Júlio Filipe Oliveira Lisboa da Silva

Estimação da taxa de incidência da infeção por VIH em Portugal

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
2015

Júlio Filipe Oliveira Lisboa da Silva

Estimação da taxa de incidência da infeção por VIH em Portugal

Tese submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Mestre em Engenharia Matemática, com a orientação de Prof. Dra. Rita Gaio e Prof. Dr. Joaquim Costa

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
2015

Agradecimentos

A realização desta dissertação marca o fim de uma das mais importantes fases da minha vida. Deixo algumas palavras, poucas, mas com um profundo agradecimento, a quem contribuiu para a sua concretização.

À Prof.ª Doutora Rita Gaio e ao Prof. Doutor Joaquim Costa pela orientação, disponibilidade e apoio incondicionais que muito contribuíram para elevar os meus conhecimentos científicos. Hoje, sinto-me uma pessoa com maior vontade de querer saber sempre mais, graças ao estímulo académico e profissional que sempre me transmitiram.

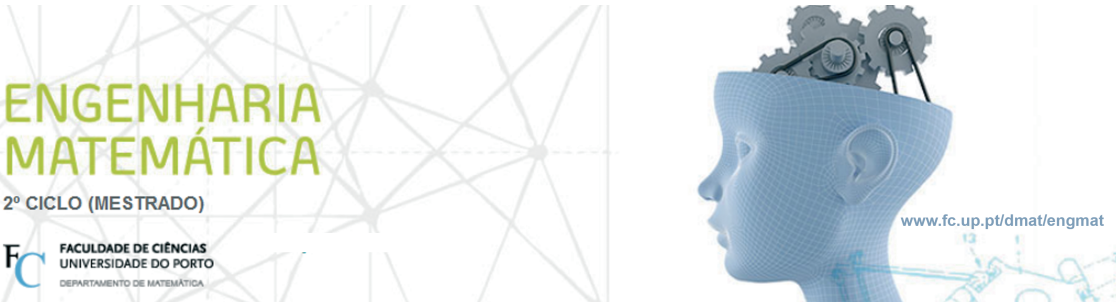
À Alexandra Oliveira, pela sua total disponibilidade que sempre revelou para comigo e pelo apoio constante, em particular, nos momentos de maior nervosismo.

A todos os professores que me transmitiram o fascínio pela Matemática, em particular aos que fazem parte do Mestrado em Engenharia Matemática, que me ensinaram a beleza e a complexidade das diversas aplicações da Matemática.

Aos meus parceiros de gabinete, pela vossa amizade, companheirismo e ajuda, fundamentais para a realização desta dissertação.

À minha namorada, um agradecimento muito especial por fazeres parte da minha vida. Sempre demonstraste compreensão e apoio, mesmo nos momentos mais difíceis. Acredito que o caminho que já percorremos juntos vai-se prolongar para o resto das nossas vidas.

À minha família, por acreditarem sempre em mim, transmitindo-me sempre confiança, força e apoio incondicionais. Sem os vossos ensinamentos que me foram transmitindo ao longo dos anos, seria impossível terminar esta fase. Espero que, de alguma forma, este trabalho possa retribuir tudo o que fizeram por mim.



Tese realizada no âmbito do mestrado em Engenharia Matemática
Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
http://www.fc.up.pt/dmat/engmat

Resumo

O vírus da imunodeficiência humana (VIH) provoca um enfraquecimento progressivo do sistema imunitário, destruindo a capacidade de defesa do organismo em relação a muitas doenças. Um indivíduo infectado por VIH diz-se sero-positivo. Se no início a infecção por VIH pode passar despercebida, no final conduz à morte. Quando o agravamento do sistema imunitário é grave (contagem de células T CD4+ abaixo de 200 células por μL de sangue ou ocorrência de doenças específicas), a infecção passa a ser designada por Síndrome de Imunodeficiência Adquirida (SIDA). Esta infeção pode ter um grande impacto nos países afetados, nomeadamente nas suas economias e níveis sociodemográficos.

Em Portugal, tal como em grande parte dos países, esta epidemia é monitorizada por um sistema de vigilância baseado em notificações da infecção feitas pelos médicos e é realizado pelo Instituto Nacional de Saúde Dr. Ricardo Jorge. Existem, pelo menos, dois problemas associados a este mecanismo: um caso de infecção diagnosticado pode demorar vários meses até ser notificado (fenómeno designado por atraso na notificação) e um caso pode nunca ser notificado (fenómeno designado por sub-notificação). Para além disso, como a infeção por VIH não é, em geral, imediatamente detetada, a informação sobre o tempo de incubação é escassa e só está disponível para certos grupos de risco, não sendo portanto uma informação representativa da população.

O objetivo deste trabalho consiste na estimação da taxa de incidência da infecção por VIH em Portugal a partir dos dados obtidos pelo sistema de vigilância Português e tendo em conta os problemas atrás mencionados. Esta informação assume uma importância elevada pois permitirá avaliar atuais e antigos planos de prevenção, bem como conduzir à definição de novos planos.

O problema dos atrasos na notificação foi abordado por outros autores usando verosimilhança condicional [Harris, 1990, Oliveira et al., 2014]. A questão da sub-notificação é feita nesta tese (tanto quanto temos conhecimento, pela primeira vez) usando a base de dados nacional do Sistema de Vigilância Epidemiológica da Tuberculose e um método de correção para a linearidade desenvolvido, noutro contexto, por [DeGruttola et al., 1991]. Fornecendo necessariamente uma sub-aproximação dos números verdadeiros, não deve ser excluída na medida em que a tuberculose é das infeções oportunistas com maior incidência nos indivíduos sero-positivos.

A estimação da taxa de incidência da infecção por VIH fez uso do método de retro-propagação [Bacchetti et al., 1993]. Fixando um mês (ou qualquer outra unidade de tempo), o número de novos infectados em meses anteriores foi estimado utilizando o número de diagnosticados nesse mês e a distribuição do tempo de incubação. Aqui, o tempo de incubação foi definido como sendo o tempo que medeia entre a data de diagnóstico de SIDA e a data de infecção por VIH. Usando a informação disponível na literatura, a sua

distribuição foi assumida com sendo uma Weibull ou Gama [Amaral et al., 2005]. Assumiu-se ainda que o número de diagnosticados e o número de infetados num certo mês segue uma distribuição de Poisson. O método baseia-se numa equação de convolução, que é adaptada para o caso discreto, [Bacchetti et al., 1993]. O número de infectados por mês corresponde a um dos parâmetros desconhecidos de uma função de verosimilhança adequada. Foram utilizados dois métodos numéricos para a maximização da verosimilhança: método de Newton-Raphson e algoritmo Estimação-Maximização (EM), [Green, 1990]. São também sugeridas duas modificações à distribuição do tempo de incubação.
O método de retro-propagação foi aplicado a dados nacionais reais. Para a sub-notificação, os resultados parecem ser congruentes com a situação real visto que até 2005 se verificou uma taxa de sub-notificação elevada, a variar entre 13 e 20 %, e a partir de 2005 (ano em que a infeção pelo VIH passou a integrar a lista das doenças de declaração obrigatória), as estimativas foram mais baixas, como seria de esperar. O método de retro-propagação previu o número de infeções de SIDA que pareceu concordante com o conhecimento epidemiológico que se tem da infeção, para o tempo de incubação independente do tempo. Aparentemente, as duas modificações efetuadas que permitiram uma dependência da distribuição do tempo de incubação com o tempo não produziram resultados satisfatórios. A metodologia inversa à retro-propagação permitiu também obter projeções para o número de casos de SIDA até 2020. Essas projeções colocam Portugal numa situação complicada abrindo portanto lugar ao desenvolvimento de novos programas estratégicos de saúde pública de combate à infeção.

Abstract

The human immunodeficiency virus (HIV) causes a gradual weakning of the immunity system, destroying the organisms defenses to many diseases. An HIV infected individual is called seropositive. If the beginning of the HIV infection can pass unnoticed, at the end it leads to death. When the immunity system is greatly affected (count of T CD4+ cells is bellow 200 cells per μL of blood or when other specific diseases occur), the infection starts to be designated Acquired Immunodeficiency Syndrome (AIDS). This infection can have a great impact on the economy and on the sociodemographics levels of the affected countries. In Portugal, like in many of the affected countries, this epidemic is monitorized by a surveillance system based on notifications of infections made by the doctors. This system is administered by Nacional Institute of Health Dr. Ricardo Jorge. However, there are two problems in this system: a diagnosed case can take up to several months to be reported (designated reporting delay) or it can never be reported at all (designated under-reporting). Besides that, as the HIV infection is usually not immediately detected, information about the incubation time is scarce, being available only to certain risk groups. This means that this information may not be representative of the population.

The purpose of this work is to estimate the incidence rate of HIV in Portugal, through data obtained from the portuguese surveillance system, taking into account the problems mentioned above. The information about the incidence rate of HIV is of great importance because it allows to evaluate both old and new prevention interventions as well as lead to creation of new interventions.

The problem about the reporting delays was addressed by other authors using conditional likelihood [Harris, 1990, Oliveira et al., 2014]. The under-reporting problem was made in this thesis (as far as we know for the first time) using the national database of Tuberculosis Epidemiological Surveillance System and a correction method for the linearity, which was developed, for another context, by [DeGruttola et al., 1991]. Providing a sub-approximation of the real number, it must not be excluded because Tuberculosis is one of the opportunistic infection with greater incidence in the seropositive individuals.

The estimation of the incidence rate of HIV infection was addressed by the backcalculation method [Bacchetti et al., 1993]. For a fixed month (or any other time unit), the number of newly infected individuals in the previous months was estimated from the number of diagnosed cases at that month and from the distribution of the incubation time (taken to be the difference between the diagnosis date and the infection date). The latter was assumed to follow a Weibull or a Gamma distribution, [Amaral et al., 2005]. The number of new diagnoses and new infected cases in any given month was assumed to follow a Poisson distribution. The method is based on a convolution equation adapted for the discrete case, [Bacchetti et al., 1993]. The number of infected individuals per month corresponds to an

unknown parameter of the adequate likelihood function. For maximixation of the likelihood, two numerical methods were used, Newton-Raphson Method and Expectation-maximization algorithm [Green, 1990]. Two changes were made in the backcalculation method.
Both methodologies mentioned above were applied to real data. The results of the under-reporting analysis are close to the real data, because till 2005 the under-reporting rate was high, between 13 and 20 %. From 2005 forward (the year which the infection of HIV became part of the list of diseases of mandatory notification) the estimated data for the under-reporting are lower, as expected. The method of backcalculation predicted the number of AIDS infections that seemed consistent with the epidemiological knowledge that we have from infection, with the incubation time distribution independent from time. Apparently, two changes that led to a dependence distribution of incubation time with the time has not produced satisfactory results. The reverse approach to the backcalculation also yielded projections for the number of AIDS cases by 2020. These projections put Portugal in a complicated situation thus opening way to the development of new strategic public health program to combat infection.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

O vírus da imunodeficiência humana (VIH) é um vírus que se espalha rapidamente pelo corpo, afetando células específicas do sangue, denominadas células CD4. Estas células têm como função coordenar a defesa imunitária, defendendo o corpo de intrusos exteriores. Ora, uma diminuição do número de células CD4 provoca consequentemente um enfraquecimento do sistema imunitário, tornado-o vulnerável a ataques exteriores. Quando o corpo deixa de ser capaz de lutar contra os ataques exteriores, significa que a infecção por VIH passou para a fase SIDA. Ao contrário de inúmeros vírus, uma vez estando infetado com VIH, é impossível remover o vírus do corpo. Para além disso, ainda não existe cura para o vírus, apesar dos inúmeros esforços da comunidade científica. Existem porém, diversos tratamentos cujo objetivo é melhorar as condições de vida dos indivíduos infetados. Ao longo dos anos, graças aos diversos tratamentos criados, foi possível aumentar bastante o tempo entre a infeção por VIH e a transição para a fase SIDA (denominado tempo de incubação), quando antes destes, essa transição ocorria em poucos anos. Esta infeção, pelo efeito que provoca, quer a nível individual, quer a nível de saúde pública, merece uma atenção especial por parte de toda a comunidade médica mundial, sendo mesmo declarado como uma epidemia. Em Portugal, o vírus foi detetado, pela primeira vez, em 1983, e até 1999, verificou-se sempre uma tendência crescente do número de casos notificados. Desde então, esse número tem vindo a decrescer. No entanto, Portugal, segundo o relatório da Organização Mundial de Saúde [ECDC/WHO Regional Office for Europe, 2012], é o sexto país com maior taxa de diagnósticos em 2011, para os países da Região Europeia da Organização Mundial da Saúde (que inclui um total de 53 Estados). Por este motivo, é essencial combater esta epidemia em Portugal de forma a que o nosso país possa, a curto prazo, atingir a média europeia.

Em 1985, foi criado o Sistema de Notificação de casos de infeção por VIH/SIDA, cujo objetivo é recolher informação referente aos novos casos de infeção nos diferentes estados e aos óbitos. A informação é obtida com base no preenchimento de um formulário, por parte dos médicos, sempre que ocorre uma das duas situações já mencionadas. Apesar disso, de 1985 até 2004, a notificação não era obrigatória, pelo que a informação obtida nesses anos pode não ser a mais fiável. Porém, em 1 de Fevereiro de 2005, pela Portaria nº 103/2005, de 25 de Janeiro, a notificação passou a ser obrigatória. Este sistema é crucial no estudo de medidas de combate à epidemia.

Uma das informações mais relevantes de apoio ao combate à epidemia de VIH é taxa de

incidência da doença. Considera-se que a taxa de incidência da infeção por VIH corresponde ao número de indivíduos infetados por VIH, num determinado espaço temporal. Existe porém um problema associado ao cálculo desta taxa. Nomeadamente, a infeção por VIH não é imediatamente detetada, isto é, a diferença entre a data de infeção e a data de diagnóstico de VIH pode ser elevada. Isto acontece pois a doença apenas apresenta sintomas ligeiros nas primeiras semanas, sendo que nas semanas seguintes não ocorrem sintomas (fase assintomática). Esta última fase pode durar alguns anos. Por esse motivo, é essencial a aplicação de metodologias estatísticas para estimar a taxa de incidência.

O objetivo desta dissertação é aplicar metodologias que permitam estimar a taxa de incidência da infeção por VIH. Para isso, é essencial a informação do número de casos diagnosticados de SIDA, por unidade temporal. Esta informação apesar de estar disponível, apresenta dois problemas que necessitam de ser resolvidos. Nomeadamente, existem casos de indivíduos diagnosticados mas que ainda não foram notificados no sistema, este problema denomina-se de atrasos na notificação. Existem também casos de indivíduos que nunca são diagnosticados, e portanto, nunca são notificados. Este problema denomina-se de sub-notificação. Estes dois problemas têm necessariamente de ser considerados nas metodologias a estudar. Para o problema dos atrasos na notificação, outros autores abordaram o problema, usando máxima verosimilhança condicional, [Harris, 1990, Oliveira et al., 2014]. Para o problema da sub-notificação, e visto que temos acesso também à base de dados nacional para a Tuberculose, utilizámos uma metodologia que permite relacionar o número de casos de Tuberculose com o número de casos de SIDA. Esta metodologia foi desenvolvida em [DeGruttola et al., 1991], e aplicada num contexto diferente.

Finda a apresentação dos dois problemas associados à estimação da taxa de incidência, apresentamos agora a metodologia aplicada a essa estimação. Esta designa-se por método de retro-propagação, e consiste em, fixando um certo mês, o número de novos infetados nos meses anteriores é estimado usando o número de diagnosticados nesse mês, bem como a distribuição do tempo de incubação. Para a distribuição do tempo de incubação, existem algumas dificuldades na sua estimação. Em primeiro lugar, dados que permitam essa estimação só estão disponíveis para certos grupos de risco, o que pode implicar que uma possível generalização da distribuição, para toda a população, não faça sentido. Apesar disso, existem também limitações nos dados disponíveis, por exemplo, podem desaparecer pessoas do estudo (em virtude das características da doença), que leva à presença de dados censurados à direita, e que podem prejudicar os métodos aplicados. Apesar disso, assumimos duas distribuições, uma distribuição de Weibull e uma distribuição Gama, baseados em [Amaral et al., 2005]. Assume-se também que o número de diagnosticados e o número de infetados num certo mês seguem uma distribuição de Poisson. O método baseia-se numa equação de convolução, que é adaptada para o caso discreto, [Bacchetti et al., 1993]. As estimativas para o número médio de novas infeções por mês foram obtidas pelo método da máxima verosimilhança, a partir dos dados sobre o número de indivíduos diagnosticados em cada mês. Foram propostas algumas modificações ao método, nomeadamente uma possível alteração da média da distribuição do tempo de incubação ao longo do tempo, e uma alteração de parâmetros da distribuição ao longo do tempo. Para a maximização da verosimilhança, foram utilizados dois métodos: o algoritmo Estimação-Maximização (EM), [Green, 1990] e o método de Newton-Raphson. Para validação dos

resultados foram utilizados algoritmos de otimização implementados na R-package Optimx [Nash and Varadhan, 2011, Nash, 2014].

De notar que todas as metodologias utilizadas nesta dissertação foram construídas em software R versão 3.1.2 (R Development Core Team, 2014) [R Core Team, 2014], algumas construídas de raiz, outras utilizadas recorrendo a bibliotecas deste software.

Apresentamos agora a estrutura da dissertação. O capítulo 1 introduz o tema e seus problemas. O capítulo 2 apresenta uma introdução aos problemas detetados no decorrer dos trabalhos, nomeadamente, faz-se uma apresentação das bases de dados utilizadas e seus problemas. O capítulo 3 refere-se a uma metodologia que procura resolver o problema da sub-notificação. São apresentados os resultados obtidos e sua interpretação. O capítulo 4 apresenta uma metodologia estatística cujo objetivo é modelar a taxa de incidência da infeção por VIH. São apresentados diversos algoritmos numéricos essenciais à metodologia. No fim, são apresentados os resultados, bem como a sua análise. Por fim, no capítulo 5 são mencionadas as principais conclusões deste trabalho e trabalho futuro.

Esperamos que o trabalho realizado e aqui apresentado venha a ser útil no combate à propagação da infecção VIH em Portugal, nomeadamente que sirva como base para a implementação/reformulação de planos estratégicos nacionais de ação no combate ao VIH.

Capítulo 2

Dados nacionais sobre o VIH

Os dados que vão ser alvo de estudo foram obtidos a partir de dois sistemas de vigilância portugueses: o sistema de vigilância intrínseco conduzido pelo Programa Nacional de Luta Contra a Tuberculose (SVIG-TB) e o sistema nacional de notificação de casos de infeção por vírus da imunodeficiência humana, levado a cabo pelo Departamento de Doenças Infeciosas do Instituto Ricardo Jorge (INSA). Em ambos os sistemas, a notificação é feita por preenchimento de um formulário, que tem de ser posteriormente enviado para o sistema em causa. Os dois formulários correspondentes são apresentados em anexo.
As bases de dados foram obtidas diretamente das entidades responsáveis, após solicitação devidamente fundamentada. Inicialmente, foi feito um trabalho de correção e desenvolvimento das bases de dados. Esse trabalho englobou vários pontos:

1. Deteção de erros e posterior correção (incluindo o contacto com a fonte de notificação);

2. Uniformização das unidades de medida das variáveis entre as diferentes bases: na base do SVIG, por exemplo, apenas constava a informação sobre o concelho de residência do indivíduo, ao passo que na base do INSA apenas constava o distrito de residência do indivíduo. Foi necessário, portanto, obter uma base de dados nacional com a informação dos concelhos por distrito, para se inserir o distrito de residência na base do SVIG. No entanto, em virtude de existirem erros de escrita nos concelhos presentes na base do SVIG, todos eles tiveram que ser corrigidos. Para além disso, a base do SVIG apresentava a informação sobre o país de origem do indivíduo (por exemplo, Portugal), enquanto que a base do INSA apresentava informação sobre a nacionalidade (por exemplo, Portuguesa). Esta contrariedade teve que ser corrigida manualmente visto não conhecermos nenhuma instrução que associasse as duas informações;

3. Identificação de indivíduos comuns a ambas as bases, *matching*: com este procedimento, o que se pretendeu foi validar (e eventualmente completar) a informação constante em cada uma das bases;

4. Recolha de informação sobre a densidade populacional por distrito e adição dessa informação às bases existentes, informação que pode ser útil para o cálculo das taxas de incidência. Para este cálculo, foi efetuado, em primeiro lugar, o cálculo da área e do número de habitantes por distrito, e posteriormente foi efetuado o cálculo da densidade populacional por distrito.

Em relação ao ponto 3, surgiram algumas questões que nos deixaram apreensivos para com os problemas que podem surgir de se estar a trabalhar com duas bases de dados obtidas a partir de sistemas de vigilância nacionais diferentes. Convém primeiro referir que, na base do SVIG-TB, é possível saber se um indivíduo com tuberculose ativa (TB) está ou não infetado com o vírus da imunodeficiência humana (VIH); no caso de estar infetado, esse indivíduo estará necessariamente na fase SIDA por apresentar uma doença definidora de SIDA. Na base do INSA, é também possível saber qual a doença que determinou a passagem para a fase SIDA de um certo indivíduo. Posto isto, vamos explicar como foi realizado o processo de *matching*. Inicialmente filtrámos a base do SVIG-TB, apenas considerando os indivíduos infetados com VIH. Daqui por diante, vamos designar esta base por base filtrada. Naturalmente, se tudo funcionasse na perfeição, seria de esperar que todos os indivíduos na base do SVIG-TB filtrada constassem na base do INSA. No entanto, isso não aconteceu. Para procedermos ao processo de identificação dos indivíduos comuns às duas bases, usámos como variáveis comuns: país de origem, sexo, data de nascimento e distrito de residência. O resultado desta procura de concordância entre os casos de ambas as bases encontra-se na Tabela 2.1.

