Quantifying biometric life insurance risks with non-parametric smoothing methods

Tomas, J.

Citation for published version (APA):
Chapter 4

Adaptive local kernel-weighted log-likelihood methods

This chapter is based on Tomas and Planchet (2012), Multidimensional smoothing by adaptive local kernel-weighted log-likelihood with application to long-term care insurance, ISFA - Laboratoire SAF Working paper 2012.8, submitted to Insurance: Mathematics & Economics, 1-28; and on Tomas (2012b), Essays on boundaries effects and practical considerations for univariate graduation of mortality by local likelihood models, Insurance and Risk Management, forthcoming.

4.1 Introduction

Local fitting techniques combine excellent theoretical properties with conceptual simplicity. They are very adaptable and also convenient statistically. In Chapter 3, we have seen the applicability of local kernel-weighted log-likelihoods to model the relation between the forces of mortality - or the crude death rates - and attained age.

Unfortunately, as we face situations where data have a low noise and a large amount of structure, the simplicity of a local modeling has flaws. For instance, at the boundary, the smoothing weights are asymmetric and the estimate may have substantial bias. Bias can be a problem if the regression function has a high curvature in the boundary. It may force the criteria to select a smaller bandwidth at the boundary to reduce the bias, but this may lead to under-smoothing in the middle of the table, see Section 4.2. As a consequence, in some cases no global smoothing parameter or degree of local polynomial provides an adequate fit to the data. Rather than restricting the smoothing parameters to a fixed value, a more flexible approach is to allow the constellation of smoothing parameters to vary across the observations.
We can restrict the observations contributing to the criteria to the central region and apply weights according to the reliability of the data to enhance the optimization criteria and refine the choice of the smoothing parameters. But weighting the criteria is an illustration for a need of an adaptive smoothing procedure. Rather than weighting the criteria and restricting the observations to the central region, we would use a more flexible approach. It would be to vary the amount of smoothing in a location dependent manner and to allow adjustment based on the reliability of the data. It may be advantageous for several reasons. The estimator can adapt the reliability of the data to take into account the nature of the risk, smoothing more when the volume of observations is low, and less when the corresponding amount of observations is large. We distinguish a locally adaptive smoothing point-wise method using the intersection of confidence intervals rule, as well as a global method using local bandwidth correction factors. Part of our work is an extension of the adaptive kernel methods proposed by Gavin et al. (1995) to adaptive local kernel-weighted log-likelihoods techniques. The techniques can be implemented without difficulty in standard statistical software such as R, R Development Core Team (2012).

Section 4.2 presents the motivation for adaptive smoothing, studying the influence of the boundaries on the choice of the smoothing parameters and the possibility of taking the nature of the risk into account. The adaptive methods are introduced in Section 4.3. Section 4.4 illustrates how the methods can be applied to multidimensional smoothing. We are interested in the variation of mortality of individuals subscribing long-term care insurance. We analyze the incidence of mortality as a function of both the age of occurrence of the pathology and the duration of the care. Tests and single indices summarizing the lifetime probability distribution are used to compare the graduated series with those obtained from global non-parametric approaches, p-splines and local likelihood, in Section 4.5. Finally, Section 4.6 summarizes the conclusions drawn in this chapter.

4.2 Motivations for an adaptive smoothing

4.2.1 Influence of the boundaries on a global criterion

Table 4.1 presents the proportion of the contribution to the AIC criterion in the boundaries given by the local likelihood models for the Dutch male and female data, computed with a triweight weight function and \( \lambda = 19 \), see Section 2.3. It is apparent that the contribution varies with the underlying structure of the data. The mortality patterns of females are less pronounced than those of the males, and thus the resulting contribution to the criterion is smaller. The local Poisson model is less influenced by the boundaries than the local Binomial model as most of the curvature appears in the central region.
Correction type 1 leads to the highest contributions to the AIC. This treatment induces the highest amount of smoothing in the boundaries and thus leads to the highest disturbance when choosing the constellation of smoothing parameters. When summing the contribution coming from the left and right boundaries, we observe that the boundaries represent at least 52.02% and 57.51% to 62.03% and 81.01% of the AIC, respectively, for the local Poisson and Binomial models. It is obvious that the selection of the smoothing parameter is driven by minimizing the criterion in the boundaries rather than for the whole set of data points.

The disturbance is reduced when treatment type 2 is used. However the contribution to the AIC is still relatively high with at most 51.74% and 32.94%, for the local Poisson and Binomial models respectively.

Correction type 3 implies smooth weights having the smallest dispersion around the central value. In consequence, it leads to the smallest disturbing nuisance. The contribution to the criterion for observations in the left boundary has strongly reduced while the contribution in the right boundary has inflated. This type of treatment leads to under-smoothed figures in the left boundary and its merit would depend on the underlying smoothness of the data.

It shows the resulting difficulty of applying a global smoothing approach when the true curve presents rapid changes in the curvature.

A solution for a homogeneous contribution of the design space to the criterion would be to modify the AIC by taking the logarithm of the deviance or weighting the criterion by the variance of the fitted values. It would lead, as for the criteria for linear smoothers, to a reduction in the variability, and the criterion would be less affected by the boundaries. Further, we consider restricting the computation of the criteria to observations in the central region and study where the contribution to these criteria are coming from in the design space.
CHAPTER 4. ADAPTIVE LOCAL KERNEL-WEIGHTED LOG-LIKELIHOOD

Figure 4.1: Pointwise contribution to the criteria when restricting and weighting the observations for the local Poisson model targeting the number of deaths, $d_i$, Dutch male population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.

Figure 4.2: Pointwise contribution to Rice's $T_c$ criterion when restricting and weighting the observations for the Whittaker-Henderson model targeting the number of deaths, $d_i$, Dutch male population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.

Figure 4.3: Pointwise contribution to the criteria when restricting and weighting the observations for the local Binomial model targeting the mortality rate, $q_i$, Dutch male population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.

Figure 4.4: Pointwise contribution to Rice's $T_c$ criterion when restricting and weighting the observations for the Whittaker-Henderson model targeting the mortality rate, $q_i$, Dutch male population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.
Figure 4.5: Pointwise contribution to the criteria when restricting and weighting the observations for the local Poisson model targeting the number of deaths, $d_i$, Dutch female population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.

Figure 4.6: Pointwise contribution to Rice’s $T_c$ criterion when restricting and weighting the observations for the Whittaker-Henderson model targeting the number of deaths, $d_i$, Dutch female population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.

Figure 4.7: Pointwise contribution to the criteria when restricting and weighting the observations for the local Binomial model targeting the mortality rate, $q_i$, Dutch female population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.

Figure 4.8: Pointwise contribution to Rice’s $T_c$ criterion when restricting and weighting the observations for the Whittaker-Henderson model targeting the mortality rate, $q_i$, Dutch female population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.
Restricting the observations participating in the computation of the criteria helps to reduce the boundary effects, see Fan et al. (1998). At the boundaries, the pointwise contributions are too large because of numerical instabilities, underlying structure and scarcity of the data. Figures 4.1, 4.3 and 4.5, 4.7, first row, show the pointwise contributions to the criteria when restricting the contribution to observations in the central region for the local Poisson model and the local Binomial model respectively, for the Dutch male and female population.

