Abstract

We propose the use of statistical emulators for the purpose of valuing mortality-linked contracts in stochastic mortality models. Such models typically require (nested) evaluation of expected values of nonlinear functionals of multi-dimensional stochastic processes. Except in the simplest cases, no closed-form expressions are available, necessitating numerical approximation. Rather than building ad hoc analytic approximations, we advocate the use of modern statistical tools from machine learning to generate a flexible, non-parametric surrogate for the true mappings. This method allows performance guarantees regarding approximation accuracy and removes the need for nested simulation which is the current gold standard. We illustrate our approach with a case study on index-based hedging based on Cairns *et al.* (2014), as well as an example related to a three-factor stochastic mortality model from Chen & Cox (2009).

Statistical Emulators for Pricing and Hedging Longevity Risk Products

J. Risk M. Ludkovski*

April 30, 2015

Disclaimer: The following is a work in progress, and some sections are to be amended. Please do not distribute beyond the conference organizers. Final version of the article will be ready in Summer 2015.

1 Introduction

Longevity risk has emerged as a key research topic in the past decade. Since the seminal work of Lee and Carter there has been a particular interest in building stochastic models of mortality. Stochastic mortality allows to generate a range of future longevity forecasts, and permits the modeler to pinpoint sources of randomness. This framework creates a natural marriage between the statistical problem of calibration, i.e. fitting to past mortality data, and the financial problems of pricing and hedging future longevity risk. At its core, the latter problems reduce to computing expected values of certain functionals of the underlying stochastic processes. For example, the fundamental survival probability for t years for an individual currently aged x can be expressed as

$$P(0,t,x) = \mathbb{E}_0 \left[\exp(-\int_0^t m(s, x+s)\, ds) \right] \tag{1}$$

where $m(s, x+s)$ is the central force of mortality at date s for an individual aged $x+s$. In the stochastic mortality paradigm $m(s, x+s)$ is random, and so one is necessarily confronted with the need to evaluate the corresponding distributions and expectations.

The past decade has witnessed a strong trend towards complexity at both ends of the problem. On the one hand, driven by the desire to provide faithful fits (and forecasts) to existing mortality data, increasingly complex mortality models have been proposed. The latest generation of models feature multi-dimensional, nonlinear stochastic state processes. These models are effective at calibration and emitting desirable forecasts, but lack tractability in terms of closed-form formulas. On the other hand, complex insurance products, such as variable annuities, make valuation and hedging highly nontrivial, and typically call for sophisticated numerical approaches. Taken together, pricing of mortality-linked contracts

*The authors are with the Department of Statistics & Applied Probability, University of California, Santa Barbara CA 93106-3110;

becomes a complex system, feeding multi-dimensional stochastic inputs through a "black box" which can only be understood through a simulation paradigm.

These developments have created a tension between the complexity of mortality models that do not admit explicit computations and the need to price, hedge and risk manage complicated contracts based on such models. Two outcomes have been a growing reliance on Monte Carlo simulation tools and the gap between the academic mortality modeling and the implemented models by the longevity risk practitioners. This development has been accompanied by exploding computational needs. For example, many emerging problem require *nested simulations* which can easily take days to complete. Similarly, many portfolios contain millions of heterogeneous products that must be accurately priced and managed. In this paper we propose to apply modern statistical methods to address this issue. Our approach is to bridge between the mortality modeling and the desired pricing/hedging needs through an intermediate *statistical emulator*. The emulator provides a computationally efficient, high-fidelity surrogate to the actual mortality model. Moreover, the emulator allows a plug-and-play strategy, so that the end user who is in charge of pricing/risk-management can straightforwardly swap one mortality model for another. This modular approach allows a flexible solution to compare and robustify the model-based longevity risk.

Use of *emulators* is a natural solution to handle complex underlying stochastic simulators and has become commonplace in the simulation and machine learning communities. Here we propose to apply this tool to insurance applications, creating a new domain for this statistical theory. Insurance applications require several adjustments, including the need for functional-regression tools. To fix ideas in this article we pursue the fundamental problem of pricing/hedging vanilla life annuities, a foundational problem in life insurance. However, we stress that our method is very general and can be applied in a variety of actuarial contexts. In particular, in future work we plan to extend it to microscopic agent-based models of mortality Barrieu *et al.* (2012) which offer a canonical "complex system" representation of population longevity. We believe that emulators could significantly simplify predictions in these types of models by providing a tractable, statistical representation of demographic interactions within a stochastic dynamic population framework.

The paper is organized as follows: First we formally introduce the concept of a mortality rate and discuss how it can be used in analyzing longevity risk. Next, we discuss several models for mortality rates. Lastly, we discuss one particular model in detail and its applications to index based hedging, with an analysis using data from the Continuous Mortality Investigation (CMI)[1] and England & Wales population[2]

2 Main Objective

We consider a stochastic system with Markov state process $(Y(t))$. Throughout the paper we will identify Y with the underlying stochastic mortality factors. In Section 2.3 we review some of the existing such models and explicit the respective structure of Y. Typically, Y

[1]Data supplied by UK life insurance companies and by actuarial consultancies. See http://www.actuaries.org.uk/research-and-resources/pages/continuous-mortality-investigation

[2]Obtained from Life & Longevity Markets Association (LLMA) mortality indices. See http://www.llma.org/home.html.

is a multivariate stochastic process based on either a stochastic differential equation (SDE) framework or a time-series ARIMA framework. For example, Y may be a Brownian-motion type process or an auto-regressive process.

The goal of the controller is to evaluate the future distribution of functional $F_T(Y(\cdot))$. Our notation indicates that this functional is potentially path-dependent, such as $F_T(Y(\cdot)) = \mathbb{E}[\exp(-\int_T^\infty f(Y(t))\,dt)|Y(T)]$. Moreover, rather than fixing the initial condition, one wishes to know some summary statistic based on the system at time T, such as

- Expected value $E[F_T(Y(T))|Y(0)]$;
- Quantile $q(\alpha; F_T(Y(\cdot)))$;
- Correlation between two functionals, $Corr(F_T^1(Y(\cdot)), F_T^2(Y(\cdot)))$.

Crucially, $F(Y(\cdot))$ is not available explicitly, and can only be sampled using a simulator. Moreover, this simulator is complex, in the sense that it is (i) black-box, i.e. there is no simple or explicit way to describe its form; (ii) simulatable but expensive, i.e. there exists an engine that after some expense can generate independent, identically distributed samples $F(y^n)$. The expensive aspect implies that computational efficiency is desired in using this simulator.

2.1 Life Annuities

In longevity modeling, Y represents the stochastic factors driving the central force of mortality $m(t,x)$. A typical state-of-the-art model decomposes $m(t,x)$ into a longevity trend, an age effect and a cohort effect. Each the above may be modeled in turn by one or more stochastic factors. Both discrete-time ARIMA and continuous-time SDE frameworks have been investigated. Starting from the model for $m(t,x)$ one can classically derive survival probabilities.

As a canonical actuarial contract, we henceforth focus on deferred life annuities. Denote by $B(T,T+s)$ the price of an s-bond at date T with maturity at $T+s$. If τ_x the future lifetime random variable of an individual aged x, assuming that interest rate and longevity risks are independent, the T-years deferred annuity has value

$$a(Y(T);T,x) = \sum_{s=1}^{\infty} B(T,T+s)\mathbb{P}(\tau_x \geq T+s \mid \mathcal{F}_T) \tag{2}$$

with some underlying filtration $\mathcal{F}_T$.

Practically, the above sum is truncated at some pre-specified upper age, say $x = 110$ as in the CMI tables. Except for very simple models, the survival probability is not analytically known and hence neither is (2). Expected value $\mathbb{E}[a(Y(T))]$ is needed for pricing deferred life annuities, e.g. as part as computing existing pension plan liabilities. Quantiles of $a(Y(T))$ are necessary for assessing the value-at-risk as needed for capital requirements (SCR). Correlation of $a(Y(T))$ with prices of other insurance products arises from cross-hedging as is recommended for mitigating raw longevity risk. In all the above cases, without a formula for $a(Y(T))$ one is forced to resort to nested simulation, first evaluating $a(Y(T))$ for some

representative scenarios, and then further manipulating the resulting "empirical" distribution of $a(y^{(n)}(T))$. Emulation provides a statistical framework for optimizing, assessing and improving such two-level simulations.

