon$ sufficiently close to zero. Thus for any of these $\varepsilon$

$$\varepsilon^{-1}\left(\Psi(\hat{F}_n + \varepsilon(F_x - \hat{F}_n)) - \Psi(\hat{F}_n)\right) \geq 0.$$

This implies $\hat{C}_n(x) \geq 1$. Together with the first part of the proof we have $\hat{C}_n(x) = 1$. □

Interpreting characterization (2.2.1) in a geometric way, allows us to see $\hat{F}_n$ as the derivative of a convex minorant (see Groeneboom and Wellner, 1992). However, in our particular situation this cannot be used to explicitly construct $\hat{F}_n$ (contrary to van Es et al., 1998), since the weights in the convex minorant interpretation are self induced, i.e. they depend on $\hat{F}_n$ itself. However, for computational issues as addressed in the following section, characterization (2.2.1) can still be used to check whether a certain candidate $F \in \mathcal{F}_1$ maximizes $\Psi_n$ or not.

## 2.3 COMPUTATION

In order to compute the MLE based on a data set $z_1 < z_2 < \cdots < z_n$, one needs to solve a convex optimization problem (see Definition 2.1.1), which we formulate for the moment as a minimization problem. We suggest the following approach based on a Newton algorithm with iterative minimization of the quadratic approximation. This general procedure can be classified as Sequential Quadratic Minimization and in our experience it is a fast and stable way to compute $\hat{F}_n$.

In order to apply this method, the parametrization $a_i = F(z_i) - F(z_{i-1})$ for $F \in \mathcal{F}_{[0,\infty)}$, $i = 1, \ldots, n$ (where $z_0 = 0$) is used, leading to the representation

$$h_F(z) = \sum_{i=1}^{n} a_i g(z - z_i),\ z \geq 0.$$

A relaxation of the condition $\sum_{i=1}^{n} a_i = 1$ and the above parametrization allow us to

rewrite the original optimization problem of Definition 2.1.1 to that of minimizing

$$
\begin{aligned}
&-\int_{[\,0,\infty)} \log h_F(z)\, dH_n(z) + \int_0^\infty h_F(z)\, dz \\
&= \quad -\int_{[\,0,\infty)} \log\left(\sum_{i=1}^n a_i g(z-z_i)\right) dH_n(z) + \sum_{i=1}^n a_i \;=:\; \Psi_n^1(a)
\end{aligned}
$$

over the cone $\mathcal{C} = \{a \in \mathbb{R}^n \,:\, a_i \geq 0,\ i = 1, 2, \ldots, n\}$. To see that, let $\hat{a} \in \mathcal{C}$ correspond to the MLE $\hat{F}_n$. Then $\hat{a}$ is the unique minimizer of $\Psi_n^1$ over $\mathcal{C}$ by the following argument. Let $a \in \mathcal{C}$ with $\sum_{i=1}^n a_i = c$ for some constant $c \in (0, \infty)$ and let the distribution function $F^{a/c}$ correspond to the vector $a/c \in \mathcal{C}$ in terms of the above parametrization. Then $F^{a/c} \in \mathcal{F}_1$ and hence

$$
\begin{aligned}
&\Psi_n^1(\hat{a}) - \Psi_n^1(a) \\
&= \quad -\Psi_n(\hat{F}_n) + 1 + \int_{[\,0,\infty)} \log\left(c \cdot \sum_{i=1}^n \frac{a_i}{c} g(z-z_i)\right) dH_n(z) - \sum_{i=1}^n a_i \\
&= \quad -\Psi_n(\hat{F}_n) + \Psi_n(F^{a/c}) + 1 + \log(c) - c \;\leq\; 0
\end{aligned}
$$

where strict inequality holds if $\hat{a} \neq a$.

According to Newtons procedure we choose a starting value and approximate $\Psi_n^1$ locally at that value by a second order Taylor expansion. In the sequel, quadratic minimization problems are solved until sufficient optimality conditions are satisfied (up to some accuracy parameter). For the quadratic minimization we use the support reduction algorithm (SR) of Groeneboom et al. (2008a).
This leads to a converging procedure because of Theorem 8.2.3 in Bazaraa and Shetty (1979) and Theorem 1 in Groeneboom et al. (2008a).
We first give a skeleton of the actual algorithm before explaining in more detail the initialization of the Newton step and the Support Reduction step.