The pointwise contributions to the criteria differ due to the underlying structure of the data as the mortality patterns are more pronounced for the male than the female population. We observe that observations around age 18, corresponding to the accident hump, as well as observations around 60, corresponding to a cohort effect, contribute more to the criteria for the male population when fitting both of the local likelihood models. By fitting the local Poisson model, we notice an increase of the pointwise contribution with the number of deaths. This is particularly visible for the ERSC and \( \log(MSE) \). On the other hand, in case of a local Binomial model, the pointwise contribution to the ERSC and \( \log(MSE) \) tends to decrease as the curvature of the observed mortality rates increases.

These features can be also seen in the pointwise contribution to Rice’s \( T \) criterion used for linear smoothing, shown in Figures 4.2, 4.4 and 4.6, 4.8, for the Whittaker-Henderson model targeting the number of deaths and the mortality rates on the original scale, for the Dutch male and female population respectively.

In graduating the mortality rates, however, the decrease of the pointwise contribution with the increasing curvature can be a problem. It may force the criterion to select a larger bandwidth and this may lead to over-smoothing at the end of the table. It results in underestimating the mortality rates and in missing the mortality pattern of the oldest ages.

### 4.2.2 The nature of the risk

In practice, the search for an optimal criterion depends not only on statistical considerations but also on the nature of the risk considered. A smoothing method well suited for annuities may not be suited for death benefits. In the first case, we have to represent effectively the remaining life expectancy in the regions where the exposure is high. In the second case, we have to represent the observed deaths well where the number of deaths is large, and these regions may not necessarily be those where there is more exposure, such as the female population. Therefore, the criteria can be weighted according to the nature of the risk considered to refine the choice of the smoothing parameters:

i. by \( l_i / \sum_j l_j \) in case of annuities, and

ii. by \( d_i / \sum_j d_j \) in case of death benefits.
Table 4.2 presents the contribution to the criteria for observations in the age range representing 80% of the exposure and number of deaths for the Dutch male and female population after weighting the criteria according to the nature of the risk considered.

<table>
<thead>
<tr>
<th>Population</th>
<th>Age range</th>
<th>l_i</th>
<th>Rice’s T</th>
<th>AIC</th>
<th>ERSC</th>
<th>log(MSE)</th>
<th>AIC</th>
<th>ERSC</th>
<th>log(MSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>8-67</td>
<td>80</td>
<td>91.27</td>
<td>85.85</td>
<td>72.87</td>
<td>82.92</td>
<td>90.06</td>
<td>89.20</td>
<td>89.84</td>
</tr>
<tr>
<td>Female</td>
<td>8-70</td>
<td>80</td>
<td>90.38</td>
<td>81.86</td>
<td>67.98</td>
<td>79.54</td>
<td>84.74</td>
<td>88.36</td>
<td>86.40</td>
</tr>
<tr>
<td>Male</td>
<td>59-90</td>
<td>80</td>
<td>89.76</td>
<td>90.07</td>
<td>93.09</td>
<td>89.91</td>
<td>86.83</td>
<td>85.05</td>
<td>83.49</td>
</tr>
<tr>
<td>Female</td>
<td>46-90</td>
<td>80</td>
<td>98.53</td>
<td>99.26</td>
<td>99.34</td>
<td>98.52</td>
<td>98.21</td>
<td>96.34</td>
<td>96.28</td>
</tr>
</tbody>
</table>

Table 4.2: Contribution to the criteria (in %) for observations in the age range representing 80% of the exposure and number of deaths for the Dutch male and female population, 2008. Computed with a cubic fit and a triweight weight function. Source: HMD.

For the male population, 80% of the exposure appears in the age range 8 – 67. For the female population the age range corresponds to 8 – 70. In case of annuities, by weighting by \( l_i / \sum_j l_j \) most of the criteria applied to the local Poisson model (force of mortality) and to the Binomial model (probability of death) provides a good representation. The contribution to these criteria, for observations in the age range considered, are mostly above 80%. Only the ERSC provides a poor representation when fitting the local Poisson model, due to the distribution of the criterion following broadly the observed number of deaths.

For the male and female population, 80% of the deaths appears in the age ranges 59 – 90 and 46 – 90, respectively. In case of death benefits, by weighting the criteria by \( d_i / \sum_j d_j \), the proportion of the contributions from observations in the age range are above 80% showing a good representation of the risk considered.

For linear smoothers, the representation of the risk given by Rice’s T and variations of the classical criteria is satisfactory.

In consequence, weighting the criteria by the reliability of the data leads to a better representation of the nature of the risk considered whatever the model used for graduating the forces of mortality or the probabilities of death.

Figures 4.1, 4.3 and 4.5, 4.7, second and third row, show the pointwise contribution to the criteria for the local Poisson model and the local Binomial model respectively, for the Dutch male and female population, when restricting the contribution to observations in the central region and weighting according the reliability of the data.

We have restricted the observations contributing to the criteria to the central region and applied weights according to the reliability of the data.
These practical considerations enhance clearly the optimization criteria, and the choice of the constellation of the smoothing parameters is refined, leading to a good representation of the risk considered.

It should be noted that graduating mortality data through the Whittaker-Henderson model by selecting the parameters through the log transform of classical criteria, see Section 2.5, performs relatively well, and taking into account the nature of the risk improves the smoothing.

Weighting the criteria according to the reliability of the data illustrates the need of an adaptive smoothing procedure. Rather than weighting the criteria and restricting the observations to the central region, we would use a more flexible approach. It would be to allow the constellation of smoothing parameters to vary across the observations to vary the amount of smoothing in a location dependent manner. It would allow adjustments based on the reliability of the data and on the nature of the risk considered. It may be advantageous for several reasons. The estimator could adapt to the reliability of the data to take into account the nature of the risk, smoothing more when the volume of observations is low, and less when the corresponding amount of observations is large.

4.3 Adaptive Methods

This section presents the adaptive methods and covers model selection issues. We treat the choices of bandwidth, polynomial degree and weight function as modeling the data and choose the constellation of smoothing parameters to balance the trade-off between bias and variance. We distinguish a locally adaptive pointwise smoothing method using the intersection of confidence intervals rule and a global method using local bandwidth correction factors. We vary the amount of smoothing in a location dependent manner and allow adjustments based on the reliability of the data.

It is well known that of the smoothing parameters, the weight function has much less influence on the bias and variance trade-off than the bandwidth or the order of approximation. The choice is not too crucial, at best it changes the visual quality of the regression curve. For convenience, we use the Epanechnikov weight function in expression (4.6) throughout this chapter, as it is computationally cheaper to use a truncated kernel. Moreover, it has been shown that the Epanechnikov kernel is optimal in minimizing the mean squared errors for local polynomial regression, see Fan et al. (1997). The biweight and triweight kernel, which behave very similarly, could have also been chosen. The choice remains subjective.

The data used for the following illustrations are presented in Section 4.4.4. In brief, the data come from observations of individuals subscribing to Long-Term Care (LTC) insurance policies originating from a portfolio of a French
insurance company. We focus on measuring the forces of mortality as a function of the age $v$ of occurrence of the pathologies and the duration $u$ of the care.

### 4.3.1 Intersection of confidence intervals

The intersection of confidence intervals was introduced by Goldenshulger and Nemirovski (1997) and further developed by Katkovnik (1999). Application of the ICI rule in case of Poisson local likelihood for adaptive scale image restoration has been studied in Katkovnik et al. (2005). Chichignoud (2010) in his Ph.D. thesis presents a comprehensive illustration of the method. The intersection of confidence intervals (ICI) provides an alternative method of assessing local goodness of fit.