Nº indivíduos semelhantes	Nº de casos
0	2174
1	1637
2	234
3	45
4	5

Tabela 2.1: Resultados do processo de *matching*.

Podemos concluir que, na base do SVIG-TB filtrada: existiam 2174 indivíduos sem nenhuma correspondência na base do INSA relativamente às quatro variáveis escolhidas, existiam 1637 indivíduos com uma correspondência na base do INSA que partilhava a mesma resposta nas quatro variáveis, etc. Isto significa que aproximadamente 53% dos indivíduos com TB, VIH na fase SIDA e notificados à base do SVIG-TB não constavam na base do INSA! Esta questão pertinente foi debatida junto da comunidade médica quase diretamente responsável pelas bases de dados em causa, mas não foi possível obter uma justificação coerente das discrepâncias observadas.

Surgiu também uma questão curiosa relacionada com o assunto anterior. Para os indivíduos da base do SVIG-TB filtrada para os quais se encontraram uma correspondência na base do INSA, estudámos a diferença entre a data de diagnóstico de TB e a data de diagnóstico de SIDA. O resultado pode ser visualizado na Figura 2.1. Teoricamente, existem duas possibilidades, ou a diferença é positiva, o que significa que o indivíduo passou para a fase SIDA por uma outra doença definidora de SIDA e posteriormente foi diagnosticado com Tuberculose, ou é negativa, o que implica que o indivíduo foi diagnosticado com Tuberculose e com VIH, e só depois foi notificado na base do INSA como estando na fase SIDA. Neste último caso, essa diferença negativa deveria ser muito próxima de zero, o que não acontece. De facto, observando a Tabela 2.2, podemos verificar que aproximadamente 19% dos casos são diagnosticados na base do INSA com SIDA um ano depois de terem sido diagnosticados com TB e VIH. Esta nossa constatação aponta para a existência de erros nos sistemas

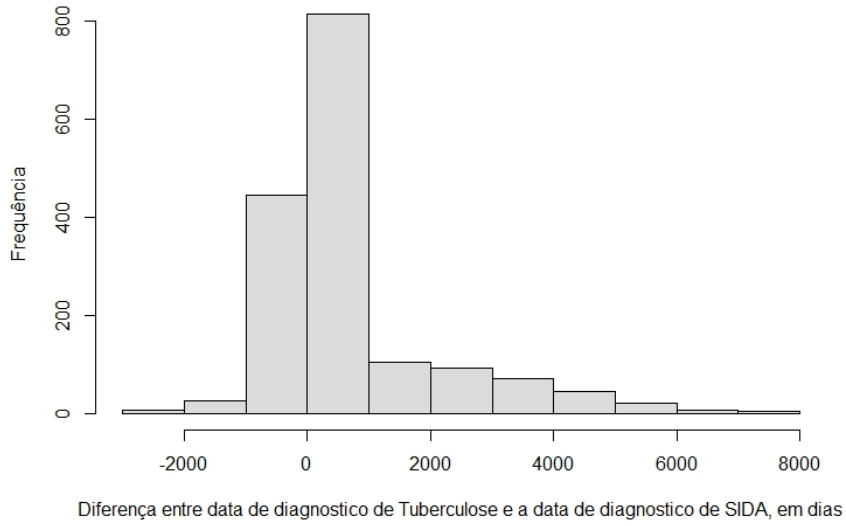


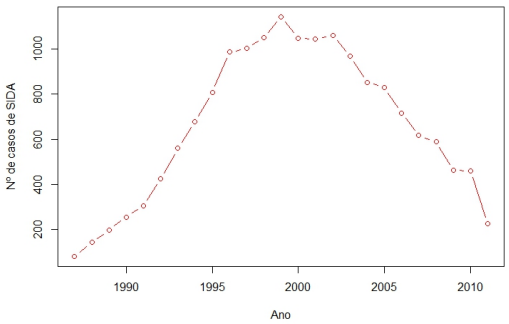
Figura 2.1: Diferença entre a data de diagnóstico de TB e a data de diagnóstico de SIDA.

Nº dias entre diagnósticos	% aproximada de casos
>30	48%
>90	34%
>150	26%
>365	19%

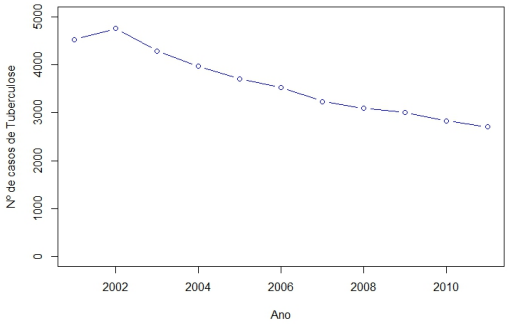
Tabela 2.2: Descrição do número de dias entre o diagnóstico de TB e o diagnóstico de SIDA, para os indivíduos com diferença entre os diagnósticos negativa.

nacionais de vigilância epidemiológica e sugere naturalmente a criação de uma colaboração mais eficaz entre os vários sistemas.

Após os trabalhos de limpeza e aprimoramento das bases de dados, foi feita uma análise às curvas empíricas de incidência de casos de SIDA entre 1987 e 2011, de casos de TB entre 2001 e 2011 e da co-infeção VIH/TB entre 2001 e 2011. Aqui designamos por co-infeção VIH/TB a infeção simultânea por VIH e TB. Relativamente à evolução temporal do número de novos casos de SIDA efetivamente notificados na base do INSA, podemos referir que, considerando a Figura 2.2(a), a epidemia atingiu o pico máximo em 1999, e a partir dessa data, ocorreu uma queda acentuada até ao ano de 2011. Para a TB, e observando a Figura 2.2(b), podemos verificar que o dois primeiros anos presentes na base do SVIG-TB correspondem aos anos com o maior número de casos diagnosticados de Tuberculose. A tendência nos restantes anos é de descida pouco acentuada.
Considerando a informação anterior, mas por trimestre, pouco se pode inferir, no entanto, convém referir que parece existir alguma sazonalidade visto que, na Figura 2.3, podemos constatar uma grande variabilidade entre trimestres. Esta informação foi apresentada junto da comunidade médica, que não conseguiu apresentar, com certeza, uma justificação para



(a) Nº de casos de SIDA, entre 1987 e 2011.



(b) Nº de casos de Tuberculose, entre 2001 e 2011.

Figura 2.2: Comparação entre o nº de casos de SIDA e o nº de casos de Tuberculose.

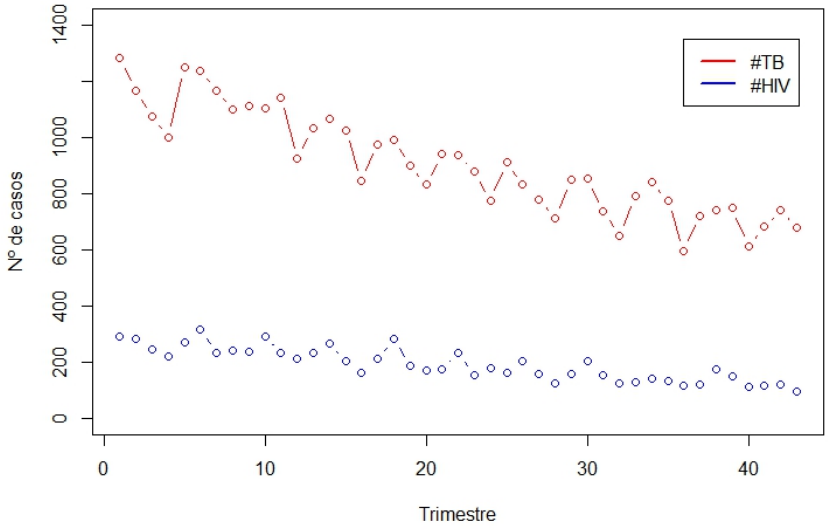


Figura 2.3: Número de casos de Tuberculose e SIDA por trimestre, entre 2001 e 2011.

este efeito. As únicas justificações dadas (com um elevado grau de incerteza) foram que podem existir médicos que acumulem formulários e apenas os preencham em determinadas épocas do ano ou que os serviços administrativos tenham a tarefa das notificações bem calendarizada no tempo.

Finda a análise das curvas de incidência, exploraram-se as relações entre as duas curvas, usando teoria de correlação. Foram consideradas as seguintes comparações:

1: Comparação entre a curva de evolução isolada da TB e a curva de evolução isolada da SIDA, por trimestre;

2: Comparação entre as diferenças entre anos sucessivos para a curva da TB e a diferença entre anos sucessivos para a curva da SIDA, por trimestre.

Matematicamente, estas ideias foram exploradas da seguinte forma:

1: Sejam $(TB_1, \ldots, TB_n)$ e $(SIDA_1, \ldots, SIDA_n)$ respetivamente o vetor do número de casos de TB e o vetor do número de casos de SIDA, no total dos trimestres $1, \ldots, n$ observados. Consideraram-se os dados agrupados por trimestre: $(TB_1, SIDA_1), \ldots, (TB_n, SIDA_n)$ e os coeficientes de correlação amostral de Pearson e de Spearman foram calculados;

2: Sejam $(TB_1, \ldots, TB_n)$ e $(SIDA_1, \ldots, SIDA_n)$ como definidos na Comparação **1**. Consideraram-se os dados na forma: $(TB_1 - TB_2, SIDA_1 - SIDA_2), \ldots, (TB_{n-1} - TB_n, SIDA_{n-1} - SIDA_n)$ e os coeficientes de correlação amostral de Pearson e de Spearman foram calculados.

Observando os resultados da Tabela 2.3, o mais relevante a referir é que, quando comparámos as duas curvas, verificámos que a Comparação **1** apresenta uma correlação forte quer na correlação de Pearson, quer na correlação de Spearman. Estes resultados serviram de inspiração e motivação para o capítulo seguinte.

	Pearson	Spearman
1	0.92	0.93
2	0.49	0.53

Tabela 2.3: Correlação entre as comparações **1** e **2**.

Capítulo 3

Modelação da Sub-notificação

3.1 Introdução

O objetivo deste capítulo consiste do estudo do fenómeno de sub-notificação. A motivação das abordagens usadas proveio do artigo [DeGruttola et al., 1991], que desenvolve e implementa uma metodologia cujo objetivo é modelar a variação de células CD4 ao longo do tempo. Existe uma grande falta de informação sobre a percentagem de casos de SIDA que não são notificados ao sistema. Por exemplo, existem casos de toxicodependentes com SIDA que nunca se deslocaram ao médico, casos de pacientes que simplesmente não são notificados ao sistema por falhas nos serviços administrativos, por falha no preenchimento dos formulários por parte dos médicos, entre outras situações. Por esse motivo, é muito complicado estimar o número de casos não notificados de SIDA, e as estimativas que existem (é assumido em [ECDC/WHO Regional Office for Europe, 2012] que a sub-notificação de casos SIDA pode variar entre 0% a 25%) parecem não ter grande base teórica que as defenda. Neste trabalho, desenvolvemos uma abordagem inovadora: identificámos a relação que existe entre o vírus do VIH e a bactéria da TB, e obrigámos a que o número de casos de SIDA fosse o mais fiel possível a essa relação. Neste processo, assumimos que o número de casos de TB está corretamente identificado. Quando um indivíduo é infetado pelo vírus do VIH, este passa por diversas fases. A fase em estudo nesta tese é a fase SIDA. Esta fase é a mais agressiva pois possui sintomas que são mais severos para a saúde do indivíduo. Uma das condições para que um indivíduo com VIH passe para a fase SIDA é que lhe seja detetada um doença definidora de SIDA. Ora, a TB é uma doença definidora de SIDA. Isto acontece porque a bactéria da TB está latente em aproximadamente 1/3 da população, [World Health Organization, 2015]. Ora, quando um indivíduo é infetado por VIH, com o passar do tempo, a sua resposta imunitária vai enfraquecendo. Por isso, se o indivíduo tem a bactéria da TB latente, esta, com o enfraquecimento do sistema imunitário, poderá passar para o estado activo, e por conseguinte, esse indivíduo irá passar para a fase SIDA.
Perante isto, e servindo como motivação, fomos comparar o número de casos de SIDA com o número de casos de TB, por trimestre. Na Figura 3.1, podemos observar que existe uma tendência linear. Perante isto, considerámos que a metodologia descrita no artigo [DeGruttola et al., 1991] poderia ser usada no contexto deste problema, para corrigir o número de casos de SIDA, assumindo que o número de casos de TB está correto.

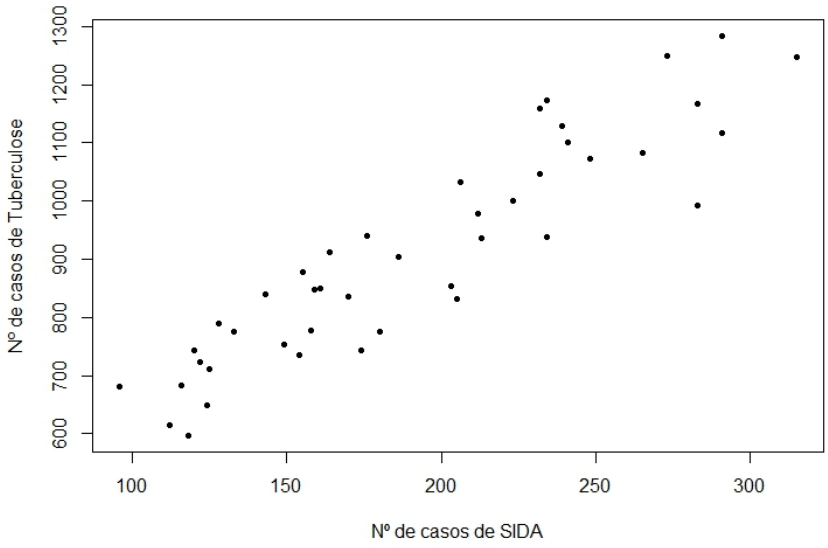


Figura 3.1: Nº de casos de Tuberculose vs Nº de casos de SIDA, por trimestre.

3.2 Regressão para preditores com erro

Nesta secção descrevemos o modelo usado para resolver o problema da sub-notificação. Tal como já foi referido atrás, vamos trabalhar com contagens por trimestre. Considere-se y_i como sendo o número de casos de TB no trimestre i e x_i o número de casos de SIDA no trimestre i, onde $i = 1, ..., n$ e n representa o número de trimestres.
Considere-se a seguinte notação:

$$\mathbf{y} \equiv (y_1, ..., y_n)^T \qquad \mathbf{x} \equiv (x_1, ..., x_n)^T \qquad \mathbf{X} \equiv [\mathbf{1}, \mathbf{x}]$$

onde

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

O modelo é o seguinte:

$$\mathbf{y} = \mathbf{X}^* \boldsymbol{\alpha} + \boldsymbol{\epsilon}^* \tag{3.1}$$

com as seguintes condições

$$\boldsymbol{\epsilon}^* \equiv (\epsilon_1^*, ..., \epsilon_n^*)^T, \qquad \boldsymbol{\alpha} \equiv (\alpha_0, \alpha_1)^T, \qquad x_i^* = x_i + \tau_i, \qquad \mathbf{x}^* \equiv (x_1^*, ..., x_n^*)^T, \qquad \mathbf{X}^* \equiv [\mathbf{1}, \mathbf{x}^*],$$
$$\epsilon^* \sim N(0, {\sigma^*}^2 I), \qquad \tau_i \sim H_i, \qquad \tau_i \text{ independente de } \epsilon_i^*, \quad i = 1, \ldots, n$$

Sendo desconhecido, o vetor de parâmetros (populacionais) $\boldsymbol{\alpha}$ terá de ser estimado. A variável aleatória τ_i corresponde ao número de casos de SIDA que não são notificados no

trimestre i, ou seja, funcionará para estimar o número de casos de sub-notificação no trimestre i. As distribuições H_i são assumidas como sendo conhecidas. Neste trabalho, usámos o relatório [ECDC/WHO Regional Office for Europe, 2012], que refere que a estimativa para a sub-notificação dos países europeus não deve ultrapassar os 25% do valor observado para a unidade temporal em causa. Posto isto, optámos por assumir as seguintes distribuições uniformes: considerando um certo x_i, correspondendo ao número observado de casos SIDA no trimestre i, e $C_i = \{0, 1, \ldots, \lceil 0.25x_i \rceil\}$:

$$\tau_i \sim H_i \Rightarrow P(\tau_i = t_s) = \tfrac{1}{|C_i|} \text{ se } t_s \in C_i.$$

Aqui usámos $|C_i|$ para representar a cardinalidade do conjunto C_i.
Visto que este modelo não corresponde a um modelo de regressão linear (pois $\mathbf{X}^*$ é não observado), é necessário uma adaptação para se obter as estimativas do parâmetros desconhecidos usando o método dos mínimos quadrados pesados. Vamos então explicar como foi resolvido o problema. Considerámos uma nova base construída da seguinte forma:

y_i	τ_i	pesos
y_1	$\tau_{11} = 0$	w_{11}
y_1	$\tau_{12} = 1$	w_{12}
y_1	$\tau_{13} = 2$	w_{13}
y_1	$\tau_{14} = 3$	w_{14}
y_1	$\tau_{15} = 4$	w_{15}
y_2	$\tau_{21} = 0$	w_{21}
y_2	$\tau_{22} = 1$	w_{22}
y_2	$\tau_{23} = 2$	w_{23}
y_2	$\tau_{24} = 3$	w_{24}
y_3	$\tau_{31} = 0$	w_{31}
y_3	$\tau_{32} = 1$	w_{32}
y_3	$\tau_{33} = 2$	w_{33}
y_3	$\tau_{34} = 3$	w_{34}
y_3	$\tau_{35} = 4$	w_{35}
y_3	$\tau_{36} = 5$	w_{36}
:	:	:

Tabela 3.1: Replicação dos dados para cada ponto de suporte H_i, considerando que $\lceil 0.25x_1 \rceil = 4$, $\lceil 0.25x_2 \rceil = 3$ e $\lceil 0.25x_3 \rceil = 5$, por exemplo.

A ideia consiste de um algoritmo iterativo, onde a cada iteração, os parâmetros são estimados usando o método dos mínimos quadrados pesados, e os pesos são recalculados.
O método é resumido do seguinte modo:

Passo 1: Construir uma nova base usando a ideia da tabela 3.1.

Passo 2: Considerar os pesos iniciais como $w_{si}^0 \equiv H_i(\tau_{is})$, para $i = 1, ..., n$ e $s = 1, ..., |C_i|$.

Passo 3: Obter as estimativas para os parâmetros, usando o método dos mínimos quadrados pesados.

Passo 4: Atualizar os pesos usando os parâmetros estimados.

Passo 5: Iterar entre os Passos 3 e 4 até convergir.

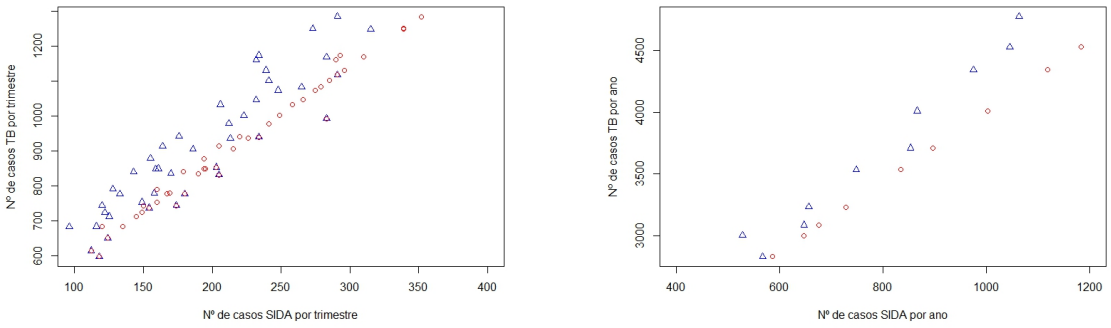
No passo 4, a fórmula para a atualização dos pesos foi encontrada em [DeGruttola et al., 1991], e é dada por:

$$w_{si}^{(p+1)} = \frac{g(y_i|\tau_{is}, \alpha^{(p)}, \sigma^{*2(p)})H_i(\tau_{is})}{\sum_s g(y_i|\tau_{is}, \alpha^{(p)}, \sigma^{*2(p)})H_i(\tau_{is})},$$

onde $g(.)$ corresponde à função densidade de probabilidade condicional de y_i dado τ_{is}, $\alpha^{(p)}$ e $\sigma^{*2(p)}$.

3.3 Resultados

A metodologia descrita na secção anterior foi aplicada aos dados considerando um agrupamento das contagens de casos de SIDA por trimestre e por ano. O facto de se terem considerado os dois agrupamentos funcionou um pouco como uma análise de sensibilidade, dado que à partida não havia motivos suficientes para escolher um agrupamento em detrimento do outro. Os resultados são apresentados nas Figuras 3.2(a) e 3.2(b). Os triângulos azuis correspondem aos valores observados, enquanto que os círculos vermelhos correspondem à correção feita pelo método. De notar que as correções apenas podem ser feitas na horizontal, nomeadamente para a direita dos triângulos azuis, ou então não podem ser feitas de todo. Isto porque as correções são necessariamente positivas. Verificámos que os círculos vermelhos que estão à direita da tendência linear dos restantes círculos coincidem com os triângulos azuis, ou seja, nesse trimestre/ano não são feitas nenhumas correções. É também importante referir que existem círculos vermelhos que não coincidem com a tendência linear dos restantes, estando mais à esquerda. Isto acontece porque as movimentações para a direita foram limitadas até 25% do valor observado. As estimativas obtidas para a sub-notificação, por trimestre e por ano, são apresentadas na Tabela 3.2.



