Modelling Phenotypic Evolution
by
Stochastic Differential Equations
Tore Schweder and Trond Reitan
University of Oslo
Jorijntje Henderiks
University of Uppsala
Combining statistical timeseries with fossil measurements.
Overview

1. Introduction:
   • Motivating example:
     ➢ Phenotypic evolution: irregular time series related by common latent processes
     ➢ Coccolith data (microfossils)
     ➢ 205 data points in space (6 sites) and time (60 million years)

2. Stochastic differential equation vector processes
   • Ito representation and diagonalization
   • Tracking processes and hidden layers
   • Kalman filtering

3. Analysis of coccolith data
   • Results for original model
   • 723 different models – Bayesian model selection and inference

4. Second application: Phenotypic evolution on a phylogenetic tree
   • Primates - preliminary results

5. Conclusion
The size of a single cell algae (*Coccolithus*) is measured by the diameter of its fossilized coccoliths (calcite platelets). Want to model the evolution of a lineage found at six sites.

- In continuous time and "continuously".
- By tracking a changing fitness optimum.
- Fitness might be influenced by observed (temperature) and unobserved processes.
- Both fitness and underlying processes might be correlated across sites.

19,899 coccolith measurements, 205 sediment samples (1 < n < 400) of body size by site and time (0 to -60 my).
Our data – Coccolith size measurements

205 Sample mean log coccolith size (1 < n < 400) by time and site.
Individual samples

Bi-modality rather common. Speciation? Not studied here!
Evolution of size distribution

Fitness = expected number of reproducing offspring.

The population tracks the fitness curve (natural selection). The fitness curve moves about, and the population follow.

With a known fitness, $\mu$, the mean phenotype should be an Ornstein-Uhlenbeck process (Lande 1976).

$$dX(t) = -\alpha X(t) - \mu \frac{d}{dt} + \sigma dW(t)$$

With fitness as a process, $\mu(t)$, we can make a tracking model:

$$dX(t) = -\alpha X(t) - \mu(t) \frac{d}{dt} + \sigma dW(t)$$
The Ornstein-Uhlenbeck process

\[ dX(t) = -\alpha X(t) - \mu \, dt + \sigma dW(t) \]

Attributes:
- Normally distributed
- Markovian
- Long-term level: \( \mu \)
- Standard deviation: \( s = \sigma / \sqrt{2\alpha} \)
- \( \alpha \): pull
- Time for the correlation to drop to \( 1/e \): \( \Delta t = 1/\alpha \)

The parameters \((\mu, \Delta t, s)\) can be estimated from the data. In this case: \( \mu \approx 1.99, \Delta t = 1/\alpha \approx 0.80 \text{Myr}, s \approx 0.12. \)
One layer tracking another

\[ dX_1(t) = -\alpha_1 X_1(t) - X_2(t) \, dt + \sigma_1 \, dW_1(t) \]
\[ dX_2(t) = -\alpha_2 X_2(t) - \mu \, dt + \sigma_2 \, dW_2(t) \]

Red process \((\Delta t_2 = 1/\alpha_2 = 0.2, s_2 = 2)\) tracking black process \((\Delta t_1 = 1/\alpha_1 = 2, s = 1)\)

Auto-correlation of the upper (black) process, compared to a one-layered SDE model.

A slow-tracking-fast can always be re-scaled to a fast-tracking slow process. Impose identifying restriction: \(\alpha_1 \geq \alpha_2\)
Process layers - illustration

Layer 1 – local phenotypic expression
Layer 2 – local fitness optimum
Layer 3 – hidden climate variations or primary optimum

Observations

External series
Coccolith model

\( X_{s,l}(t) = \text{state of the } l^{th} \text{ process at site } s \text{ at time } t \) (log size).

\[
dX_{s,1}(t) = -\alpha_{s,1} X_{s,1}(t) - X_{s,2}(t) \, dt + \text{noise}
\]

\[
dX_{s,2}(t) = -\alpha_{s,2} X_{s,2}(t) - X_{s,3}(t) - \beta_{s} u(t) \, dt + \text{noise}
\]

\[
dX_{s,3}(t) = -\alpha_{s,3} X_{s,3}(t) - \mu \, dt + \text{noise}
\]

\( u(t) = \text{observed covariate process (temperature)} \)

\( \alpha_{s,l} \geq 0 \) is the tracking speed of site \( s \) belonging to the \( l^{th} \) layer

