Bayesian Forecasting of Cohort Fertility

Understanding of Methods

Yichen He[1]

[1]Department of Statistics and Data Science
National University of Singapore

All methods and charts come from Schmertmann et.al[1]

Table of Contents

1. Notation
2. General Idea
3. Shape Penalty
4. Time Penalty
5. Weighting
6. Final result
7. Proof
8. Reference

Table of Contents

1. Notation
2. General Idea
3. Shape Penalty
4. Time Penalty
5. Weighting
6. Final result
7. Proof
8. Reference

Notation

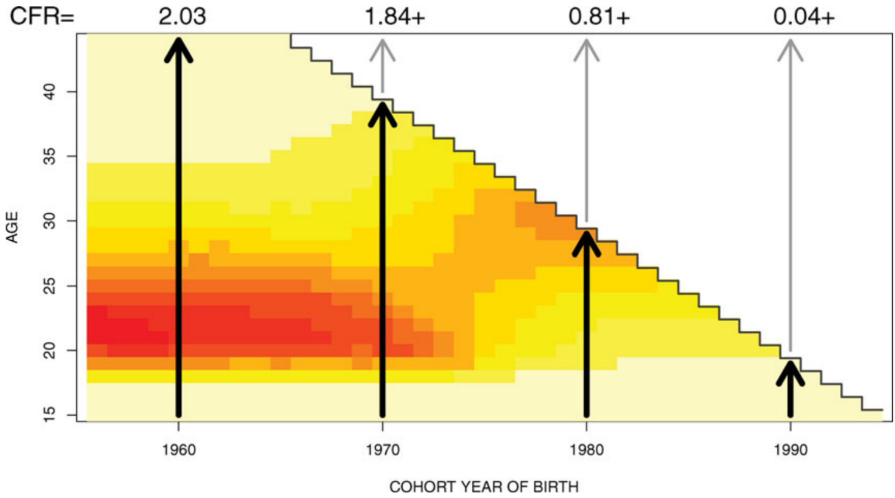
For a specific country, we have C birth cohorts of interest ($c = 1 \dots C$) over A reproductive ages ($a = 1 \dots A$).

- $\theta_{ca} \in \mathbb{R}$, the true fertility rate for cohort c between exact ages a and $a + 1$;
- $\theta_c = (\theta_{c1} \dots \theta_{cA})^\top \in \mathbb{R}^A$, the fertility schedule for cohort c;
- $\theta_a = (\theta_{1a} \dots \theta_{Ca})^\top \in \mathbb{R}^C$, the time series of rates at age a;
- $\theta = \left(\theta_1^\top \dots \theta_C^\top\right)^\top \in \mathbb{R}^{CA}$, **the vector of all rates, sorted by age within cohort;**

Table of Contents

1. Notation
2. General Idea
3. Shape Penalty
4. Time Penalty
5. Weighting
6. Final result
7. Proof
8. Reference

General Idea



General Idea

Notation

1. $y \in \mathbb{R}^n$, a vector of published data for some subset of θ;
2. $\mathrm{V} \in \mathbb{R}^{n \times CA}$, a matrix of ones and zeroes such that $\mathrm{V}\theta \in \mathbb{R}^n$ is the subset of parameters corresponding to y.

Assume our historical data y generated from the normal distribution:

$$y \mid \theta \sim N_n (V\theta, \Psi)$$

where $\Psi = \mathrm{diag}_{i=1\ldots n} \left[y_i (1 - y_i) / W_i\right]$ and W_i is the number of a-year-old women in the (c, a) cell corresponding to the i-th rate.
Hence, we can based on the distribution, give out the log-liklihood function:

$$\ln L(y \mid \theta) = \text{ const } - \frac{1}{2}(y - \mathrm{V}\theta)^\top \Psi^{-1}(y - \mathrm{V}\theta)$$

General Idea

Notation

1. $y \in \mathbb{R}^n$, a vector of published data for some subset of θ;
2. $\mathrm{V} \in \mathbb{R}^{n \times CA}$, a matrix of ones and zeroes such that $\mathrm{V}\theta \in \mathbb{R}^n$ is the subset of parameters corresponding to y.