(a) Por trimestre.

(b) Por ano.

Figura 3.2: Resultado da metodologia aplicada à sub-notificação.

Os resultados parecem ser congruentes com a situação real, na medida em que, como já foi referido, em 2005 a infeção por VIH passou a integrar a lista das doenças de declaração

Ano	Trimestre	Estimativa por trimestre	Estimativa por ano
2001	1	20.8	13.3
	2	9.4	
	3	11.0	
	4	11.8	
2002	1	24.3	19.6
	2	7.5	
	3	25.2	
	4	18.5	
2003	1	23.8	14.8
	2	0.3	
	3	25.0	
	4	6.1	
2004	1	14.5	15.7
	2	5.3	
	3	25.1	
	4	20.9	
2005	1	13.7	5.0
	2	0.0	
	3	15.5	
	4	11.5	
2006	1	25.0	11.5
	2	0.0	
	3	25.1	
	4	0.0	
2007	1	25.0	11.0
	2	0.0	
	3	6.9	
	4	16.9	
2008	1	22.1	4.5
	2	0.0	
	3	0.5	
	4	0.4	
2009	1	25.0	22.4
	2	25.2	
	3	25.2	
	4	0.0	
2010	1	22.2	3.4
	2	0.0	
	3	7.3	
	4	0.3	

Tabela 3.2: Estimativas dadas pela metodologia para a sub-notificação, por trimestre e por ano.

obrigatória, e na verdade, a estimativa para a sub-notificação parece ser inferior a partir desse ano. Subsiste contudo a dúvida quanto ao valor estimado para 2009, que é extremamente elevado e que contradiz a tendência decrescente definida pelas estimativas anteriores e posteriores. Para as estimativas por trimestre, observámos que, em quase todos os anos, a estimativa para o segundo trimestre é inferior à dos restantes três trimestres. Esta constatação não está de acordo com as flutuações observadas nos dados por trimestre. Procurámos uma explicação para este facto junto da comunidade médica especializada no tema, no entanto, como existem muitos problemas associados ao registo dos casos de SIDA, não nos foram dadas respostas esclarecedoras.

As hipóteses assumidas pela metodologia aplicada acima têm algumas limitações. Em primeiro lugar, a condição de que os erros são independentes entre os anos pode não ser satisfeita. Eventualmente poder-se-ia alterar o método incorporando efeitos aleatórios, e dessa forma, relaxar as condições impostas sobre os erros. Outro ponto sensível consiste na assunção de que a base da TB é totalmente correta. Ora, inicialmente tudo indicava que sim, até pelas indicações que nos foram dadas pelo grupo responsável pela base. No entanto, fomos detetando erros, como por exemplo, a duplicação de dados, dados introduzidos com erro, entre outros.

Perante tudo isto, não existe total confiança nos resultados aqui obtidos. Acresce que, o facto de não existirem estimativas anteriores para o fenómeno da sub-notificação de SIDA em Portugal, não permite validar nem criticar objetivamente os valores acima estimados.

Capítulo 4

O método de retro-propagação

4.1 Introdução

O objetivo principal desta tese consiste da estimação da taxa de incidência da infeção por VIH em Portugal. Para tal, usámos o método de retro-propagação. Este método foi introduzido pela primeira vez em [Brookmeyer and Gail, 1988] para reconstruir o passado da epidemia do VIH nos EUA, bem como para prever o número futuro de casos de SIDA. A metodologia foi proposta em 1988 e desde então tem sido amplamente usada, [Amaral et al., 2005, Deuffic and Costagliola, 1999, Mariotti and Cascioli, 1996].
O método de retro-propagação descrito em [Brookmeyer and Gail, 1988] usa uma equação de convolução contínua. Nesta dissertação, considerámos uma abordagem discreta a essa equação de convolução [Bacchetti et al., 1993]. O método calcula o número de infetados por VIH no passado, a partir do número de casos diagnosticados de SIDA e da distribuição do tempo de incubação. Assume-se que a distribuição do tempo de incubação é conhecida, que o número de casos de SIDA está disponível (por unidade temporal) e que existe um modelo para a distribuição do número de infetados por VIH. Para o primeiro ponto, vamos considerar ao longo desta dissertação as duas seguintes distribuições para o tempo de incubação (tal como foi considerado em [Amaral et al., 2005]):

- Distribuição $Weibull(2.5, 12.818)$:

$$\begin{aligned} X &\sim \text{Weibull}(\lambda, k), \\ f(x) &= \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} e^{(-x/\lambda)^k}, \qquad x > 0 \\ E(X) = \lambda\Gamma(1+\frac{1}{k}), &\qquad V(X) = \lambda^2[\Gamma(1+\frac{2}{k}) - (\Gamma(1+\frac{1}{k}))^2]. \end{aligned}$$

- Distribuição $\Gamma(5.7, 2.05)$:

$$\begin{aligned} X &\sim \Gamma(k, \theta), \\ f(x) &= \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}, \qquad x > 0 \\ E(X) = k\theta, &\qquad V(X) = k\theta^2. \end{aligned}$$

Relativamente ao número de casos diagnosticados de SIDA num dado instante/período de tempo, este número foi obtido da base do INSA, mas existem dois problemas subjacentes: a sub-notificação (já considerada no Capítulo 3) e os atrasos na notificação. Apesar de ambas as situações poderem ser salvaguardadas na metodologia, apenas os atrasos na notificação serão considerados visto que não existe total confiança nos resultados obtidos no Capítulo 3. Finalmente, assume-se que o número de infetados por VIH pode ser bem modelado usando uma distribuição de Poisson. Esta hipótese é suportada por várias publicações na área, [Bacchetti et al., 1993, Amaral et al., 2005].

4.2 O método da máxima verosimilhança

Nesta secção, vamos introduzir o método da máxima verosimilhança, fundamental para o estudo do método de retro-propagação. Em primeiro lugar, começamos com algumas definições [Pestana and Velosa, 2008, Arnold, 1990]:

Definição 4.2.1. Uma estatística é uma função de uma ou mais amostras aleatórias que não envolve parâmetros desconhecidos.

Definição 4.2.2. Dada uma amostra aleatória $X_1, X_2, \cdots, X_n$, um estimador $\hat{\theta}$ de θ diz-se suficiente se a distribuição condicional (conjunta) f de $X_1, X_2, \cdots, X_n$ dado $\hat{\theta}$ não depende de θ. Formalmente,

$$f(x_1, x_2, \cdots, x_n|\hat{\theta}) = \frac{f(x_1, x_2, \ldots, x_n)}{g(\hat{\theta})}$$

onde g é uma função densidade de probabilidade.

Definição 4.2.3. Suponhamos que uma variável aleatória X tem função densidade de probabilidade $f(x|\theta)$. Dado o valor observado x de X, a função de verosimilhança é dada por

$$l(\theta|x) = f(x \mid \theta)$$

Ou seja, estamos a considerar a função densidade de probabilidade como função de θ e não de x, dado que se observou x. No caso multivariado, $x = (x_1, x_2, \cdots, x_n)$ é o vetor dos valores observados das variáveis aleatórias $(X_1, X_2, \cdots, X_n)$. No caso de uma realização $x_1, \ldots, x_n$ de uma amostra aleatória. $X_1, \ldots, X_n$, a função de verosimilhança é dada por:

$$l(\theta|X) = f(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\theta) = \prod_{i=1}^{n} f(X_i = x_i|\theta) \tag{4.1}$$

No caso discreto, a função de verosimilhança corresponde portanto à probabilidade (conjunta) de observação de $(x_1, \ldots, x_n)$. Faz então todo o sentido estimar θ como sendo o valor do parâmetro que maximiza a probabilidade de ocorrência dos dados no modelo estabelecido. Um tal estimador $\hat{\theta}(X)$ é denominado por estimador de máxima verosimilhança. Na prática, o estimador de máxima verosimilhança obtém-se maximizando $log(l(\theta))$, que é conhecida como a função de log-verosimilhança, para simplificação dos cálculos envolvidos.

Exemplo 4.2.4. Suponhamos que $(X_1, \ldots, X_n)$ é uma amostra aleatória, tal que que $X_i \sim P(\lambda)$, para $i = 1, \cdots, n$, e que se pretende estimar o parâmetro λ a partir de uma realização $x_1, \ldots, x_n$ da amostra aleatória. Da equação (4.1) temos que,

$$l(\theta|x) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

O estimador de máxima verosimilhança é calculado da seguinte forma:

$$\frac{d \log l(\theta|x)}{d\lambda}|_{\lambda=\hat{\lambda}} = 0 \Leftrightarrow \frac{\sum_{i=1}^{n} x_i}{\hat{\lambda}} - n = 0 \Leftrightarrow \hat{\lambda} = \bar{X}$$

Exemplo 4.2.5. Suponhamos que $(X_1, \ldots, X_n)$ é uma amostra aleatória, tal que que $X_i \sim P(\lambda_i)$, para $i = 1, \cdots, n$, e que se pretende estimar o vetor de parâmetros λ a partir de uma realização $x_1, \ldots, x_n$ da amostra aleatória. Da equação (4.1) temos que,

$$l(\theta|x) = \prod_{i=1}^{n} \frac{\lambda_i^{x_i} e^{-\lambda_i}}{x_i!}$$

O estimador de máxima verosimilhança é calculado, resolvendo o seguinte sistema:

$$\begin{cases} \frac{\partial \log l(\theta|x)}{\partial \lambda_1}|_{\lambda_1=\hat{\lambda}_1} = 0 \\ \vdots \\ \frac{\partial \log l(\theta|x)}{\partial \lambda_n}|_{\lambda_n=\hat{\lambda}_n} = 0 \end{cases} \Leftrightarrow \begin{cases} \frac{x_1}{\hat{\lambda}_1} - 1 = 0 \\ \vdots \\ \frac{x_n}{\hat{\lambda}_n} - 1 = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{\lambda}_1 = x_1 \\ \vdots \\ \hat{\lambda}_n = x_n \end{cases}$$

4.3 Descrição do método de retro-propagação

Vamos começar por descrever a equação de convolução que constitui a base do método de retro-propagação. Sejam t_{j-1} e t_j dois tempos quaisquer com $t_{j-1} \leq t_j$. No que se segue iremos apenas considerar instantes de tempo não negativos, assumindo que o zero corresponde ao início da epidemia. Definimos:

- Y_j como o número de novos casos de SIDA que ocorreu em (t_{j-1}, t_j).
- a curva de infeção $I(.)$, dependente do tempo, de tal forma que $\int_0^s I(u)du$ corresponde ao número total de infeções que ocorreram até ao tempo s.
- $F(t)$ como a probabilidade de o diagnóstico de SIDA ocorrer até t unidades de tempo após a infeção; isto é, a probabilidade de o tempo de incubação ser inferior a t unidades.
- $F(t \mid s)$ como a distribuição do tempo de incubação para os indivíduos infetados no tempo s.

Com esta notação, o número esperado de casos de SIDA para o intervalo (t_{j-1}, t_j) é

$$E(Y_j) = \int_0^{t_j} I(s)[F(t_j - s \mid s) - F(t_{j-1} - s \mid s)]ds, \tag{4.2}$$

A ideia principal do método de retro-propagação é, a partir do conhecimento do vetor $(Y_1, Y_2, ..., Y_n)$, onde a média de cada componente satisfaz (4.2), e da distribuição do tempo de incubação F, estimar a curva do número de infetados I.
Neste trabalho vamos considerar uma abordagem discreta ao problema. Faz todo o sentido essa abordagem visto que a informação para o número de diagnósticos de SIDA está disponível por mês. Sejam então i, j e k três meses quaisquer do estudo, com $i \leq j \leq k$. Definimos agora

- Y_j como o número de novos casos de SIDA que ocorreu no mês j.
- $P(\text{diagnosticado mês } j \mid \text{infetado mês } i) \equiv P(j \mid i) \equiv D_{ij}$
- $P(\text{notificado mês } k \mid \text{diagnosticado mês } j) \equiv P(k \mid j) \equiv R_{jk}$

$$0 \leq i \leq j,\ i \leq j \leq n \text{ e } j \leq k \leq n^*$$

- n corresponde ao último mês de diagnóstico e n^* corresponde ao último mês de notificação, presentes na base de dados,

e assumimos que o número de novas infeções no mês i segue uma distribuição de Poisson com média θ_i. O objetivo do trabalho é, portanto, estimar $\theta = (\theta_1, \theta_2, ..., \theta_n)$.
A equação (4.2) pode ser re-escrita na forma:

$$E(Y_j) = \sum_{i=0}^{j} \theta_i D_{ij}. \tag{4.3}$$

No entanto, como apenas temos acesso ao número de diagnosticados efetivamente notificados (devido a atrasos na notificação e a sub-notificação dos dados), vamos introduzir a seguinte notação:

$$\begin{aligned} R_j &\equiv \sum_{k=j}^{n^*} R_{jk} \\ &= P(\text{notificado até ao mês } n^* \mid \text{diagnosticado no mês } j) \end{aligned}$$

Ou seja,

$$Y_j^* = Y_j R_j$$

onde Y_j^* corresponde ao número observado de diagnósticos no mês j. A equação (4.3) pode ser reescrita da seguinte forma:

$$E(Y_j^*) = \sum_{i=0}^{j} \theta_i D_{ij} R_j = \sum_{i=0}^{j} \sum_{k=j}^{n^*} \theta_i D_{ij} R_{jk} \tag{4.4}$$

Posto isto, vamos considerar que

$$Y_j^* = \sum_{k=j}^{n^*} Y_{jk}^*, \text{ com } Y_{jk}^* \sim P(\sum_{i=0}^{j} \theta_i D_{ij} R_{jk}), \tag{4.5}$$

onde Y^*_{jk} corresponde ao número observado de indivíduos diagnosticados no mês j e notificados no mês k. É importante esclarecer a consideração (4.5), nomeadamente, na média considerada para a variável aleatória Y^*_{jk}.
Seja $a_{ijk} = P(\text{notificado no mês } k, \text{ diagnosticado no mês } j \mid \text{infetado no mês } i)$ e assumindo que:

$$\begin{aligned} a_{ijk} \equiv P(j,k|i) &= P(j|i)P(k|j) \Leftrightarrow \\ &= D_{ij}R_{jk} \end{aligned}$$

temos que:

$$E(Y^*_{jk}) = \sum_{i=0}^{j} \theta_i a_{ijk}$$

A suposição anterior corresponde matematicamente a:

$$\begin{aligned} P(j,k|i) &= P(j|i)P(k|j) \Leftrightarrow \\ \frac{P(j,k,i)}{P(i)} &= P(j|i)P(k|j) \Leftrightarrow \\ \frac{P(j,k,i)}{P(j,i)}\frac{P(j,i)}{P(i)} &= P(j|i)P(k|j) \Leftrightarrow \\ P(k|j,i)P(j|i) &= P(j|i)P(k|j) \Leftrightarrow \\ P(k|j,i) &= P(k|j), \end{aligned}$$

ou seja, assumimos que o tempo de incubação não influencia o atraso nas notificações. Posto isto, é fácil a interpretação para a média de Y^*_{jk}. Corresponde de facto à soma sobre todos os meses de infecção i, do produto do número médio de infetados no mês i pela probabilidade de ser diagnosticado no mês j e notificado no mês k dado que foi infetado no mês i. Também se verifica que a equação (4.4) continua válida para a consideração (4.5).
Usando Y^*_{jk}, podemos fatorizar a função de verosimilhança da seguinte forma:

$$\begin{aligned} P(y^*_{jj}, ..., y^*_{jn^*}, y^*_j) &= P(y^*_{jj}, y^*_{jj+1}, ..., y^*_{jn^*}|y^*_j)P(y^*_j) \Leftrightarrow \\ P(y^*_{jj}, ..., y^*_{jn^*}) &= P(y^*_{jj}, y^*_{jj+1}, ..., y^*_{jn^*}|y^*_j)P(y^*_j) \Leftrightarrow \\ L(\theta|y^*_{jj}, ..., y^*_{jn^*}) &= L_c \cdot L_m \end{aligned}$$

onde L_c corresponde à função de verosimilhança condicional de $\{y^*_{jk}\}$ dado $\{y^*_j\}$ e L_m corresponde à função de verosimilhança marginal de $\{y^*_j\}$. No primeiro passo usa-se o facto de que $Y^*_j = \sum_{k=j}^{n^*} Y^*_{jk}$. Ora, sabe-se que [Kaas et al., 2008], fixando j, a distribuição condicional de $(y^*_{jj}, ..., y^*_{jn^*})$ dado y^*_j segue uma distribuição multinomial com y^*_j tentativas e probabilidades $P_{jj}, ..., P_{jn^*}$, onde:

$$P_{jk} = \frac{\sum_{i=0}^{j} \theta_i D_{ij} R_{jk}}{\sum_{i=0}^{j} \sum_{k=j}^{n^*} \theta_i D_{ij} R_{jk}}$$

Podemos agora escrever a expressão para L_c:

$$L_c = \prod_{j=0}^{n}\prod_{k=j}^{n^*}\left(\frac{\sum_{i=0}^{j}\theta_i D_{ij}R_{jk}}{\sum_{i=0}^{j}\sum_{k=j}^{n^*}\theta_i D_{ij}R_{jk}}\right)^{y_j^*} \times \text{constante}$$
$$= \prod_{j=0}^{n}\prod_{k=j}^{n^*}\left(\frac{R_{jk}}{\sum_{k=j}^{n^*}R_{jk}}\right)^{y_j^*} \times \text{constante}$$

Ou seja, L_c não depende de θ. Por esse motivo, apenas temos que maximizar $log(L_m)$, cuja expressão é dada por:

$$log(L_m) = \sum_{j=0}^{n}(y_j^* log(\sum_{i=0}^{j}\theta_i D_{ij}R_j) - \sum_{i=0}^{j}(\theta_i D_{ij}R_j)) + \text{constante} \tag{4.6}$$

A expressão (4.6) pode ser estimada usando um de dois métodos, [Bacchetti et al., 1993]:

- Método de Newton-Raphson
- Algoritmo Estimação-Maximização (EM),

cujos detalhes matemáticos serão providenciados na próxima secção.

Relativamente à metodologia geral do método da retro-propagação que descrevemos acima, temos, desde já, uma crítica a fazer. É assumido que a média da distribuição do tempo de incubação é constante ao longo do tempo. De facto, não existem estimativas para a variação da média da distribuição do tempo de incubação ao longo do tempo. Contudo, esta hipótese não nos parece razoável, dado que, com o desenvolvimento científico e farmacológico, o tempo de incubação é hoje superior ao verificado no passado. Surgiram então duas ideias alternativas:

Ideia 1: Considerar uma modificação num dos parâmetros da distribuição do tempo de incubação permitindo que este varie ao longo do tempo. Aqui começamos por considerar a relação mais simples, do tipo linear: $a\ i + b$ onde i é o trimestre;

Ideia 2: Alterar a média da distribuição do tempo de incubação, multiplicando-a por uma constante positiva que pode ser diferente conforme o tempo.

