Gaussian Process Models for Mortality Rates & Improvement Factors

Insurance, Actuarial Science, Data and Models: Paris June 11, 2018

Mike Ludkovski

Dept of Statistics & Applied Probability UC Santa Barbara

STATISTICS & APPLIED PROBABILITY CENTER FOR FINANCIAL MATHEMATICS & ACTUARIAL RESEARCH

Joint with J. Risk (Cal Poly Pomona) and H. Zail (Elucidor)

Mortality modeling

Inputs x^n		Log Mortality Rate y^n		Mortality Rate $\exp(y^n)$	
Age (x^n_{ag})	Year (x^n_{yr})	Male	Female	Male	Female
50	2011	-4.931	-5.437	0.00722	0.00435
64	2011	-4.264	-4.707	0.01406	0.00901
74	2011	-3.435	-3.821	0.03222	0.02191
84	2011	-2.408	-2.714	0.08999	0.06625

Excerpt of CDC mortality data to compare exposures and mortality rates over Ages and gender for calendar year 2011. *Mortality* is the observed proportion D^n/L^n of the deceased during the Year relative to the mid-year population.

Objectives:

- Smooth observed mortality experience
- Forecast unobserved mortality (future years, additional ages)
- Quantify uncertainty (data-driven vs model-driven)
- Interpret mortality evolution, especially mortality **improvement factors**

Main Message

- A *statistical* framework based on Gaussian Process regression
- Nonparametric method that treats Age/Calendar Year equally
- Bayesian paradigm to quantify model/predictive uncertainty
- A spatial/"machine learning" alternative to traditional APC/Lee-Carter frameworks
- Coherently handles smoothing and extrapolation
- A burgeoning ecosystem amenable to many extensions

Mortality Dataset

- 2-D table indexed by Age and Year: $x = (x^n_{ag}, x^n_{yr})$
- Raw deceased count D^n; Exposures E^n; Mid-Year Lives L^n
- Noisily observed in raw log-rates $Y(x^n) \equiv \mu^n = \log \frac{D^n}{L^n}$
- $Y(x) = f(x) + \varepsilon$ where $Var(\varepsilon(x)) = \sigma^2(x)$ (additive noise) and $f(\cdot)$ is the latent log-mortality surface
- Illustrate with US CDC dataset: years 1999-2014, ages 50-84
- Also UK & Japan using HMD datasets
- (Ignore old data or Young ages since focus on annuity/pension plan projections)

Statistical Framework

- Treat the true mortality surface f as a random function
- Specify prior distribution and then compute conditional distribution given the data $p(f|\mathcal{D}) \propto p(\boldsymbol{y}|f, \boldsymbol{x})p(f) = \{likelihood\} \cdot \{prior\}$
- Covariance structure: knowing mortality at x will greatly influence mortality at “neighboring” x’s: the mortality rate for a 60 year old in 2015 will be closer to that of a 61 year old in 2016, than that of a 20 year old in 1990
- Output is a posterior distribution: provides both the point estimate and credible bands
- Gaussian prior + Gaussian likelihood $\Rightarrow$ Gaussian posterior

Gaussian Processes

- Gaussian random field w/prior $f \sim GP(\boldsymbol{m}(\mathbf{x}), \boldsymbol{C}(\mathbf{x}, \mathbf{x}))$
- Mean function $m(x^i) = \mathbb{E}[f(x^i)]$
- Covariance $C(x^i, x^j) = \mathbb{E}[\left(f(x^i) - m(x^i)\right) \left(f(x^j) - m(x^j)\right)]$ controls the spatial smoothness
- Squared-Exponential kernel

$$C(x, x') = \eta^2 \exp\left(-\frac{(x_{ag} - x'_{ag})^2}{2\theta_{ag}^2} - \frac{(x_{yr} - x'_{yr})^2}{2\theta_{yr}^2}\right)$$