According to the Lexis surface, we know that θ is not complete, so we still need a priori information for our θ:

$$\theta \sim N_{CA}\left(\underline{0}, K^{-1}\right)$$

Therefore if we combine the prior information with log-liklihood together, we can get the expression for posterior for our θ in a Bayesian framework:

$$\begin{aligned} \ln P(\theta \mid y) &= \text{const} + \ln L(y \mid \theta) + \ln f(\theta) && (1) \\ &= \text{const} - \frac{1}{2}(y - \mathrm{V}\theta)^\top \Psi^{-1}(y - \mathrm{V}\theta) - \frac{1}{2}\theta^\top \mathrm{K}\theta && (2) \end{aligned}$$

General Idea

Recall

Posterior: $\ln P(\theta \mid y) = \text{const} - \frac{1}{2}(y - \mathrm{V}\theta)^\top \Psi^{-1}(y - \mathrm{V}\theta) - \frac{1}{2}\theta^\top \mathrm{K}\theta$

As we can tell from the above formula, the most important thing is the covariance matrix K. This part is aggregated by two different penalty terms, **Shape penalty** and **Time penalty**. We will explain these two specially-designed terms in the next two sections in detail.

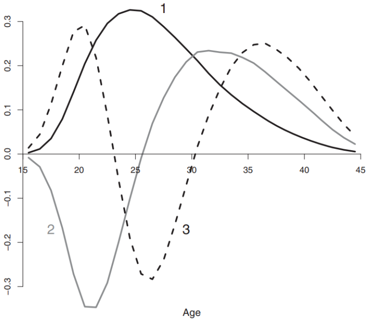
Table of Contents

1. Notation
2. General Idea
3. Shape Penalty
4. Time Penalty
5. Weighting
6. Final result
7. Proof
8. Reference

SVD

In the real case, we can set the number of ages and cohorts of interest as $A = 30$ ages $(15, \cdots, 44)$ and $C = 40$ cohorts $(1956, \cdots, 1995)$. In this way, for a specific country, we can define the historical array Φ.

To extract its rough pattern of previous data, we apply **Singular Value Decomposition (SVD)** on Φ, and choose the first three principle components, denote them as $\mathrm{X} \in \mathbb{R}^{30 \times 3}$.



Projection matrix

Notation

1. $\theta_c = (\theta_{c1} \ldots \theta_{cA})^\top \in \mathbb{R}^A$, the fertility schedule for cohort c

To see to what extent our real rates θ_c could be explained by the first three principle components, we can do the projection on it:

$$\theta_c = \mathrm{X}\left(\mathrm{X}^\top \mathrm{X}\right)^{-1} \mathrm{X}^\top \theta_c + \varepsilon_c$$

After projetion, we only need ε_c to measure to what degree the cohort matches with our historical pattern. Note that redisual vector ε_c is:

$$\varepsilon_c = \left[\mathrm{I}_A - \mathrm{X}\left(\mathrm{X}^\top \mathrm{X}\right)^{-1} \mathrm{X}^\top\right] \theta_c = \mathrm{M}\theta_c$$

Construct Shape penalty

Recall

1. $\varepsilon_c = \left[\mathrm{I}_A - \mathrm{X}\left(\mathrm{X}^\top \mathrm{X}\right)^{-1}\mathrm{X}^\top\right]\theta_c = \mathrm{M}\theta_c$

To further quantify 'small' of residual vector, we consider calculating their average outer product :

$$\overline{\Omega} = \frac{1}{s}\sum_{s} \varepsilon_s \varepsilon_s^\top$$

Then, these historical data allow us to establish a scalar penalty for the "badness" of each cohort schedule's shape:

$$\begin{aligned}
\pi_c &= \varepsilon_c^\top \overline{\Omega}^\dagger \varepsilon_c \\
&= \theta_c^\top \left[\mathrm{M}\overline{\Omega}^\dagger \mathrm{M}\right]\theta_c \\
&= \theta^\top \left[\mathrm{G}_c^\top \mathrm{M}\overline{\Omega}^\dagger \mathrm{M}\mathrm{G}_c\right]\theta \\
&= \theta^\top \mathrm{K}_c \theta,
\end{aligned}$$