2.2 Mathematical Background

We now formally define Y. We assume the existence of an index based fund that is available for trading purposes. Let $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t))$ be a complete filtered probability space on which all relevant quantities exist, where Ω is the sample space of all possible states of mortality rates, and $\mathbb{P}$ is the physical probability measure. $(\mathcal{F}_t)$ is the information up to time t of the evolution of the mortality processes.

Let $Y = (Y(t))_{t\in[0,\infty)} = (Y(t_1), \ldots, Y(t_d))_{t\in[0,\infty)}$ be the $d-$dimensional $(\mathcal{F}_t)$ measurable Markov process, the so called *state process*, which models the uncertainty of the evolution of mortality. In this paper we assume existence of a risk-free asset (a bond) $(B_t)_{t\in[0,\infty)}$, with $B_t = \exp(\int_0^t r_u du)$, where $r_t = r(Y(t))$ is the instantaneous risk-free interest rate at time t.

Based on this model, we assume there exists cash flow functionals $F_1, \ldots, F_T$ that give future profits at time t from the evolution of the state process up to time $t, t = 1, \ldots, T$. These cash flow functionals reflect all necessary legal and regulatory requirements. The profit vector can then be recognized as $X = (X_1, \ldots, X_T)$, where $X_t = F_t(Y(s), s \in [0, t]), t = 1, \ldots, T$.

Consider an individual aged x at time 0 whose remaining lifetime random variable is denoted as τ_x. The state process Y captures $\mu = (\mu(t, x))_{t\in[0,\infty),x\in[0,\infty)}$, the mortality rate (instantaneous hazard rate) process for τ_x, where $\mu(t, x)$ is the mortality rate[3] for τ_x at time t. This is interpreted as the instantaneous death rate at time t for an individual aged x. For small dt, the probability of death is approximately $\mu(t, x)dt$. When necessary, i.e. in the case of a longevity index, we will attach a subscript to μ, that is, we write $\mu_k, k = 1, 2, \ldots$.

Let the random survival function of τ_x with mortality rate μ be

$$S(\mu; t, x) := \exp\left(-\sum_{s=1}^{t} \mu_k(s, x + s)\right).$$

i.e. the probability for τ_x to survive from time t to T measured at time 0. Note that

$$\begin{aligned}
\mathbb{P}(\tau_x > T \mid \tau_x > t, \mathcal{F}_u) &= \mathbb{E}[\mathbb{1}_{\tau_x>T} \mid \tau_x > t, \mathcal{F}_u] \\
&= \mathbb{E}\left[\mathbb{E}[\mathbb{1}_{\tau_x>T} | \tau_x > t, \mathcal{F}_T] \mid \mathcal{F}_u\right] \\
&= \mathbb{E}\left[\frac{S(T, x)}{S(t, x)}\middle| \mathcal{F}_u\right] \\
&= \mathbb{E}\left[\exp\left(-\sum_{s=t+1}^{T} \mu(s, x + s)\right)\middle| \mathcal{F}_u\right] \\
&= \mathbb{E}\left[\exp\left(-\sum_{s=t+1}^{T} \mu(s, x + s)\right)\middle| Y(u)\right],
\end{aligned}$$

[3]In the preceding we discussed the central force of mortality $m(t, x)$; under reasonable assumptions this is in fact equal to $\mu(t, x)$. This is shown at the end of this section.

where the last equality follows from the Markov property. We denote this as

$$P(Y(u);t,T,x) := \mathbb{E}\left[\exp\left(-\sum_{s=t+1}^{T}\mu(s,x+s)\right)\middle| Y(u)\right], \tag{3}$$

which is interpreted as the probability of an individual aged x to survive between t and T additional years, given the information at time u. The deterministic analogue of $P(Y(u),t,T,x)$ in actuarial literature is ${}_{T-t}p_{x+t}$, when $Y(u)$ is non-random.

Define the term *calendar year* to be the time from t to $t+1$. We let

$$m(t,x) := \frac{D(t,x)}{E(t,x)},$$

be the central death rate, where $D(t,x)$ is the number of deaths in calendar year t for individuals aged x last birthday, and $E(t,x)$ is the average population during calendar year t for individuals aged x last birthday. We note here that $E(t,x)$ is often approximated by the population aged x last birthday in the middle of the calendar year t. Two common assumptions are made with these ideas in mind (Cairns *et al.*, 2009):

(i) For integers t and x and for all $0 \leq s,u \leq 1$, we have $\mu(t+s,x+u) = \mu(t,x)$; that is, the force of mortality remains constant over calendar year and integer age.

(ii) The sizes of populations for each age remain constant.

Assumption (ii) implies $E(t,x)$ can be measured at any time during calendar year t. With this in mind, assumption (i) implies $\mu_k(t,x) = m_k(t,x)$. Therefore

$$P(Y(u);t,T,x) = \mathbb{E}\left[\exp\left(-\sum_{s=t+1}^{T}m_k(s,x+s)\right)\middle| Y(u)\right].$$

A quantity of important interest that uses deferred survival probabilities is the $T-$year deferred life annuity that pays \$1 annually for life beginning in T years. Let $B(T,T+s)$ be the discount factor (i.e. a zero-coupon bond price) between the dates T and $T+s$. In the basic situation, $B(T,T+s) = (1+r)^{-s}$ there is a constant risk-free interest rate[4] r. Then the value at time T of a life annuity is

$$A(Y(T);T,x) = \sum_{s=1}^{\infty}(1+r)^{-s}P(Y(T);T,T+s,x+T).$$

The popular age-period-cohort (APC) mortality models assume that (see Appendix A for more details)

$$\log m(t,x) = \beta_x^{(1)} + \frac{1}{n_a}\kappa_t^{(2)} + \frac{1}{n_a}\gamma_{t-x}^{(3)}, \tag{4}$$

[4]The ideas in this paper could easily be extended to have the interest rate be a stochastic process. Furthermore, it is worth considering that Jalen & Mamon (2009) discusses correlation structure between mortality and interest rates.

where $Y(t) \equiv (\kappa_t^{(2)}, \gamma^{(3)})$ themselves are stochastic processes and n_a is the number of ages that x can range from. The combination of (3) and (4) offers a glean into the complexity involved in attempting to understand the distribution (and expected value) of $A(Y(T); T, x)$. Our goal is to use statistical emulators to estimate important summary statistics of functionals of a Markov mortality process, such as $A(Y(T); T, x)$.

2.3 Stochastic Mortality

Three major approaches to stochastic mortality have been put forward in the literature. The first approach, pioneered by Lee & Carter (1992), directly treats $m(t, x)$ as a stochastic process, e.g. an ARIMA time-series. This setup allows incorporating demographic insights, as well as disentangling age- and cohort- effects in future forecasts. The second approach, due to Cairns *et al.* (2006) (CBD), generates a stochastic model for the survival probability $P(t, T, x)$, allowing for straightforward pricing of longevity-linked products. However, it is more difficult to calibrate and to obtain reasonable forecasts for future mortality experience in a population as a whole. The third approach works with forward mortality rates, borrowing ideas from fixed income markets. Forward models give a holistic view of how the mortality curves can evolve over time, and present a dynamically consistent structure for mortality forecasting. Once again however, they do not provide easy expressions for $P(t, T, x)$ and hence require further manipulation for pricing purposes.

The flexibility and accuracy of stochastic mortality models has caused a growing desire to create, analyze, and improve various models. Attempts to understand the statistical validity of these models have been done by, for example, Lee & Miller (2001), Brouhns *et al.* (2002), Booth *et al.* (2002), Czado *et al.* (2005), Delwarde *et al.* (2007), and Li & Tan (2009). In addition, there have been several developments extensions of the Lee Carter model by Renshaw & Haberman (2006), Hyndman & Ullah (2007), Plat (2009), Debonneuil (2010), and Cairns *et al.* (2011a). See Appendix A for a brief discussion of the Lee-Carter and CBD models.

Most of these models do not admit closed form expressions for survival probabilities. Consequently, several papers have proposed approximation methods, see Coughlan *et al.* (2011) and Cairns *et al.* (2014). Coughlan *et al.* (2011) uses a bootstrapping approach, while Cairns *et al.* (2014) derives a deterministic approximation, commenting that it is common industry practice. The more flexible tool of Monte Carlo simulations has been applied in Bauer *et al.* (2012) among others. In section 3 we discuss spline and Kriging approaches.

As mentioned, the building block of pricing a life annuity is usually embedded in a larger setting which requires repeated evaluation of the former quantity. For example, the estimation of a deferred life annuity will require a nested simulation, which can be computationally intensive. Bauer *et al.* (2012) addresses nested Monte Carlo simulations in a similar situation, which is calculating the present value of life annuity like instruments in calculating solvency capital requirements.

