

Modelling covariance kernels for nonstationary random fields

Christopher G. Small
University of Waterloo

University of Guelph, October 2007

1. Random fields and covariance kernels
2. The role of covariance kernels in semiparametric inference
3. The Karhunen-Loève expansion
4. The estimation problem reconsidered
5. An application

1. Random fields and covariance kernels

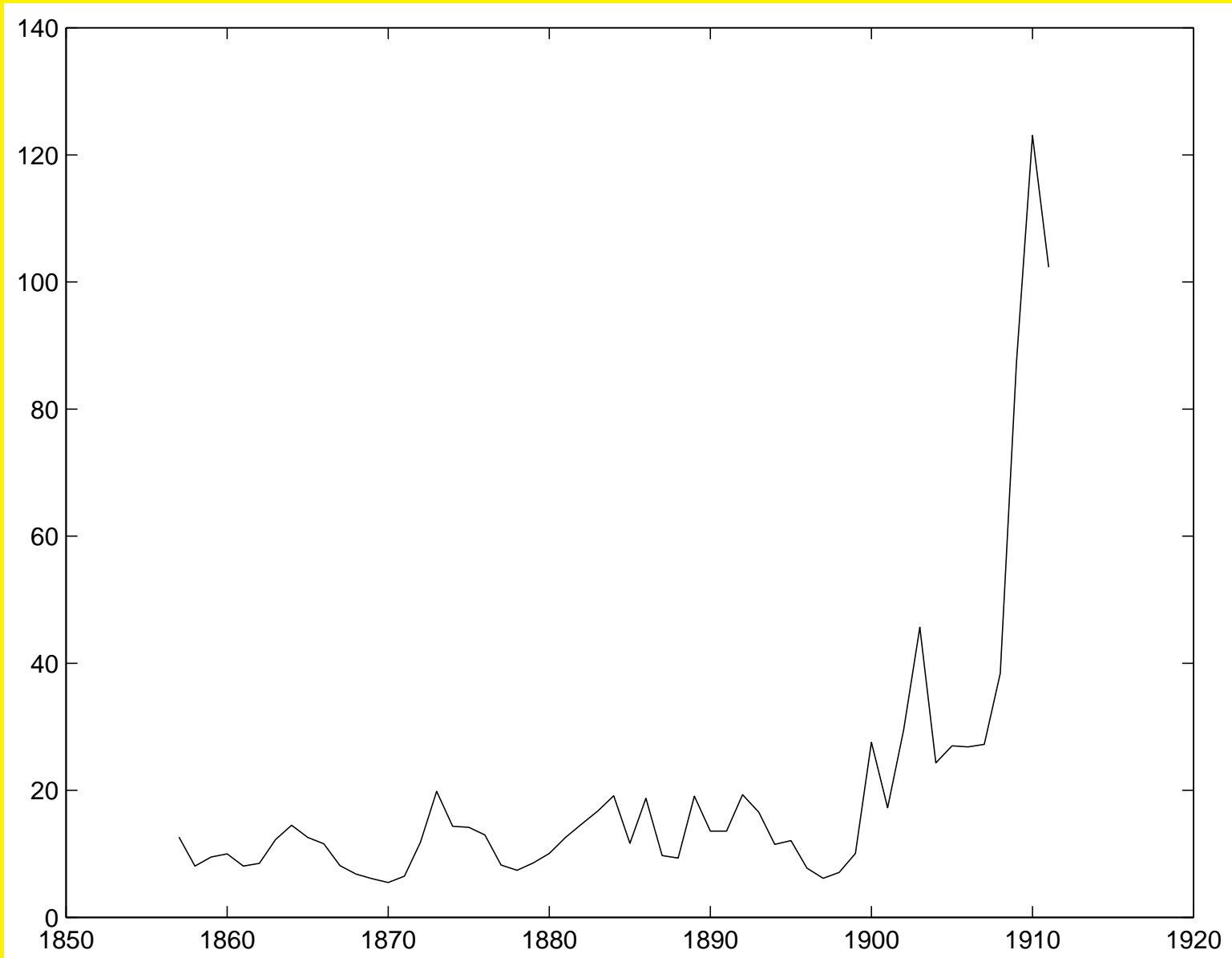
- For a random sample, inference about the mean μ of the variates depends upon knowledge or estimation of the **variance** σ^2 . In practice the variance is harder to estimate than the mean.
- Moreover, **heteroscedasticity** is a problem for the optimal estimation of μ . To optimally estimate μ we must model or estimate the variance function.
- For a random field $X(\mathbf{t})$, inference about the mean function $\mu(\mathbf{t})$ requires knowledge or estimation of the **covariance kernel** $\Gamma(\mathbf{s}, \mathbf{t}) = \text{Cov}[X(\mathbf{s}), X(\mathbf{t})]$.
- When Γ is unknown we need to estimate it. What methods are available when the process is not second order stationary, *i.e.*, when $\Gamma(\mathbf{s}, \mathbf{t}) \neq \gamma(\mathbf{s} - \mathbf{t})$?

- By a **random field** we shall mean a family of random variables $X(\mathbf{t})$ indexed by some parameter $\mathbf{t} \in \mathbb{R}^q$.
- In practice, we only observe a finite “piece” of this random field. So we shall assume that \mathbf{t} lies in some bounded subset \mathcal{R} of \mathbb{R}^q .
- When \mathcal{R} is a countable set—finite or denumerable, usually a lattice—then we say that $X(\mathbf{t})$ is a **discrete random field**.
- When \mathcal{R} is an open subset of \mathbb{R}^q then $X(\mathbf{t})$ is said to be a **continuous random field**.