Results

- An important feature of this π_c is that it's improper;
- By construction, the empirical average of π_c across the historical cohort schedules in Φ equals 27;
- The shape penalty term will give penalty on those ones which don't share similar trend of historic data, but will not be too heavy;

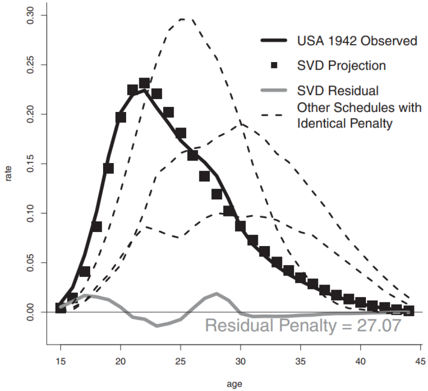


Table of Contents

1. Notation
2. General Idea
3. Shape Penalty
4. Time Penalty
5. Weighting
6. Final result
7. Proof
8. Reference

Freeze rate

The freeze-rate method assumes that the most likely future value for the fertility rate at age a is simply the last observed rate at that age. It suggests that

$$\theta_{a,c+1} \approx \theta_{a,c}$$

which can be expressed in the numerical way, note that at each age on the Lexis surface, we define a vector of 30 freeze-rate residuals for cohorts 1966–1995:

$$\begin{aligned} u_a &= \begin{bmatrix} \theta_{a,1966} - \theta_{a,1965} \\ \vdots \\ \theta_{a,1995} - \theta_{a,1994} \end{bmatrix} \\ &= \begin{bmatrix} 0 & \cdots & -1 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & -1 & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \theta_a \\ &= \mathrm{W}_R \theta_a = \mathrm{W}_R \mathrm{H}_a \theta \end{aligned}$$

Freeze slope

Similarly, the freeze-slope method assumes that trends, measured as fitted slopes over some recent period, will continue into the future. It suggests that

$$\theta_{a,c+1} \approx \theta_{a,c} + \hat{\Delta}_c$$

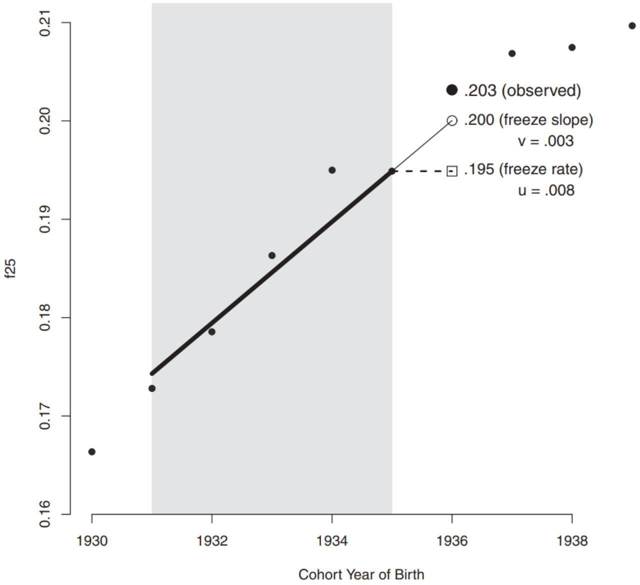
which can be expressed in the numerical way:

$$v_a = \begin{bmatrix} \theta_{a,1966} - \left(\theta_{a,1965} + \hat{\Delta}_{1965}\right) \\ \vdots \\ \theta_{a,1995} - \left(\theta_{a,1994} + \hat{\Delta}_{1994}\right) \end{bmatrix} = \mathrm{W}_S \theta_a = \mathrm{W}_S \mathrm{H}_a \theta$$

where under this context, $\hat{\Delta}$ could be expressed by:

$$\hat{\Delta}_c = \frac{1}{30}\left(10\theta_{a,c} - \theta_{a,c-1} - 2\theta_{a,c-2} - 3\theta_{a,c-3} - 4\theta_{a,c-4}\right)$$