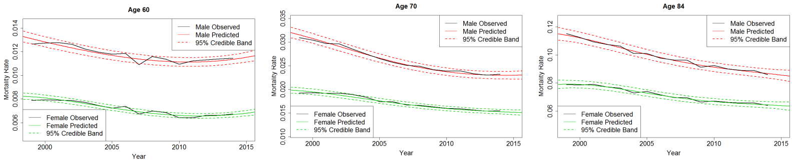
- Lengthscales θ's and amplitude/process variance η.
- Observation likelihood $p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \Sigma)$ (Gaussian conjugate!) w/$\boldsymbol{\Sigma} = diag(\sigma^2(x^i))$

GP Modeling

- The posterior is Gaussian $f(x)|\boldsymbol{x}, \boldsymbol{y} \sim \mathcal{N}(m_*(x), s_*^2(x))$

$$m_*(x) = \vec{c}(x)^T(\mathbf{C} + \Sigma)^{-1}\vec{y}$$
$$s_*(x, x') = K(x, x') - \vec{c}(x)^T(\mathbf{C} + \Sigma)^{-1}\vec{c}(x')$$

- $C_{ij} = C(x^i, x^j)$, $c_i = C(x, x^i)$



Mortality rates for Males (top curves) and Females (bottom curves) aged 60, 70 and 84 over time. The plots show **raw mortality** rates for years 1999-2014, as well as predicted mean of the smoothed mortality surface and its 95% credible band. Models are fit to 1999–2014 CDC data for Ages 50–84 (All data).

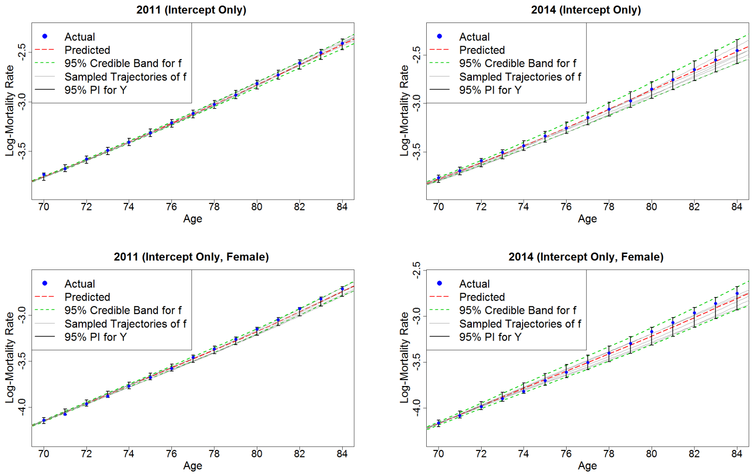
Working with GP Predictive Equations

- The predictive equations apply to any **vector** $\boldsymbol{x}_*$
- Consistently provide a point estimate m_* and related uncertainty quantification s_* which gives a credible band $(m_* - z_\alpha s_*, m_* + z_\alpha s_*)$

⇒ Estimate the historic smoothed mortality curves by calendar year $(\boldsymbol{m}_*(x^{1:N})$;

- Project the curves forward: $\boldsymbol{m}_*(\mathbf{x}_*)$ for future inputs $\mathbf{x}_*$;
- Generate stochastic scenarios (sample from the random vector $\mathbf{f}_*(\mathbf{x}_*)$)
- Impute missing data (e.g. at the edge of the dataset)
- Could also forecast future mortality observations $y_* = f_* + \varepsilon(x_*)$

Mortality Credible Bands



Mortality rate prediction for years 2011 and 2014 and ages 71–84. Model is fit with Subset II data with intercept-only mean functions and squared-exponential kernel. “Simulated paths of *f*” refers to simulated trajectories of the latent $\mathbf{f}_*$. Credible bands are for the mortality surface $\mathbf{f}_*$; vertical intervals are for predicted observable mortality experience $\mathbf{y}_*$.