Stochastic processes $X(t), t \geq 0$ are random fields

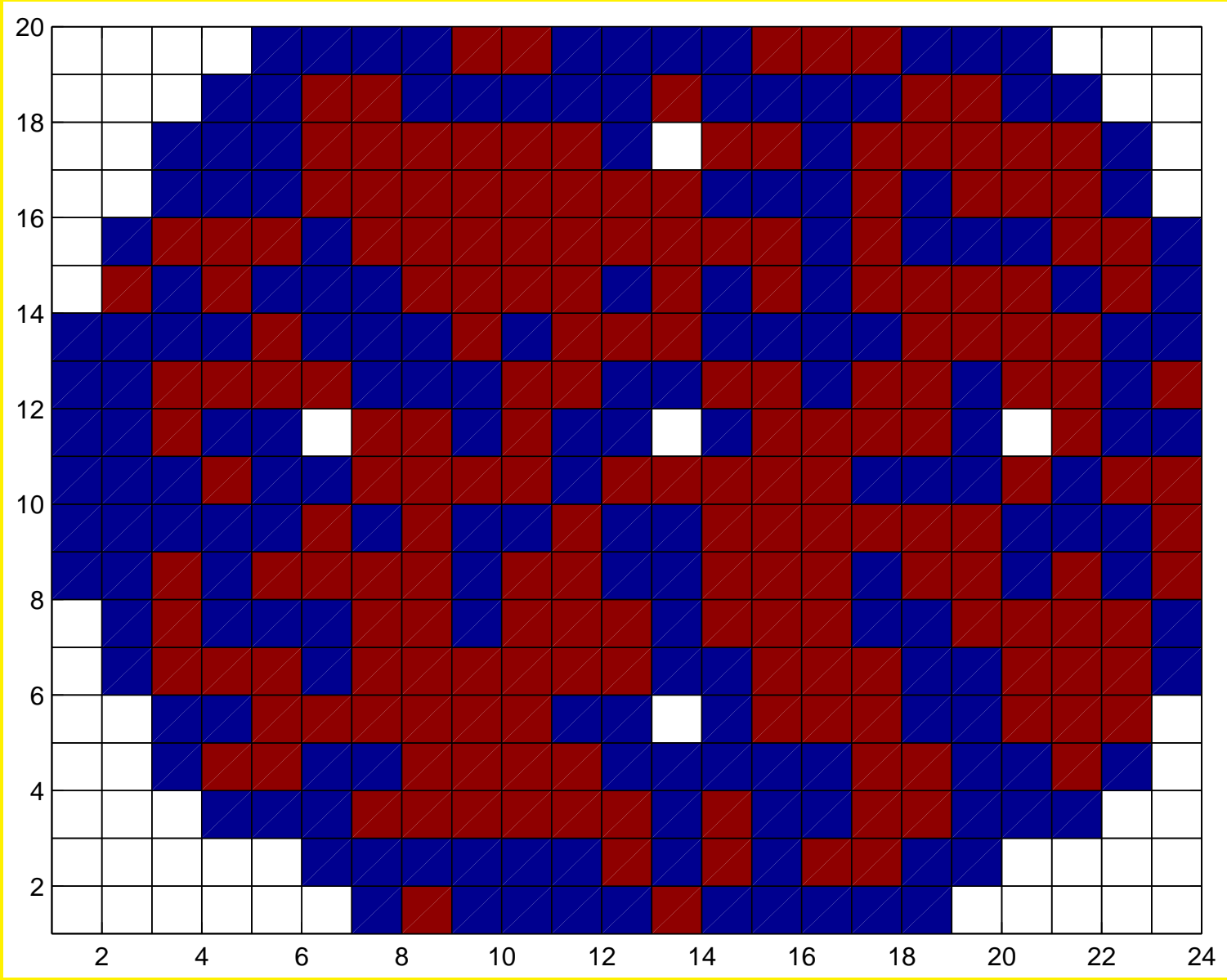
Example: Lynx Pelt Prices, HBC 1857-1911.

Elton & Nicholson (1942).



Random sets are also special cases where $X(\mathbf{t}) \in \{0, 1\}$

Example: Two-dimensional random set.
Integrated circuit data, Mallory et al. (1983).



2. Role of covariance kernels in semiparametric inference

- Let

$$E[X(\mathbf{t})] = \mu_{\theta}(\mathbf{t}) \text{ and}$$

$$\text{Cov}[X(\mathbf{s}), X(\mathbf{t})] = \Gamma_{\theta}(\mathbf{s}, \mathbf{t})$$

be the mean function and covariance kernel respectively, where $\mathbf{s}, \mathbf{t} \in \mathcal{R}$ and $\boldsymbol{\theta} \in \mathbb{R}^k$.

- Both μ_{θ} and Γ_{θ} are assumed to be known real-valued functions of the unknown parameter $\boldsymbol{\theta} \in \mathbb{R}^k$.

- With these **semiparametric** assumptions, $\boldsymbol{\theta}$ can be estimated by a **linear functional estimating equation** of the form

$$\mathbb{L}(X; \hat{\boldsymbol{\theta}}) = \mathbf{0}$$

where

$$\mathbb{L}(X; \boldsymbol{\theta}) = \int_{\mathcal{R}} [X(\mathbf{t}) - \mu_{\boldsymbol{\theta}}(\mathbf{t})] d\mathbb{A}_{\boldsymbol{\theta}}(\mathbf{t}),$$

where $\mathbb{A}_{\boldsymbol{\theta}}$ is a vector-valued measure on \mathcal{R} taking values in \mathbb{R}^k .

- For a **discrete random field** where \mathbf{t} typically lies in a lattice, this reduces to an estimating function of the form

$$\mathbb{L}(X; \boldsymbol{\theta}) = \sum_{\mathbf{t} \in \mathcal{R}} \mathbf{a}_{\boldsymbol{\theta}}(\mathbf{t}) [X(\mathbf{t}) - \mu_{\boldsymbol{\theta}}(\mathbf{t})].$$

- For a **continuous random field**, typically $d\mathbb{A}_{\boldsymbol{\theta}}(\mathbf{t}) = \mathbf{a}_{\boldsymbol{\theta}}(\mathbf{t}) d\mathbf{t}$, so that

$$\mathbb{L}(X; \boldsymbol{\theta}) = \int_{\mathcal{R}} \mathbf{a}_{\boldsymbol{\theta}}(\mathbf{t}) [X(\mathbf{t}) - \mu_{\boldsymbol{\theta}}(\mathbf{t})] d\mathbf{t}.$$

- In both cases, $\mathbf{a}_{\boldsymbol{\theta}} : \mathbb{R}^q \rightarrow \mathbb{R}^k$ is a vector-valued coefficient function which is functionally independent of X .
- In this talk, we will emphasize the continuous case. However, most remarks apply with appropriate modification to other types of random fields.