---

Select $a^{(0)} \in \mathcal{C}$ with $\Psi_n^1(a^{(0)}) < \infty$, $\eta > 0$ (accuracy parameter), $k := 0$ (Initialization)
While [not optimal, involving $\eta$]
$\Psi_{n,k}^1(a) := \Psi_n^1(a^{(k)}) + \nabla\Psi_n^1(a^{(k)})(a - a^{(k)}) + 1/2(a - a^{(k)})^T \nabla^2 \Psi_n^1(a^{(k)})(a - a^{(k)})$ (quadratic approximation)
$b^{(k)} := \arg\min_{a \in \mathcal{C}} \Psi_{n,k}^1(a)$ (quadratic minimization by SR)
$a^{(k+1)} = a^{(k)} + \lambda^*(b^{(k)} - a^{(k)})$ (line search by Armijo's rule)
$k := k + 1$
End

---

In the initialization step, an appropriate starting point $a^{(0)}$ has to be chosen. It will become clear later on why it is convenient to have a vector with many zeros. In case of $g$ having support $[\,0,\infty)$, one can always take $a^{(0)} = (1,0,\ldots,0)$, since then $h_F(z_i) = g(z_i - z_1) > 0$ for all $i$ and hence $\Psi_n^1(a^{(0)}) < \infty$. If $g$ has bounded support, say $[\,0, S_g\,]$ for $0 < S_g < \infty$, this $a^{(0)}$ cannot be used in general. One could take $a_1^{(0)} = 1$ and find the smallest $i$ for which $z_i - z_1 \geq S_g$, also setting $a_i = 1$ and $a_j = 0$ for $1 < j < i$ and continue this procedure in order to guarantee that $h_F(z_i) > 0$ for all $i = 1,\ldots,n$, implying finiteness of $\Psi_n^1(a^{(0)})$.
The quadratic minimization within the Newton step is solved iteratively by the SR algorithm. In doing so, the *restricted* quadratic minimization problem is solved using a sequence of *unrestricted* quadratic minimization problems. Thus, we are left with solving linear systems of equations. The number of nonzero elements in an iterate $a^{(k)}$ reflects the dimension of this linear system of equations, which is the reason for trying to have as many zeros as possible in the vectors $a^{(k)}$.
The line search parameter $\lambda^*$ is chosen according to Armijo's rule (see Luenberger, 2005, p. 212). This yields a $\lambda^* \in (0,1\,]$ such that global convergence of the algorithm is guaranteed.

Example (Exponential Deconvolution).
We consider the same situation as in the previous examples, i.e. $g$ is the exponential density and $F_0 = U[\,0,5\,]$. For $n = 250$ we compute $\hat{F}_{250}$ and show various intermediate iterates for the MLE, the behavior of $\Psi_{250}^1$ and the number of support points per iterate. The Newton procedure needed 14 outer iterations. In Figure 2.7 and 2.8 we show the second and seventh iterate, respectively. The decreasing function in Figure 2.9 illustrates the likelihood function $\Psi_{250}^1$ and the number of support points per iterate can be seen in Figure 2.10.

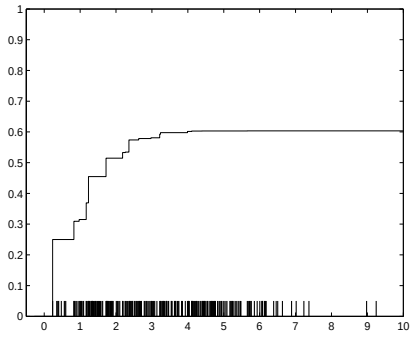


Figure 2.7: Second Iterate.

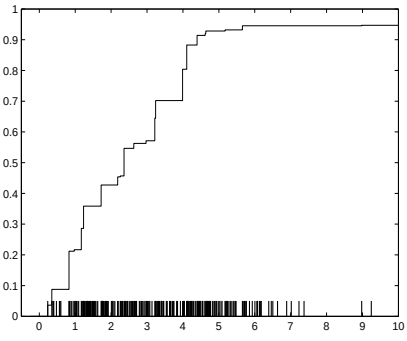


Figure 2.8: Seventh Iterate.

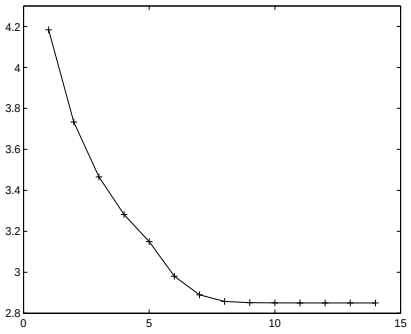


FIGURE 2.9: Log likelihood $\Psi^1_{250}$.

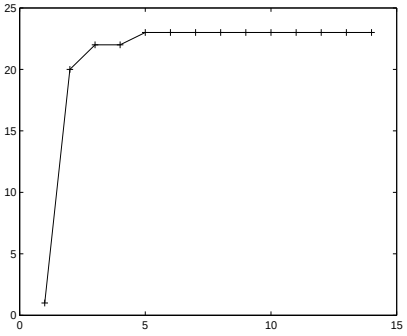


FIGURE 2.10: Number of support points.

Alternative procedures, also iterative algorithms, for computing $\hat{F}_n$ are based on a different classification of the deconvolution model. Interpreting the estimation problem as a missing data problem, the Expectation Maximization algorithm described in Dempster et al. (1977) is a natural candidate. Parameterizing the problem in terms of $(F(z_1), \ldots, F(z_n))$ the iterative convex minorant algorithm of Groeneboom and Wellner (1992), further studied in Jongbloed (1998b), is a useful algorithm. Also more general minimization methods as interior point methods can be applied (see for example Wright, 1997).