Lee-Carter Models

GP is a very different paradigm compared to Lee-Carter:

- LC postulates that $\mu_{ag,yr} = \alpha_{ag} + \beta_{ag}\kappa_{yr} + \varepsilon_{ag,yr}$
- Two-step: first fit the factors α, β, κ using *the entire* dataset
- Then build a time-series model for time-dependent κ to make forecasts (eg AR(1))
- Postulate a priori the structural dependence; no direct link between calibration and forecasting
- Smoothing is via a point estimator, no credible bands

Fitting a GP

- Must specify the kernel family (other popular choices include Matern-5/2, et cetera)
- Need to learn the kernel hyperparameters $\Theta - \eta, \theta$'s, et cetera.
- Use MLE
    - Care is needed for doing the nonlinear optimization
    - We used `DiceKriging` package in `R`
- Alternatively: hierarchical approach through specifying priors on Θ
    - Posterior is a mixture of Gaussians
    - Sample from the posterior using MCMC
    - Hamiltonian MCMC implemented in `Stan`

Estimated Mortality Covariance Structure

	DiceKriging	Stan		
	MLE	MAP	MCMC Mean	MCMC 80% Posterior CI
θ_{ag}	15.8384	14.7988	11.0401	(6.3369, 17.0395)
θ_{yr}	15.5308	15.7910	25.8306	(14.6287, 39.4763)
η^2	1.8468	1.2365	1.9920	(0.8744, 3.3930)
σ^2	2.808e-04	2.753e-04	2.760e-04	(2.536e-04, 2.998e-04)
β_0	-3.8710	-3.8003	-3.8302	(-4.7305, -2.9350)

Hyperparameter estimates based on maximum likelihood (`DiceKriging`) and maximum a posteriori probability (Stan), along with MCMC summary statistics. The GP is fitted to all data and uses squared-exponential covariance kernel (**??**) with prior mean $m(x) = \beta_0$. Stan hyper-priors (on standardized data) were $\log \theta_{ag}, \log \theta_{yr}, \log \eta^2 \sim \mathcal{N}(0, 1)$ i.i.d., $\sigma^2 \sim \mathcal{N}_+(0, 0.2)$, $\beta_0 \sim \mathcal{N}(-4, 5)$.

GP Observation Noise

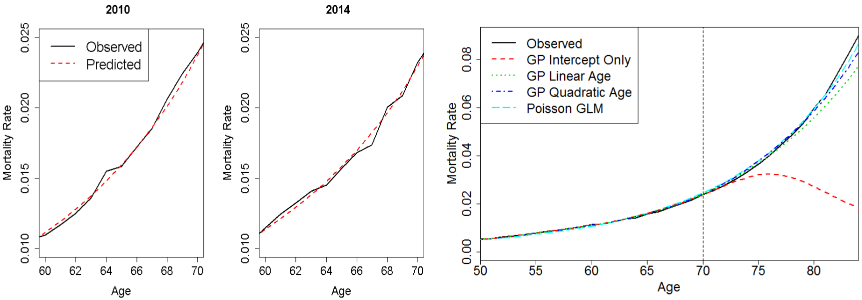
- Instead of additive noise it's more correct to use $D^n \sim Bin(L^n, e^{f^n})$ or $D^n \sim Poisson(L^n e^{f^n})$
- Replace the Gaussian likelihood with a Binomial or Poisson (a la GLM), then continue as before
- Posterior is no longer Gaussian (so either use MCMC or Laplace Approximation)
- Empirically there is overdispersion (partly because e^{f^n} is unknown)
- Can also directly assume heteroskedasticity: $\varepsilon(x^n)$ has state-dependent variance (either based on Binomial or latent)
- Tried both of above and saw minimal effect on results/goodness-of-fit
- Empirically, recommend sticking with a constant "nugget" σ that is estimated by MLE
- Might need to be refined depending on context

Mean Function

- The shape of f is a blend of the prior and the influence of the data
- At edges/beyond the dataset, f is driven by the prior
- Important to correctly specify the mean function $m(x)$: captures the Age-shape of mortality and long-term Year trend
- In age, log-mortality is increasing (super-linearly): $m(x) = \beta_0 + \beta_1^{ag} x_{ag} + \beta_1^{yr} x_{yr}$ or $m(x) = \beta_0 + \beta_1^{ag} x_{ag} + \beta_1^{yr} x_{yr} + \beta_2^{ag} x_{ag}^2$