- The **optimal estimating function** is that which has a vector-valued measure \mathbb{A}_θ satisfying

$$\int_{\mathcal{R}} \Gamma_\theta(\mathbf{s}, \mathbf{t}) d\mathbb{A}_\theta(\mathbf{s}) = \dot{\boldsymbol{\mu}}_\theta(\mathbf{t}).$$

where $\dot{\boldsymbol{\mu}}_\theta(\mathbf{t})$ is the vector-valued partial derivative of $\mu_\theta(\mathbf{t})$ with respect to $\boldsymbol{\theta}$.

There are two problems with implementing this optimal solution:

- **Problem 1.** Note that the equation for \mathbb{A}_θ must be solved for **each value** of the parameter θ , iteratively used **within** any algorithm that solves the equation $\mathbb{L}(\hat{\theta}) = \mathbf{0}$.
 - For example, when $\theta \in \mathbb{R}^2$, a discrete random field on a 20×20 lattice requiring as little as ten iterations over θ , will need the solution to 800 simultaneous non-sparse linear equations, ten successive times in a row, just to produce a single approximation to $\hat{\theta}$.
- **Problem 2.** In practice, we do not know Γ_θ . This must usually be estimated as well!!

3. The Karhunen-Loève expansion

- The solution to both of these problems can be obtained using the **Karhunen-Loève expansion**.
- Let $b_1(\mathbf{t}), b_2(\mathbf{t}), \dots$ be the set of **eigenfunctions** for the kernel Γ satisfying

$$\int_{\mathcal{R}} b_j(\mathbf{s})\Gamma(\mathbf{s}, \mathbf{t}) d\mathbf{s} = \sigma_j^2 b_j(\mathbf{t})$$

for $j = 1, 2, \dots$. Here, the parameter θ is suppressed in the notation for simplicity. Since Γ is symmetric, the **eigenfunctions** b_j can be chosen to be **real and orthonormal**.

- Since Γ is positive definite, the **eigenvalues** will be also be positive. So we can write the j^{th} eigenvalue as σ_j^2 .
- Provided that the kernel function Γ is **complete**, the set of standardised eigenfunctions of Γ will form an orthonormal basis for $\mathcal{L}^2([\mathcal{R}])$.

- Using the completeness condition, we may write

$$X(\mathbf{t}) = \sum_{j=1}^{\infty} Y_j b_j(\mathbf{t}),$$

where Y_1, Y_2, \dots satisfy

$$Y_j = \int_{\mathcal{R}} X(\mathbf{t}) b_j(\mathbf{t}) d\mathbf{t}.$$

- Let $E(Y_j) = \mu_j$ for all j .
- We have $\text{Var}(Y_j) = \sigma_j^2$.
- We will also need

$$\dot{\boldsymbol{\mu}}(\mathbf{t}) = \sum_{j=1}^{\infty} \dot{\boldsymbol{\mu}}_j b_j(\mathbf{t})$$

where $\dot{\boldsymbol{\mu}}_j = \int \dot{\boldsymbol{\mu}}(\mathbf{t}) b_j(\mathbf{t}) d\mathbf{t}$.

- Writing out X in terms of the Karhunen-Loève expansion, we obtain an equivalent expression for $\mathbb{L}(\boldsymbol{\theta})$, namely

$$\mathbb{L}(\boldsymbol{\theta}) = \sum_{j=1}^{\infty} \sigma_j^{-2}(\boldsymbol{\theta}) \dot{\mu}_j(\boldsymbol{\theta}) [Y_j(\boldsymbol{\theta}) - \mu_j(\boldsymbol{\theta})] ,$$

which is a rather standard looking **quasi-likelihood equation**, with the exception that the random variables Y_j are also functions of the parameter $\boldsymbol{\theta}$.

4. The estimation problem reconsidered

Proposed solution to Problem 1:

- We need only sum the first few terms of the K.-L. expansion. Since $\sum_j \sigma_j^2 < \infty$, we choose terms with the most significant leading eigenvalues. Say, the first m terms.
- Instead, choose $Y_j^* = Y_j(\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ is some simple consistent approximation to $\boldsymbol{\theta}$ —possibly, but not necessarily an estimator. However, consider $\boldsymbol{\theta}^*$ as fixed, not random.

- Reduce the problem of estimating $\boldsymbol{\theta}$, to that of estimation given $Y_1^*, Y_2^*, \dots, Y_m^*$ as **data**. The GEE has the form

$$\sum_{j=1}^m [\sigma_j^*(\boldsymbol{\theta})]^{-2} \dot{\boldsymbol{\mu}}_j^*(\boldsymbol{\theta}) [Y_j^* - \boldsymbol{\mu}_j^*(\boldsymbol{\theta})] = 0.$$

where

$$\boldsymbol{\mu}_j^*(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(Y_j^*), \quad \sigma_j^*(\boldsymbol{\theta}) = \text{Var}_{\boldsymbol{\theta}}(Y_j^*),$$