2.3.1 Two-Population Model

Multi-population mortality modeling is an important problem for insurance practitioners. The creation of quoted and standardized longevity indices implies that any hedging strategy

for longevity risk in a bespoke pool must necessarily contend with basis risk. Consequently, it is necessary to create a model to capture the link between the index and the insured population. From a different angle, some longevity products explicitly integrate mortality experience in several regions, for example across different countries (UK, Germany, Netherlands) or across different constituencies (England vis-a-vis Great Britain).

The correlation structure for mortality across populations is complex. One notable recent contribution is by Cairns *et al.* (2011a, 2014) who considered a hedging problem between the index pool and the insured sub-pool. Cairns *et al.* (2011a) introduces a cointegrated two-population Bayesian model that is based on the Lee-Carter framework (see 2.3.1 for details). Cairns *et al.* (2014) then applies this model to the England & Wales population along with the sub-population from the CMI data which represents an insured population. The main purpose of this application is to analyze the effectiveness of an index based longevity hedge for pensioners or life insurance companies.

This was modeled by building a co-integration model for the respective mortality rates $m_k(t,x)$, similar to (M3) (see Appendix 2.3) but now $\kappa_1(t)$ is a driving force in the sense that $\kappa_1(t)$ is modeled as a random walk with drift

$$\kappa_1(t) = \kappa_1(t-1) + \mu_1 + \sigma_1 Z_1(t), \quad Z_1(t) \overset{\text{iid}}{\sim} N(0,1),$$

while the smaller (insured) population has κ_2 such that

$$S(t) = \kappa_1(t) - \kappa_2(t) = \mu_2 + \phi(S(t-1) - \mu_2) + \sigma_2 Z_2(t-1) + cZ_1(t-1)$$

is modelled as an AR(1) process with the Z as iid $N(0,1)$ random variables, and $c = \sigma_1 - \rho\sigma_2$ captures the covariance (notice that the noise is Z_1 which is the same as the one used in κ_1). In both models γ_k is an AR(2) process.

3 Statistical Emulation

The idea of emulation is to replace the computationally expensive process of running a MC sub-routine to compute $F(y_0)$ with a cheap-to-evaluate surrogate model that statistically predicts $F(y_0)$ based on results from a training dataset. At the heart of emulation is regression. Namely, the above predictions are based on regressing $F(y^n)$ against y^n. If one is able to produce exact values for $F(y^n)$ then the goal of emulation is interpolation, i.e. using existing data to make predictions at new, not yet visited state sites. More typically, if $F(y^n)$ was itself approximated, then the goal of emulator is not only to interpolate but also to *smooth* the noise in these approximations.

Because the structure of F is unknown, the regression method must be nonparametric, in the sense of having a sufficiently rich architecture to approximate F to any arbitrage degree of accuracy. Such frameworks include kernel regressions, splines or Gaussian processes (kriging) methods. Hastie *et al.* (2009) offers a broad overview of spline models, and Roustant *et al.* (2012) introduces kriging with several examples in R.

Formally, the statistical problem of emulation deals with a sampler (or oracle)

$$Z = F(Y) + \epsilon, \tag{5}$$

where F is the unknown *response surface* and ϵ is the sampling noise, assumed to be independent and identically distributed across different calls to the oracle. The aim is to propose a design $\mathcal{D} = \{y^n : n = 1, \ldots, N\}$ and an approximation procedure that would use the queried results $(y, z)^{1:N} = \{(y^n, z^n) : n = 1, \ldots, N\}$, with $z^n = F(y^n) + \epsilon^n$, to construct a fitted response surface $\hat{F}$. The accuracy of $\hat{F}$ is measured using a loss function $L(\hat{F}, F)$. In this paper we will focus on the weighted mean-squared approximation error

$$L(\hat{F}, F) = \int (\hat{F}(y) - F(y))^2 w(y) dy. \tag{6}$$

Because the true F is unknown, the above definition is of course cannot be operationalized and instead a proxy based on the uncertainty (such as Bayesian posterior uncertainty or standard errors) surrounding $\hat{F}$ is applied.

Remark. In this paper we focus on the original task of producing an accurate approximation to F everywhere. In some contexts, accuracy is judged not globally, but locally, so that a differentiated accuracy measure is used. For example, in VaR applications, the model for F must be accurate in the left-tail, but can be rather rough in the right-panel.

3.1 Emulators based on Spline Models

A spline acts similar to a linear model in the sense that it produces a scalar or vector valued response given some input. A motivating example as to why splines can be useful is the computational issue of nested simulation. In the two-population annuity problem studied in section 4, achieving an annuity value with Monte Carlo simulation requires us to simulate two processes $\kappa_1(t)$ and $\kappa_2(t)$ to some deferral period T, and then perform additional simulations from time T. Here we can use the simulated state process $Y(T)$ at time T to produce an annuity value through a spline, reducing the number of simulations significantly.

We give a brief introduction to splines and how they can be used. In the most general case, for a functional F and data X ($p-$dimensional) we produce a linear basis expansion in X by taking M transformations of X via functionals $h_m : \mathbb{R}^p \to \mathbb{R}$ and model

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X).$$

The functions h_m can be chosen to take whatever form is desired. For example if we choose $h_m(X) = X_m, m = 1, \ldots, p$ then we have a linear model in $X = (X_1, \ldots, X_p)$. A piecewise polynomial can be created by letting the h_m be indicator functions; for example, for $\alpha_1 < \alpha_2$ if we let

$$h_1(X) = I(X < \alpha_1), \quad h_2(X) = I(\alpha_1 \leq X < \alpha_2), \quad h_3(X) = I(\alpha_2 \leq X),$$

then we have a piecewise constant expansion divided into three regions.

A cubic spline is a piecewise cubic polynomial that has continuous first and second derivatives at the points of continuity (also referred to as knots). See Hastie *et al.* (2009) for more details including the basis functions involved. A consequence of standard cubic splines is larger variance near the boundaries, since optimization is done through data points and the

constraints at knots reduce uncertainty. Natural cubic splines add the additional constraint that the function is linear outside of the boundary knots, which removes the constraints of continuity of derivatives at these knots, giving additional knots that can be placed in other locations.

3.1.1 One-Dimensional Smoothing Splines

A smoothing spline avoids the knot issue altogether. First consider the one-dimensional case where we have X explaining Y with N points, i.e. $((x_1, y_1), \ldots, (x_N, y_N))$. Here we say out of all functions $f(x)$ with two continuous derivatives, find one that minimizes the penalized residual sum of squares

$$RSS(f, \lambda) = \sum_{i=1}^{N} \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 dt,$$

with a fixed smoothing parameter λ. The first term is a measure of closeness of data, while the second penalizes the fluctuation of the function f. Notice that $\lambda = \infty$, then we have the traditional least squares line fit, since it requires $f''(t) = 0$.

It can be shown that there is a unique minimizer f of the above equation which is a natural cubic spline with knots at the values of the $x_i, i = 1, \ldots, N$. Then we can write

$$f(x) = \sum_{j=1}^{N} N_j(x)\theta_j,$$

where the $N_j(x)$ are the N basis functions corresponding to this family of natural cubic splines. The minimized values of θ satisfy

$$\hat{\theta} = (\mathbf{N}^T\mathbf{N} + \lambda\Omega_N)^{-1}\mathbf{N}^T\mathbf{y},$$

where $\{\mathbf{N}\}_{ij} = N_j(x_i)$ and $\{\Omega_N\}_{jk} = \int N_j''(t)N_k''(t)dt$. The fitted smoothing spline then has representation

$$\hat{f}(x) = \sum_{j=1}^{N} N_j(x)\hat{\theta}_j.$$

There is freedom in choosing λ. As discussed before, $\lambda = \infty$ creates a linear fit, and $\lambda = 0$ gives no restriction as long as f interpolates the data. The parameter λ can be chosen subjectively to meet some desired criteria, or chosen objectively based on a measure like squared prediction error which creates a tradeoff between bias and variance. If we define $\hat{\mathbf{f}}$ by the vector of the fitted $\hat{f}(x_i)$, then we have

$$\hat{\mathbf{f}} = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\Omega_N)^{-1}\mathbf{N}^T\mathbf{y}.$$

If we define $\mathbf{S}_\lambda := \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\Omega_N)^{-1}\mathbf{N}^T$, we have

$$\hat{\mathbf{f}} = \mathbf{S}_\lambda\mathbf{y}.$$

The finite linear operator $\mathbf{S}_\lambda$ is called the smoother matrix. The definition of the effective degrees of freedom of a smoothing spline is

$$\mathrm{df}_\lambda = \mathrm{trace}(\mathbf{S}_\lambda). \tag{7}$$

This allows a more intuitive way to parameterize the smoothing spline. Following this procedure, to find λ we can set df_λ to a specified number and numerically solve (7). Different choices of df_λ will result in different squared prediction error values, so it can be chosen accordingly to minimize this (or any other quantity of interest). A specific definition and consideration of df_λ rather than λ itself is important since it can be used to compare other types of smoothing methods (see Chapter 9 in Hastie *et al.* (2009)).