We start by defining a finite set of window sizes $\Lambda = \{\lambda_1 < \lambda_2 < \ldots < \lambda_K\}$, and determines the optimal bandwidth by evaluating the fitting results. Let $\hat{\psi}(x_i, \lambda_k)$ be the estimate at $x_i$ for the window $\lambda_k$. To select the optimal bandwidth, the ICI rule examines a sequence of confidence intervals of the estimates $\hat{\psi}(x_i, \lambda_k)$:

\[
\hat{I}(x_i, \lambda_k) = \left[ \hat{L}(x_i, \lambda_k), \hat{U}(x_i, \lambda_k) \right],
\]

\[
\hat{U}(x_i, \lambda_k) = \hat{\psi}(x_i, \lambda_k) + c \hat{\sigma}(x_i) \|s(x_i, \lambda_k)\|,
\]

\[
\hat{L}(x_i, \lambda_k) = \hat{\psi}(x_i, \lambda_k) - c \hat{\sigma}(x_i) \|s(x_i, \lambda_k)\|,
\]

where $c$ is a threshold parameter of the confidence interval. Then, from the confidence intervals, we define

\[
\hat{L}(x_i, \lambda_k) = \max \left[ \hat{L}(x_i, \lambda_{k-1}), \hat{L}(x_i, \lambda_k) \right],
\]

\[
\hat{U}(x_i, \lambda_k) = \min \left[ \hat{U}(x_i, \lambda_{k-1}), \hat{U}(x_i, \lambda_k) \right],
\]

$k = 1, 2, \ldots, K$ and $\hat{L}(x_i, \lambda_0) = \hat{U}(x_i, \lambda_0) = 0$.

The largest value for these $k$ for which $\hat{U}(x_i, \lambda_k) \geq \hat{L}(x_i, \lambda_k)$ gives $k^*$, and it yields a bandwidth $\lambda_k^*$, that is the required optimal ICI bandwidth.

In other words, denoting $I_j = \bigcap_{j=k}^{K} \hat{I}(x_i, \lambda_j)$ for $k = 1, 2, \ldots, K$, we choose $k^*$ such that

\[
\begin{cases}
I_j \neq \emptyset, & \forall j \geq k^*, \\
I_{k^*-1} = \emptyset.
\end{cases}
\]

As the bandwidth $\lambda_k$ is increased, the standard deviation of $\hat{\psi}(x_i, \lambda_k)$, and hence $\|s(x_i, \lambda_k)\|$, decreases. The confidence intervals become narrower. If $\lambda_k$ is increased too far, the estimate $\hat{\psi}(x_i, \lambda_k)$ will become heavily biased,
and the confidence intervals will become inconsistent in the sense that the intervals constructed at different bandwidths have no common intersection. The optimal bandwidth $\lambda_{k^*}$ is the largest $k$ when $\hat{U}(x_i, \lambda_k) \geq \hat{L}(x_i, \lambda_k)$ is still satisfied, i.e. when $I_j \neq \emptyset$.

Because the optimal bandwidth is decided by $c$, this parameter plays a crucial part in the performance of the algorithm. When $c$ is large, the segment $\hat{I}(x_i, \lambda_k)$ becomes wide and it leads to a larger value of $\lambda_{k^*}$. This results in over-smoothing. On the contrary, when $c$ is small, the segment $\hat{I}(x_i, \lambda_k)$ would become narrow and it leads to a small value of $\lambda_{k^*}$ so that the volatility cannot be removed effectively. In theory, we could apply the criteria presented in Section 3.4 to determine a reasonable value $c$. However, because of practical constraints, the choice of $c$ is done subjectively.

### 4.3.2 Local bandwidth factor methods

Instead of having a pointwise procedure, other types of adaptive approaches could be performed by using a global criterion. We could incorporate additional information into a global procedure by allowing the bandwidth to vary according to the reliability of the data, such as the variable kernel estimator proposed in Gavin et al. (1995, pp.190-193). We can calculate a different bandwidth for each age at which the curve has to be estimated. The local bandwidth at each age is simply the global bandwidth multiplied by a local bandwidth factor to allow explicit dependence on this information. As we already obtained the local bandwidth factors, the process of using a global criterion decides the global value at which the bandwidth curve is located.

The aim is to allow the bandwidth to vary according to the reliability of the data, and to take into account the nature of the risk considered. The local bandwidth factors could depend on the exposure or the number of deaths per attained age, in case of annuities and death benefits, respectively. For regions in which the exposure is large, a low value for the bandwidth results in an estimate that more closely reflects the crude rates. On the other hand, for regions in which the exposure is small, such as long duration, a higher value for the bandwidth allows the estimate of the true forces of mortality to progress more smoothly. This means that for long duration we are calculating local averages over a greater number of observations, which reduces the variance of the graduated rates but at the cost of a potentially higher bias.

The local bandwidth at each age is the global bandwidth multiplied by a local bandwidth factor, $h_i = h \times \delta_i^s$ for $i = 1, \ldots, n$. The variation in exposure or in deaths within a dataset can be enormous. To dampen the effect of this variation we choose

$$\delta_i^s \propto \hat{\xi}_i^{-s}, \quad \text{for } i = 1, \ldots, n \text{ and } 0 \leq s \leq 1,$$

(4.1)
where \( s \) is a sensitivity parameter and

\[
\hat{\xi}_i = \begin{cases} 
E_i / \sum_{j=1}^{n} E_j & \text{for } i = 1, \ldots, n \text{ in case of annuities}, \\
d_i / \sum_{j=1}^{n} d_j & \text{for } i = 1, \ldots, n \text{ in case of death benefits}.
\end{cases}
\]  

Choosing \( s = 0 \) reduces both models to the fixed parameter case, while \( s = 1 \) may result in very large smoothness variation depending on the particular dataset. We choose the reciprocal of \( \max\{\xi_i^{-s}; \; i = 1, \ldots, n\} \) as the constant of proportionality in (4.1), so that \( 0 < \delta^s_i \leq 1 \), for \( i = 1, \ldots, n \). The observed exposure, or the observed deaths, decides the shape of the local bandwidth factor but the sensitivity parameter \( s \) determines the magnification of that shape, becoming more pronounced as \( s \) tends to 1. Figure 4.9a shows the exposure for the age of occurrence 70 and Figure 4.9b displays the resulting smoothness tuning parameter for values of the sensitivity parameter of 0, 0.05, 0.1, 0.15, 0.25, 0.5 and 1.

For \( s = 0.15 \), the minimum smoothness tuning parameter is about 0.5, at duration 0. This means that the bandwidth at the longest duration is about twice that at the shortest duration.

Figure 4.10 presents the value of \( \delta_{u,v} \) for \( s = 0.15 \) and local bandwidth values (radius) derived. If there is a small exposure, then \( \delta_{u,v}^s \) is large. It increases the smoothness tuning parameter and allows to apply more smoothing. The other way around if the amount of exposure is large.

Similarly to the global approach, we can apply the criteria presented in Section 3.4 to select the optimal constellation of smoothing parameters. As
we already obtained the shape and the magnification of the local bandwidth factors, this process decides the global value at which the bandwidth curve is located.