and

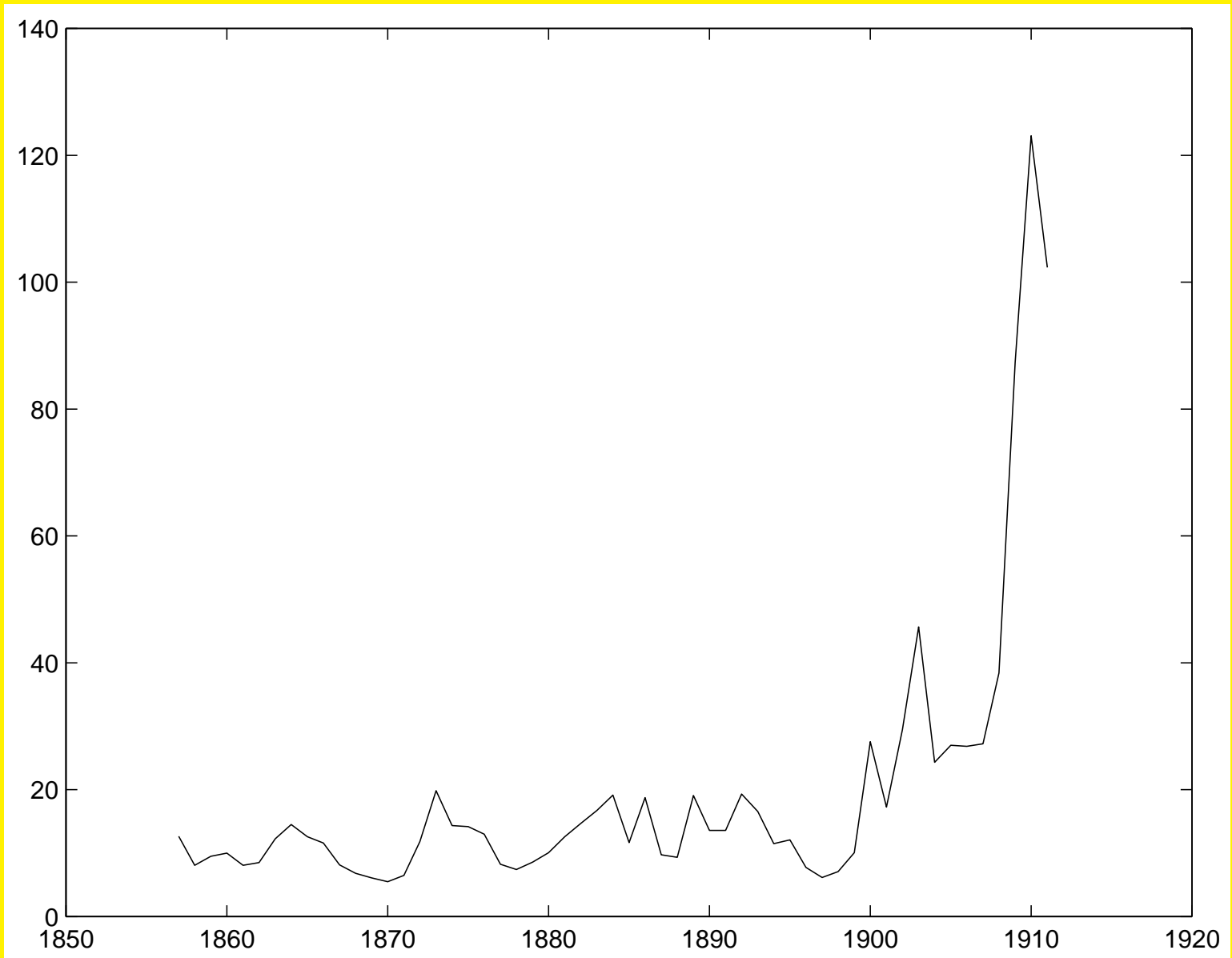
$$\dot{\boldsymbol{\mu}}_j^* = \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\mu}_j^*(\boldsymbol{\theta}).$$

Proposed solution to Problem 2:

- If the covariance kernel Γ is an unknown function of $\boldsymbol{\theta}$, we can estimate it directly.
- This is often done by assuming that $\Gamma(\mathbf{s}, \mathbf{t}) = \gamma(\mathbf{s} - \mathbf{t})$.
- But such a **stationarity assumption** is
 - artificial if $\mu(\mathbf{t})$ is not constant;
 - requires special constraints on γ to make Γ nonnegative definite.
- An alternative is to use a **working product kernel** with unknown coefficients

5. An application

Example: Lynx Pelt Prices (Continued).



- Lynx populations rose and fell on a 10 year cycle.

Lynx Pelt Prices (Continued).

- The prices looks stationary up to 1899.
- There is also the 10-year oscillation in the lynx population which may have influenced lynx pelt prices. This 10-year cycle of the lynx population can be explained by the **predator-prey** equations for the populations of **lynx** and its main prey, the **snowshoe rabbit**.
- Stationarity appears to make sense.
- However

Lynx Pelt Prices (Continued).

- By 1900 and after, prices increased dramatically. This is associated with reduced catches of lynx.
- “The **smallpox**, killing off a large fraction of the Indian population, accounts for the greatly reduced catches of the fifteen years that followed [the years 1878 to 1890].”
 - Elton and Nicholson (1942).

- It is always dangerous to assume stationarity for **socio-historical** data.
- Unlike the predator-prey relationships that govern the 10-year cycle of the lynx and the snowshoe rabbit, socio-historical data are influenced by time-irreversible historical events.

- What can we deduce without assuming stationarity?
- We propose a **working covariance kernel**

- Let $b_1(\mathbf{t}), \dots, b_m(\mathbf{t})$ be orthonormal functions.
- We fit a covariance kernel of the form

$$\Gamma(\mathbf{s}, \mathbf{t}) = \hat{\sigma}_1^2 b_1(\mathbf{s}) b_1(\mathbf{t}) + \dots + \hat{\sigma}_m^2 b_m(\mathbf{s}) b_m(\mathbf{t}).$$

- The eigenvectors $\hat{\sigma}_j^2$ are **estimated** from the data.
- We **choose** the functions b_j by using a mathematically **tractable class** of functions, e.g., trigonometric functions (which arise from the Laplacian kernel for example).
- The class of covariance kernels so defined can be called **working product kernels**.