Enquanto que a Ideia 1 apenas acrescenta dois parâmetros ao número de parâmetros a estimar, a Ideia 2 implica que o número de parâmetros a estimar duplique.
Vamos em primeiro lugar considerar a Ideia 1.
Tal como já foi referido, temos duas distribuições possíveis para o tempo de incubação, uma distribuição de $Weibull(2.5, 12.818)$ e uma distribuição $\Gamma(5.7, 2.05)$, [Amaral et al., 2005]. A ideia é considerar, supondo que estamos a trabalhar com a distribuição de Weibull,

$$D_{ij} = P(j-i|i) = f(j-i; 2.5, a + b\ i) \equiv D_{ij}(a, b).$$

Supondo que estamos a trabalhar com a distribuição Gama, temos

$$D_{ij} = P(j-i|i) = f(j-i; a + b\ i, 2.05) \equiv D_{ij}(a, b).$$

Podemos então reescrever a equação (4.6) na forma:

$$log(L_m) = \sum_{j=0}^{n}(y_j^* log(\sum_{i=0}^{j}\theta_i D_{ij}(a,b)R_j) - \sum_{i=0}^{j}(\theta_i D_{ij}(a,b)R_j)) + \text{constante}, \tag{4.7}$$

sendo que agora o objetivo consiste na estimação do vetor $(\theta_1, \theta_2, \ldots, \theta_n, a, b)$. Para isso, usámos o método de Newton-Raphson.
Considerando agora a Ideia 2, e tal como foi referido, pretende-se fazer variar a média por ano. Como considerámos duas distribuições, teremos que as trabalhar separadamente. Comecemos pela distribuição Gama. Relembrando, se $X \sim \Gamma(k, \theta)$, então

$$D_{ij} = f(j-i; k, \theta), \qquad E(X) = k\theta, \qquad V(X) = k\theta^2.$$

Ora, considerando a base da Ideia 2, multiplicámos a média por uma constante positiva dependente do tempo e^{γ_i}, isto é

$$\begin{cases} E(X) = \hat{k}\hat{\theta}e^{\gamma_i} \\ V(X) = \hat{k}\hat{\theta}^2 \end{cases} \Leftrightarrow \begin{cases} \hat{k} = \dfrac{E(X)}{\hat{\theta}e^{\gamma_i}} \\ V(X) = \dfrac{E(X)}{\hat{\theta}e^{\gamma_i}}\hat{\theta}^2 \end{cases} \Leftrightarrow \begin{cases} \hat{k} = \dfrac{E(X)^2}{V(X)e^{2\gamma_i}} \\ \hat{\theta} = \dfrac{V(X)e^{\gamma_i}}{E(X)} \end{cases}$$

A ideia é semelhante ao método dos momentos: para $k = 5.7$ e $\theta = 2.05$, calculámos a média e a variância; depois, aplicámos o método dos momentos para calcular $\hat{k}$ e $\hat{\theta}$ tal que a média seja igual ao produto da média (com os parâmetros iniciais) com a constante positiva dependente do tempo. Resumindo, ficámos com

$$D_{ij}(\gamma_i) = f(j-i; \hat{k}, \hat{\theta})$$

Podemos então reescrever a equação (4.6) na forma:

$$log(L_m) = \sum_{j=0}^{n}(y_j^* log(\sum_{i=0}^{j}\theta_i D_{ij}(\gamma_i)R_j) - \sum_{i=0}^{j}(\theta_i D_{ij}(\gamma_i)R_j)) + \text{constante} \tag{4.8}$$

e o objetivo é estimar o vetor $(\theta_1, \theta_2, \ldots, \theta_n, \gamma_1, \gamma_2, \ldots, \gamma_n)$.
Considerando agora a distribuição de Weibull, se $X \sim Weibull(\lambda, k)$, então

$$D_{ij} = f(j-i; \lambda, k), \qquad E(X) = \lambda\Gamma(1+\frac{1}{k}), \qquad V(X) = \lambda^2[\Gamma(1+\frac{2}{k}) - (\Gamma(1+\frac{1}{k}))^2]$$

Ora, considerando, mais uma vez, a base da Ideia 2, multiplicámos a média por uma constante positiva dependente do tempo e^{γ_i}, isto é

$$E(X) = \hat{\lambda}\Gamma(1+\frac{1}{\hat{k}})e^{\gamma_i}, \qquad V(X) = \hat{\lambda}^2[\Gamma(1+\frac{2}{\hat{k}}) - (\Gamma(1+\frac{1}{\hat{k}}))^2], \qquad E(X^2) = \hat{\lambda}^2\Gamma(1+\frac{2}{\hat{k}})$$

$$\frac{E(X^2)}{E(X)^2} = \frac{V(X)+E(X)^2}{E(X)^2} = \frac{V(X)}{E(X)^2} + 1 = \frac{\Gamma(1+\frac{2}{\hat{k}})}{\Gamma(1+\frac{1}{\hat{k}})e^{2\gamma_i}} \tag{4.9}$$

A ideia é semelhante à usada para a distribuição Gama: para $\lambda = 2.5$ e $k = 12.818$, calculámos a média e a variância; depois, aplicámos o método dos momentos para calcular

$\hat{k}$, e neste caso, é necessário aplicar um algoritmo para resolução da equação (4.9). Usámos o método das bissecções, que se trata de um algoritmo iterativo. Calculado $\hat{k}$, rapidamente se obtém $\hat{\lambda}$. Resumindo, ficámos com

$$D_{ij}(\gamma_i) = f(j-i; \hat{\lambda}, \hat{k})$$

Podemos então reescrever a equação (4.6) na forma:

$$log(L_m) = \sum_{j=0}^{n}(y_j^* log(\sum_{i=0}^{j}\theta_i D_{ij}(\gamma_i)R_j) - \sum_{i=0}^{j}(\theta_i D_{ij}(\gamma_i)R_j)) + \text{constante}, \tag{4.10}$$

e o objetivo passa a consistir na estimação do vetor $(\theta_1, \theta_2, \ldots, \theta_n, \gamma_1, \gamma_2, \ldots, \gamma_n)$. De notar que o uso da distribuição de Weibull para a ideia 2 não parece ser exequível, visto que sempre que a função $D_{ij}(\gamma_i)$ tem que ser calculada, um algoritmo iterativo tem que ser executado. Ora, embora o método das bissecções seja de muito rápida execução, temos que relembrar que a função $D_{ij}(\gamma_i)$ é calculada várias vezes ao longo do método de estimação, pelo que o uso do método das bissecções várias vezes pode provocar um aumento significativo no tempo de execução do método de estimação.

4.4 Métodos de Estimação

4.4.1 Introdução

Para se conseguir obter resultados a partir do método de retro-propagação, são fundamentais os métodos de estimação. Inicialmente, implementámos o método de Newton-Raphson, para maximização da log-verosimilhança apresentada em (4.6), com uma penalização. Em [Cole et al., 2013], a presença de uma penalização é justificada com um conhecimento à priori do comportamento da solução. No nosso caso, é importante que não exista uma variabilidade fora do normal entre o número de infetados por VIH em anos consecutivos. Para além disso, a penalização pode ser interpretada como um método que introduz um pequeno aumento na desviância, em troca de uma redução na variabilidade das estimativas. A escolha para a penalização, nesta dissertação, foi baseada em [Bacchetti et al., 1993]:

$$Pen(\theta) = \frac{\lambda_\theta}{2}\sum_{i=0}^{n-2}(log(\theta_i) - 2log(\theta_{i+1}) + log(\theta_{i+2}))^2 \tag{4.11}$$

onde λ_θ é denominado por parâmetro de suavidade. De notar que esta penalização faz todo o sentido no nosso caso. De facto, para um certo i, temos

$$(log(\theta_i) - 2log(\theta_{i+1}) + log(\theta_{i+2}))^2 = \left(log\left(\frac{\theta_i}{\theta_{i+1}}\right) + log\left(\frac{\theta_{i+2}}{\theta_{i+1}}\right)\right)^2$$

o que implica que, quanto maior for a variabilidade entre os anos, maior será o valor para a penalização, tal como seria de esperar.
O método de Newton-Raphson é usado então para maximizar a expressão:

$$log(L_m) - Pen(\theta) \tag{4.12}$$

Para a equação (4.6), foi também usado o algoritmo EM, que mais à frente será detalhado. Por fim, para otimização das equações (4.7), (4.8) e (4.10), foi necessária uma modificação no método de Newton-Raphson.

4.4.2 Método de Newton-Raphson

Considere-se a equação

$$f(x) = 0 \Leftrightarrow f_i(x_1, \dots x_n) = 0, \quad i = 1, \dots n$$

O desenvolvimento em série de Taylor de ordem 1 de f em torno de uma aproximação $x^{(v)}$, da solução $x^\star$, permite reescrever f na forma

$$f_i(x) \simeq f_i(x^{(v)}) + \sum_{j=1}^{n} \frac{\partial f_i(x^{(v)})}{\partial x_j}(x_j - x_j^{(v)}), \quad i = 1, \dots n$$

Fazendo $x = x^{(v+1)}$, obtemos a relação de recorrência do método de Newton:

$$f_i(x^{(v)}) + \sum_{j=1}^{n} \frac{\partial f_i(x^{(v)})}{\partial x_j}(x_j^{(v+1)} - x_j^{(v)}) = 0, \quad i = 1, \dots n \tag{4.13}$$

Usando a matriz jacobiana de f, $J(x)$, podemos escrever a relação (4.13) na forma:

$$f(x^{(v)}) + J(x^{(v)})(x^{(v+1)} - x^{(v)}) = 0$$

Assumindo que $J(x)$ é regular numa vizinhança da solução, podemos então escrever

$$x^{(v+1)} = x^{(v)} - J^{-1}(x^{(v)})f(x^{(v)}), \quad v = 0, 1, \dots \tag{4.14}$$

Contudo, o iterado $x^{(v+1)}$ do método de Newton pode, e deve, obter-se sem recurso ao cálculo da inversa de $J(x^{(v)})$, resolvendo o sistema

$$J(x^{(v)})h^{(v)} = -f(x^{(v)}) \tag{4.15}$$

e considerando $x^{(v+1)} = x^{(v)} + h^{(v)}$ pela equação (4.14).
Naturalmente, em todos os algoritmos iterativos é necessário fixar um critério de paragem. Para o método de Newton-Raphson, optámos pelo seguinte critério:

$$||f(x^{(v)})||_2 < \epsilon \tag{4.16}$$

com $\epsilon = 10^{-4}$.

Falta apenas referir como é resolvido o sistema (4.15). Existem diversos algoritmos para resolução de sistemas de equações lineares: nesta tese, optámos por utilizar o método de Cholesky.

4.4.3 Método de Cholesky

Vamos começar por apresentar um lema e um teorema fundamentais para o melhor entendimento do método em causa.

Definição 4.4.1. Uma matriz A é definida positiva se e só se $x^T Ax > 0, \forall x \neq 0$

Lema 4.4.2. Se $A \in \mathbb{R}^{n\times n}$ é definida positiva, então:

1. A é regular
2. Todas as sub-matrizes principais de A, $A_k, k = 1, \ldots, n$, são definidas positivas e portanto regulares

Demonstração.

1. Se A é singular, $\exists x \neq 0 : Ax = 0$ portanto $x^T Ax = 0$, logo A não seria definida positiva.
2. Seja $x = (x_1, \ldots, x_k, 0, \ldots, 0)^T = (\mathbf{y}, \mathbf{0})$ e $\mathbf{y} \in \mathbb{R}^k, \mathbf{y} \neq 0$. Pode-se concluir que: $0 < x^T Ax = \mathbf{y}^T A_k \mathbf{y}$.

□

Teorema 4.4.3. *Uma matriz $A \in \mathbb{R}^{n\times n}$ simétrica é factorizável na forma $A = GG^T$, sendo G uma matriz triangular inferior, real e regular, se e só se A for definida positiva. Além disso, aquela fatorização é única desde que se tomem para G elementos diagonais todos positivos.*

Demonstração.

- ($\Rightarrow$) Sendo $A = GG^T$, então $x^T Ax = x^T GG^T x = y^T y$, com $y = G^T x$. Como $x \neq 0$ e G é regular, então, se $y \neq 0$, $x^T Ax = y^T y = \| y \|_2^2 > 0$. Isto significa que A é definida positiva.
- ($\Leftarrow$) Definindo $D^{1/2} = diag(\sqrt{d_1}, \ldots, \sqrt{d_n})$, e sob as hipóteses, a fatorização $A = LDL^T = (LD^{1/2})(D^{1/2}L^T) = GG^T$ existe e é única. Vamos mostrar que $a_{ii} = d_i > 0, \quad i = 1, \ldots n$. Seja x a solução do sistema $L^T x = e_i$, então: $0 < x^T Ax = x^T LDL^T x = (L^T x)^T D(L^T x) = e_i^T De_i = d_i, \quad i = 1, \ldots n$. Assim, $D^{1/2} = diag(\sqrt{d_1}, \ldots, \sqrt{d_n})$ é real, e obtém-se $A = GG^T$, com $G = LD^{1/2}$, como vimos anteriormente.

□

Posto isto, falta apenas apresentar os cálculos necessários à fatorização de Cholesky, coluna a coluna.

$$A = GG^T \Leftrightarrow \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix} = \begin{pmatrix} g_{11} & & \\ \vdots & \ddots & \\ g_{n1} & \ldots & g_{nn} \end{pmatrix} \begin{pmatrix} g_{11} & \ldots & g_{1n} \\ & \ddots & \vdots \\ & & g_{nn} \end{pmatrix}$$

Para a determinação da primeira coluna de G, consideremos os seguintes cálculos:

$$\begin{aligned} &Ae_1 = GG^T e_1 = G(g_{11}e_1) = g_{11}Ge_1 \Rightarrow \\ &e_1^T Ae_1 = e_1^T g_{11} Ge_1 \Rightarrow a_{11} = g_{11}^2 \Rightarrow g_{11} = \sqrt{a_{11}} \\ &Ge_1 = \frac{Ae_1}{g_{11}} \Rightarrow e_i^T Ge_1 = \frac{e_i^T Ae_1}{g_{11}} \Leftrightarrow g_{i1} = \frac{a_{i1}}{g_{11}}, \quad i = 2, \ldots, n, \end{aligned}$$

onde e_i corresponde a um vetor coluna com entradas todas iguais a zero, excepto na entrada i, onde é igual a 1.
Supondo que as $j-1$ primeiras colunas de G estão determinadas, o cálculo da coluna j ($j \neq 1$) pode ser feito da seguinte forma:

$$Ae_j = GG^T e_j = G(\sum_{k=1}^{j-1} g_{jk}e_k + g_{jj}e_j) = \sum_{k=1}^{j-1} g_{jk}Ge_k + g_{jj}Ge_j \Rightarrow$$

$$a_{jj} = \sum_{k=1}^{j-1} g_{jk}^2 + g_{jj}^2 \Rightarrow g_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} g_{jk}^2}$$

$$Ge_j = \frac{Ae_j - \sum_{k=1}^{j-1} g_{jk}Ge_k}{g_{jj}} \Leftrightarrow$$

$$g_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} g_{jk}g_{ik}}{g_{jj}}, \quad i = j+1, \ldots, n.$$

Estamos agora em condições de apresentar o pseudocódigo para o cálculo da fatorização de Cholesky:

for $j = 1, \ldots, n$ **do**
  $g_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} g_{jk}^2}$
  for $i = j+1, \ldots, n$ **do**
    $g_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} g_{jk}g_{ik}}{g_{jj}}$
  end for
end for

Finda a apresentação do método da fatorização de Cholesky, falta agora mostrar como é que este método é utilizado na resolução de um sistema de equações lineares $Ax = b$. Para uma matriz A que satisfaz as condições de aplicação da fatorização de Cholesky, tem-se

$$Ax = b \Leftrightarrow GG^T x = b \Leftrightarrow Gy = b \wedge G^T x = y$$

Ou seja, usando o método, podemos resolver um sistema de equações lineares resolvendo dois sistemas, respetivamente, triangular inferior e superior. É importante também referir que o determinante de uma matriz (nas condições necessárias para aplicação do método de Cholesky) pode ser calculado da seguinte forma:

$$det(A) = det(GG^T) = det(G)det(G^T) = det(G)^2 = (\prod_{j=1}^{n} g_{jj})^2$$

Apresentámos agora três pequenos exemplos que usámos para validar a implementação computacional, não só da decomposição de Cholesky, mas também da resolução de sistemas triangulares inferiores e superiores.

Exemplo 4.4.4. Considere-se a seguinte matriz:

$$A = \begin{pmatrix} 6 & 3 & 4 & 8 \\ 3 & 6 & 5 & 1 \\ 4 & 5 & 10 & 7 \\ 8 & 1 & 7 & 25 \end{pmatrix}$$

Aplicando o algoritmo implementado, o resultado foi o seguinte:

$$\hat{G} = \begin{pmatrix} 2.449 & 0.000 & 0.000 & 0.000 \\ 1.225 & 2.121 & 0.000 & 0.000 \\ 1.633 & 1.414 & 2.309 & 0.000 \\ 3.266 & -1.414 & 1.588 & 3.132 \end{pmatrix}$$

Apresentámos também a Tabela 4.1 para validação da implementação, onde $\hat{A} = \hat{G} \times \hat{G}^T$, $\hat{e} = A - \hat{A}$, $ea(\hat{A}) = ||\hat{e}||_2$ e $er(\hat{A}) = ||\hat{e}||_2 / ||A||_\infty$:

Erro absoluto de $\hat{A}$ $(ea(\hat{A}))$	9×10^{-16}
Erro relativo de $\hat{A}$ $(er(\hat{A}))$	0

Tabela 4.1: Erro da solução para o Exemplo 4.4.3.

Exemplo 4.4.5. Considere-se a função de Rosenbrock (função Banana):

$$f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \tag{4.17}$$

De seguida, minimizámos a função usando o método de Newton-Raphson. Inicialmente calculámos o gradiente e a matriz Jacobiana de f num vetor genérico:

$$\bigtriangledown f(x_1, x_2) = (2(-1 + x_1 + 200x_1^3 - 200x_1x_2), 200(-x_1^2 + x_2)) \tag{4.18}$$

$$J(x_1, x_2) = \begin{pmatrix} 2(1 + 600x_1^2 - 200x_2) & -400x_1 \\ -400x_1 & 200 \end{pmatrix}$$

Tomando como vetor de arranque para o método de Newton-Raphson $(x_1^{(0)}, x_2^{(0)}) = (-1, -1)$, a matriz Jacobiana nesse ponto é:

$$J(-1, -1) = \begin{pmatrix} 1602 & 400 \\ 400 & 200 \end{pmatrix}$$

O passo seguinte consiste da resolução do sistema $J(-1, -1)\, x = -\bigtriangledown f(-1, -1) = (804, 400)^T$, usando o método de Cholesky. O resultado é o seguinte:

$$x = \begin{pmatrix} 0.005 \\ 1.990 \end{pmatrix}$$

Posto isto, o valor para o resultado da primeira iteração é:

$$x^{(1)} = x + x^{(0)} = \begin{pmatrix} 0.005 \\ 1.990 \end{pmatrix} + \begin{pmatrix} -1 \\ -1 \end{pmatrix} = \begin{pmatrix} -0.995 \\ 0.990 \end{pmatrix}$$

O algoritmo é iterado, até o critério de paragem (4.16) ser atingido. Foram necessárias 5 iterações até ser calculado o mínimo da função,

$$x^{(4)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Exemplo 4.4.6. Considere-se a seguinte matriz e o problema da determinação da sua decomposição de Cholesky:

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

Esta matriz tem valores próprios aproximadamente iguais a 5.11, 0.09 e -2.20. Como apresenta um valor próprio negativo, não é definida positiva, pelo que não é possível a aplicação do Método de Cholesky, e por conseguinte, não é recomendável usar o Método de Newton-Raphson para um sistema com matriz Jacobiana igual à matriz A (nem possível visto que a nossa implementação usa o Método de Cholesky). Posto isto, é necessário estudar e implementar um algoritmo que seja capaz de resolver esta nova questão. Aliás, ambas as modificações propostas (Ideia 1 e Ideia 2) ao método da retro-propagação levam a que, aquando da otimização da log-verosimilhança, ocorra uma situação semelhante.

4.4.4 Método de Cholesky modificado

O algoritmo que aqui apresentámos foi proposto em [Gill and Murray, 1974], e por conseguinte, o conteúdo desta secção é baseado no mesmo artigo. Considere-se uma matriz A qualquer. Nesta secção o método de Cholesky clássico é alterado de forma a que a questão levantada no Exemplo 4.4.6 seja resolvida. Sempre que A não for definida positiva, a ideia consiste na construção de uma matriz $\bar{\mathrm{A}} = A + E$ tal que $\bar{\mathrm{A}}$ é definida positiva e E é uma matriz diagonal. Antes de apresentar o algoritmo, vamos expor as fórmulas para cada uma das entradas das matrizes L e D, considerando a fatorização $A = GG^T = LDL^T$, para facilitar a compreensão do método. Ora, visto que $G = LD^{1/2}$, onde L é uma matriz triangular inferior com os elementos da diagonal iguais à unidade e $D^{1/2}$ é uma matriz diagonal, e designando por l_{ij} a entrada (i, j) da matriz L, e por d_j a entrada (j, j) da matriz D, temos que

$$\begin{aligned} l_{ij} d_j^{1/2} &= g_{ij} \\ l_{jj} &= 1. \end{aligned}$$

Relembrando que

$$\begin{aligned} g_{jj} &= \sqrt{a_{jj} - \sum_{k=1}^{j-1} g_{jk}^2} \\ g_{ij} &= \frac{a_{ij} - \sum_{k=1}^{j-1} g_{jk} g_{ik}}{g_{jj}} \end{aligned}$$

temos então as seguintes fórmulas:

$$l_{jj}d_j^{1/2} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 d_k}$$

$$\Leftrightarrow d_j = a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 d_k \tag{4.19}$$

$$l_{ij}d_j^{1/2} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{jk}d_k^{1/2} l_{ik}d_k^{1/2}}{l_{jj}d_j^{1/2}}$$

$$\Leftrightarrow l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{jk}l_{ik}d_k}{d_j}. \tag{4.20}$$

Seja β uma constante positiva que será explicada e calculada mais à frente. Esta modificação tem como base duas suposições: após a aplicação da modificação, as entradas da matriz D são todas positivas e os elementos da matriz G são inferiores a β, ou seja, para $k = 1, \ldots, n$

$$d_k > \delta \quad \text{e} \quad |l_{ik}d_k^{1/2}| \leq \beta, \quad i > k \tag{4.21}$$

A ideia básica do método é aumentar os elementos da matriz D, durante a fatorização, para que as condições (4.21) sejam verificadas.
Vamos agora descrever o cálculo da coluna j da fatorização, assumindo que as primeiras $j-1$ colunas estão calculadas, e que a condição (4.21) é verificada para $k = 1, \ldots, j-1$. Em primeiro lugar, considere-se

$$\phi_j = a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 d_k \tag{4.22}$$
$$\bar{d}_j = max(\delta, |\phi_j|)$$

O parâmetro δ corresponde a uma quantidade pequena, introduzida para evitar problemas numéricos. Para testar se $\bar{d}_j$ é um valor aceitável para a entrada (j, j) de D, vamos verificar quais os valores de $l_{ik}\bar{d}_k^{1/2}$ que satisfazem a condição (4.21). Seja $\theta = max\{|l_{ik}d_k|, i = k+1, \ldots, n\}$, então

$$\begin{aligned}
&|l_{ij}\bar{d}_j^{1/2}| \leq \beta \\
&\Leftrightarrow l_{ij}^2\bar{d}_j \leq \beta^2 \\
&\Leftrightarrow l_{ij}^2\bar{d}_j^2 \leq \beta^2\bar{d}_j \\
&\Rightarrow l_{ij}^2\bar{d}_j^2 \leq \theta^2 \leq \beta^2\bar{d}_j \\
&\Rightarrow \quad \frac{\theta^2}{\beta^2} \leq \bar{d}_j
\end{aligned} \tag{4.23}$$