4.4 Application

To illustrate the adaptive local kernel-weighted log-likelihood approaches, we discuss an application concerning the mortality of individuals having a long-term care (LTC) insurance contract. LTC is a mix of social and health care provided on a daily basis, formally or informally, at home or in institutions, to people suffering from a loss of mobility and autonomy in their activity of daily living. Although loss of autonomy may occur at any age, its frequency rises with age. LTC insurance contracts are individual or collective and guarantee the payment of a fixed allowance, in the form of monthly cash benefit, possibly proportional to the degree of dependency, see Kessler (2008) and Courbage and Roudaut (2011) for studies on the French LTC insurance market.

Most of the actuarial publications on this topic focus on the construction of models of projected benefits, see Gauzère et al. (1999) and Deléglise et al. (2009). Here we are concerned about the construction of the survival distribution of LTC insurance policyholders. The pricing and reserving as well as the management of LTC portfolios are very sensitive to the choice of the mortality table adopted. In addition, the construction of such table is a difficult exercise due to the following features:

i. French LTC portfolios are relatively small and the estimation of crude death rates is very volatile;
ii. because of the strong link between the age at subscription of LTC insurance policy and the related pathology, it is usual to construct a mortality table based on both age of occurrence of the pathologies, which is an explanatory variable, and duration of the care (or seniority), which is the duration variable. Hence, it is necessary to construct a mortality surface;

iii. mortality rates decrease very rapidly with the duration of the care. In consequence, the first year is often difficult to integrate into the usual (parametric) models.

Thus practitioners often use empirical methods that rely heavily on experts opinion. We therefore propose, in this chapter, more rigorous methods for the graduation of mortality tables not depending on experts advice.

4.4.1 Analysis of the changes in mortality

We analyze the changes in mortality of individuals subscribing LTC insurance policies as a function of both the duration of the care and the age of occurrence of the pathology. Let $T_u(v)$ be the remaining lifetime of an individual when the pathology occurred at age $v$, for the duration of the care $u$, with $v$ and $u$ being integers. We are working with two temporal dimensions $u$ and $v$, however, they do not have the same status: $v$ is a variable denoting the heterogeneity while $u$ represents the variable linked with the duration. The distribution function of $T_u(v)$ is denoted as $\tau q_u(v) = \Pr \{ T_u(v) \leq \tau \} = 1 - \tau p_u(v)$. The force of mortality at duration $u + \tau$ for the age of occurrence $v$, denoted by $\varphi_{u+\tau}(v)$ is defined by

$$
\varphi_{u+\tau}(v) = \lim_{\Delta \tau \to 0^+} \frac{\Pr \{ \tau < T_u(v) \leq \tau + \Delta \tau | T_u(v) > \tau \}}{\Delta \tau} = \frac{1}{\tau p_u(v)} \frac{\partial}{\partial \tau} \tau q_u(v),
$$

and, $\tau p_u(v) = \exp \left( - \int_0^\tau \varphi_{u+\xi} (v + \xi) \, d\xi \right)$.

We assume that the duration-specific forces of mortality are piecewise constant in each unit square, but allowed to vary from one unit square to the next, $\varphi_{u+\tau}(v + \xi) = \varphi_u(v)$ for $0 \leq \tau < 1$ and $0 \leq \xi < 1$. Under this assumption, $p_u(v) = \exp(-\varphi_u(v)) \iff \varphi_u(v) = -\log(p_u(v))$.

We define the exposure-to-risk ($E_{u,v}$), measuring the time during which individuals are exposed to the risk of dying. It is the total time lived by these individuals. Assume that we have $L_{u,v}$ individuals at duration $u$ and age of occurrence $v$. Using the notation of Gschlössl et al. (2011), we associate to each of these $L_{u,v}$ individuals the dummy variable

$$
\delta_i = \begin{cases} 
1 & \text{if individual } i \text{ dies,} \\
0 & \text{otherwise},
\end{cases}
$$
psi for \( i = 1, 2, \ldots, L_{u,v} \). We define the time lived by individual \( i \) before \((u + 1)\text{st}\) duration when the pathology occurred at age \( v \) by \( \tau_i \). We assume that we have at our disposal iid observations \((\delta_i, \tau_i)\) for each of the \( L_{u,v} \) individuals. The contribution of individual \( i \) to the likelihood equals \( \exp(-\tau_i \varphi_u(v))(\varphi_u(v))^\delta_i \). Finally we define

\[
\sum_{i=1}^{L_{u,v}} \tau_i = E_{u,v} \quad \text{and} \quad \sum_{i=1}^{L_{u,v}} \delta_i = D_{u,v}.
\]

Under these assumptions, the likelihood becomes

\[
\mathcal{L}(\varphi_u(v)) = \prod_{i=1}^{L_{u,v}} \exp(-\tau_i \varphi_u(v))(\varphi_u(v))^\delta_i = \exp(-E_{u,v} \varphi_u(v))(\varphi_u(v))^{D_{u,v}}.
\]

The associated log-likelihood is \( \ell(\varphi_u(v)) = \log \mathcal{L}(\varphi_u(v)) = -E_{u,v} \varphi_u(v) + D_{u,v} \log \varphi_u(v) \). Maximizing the log-likelihood \( \ell(\varphi_u(v)) \) gives \( \hat{\varphi}_u(v) = D_{u,v}/E_{u,v} \) which coincides with the central death rates \( \hat{m}_u(v) \). Then it is apparent that the likelihood \( \ell(\varphi_u(v)) \) is proportional to the Poisson likelihood based on \( D_{u,v} \sim \text{Poisson}(E_{u,v} \varphi_u(v)) \) and it is equivalent to work on the basis of the true likelihood or on the basis of the Poisson likelihood, as recalled in Gschlössl et al. (2011). Thus, under the assumption of constant forces of mortality between non-integer values of \( u \) and \( v \), we consider

\[
D_{u,v} \sim \text{Poisson}(E_{u,v} \varphi_u(v)),
\]

(4.3)

to take advantage of the Generalized Linear Models (GLMs) framework.

### 4.4.2 Bi-dimensional local likelihood

We present briefly the generalization to two predictors. Suppose we have \( n \) independent realizations \( y_1, y_2, \ldots, y_n \) of the random variable \( Y \) with

\[
Y_i \sim f(Y|\theta(x_i)), \quad \text{for } i = 1, 2, \ldots, n,
\]

where \( f(\cdot|\theta(x_i)) \) is a probability mass/density function in the exponential dispersion family and \( \theta(x_i) \), the natural parameter in the GLMs framework, is an unspecified smooth function \( \psi(x_i) \). For simplicity, we use \( x_i = (u_i, v_i) \) to denote the vector of the predictor variables. The bivariate local likelihood fits a polynomial model locally within a bivariate smoothing window. Suppose that the function \( \psi \) has a \((p + 1)\text{st}\) continuous derivative at the point \( x_i = (u_i, v_i) \). For data point \( x_j = (u_j, v_j) \) in a neighborhood of \( x_i = (u_i, v_i) \) we approximate \( \psi(x_j) \) via a Taylor expansion by a polynomial of degree \( p \). If locally linear fitting is used, the fitting variables are just the independent variables. If locally quadratic fitting is used, the fitting variables are the independent variables, their squares and their cross-products. For example, a local quadratic approximation is:

\[
\psi(x_j) = \psi(u_j, v_j) \approx \beta_0(x_i) + \beta_1(x_i)(u_j - u_i) + \beta_2(x_i)(v_j - v_i) \\
+ \frac{1}{2}\beta_3(x_i)(u_j - u_i)^2 + \beta_4(x_i)(u_j - u_i)(v_j - v_i) + \frac{1}{2}\beta_5(x_i)(v_j - v_i)^2.
\]
The local log-likelihood can be written as

\[ L(\beta | \lambda, x_i) = \sum_{j=1}^{n} l(y_j, x^T \beta) w_j, \]  

(4.4)

where, in the case of locally quadratic fitting, \( x = (1, u_j - u_i, v_j - v_i, (u_j - u_i)^2, (v_j - v_i)(u_j - u_i), (v_j - v_i)^2)^T \) and \( \beta = (\beta_0, \ldots, \beta_5)^T \).