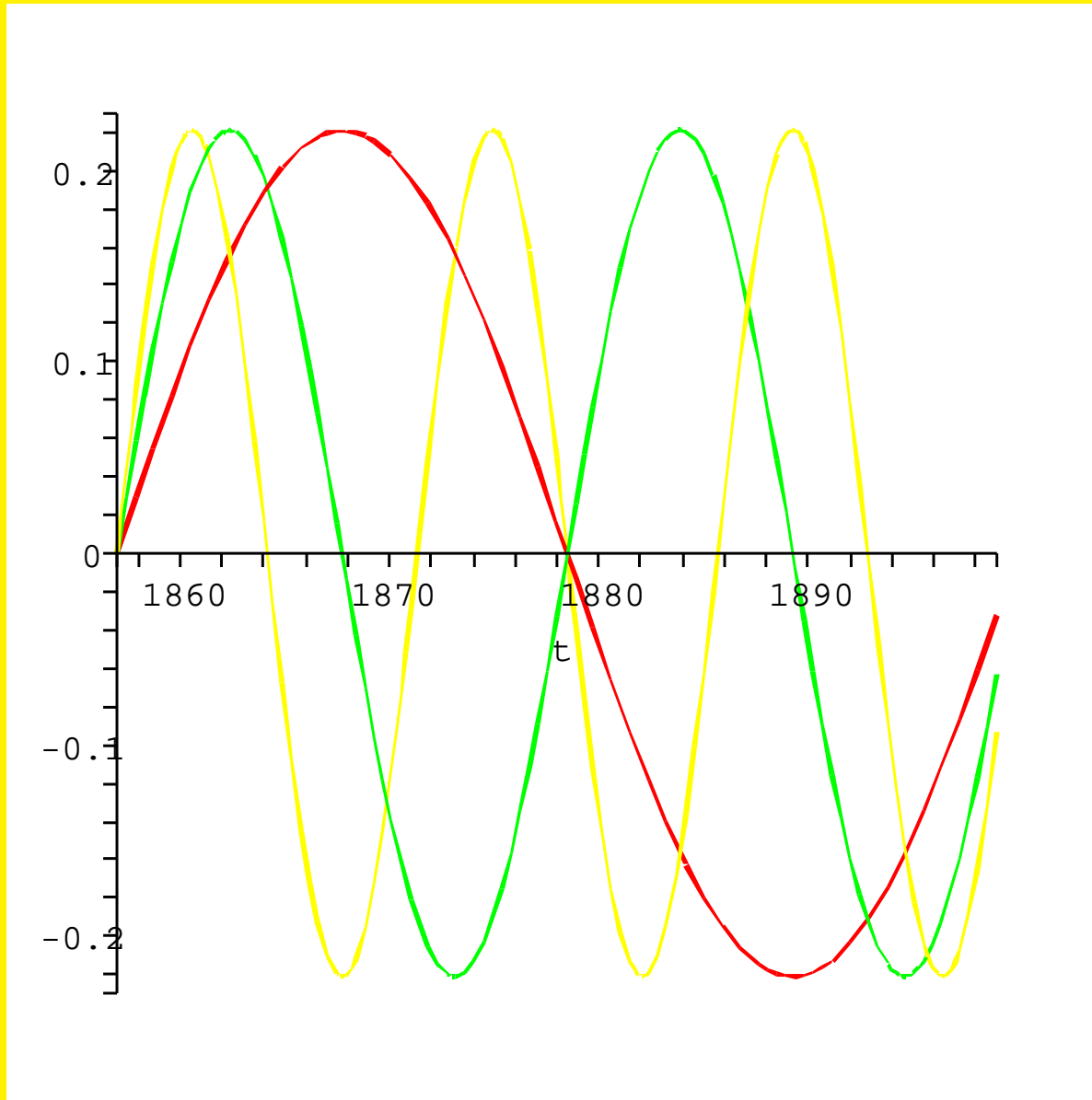
Fitting of Eigenvalues:

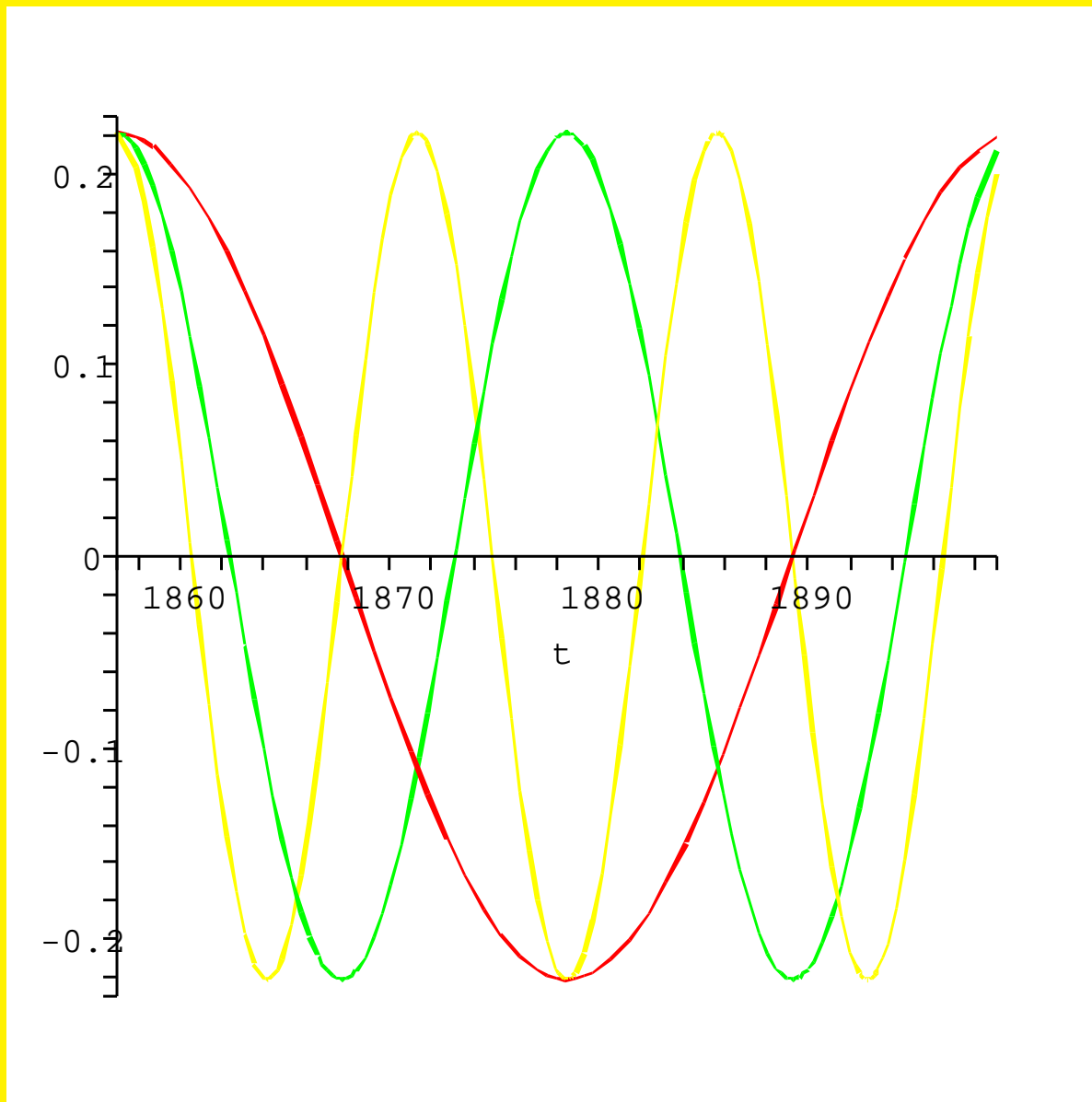
- Estimate $\mu(\mathbf{t})$ by some “rough” estimate $\hat{\mu}(\mathbf{t})$, such as a moving average of $X(\mathbf{t})$, or
- as $\hat{\mu}(\mathbf{t}) = \mu_{\theta^*}(\mathbf{t})$ if this information is available.
- Set

$$\hat{\sigma}_j^2 = \left\{ \int_{\mathcal{R}} [X(\mathbf{t}) - \hat{\mu}(\mathbf{t})] b_j(\mathbf{t}) dt \right\}^2 .$$

Lynx Pelt Prices (Continued).