\( \alpha_{s,l} = 0 \Rightarrow \text{random walk, } \alpha_{s,l} = \infty \Rightarrow \text{perfect tracking} \)

The noise is white, possibly correlated inside a layer

\( X_3, u \rightarrow X_2 \rightarrow X_1 \)

Observations \( Z_{s,t} \) are \( N(X_{s,1}(t), \sigma_{s,t}^2) \)
Stochastic differential equation (SDE) vector processes

\[ X(t) : p \text{- dimensional vector state process} \]
\[ W(t) : q \text{- dimensional vector Wiener process (Brownian motion)} \]
\[ \mu(t) = \beta u(t) : p \text{- dimensional level process, } \beta \text{ a } p \times r \text{ regression matrix} \]
\[ A : p \times p \text{ dimensional pull matrix} \]
\[ \Sigma : p \times q \text{ dimensional diffusion matrix} \]

\[ dX(t) = (AX(t) + \mu(t))dt + \Sigma dW(t) \]

Itô solution:

\[ X(t) = B(t)^{-1} \left\{ X(0) + \int_0^t B(u)\mu(u)du + \int_0^t B(u)\Sigma dW(u) \right\} \]

\[ B(t) = e^{-At} = \sum_{n=0}^\infty A^n (-t)^n / n! = V^{-1} e^{-At} V \text{ (when existing)} \]

\[ A \text{ is a diagonal eigenvalue matrix of } A, V \text{ is the left eigenvector matrix.} \]
The individual size of the algae Coccolithus has evolved over time in the world oceans. What can we say about this evolution?

- How fast do the populations track its fitness optimum?
- Are the fitness optima the same/correlated across oceans, do they vary in concert?
- Does fitness depend on global temperature? How fast does fitness vary over time?
- Are there unmeasured processes influencing fitness?

Model variations:
- 1, 2 or 3 layers.
- Inclusion of external timeseries
- In a single layer:
  - Local or global parameters
  - Correlation between sites (inter-regional correlation)
  - Deterministic response to the lower layer
  - Random walk (no tracking)
Likelihood: Kalman filter

Need a linear, normal Markov chain with independent normal observations:

\[ X(t_1) \rightarrow X(t_2) \rightarrow \cdots \rightarrow X(t_n) \]

Process

\[ \downarrow \quad \downarrow \quad \downarrow \]

\[ Z_1 \quad Z_2 \quad Z_n \]

Observations

The Ito solution

\[
X(t) = B(t)^{-1} \left\{ X(0) + \int_0^t B(u) \mu(u) du + \int_0^t B(u) \Sigma dW(u) \right\}
\]

gives, together with measurement variances, what is needed to calculate the likelihood using the Kalman filter:

\[ E(z_i \mid z_1, \ldots, z_{i-1}) \equiv m_i \quad \text{and} \quad Var(z_i \mid z_1, \ldots, z_{i-1}) \equiv v_i \]

\[
f(z \mid \theta) = \prod_{i=1}^n f_N(z_i \mid m_i, v_i, \theta)
\]
Kalman smoothing (state estimation)
ML fit of a tracking model with 3 layers

North Atlantic
Red curve: expectancy
Black curve: realization
Green curve: uncertainty

Snapshot:
Inference

- Exact Gaussian likelihood, multi-modal

- Maximum likelihood by hill climbing from 50 starting points
  - BIC for model comparison.

- Bayesian:
  - Wide but informative prior distributions respecting identifying restrictions
  - MCMC (with parallel tempering) + Importance sampling (for model likelihood)
  - Bayes factor for model comparison and posterior probabilities
  - Posterior weight of a property C from posterior model probabilities

\[
W(C) = \frac{1}{n_C} \sum_{M \in C} f(data \mid M) \\
\sum_{C'} \frac{1}{n_{C'}} \sum_{M \in C'} f(data \mid M)
\]
Model inference concerns

Concerns:
- Identification restriction: increasing tracking speed up the layers
- In total: 723 models when equivalent models are pruned out

Enough data for model selection?
- Data Simulated from the ML fit of the original model
- Model selection over original model plus 25 likely suspects.
  - correct number of layers generally found with the Bayesian approach.
  - BIC seems too stingy on the number of layers.
**Results, original model**

3 layers, no regionality, no correlation

---

Table 1. Results from the original model (10). Maximum likelihood (ML), Bayesian posterior median (B.median), 95% prior and posterior credibility intervals. The prior credibility intervals for Δt parameters are given before the identification requirement Δt_i < Δt_{i+1} is imposed.