	Mean Function Parameter MLE's				GP Hyperparameter MLE's			
	β_0	β_1^{ag}	β_2^{ag}	β_1^{yr}	η^2	σ^2	θ_{ag}	θ_{yr}
Intercept	-4.526	-	-	-	6.213e-01	3.428e-04	8.384	12.746
Linear	18.737	0.081	-	-1.397e-02	8.521e-04	1.761e-04	3.610	3.543
Quadratic	19.641	0.064	1.459e-04	-1.417e-02	1.403e-03	2.998e-04	3.629	3.475

Mortality Smoothing

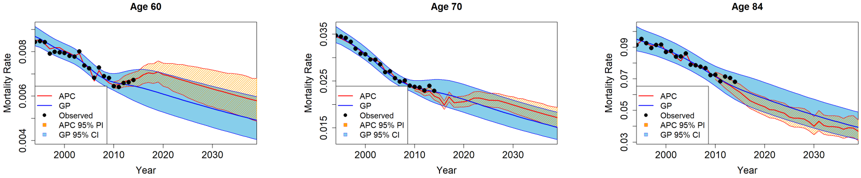


Mortality rates for US Males aged 60-70 during the years 2010-2014. Raw vs. estimated smoothed mortality curves. Models are fit to 1999–2014 CDC data for Ages 50–84 (All data).

- Results are stable across different implementations of GPs
- In-sample prediction is data-driven, priors/covariance structure/observation noise plays secondary role

Long Range Forecasts

Beyond the dataset revert to the prior $m_*(x) \to m(x)$: transition controlled by θ. We find that can forecast $\sim 5 - 7$ years into the future

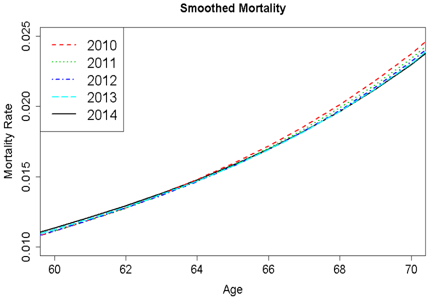


US Females US Males UK Females

Observed and predicted mortality rates for 1994-2040 for three representative datasets. GP model uses quadratic mean function $m(x) = \beta_0 + \beta_1^{ag} x_{ag} + \beta_1^{yr} x_{yr} + \beta_2^{ag} x_{ag}^2$, and the APC model is $\mu_{ag,yr} = \alpha_{ag} + \frac{1}{na}\kappa_{yr} + \frac{1}{na}\gamma_{yr-ag} + \varepsilon_{ag,yr}$. Models fit to HMD data for 1994–2009 and ages 50–84.

Forecasting

- The GP model is not Markovian in time: directly sample full trajectories for future mortality evolution (not just marginally, but the full fitted dependence in Ages or Years)
- Maintain Gaussian structure which allows explicit quantification of risk; very useful for risk management (eg assessing VaR of annuity portfolio)
- The GP model is updateable: directly shows the impact of new/additional data on the projections
- Salhi et al (2017) present a different formulation that has Markov structure in t

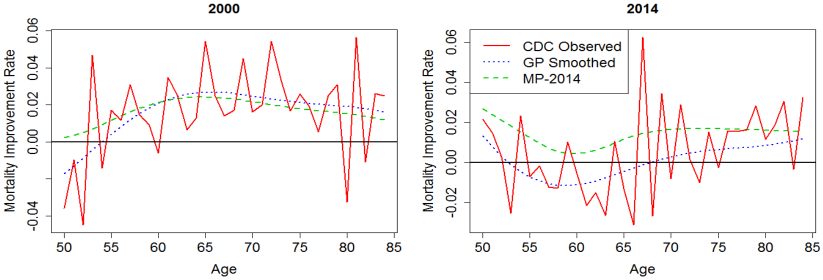


Discussion

- Choose to model log-mortality, but can also model other quantities
- May be desirable to impose constraints:
  - monotonicity in Age
  - long-range trends
  - hierarchical relationships between multiple populations