Assim: se um valor de $\bar{d}_j$ é tal que a condição (4.23) é verificada, esse valor não é alterado e temos $d_j = \bar{d}_j$; caso contrário, deve-se fazer d_j igual ao menor valor que satisfaz (4.23). Isto é,

$$d_j = max(\bar{d}_j, \frac{\theta^2}{\beta^2})$$

Depois de calculado d_j, os elementos da coluna j da matriz L são calculados usando o método de Cholesky usual.
Notamos ainda que o elemento d_j pode ser escrita da seguinte forma:

$$d_j = \begin{cases} \delta & \text{se } \delta \geq max(|\phi_j|, \frac{\theta^2}{\beta^2}) \\ |\phi_j| & \text{se } |\phi_j| \geq max(\delta, \frac{\theta^2}{\beta^2}) \\ \frac{\theta^2}{\beta^2} & \text{se } \frac{\theta^2}{\beta^2} \geq max(|\phi_j|, \delta) \end{cases}$$

e isto implica que:

$$d_j = a_{jj} + E_j - \sum_{k=1}^{j-1} l_{jk}^2 d_k$$

onde

$$E_j = \begin{cases} \delta - \phi_j & \text{se } \delta \geq max(|\phi_j|, \frac{\theta^2}{\beta^2}) \\ |\phi_j| - \phi_j & \text{se } |\phi_j| \geq max(\delta, \frac{\theta^2}{\beta^2}) \\ \frac{\theta^2}{\beta^2} - \phi_j & \text{se } \frac{\theta^2}{\beta^2} \geq max(|\phi_j|, \delta) \end{cases} \tag{4.24}$$

Posto isto, a fatorização resultante da modificação é a seguinte:

$$\bar{A} = A + E$$

onde E é uma matriz diagonal com entrada (j, j) igual a E_j. De notar que se a matriz A for definida positiva, então $E = 0_{n,n}$, e caso contrário, $E_j > 0, \forall j$. A fatorização de Cholesky é posteriormente aplicada à matriz $\bar{A}$.
Falta apenas determinar o valor de β. Para isso considere-se o seguinte teorema:

Teorema 4.4.7. *Seja A uma matriz simétrica com entradas que são limitadas por uma constante. Então o elemento j da diagonal da matriz E calculada usando o método de Cholesky modificado é limitado e satisfaz a seguinte inequação:*

$$0 \leq E_j \leq (\xi_j/\beta + (j-1)\beta)^2 + 2(|a_{jj}| + (j-1)\beta^2) + \delta$$

onde

$$\xi_j = max\{|a_{ij}| : i = j+1, \ldots, n\}$$

Demonstração. Relembrando (4.24), naturalmente E_j satisfaz a seguinte condição:

$$E_j \leq \theta^2/\beta^2 + 2|\phi_j| + \delta. \tag{4.25}$$

Considerando também a equação (4.20), e relembrando que $|l_{ik} d_k^{1/2}| \leq \beta$, então:

$$\begin{aligned} |l_{ij} d_j| &\leq |a_{ij}| + |\sum_{k=1}^{j-1} l_{jk} l_{ik} d_k| \\ &\leq \xi_j + \sum_{k=1}^{j-1} |l_{jk} l_{ik} d_k| \\ &\leq \xi_j + \sum_{k=1}^{j-1} \beta^2 \\ &= \xi_j + (j-1)\beta^2 \end{aligned} \tag{4.26}$$

Relembrando, mais uma vez, que $\theta = max\{|l_{ij}d_j|, i = j+1, \ldots, n\}$, verifica-se que θ satisfaz

$$\theta \leq \xi_j + (j-1)\beta^2$$

Para além disso, considerando a definição de ϕ_j em (4.22), temos que

$$\begin{aligned} |\phi_j| &\leq |a_{jj}| + |\sum_{k=1}^{j-1} l_{jk}^2 d_k| \\ &\leq |a_{jj}| + (j-1)\beta^2 \end{aligned} \tag{4.27}$$

Se substituirmos as inequações (4.26) e (4.27) na inequação (4.25), temos:

$$E_j \leq (\xi_j/\beta + (j-1)\beta)^2 + 2(|a_{jj}| + (j-1)\beta^2) + \delta.$$

□

Da inequação (4.25), podemos concluir que:

$$||E||_\infty \leq (\xi/\beta + (j-1)\beta)^2 + 2(\gamma + (j-1)\beta^2) + \delta \equiv h(\beta)$$

onde ξ é a maior entrada, fora da diagonal, em módulo, da matriz E, e γ é a maior entrada, na diagonal, em módulo, da matriz E, e $||E||_\infty = max\{|E_j|\}_{1\leq j\leq n}$.
Agora, resta-nos apenas determinar β que minimize $h(\beta)$ e que, ao mesmo tempo, não modifique A caso este seja uma matriz definida positiva. Vamos começar por calcular β que minimiza $h(\beta)$. Tem-se:

$$\begin{aligned} \frac{dh}{d\beta} &= 2(\frac{\xi}{\beta} + (n-1)\beta)(-\frac{\xi}{\beta^2} + (n-1)) + 4((n-1)\beta) = 0 \Leftrightarrow \\ &\Leftrightarrow (\frac{\xi}{\beta} + (n-1)\beta)(-\frac{\xi}{\beta^2} + (n-1)) = -2((n-1)\beta) \\ &\Leftrightarrow -\frac{\xi^2}{\beta^3} + \frac{\xi}{\beta}(n-1) - \frac{\xi}{\beta}(n-1) + (n-1)^2\beta = -2(n-1)\beta \\ &\Leftrightarrow \frac{\xi^2}{\beta^3} = 2(n-1)\beta + (n-1)^2\beta \\ &\Leftrightarrow \frac{\xi^2}{\beta^4} = (n-1)[2 + (n-1)] \\ &\Leftrightarrow \frac{\xi^2}{\beta^4} = (n-1)(n+1) \\ &\Leftrightarrow \frac{\xi^2}{\beta^4} = n^2 - 1 \\ &\Leftrightarrow \beta^2 = \frac{\xi^2}{\sqrt{(n^2-1)}} \end{aligned}$$

Calculado o extremo, falta verificar se a função é convexa.
Ora

$$\begin{aligned} \frac{d^2h}{d\beta^2} &= 2[(-\frac{\xi}{\beta^2} + (n-1))^2 + (2\frac{\xi}{\beta^3})(\frac{\xi}{\beta} + (n-1)\beta)] + 4(n-1) \\ &= \underbrace{2(-\frac{\xi}{\beta^2} + (n-1))^2}_{>0} + \underbrace{4\frac{\xi^2}{\beta^4}}_{>0} + \underbrace{(n-1)\frac{\xi}{\beta^2}}_{>0} > 0 \end{aligned}$$

Falta agora calcular β tal que, se A for definida positiva, então este não altera a matriz A. Ora, se A for definida positiva, então, da fórmula (4.19), podemos afirmar que:

$$\begin{aligned} d_j = a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 d_k > 0 \\ \Leftrightarrow a_{jj} > \sum_{k=1}^{j-1} l_{jk}^2 d_k \\ \Leftrightarrow a_{jj} > l_{jk}^2 d_k, \qquad \forall j,k \end{aligned}$$

Se assumirmos que

$$\beta^2 \geq max\{|g_{jj}| : j = 1, \ldots, n\} = \gamma$$

temos que $l_{jk}^2 d_k \leq \beta^2$ e portanto resolvemos o problema.
Posto isto, a escolha final para β é:

$$\beta^2 = max\{\gamma, \frac{\xi^2}{\sqrt{(n^2-1)}}, \delta\}.$$

A constante $\delta > 0$ foi introduzida apenas para evitar problemas numéricos.

Exemplo 4.4.8. Considere-se a função presente no Exemplo 4.4.5. Vamos agora resolver o problema de minimização dessa função usando o Método de Cholesky modificado. Tal como no Exemplo 4.4.5, o mínimo da função é corretamente calculado, no entanto, com um maior número de iterações (13). A Figura 4.1 mostra o comportamento deste método aplicado à função de Rosenbrock. Podemos verificar que o percurso da solução, para as 13 iterações, excepto na primeira, percorre a curva de forma quase perfeita.

Exemplo 4.4.9. Considere-se a matriz presente no Exemplo 4.4.6. Tal como foi referido, a matriz não é definida positiva. Por esse motivo, vamos aplicar o método de Cholesky modificado e validar os resultados. Aplicando o algoritmo numérico implementado, os resultados foram os seguintes:

$$\hat{L} = \begin{pmatrix} 1.000 & 0.000 & 0.000 \\ 0.265 & 1.000 & 0.000 \\ 0.530 & 0.429 & 1.000 \end{pmatrix}, \hat{D} = \begin{pmatrix} 3.771 & 0.000 & 0.000 \\ 0.000 & 5.750 & 0.000 \\ 0.000 & 0.000 & 1.121 \end{pmatrix}, \hat{E} = \begin{pmatrix} 2.771 & 0.000 & 0.000 \\ 0.000 & 5.016 & 0.000 \\ 0.000 & 0.000 & 2.243 \end{pmatrix}$$

Podemos verificar que, em virtude dos elementos da diagonal de D serem todos positivos, então a matriz $A + E$ é definida positiva, tal como seria de esperar.
De notar que o erro absoluto de $\bar{\text{A}}$ e o erro relativo de $\bar{\text{A}}$ foram determinados, e ambos foram nulos, validando novamente a implementação.

4.4.5 Algoritmo de busca em linha

Quando o método de Cholesky modificado é utilizado no método de Newton-Raphson, é fundamental introduzir um algoritmo de busca em linha, depois do cálculo de $h^{(k)}$, [Dennis and Schnabel, 1996].

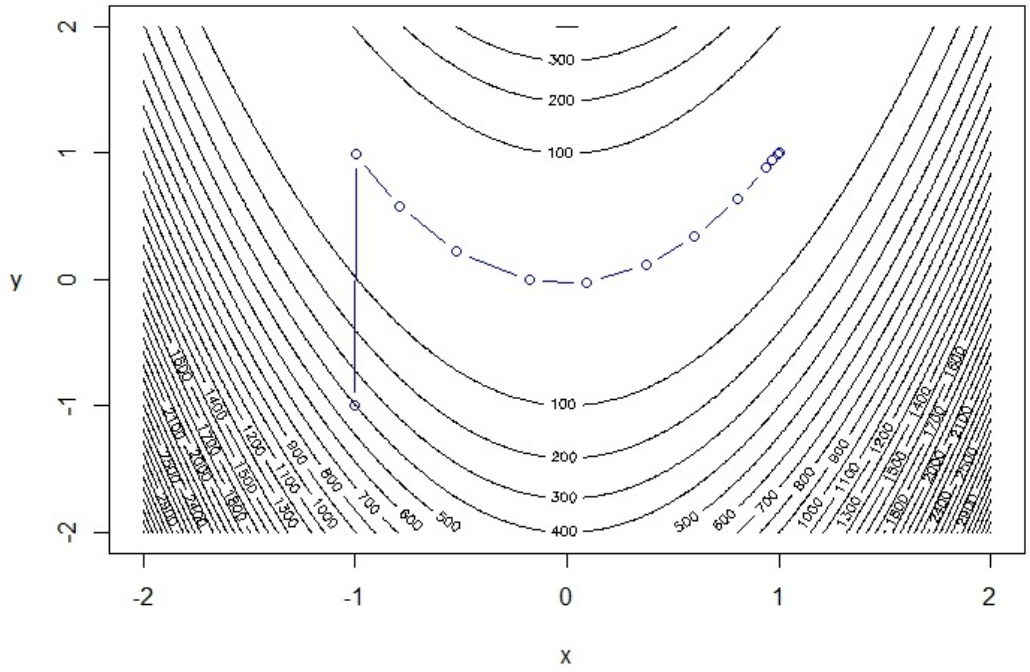


Figura 4.1: Percurso das iterações calculadas pelo método de Cholesky modificado (círculos azuis). As curvas correspondem às coordenadas que partilham o mesmo valor da função de Rosenbrock.

Numa certa iteração k do método de Newton-Raphson, estamos num ponto $x^{(k)}$ e o método calcula uma certa direção $h^{(k)}$. Considere-se a função de uma variável definida por:

$$g(t) = f(x^{(k)} + t\ h^{(k)}) \tag{4.28}$$

Um algoritmo de busca em linha tem como objetivo calcular $t^{(k)}$, minimizando a função (4.28):

$$t^{(k)} = min_t\{f(x^{(k)} + t\ h^{(k)})\}$$

Este tipo de algoritmo permite calcular o melhor benefício ao deslocarmo-nos ao longo de $h^{(k)}$. No fim, tomamos

$$x^{(k+1)} = x^{(k)} + t^{(k)}h^{(k)}.$$

O cálculo exato de $t^{(k)}$ pode, por exemplo, ser efetuado utilizando o método de Newton-Raphson. No entanto, diversas experiências práticas revelaram que esse cálculo exato não é necessário, e que devem ser usadas metodologias mais eficientes, [Quarteroni et al., 2000]. Posto isto, decidimos usar a condição de Armijo.
Usando a regra da cadeia, podemos escrever:

$$g'(t) = \nabla f(x^{(k)} + t\ h^{(k)})h^{(k)}$$

A expansão em série de Taylor de ordem 1 da função g em redor do ponto 0 é:

$$g(t) \approx g(0) + tg'(0) \Rightarrow$$
$$f(x^{(k)} + t\ h^{(k)}) - f(x^{(k)}) \approx t \triangledown f(x^{(k)}) h^{(k)} \tag{4.29}$$

Ou seja, obtivemos uma quantificação do decrescimento aproximado de f ao deslocarmo-nos sobre a reta $x^{(k)} + t\ h^{(k)}$, $t \in \mathbf{R}^+$. Podemos notar que $f(x^{(k)} + t\ h^{(k)}) - f(x^{(k)}) < 0$.
Seja $t = t^{(k)}$ um valor de t que satisfaz a seguinte condição, denominada condição de Armijo:

$$f(x^{(k)} + t^{(k)}\ h^{(k)}) - f(x^{(k)}) < \mu t^{(k)} \triangledown f(x^{(k)}) h^{(k)} \tag{4.30}$$

onde $\mu \in]0, 1[$. Ou seja, o decrescimento obtido da transição em $t = 0$ para $t = t^{(k)}$ não pode apenas ser inferior ao valor aproximado dado em (4.29); tem que ser inferior a uma fração deste. Tipicamente, opta-se por escolher μ tal que este seja inferior a 0.5, [Quarteroni et al., 2000, Armijo et al., 1966].
Posto isto, o algoritmo consiste em averiguar se, com $t = 1$, a condição de Armijo (4.30) é verificada. Caso não seja verificada, toma-se $t = 1/2$ e verifica-se se (4.30) é verificada. Sempre que a condição não for verificada, toma-se $t = 2^{-j}$, $j = 0, 1, \ldots$, até que algum t satisfaça a condição. Quando a condição é finalmente verificada, considera-se $x^{(k+1)} = x^{(k)} + t\ h^{(k)}$. O algoritmo é o seguinte:

$j = 0$
$t = 1$
while $f(x^{(k)} + t^{(k)}\ h^{(k)})h^{(k)} - f(x^{(k)}) > \mu t \triangledown f(x^{(k)}) h^{(k)}$ **do**
    $j = j + 1$
    $t = 2^{-j}$
end while

4.4.6 Algoritmo EM

Por vezes o cálculo da estimativa dada pelo método da máxima verosimilhança pode ser complicado. O algoritmo EM pode ser usado para resolver esse problema.
Antes de começarmos a explicar o método, é essencial introduzir um pequeno teorema:

Teorema 4.4.10 (Desigualdade de Jensen)**.** *Seja f uma função convexa e X uma variável aleatória. Então:*

$$E[f(X)] \geq f(E[X])$$

Se f for uma função côncava, então o contrário é verificado, ou seja:

$$E[f(X)] \leq f(E[X])$$

Para além disso, se f for uma função estritamente convexa (ou estritamente côncava), então $E[f(X)] = f(E[X])$ se e só se $X = E[X]$ com probabilidade 1, ou seja, se X é uma constante.

É também comum [Green, 1990] assumir-se que $\exists\, h : \mathbf{R^m} \longrightarrow \mathbf{R^n}$:

$$X = h(Y)$$

e

$$p(x;\theta) = \sum_{y \in Y(x)} g(y;\theta)$$

$$\text{onde } Y(x) = \{y : h(y) = x\}$$

Vamos agora supor que pretendemos estimar um conjunto de parâmetros θ, usando uma realização $x = (x_1, x_2, \cdots, x_n)$, de uma variável aleatória X seguindo uma distribuição de Poisson. Tal como já foi referido, podemos usar o método da máxima verosimilhança. A função de log-verosimilhança é dada por, usando o resultado anterior:

$$\begin{aligned} log(l(\theta|x)) &= \sum_{i=1}^{n} \log p(x_i;\theta) \\ &= \sum_{i=1}^{n} \log \sum_{y \in Y(x_i)} g(y;\theta) \end{aligned}$$

No entanto, a maximização da equação anterior pode ser complicada. Na situação em que o vetor $y = (y_1, y_2, \cdots, y_m)$ é observado, o algoritmo EM permite, de um modo eficiente, maximizar a função de log-verosimilhança:

$$\begin{aligned} \sum_{i=1}^{n} \log p(x_i;\theta) &= \sum_{i=1}^{n} \log \sum_{y \in Y(x_i)} g(y;\theta) \\ &= \sum_{i=1}^{n} \log \sum_{y} h(y|x_i;\theta) \frac{g(y;\theta)}{h(y|x_i;\theta)} \\ &\geq \sum_{i=1}^{n} \sum_{y} h(y|x_i;\theta) \log \frac{g(y;\theta)}{h(y|x_i;\theta)} \end{aligned} \tag{4.31}$$

onde $h(y|x_i;\theta)$ é uma função densidade de probabilidade arbitrária de y.
Vamos explicar a inequação (4.31). Considerando a desigualdade de Jensen, e como a função $f(x) = log(x)$ é estritamente côncava, e por conseguinte, côncava, visto que $f''(x) = -1/x^2 < 0$, e observando que

$$\sum_{y} h(y|x_i;\theta) \log \frac{g(y;\theta)}{h(y|x_i;\theta)} = E_{h(y|x_i;\theta)}[\log \frac{g(y;\theta)}{h(y|x_i;\theta)}]$$

temos imediatamente que:

$$log(E_{h(y|x_i;\theta)}[\frac{g(y;\theta)}{h(y|x_i;\theta)}]) \geq E_{h(y|x_i;\theta)}[log(\frac{g(y;\theta)}{h(y|x_i;\theta)}])) = \sum_{y} h(y|x_i;\theta) log \frac{g(y;\theta)}{h(y|x_i;\theta)} \equiv F(h,\theta) \tag{4.32}$$

Ao invés de otimizar $log(l(\theta))$, o algoritmo EM otimiza o limite inferior $F(h,\theta)$ da seguinte forma:

Passo E: $h^{(t+1)} = arg\ \max_h F(h, \theta^{(t)})$

Passo M: $\theta^{(t+1)} = arg\ \max_\theta F(h^{(t+1)}, \theta)$.