- Let us perform a nonparametric fit to the covariance kernel of the lynx pelt data.
- We need to choose some sensible basis functions.....
- The trigonometric functions can form an orthonormal basis for the interval:

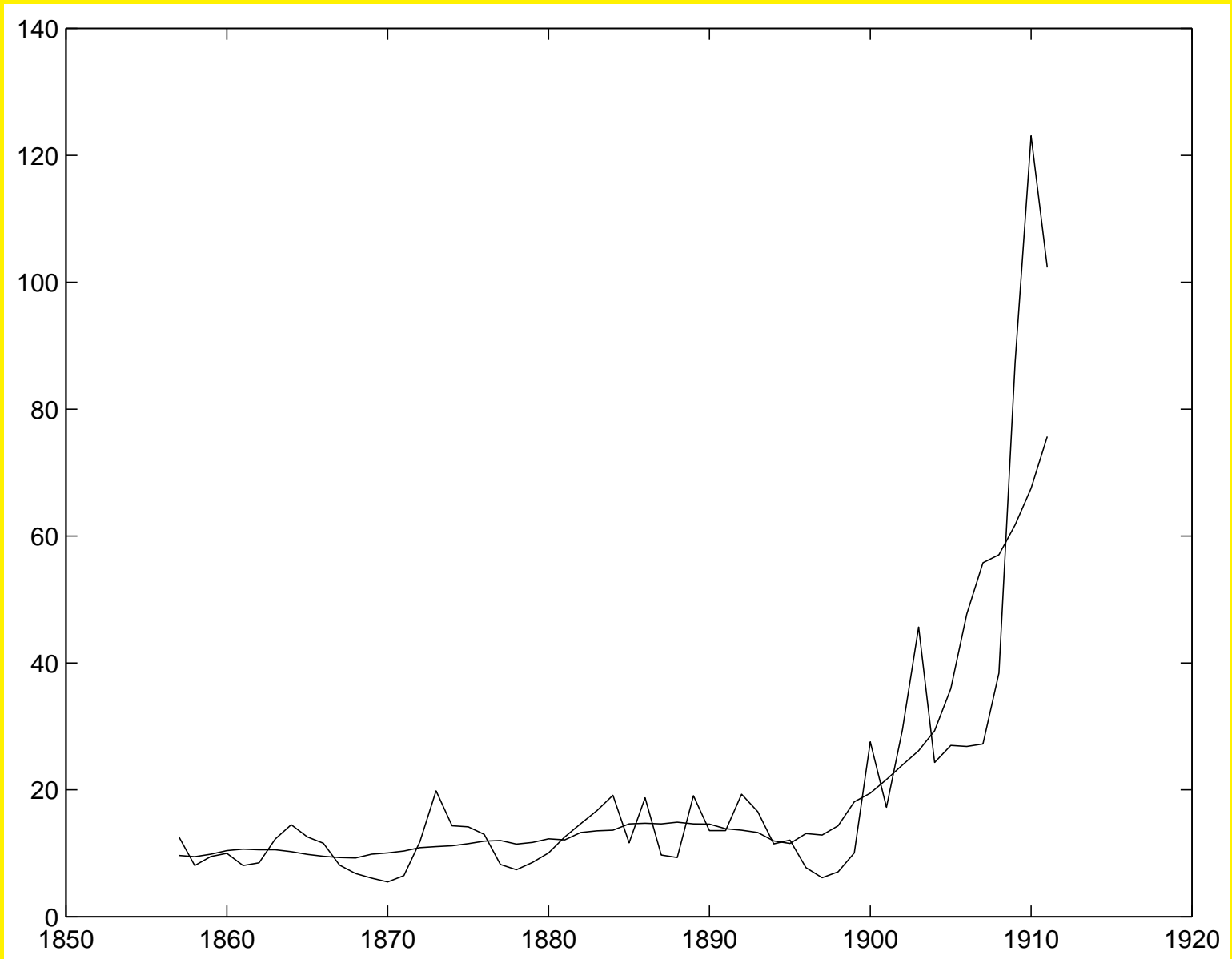




Lynx Pelt Prices (Continued).

- Let us try to find estimates for μ and Γ for the Lynx Pelt Price Data.
- It is a straightforward matter to estimate μ by a moving average, which will be appropriate if μ does not oscillate too much

Lynx Pelt Prices (Continued).