Mortality Improvement

- MI: change in f as a function of x_{yr}
- Raw YoY changes are **extremely** noisy; smoothing is essential

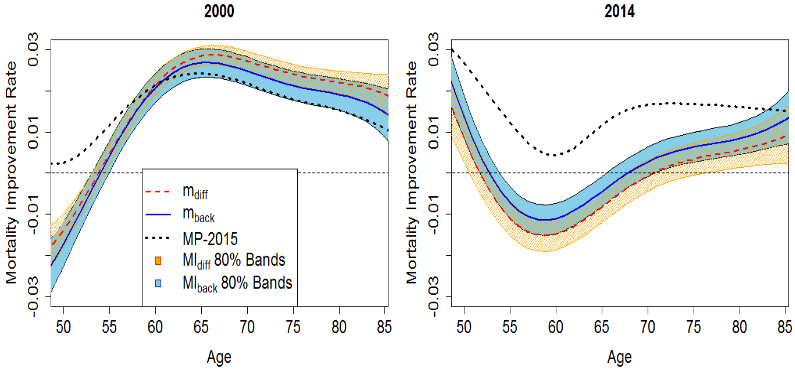


Mortality improvement factors for US Males using All Data. Solid lines indicate the empirical mortality experience $MI^{obs}_{back}(\cdot; yr)$ for years $yr \in \{2000, 2014\}$, the dotted and dashed lines are $\partial m^{GP}_{back}(\cdot; yr)$, and the MP-2015 improvement scale $MI^{MP}_{back}(\cdot; yr)$.

Improvement Factors

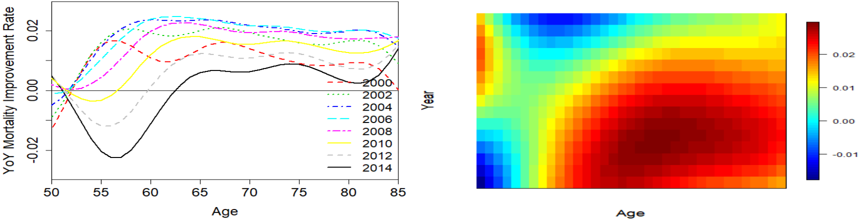
- Traditional is backward-looking: $f(ag, yr) - f(ag, yr-1) =: MI_{back}$
- Instantaneous: $MI_{diff} := \partial_{yr} f(ag, yr)$
- Can analytically differentiate the latent log-mortality surface; $\partial_{x_{yr}} f$ is again a GP!
- Gradient is $\partial m_{diff}(x_*) = \mathbb{E}\left[\frac{\partial f_*}{\partial x_{yr}}(x_*) \middle| \mathbf{x}, \mathbf{y} \right] = \frac{\partial C}{\partial x'_{yr}}(\mathbf{x}, x_*)(\boldsymbol{C} + \boldsymbol{\Sigma})^{-1}\mathbf{y}$, similar formula for $s^2_{diff}(x_*)$
- Can quantify uncertainty on improvement factors – not possible in other approaches

Improvement Factors



Estimated male mortality improvement using the differential GP model (instantaneous improvement) and the YoY improvement from the original GP model. We show the means and 80% uncertainty bands for MI^{GP}_{diff} and MI^{GP}_{back} for males aged 50–84 and years 2000 & 2014. Models used are fit to All data.