Iniciando o algoritmo com um valor inicial $\theta^{(0)}$ para θ, no Passo E obtemos uma função $h^{(1)}$ que depende de $\theta^{(0)}$. No Passo M, substituímos a função calculada anteriormente, e calculamos $\theta^{(1)} = \theta$ que maximize $F(h^{(1)}, \theta)$. O método itera entre os dois passos até se obter convergência. Convém, no entanto, referir que podemos reescrever a equação (4.32) na seguinte forma:

$$\begin{aligned} F(h^{(t+1)}, \theta) &= \sum_y h^{(t+1)}(y|x_i;\theta^{(t)}) \log \frac{g(y;\theta)}{h^{(t+1)}(y|x_i;\theta^{(t)})} \\ &= \sum_y h^{(t+1)}(y|x_i;\theta^{(t)}) \log g(y;\theta) - h^{(t+1)}(y|x_i;\theta^{(t)}) \log h^{(t+1)}(y|x_i;\theta^{(t)}) \\ &= Q(\theta|\theta^{(t)}) + K(\theta^{(t)}) \end{aligned}$$

Ou seja, maximizar $F(h, \theta)$ é equivalente a maximizar $Q(\theta|\theta^{(t)}) = E_{h(y|x_i;\theta^{(t)})}[\log g(y;\theta)]$. Posto isto, podemos reescrever o **Passo E** e o **Passo M** da seguinte forma:

Passo E: Exprimir $Q(\theta|\theta^{(t)}) = E_{h(y|x_i;\theta^{(t)})}[\log g(y;\theta)]$

Passo M: $\theta^{(t+1)} = arg\ \max_\theta E_{h(y|x_i;\theta^{(t)})}[\log g(y;\theta)]$

Exemplo 4.4.11. Consideremos x_{ijk} como sendo o número de infetados no mês i, diagnosticados no mês j e notificados no mês k. Vamos assumir que $X_{ijk} \sim P(\theta_i R_{jk} D_{ij})$, e x é o vetor dos valores observados de X_{ijk}. Relembramos também que $Y_j^* \sim P(\sum_{i=0}^{j} D_{ij} R_j \theta_i)$. A função densidade de probabilidade conjunta de x é dada pela seguinte expressão:

$$f(x|\theta) = \prod_{i=1}^{j} \prod_{j=i}^{n} \prod_{k=j}^{n^*} e^{-\theta_i R_{jk} D_{ij}} \frac{(\theta_i R_{jk} D_{ij})^{x_{ijk}}}{x_{ijk}!}. \tag{4.33}$$

Considerando que $\sum_{i=1}^{j} \sum_{k=j}^{n^*} X_{ijk} = Y_j^*$, temos que a distribuição das variáveis aleatórias $X_{1j1}, X_{1j2}, ..., X_{jjn^*}$ corresponde a uma distribuição multinomial com

$$p_{ik} = \frac{\theta_i R_{jk} D_{ij}}{\sum_{i=0}^{j} D_{ij} R_j \theta_i},$$

logo

$$E(x_{ijk}|y_j^*, \theta) = y_j^* \frac{\theta_i R_{jk} D_{ij}}{\sum_{i=0}^{j} D_{ij} R_j \theta_i}. \tag{4.34}$$

Aplicando a função logaritmo à equação (4.33), temos:

$$log(f(x|\theta)) = \sum_{i=1}^{j} \sum_{j=i}^{n} \sum_{k=j}^{n^*} x_{ijk} log((\theta_i R_{jk} D_{ij})) - \theta_i R_{jk} D_{ij} + \text{constante}$$

e usando a equação (4.34), obtemos:

$$E(log(f(x|\theta))|y_j,\theta) = \sum_{i=1}^{j}\sum_{j=i}^{n}\sum_{k=j}^{n^*}[y_j \frac{\theta_i R_{jk} D_{ij}}{\sum_{i=0}^{j} D_{ij} R_j \theta_i} log(\theta_i R_{jk} D_{ij}) - \theta_i R_{jk} D_{ij} + \text{constante}]$$
$$= \sum_{i=1}^{j}\sum_{j=i}^{n}[y_j \frac{\theta_i D_{ij}}{\sum_{i=0}^{j} D_{ij}\theta_i} log(\theta_i) - \theta_i R_j D_{ij} + \text{constante}].$$

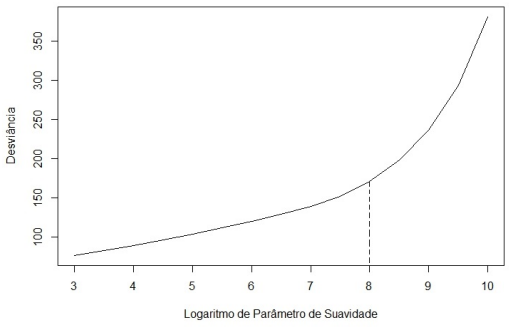
4.5 Resultados

O método de retro-propagação foi aplicado aos dados considerando um agrupamento das contagens de casos de SIDA por ano, de 1986 a 2010, já ajustados para os atrasos na notificação.
Apresentamos os resultados em três partes: na primeira parte descrevemos os resultados obtidos otimizando a expressão (4.6) com a função de penalização (4.11), utilizando o método de Newton-Raphson e o algoritmo EM (designada por metodologia geral); na segunda parte, apresentamos os resultados obtidos pela aplicação da Ideia 1; na terceira e última parte, considerámos uma aplicação da Ideia 2 e apresentamos os seus resultados.
Convém referir que, ao longo desta secção, serão também apresentados os resultados obtidos pela aplicação da R-package Optimx, [Nash and Varadhan, 2011, Nash, 2014], para validação dos resultados. Para além disso, todas as derivadas e matrizes hessianas foram determinadas usando a R-package numDeriv [Gilbert and Varadhan, 2012].
Em primeiro lugar, e antes da apresentação dos resultados, vamos apresentar duas fórmulas fundamentais para os mesmos: a dos valores ajustados e a da desviância, dadas respetivamente por

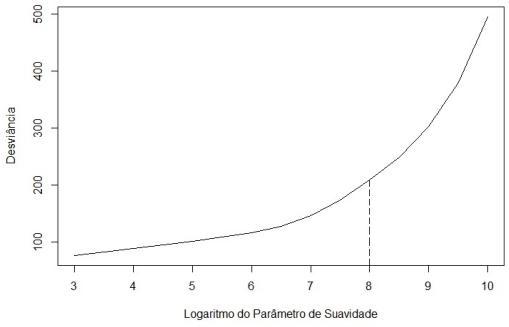
$$\hat{y_j}^* = \sum_{i=0}^{j} \hat{\theta}_i D_{ij} R_j,$$
$$D = -2\times \sum_{j=i}^{n} \left(y_j^* log \frac{\hat{y_j}^*}{y_j^*} - (\hat{y_j}^* - y_j^*) \right)$$

4.5.1 Resultados da aplicação da metodologia geral

Começamos por descrever a forma como λ_θ presente nas expressões (4.11) e (4.12) foi calculado. A ideia consiste da maximização da expressão (4.6) com a função de penalização (4.11), para vários possíveis λ_θ's. Depois, foi calculada a desviância para cada resultado, e considerando o gráfico que apresenta a desviância para cada λ_θ considerado. O λ_θ foi escolhido usando a Regra do Cotovelo. Os resultados podem ser visualizados nas Figuras 4.2(a) e 4.2(b), e em ambos os casos o λ_θ escolhido foi $exp(8)$. Existem algoritmos automáticos para o cálculo de λ_θ, [Bacchetti et al., 1993], no entanto, não foram considerados devido ao elevado peso computacional exigido e ao facto de pequenas perturbações da solução escolhida produzirem resultados bastante semelhantes.
Considerando agora $\lambda_\theta = exp(8)$, vamos maximizar a expressão (4.6) com a função de penalização (4.11). Os resultados (Figura 4.3) sugerem que as duas possíveis distribuições



(a) Considerando a distribuição do tempo de incubação Gama.



(b) Considerando a distribuição do tempo de incubação Weibull.

Figura 4.2: Valor da desviância do modelo para diferentes parâmetros de suavidade.

do tempo de incubação conduzem a estimativas para o número de novos casos de VIH por ano semelhantes. No entanto, a desviância para a distribuição Gama é ligeiramente inferior à da distribuição de Weibull.

Vamos agora proceder à validação dos resultados, usando a R-package Optimx. Na Tabela 4.2, os primeiros 12 algoritmos de otimização correspondem aos algoritmos presentes na Package Optimx, e o último método (Newton-Raphson) corresponde à implementação de raiz realizada nesta dissertação. A Tabela apresenta os diferentes máximos obtidos, bem como a informação sobre se foi atingida convergência ou não. Podemos concluir que o método de Newton-Raphson implementado está a produzir bons resultados, muito semelhantes aos obtidos pelos algoritmos de otimização (da R-package Optimx) BFGS, CG, nlm e ucminf. Podemos, por fim, utilizar as estimativas obtidas pelo modelo para estimar o número futuro

Método	Máximo	Convergência
Nelder-Mead	85960.22	Não
BFGS	**94127.17**	**Sim**
CG	**94127.17**	**Não**
L-BFGS-B	94127.07	Sim
nlm	**94127.17**	**Sim**
nlminb	94125.98	Não
spg	94125.98	Não
ucminf	**94127.17**	**Sim**
newuoa	92977.69	Sim
bobyqa	92357.76	Sim
nmkb	92821.09	Não
hjkb	58712.57	Não
Newton-Raphson	**94127.18**	**Sim**

Tabela 4.2: Resultados obtidos pelos diferentes métodos de otimização, para a metodologia geral e distribuição do tempo de incubação Gama.

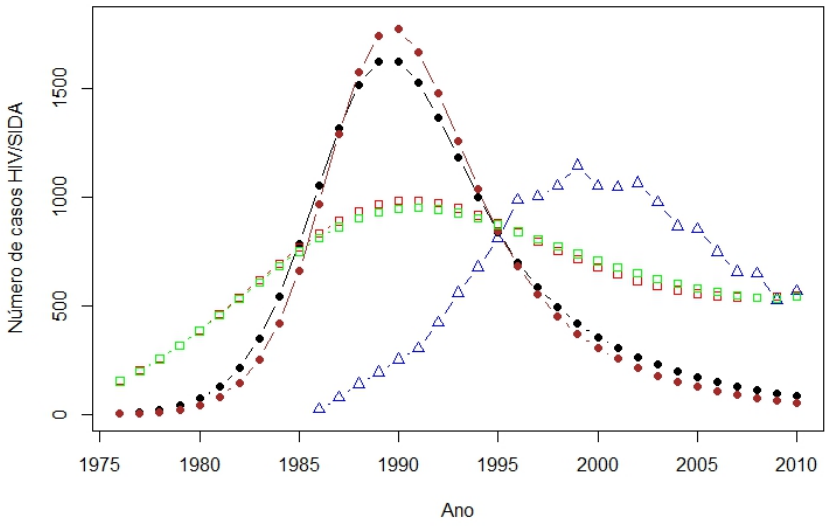


Figura 4.3: Gráfico das estimativas calculadas maximizando a expressão (4.6) com a função de penalização (4.11) e $\lambda_\theta = exp(8)$. Os círculos pretos correspondem às estimativas obtidas pelo método de Newton-Raphson, considerando a distribuição Gama como a distribuição do tempo de incubação; os círculos castanhos correspondem às estimativas obtidos pelo método de Newton-Raphson, considerando a distribuição Weibull como a distribuição do tempo de incubação; os quadrados vermelhos correspondem às estimativas obtidos pelo algoritmo EM, considerando a distribuição Gama como a distribuição do tempo de incubação; os quadrados verdes correspondem às estimativas obtidos pelo algoritmo EM, considerando a distribuição Weibull como a distribuição do tempo de incubação e os triângulos azuis correspondem aos casos de SIDA observados.

de casos de SIDA. Para isso, utilizámos a fórmula:

$$E(Y_j^*) = \sum_{i=0}^{n} \hat{\theta}_i D_{ij} R_j, \qquad j \geq n \tag{4.35}$$

Tal como podemos verificar na Figura 4.4 e na Tabela 4.3, prevê-se uma tendência decrescente do número de casos de SIDA, prevendo-se também que será atingindo um valor inferior a 300 casos no ano de 2014. Neste momento, para que Portugal possa atingir a média Europeia, tem que conseguir reduzir o número de casos de SIDA, atingindo aproximadamente 80 por ano. Os nossos resultados indicam que as medidas tomadas no sentido de controlar a epidemia em Portugal estão a surtir efeito, no entanto, não parecem ser suficientes para que Portugal consiga atingir a média Europeia a curto prazo.

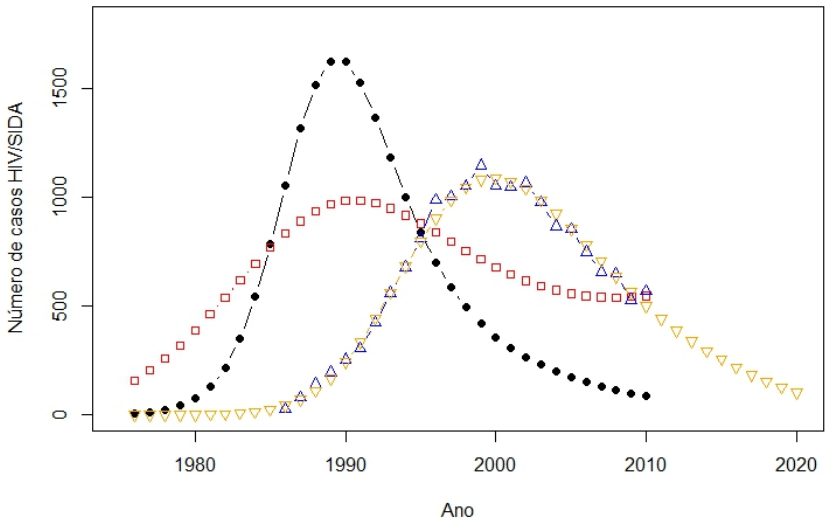


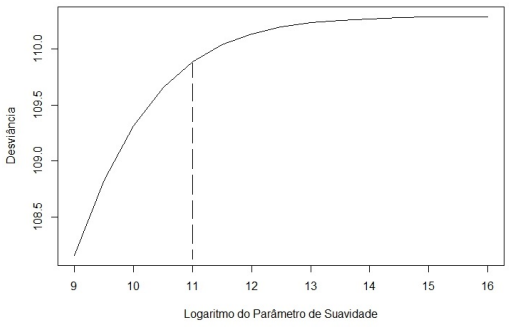
Figura 4.4: Gráfico das estimativas calculadas maximizando a expressão (4.6) com a função de penalização (4.11) e $\lambda_\theta = exp(8)$. Os círculos pretos correspondem às estimativas obtidos pelo método de Newton-Raphson, considerando a distribuição Gama como a distribuição do tempo de incubação; os quadrados vermelhos correspondem às estimativas obtidos pelo algoritmo EM, considerando a distribuição Gama como a distribuição do tempo de incubação; os triângulos azuis correspondem aos casos de SIDA observados e os triângulos laranjas correspondem aos valores previstos para os casos de SIDA utilizando as estimativas dadas pelos círculos pretos.

Ano	Estimativa
2011	441
2012	387
2013	338
2014	294
2015	253
2016	217
2017	183
2018	153
2019	126
2020	103

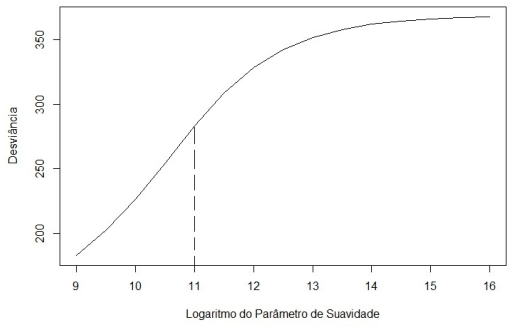
Tabela 4.3: Estimativas para a previsão do número de casos de SIDA, de 2011 a 2020.

4.5.2 Resultados da aplicação da Ideia 1

Para a segunda parte dos resultados, utilizámos também o método descrito na primeira parte para o cálculo de λ_θ. Tendo em conta as Figuras 4.5(a) e 4.5(b), decidimos considerar $\lambda_\theta = exp(11)$ para ambas as distribuições. Ao contrário dos resultados obtidos na secção anterior, os resultados para a Ideia 1 indicam que as duas distribuições produzem resultados



(a) Considerando a distribuição do tempo de incubação Gama



(b) Considerando a distribuição do tempo de incubação Weibull

Figura 4.5: Valor da desviância do modelo para diferentes parâmetros de suavidade, para a aplicação da Ideia 1.

relativamente diferentes. Por esse motivo, vamos apresentá-los separadamente.
Considerando a distribuição Gama, os resultados presentes na Figura 4.6 indicam que o número de infeções tem vindo a decrescer. A figura sugere que o pico de infeções por VIH tenha ocorrido em 1975 (ou antes), o que não faz sentido visto que o primeiro caso identificado de SIDA apenas se verificou em 1981 (a nível mundial). Para além disso, foram obtidas as seguintes estimativas: $\hat{a} = 2.49$ e $\hat{b} = -0.06$. Isto implica que a média da distribuição do tempo de incubação era de aproximadamente 14 anos em 1985, e de 4 anos em 2005. Ora, seria de esperar que ocorresse o contrário, ou seja, que o tempo desde a infeção por VIH até à fase SIDA fosse hoje maior do que era em 1985 em virtude da melhoria dos cuidados de saúde, e consequentemente, da criação de diversos tratamentos que prolongam a esperança média de vida de portadores de VIH.

Vamos agora proceder à validação dos resultados, usando novamente a R-package Optimx. A Tabela 4.4, sugere que o algoritmo Newton-Raphson implementado está a produzir bons resultados, muito semelhantes aos obtidos pelos algoritmos de otimização nlm e BFGS, que são os algoritmos que obtêm o valor mais alto para o máximo.

Considerando a distribuição de Weibull, os resultados presentes na Figura 4.7 indicam que o número de infeções aumentou gradualmente até 2002, começando a decrescer a partir desse ano. Mais uma vez, estes resultados não parecem fazer sentido pois foram obtidas as seguintes estimativas: $\hat{a} = 3.40$ e $\hat{b} = -0.11$, que implicam que a média da distribuição do tempo de incubação é de aproximadamente 9 anos em 1985, e de 1 ano em 2005. Já foi verificado atrás o porquê destas estimativas não fazerem sentido.
Vamos agora proceder à validação dos resultados. A partir da Tabela 4.5, conclui-se que o método de Newton-Raphson implementado está, mais uma vez, a produzir resultados coerentes, muito semelhantes aos obtidos pelo algoritmo de otimização nlm, que é o algoritmo que obtém o valor mais alto para o máximo.

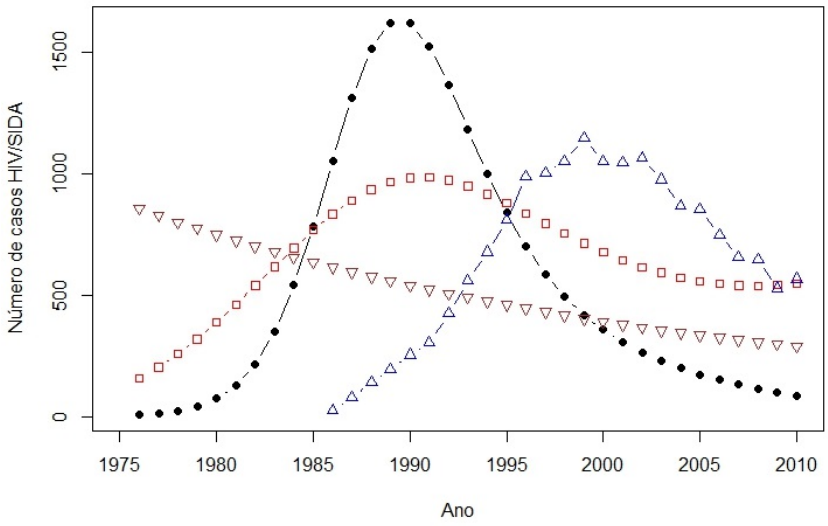


Figura 4.6: Estimativas obtidas maximizando a expressão (4.7) com a função de penalização (4.11) e $\lambda_\theta = exp(11)$ e considerando a distribuição Gama para o tempo de incubação (triângulos castanhos). Os círculos pretos correspondem às estimativas obtidas pelo método de Newton-Raphson na metodologia geral, considerando a distribuição Gama para o tempo de incubação. Os quadrados vermelhos correspondem às estimativas obtidos pelo algoritmo EM na metodologia geral, considerando a distribuição Gama para o tempo de incubação. Os triângulos azuis correspondem aos casos de SIDA observados.

Método	Máximo	Convergência
Nelder-Mead	93621.73	Não
BFGS	**94224.80**	**Sim**
CG	94215.57	Não
L-BFGS-B	94219.51	Sim
nlm	**94224.80**	**Sim**
nlminb	94202.15	Não
spg	94199.25	Não
ucminf	94224.64	Sim
newuoa	93805.39	Sim
bobyqa	93681.01	Sim
nmkb	94010.48	Não
hjkb	93104.38	Não
Newton-Raphson	**94224.80**	**Sim**

Tabela 4.4: Resultados obtidos pelos diferentes métodos de otimização, para a Ideia 1 e distribuição Gama para o tempo de incubação.

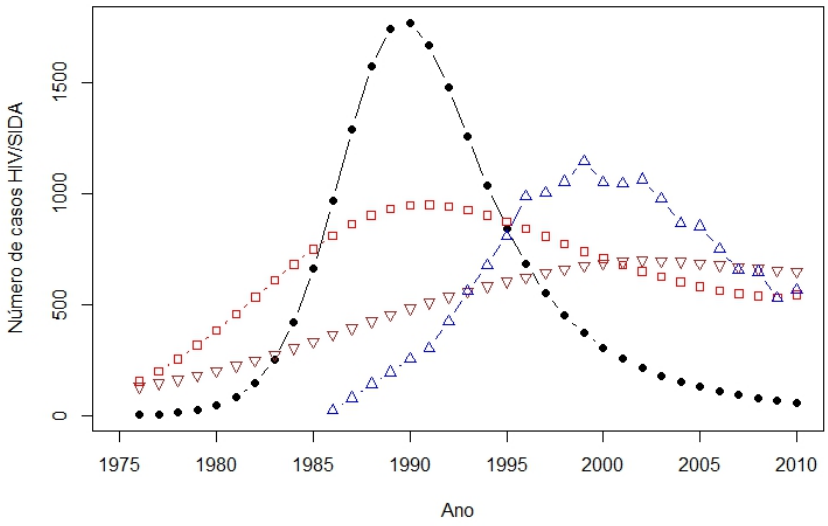


Figura 4.7: Estimativas obtidas maximizando a expressão (4.7) com a função de penalização (4.11) e $\lambda_\theta = exp(11)$ considerando a distribuição de Weibull como a distribuição do tempo de incubação (triângulos castanhos). Os círculos pretos correspondem às estimativas obtidas pelo método de Newton-Raphson na metodologia geral, considerando a distribuição de Weibull como a distribuição do tempo de incubação, os quadrados vermelhos correspondem às estimativas obtidos pelo algoritmo EM na metodologia geral, considerando a distribuição de Weibull como a distribuição do tempo de incubação e os triângulos azuis correspondem aos casos de SIDA observados.

Método	Máximo	Convergência
Nelder-Mead	93156.53	Não
BFGS	94119.62	Sim
CG	94110.58	Sim
L-BFGS-B	94115.09	Sim
nlm	**94119.84**	**Sim**
nlminb	94091.53	Não
spg	94095.53	Não
ucminf	94119.56	Sim
newuoa	93507.94	Sim
bobyqa	93341.98	Sim
nmkb	93405.83	Não
hjkb	92890.58	Não
Newton-Raphson	**94119.84**	**Sim**

Tabela 4.5: Resultados obtidos pelos diferentes métodos de otimização, para a Ideia 1 e distribuição Weibull para o tempo de incubação.