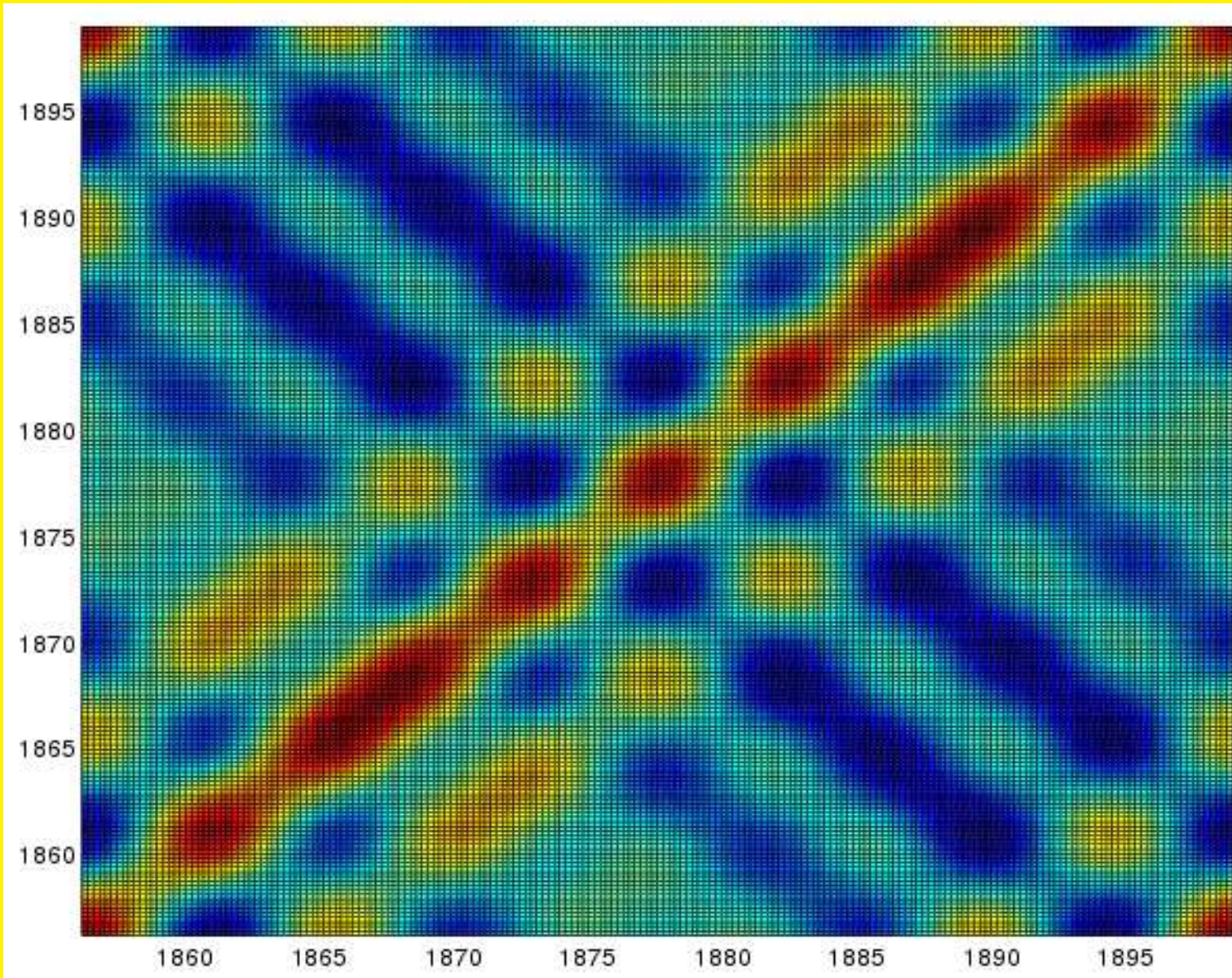
Lynx Pelt Prices (Continued).

- But what about the kernel Γ ?

Lynx Pelt Prices (Continued).

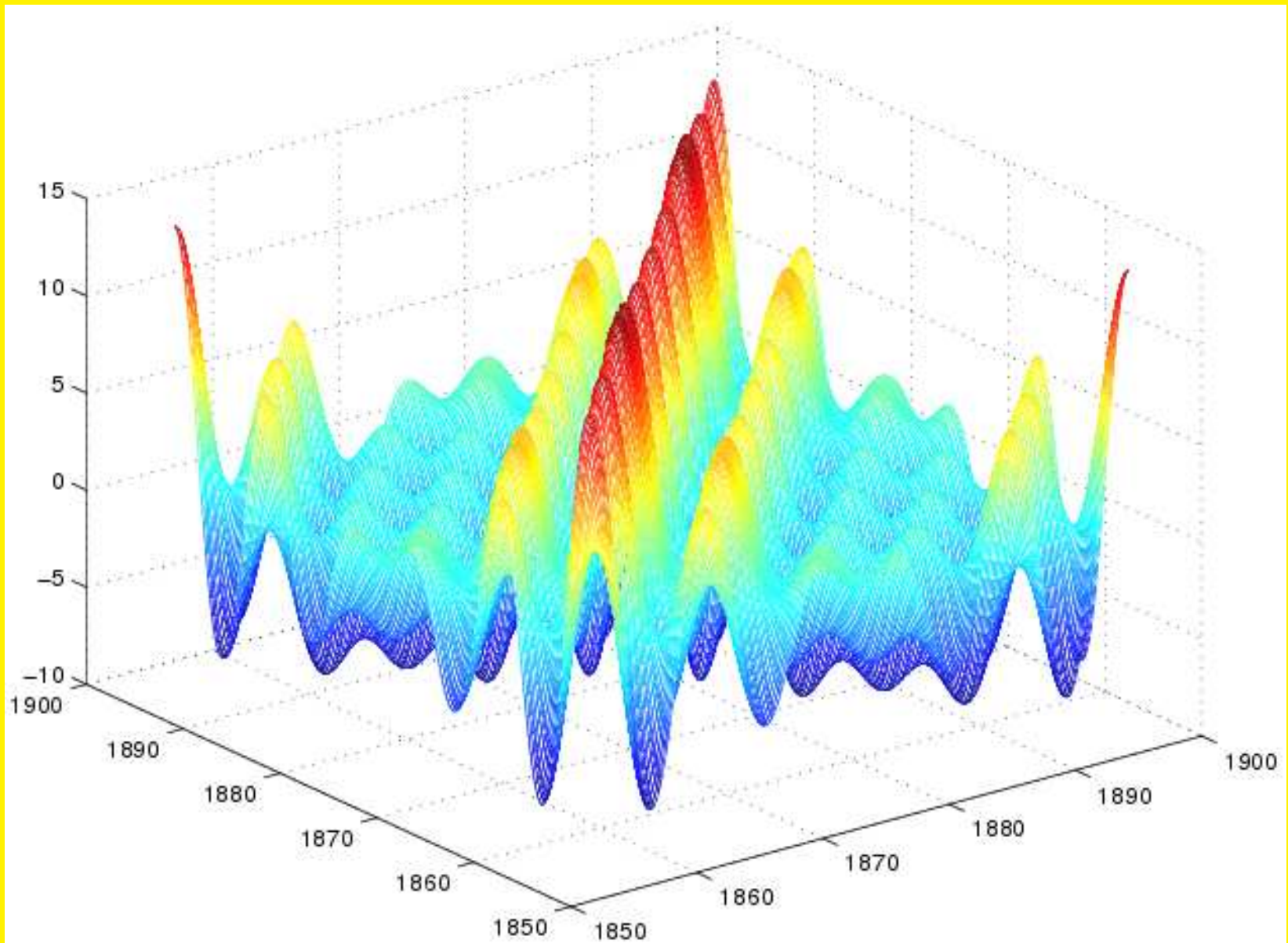
- We can estimate the coefficients σ_j^2 of the product kernel. We rescale the time axis to $[0, 2\pi]$ for simplicity to obtain:

j	b_j	$\hat{\sigma}_j^2$	j	b_j	$\hat{\sigma}_j^2$
1	1	0	7	$\cos 3t$	1.110
2	$\sin t$	3.486	8	$\sin 4t$	15.571
3	$\cos t$	0.587	9	$\cos 4t$	1.914
4	$\sin 2t$	4.193	10	$\sin 5t$	14.870
5	$\cos 2t$	3.638	11	$\cos 5t$	1.093
6	$\sin 3t$	1.555	12	$\sin 6t$	0.148