Visualization



Construct Time penalty

Recall

1. Freeze rate: $u_a = \mathrm{W}_R \theta_a = \mathrm{W}_R \mathrm{H}_a \theta$
2. Freeze slope: $v_a = \mathrm{W}_S \theta_a = \mathrm{W}_S \mathrm{H}_a \theta$

As we did in Shape Penalty, to extrapolate the previous time trend into future in the horizontal direction, we also constrcu the time penalty by applying stardization and aggregate them into quadratic penalty term:

Freeze rate

$$\begin{aligned}\pi_{Ra} &= s_{Ra}^{-2} u_a^\top u_a \\ &= \theta^\top \left[s_{Ra}^{-2} \mathrm{H}_a^\top \mathrm{W}_R^\top \mathrm{W}_R \mathrm{H}_a \right] \theta \\ &= \theta^\top \mathrm{K}_{Ra} \theta\end{aligned}$$

Freeze slope

$$\begin{aligned}\pi_{Sa} &= s_{Sa}^{-2} v_a^\top v_a \\ &= \theta^\top \left[s_{Sa}^{-2} \mathrm{H}_a^\top \mathrm{W}_S^\top \mathrm{W}_S \mathrm{H}_a \right] \theta \\ &= \theta^\top \mathrm{K}_{Sa} \theta\end{aligned}$$

where s_{Ra}^{-2}, s_{Sa}^{-2} are the average (mean) squared residuals of freeze rate and freeze slope for each (age, method) combination respectively.

Table of Contents

1. Notation
2. General Idea
3. Shape Penalty
4. Time Penalty
5. Weighting
6. Final result
7. Proof
8. Reference

Weighting

Note that there are actually 90 penalty terms in total, including 30 **Shape Penalty** terms, 30 **Freeze Rate Penalty** terms and 30 **Freeze Slope Penalty** terms.

Since the residuals on which we base the penalties are not mutually independent, we need to assign non-unit weights to each term to better modify our prior distribution. We first assume the covariance matrix K to be constructed as:

$$\mathrm{K} = \sum_{j=1}^{90} w_j \mathrm{K}_j$$

where $K_j \in \mathbb{R}^{CA \times CA}$, are penalty matrices.

Weighting

Theorem

Using notations, condition on θ is restricted to the column space of K, and given weights, we can calculate expectation of π_j based on the trace of multiplication of corresponding matrices and K, which is

$$E^*\left(\pi_j \mid w\right) = \text{trace}\left(\mathrm{K}_j \mathrm{K}^{\dagger}\right)$$

Table 1. Summary of penalties for rate surfaces over birth cohorts 1956–1995 ($C = 40$) and ages 15–44 ($A = 30$)

	Schedule shapes	Time-series (freeze rate)	Time-series (freeze slope)
Number of penalties	30	30	30
Penalty terms	$\pi_{1966} \dots \pi_{1995}$	$\pi_{R,15} \dots \pi_{R,44}$	$\pi_{S,15} \dots \pi_{S,44}$
Residuals	$\varepsilon_c = \mathbf{M}\,\theta_c$	$\mathbf{u}_a = \mathbf{W}_R\theta_a$	$\mathbf{v}_a = \mathbf{W}_S\theta_a$
Penalty matrices	$\mathbf{K}_{1966} \dots \mathbf{K}_{1995}$	$\mathbf{K}_{\mathrm{R},15} \dots \mathbf{K}_{\mathrm{R},44}$	$\mathbf{K}_{\mathrm{S},15} \dots \mathbf{K}_{\mathrm{S},44}$
A priori assumption	Incomplete schedules well approximated by SVD basis functions **X**	Next cohort's rate at age *a* well predicted by current rate	Next cohort's rate at age *a* well predicted by recent trend
Calibration information from historical data	Projection errors from **X**	One-ahead freeze-rate prediction errors	One-ahead freeze-slope prediction errors
Number of elements in each residual	30	30	30
Expected value of each penalty (= rank of **M** or **W**)	27	30	30

Attributing procedure

Recall

1.$\text{target}_j = 27, j = 1, \cdots, 30$
2.$\text{target}_j = 30, j = 31, \cdots, 90$