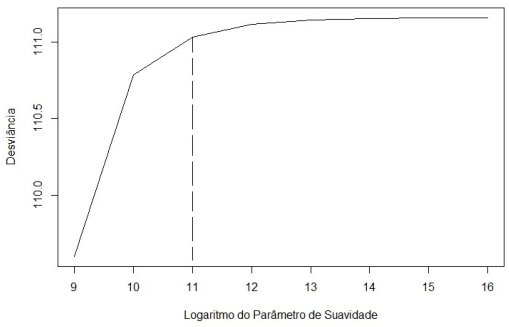
4.5.3 Resultados da aplicação da Ideia 2

Nesta terceira parte dos resultados, considerámos a função a maximizar dada em (4.8) com as seguintes penalizações:

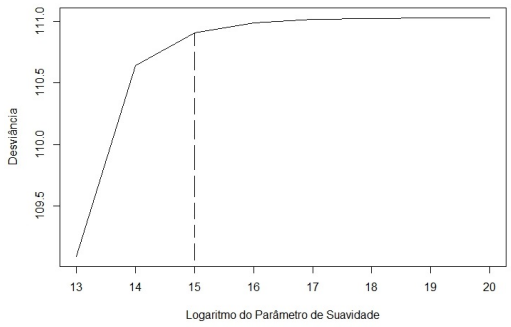
$$log(L_m) - \frac{\lambda_\theta}{2}\sum_{i=0}^{n-2}(log(\theta_i) - 2log(\theta_{i+1}) + log(\theta_{i+2}))^2 - \frac{\lambda_\gamma}{2}\sum_{i=0}^{n-2}(log(e^{\gamma_i}) - 2log(e^{\gamma_{i+1}}) + log(e^{\gamma_{i+2}}))^2. \tag{4.36}$$

Ou seja, para além de assumirmos que as estimativas por ano não podem apresentar uma variação muito elevada de ano para ano, assumimos o mesmo para a média da distribuição do tempo de incubação. Trabalhámos com a distribuição do tempo de incubação Gama, visto que, como já foi referido, assumir uma distribuição de Weibull para o tempo de incubação, na Ideia 2, não parece ser exequível.

Perante isto, é necessário um método diferente do que tem sido usado para o cálculo dos parâmetros de suavidade ótimos. A ideia foi inspirada no algoritmo de cálculo automático presente em [Bacchetti et al., 1993], que começa por atribuir um valor muito elevado ($exp(20)$) para λ_γ, e estima λ_θ usando o algoritmo de cálculo automático. Depois de calculado λ_θ, o algoritmo de cálculo automático é aplicado para o cálculo de λ_γ. Nesta dissertação aplicámos uma regra semelhante, ou seja, usámos a regra do Cotovelo fixando λ_γ, ou seja calculámos λ_θ, e de seguida, fixando λ_θ, calculámos λ_γ. O resultado pode ser verificado nas Figuras 4.8(a) e 4.8(b). As Figuras 4.9(a) e 4.9(b) contêm os resultados da



(a) Estimação de λ_θ, considerando $\lambda_\gamma = exp(20)$.

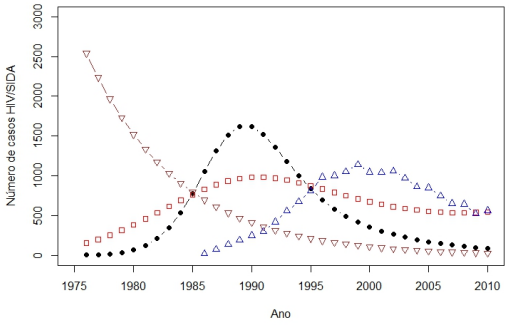


(b) Estimação de λ_γ, considerando $\lambda_\theta = exp(11)$.

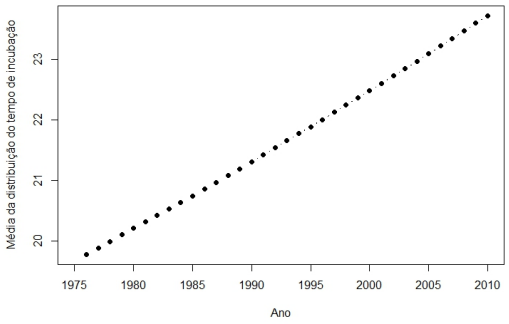
Figura 4.8: Valor da desviância do modelo para diferentes parâmetros de suavidade, para a aplicação da Ideia 2.

aplicação da Ideia 2. Podemos constatar que estes não fazem sentido, em primeiro lugar, pela estimativa obtida para o ano de 1975 (já justificado atrás o porquê desta estimativa não fazer sentido), e em segundo lugar, não há indícios de que a média da distribuição do tempo de incubação seja superior a 20 anos.

Vamos agora proceder à validação dos resultados. Consultando a Tabela 4.6, conclui-se que, o algoritmo de Newton-Raphson implementado produziu resultados válidos, muito semelhantes aos obtidos pelos algoritmos de otimização nlm e BFGS, que foram os algoritmos que obtiveram o valor mais alto para o máximo.



(a) Gráfico da estimativa calculada maximizando a expressão (4.36), com $\lambda_\theta = exp(11)$ e $\lambda_\gamma = exp(15)$ considerando a distribuição Gama para o tempo de incubação (triângulos castanhos). Os círculos pretos correspondem às estimativas obtidas pelo método de Newton-Raphson. Os quadrados vermelhos correspondem às estimativas obtidos pelo algoritmo EM, ambos na metodologia geral considerando a distribuição Gama para o tempo de incubação e os triângulos azuis correspondem aos casos de SIDA observados.



(b) Estimativas para a média da distribuição do tempo de incubação.

Figura 4.9: Estimativas obtidas pela aplicação da Ideia 2.

Método	Máximo	Convergência
Nelder-Mead	92852.37	Não
BFGS	**94224.32**	**Sim**
CG	93262.83	Não
L-BFGS-B	94134.79	Não
nlm	**94224.33**	**Sim**
nlminb	92981.18	Não
spg	93073.45	Não
ucminf	94169.09	Sim
newuoa	92857.28	Sim
bobyqa	92860.17	Sim
nmkb	92848.60	Não
hjkb	92848.60	Não
Newton-Raphson	**94224.33**	**Sim**

Tabela 4.6: Resultados obtidos pelos diferentes métodos de otimização, para a Ideia 2 e distribuição Gama para o tempo de incubação.

Capítulo 5

Conclusões e Trabalho Futuro

Esta dissertação teve como objetivo estimar a taxa de incidência da infeção por VIH em Portugal, a partir dos dados obtidos do sistema de vigilância Português, tendo em consideração os dois problemas que existem no sistema, a sub-notificação e os atrasos na notificação.

Os dados utilizados nesta dissertação foram obtidos a partir de dois sistemas de vigilância Portugueses, o sistema de vigilância intrínseco do Programa Nacional de Luta Contra a Tuberculose (SVIG-TB) e o sistema nacional de notificação de casos de infeção por vírus da imunodeficiência humana, levado a cabo pelo Departamento de Doenças Infeciosas do Instituto Ricardo Jorge (INSA). Foram realizados trabalhos de correção e desenvolvimento das bases de dados, fundamentais para as diversas metodologias consideradas. No âmbito desses trabalhos, foram detetados erros em ambas as bases, bem como incongruências, impossíveis contudo de corrigir. Sugerimos, portanto, que seja criado um mecanismo de colaboração entre os diferentes sistemas de notificação.

Foram várias as metodologias implementadas de raiz, fundamentais para a resolução do objetivo principal desta dissertação. Em primeiro lugar, para o problema da sub-notificação, utilizámos uma metodologia que relaciona o número de casos de Tuberculose com o número de casos de SIDA. Esta metodologia foi desenvolvida em [DeGruttola et al., 1991], embora aplicada num contexto diferente. Para a estimação da taxa de incidência da infeção por VIH, usámos o método de retro-propagação. Para resolver o problema de otimização naturalmente presente no método, implementámos, de raiz, o método de Newton-Raphson. Este, por sua vez, necessitou de vários algoritmos internos, nomeadamente: o método de Cholesky; métodos de resolução de sistemas lineares triangulares superiores e inferiores; o método de Cholesky modificado para os casos em que a matriz Jacobiana não é definida positiva; e o algoritmo de busca em linha. Todos estes algoritmos foram testados e validados usando pequenos exemplos. Para além do método de Newton-Raphson, implementámos também o algoritmo EM, que fornece uma solução utilizada como aproximação inicial para algoritmos mais complexos.

Para a sub-notificação, os resultados pareceram ser congruentes com a situação real, visto que as estimativas decrescem a partir de 2005, ano em que a infeção por VIH passou a integrar a lista das doenças de declaração obrigatória em Portugal. Contudo, o valor

estimado para 2009 é extremamente elevado e contradiz a tendência decrescente definida pelas estimativas anteriores e posteriores. Para a estimação da taxa de incidência da infeção por VIH, a metodologia de retro-propagação prevê uma tendência decrescente do número de casos SIDA, atingindo um valor inferior a 300 casos no ano de 2014. Estes resultados sugerem que as medidas tomadas no sentido de controlar a epidemia em Portugal estão a surtir efeito, no entanto, não parecem ser suficientes para que Portugal atinga a média Europeia a curto prazo. As duas modificações propostas não produziram resultados realistas.

É importante também referir que existem algumas limitações às metodologias utilizadas nesta dissertação. Para a metodologia aplicada ao problema da sub-notificação, as hipóteses assumidas pelo método têm algumas limitações e críticas. Em primeiro lugar, a condição de que os erros são independentes entre os anos pode não ser satisfeita. Em segundo lugar, a assunção de que a base da TB é totalmente correta pode não ser verificada. De facto, tal como já foi referido, essa assunção não é de facto verificada. Para a estimação da taxa de incidência da infeção por VIH, o método de retro-propagação apresenta também algumas limitações. Nomeadamente, supõe-se que a distribuição do tempo de incubação é totalmente conhecida, o que não é verdade. Aliás, é reconhecido que a incerteza associada à distribuição do tempo de incubação é a maior fonte de erro associada aos resultados, [Bacchetti et al., 1993]. Para além disso, não considerámos qualquer correção para a sub-notificação, o que implica que as estimativas obtidas estão subestimadas.

Planeamos, como trabalho futuro, continuar a explorar as potencialidades do método, tentando melhorá-lo no que diz respeito às fragilidades da distribuição do tempo de incubação.

Bibliografia

[Amaral et al., 2005] Amaral, J. A., Pereira, E. P., and Paixão, M. T. (2005). Data and projections of HIV/AIDS cases in Portugal: an unstoppable epidemic? *Journal of Applied Statistics*, 32(2):127–140.

[Armijo et al., 1966] Armijo, L. et al. (1966). Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3.

[Arnold, 1990] Arnold, S. (1990). *Mathematical Statistics*. Prentice-Hall.

[Bacchetti et al., 1993] Bacchetti, P., Segal, M. R., and Jewell, N. P. (1993). Backcalculation of HIV Infection Rates. *Journal of Applied Statistics*, 8(2):pp. 82–101.

[Brookmeyer and Gail, 1988] Brookmeyer, R. and Gail, M. H. (1988). A Method for Obtaining Short-Term Projections and Lower Bounds on the Size of the AIDS Epidemic. *Journal of the American Statistical Association*, 83(402):pp. 301–308.

[Cole et al., 2013] Cole, S. R., Chu, H., and Greenland, S. (2013). Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer. *American Journal of Epidemiology*.

[DeGruttola et al., 1991] DeGruttola, V., Lange, N., and Dafni, U. (1991). Modeling the Progression of HIV Infection. *Journal of the American Statistical Association*, 86(415):pp. 569–577.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38.

[Dennis and Schnabel, 1996] Dennis, Jr., J. E. and Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations (Classics in Applied Mathematics, 16)*. Soc for Industrial & Applied Math.

[Deuffic and Costagliola, 1999] Deuffic, S. and Costagliola, D. (1999). Is the aids incubation time changing? a back-calculation approach. *Statistics in Medicine*, 18(9):1031–1047.

[ECDC/WHO Regional Office for Europe, 2012] ECDC/WHO Regional Office for Europe (2012). HIV/AIDS surveillance in Europe 2011.

[Gilbert and Varadhan, 2012] Gilbert, P. and Varadhan, R. (2012). *numDeriv: Accurate Numerical Derivatives*. R package version 2012.9-1.

[Gill and Murray, 1974] Gill, P. and Murray, W. (1974). Newton-type methods for unconstrained and linearly constrained optimization. *Mathematical Programming*, 7(1):311–350.

[Gill et al., 1981] Gill, P., Murray, W., and Wright, M. (1981). *Practical optimization*. Academic Press.

[Green, 1990] Green, P. J. (1990). On Use of the EM for Penalized Likelihood Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):pp. 443–452.

[Harris, 1990] Harris, J. E. (1990). Reporting Delays and the Incidence of AIDS. *Journal of the American Statistical Association*, 85(412):pp. 915–924.

[Kaas et al., 2008] Kaas, R., Goovaerts, M., Dhaene, J., and Denuit, M. (2008). *Modern Actuarial Risk Theory: Using R*. SpringerLink : Bücher. Springer Berlin Heidelberg.

[Laird et al., 1987] Laird, N., Lange, N., and Stram, D. (1987). Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm. *Journal of the American Statistical Association*, 82(397):pp. 97–105.

[Laird and Ware, 1982] Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4):pp. 963–974.

[Mariotti and Cascioli, 1996] Mariotti, S. and Cascioli, R. (1996). Sources of Uncertainty in Estimating HIV Infection Rates by Back-Calculation: An Application to Italian Data. *Statistics in Medicine*, 15(24):2669–2687.

[Nash, 2014] Nash, J. C. (2014). On Best Practice Optimization Methods in R. *Journal of Statistical Software*, 60(2):1–14.

[Nash and Varadhan, 2011] Nash, J. C. and Varadhan, R. (2011). Unifying Optimization Algorithms to Aid Software System Users: optimx for R. *Journal of Statistical Software*, 43(9):1–14.

[Oliveira et al., 2014] Oliveira, A., Costa, J., and Gaio, R. (2014). The incidence of AIDS in Portugal adjusted for reporting delay and underreporting. *Iberian Conference on Information Systems and Technologies, CISTI*.

[Pestana and Velosa, 2008] Pestana, D. and Velosa, S. (2008). *Introdução à probabilidade e à estatística*. CALOUSTE GULBENKIAN.

[Quarteroni et al., 2000] Quarteroni, A., Sacco, R., and Saleri, F. (2000). *Numerical Mathematics*. Texts in applied mathematics. Springer.

[R Core Team, 2014] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[World Health Organization, 2015] World Health Organization (2015 (visitado em Setembro 2015)). *Tuberculosis*. http://www.who.int/mediacentre/factsheets/fs104/en/.

Apêndice A

Formulário do sistema nacional de notificação de casos de infeção por VIH

Centro de Vigilância Epidemiológica das Doenças Transmissíveis

Vigilância Epidemiológica da Infecção pelo VIH

Folha de Notificação (ver instruções no verso, s.f.f.)

N.º / SIDA *

** A prencher pelo CVEDT*

Ministério da Saúde

1. tipo / classificação

SIDA ☐

CRS-LGP ☐

PA ☐

CDC*

	A	B	C
1			
2			
3			

* Se também disponível

2. datas

Ano provável de infecção _____

Notificação _aa /mm/ dd Diagnóstico _aa /mm/ dd

1.os Sintomas _aa /mm/ dd Falecimento _aa /mm/ dd

3. dados de codificação

Último apelido *(3 prim.as consoantes)* _____

Primeiro nome próprio *(2 prim.as consoantes)* _____

Sexo *(M/F)* ____ Data de nasc. _aa/ mm/ dd Idade _____

Naturalidade_____

Nacionalidade _____

4. residência

Distrito _____ Concelho _____

País de resid.a no provável contágio _____

País de resid.a nos 1.os sintomas _____

5. motivo

Motivo da consulta/internamento ou do teste _____

_____

_____

6. gravidez

Gravidez à data de diagnóstico? SIM ☐ NÃO ☐

Categoria de transmissão da mãe nos casos de mãe-para-filho

☐ Toxicodependente IV

☐ Heterossexual

☐ Transfundida Data aa /mm/ dd País _____

☐ Outras/Indeterminada

7. Viagens/estadas no estrangeiro c/ possibilidades de contágio

País	Datas	Tipo de contágio
_____	aa /mm/ dd	_____
_____	aa /mm/ dd	_____
_____	aa /mm/ dd	_____
_____	aa /mm/ dd	_____

Serviço militar fora de Portugal

_____	aa /mm/ dd	_____
_____	aa /mm/ dd	_____

8. categorias de transmissão

☐ Bissexual ☐ Heterossexual

☐ Homossexual ☐ Toxicodep. IV

☐ Diálise renal ☐ Hemofílico tratado c/ concentrados

☐ Hemofílico tratado/crioprecipitados/plasma

☐ Infecção nosocomial

☐ Transfundido Data _aa /mm/ dd País _____

☐ Transplantado Data _aa /mm/ dd

☐ Trab. sexo ☐ Transmissão mãe-para-filho

☐Outras categorias (especificar)_____

_____

9. características do parceiro no contacto heterossexual

☐ Desconhecido ☐ Hemofílico

☐ HIV 1 positivo ☐ HIV 2 positivo

☐ Homem Bissexual

☐ Originário/residente de país estrang. Qual? _____

☐ Trab. sexo ☐ Toxicodependente IV

☐ Transfundido ☐ Nenhum dos grupos mencionados

10. doenças indicadoras de SIDA

1. Doença _____
   Método de diagnóstico _____
   Data _aa/ mm/ dd Serviço _____
2. Doença _____
   Método de diagnóstico _____
   Data _aa/ mm/ dd Serviço _____
3. Doença _____
   Método de diagnóstico _____
   Data _aa/ mm/ dd Serviço _____
4. Doença _____
   Método de diagnóstico _____
   Data _aa/ mm/ dd Serviço _____

11. serologia VIH

	Data	Data 1.º teste VIH+
☐ Anti-VIH 1	_aa/ mm/ dd	_aa/ mm/ dd
Obs. _____		
☐ Anti-VIH 2	_aa/ mm/ dd	_aa/ mm/ dd
Obs. _____		
☐ Anti-VIH 1+VIH 2	_aa/ mm/ dd	_aa/ mm/ dd
Obs. _____		
☐ WBlot 1	_aa/ mm/ dd	_aa/ mm/ dd
Obs. _____		
☐ WBlot 2	_aa/ mm/ dd	_aa/ mm/ dd
Obs. _____		
☐ Antigénio	_aa/ mm/ dd	_aa/ mm/ dd
Obs. _____		
☐ Outros	_aa/ mm/ dd	_aa/ mm/ dd
Obs. _____		
_____		
_____		

12. entidade que notifica

Nome _____

Serviço _____

Hospital _____

13. Outros Serviços que Contactam ou Contactaram com o doente

_____

_____

_____

_____

Data ___/___/___

Assinatura _____

Vigilância Epidemiológica da Infecção pelo VIH
Folha de Notificação

Instruções para o preenchimento

■ Escrever legivelmente com letra de imprensa.

■ **Ponto 8** – ***Categorias de Transmissão*** –

- pode ser assinalada mais do que uma categoria de transmissão;
- a opção ***Outras categorias*** refere-se a qualquer modo de transmissão não mencionado anteriormente como, por exemplo, corte, picada involuntária por agulha ou contactos com líquidos orgânicos.

■ **Ponto 10** – deve seguir-se a ***"Definição de Casos de SIDA para Fins de Vigilância Epidemiológica, Revisão de 1993"*** (Doc. 77 do C.V.E.D.T./Comissão Nacional de Luta Conta a SIDA, Junho de 1994).

■ **Mais Informações em www.sida.pt**

Envio da Folha de Notificação

■ Enviar a Folha de Notificação para:

Instituto Nacional de Saúde
Centro de Vigilância Epidemiológica
das Doenças Transmissíveis
Av. Padre Cruz
1649 – 016 LISBOA

Tel. 217 519 200
Fax. 217 590 441

Apêndice B

Formulário do sistema de vigilância intrínseco do Programa Nacional de Luta Contra a Tuberculose

Programa Nacional de Luta Contra a Tuberculose
Direcção Geral da Saúde
Ministério da Saúde
SVIG-TB
Sistema de Vigilância da Tuberculose

O Médico

Data

1 U. de Saúde

Nº de Processo

Nº Cartão Utente

Transferido, **já registado**, de outra U. Saúde ☐

Formulário 1 Registo de um caso de Tuberculose, caso novo ou retratamento

2 Identificação

Nome

Sexo M ◯ F ◯ Data Nasc.