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>B.median</th>
<th>prior 95%</th>
<th>posterior 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper</td>
<td>Δt_1</td>
<td>200 kyr</td>
<td>140 kyr</td>
<td>3 yr - 430 kyr</td>
</tr>
<tr>
<td>layer</td>
<td>σ_1</td>
<td>0.21</td>
<td>0.23</td>
<td>0.080 - 4.3</td>
</tr>
<tr>
<td></td>
<td>Δt_2</td>
<td>1.9 My</td>
<td>1.03 My</td>
<td>36 yr - 4.2 My</td>
</tr>
<tr>
<td>layer</td>
<td>σ_2</td>
<td>0.072</td>
<td>0.087</td>
<td>0.018 - 3.7</td>
</tr>
<tr>
<td>Bottom</td>
<td>Δt_3</td>
<td>2.5 My</td>
<td>2.8 My</td>
<td>0.88 My - 14 My</td>
</tr>
<tr>
<td>layer</td>
<td>σ_3</td>
<td>0.075</td>
<td>0.080</td>
<td>0.026 - 0.17</td>
</tr>
<tr>
<td></td>
<td>μ</td>
<td>1.90</td>
<td>1.92</td>
<td>1.77 - 2.05</td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>0.052</td>
<td>0.046</td>
<td>-0.023 - 0.84</td>
</tr>
</tbody>
</table>
Bayesian inference on the 723 models

Table 2. Posterior weights for different properties. Number of models in parenthesis.

<table>
<thead>
<tr>
<th>Property</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Layers:</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>4.3% (19)</td>
</tr>
<tr>
<td>Regionality in:</td>
<td>none or $\mu_0$</td>
</tr>
<tr>
<td></td>
<td>0.15% (252)</td>
</tr>
<tr>
<td>Regional parameters in:</td>
<td>no layer</td>
</tr>
<tr>
<td></td>
<td>0.09% (252)</td>
</tr>
<tr>
<td>Inter-regional correlations:</td>
<td>none</td>
</tr>
<tr>
<td></td>
<td>19.0% (39)</td>
</tr>
<tr>
<td>Random walk in lowest layer:</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td>90.4% (393)</td>
</tr>
</tbody>
</table>

\[
W(C) = \frac{1}{\frac{1}{n_{C}} \sum_{M \in C} f(data | M)} \sum_{C'} \frac{1}{n_{C'}} \sum_{M \in C'} f(data | M)
\]
95% credibility bands for the 5 most probable models

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e^\mu$</td>
<td>7.32-7.55</td>
<td>7.32-7.55</td>
<td>7.31-7.55</td>
<td>7.31-7.54</td>
<td>7.31-7.55</td>
</tr>
<tr>
<td>$\Delta t_1$</td>
<td>0.0-8.1</td>
<td>0.0-11.6</td>
<td>0.0-13.2</td>
<td>0.0-9.7</td>
<td>0.0-8.4</td>
</tr>
<tr>
<td>$\Delta t_2$</td>
<td>0.1-49</td>
<td>0.2-54</td>
<td>0.2-50</td>
<td>0.1-32</td>
<td>0.2-68</td>
</tr>
<tr>
<td>$\Delta t_{525,3}$</td>
<td>0.6-4.0</td>
<td>0.5-3.1</td>
<td>0.6-3.5</td>
<td>0.6-3.3</td>
<td>0.6-4.2</td>
</tr>
<tr>
<td>$\Delta t_{612,3}$</td>
<td>0.00-0.46</td>
<td>0.00-0.34</td>
<td>0.00-0.46</td>
<td>0.00-0.43</td>
<td>0.00-0.62</td>
</tr>
<tr>
<td>$\Delta t_{516,3}$</td>
<td>0.5-4.7</td>
<td>0.5-5.0</td>
<td>0.5-4.8</td>
<td>0.5-4.2</td>
<td>0.5-5.5</td>
</tr>
<tr>
<td>$\Delta t_{752,3}$</td>
<td>0.5-11.8</td>
<td>0.5-6.6</td>
<td>0.5-9.6</td>
<td>0.5-9.3</td>
<td>0.5-11.8</td>
</tr>
<tr>
<td>$\Delta t_{806,3}$</td>
<td>0.00-0.16</td>
<td>0.00-0.18</td>
<td>0.00-0.16</td>
<td>0.00-0.15</td>
<td>0.00-0.20</td>
</tr>
<tr>
<td>$\Delta t_{982,3}$</td>
<td>1.3-97.0</td>
<td>1.3-89.2</td>
<td>1.4-96.3</td>
<td>1.3-83.8</td>
<td>1.4-109</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>0.01-6.9</td>
<td>0.01-6.0</td>
<td>0.01-5.4</td>
<td>0 (exact)</td>
<td>0.01-4.5</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td>0.02-3.9</td>
<td>0.02-3.3</td>
<td>0.04-4.6</td>
<td>0.35-7.9</td>
<td>0.06-4.1</td>
</tr>
<tr>
<td>$\sigma_3$</td>
<td>0.09-0.18</td>
<td>0.09-0.19</td>
<td>0.09-0.18</td>
<td>0.09-0.18</td>
<td>0.08-0.19</td>
</tr>
<tr>
<td>$\rho_1$</td>
<td>0 (exact)</td>
<td>0 (exact)</td>
<td>-0.19-0.96</td>
<td>0 (exact)</td>
<td>1 (exact)</td>
</tr>
<tr>
<td>$\rho_2$</td>
<td>0 (exact)</td>
<td>-0.19-0.96</td>
<td>0 (exact)</td>
<td>0 (exact)</td>
<td>0 (exact)</td>
</tr>
<tr>
<td>$\rho_3$</td>
<td>0.08-0.87</td>
<td>0.09-0.88</td>
<td>0.11-0.88</td>
<td>0.12-0.89</td>
<td>0.09-0.91</td>
</tr>
</tbody>
</table>
Summary