- Initialize all weights at unity: $w_1 = w_2 = \cdots = w_{90} = 1$;
- Calculate $\mathrm{K} = \sum w_j \mathrm{K}_j$, and its generalized inverse $\mathrm{K}^\dagger$;
- Calculate $E^*(\pi_j \mid w) = \operatorname{trace}\left(\mathrm{K}_j \mathrm{K}^\dagger\right)$, for all $j = 1 \ldots 90$;
- Update weights as $w_j^{\text{new}} = w_j \cdot \frac{E^*\left(\pi_j \mid w\right)}{\text{target}_j}, j = 1 \ldots 90$;
- Stop if converged; otherwise return to Step 1.

Final posterior expression

Finally, after calculating the K matrix for prior distribution, we can write the expression for posterior distribution based on the formula:

$$\ln P(\theta \mid y) = \text{const} - \frac{1}{2}(y - \mathrm{V}\theta)^\top \Psi^{-1}(y - \mathrm{V}\theta) - \frac{1}{2}\theta^\top \mathrm{K}\theta$$

And through the procedure of maximizing the posterior distribution, we can get the optimal μ_{post} and Σ_{post}, it suggests that:

$$\theta \mid y \sim \mathrm{N}\left\{ \begin{array}{ll} \mu_{\text{post}} & = \left[\mathrm{V}^\top \Psi^{-1} \mathrm{V} + \mathrm{K}\right]^{-1} \left[\mathrm{V}^\top \Psi^{-1} y\right] \\ \Sigma_{\text{post}} & = \left[\mathrm{V}^\top \Psi^{-1} \mathrm{V} + \mathrm{K}\right]^{-1} \end{array} \right\}$$

which is our final result.

Table of Contents

1. Notation
2. General Idea
3. Shape Penalty
4. Time Penalty
5. Weighting
6. Final result
7. Proof
8. Reference

An example for Singapore cohort fertility

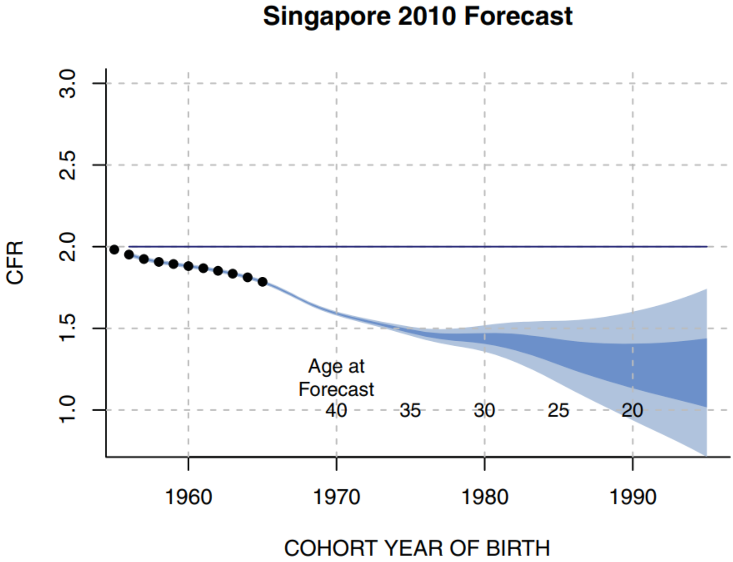


Table of Contents

1. Notation
2. General Idea
3. Shape Penalty
4. Time Penalty
5. Weighting
6. Final result
7. Proof
8. Reference

Proof

Penalty Terms

Why expected value of shape penalty, freeze-rate penalty and freeze-slope penalty equal to 27, 30 and 30 respectively? (As in table 1)

Pf: 1) **Shape Penalty**

According to the paper, X is the first three principal components of historical array $\Phi \in \mathbb{R}^{A \times S}$, where S is the number of historical complete cohorts in 1900 to 1949, i.e.

$$\text{Denote } \Phi = \mathrm{UDV}^\top, \quad \mathrm{X} = \mathrm{U}(:, 1:3) \in \mathbb{R}^{A \times 3}$$