País Origem | Desc ◯ | Cidadania | Desc ◯ | Data Entrada em Portugal (mm/aaaa)

Cod Postal – | Concelho | Freguesia

3 Profissão

Profissão/Ocupação | Desc ☐ | **Desempregado** há mais de 24 meses ☐

Área de Actividade
Instituição de Saúde (SNS) ◯ | Residência Comunitária ◯ | Establecimento Prisional ◯ | Outros ◯ | Outros Prestadores de Cuidados de saúde ◯

Especifique a instituição de Saúde em que o doente trabalha (se for profissional do SNS)

4 Detecção

Meio de Detecção
Rastreio Passivo (Sintomas) ◯ | Diagnóstico Pós-Mortem ◯
Rastreio de Contactos ◯ | Outra ◯
Rastreio de Outros Grupos ◯ | Desconhecido ◯

Estado Vital à data do Registo Vivo ◯ Falecido ◯

Critérios Clínico-Imagiológicos Tem ◯ Não Tem ◯ Desconhecido ◯

Ínicio dos Sintomas (mm/aaaa) | Desconhecido ☐

1ª Consulta - qualquer serviço | Desconhecido ☐

Rastreio dos Contactos - Número de cohabitantes seleccionados | Desconhecido ☐

5 Patologias Antes da TB

Infecção por VIH ☐ | Diabetes ☐ | Neoplasia do Pulmão ☐ | Neoplasia de Outros Orgãos ☐
Insuficiência Renal em Diálise ☐ | Doença Inflamatória Articular ☐ | Silicose ☐ | Sarcoidose ☐ | DPOC ☐
Linfomas ou D. Mieloproliferativas ☐ | Outra Doença do Interstício ☐ | Doença Hepática ☐ | Desconhecida ☐ | Outra ☐

6 Grupos de Risco

	Sim	Não	Desc
Dependência Alcoólica	◯	◯	◯
Depedência de Drogas IV	◯	◯	◯
Depedência de outras drogas	◯	◯	◯
Reclusão	◯	◯	◯
Sem Abrigo	◯	◯	◯
Residência Comunitária	◯	◯	◯

Outro Factor de Risco Actual

7 BCG

1ª Inoculação Tem ◯ | Não Tem ◯ | Desc ◯
Última Revacinação Tem ◯ | Não Tem ◯ | Desc ◯
Cicatriz Vacinal Tem ◯ | Não Tem ◯ | Desc ◯

8 Mantoux e IGRA Actuais

Mantoux Tem ◯ Não Tem ◯ Desconhecido ◯ Resultado mm

Teste IGRA Positivo ◯ Negativo ◯ Indeterminado ◯ Não Tem ◯

9 Apresentação Clínica

Tuberculose não Activa
TB - Infecção ☐

TB Doença - Localização	Principal	Secundária
Pulmonar	◯	
Pleural	◯	◯
Linfática Intratorácica	◯	◯
Linfática Extratorácica	◯	◯
Vertebral	◯	◯
Osteoarticular não Vertebral	◯	◯
Meningite	◯	◯

TB Doença - Localização	Principal	Secundária
SNC (não Meningite)	◯	◯
Génito / Urinária	◯	◯
Peritoneal / Digestiva	◯	◯
Disseminada	◯	◯
Outra	◯	◯
Desconhecida	◯	◯

Radiografia do Tórax
Normal ◯
Cavitada ◯
Não Cavitada ◯
Desconhecida ◯

10 Exames

	Data Exame	Produto Biológico	Resultado
Microscopia Inicial			+ ◯ - ◯ Aguarda ◯ Desconhecido ◯
Cultura Inicial			+ ◯ - ◯ Aguarda ◯ Desconhecido ◯
Exame Anátomo Patológico			+ ◯ - ◯ Aguarda ◯ Desconhecido ◯
Teste de Amplificação DNA			+ ◯ - ◯ Aguarda ◯ Desconhecido ◯

11 Tratamento

Nº de Tratamentos Anteriores

Tratamento Anterior Ano | Completado ◯ | Interrompido ou Abandonado ◯ | Insucesso ◯ | Crónico ◯ | Desconhecido ◯

Tratamento Inicial Data de Inicio ou Diagnóstico | **Toma Observada Directamente (TOD)** Sim ◯ Não ◯ Desconhecido ◯

H	R	Z	E	S	Tiac	Km	Am	Cm	Et	Cx	O	Cs	PAS	Rfb	Clo	Pt	Levo	Mox	Gati	Clar	AClav	Lnz	Outra
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

3.0.7.0.v4

Formulário 1 Registo de um Caso de Tuberculose, caso novo ou retratamento.

Definição de Caso de Tuberculose (TB) – para efeitos de notificação no Sistema de Vigilância do Programa de Tuberculose (SVIG-TB) a definição e classificação de caso de TB rege-se pela *Decisão da Comissão Europeia de 28/04/2008 (2008/426/CE).* A notificação de um caso não exige o conhecimento exaustivo da definição, uma vez que, introduzidos os dados disponíveis o SVIG-TB classificará o caso como: **possível, provável ou confirmado.**

INSTRUÇÕES DE PREENCHIMENTO – A numeração das caixas está de acordo com a aplicação informática SVIG-TB 3.0

1. Unidade de Saúde – O código é de preenchimento obrigatório e refere-se à entidade que regista o caso. Corresponde ao serviço que assume o tratamento e/ou a produção de informação relativa ao caso. Preferencialmente deve ser usada a respectiva vinheta. **O processo** tem um número próprio para cada registo de TB em cada Unidade de Saúde. O mesmo indivíduo, se tiver registos sucessivos no mesmo serviço, terá um número de processo diferente em cada registo. O **Número do Cartão de Utente do SNS** é um código nacional imprescindível ao cruzamento deste com outros sistemas de vigilância. Assinale-se **já registado, transferido de outra U. de Saúde** sempre que o doente tenha transitado de outro serviço (CDP ou extensão) da mesma ou outra região de saúde, desde que tenha sido registado no sistema antes da transferência. Para o efeito deste registo não são considerados "transferidos de outra U. de Saúde" os doentes que iniciam o tratamento no hospital e passam a ser seguidos no CDP sem registo prévio.

2. Identificação – Os dados de identificação são de preenchimento obrigatório, assinalando-se "**desconhecido"** quando não houver informação fiável. No espaço: **Pais de origem**, pretende-se que se indique o país de nascimento, se o país de origem é diferente de Portugal, é mandatório o registo da **data de entrada em Portugal**. Se o país de origem ou a cidadania forem desconhecidos, assinale "Desc,". **O código postal** da residência deverá ser preenchido, completo se possível, condição necessária para referenciação geográfica. O **Concelho** e a **Freguesia** de residência são de registo obrigatório. Os dados de residência devem referir-se ao local de permanência com maior significado epidemiológico.

3. Profissão – Se o doente estiver **desempregado há mais de 24 meses**, a informação resume-se a assinalar **"Desempregado".** Caso contrário, é necessário optar por uma das **5 áreas de actividade**. Se a área de actividade for o **Serviço Nacional de Saúde** (SNS), especifique qual a instituição de saúde em que o doente exerce. Qualquer que seja a área de actividade, especifique concretamente a profissão/ocupação, incluindo reformado, estudante ou sem profissão.

4. Detecção – **O meio de detecção** do caso terá sempre que ser referido entre as hipóteses apresentadas. O **estado vita**l, à data do registo, é vivo se tiver iniciado tratamento com dois ou mais antibióticos. A existência de **critérios clínico-imagiológicos** serve para a definição de caso de TB; interessa apenas referir se a decisão de tratamento anti-TB se fundamenta em dados clínicos para além dos laboratoriais. A data da **1ª consulta** refere-se ao atendimento em qualquer serviço de saúde, público ou privado. O **rastreio** deve ser exaustivo nos contactos próximos (convívio acumulado, em espaço confinado por mais de 8 horas), mas, para efeitos de registo, consideram-se os **coabitantes**, i.e., os contactos próximos residentes na mesma habitação (familiar ou outra). Se houver informação fiável, assinale o número de pessoas coabitantes que foram seleccionadas para rastreio (não havendo informação, deduz-se que não houve inventário organizado, e portanto considera-se zero).

5. Patologias Antes da TB – A selecção das **Patologias anteriores à TB** pode ser por escolha múltipla e/ou especificada no espaço disponível. Refere-se apenas a doenças já existentes à data do diagnóstico, mesmo que diagnosticadas posteriormente.

6. Factores de risco – Todas das hipóteses de risco acrescido contempladas carecem de resposta: sim, não ou desconhecido. Outra situação de risco não prevista, deve ser descriminada. A **dependência alcoólica**, é uma informação subjectiva, baseada no score de CAGE. Positivo se o doente tem necessidade de ingerir álcool logo pela manhã ("eye opener"), ou se preencher pelo menos 2 critérios entre 3 seguintes: sentir a necessidade de deixar o consumo de álcool; sentir-se irritado por receber criticas relativas ao álcool; sentir sentimento de culpa por beber. A **dependência de drogas**, endovenosas ou outras, exclui o consumo ocasional, subentendendo-se que haja fenómenos de tolerância e ou sintomas de privação.

7. BCG – A **1ª inoculação pelo BCG** só deve ser considerada como "tem" se estiver documentada. Nesse caso é obrigatório registar a data da vacinação. Caso contrário assinala-se "Desc". Se a resposta for "não" ou "Desc", torna-se desnecessária referência à **ultima revacinação.**

8. Prova de Mantoux Actual – O resultado da prova de **Mantoux** a registar se for assinalado "tem", refere-se ao teste actual. **O Teste IGRA**, se for actual, deve ser assinalada numa das três opções: positivo, negativo ou indeterminado.

9. Apresentação clínica – A apresentação clínica contempla as situações de **tuberculose-infecção**, as situações de **tuberculose activa** (assinale a localização das lesões principais e secundárias) e aspectos da **radiografia do tórax**. A **localização pulmonar**, se existir, será sempre a **principal**. Se houver mais do que duas localizações, sem lesões pulmonares, assinale *Disseminada* na localização principal. Se houver mais do que duas localizações, com lesões pulmonares, assinale *Pulmonar* na localização principal e *Disseminada* na **secundária**. A TB Disseminada inclui ainda a polisserosite e a TB miliar aguda e os casos com isolamentos do *Mt* no sangue (CID 10: A19). Nos casos de TB nas crianças com envolvimento do parênquima pulmonar e linfático locorregional (Complexo Primário), deve assinalar-se *pulmonar* na localização principal e *Linfática intratorácica* na secundária (CID 10: A15.4; A16.3; A16.7). A **classificação radiológica** só é exigida quando a localização for pulmonar (ATS 1980): **Normal**; **Cavitada** – Se houver evidência de cavitação no seio das lesões pulmonares; **Não cavitada** – lesões em qualquer segmento, infiltrados nodulares, densas homogéneas e ou com evidência de atelectasia sem cavitação.

10. Exames - Se tiverem sido efectuados, os resultados dos **exames microscópicos** (directos), **cultural, anatomo-patológico ou do Teste de Amplificação do DNA** são assinalados obrigatoriamente com o produto biológico e data de colheita. Se não houver registo de exame e respectiva data, considera-se que não foi efectuado ou é desconhecido. Se houver registo da data do exame mas não houver resultado, assume-se "Aguarda".

11. Tratamento – Considera-se **tratamento anterior**, a toma de 2 ou mais antibióticos antituberculosos por um período superior a 1 mês. Nos casos em que o estado vital é falecido, à data do registo, sem ter iniciado tratamento, deve registar-se a data do diagnóstico ou do óbito. Nos casos com tratamento/s anterior/es, o **último tratamento** será obrigatoriamente classificado conforme o resultado: Completado – doente tratado anteriormente e declarado curado; Interrompido ou abandono – doente que, em qualquer altura depois de registado, interrompeu o tratamento por 2 meses ou mais e regressa com critérios de doença; Insucesso Terapêutico - doente que anteriormente tinha microscopia ou cultura positiva e que permanece, ou se torna positivo, 5 meses ou mais após o começo do tratamento; Crónico – doente que, após um retratamento completo, permanece com exames bacteriológicos positivos; Desconhecido – doente com tratamento anterior, cujo resultado é desconhecido. Episódio de quimioprofilaxia não é considerado tratamento anterior. A **data do inicio do tratamento** refere-se ao tratamento do episódio actual. Não sendo possível especificar precisamente o inicio do tratamento pode registar-se, como alternativa, a data do diagnóstico. O ano do **último tratamento** deverá ser assinalado se houver tratamentos anteriores. O **regime inicial** de tratamento é o esquema preconizado para o doente, não incluindo as alterações que eventualmente ocorram. (estas registam-se no Formulário 2 Caixa 11A).
A **toma observada directamente** (TOD) é assinalada conforme foi programada na fase inicial do tratamento, independentemente do período por que se vier a prolongar.

Programa Nacional de Luta Contra a Tuberculose
Direcção Geral da Saúde
Ministério da Saúde
SVIG-TB
Sistema de Vigilância da Tuberculose

O Médico

Data

1 U. de Saúde

Nº de Processo

Nº Cartão Utente

Transferido, **já registado**, de outra U. Saúde ☐

Formulário 2 Dados complementares ao registo de caso e declaração do termo de tratamento

Nome

10A Exames

Referente apenas aos casos com microscopia ou cultura positivas na expectoração

Microscopia - Fase da Microscopia Positiva (M+)

Data da Primeira Positiva

Data da Última Positiva

Microscopia - Fase da Microscopia Negativa (M-)

Data da Primeira Negativa — Não Tem ◯

Data da Negativa no Último Mês — Não Tem ◯

Cultura - Fase de Cultura Positiva (C+)

Data da Primeira Positiva

Data da Última Positiva

Cultura - Fase de Cultura Negativa (C-)

Data da Primeira Negativa — Não Tem ◯

Data da Negativa no Último Mês — Não Tem ◯

11A Tratamento

Alteração do Tratamento

Fase Manutenção ◯ Pós TSA ◯ Pós Toxicidade ◯ Desconhecido ◯ Data

H	R	Z	E	S	Tiac	Km	Am	Cm	Et	Cx	O	Cs	PAS	Rfb	Clo	Pt	Levo	Mox	Gati	Clar	AClav	Lnz	Outra
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Fase Manutenção ◯ Pós TSA ◯ Pós Toxicidade ◯ Desconhecido ◯ Data

H	R	Z	E	S	Tiac	Km	Am	Cm	Et	Cx	O	Cs	PAS	Rfb	Clo	Pt	Levo	Mox	Gati	Clar	AClav	Lnz	Outra
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

12 Espécie e Antibiograma

Teste Rápido TB-MR Tem ◯ Não Tem ◯ Desconhecido ◯

Isoniazida: Resistente ◯ Sensível ◯ Indeterminado ◯ **Rifampicina:** Resistente ◯ Sensível ◯ Indeterminado ◯

Antibiograma Convencional Inicial Tem ◯ Não Tem ◯ Desconhecido ◯ Data

	H	R	Z	E	S	Tiac	Km	Am	Cm	Et	Cx	O	Cs	PAS	Rfb	Clo	Pt	Levo	Mox	Gati	Clar	AClav	Lnz	Outra
Sensibilidade	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Resistência	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Último Antibiograma Convencional de Controlo Tem ◯ Não Tem ◯ Desconhecido ◯ Data

	H	R	Z	E	S	Tiac	Km	Am	Cm	Et	Cx	O	Cs	PAS	Rfb	Clo	Pt	Levo	Mox	Gati	Clar	AClav	Lnz	Outra
Sensibilidade	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Resistência	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Identificação da Espécie Tem ◯ Não Tem ◯ Desconhecido ◯

tuberculosis complex ◯ *bovis* não BCG ◯ *avium intracelullare* ◯ *xenopi* ◯ *chelonae* ◯ cultura contaminada ◯ outras ◯

tuberculosis ◯ *africanum* ◯ *gordonae* ◯ *kansasii* ◯ *fortuitum* ◯ cultura mista indissociável ◯

13 Genotipagem

Data — Nº da Estirpe

ETR A	MIRU 04 (ETR D)	MIRU 23	MIRU 31	Mtub 21	Mtub 39
ETR B	MIRU 10	MIRU 24	MIRU 39	Mtub 29	QUB 11
ETR C	MIRU 16	MIRU 26	MIRU 40	Mtub 30	QUB 26
MIRU 02	MIRU 20	MIRU 27	Mtub 04	Mtub 34	QUB 4156

14 Serologia VIH

Positivo ◯

Negativo ◯

Desconhecido ◯

15 Final do Tratamento

Toxicidade Fatal (Morte por Toxicidade dos Antituberculosos) Sim ◯ Não ◯ Desconhecida ◯

Termo do Tratamento Data — Tranferência ou Emigração ☐

Motivo do Termo do Tratamento Tratamento Completado ◯ Insucesso Terapêutico ◯ Crónico ◯

Interrupção ou Abandono ◯ Diagnóstico não sustentado ◯ Morte ◯

Rastreio de Contactos Nº de Cohabitantes Rastreados — Desconhecido ☐

3.0.7.0.v4

Formulário 2 Dados complementares ao registo de caso e declaração de termo de tratamento

Qualquer registo conforme o formulário 1 carece de informação de seguimento, sendo indispensável a relativa ao final de tratamento (caixa 15), única forma de poder integrar coortes para análise do resultado do tratamento.

INSTRUÇÕES DE PREENCHIMENTO – A numeração das caixas está de acordo com a aplicação informática SVIG-TB 3.0

1. Unidade de Saúde – O código da Unidade de Saúde e o número do **Processo** referem-se ao serviço que trata presentemente o doente. Se tiver havido transferência de outra Unidade de Saúde, assinale "transferido de outra Unidade de Saúde". Na Unidade de Saúde de origem, ao ser transferido, foi declarado o termo do tratamento por motivo do "transferência ou emigração". Para efeito de cálculo de incidência, um doente transferido, se se observarem as indicações anteriores, conta apenas para a Unidade de Saúde de origem; para efeitos de avaliação dos recursos envolvidos e do resultado do tratamento, conta para a Unidade de Saúde de origem e para a Unidade de Saúde receptora.

10A. Exames – Este quadro destina-se apenas aos casos com exames bacteriológicos positivos na expectoração. Neste quadro resume-se a evolução dos resultados desde a 1ª amostra positiva até à 1ª negativa, referência para o sucesso terapêutico, e à do controlo no último mês, referência para a cura. A referência à primeira amostra com exame directo ou cultural positivo, deve ser feita com base na informação actualizada, não tendo que coincidir com a data da amostra assinalada na ficha de registo de caso. A determinação da negativação comprovada, quer em microscopia, quer em cultura, pode estar dificultada por ausência de produto biológico viável – neste caso assinale "não tem" e, obviamente, omite-se a data.

11A. A alteração do tratamento que se pretende assinalar é qualquer modificação significativa do regime inicial assinado na caixa 11 do Formulário 1, qualquer que seja o motivo a discriminar: 1. Passagem à fase de manutenção; 2. Ajuste em função do TSA; 3. ajuste por toxicidade, ou 4. desconhecido. Deve registar-se a data da alteração e, em cada formulário, podem assinalar-se duas alterações (para mais alterações, em número ilimitado, usar cópia do formulário).

12. Espécie e Antibiograma - Os resultados dos **Antibiogramas (TSA)** incluem o espectro de resistência e de sensibilidade. Se foi efectuado mas ainda não há acesso ao resultado, assinale "**Desc**". O primeiro TSA registado é o **Teste Rápido TB-MR.** É um **teste molecular, geralmente feito no início do tratamento. Se tiver sido feito, analisar o resultado para Isoniazida e para Rifampicina.** O **Antibiograma Convencional Inicial** diz respeito ao isolado antes do início do tratamento actual e deve ser efectuado em todos os casos com isolamento do *Mt*. O **Último Antibiograma de controlo** pressupõe algum tempo de tratamento actual, a avaliar pela data do isolado correspondente. A aplicação SVIG-TB 3.0 permite inserir e guardar o histórico dos TSA que se pretendam registar, em número ilimitado. **Atenção: as datas dos antibiogramas são as datas das culturas correspondentes, e a data da cultura que deu origem ao Antibiograma Inicial tem de ser inferior a 30 dias depois do início do tratamento.** Quanto à **Identificação da espécie**, a base de dados apresenta o grupo *Mybacterium tuberculosis Complex* por defeito. Se houver informação sobre a espécie, deste ou outro grupo, especifique. Consideram-se também válidas as informações "Cultura contaminada" e "Cultura com micobactérias não dissociáveis", devendo-se registar esta informação, se for o caso**.**

13. Genotipagem - Preencher com a **data** da respectiva cultura, com o **número da estirpe** dado pelo laboratório e com os valores de cada um dos loci do conjunto de MIRU-VNTR analisado (conjunto de 12, 15 ou 24).

14. Serologia VIH – A informação ao estado serológico para o VIH pode ser diferente da referida na ficha de registo do caso (Formulário 1, caixa 5 – Patologias antes da TB). Pode tratar-se de um dado conhecido no decurso do tratamento.

15. Final do tratamento – A **Toxicidade** aos fármacos é considerada relevante se implicar alteração do esquema terapêutico. O **Rastreio de Contactos** refere-se aqui ao nº de coabitantes que foram de facto rastreados. Não pode ser maior do que o número de seleccionados registados na caixa 4 do Formulário 1. Caso venha a ser maior, o número de seleccionados da caixa 4 deverá ser rectificado. A **data do termo do tratamento** marca o momento em que o doente pára o tratamento no serviço responsável pela informação. O termo do tratamento pode ser definitivo (ex. Trat. Completado) ou corresponder a um procedimento de alteração de definição de caso com consequente abertura de novo registo (vide Circular Normativa 8 DT – 21\05\00 DGS). Os motivos da paragem do tratamento podem ser: Transferência ou emigração; tratamento completado; interrupção ou abandono; insucesso terapêutico; caso crónico; morte ou diagnóstico não sustentado. Se tiver sido **transferido ou emigrado**, além de assinalar "transferido ou emigrado", deverá ser assinalado também, se for conhecido, o resultado final, ou seja, uma das hipóteses do **motivo do termo do tratamento**. O motivo do termo do tratamento irá definir o caso quanto aos resultados do tratamento: Completado – doente tratado anteriormente e declarado curado. Insucesso Terapêutico – doente que anteriormente tinha microscopia ou cultura positiva que permanece ou se torna positivo, 5 meses ou mais, após o começo do tratamento. Interrompido ou abandonado – doente que em qualquer altura depois de registado, interrompeu o tratamento por 2 meses ou mais, e regressa com critérios de doença. Morte – doente com tuberculose que faleceu antes ou depois do inicio do tratamento independentemente da causa da morte. No caso de a morte ser atribuída a toxicidade do tratamento, assinale-se Sim no campo Morte por Toxicidade dos AT (antituberculosos). Caso contrário, assinale Não; Diagnóstico não sustentado – doente já registado como tuberculoso e que teve evolução que levou à decisão médica de suspensão do tratamento por discordância com o diagnóstico inicial. Só é possível em casos com cultura negativa, desconhecida ou não efectuada.

Nota: *Este formulário serve de suporte para a actualização da informação no decurso da evolução do caso, com periodicidade indeterminada. É obrigatório a comunicação ao fim de cada 3 meses de tratamento e no termo do tratamento, por transferência, cura ou outro motivo. Enquanto ficha de actualização periódica carece apenas da informação nova, mas aquando o termo do tratamento, os dados deverão ser totalmente revistos.*

A **assinatura do médico responsável** pela informação**,** no topo da página, deve ser acompanhada do número da ordem do médico ou, de preferência, da vinheta com o código de barras.