3.1.2 Multidimensional Splines

One-dimensional smoothing splines generalize to higher dimensions. If we have data $(X, Y) = (x_i, y_i)_{i=1}^N$ with $x_i \in \mathbb{R}^d$, then we consider the minimization problem

$$\min_f \sum_{i=1}^N \{y_i - f(x_i)\}^2 - \lambda J[f] \tag{8}$$

similar to the one-dimensional case but now J is a penalty functional for stabilizing a function in $\mathbb{R}^d$. In the one-dimensional case we had $\int \{f''(t)\}^2 dt$. One example in the two-dimensional case could be

$$J[f] = \int\int_{\mathbb{R}^2} \left[\left(\frac{\partial^2 f(x)}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f(x)}{\partial x_2^2} \right)^2 \right] dx_1 dx_2. \tag{9}$$

The optimization of (8) along with (9) gives a smooth two-dimensional surface which is called a thin-plate spline (TPS). The solution has the form

$$f(x) = \beta_0 + \beta^T x + \sum_{j=1}^N \alpha_j h_j, \tag{10}$$

where $h_j(x) = \|x - x_j\|^2 \log \|x - x_j\|$. The coefficients are found by plugging (10) into (8). With N observation points, the computational complexity of fitting the TPS is $O(N^3)$. In practice, one often creates a lattice of knots covering the domain that resides inside the convex hull of the points. Using K knots reduces the computations to $O(NK^2 + K^3)$. Hastie *et al.* (2009) states that in practice, it is usually sufficient to work with a lattice of knots covering the domain.

3.1.3 Advantages to the spline approach

The spline approach has several appealing properties. One is the concept of interpretability. Given a scatterplot, anyone can draw a curve connecting the dots. A smoothing spline is a mathematically correct way of doing this (in multi-dimensional we have a surface instead of a line, but it still is easy to understand and explain). In addition, it is easy to predict

values using splines, since a spline simply takes an input x and gives an output (estimate) y by plugging x into a fitted function whose form is known after fitting. Splines also have a smoothness property: it maintains a smooth structure that, depending on the choice of the smoothing parameter λ, will give a more natural fit to the data than a line of best fit or a pure interpolation approach. Furthermore, the spline approach offers flexibility in the sense that one can choose λ to make the curve or surface more or less rigid depending on what applies to the data.

Spline models are also easy to fit. In the one-dimensional case, `R` has a built in function `smooth.spline`. While `smooth.spline` gives many optional arguments allowing for additional flexibility, in the simplest case with data (x, y), if the df_λ is chosen for example to be 5, one simply types `smooth.spline(x,y,df=5)`, and `R` returns a function that can now predict values for new x values. In the multi-dimensional case, the `R` package "fields" contains a function `Tps` which functions similarly to `smooth.spline`. Detailed documentation is available for both of these functions.

Lastly, splines have the advantage of being deterministic and nonparametric. According to Cairns *et al.* (2014), pension industries calculate future values using deterministic projections. Furthermore, no assumptions on distribution are necessary for spline estimates.

3.2 Kriging Surrogates

Suppose we have an input $\mathbf{x}$ on a domain $D \subset \mathbb{R}^d$ and an unknown function $y : D \to \mathbb{R}$ that is one sample of a real-valued random field $(Y(\mathbf{x}))_{\mathbf{x}\in D}$. Continuing the examples in this paper, y could be the $\mathcal{F}_T$ annuity functional and $\mathbf{x}$ the value of the state process $Y(T)$. Kriging consists of making predictions of values of $y(\mathbf{x}^{(0)})$ based on the conditional distribution of $Y(\mathbf{x}^{(0)})$ given the realizations $\left(y(\mathbf{x}^{(1)}), \ldots, y(\mathbf{x}^{(n)})\right)$.

3.2.1 Simple Kriging

In simple Kriging (SK), we assume Y has the form

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + Z(\mathbf{x}),$$

where $\mu : D \to \mathbb{R}$ is a deterministic trend function and Z is a mean zero square-integrable process. Z has covariance kernel $C : D^2 \to \mathbb{R}$ that is assumed to be known and will be referred to as the kernel from henceforth. By considering the process $Y(\mathbf{x}) - \mu(\mathbf{x})$, without loss of generality assume that Y is centered. Again the goal is predict the value of $y(\mathbf{x})$ for some new observation $\mathbf{x}$. In terms of minimizing mean square error, we would like to minimize the mean squared error $\text{MSE}(\mathbf{x}) := \mathbb{E}\left[\left(Y(\mathbf{x}) - \boldsymbol{\lambda}(\mathbf{x})^T Y(\mathbf{X})\right)^2\right]$ with respect to $\boldsymbol{\lambda}$. The derivation of the solution is simple and can be found in Roustant *et al.* (2012). It yields the optimal $\boldsymbol{\lambda}$ is $\boldsymbol{\lambda}^*(\mathbf{x}) = \mathbf{C}^{-1}\mathbf{c}(\mathbf{x})$, where $\mathbf{C} = \left(C(x^{(i)}), C(x^{(j)})\right)_{1\le i,j\le n}$ is the covariance matrix of $Y(\mathbf{X})$, and $\mathbf{c}(\mathbf{x}) = \left(C(\mathbf{x}, \mathbf{x}^{(i)})\right)_{1\le i\le n}$ is the vector of covariances between $Y(\mathbf{x})$ and $Y(\mathbf{X})$. By plugging $\boldsymbol{\lambda}^*(\mathbf{x})$ back into the MSE equation, we get the SK variance at $\mathbf{x}$: $s^2_{SK}(\mathbf{x}) := C(\mathbf{x}, \mathbf{x}) - \mathbf{c}(\mathbf{x})^T\mathbf{C}^{-1}\mathbf{c}(\mathbf{x})$. We can also plug in $\boldsymbol{\lambda}^*(\mathbf{x})$ into $\boldsymbol{\lambda}(\mathbf{x})^T Y(\mathbf{X})$ to get the mean prediction at $\mathbf{x}, m_{SK}(\mathbf{x}) := \mathbf{c}(\mathbf{x})^T\mathbf{C}^{-1}\mathbf{y}$. Some nice properties of these values are that m_{SK} interpolates the data $(\mathbf{X}, \mathbf{y})$, and that s^2_{SK} is zero at the points $\mathbf{X}$.

If Z is assumed to be Gaussian, it turns out that the conditional distribution of $Y(\mathbf{x})$ given the observed $Y(\mathbf{X}) = \mathbf{y}$ is

$$Y_{\mathbf{x}} \mid Y(\mathbf{X}) = \mathbf{y} \sim N(m_{SK}(\mathbf{x}, s^2_{SK}(\mathbf{x})).$$

See Roustant *et al.* (2012) for details.

3.2.2 Ordinary and universal Kriging

If the mean function is of the form $\mu(\mathbf{x}) = \sum_{j=1}^{p} \beta_j h_j(\mathbf{x})$ where β_j are unknown constants and h_j are fixed basis functions, then Universal Kriging (UK) is the process of deriving best linear predictions of Y based on the observations $Y(\mathbf{X})$ while simultaneously estimating the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. The case where the basis functions are constants is referred to ordinary Kriging (OK). The UK equations are given by

$$\begin{aligned} m_{UK}(\mathbf{x}) &= \mathbf{f}(\mathbf{x})^T \hat{\boldsymbol{\beta}} + \mathbf{c}(\mathbf{x})^T \mathbf{C}^{-1}(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}) \\ s^2_{UK}(\mathbf{x}) &= s^2_{SK}(\mathbf{x}) + \left(\mathbf{f}(\mathbf{x}^T) - \mathbf{c}(\mathbf{x}^T)\mathbf{C}^{-1}\mathbf{F}\right)^T \left(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F})^{-1}\right) \left(\mathbf{f}(\mathbf{x})^T - \mathbf{c}(\mathbf{x})^T\mathbf{C}^{-1}\mathbf{F}\right), \end{aligned}$$

where $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}, \ldots, f_p(\mathbf{x}),\ \mathbf{F} = \left(\mathbf{f}(\mathbf{x}^{(1)}), \ldots, \mathbf{f}(\mathbf{x}^{(n)})\right)^T$ is the $n \times p$ experimental matrix, and the best linear estimator of $\boldsymbol{\beta}$ is given by the usual formula $\hat{\boldsymbol{\beta}} := \left(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}\right)^{-1} \mathbf{F}^T\mathbf{C}^{-1}\mathbf{y}$.