Then we consider the residual vector between θ_c and $\theta_c \dot{s}$ linear projection on X 's column space $(c = 1, \cdots, S)$, ie.

$$\varepsilon_c = \left[\mathrm{I}_A - \mathrm{X} \left(\mathrm{X}^\top \mathrm{X} \right)^{-1} \mathrm{X}^\top \right] \theta_c \triangleq \mathrm{M} \theta_c$$

Proof

Note that M $\left(= \mathrm{I}_A - \mathrm{X}\left(\mathrm{X}^\top \mathrm{X}\right)^{-1}\mathrm{X}^\top\right)$ is an idempotent matrix, therefore,

$$
\begin{aligned}
\mathrm{rank}(\mathrm{M}) &= \mathrm{trace}(\mathrm{M}) \\
&= A - \mathrm{trace}\left(\mathrm{X}\left(\mathrm{X}^\top \mathrm{X}\right)^{-1}\mathrm{X}^\top\right) \\
&= A - \mathrm{trace}\left(\left(\mathrm{X}^\top \mathrm{X}\right)^{-1}\mathrm{X}^\top \mathrm{X}\right) \\
&= A - 3
\end{aligned}
$$

Still, according to the paper, $\bar{\Omega} = \frac{1}{S}\sum_{c=1}^{S} \varepsilon_c \varepsilon_c^\top$ is covariance matrix for complete cohorts between 1900-1949, then,

$$
\begin{aligned}
\frac{1}{S}\sum_{c=1}^{S} \varepsilon_c^\top \bar{\Omega}^\dagger \varepsilon_c &= \frac{1}{S}\ \mathrm{trace}\ \left(\bar{\Omega}^\dagger \sum_{c=1}^{S} \varepsilon_c \varepsilon_c^\top\right) \\
&= \mathrm{trace}\left(\bar{\Omega}^\dagger \bar{\Omega}\right) = \mathrm{rank}(\mathrm{M}) \\
&= A - 3
\end{aligned}
$$

Proof

2) **Freeze-rate Penalty**

$$\frac{1}{A}\sum_{a=1}^{A} s_{R_a}^{-2} u_a^{\top} u_a = \frac{1}{A}\sum_{a=1}^{A}\left(\frac{\sum_{c=1}^{S} u_{a,c}^2}{S_{Ra}^2}\right)$$
$$= \frac{1}{A}\sum_{a=1}^{A} S$$
$$= S$$

3) **Freeze-slope Penalty**

Similarly,

$$\frac{1}{A}\sum_{a=1}^{A} s_{s_a}^{-2} v_a^{\top} v_a = S$$

Proof

To conclude, by designing prior above, during forecasting procedure, we believe they generate from
for shape penalty,

$$\varepsilon_c \sim N(\underline{0}, \bar{\Omega})$$

for freeze rate penalty,

$$u_a \sim N\left(\underline{0}, \begin{pmatrix} s_{R_1}^2 & & \\ & \ddots & s_{Ra}^2 \end{pmatrix}\right)$$

for freeze rate penalty,

$$v_a \sim N\left(\underline{0}, \begin{pmatrix} s_{S_1}^2 & & \\ & \ddots & \\ & & s_{S_A}^2 \end{pmatrix}\right)$$

Proof

Hence, when $a = 1, \cdots, 30(15, \cdots, 44)$, $C = 1, \cdots, 30(1966, \cdots, 1995)$, we have

$$\mathbb{E}\left[\pi_c \mid \text{ history } \right] = A - 3 = 27$$

$$\mathbb{E}\left[\pi_{R_a} \mid \text{ history } \right] = C = 30$$

$$\mathbb{E}\left[\pi_{s_a} \mid \text{ history } \right] = C = 30$$

Table of Contents

1. Notation
2. General Idea
3. Shape Penalty
4. Time Penalty
5. Weighting
6. Final result
7. Proof
8. Reference

References

Schmertmann, C., Zagheni, E., Goldstein, J. R., Myrskylä, M. (2014). Bayesian forecasting of cohort fertility. Journal of the American Statistical Association, 109(506), 500-513.