The weights are defined on the bivariate space. The non-negative weight function, \( w_j = w_j(x_i) \), depends on the distance \( \rho(x_i, x_j) \) between the observations \( x_j = (u_j, v_j) \) and the fitting point \( x_i = (u_i, v_i) \) and in addition, it contains a smoothing parameter \( h = (\lambda - 1)/2 \) which determines the radius of the neighborhood of \( x_i \).

Maximizing the local log-likelihood (4.4) with respect to \( \beta \) gives the vector of estimators \( \hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_5)^T \). Estimator \( \psi(x_i) \) is given by \( \hat{\psi}(x_i) = \hat{\beta}_0 \).

We proceed by forming the local likelihood as in (4.4) and estimate the coefficients \( \beta \) based on data in the neighborhood \( x_j = (u_j, v_j) \) of the target point \( x_i = (u_i, v_i) \).

Since we want to maximize the log-likelihood, we look for a solution of the set of normal equations to be fulfilled by the maximum likelihood parameter estimates \( \beta \). In case of locally quadratic fitting,

\[ \frac{\partial L(\beta_v | y, w_j, \phi)}{\partial \beta_v} = 0 \quad \text{for } v = 0, 1, \ldots, 5. \]

These equations are usually non-linear, so the solution must be obtained through iterative methods. One way to solve those is to use Fisher’s scoring method.

The derivatives of the local Poisson log-likelihood function with respect to \( \beta \) are

\[
\begin{align*}
\frac{\partial L}{\partial \beta_0} &= \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j} \frac{\partial \mu_j}{\partial \eta_j}; \\
\frac{\partial L}{\partial \beta_1} &= \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j} \frac{\partial \mu_j}{\partial \eta_j} (u_j - u_i); \\
\frac{\partial L}{\partial \beta_2} &= \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j} \frac{\partial \mu_j}{\partial \eta_j} (v_j - v_i); \\
\frac{\partial L}{\partial \beta_3} &= \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j} \frac{\partial \mu_j}{\partial \eta_j} (u_j - u_i)^2; \\
\frac{\partial L}{\partial \beta_4} &= \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j} \frac{\partial \mu_j}{\partial \eta_j} (v_j - v_i); \\
\frac{\partial L}{\partial \beta_5} &= \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j} \frac{\partial \mu_j}{\partial \eta_j} (v_j - v_i)^2.
\end{align*}
\]

(4.5)

The Fisher information for \( \beta \) is given, in matrix notation, by

\[ I_{\beta} = \left\{ X^T W \Omega X \right\}_{\beta}. \]
where $\mathcal{I}$ denotes the Fisher information matrix, $X$ is the design matrix

$$
X = \begin{bmatrix}
1 & u_1 - u_i & v_1 - v_i & (u_1 - u_i)^2 & (u_1 - u_i)(v_1 - v_i) & (v_1 - v_i)^2 \\
1 & u_2 - u_i & v_2 - v_i & (u_2 - u_i)^2 & (u_2 - u_i)(v_1 - v_i) & (v_1 - v_i)^2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & u_n - u_i & v_n - v_i & (u_n - u_i)^2 & (u_n - u_i)(v_1 - v_i) & (v_1 - v_i)^2 
\end{bmatrix}
$$

and $\Omega$ is the matrix of the working weights just as in (3.11), while $W$ is a diagonal matrix, with entries $\{w_j\}_{j=1}^n$, such that

$$
w_j = \begin{cases}
W(\rho(x_i, x_j)/h) & \text{if } \rho(x_i, x_j)/h \leq 1, \\
0 & \text{otherwise.}
\end{cases} \quad (4.6)
$$

$W(\cdot)$ denotes a non-negative weight function depending on the distance $\rho(x_i, x_j)$. A common choice is the Euclidean distance,

$$
\rho(x_i, x_j) = \sqrt{(u_j - u_i)^2 + (v_j - v_i)^2}.
$$

In addition, it contains a smoothing parameter $h = (\lambda - 1)/2$ which determines the radius of the neighborhood of $x_i$. The two components of the Euclidean distance can be scaled in order to apply more smoothing in one direction than the other.

Following the general Fisher scoring procedure, see Section 3.2.4, we obtain the estimates.

When modeling experience data from life-insurance, we wish generally to take into account the exposure in the setting. Specifically, we are looking for a smooth estimate of the observed forces of mortality and from equation (4.3) the linear predictor $\eta_j$ can be written as

$$
\eta_j = \log(\mathbb{E}[Y | X = x_j]) = \log(\mu_j) = \log(E_j \varphi_j) = \log(E_j) + \log(\varphi_j)
$$

The term $E_j$ called the offset can be easily incorporated.

### 4.4.3 $p$-splines framework for count data

In this section, we present the essential background material on $p$-splines methodology for count data. Descriptions of the $p$-splines method can be found in the seminal paper of Eilers and Marx (1996), as well as in Marx and Eilers (1998), Eilers and Marx (2002), and in Currie and Durbán (2002). Currie et al. (2006) present a comprehensive study of the methodology. Applications covering mortality can be found in Currie et al. (2004), Richards et al. (2006), Kirkby and Currie (2010) and in the Ph.D. thesis of Camarda (2008). Planchet and Winter (2007) use the same framework to discuss an application concerning sick leave retentions.
Again, we suppose that the data can be arranged as a column vector, $y = \text{vec}(Y) = (y_1, y_2, \ldots, y_n)^T$. Let $B_u = B(u)$ and $B_v = B(v)$, be regression matrices, of dimensions $n_u \times k_u$ and $n_v \times k_v$, of B-splines based on the duration $u$ and age of occurrence $v$, respectively, with $k$ denoting the number of internal knots.

Specifically, B-splines are bell-shaped curves composed of smoothly joint polynomial pieces. Polynomials of degree 3 are used in the following. The positions on the horizontal axis where the pieces come together are called knots. We use equally spaced knots. The numbers of columns of $B_u$ and $B_v$ are related to the number of knots chosen for the B-splines. Details on B-splines can be found in de Boor (2001).