The properties of universal Kriging are the same as ordinary Kriging, with the additional property that $m_{UK}(\mathbf{x})$ tends to the best linear $\mathbf{f}(\mathbf{x})^T\hat{\boldsymbol{\beta}}$ whenever the covariances $\mathbf{c}(\mathbf{x})$ vanishes, which happens for example when $\mathbf{x}$ is very far from any observed point with respet to some norm $\|\cdot\|$.

With some strong assumptions, universal Kriging can be interpreted as conditional Gaussian distribution. In fact, if $\boldsymbol{\beta}$ has an improper uniform prior over $\mathbb{R}^p$, then given $Y(\mathbf{X}) = \mathbf{y}$, $Y(\mathbf{x})$ has a Gaussian posterior distribution. See Omre (1987) for more details regarding Bayesian Kriging.

3.2.3 Filtering noisy observations with Kriging

This example is of particular importance, since one typically does not have the value of y at the values of $\mathbf{x}$, but instead could produce them using say Monte Carlo methods. In this case for $\mathbf{x} \in D$, we do not have access to $y(\mathbf{x})$ but instead to $y(\mathbf{x}) + \epsilon$ for some noise ϵ. Under this framework it is still possible to derive Kriging approximations. We denote the sequence of noisy measurements as $\tilde{y}_i = y(\mathbf{x}^{(i)}) + \epsilon_i$, and denote $\tau_i^2 = \mathrm{var}(\epsilon_i)$. Here it is not necessary that each $\mathbf{x}^{(i)}$ is distinct. Following the Monte Carlo simulation case, we can assume $\epsilon_i \sim N(0, \tau_i^2)$ and are independent.

If we still say that y is a realization of a Gaussian process, then it is clear that $\tilde{y}_i$ is the realization of $\tilde{Y}_i = Y(\mathbf{x}^{(i)}) + \epsilon_i$. If Y and ϵ_i are independent, the process Y is still Gaussian conditional on $(\tilde{Y}_1, \ldots, \tilde{Y}_n) = (\tilde{y}_1, \ldots, \tilde{y}_n)$, and its mean and variance are given by similar simple Kriging equations,

$$\begin{aligned} m_{SK}(x) &= \mu(\mathbf{x}) + \mathbf{c}(\mathbf{x})^T(\mathbf{C} + \boldsymbol{\Delta})^{-1}(\tilde{\mathbf{y}} - \boldsymbol{\mu}) \\ s^2_{SK}(\mathbf{x}) &= C(\mathbf{x}, \mathbf{x}) - \mathbf{c}(\mathbf{x})^T(\mathbf{C} + \boldsymbol{\Delta})^{-1}\mathbf{c}(\mathbf{x}), \end{aligned}$$

Table 1: Taken from Roustant *et al.* (2012); Covariance kernels available in the `R` package **DiceKriging**.

Gaussian:	$g(h) = \exp\left(-\frac{h^2}{2\theta^2}\right).$
Matérn $\nu = 5/2$:	$g(h) = \left(1 + \frac{\sqrt{5}\lvert h\rvert}{\theta} + \frac{5h^2}{3\theta^2}\right)\exp\left(-\frac{-5\lvert h\rvert}{\theta}\right).$
Matérn $\nu = 3/2$:	$g(h) = \left(1 + \frac{\sqrt{3}\lvert h\rvert}{\theta}\right)\exp\left(-\frac{-3\lvert h\rvert}{\theta}\right).$
Exponential:	$g(h) = \exp\left(-\frac{\lvert h\rvert}{\theta}\right).$
Power-Exponential:	$g(h) = \exp\left(-\left(\frac{\lvert h\rvert}{\theta}\right)^p\right).$

where $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T$, and $\mathbf{\Delta}$ is the diagonal matrix with terms $\tau_1^2, \ldots, \tau_n^2$. Unlike the simple Kriging case, $m_{SK}(\cdot)$ does not interpolate the noisy observations and similarly does not have zero variance at these points.

3.2.4 Covariance kernels and parameter estimation

The covariance function is a crucial part of the Kriging model. In practice, one usually considers stationary kernels, so we instead consider the function covariance function $c(\mathbf{h}) := C(\mathbf{u}, \mathbf{v})$ where $\mathbf{h} := \mathbf{u} - \mathbf{v}$. The `R` package **DiceKriging** has five available one-dimensional kernels which are listed in table 1. For the d-dimensional case, it uses

$$c(\mathbf{h}) = \sigma^2 \prod_{j=1}^{d} g(h_j; \boldsymbol{\theta}_j),$$

where $\mathbf{h} = (h_1, \ldots, h_d)$, and g is a 1-dimensional kernel. The parameters $\boldsymbol{\theta}_j$ are called characteristic length-scales and are discussed in Rasmussen & Williams (2006). This book also discusses the kernels listed in table 1 in detail. The physical interpretation of θ is a universal scaling factor, since it contracts or extends the input h depending on the value of θ.

The choice of kernel results in a different level of smoothness for the associated random processes. For example, a Gaussian kernel gives sample paths with derivatives of all orders.

Currently **DiceKriging** allows the user to choose the covariance parameters θ_j or have them estimated. Two esimation methods are provided which are maximum likelihood estimation, and penalized MLE. More information on this including likelihood functions using the above kernels and also optimization algorithms can be found in Roustant *et al.* (2012).

4 Case Study: Hedging an Index-Based Fund in a Two-Population Model

There has been a lot of recent discussion regarding index-based funds. Information on the death rates of the general public is widely available. If one was to create a market fund that

had the death rates as its price index, a pension fund can initiate a swap to receive floating payments on the index and pay fixed payments. If it were the case that the two populations were perfectly correlated, the pension fund could buy as many units of the swap as it has to pay out to its annuitants, resulting in a situation where the amount paid is nearly equal to the amount received from the swap (it would be exactly equal if the insured population size was equal to the index population size).

We consider a case study from Cairns *et al.* (2014) related to hedging deferred annuities. In our case the two populations are not perfectly correlated, and we seek out how effective such a hedge would be. For an individual aged x, at time T a pension fund will begin making payments of $a_2(Y(T); T, r)$, a negative cash flow. An index fund can provide payments in the form of $\delta a_1(Y(T); T, r)$, where δ is the amount of annuity purchased. For now we ignore what would be a fixed premium. The main quantity of interest is then the value of the hedge portfolio, or $\delta a_1(Y(T); T, r) - a_2(Y(T); T, r)$.

Several risk measures can be used to determine the effectiveness. It is clear that a pension fund would not want this quantity to have high variance, for example. In addition, the value at risk (or similar measures) should not be too high. Until recently, little has been published on the impact of index-based hedges. See Coughlan *et al.* (2011) and Cairns *et al.* (2014) with alternative but similar treatments of this problem.

4.1 Annuity Value Emulation and Approximation

4.1.1 Deterministic Approximations

Because (3) is not available in closed-form for Lee-Carter type models, approximations are needed for the latter computation. As mentioned in Cairns *et al.* (2014), it is usual practice in industry to use a deterministic projection of mortality rates rather than use a simulation approach. There are various reasons to this, one being that with specialized populations is it hard to find a good model to fit the data.