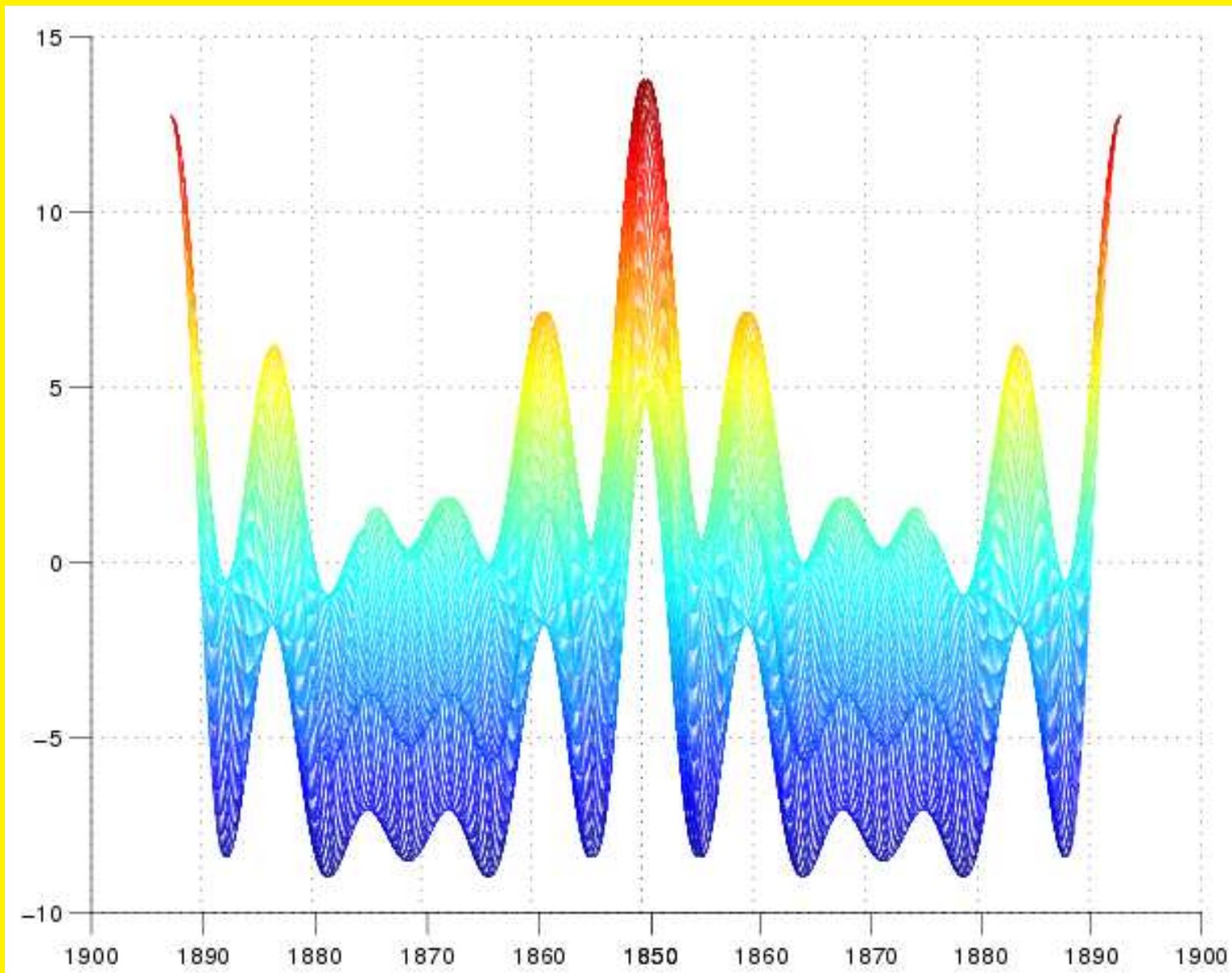


Pseudocolour plot of the covariance kernel

- We “read” the variance function along the main diagonal from lower left to upper right.
- Bright bands parallel and on each side of the main diagonal indicate high autocorrelation with a lag of about ≈ 10 years—the same lag as the predator-prey cycle for the lynx.



- Surface plot of the same working product kernel as seen from above at an angle.
- The high autocorrelation “detected” at the extremes is an anomalous consequence of the periodicity of the trigonometric basis functions—it should be discounted.



- The same working product kernel from the side looking down the main diagonal. The **10 year cycle** –actually 9.6 – is clearly seen.
- The “thickness” of this graph is a measure of the departure of the kernel from **stationarity**: i.e., a kernel of the form

$$\Gamma(s, t) = \gamma(t - s).$$

- With a fit to the covariance kernel, we can estimate variation.
- For example, in dimension one, if the estimate $\hat{\theta}$ can be written as

$$\hat{\theta} = \int_{\mathcal{R}} a(t) X(t) dt ,$$

then

$$\begin{aligned} \widehat{\text{Var}}(\hat{\theta}) &= \int_{\mathcal{R}} \int_{\mathcal{R}} a(s) a(t) \hat{\Gamma}(s, t) ds dt \\ &= \sum_{j=1}^m \hat{\sigma}_j^2 \left[\int_{\mathcal{R}} a(t) b_j(t) dt \right]^2 . \end{aligned}$$

Selected References:

- Cressie (1993). *Statistics for Spatial Data*. Wiley.
- Elton & Nicholson (1942). *J. Animal Ecology*, 215-244.
- Heyde (1997). *Quasi-Likelihood and Its Application*. Springer.
- Small & Wang (2003). *Numerical Methods for Nonlinear Estimating Equations*. Oxford U.
- Verwoerd & Kulasiri (1999). *Proc. Int. Congress Modelling & Simulation*.