The regression matrix for our two dimensional model is the Kronecker product

$$B = B_u \otimes B_v$$

The matrix $B$ has an associated vector of regression coefficients $a$ of length $k_u k_v$. As in the GLM framework, the linear predictors $\eta$ is linked to the expectation of $y$ by a link function $g(.)$.

$$\eta = g(\mathbb{E}[y]) = \log(\mu) = B a = (B_u \otimes B_v) a,$$

(4.7)

The elements of $a$ can be arranged in a $k_u \times k_v$ matrix $A$, where $a = \text{vec}(A)$. The columns and rows of $A$ are then given by $A = (a_1, \ldots, a_u)$ and $A^T = (a_1, \ldots, a_v)$. Then instead of computing equation (4.7) as a vector, it can be written as

$$\log(\mathbb{E}[y]) = \log(M) = B_u A B_v^T.$$ 

(4.8)

From the definition of the Kronecker product, the linear predictor of the columns of $Y$ can be written as linear combinations of $k_v$ smooths in the duration $u$. The linear predictors corresponding to the $j$th column of $Y$ can be expressed as

$$\sum_{k=1}^{k_v} b_{jk}^v B_u a_k,$$

where $B_v = b_{ij}^v$. We apply a roughness penalty to each of the columns of $A$. The penalty is given by

$$\sum_{j=1}^{k_v} a_j^T D_u^T D_u a_j = a^T \left( I_{k_v} \otimes D_u^T D_u \right) a,$$

where $D_u$ is the second order difference matrix acting on the columns of $A$. Similarly by considering the linear predictor corresponding to the $i$th row of $Y$,

$$\sum_{i=1}^{k_u} a_i^T D_v^T D_v a_i = a^T \left( D_v^T D_v \otimes I_{k_u} \right) a,$$
where $D_v$ is the second order difference matrix acting on the rows of $A$.

The penalized log-likelihood to be maximized can be written as

$$\ell^* = \ell(a; B; y) - \frac{1}{2} a^T P a,$$

(4.9)

where $\ell(a; B; y)$ is the usual log-likelihood for a GLM and the penalty term $P$ is given by

$$P = \lambda_u \left( I_{k_u} \otimes D_u^T D_u \right) + \lambda_v \left( D_v^T D_v \otimes I_{k_v} \right),$$

where $\lambda_u$ and $\lambda_v$ are the smoothing parameters used for the duration and the age of occurrence respectively, $I_{k_u}$ and $I_{k_v}$ being identity matrices of dimension $k_u$ and $k_v$ respectively. More details can be found in Currie et al. (2004).

Then maximizing equation (4.9) gives the penalized likelihood equations

$$B^T (y - M) = Pa,$$

which can be solved by a penalized version of the IRWLS algorithm,

$$\left( B^T \Omega B + P \right) a = B^T \Omega z,$$

(4.10)

where $\Omega$ is the matrix of the working weights similar to (3.11). Again in case of Poisson errors, $\Omega = \text{diag}(\mu)$. The working dependent variable $z$ is defined by

$$z = B a + \frac{y - \mu}{\mu}.$$

Hence, a maximum likelihood estimate of $a$ is found by a penalized version of IRWLS algorithm:

Repeat $a^* := B(B^T \Omega B + P)^{-1} B^T \Omega z$;

using $a^*$, update the working weights $\Omega$, as well as the working dependent variables $z$ until convergence.

Again when modeling mortality data, we may take into account the exposure in the setting. The linear predictor $\eta$ can be written as

$$\eta = g(\mathbb{E}[y]) = \log(\mu) = \log(e) + \log(\varphi) = \log(e) + B a = \log(e) + (B_u \otimes B_v) a,$$

where $e$ denotes the vector of exposure. Similarly to Section 4.4.2, the offset can be easily incorporated in the regression system (4.10).

The penalized IRWLS would be efficient only in moderate-sized problems. For our application, the parameter vector $a$ has length 2520 and this required the usage of $2520 \times 2520$ matrices. The size is moderate, but for larger dimensional matrices the penalized IRWLS algorithm can run into storage and computational difficulties. Currie et al. (2006) and Eilers et al. (2006) proposed an algorithm that takes advantage of the special structure of both
the data as a rectangular array and the model matrix as a tensor product. The idea of this algorithm can be seen in the computation of the mean $\mu = \text{vec}(M)$ in two dimensions, as in (4.8). This avoids having to construct a large Kronecker product basis, saving time and space. For the presentation of this algorithm we refer to the mentioned articles Currie et al. (2006) and Eilers et al. (2006).

The smoothing parameters for $p$-splines method are chosen according the Bayesian information criterion (BIC) which penalizes heavily the model complexity particularly when $n$ is large,

$$BIC = \sum_{i=1}^{n} D(y_i, (\theta(\hat{\mu}_i))) + \log(n) \upsilon.$$

### 4.4.4 The data

The data come from observations of individuals subscribing to LTC insurance policies originating from a portfolio of a French insurance company. For these applications, we focus on measuring the forces of mortality as a function of the age $v$ of occurrence of the pathologies and the duration $u$ of the care.

The range of ages of occurrence is $70 - 90$ and the maximum duration of the pathologies is 119 months. The period of observation stretches from 01/01/1998 to 31/12/2010. The data have been aggregated according to the age of occurrence and the duration. The pathologies are composed, among others, by dementia, neurological illness and terminal cancer. The data consist for $2/3$ of women and $1/3$ of men. Figures 4.11a, 4.11b, and 4.11c display the observed statistics of the dataset.

Moreover, we have at our disposal the adjusted surface obtained from the technical report Planchet (2012), Figure 4.11d. It gives an idea about the desirable shape that we aim to retain, and the adjusted forces of mortality will be useful when assessing the comparisons of the models. This surface has been obtained by treating separately the first month of duration from the others and applying a Whittaker-Henderson model to adjust the crude surface.

### 4.4.5 Smoothed surfaces and fits

Figure 4.11e presents the smoothed surface obtained with the local likelihood model with an Epanechnikov weight function, a polynomial of degree 2 and a bandwidth (radius) of 13 observations. The corresponding degrees of freedom $\upsilon$ are 29.25. The order of polynomial and the bandwidth have been chosen by minimizing the $AIC$ criterion. The surface is relatively wiggly showing an inappropriate variance.
Figure 4.11: Observed statistics: $E_{u,v}$, $D_{u,v}$, and smoothed forces of mortality $\hat{\psi}_{u,v}$ according to Planchet (2012), local likelihood, $p$-splines, ICI rule and local bandwidth factors methods.

(a) Number of exposures to the risk, $E_{u,v}$
(b) Number of death, $D_{u,v}$
(c) Crude forces of mortality, $\psi_{u,v}$
(d) $\hat{\psi}_{u,v}$, Planchet (2012)
(e) $\hat{\psi}_{u,v}$, local likelihood
(f) $\hat{\psi}_{u,v}$, $p$-splines
(g) $\hat{\psi}_{u,v}$, ICI
(h) $\hat{\psi}_{u,v}$, local bandwidth factors
Figure 4.11f displays the smoothed surface obtained when fitting p-splines. The smoothing parameters $\lambda_u = 31.6, \lambda_v = 31.6$, have been chosen by minimizing the BIC criterion. It leads to $k_u = 24, k_v = 4$ for $\nu = 18.11$. The surface seems satisfactory, though the increase in the upper right corner (highest age of occurrence and longest duration) is not present as in the surface adjusted from Planchet (2012).

Figures 4.11g and 4.11h present the smoothed surface obtained with the adaptive local likelihood methods. For these applications, only the bandwidth is varying. The order of polynomial is still fixed at 2 and we use an Epanechnikov weight function. The fitted degrees of freedom $\nu$ are 10.05, 10.76 and 16.16 respectively.