So far we have discussed interest in conditional survival probabilities and annuity values. Since annuity values can be calculated from conditional survival probabilities, our focus for now is to approximate the survival probabilities. The basic idea for the deterministic approximations is that if $\tilde{m}_k(t, x)$ is an unbiased estimate for $m_k(t, x)$, then

$$\begin{aligned} P(Y(u); t, T, x) &= \mathbb{E}\left[\exp\left(-\sum_{s=t+1}^{T} m(s, x+s)\right)\middle| Y(u)\right] \\ &\approx \exp\left(-\sum_{s=t+1}^{T} \mathbb{E}\left(m(s, x+s) \mid Y(u)\right)\right) \\ &= \exp\left(-\sum_{s=t+1}^{T} \tilde{m}(s, x+s)\right). \end{aligned}$$

We now discuss deterministic approximations in the two-population model discussed in this paper. Cairns *et al.* (2014) used the fact that $\mathbb{E}(\kappa_1(T+t)|\kappa_1(T)) = \kappa_1(T) + \mu_1 t$ to

introduce the approximation

$$\tilde{m}_1^{MM1}(T+s,x) = \exp\left[\beta_1(x) + \frac{1}{n_a}(\kappa_1(T) + \mu_1 s) + \frac{1}{n_a}\gamma_1(T+s-x)\right] \tag{11}$$

with MM standing for median mortality.

Since $X(t)$ is mean reverting, it is also suggested to use the approximation for the CMI population of

$$\tilde{m}_2^{MM1}(T+s,x) = \exp\left[\beta_2(x) + \frac{1}{n_a}(\kappa_2(T) + \mu_1 s) + \frac{1}{n_a}\gamma_2(T+s-x)\right], \tag{12}$$

i.e. the same drift as the general population but different starting point. With these thoughts in consideration, we introduce an additional approximation, using the fact that for the $AR(1)$ process in our two population model we have

$$\mathbb{E}(S(t)) = \mu_2(1-\phi^t) + \phi^t S_0.$$

It follows that

$$\mathbb{E}(\kappa_2(t)) = \mathbb{E}(\kappa_1(t)) - \mathbb{E}(S_2(t)) = \kappa_1(0) + \mu_1 s - \mu_2(1-\phi^t) - \phi^t(\kappa_1(0) - \kappa_2(0))$$

and

$$\mathbb{E}(\kappa_2(T+t)|\mathcal{F}_T) = \kappa_1(T) + \mu_1 t - \mu_2(1-\phi^t) - \phi^t(\kappa_1(T) - \kappa_2(T)).$$

We denote $\mathbb{E}(\kappa_2(T+t)|\mathcal{F}_T)$ as $\xi(t,T)$. A reasonable approximation for $m_2(T+s,x)$ then is

$$\tilde{m}_2^{MM2}(T+s,x) := \exp\left[\beta_2(x) + \xi(t,T) + \frac{1}{n_a}\gamma_2(T+s-x)\right]. \tag{13}$$

Remark. Considering that $\tilde{m}_k(t,x) = \mathbb{E}(m_k(s,x+s))$, Jensen's inequality yields

$$\mathbb{E}\left[\exp\left(-\sum_{s=t+1}^{T} m_k(s,x+s)\right)\right] \leq \exp\left(-\sum_{s=t+1}^{T} \tilde{m}_k(s,x+s)\right).$$

Consequently, none of the above approximations are unbiased for the survival probabilities themselves (and subsequently the annuity values).

For comparison purposes we also introduce a "frozen" estimate,

$$\tilde{m}_k^{frz}(T+s,x) = \exp\left[\beta_k(x) + \frac{1}{n_a}\kappa_k(T) + \frac{1}{n_a}\gamma_k(T+s-x)\right]. \tag{14}$$

4.1.2 Monte Carlo

Because the annuity is to be valued at time T, one must integrate $a_k(T)$ against the future distribution of the underlying factors $Y(T)$. In Monte Carlo approach, one first simulates to time T to determine $Y(T)$, and then performs additional simulations from time T to determine $(Y(T))_{t\geq T}$. Suppose the Monte Carlo simulation will consider N scenarios. Let K be the number of forward scenarios from time T and $a^{(i,j)}(Y(0);T,x)$ be the simulated value

for the ith scenario under the jth forward scenario. Then the Monte Carlo estimate for the annuity value is denoted as

$$\hat{a}^{MC}(Y(T);T,x) = \frac{1}{N}\frac{1}{K}\sum_{i=1}^{N}\sum_{j=1}^{K} a^{(i,j)}(Y(0);T,x). \tag{15}$$

Similarly, if we assume $Y(T) = y$, then through Monte Carlo we can have a time T estimated annuity value, denoted

$$\tilde{a}_T^{MC}(y;T,x) = \frac{1}{N}\sum_{i=1}^{N} a^{(i)}(y;T,x). \tag{16}$$

4.1.3 Surrogate Model

An alternative to the nested simulation/approximation strategy is to directly build a response surface model to the quantity of interest. This avoids unnecessary intermediate steps and gives predictions of what is exactly desired. As discussed in section 3.1, we can determine a range (or lattice in the multi-dimensional case) of values for $Y(T)$, and then create a training input set $\mathbf{Y}_T^{(tr)} = (Y^{(1)}(T), \ldots, Y^{(N_{tr,1})}(T))$ and given $Y(T) = Y(T)^{(i)}$, perform $N_{tr,2}$ Monte Carlo simulations to create a training output set (in the one population annuity case it would be $(a_T^{(1)}, \ldots, a_T^{(N_{tr,1})})$. The resulting training set would be used to create a surrogate model that can directly predict annuity values given new values of $Y(T)$. This completely eliminates the problem of nested simulation since we have assumed the first stage of simulation in creating the lattice. The resulting surrogate model can be used to predict values given any input.

In the case of hedging, one can directly create a surrogate model for the hedge value. This is a more accurate approach, since modeling each value and taking differences creates an additional dimension of error. This is an advantage of the surrogate model approach – we can directly model whatever is of interest.

We note an important observation: predictions from the surrogate models only depend on the value of $Y(T)$, thus they are distribution free and no model assumptions are required. In practice, one could use a bootstrap method or a projection method to determine the surrogate models.

4.2 Methods

In fitting and predicting the surrogate models, one must first analyze the true model and make any necessary assumptions. The simulation method involves finding time T values and simulating onward, so one should be aware of the additional uncertainty this creates in parameter values at time T. Depending the model and how parameters are estimated, as well as personal preference, one may wish to recalibrate the model at time T and re-estimate either all or some of the parameters. The procedure for model fitting and updating the model at a future time T should then go as follows:

1. Choose and fit a model with initial data.

2. Determine which quantities are fixed and which are dynamic.

3. Update the dynamic quantities at time T whose values are not already updated through some process.

For the two-population hedge case study, we use the two population-model which has an initial set of estimates from the data, $\beta_i^{(1)}(x), \beta_i^{(2)}(x), \kappa_i(t), \gamma_i(t-x)$, where $i = 1, 2$ for the E&W and CMI populations respectively, and the time and age respectively range from calendar years 1961 to 2005 (with 2005 treated as $t = 0$) and x from 50 to 89. Then random walk and time series models are fitted to $(\kappa_1(t))$ and $(\kappa_1(t) - \kappa_2(t))$, introducing additional estimates $\mu_1, \sigma_1, \mu_2, \phi, \sigma_2$ and c.

Following Cairns *et al.* (2014) who also uses the two-population model in their analysis of longevity hedge effectiveness, we consider three cases for parameter uncertainty, parameter certain (PC), parameter-partial certain (PPC), and parameter uncertain (PU) which all change step 2 above. In all three cases, $\kappa_i(t), i = 1, 2$ is simulated to time T so $\kappa_i(T)$ is known. In the parameter certain case, we assume all parameter values are static. The only dynamic quantities are $\kappa_1(T)$ and $\kappa_2(T)$. In the parameter partial certain case, the second level of estimated parameters are reestimated, that is, we refit the random walk and time series models using the historical and newly simulated data values. Lastly, the parameter uncertain case involves refitting of the entire model at time T using combined historical and simulated data.

Next, we use the following algorithm to fit the surrogate models.

1. Simulate the model to time T, and extend the historical data so that the simulated data is the new "history" from time 0 to time T.

2. Refit the model depending on the choice of parameter certainty.

3. Repeat steps 2 and 3 $N_1 - 1$ times to obtain N_1 realizations of $Y(T)$. This suggests the values that $Y(T)$ can take. (This step is optional if one already has a range of values in mind.)

4. According to the range of values of $Y(T)$, create a grid of values. This is the training input set $\mathbf{Y}^{(tr)}(T) = (\tilde{Y}_1(T), \ldots, \tilde{Y}_{N_{tr,1}}(T)$.