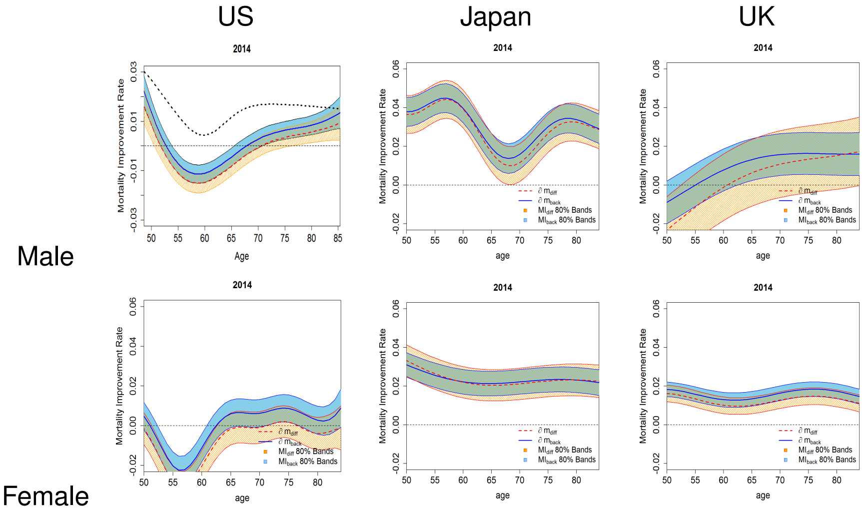
Declining Longevity in US



Smoothed yearly mortality improvement factors $\partial m^{GP}_{back}(ag, yr)$ for US Males using All data. Left: age factors for $yr = 2000, \ldots, 2014$. Right: Heatmap of the estimated YoY improvement factors.

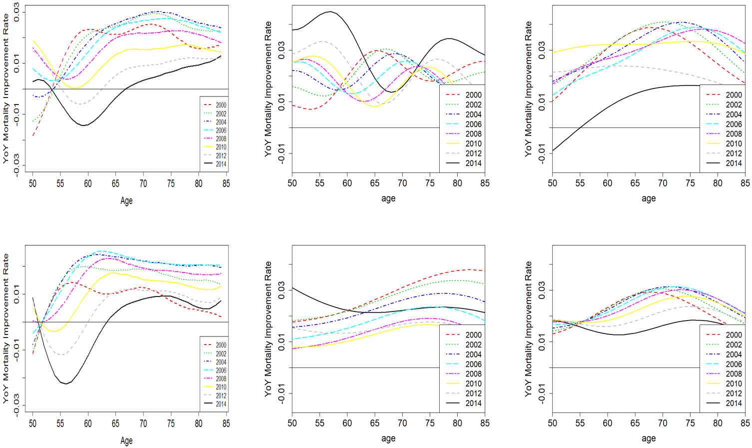
- Results strongly suggest that US longevity is now decreasing at Ages 55-70
- Diverges from the official MP-2016 scales that continue to assume improvement
- Improvement is highly age-dependent

Comparing Mortality Improvements



Comparing Mortality Improvements

US Japan UK



Male

Female

Data Relevance

- For actuaries, central issue is longevity risk: middle/older ages, today/future
- Hence, we build our models for Age 50+/1999+
- The mortality surface is non-stationary, so more distant data is detrimental (?)
- Our analysis suggests that looking at more than 20 years of data is irrelevant
- Perhaps need further segmenting over ages but seek balance over credibility/over-fitting/non-stationarity

Next Steps

- Sub-populations
- Multiple populations
- Modeling by cause of death

Next Steps

- Sub-populations
- Multiple populations
- Modeling by cause of death

THANK YOU!

References

Williams, C. K. and Rasmussen, C. E. 2006.
Gaussian processes for machine learning, the MIT Press.

Girosi, F. and King, G. 2008.
Demographic forecasting, Princeton University Press.

Currie, I. D. 2016.
On fitting generalized linear and non-linear models of mortality.
Scandinavian Actuarial Journal, 2016(4):356–383.

Roustant, O., Ginsbourger, D., Deville, Y., et al. 2012.
Dicekriging, Diceoptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization.
Journal of Statistical Software, 51(1):1–55.

M. Ludkovski, J. Risk, H. Zail
Gaussian Process Models for Mortality Rates and Improvement Factors
ASTIN Bulletin to Appear
R notebook that allows full reproduction of all the figures: github.com/jimmyrisk/GPmortalityNotebook