In general, only for the first months of the duration, the graduations are similar. After that, we obtain very different shapes according to the models. The ICI rule and the local bandwidth factors seem the most satisfying methods in modeling the monotone phenomenon at the extreme ages, Figures 4.11g and 4.11h. The fitted degrees of freedom for the local bandwidth factors are larger than the ones obtained by the ICI rule indicating that the model is slightly more flexible and shows more features. The bandwidth values depend on the amount of exposure to represent effectively the remaining life expectancy in the regions where the amount of exposure is high. The corresponding bandwidths, in the left region, are relatively low, and they increase as the amount of exposure decreases. For regions in which the amount of exposure is low, a large value for the bandwidth results in an estimate that progress more smoothly. As we already obtained the shape and the magnification of the local bandwidth factors, we used the AIC criterion to decide the global value at which the bandwidth curve is located. The sensitivity parameter $s$ for the local bandwidth factors as well as the value $c$ for the ICI rule have been chosen arbitrarily to be 0.15 and 0.1 respectively. For higher value of $s$ spurious features started to appear showing unacceptable variance, while for higher $c$, bias tends to show up.

Figures 4.12 and 4.13 present the smooth fits obtained from the different models for various ages of occurrence and durations.

The approaches produce relatively similar graduations for regions having a low amount of noise, Figure 4.12b. However, when the data are more volatile the benefits of the adaptive approaches become apparent. The fit obtained from global local likelihood, and not as strongly the p-splines, present an unacceptably high variance. It shows the inapplicability to model such datasets with global methods or to select the smoothing parameters by relying explicitly on a criterion. The local bandwidth factors method has the capability to model the forces of mortality in the first months of duration relatively well, Figure 4.12d, and the sharp increase at the highest extremes of the age of occurrence and duration, Figures 4.12c and 4.12f. The ICI rule
Figure 4.12: Observed forces of mortality and smooth fits for various ages of occurrence.

- (a) Age of occurrence $v = 70$
- (b) Age of occurrence $v = 80$
- (c) Age of occurrence $v = 90$

Figure 4.13: Observed forces of mortality and smooth fits for various durations.

- (d) Duration $u = 0$
- (e) Duration $u = 60$
- (f) Duration $u = 119$
and p-splines fail to model these features. However, all the models miss the slow increase at the age of occurrence 70 present in the fit obtained from Planchet (2012), Figure 4.12a.

4.4.6 Analysis of the residuals

Figure 4.14 presents the residuals of the 5 models for the age of occurrence 70 as well as the ones obtained from the adjusted surface from Planchet (2012).

![Figure 4.14: Response, Pearson and deviance residuals for the age of occurrence 70](image)

The pattern of the residuals displayed for each model is roughly similar. We superimposed a loess smooth curve on the response and Pearson residuals. These smooths help search for clusters of residuals that may indicate...
lack of fit. By reducing the noise, our attention may be more readily drawn to features that have been missed or not properly modeled by the smooth. Here the process is not to judge a fit adequate if a smooth curve on its residuals plot is flat. A flat curve means simply that no systematic, reproducible lack of fit has been detected. The fit may well be too noisy, and stays to close to an interpolation since trends in small parts of the data are interpreted as more widespread trends. Then for small datasets, the fit is very nearly interpolating the data which results in unacceptably high variance. Strong patterns appear in the response residuals in Figure 4.14. It indicates a lack of fit in this region. However, this is not surprising as most of the deaths at the longest durations are zero for the age of occurrence 70.

The Pearson residuals are mainly in the interval $[-2, 2]$, which indicates that the models adequately capture the variability of the dataset.

The deviance residuals present, for the longest durations, several successive residuals having the same sign. It illustrates that the forces of mortality are over-smoothed locally. As the sign is negative, from 80 to 119 months, we strongly overestimate the forces of mortality. However, we would have excepted such a pattern as we observe zero deaths at the highest extreme of the duration of the care.

### 4.5 Comparisons

#### 4.5.1 Tests to compare graduations

We continue the comparisons by applying the tests proposed by Forfar et al. (1988, p.56-58) and Debón et al. (2006, p.231). We have also obtained the values of the mean absolute percentage error $MAPE$ and $R^2$ used in Felipe et al. (2002). We compare the crude mortality rates to the graduated series to see whether the approaches lead to similar graduation. Table 4.3 presents the results.

The approaches display different results. The global local likelihood approach, having the highest degrees of freedom, has the capacity to show many features in the data. Therefore, the values of the various tests are the best. It has the lowest deviance, lowest number of standardized residuals exceeding the thresholds 2 and 3, highest number of runs, best mix of the residuals between positive and negative signs, highest value for the run test. In addition, the approach results in the minimum $\chi^2$ and $MAPE$. Conversely, the adaptive local likelihood using the ICI rule yields the smallest degrees of freedom. As a consequence, the results of the various tests and values of the $\chi^2$ and $MAPE$ are the worst.

The results for the $p$-splines and the adaptive approaches using the local bandwidth factors are similar, even though we have seen that the adaptive method has a better ability to model the mortality patterns (high mortality for the first month of duration of the care and increase at the extreme highest
of the duration and age of occurrence), even though the \( p \)-splines model has higher degrees of freedom.

### 4.5.2 Comparing figures summarizing the lifetime probability distribution

We end these comparisons by presenting some figures summarizing the lifetime probability distribution. Figures 4.15 and 4.16 display the life expectancy obtained from the different models for various ages of occurrence and durations.

At age of occurrence 70, with the exception of the adaptive local likelihood using the ICI rule and local bandwidth factors, the models are over-estimating the period life expectancy for the first months of duration (until 10 months), Figure 4.15a. This is particularly visible for the \( p \)-splines and the adjusted surface obtained in Planchet (2012). The over-estimation is general at age of occurrence 80, Figure 4.15b. The shapes of the life expectancy differ much at age of occurrence 90, Figure 4.15c, where the global local likelihood tends to estimate a more rectangular shape.

The shape and trend of the life expectancies are similar when we observe a large amount of exposure (first months of duration of the care), Figure 4.15d. The high correlation of the pathologies with the age of occurrence can explained the concave shape observed for the life expectancies during

<table>
<thead>
<tr>
<th></th>
<th>Local lik.</th>
<th>( p )-splines</th>
<th>Adapt. lik. ICI</th>
<th>Adapt. band. factors</th>
<th>Planchet (2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitted DF ( \nu )</td>
<td>29.25</td>
<td>18.11</td>
<td>10.76</td>
<td>16.16</td>
<td>NA</td>
</tr>
<tr>
<td>Deviance</td>
<td>2259.91</td>
<td>2291.89</td>
<td>2440.81</td>
<td>2311.04</td>
<td>2409.99</td>
</tr>
<tr>
<td>Standardised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>residuals</td>
<td>&gt; 2</td>
<td>108</td>
<td>117</td>
<td>127</td>
<td>115</td>
</tr>
<tr>
<td>&gt; 3</td>
<td>25</td>
<td>31</td>
<td>38</td>
<td>32</td>
<td>42</td>
</tr>
<tr>
<td>Signs test</td>
<td>+ (−)</td>
<td>910(1610)</td>
<td>907(1613)</td>
<td>891(1629)</td>
<td>900(1620)</td>
</tr>
<tr>
<td>p-value</td>
<td>&lt; 2.2e − 16</td>
<td>910(1610)</td>
<td>907(1613)</td>
<td>891(1629)</td>
<td>900(1620)</td>
</tr>
<tr>
<td>Runs test test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nb of runs</td>
<td>1028</td>
<td>1016</td>
<td>971</td>
<td>1019</td>
<td>1013</td>
</tr>
<tr>
<td>Value</td>
<td>−6.25</td>
<td>−6.71</td>
<td>−8.19</td>
<td>−6.47</td>
<td>−6.64</td>
</tr>
<tr>
<td>p-value</td>
<td>4.05e − 10</td>
<td>1.98e − 11</td>
<td>2.57e − 16</td>
<td>9.69e − 11</td>
<td>3.10e − 11</td>
</tr>
<tr>
<td>Kolmogorov test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>0.4401</td>
<td>0.5008</td>
<td>0.4829</td>
<td>0.4786</td>
<td>0.5167</td>
</tr>
<tr>
<td>p-value</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>χ²</td>
<td>2433.56</td>
<td>2487.09</td>
<td>2616.08</td>
<td>2458.71</td>
<td>2644.09</td>
</tr>
<tr>
<td>R²</td>
<td>0.2406</td>
<td>0.2340</td>
<td>0.2433</td>
<td>0.2476</td>
<td>0.2234</td>
</tr>
<tr>
<td>MAPE (%)</td>
<td>46.11</td>
<td>46.62</td>
<td>48.18</td>
<td>46.99</td>
<td>47.38</td>
</tr>
</tbody>
</table>