5. For each $\tilde{Y}_i(T) \in \tilde{\mathbf{Y}}^{(tr)}(T)$, perform $N_{tr,2}$ Monte Carlo simulations given $\tilde{Y}(T) = \tilde{Y}_i(T)$ to determine $\tilde{\Delta}_i(T) := \tilde{a}_1^{MC}(\tilde{Y}_i(T), T, x) - \tilde{a}_2^{MC}(\tilde{Y}_i(T), T, x)$, the hedge portfolio value for one unit of index annuity less the premium, where $\hat{a}^{MC}(Y(T), T, x)$ is the Monte Carlo simulated annuity value defined in 16. In addition, we record $\tilde{\sigma}_i^{(MC)}(T)$ as the standard deviation of the Monte Carlo simulations given $\tilde{Y}_i(T)$, which is used in fitting the kriging models. This gives the training set $(\tilde{Y}_1(T), \tilde{\Delta}_1(T), \tilde{\sigma}_1^{(MC)}(T)), \ldots, (\tilde{Y}_{N_{tr,1}}(T), \tilde{\Delta}_{(N_{tr,1})}(T), \tilde{\sigma}_{N_{tr,1}}^{(MC)}(T))$.

6. Fit the spline and kriging models based on the training set. The resulting values predicted by these models are denoted by superscript $(spl), (krg0), (krg1)$, and $(krg2)$ for spline, 0th-order kriging, 1st-order kriging, and 2nd-order kriging respectively.

Table 2: $\Delta^{(\cdot)} = a_1^{(\cdot)}(T) - a_2^{(\cdot)}(T)$. N_{tr_1} and N_{tr_2} are the number of training set elements and Monte Carlo simulations per training set element respectively, N is the number of testing set elements, and N_2 is the number Monte Carlo simulations per testing set element. MM1 uses (11) and (12) for m_1 and m_2 respectively, while MM2 uses (11) and (13).

Symbol	Type	# Simulations	Model Dependent
$\Delta^{(MC)}(T)$	Monte Carlo	$N_2 \times N$	N
$\Delta^{(MM1)}(T)$	Deterministic	0	Y
$\Delta^{(MM2)}(T)$	Deterministic	0	Y
$\Delta^{(spl)}(T)$	Thin Plate Spline	$N_{tr,1} \times N_{tr,2}$	N
$\Delta^{(krg0)}(T)$	0th-order Kriging	$N_{tr,1} \times N_{tr,2}$	N
$\Delta^{(krg1)}(T)$	1st-order Kriging	$N_{tr,1} \times N_{tr,2}$	N
$\Delta^{(krg2)}(T)$	2nd-order Kriging	$N_{tr,1} \times N_{tr,2}$	N

With the surrogate models available, we evaluate the different emulators and deterministic approximations on an out-of-sample test set consisting of N starting values for $Y(T)$. To benchmark, we run the classical nested Monte Carlo method, generating N_2 trajectories for each starting point and recording the resulting Δ^{MC} and σ^{MC}.

This procedure shows that the surrogate models require $N_{tr,1} \times N_{tr,2}$ (to fit) simulations (independent of N), while Monte Carlo require $N_2 \times N$ simulations. This is summarized in table 2.

4.2.1 Model Fitting

Several stochastic mortality models (see 2.3 or Cairns *et al.* (2011b) for details on models) have `R` code available[5] to fit to data sets; We use the code to fit the two-population model parameters which yields the age, period, and cohort effects. Separately the estimated period and cohort effects are modeled using random walk and time series techniques. In the case of parameter uncertainty, each training and testing set element requires refitting, which can be computationally expensive.

The method of determining the training set lattice depends on the problem at hand. In our particular example we aim to give an accurate result of the expectation of the hedge portfolio, so our lattice is equally spread throughout the values that $Y(T) = (\kappa_1(T), \kappa_2(T))$ can take (in the parameter certain case), with the end points determined through an initial simulation to estimate the joint distribution of $(\kappa_1(T), \kappa_2(T))$. In a case study where VaR for example was the quantity of interest, the lattice could be skewed toward having more extreme values that produce lower $\Delta(T)$ values, therefore improving accuracy of VaR estimations.

In this particular example we use the two-population model, but we note that the simulations up to time T and simulation on $[T, \infty)$ are completely separate, so different models could in principle be used.

[5]LifeMetrics Open Source R code for Stochastic Mortality Modelling; see http://www.macs.hw.ac.uk/ andrewc/lifemetrics/ for details

Table 3: Summary statistics in the parameter certain case for relative hedge values for $N = 1000$ randomly generated testing set values, with $N_2 = 100$ for the Monte Carlo benchmark. The lattice for the training set was generated uniformly within a plausible range of $\kappa_1(T), \kappa_2(T)$; $N_{tr,1} = 300, N_{tr,2} = 100$.

Type	μ	$VaR_{0.005}$
$\Delta^{(MC)}(T) - \Delta^{(MM1)}(T)$	4.337E-2	0.146719
$\Delta^{(MC)}(T) - \Delta^{(MM2)}(T)$	4.551E-3	-0.005879
$\Delta^{(MC)}(T) - \Delta^{(frz)}(T)$	2.9024E-3	-0.099969
$\Delta^{(MC)}(T) - \Delta^{(spl)}(T)$	-2.830E-4	-0.003831
$\Delta^{(MC)}(T) - \Delta^{(krg0)}(T)$	5.180E-4	-0.000558
$\Delta^{(MC)}(T) - \Delta^{(krg1)}(T)$	2.280E-3	-0.010764
$\Delta^{(MC)}(T) - \Delta^{(krg2)}(T)$	4.991E-3	-0.016599

4.3 Results

The goal of our results is to examine the effectiveness of statistical emulators in a real world setting, in this case the index-based hedge. Several estimators that have been discussed depend on the values of simulated values up to time T. For the two-population model (see 2.3.1), this involves understanding the behavior of $\kappa_1(t)$ and $\kappa_2(t)$, particularly at time T. We therefore provide an analysis of these processes including a plot of historical and simulated results.

For now we assume $\delta = 1$ so the value of the hedge portfolio will be $\Delta(T) := a_1(T) - a_2(T)$. We then follow the framework described in the beginning of the section, with $N_{tr,1} = 300$ lattice points and $N_{tr,2} = 100$ Monte Carlo simulations to determine the training annuity values. We then generate a random testing set of $N = 1000$ values of $Y(T)$, and for each $Y(T)$ perform 100 Monte Carlo simulations for a benchmark. For each approach, the resulting average and $VaR_{0.005}$ values are reported over the testing set, relative to the Monte Carlo benchmark. This is summarized in table 3. Box plots for each approach are shown in figure 1.

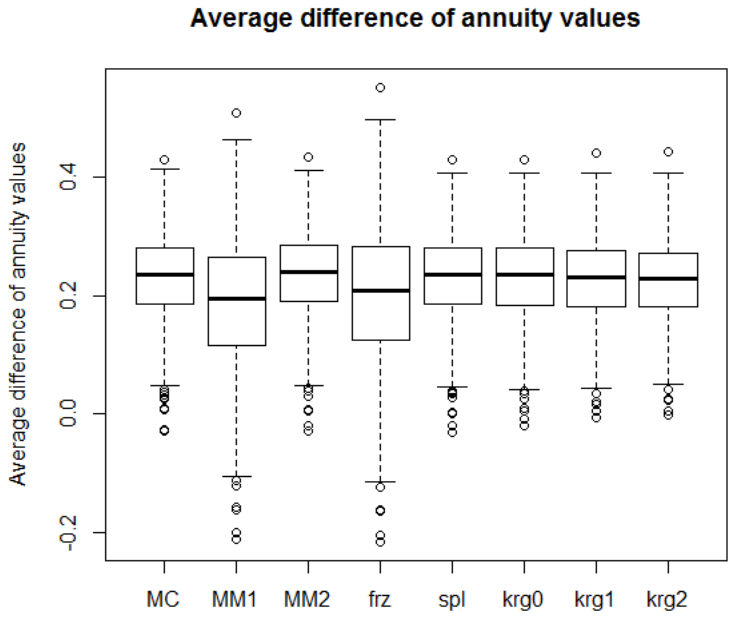
References

Barrieu, Pauline, Bensusan, Harry, El Karoui, Nicole, Hillairet, Caroline, Loisel, Stéphane, Ravanelli, Claudia, & Salhi, Yahia. 2012. Understanding, modelling and managing longevity risk: key issues and main challenges. *Scandinavian actuarial journal*, **2012**(3), 203–231.

Bauer, Daniel, Reuss, Andreas, & Singer, Daniela. 2012. On the calculation of the solvency capital requirement based on nested simulations. *Astin Bulletin*, **42**(02), 453–499.