- Best 5 models in good agreement. (together, 19.7% of summed integrated likelihood):
  - Three layers.
  - Common expectancy in bottom layer.
  - No impact of exogenous temperature series.
  - Lowest layer: Inter-regional correlations, $\rho \approx 0.5$. Site-specific pull.
  - Middle layer: Intermediate tracking.
  - Upper layer: Very fast tracking.
Layered inference and inference uncertainty

Fig. 3. Process inference for Site 525, using the original model.
(a) 3-layered process inference with Bayesian parameter sampling. Solid line: upper layer, short dashed line: middle layer, long dashed line: lower layer.
(b) Upper layer ML inference. Solid line: mean, dashed lines: limits of a 95% confidence interval.
Phenotypic evolution on a phylogenetic tree: Body size of primates

Measurements: $i \in \{1, \ldots, 209\}$ 92 extant measurements, 117 fossils

$X_{i,1}(t) = \text{mean phenotype (log mass)}$

$X_{i,2}(t) = \text{fitness optimum}$

$X_{i,3}(t) = \text{underlying environment, primary optimum}$

$X_{i,3} \rightarrow X_{i,2} \rightarrow X_{i,1} \text{ same tracking model as for coccoliths}$

Phylogenetic tree:

$T_{i,j} = \text{Time when } i \text{ split from } j.$

$X_{i,t}(t) = X_{j,t}(t) \text{ when } t \leq T_{i,j}.$

Condition on these equalities for the given tree.

Observations: $Z_{i,k} = X_{i,k}(t_k) + N(0, \sigma_k^2)$ will be multi-normally distributed with mean vector and covariance matrix as parametric expression from the SDE model.
First results for some primates
Why linear SDE processes?

- Parsimonious: Simplest way of having a stochastic continuous time process that can track something else.
- Tractable: The likelihood, $L(\theta) \equiv f(\text{Data} | \theta)$, can be analytically calculated by the Kalman filter or directly by the parameterized multivariate normal model for the observations. ($\theta$ = model parameter set)
- Great flexibility, widely applicable...
Further comments

- Many processes evolve in continuous rather than discrete time. Thinking and modelling might then be more natural in continuous time.
- Tracking SDE processes with latent layers allow rather general correlation structure within and between time series. Inference is possible.
- Continuous time models allow related processes to be observed at variable frequencies (high-frequency data analyzed along with low frequency data).
- Endogeneity, regression structure, co-integration, non-stationarity, causal structure, seasonality…. are possible in SDE processes.
- Extensions to non-linear SDE models…
- SDE processes might be driven by non-Gaussian instantaneous stochasticity (jump processes).
Bibliography


Source codes, examples files and supplementary information can be found at http://folk.uio.no/trondr/layered/.