Table 4.3: Comparisons between the smoothing approaches.
Figure 4.15: Observed and predicted life expectancy for various ages of occurrence.

(a) Age of occurrence \( v = 70 \)

(b) Age of occurrence \( v = 80 \)

(c) Age of occurrence \( v = 90 \)

Figure 4.16: Observed and predicted life expectancy for various durations.

(d) Duration \( u = 0 \)

(e) Duration \( u = 60 \)

(f) Duration \( u = 119 \)
the first months of the care. The lowest ages of occurrence are marked by a relatively high mortality mainly due to the death of the individuals suffering from terminal cancer, while the highest ages concern principally the dementia. At the 60th month of the care, the life expectancy is decreasing rapidly, Figure 4.15e. However, while the ICI rule and local bandwidth factors produce similar patterns, the shapes and trends given by the other models diverge markedly, the local likelihood predicting a rise of the life expectancy for the highest age of occurrence. This pattern is also present, although less markedly, in Figure 4.15f.

Figure 4.17 shows the median month at death, Figure 4.17a, standard deviation of the random life time, Figure 4.17b and entropy, Figure 4.17c, as a function of the age of occurrence of the pathology.

In Figure, 4.17a displaying the median month at death as a function of the age of occurrence of the pathology, we observe a concave shape similar to Figure 4.15d. This phenomenon shows, once more, the correlation between the age of occurrence and the pathologies. The adaptive local likelihood using the ICI rule, having the lowest degrees of freedom, mostly underestimates the median month at death compared to the others models.

After a steady increase, the standard deviation of the random lifetime is slowing down at age of occurrence 82, and decreases until 90 years old, Figure 4.17b. It is explained by the fact that we observe most of the deaths at the lowest age of occurrence and duration, while the number of deaths is zero, and thus stable, at the highest age and duration.

Figure 4.17c shows the entropy. The values decline as the deaths become more concentrated. We observe that the deaths predicted by the adaptive local likelihood models (ICI rule and bandwidth factors) are the most stretched. Conversely, the adjusted number of deaths obtained by Planchet (2012) are more concentrated.

Table 4.4 summarizes the indices. For the life expectancy, $0_{120}e_{70}$, $0_{120}e_{80}$, and $0_{120}e_{90}$, the observations suggest an increase with the age, which, based on our knowledge, is unrealistic. We are more likely to look for a concave shape, predicted by the models as displayed in Figure 4.15d. On average, the models agree on the same life expectancy, around 38 months and underestimate slightly the observed $0_{120}e$.

The median month at death, Med($T_0$), estimated by the models varies in average slightly from 25 to 27 months. However, for a particular age of occurrence, such as Med($T_0(70)$), the difference between the models ($p$-splines and ICI rule) can grow until 6 months.

All the models sensibly estimated the same average standard deviation of the random life time, $\sigma_0$, which corresponds to the observed standard deviation, around 0.22.

Finally, all the models agree on the estimated average entropy $H(T_0)$,
Figure 4.17: Median month of death, standard deviation of the random life time and entropy with the age of occurrence of the pathology.

(a) Median month of death
(b) Std. dev. of life time
(c) Entropy
4.6 Summary and Outlook

In this chapter, we illustrate how adaptive local likelihood methods can be used to graduate mortality tables in two dimensions. Tests and single indices summarizing the lifetime distributions are used to compare the graduated forces of mortality obtained from adaptive local likelihood to global non-parametric methods such as local likelihood and \( p \)-splines models.

Using locally adaptive parameters instead of a global smoothing one may be advantageous for several reasons. The estimator can adapt to the structure of the regression function and to the reliability of the data, smoothing more when the volume of observations is low and less when it is high.

The intersection of confidence intervals (ICI) rule has been introduced as a locally adaptive pointwise method. The critical value controls the bias-variance tradeoff. Because a larger class of estimators is available, it may in turn affect the variability. Hence, the set of window sizes contains relatively large bandwidths. The choice of the set of window sizes is done subjectively,
based on the mechanism generating the data and on the performance of the smoother used in the fitting. Another drawback in applying such methods is that they require more computer time than a global procedure. Specifically, the computational effort is multiplied by the number of observations.

A technique closely related to the ICI rule is the Lepski method. This procedure uses the standard deviation of the difference $\hat{\psi}_{\lambda_1}(x_i) - \hat{\psi}_\lambda(x_i)$ for some $\lambda \leq \lambda_1$ until a significance difference is found. Chichignoud (2010, Section 1.5) provides an extensive discussion of the technique in his recent Ph.D Thesis. The discussion and the implementation of the Lepski method for graduating experience data originating from life insurance is a topic of ongoing research.

The bandwidth correction factors method allows the estimated forces of mortality to include explicitly the extra information provided by the changing amounts of exposure. The observed exposure decided the shape of the smoothness parameter. The magnification of that shape has been determined by a sensitivity parameter which we chose subjectively for practical reasons. The global bandwidth parameter is used to control the absolute level of the bandwidth curve. We used a global criterion instead of pointwise methods. It appears that the procedure has the ability to model relatively well the mortality pattern where the other models fail to model these features.

In global procedures as well as for locally adaptive procedures, there is no deterministic method to obtain the constellation of smoothing parameters with the classical selectors. Residual analysis and goodness of fit diagnostics are just as important for locally adaptive procedures as they are for global procedures. It is important to use appropriate residuals diagnostics to look for lack of fit. The purpose for which the mortality table is required must be kept clearly in mind, and the final choice of graduation is always a matter of judgment.

The methodologies proposed adapt neatly to the complexity of mortality surface, clearly because of the appropriate data-driven choice of the adaptive smoothing parameters. The adaptive local likelihood method using the bandwidth factors models well the high mortality during the first months of duration and the increase at the extreme high duration and age of occurrence compared to the other methods. Having 13 degrees of freedom less than the local likelihood model, the adaptive bandwidth factors model is less flexible although the tests presented in Table 4.3 show relatively good results. Rather than treating the first months separately, having an adaptive model can be a benefit. However, the relative merit of the procedures would depend on the purpose for which the mortality table has been computed. If we are essentially exploring the data, then additional information derived might not justify the effort. However, the potential uses of adaptive approaches suggest that they have much to offer as part of the actuarial toolkit.