Booth, Heather, Maindonald, John, & Smith, Len. 2002. Applying Lee–Carter under conditions of variable mortality decline. *Population studies*, **56**(3), 325–336.

Figure 1: Box plots in the parameter certain case for hedge values for $N = 1000$ randomly generated testing set values, with $N_2 = 100$ for the Monte Carlo benchmark. The lattice for the training set was generated uniformly within a plausible range of $\kappa_1(T), \kappa_2(T); N_{tr,1} = 300, N_{tr,2} = 100$.



Brouhns, Natacha, Denuit, Michel, & Vermunt, Jeroen K. 2002. A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, **31**(3), 373–393.

Cairns, Andrew JG, Blake, David, & Dowd, Kevin. 2006. A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration. *Journal of Risk and Insurance*, **73**(4), 687–718.

Cairns, Andrew JG, Blake, David, Dowd, Kevin, Coughlan, Guy D, Epstein, David, Ong, Alen, & Balevich, Igor. 2009. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, **13**(1), 1–35.

Cairns, Andrew JG, Blake, David, Dowd, Kevin, Coughlan, Guy D, & Khalaf-Allah, Marwa. 2011a. Bayesian stochastic mortality modelling for two populations. *Astin Bulletin*, **41**(01), 29–59.

Cairns, Andrew JG, Blake, David, Dowd, Kevin, Coughlan, Guy D, Epstein, David, & Khalaf-Allah, Marwa. 2011b. Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, **48**(3), 355–367.

Cairns, Andrew JG, Dowd, Kevin, Blake, David, & Coughlan, Guy D. 2014. Longevity hedge effectiveness: A decomposition. *Quantitative Finance*, **14**(2), 217–235.

Chen, Hua, & Cox, Samuel H. 2009. Modeling mortality with jumps: Applications to mortality securitization. *Journal of Risk and Insurance*, **76**(3), 727–751.

Coughlan, Guy D, Khalaf-Allah, Marwa, Ye, Yijing, Kumar, Sumit, Cairns, Andrew JG, Blake, David, & Dowd, Kevin. 2011. Longevity hedging 101: A framework for longevity basis risk analysis and hedge effectiveness. *North American Actuarial Journal*, **15**(2), 150–176.

Currie, Iain D, Durban, Maria, & Eilers, Paul HC. 2004. Smoothing and forecasting mortality rates. *Statistical modelling*, **4**(4), 279–298.

Czado, Claudia, Delwarde, Antoine, & Denuit, Michel. 2005. Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics*, **36**(3), 260–284.

Debonneuil, Edouard. 2010. A simple model of mortality trends aiming at universality: Lee Carter + Cohort. *arXiv preprint arXiv:1003.1802*.

Delwarde, Antoine, Denuit, Michel, & Eilers, Paul. 2007. Smoothing the Lee–Carter and Poisson log-bilinear models for mortality forecasting A penalized log-likelihood approach. *Statistical Modelling*, **7**(1), 29–48.

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, Hastie, T, Friedman, J, & Tibshirani, R. 2009. *The elements of statistical learning*. Vol. 2. Springer.

Hyndman, Rob J, & Ullah, Md Shahid. 2007. Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, **51**(10), 4942–4956.

Jalen, Luka, & Mamon, Rogemar. 2009. Valuation of contingent claims with mortality and interest rate risks. *Mathematical and Computer Modelling*, **49**(9), 1893–1904.

Lee, Ronald, & Miller, Timothy. 2001. Evaluating the performance of the Lee–Carter method for forecasting mortality. *Demography*, **38**(4), 537–549.

Lee, Ronald D, & Carter, Lawrence R. 1992. Modeling and forecasting US mortality. *Journal of the American statistical association*, **87**(419), 659–671.

Li, J.S.-H., Hardy M.R., & Tan, K.S. 2009. Uncertainty in model forecasting: An extension to the classic Lee-Carter approach. *ASTIN Bulletin*, **39**, 137–164.

Omre, Henning. 1987. Bayesian krigingmerging observations and qualified guesses in kriging. *Mathematical Geology*, **19**(1), 25–39.

Plat, Richard. 2009. On stochastic mortality modeling. *Insurance: Mathematics and Economics*, **45**(3), 393–404.

Rasmussen, Carl Edward, & Williams, CKI. 2006. Gaussian processes for machine learning. 2006. *Cited on*, 95.

Renshaw, Arthur E, & Haberman, Steven. 2006. A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, **38**(3), 556–570.

Roustant, Olivier, Ginsbourger, David, & Deville, Yves. 2012. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization.

A Lee Carter & CBD Stochastic Mortality Models

Cairns *et al.* (2011b) performs an analysis on several mortality models. We follow their notation in this paper. In 1992 Lee & Carter (1992) introduced a model for stochastic mortality rates. Since then, most stochastic mortality modeling has been done using either the Lee-Carter model itself or an extension of it. In this paper we use a two population version of an extension introduced by Renshaw & Haberman (2006), which postulates

$$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_{t-x}^{(3)}. \tag{M2}$$

One can interpret $\beta_x^{(1)}$, $\kappa_t^{(2)}$ and $\gamma^{(3)}$ as the age, period and cohort effects respectively. The original model proposed by Lee & Carter (1992) is a special case where $\gamma^{(3)} = 0$. The age effects are non-parametric and estimated from historical data, while the period and cohort effects assume the form of a stochastic process. Lee & Carter (1992) originally suggests a random-walk process for the period effect. This choice for $\kappa_t^{(2)}$ has continued to be used in future analysis of the model. More precisely, we have

$$\kappa_t^{(i)} = \kappa_{t-1}^{(i)} + \mu_\kappa^{(i)} + \sigma_\kappa^{(i)} Z_\kappa^{(i)}(t),$$

where $\mu_\kappa^{(i)}$ is the drift, $\sigma_\kappa^{(i)}$ is the volatility, and $Z_\kappa^{(i)}$ is the standard normal driver of the random walk. In the original Lee Carter case they consider only one population and random walk model, but extensions can be (and have been) made to model multiple populations, and in this case one may have the $Z_\kappa^{(i)}(t)$ to be correlated.

On the other hand, Cairns *et al.* (2011b) goes on to say that more general $ARIMA$ models might provide a better fit depending on the data set, since in 2007 the CMI uses an ARIMA(1,1,0) process for $\kappa_t^{(2)}$ in their analysis of the Lee-Carter model.

Renshaw & Haberman (2006) suggests using ARIMA models for $\gamma_{t-x}^{(3)}$, and it is suggested in Cairns *et al.* (2011b) to use either $ARIMA(0,2,1$ or $ARIMA(1,1,0)$. Renshaw & Haberman (2006) and Cairns *et al.* (2011b) both assume $\gamma_{t-x}^{(i)}$ is independent of $\kappa_t^{(2)}$.

This model has identifiability issues, and one set of constraints could be

$$\sum_t \kappa_t^{(2)} = 0, \quad \sum_x \beta_x^{(2)} = 0, \quad \sum_{x,t} \gamma_{t-x}^{(3)} = 0, \quad \text{and} \quad \sum_x \beta_x^{(3)} = 1.$$

Alternatively, Currie *et al.* (2004) suggested use of B-splines and P-splines to create a deterministic but nonparametric forecast for future mortality

$$\log m(t,x) = \sum_{i,j} \theta_{ij} B_{ij}^{ay}(x,t), \tag{M4}$$

with smoothing of the θ_{ij} in the age and cohort directions.

For completeness we mention the model proposed by Cairns *et al.* (2006) (CBD) with different dynamics than the Lee Carter model. They fit

$$\text{logit} q(t,x) = \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(2)}\kappa_t^{(2)},$$

where $\text{logit}(t) = \log\left(\frac{t}{1-t}\right)$ and $q(t,x)$ is the probability of death within t years for someone aged x, which is analogous to the term $1 - P(Y(u); 0, t, x)$ to compare with our notation.

If we let n_a be the number of ages available in the data set for fitting, and $\bar{x} = n_a^{-1} \sum_i x_i$, they assume the forms

$$\beta_x^{(1)} = 1, \quad \text{and } \beta_x^{(2)} = (x - \bar{x}),$$

so that

$$\text{logit} q(t, x) = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)}. \tag{M5}$$

With these assumptions there are no identifiability issues. The CBD model has several plausable extensions, including ones with cohort effects, which are all discussed in more detail by Cairns *et al.* (2011b)