SUMO: WP 2 – year 2

Quality of local optima and global optimization methods

Wim Wiegerinck, Willem Burgers, Lasko Basnarkov, Ljupco Kocarev

20 November 2012
Contents

- Overview: Quality of local optima and global methods
- From small to large connection strengths
- Weather prediction versus climate projection and supermodeling
Goal of WP 2

Research and develop efficient, robust and scalable learning strategies to optimize supermodels for dynamical systems of low complexity (up to 1000 variables)
Learning connected supermodels result often in different connection matrices but with very similar dynamics.
- they all have large connections, and correspond to the same weighted supermodel (WP1)

Local optima: sometimes learning connected supermodels results in dynamics which is much less chaotic than expected (e.g. less chaotic than each of the imperfect models).
- These suboptimal supermodels often have small connections.
- What happens in the transition from small to large connections?

In more complex models, supermodeling improve on weather prediction but not always on climate.
- Why? Local optima?
- Global measures
- What to do? (Global optimization method?)
Motivation and approach

Motivation:
- If connections are too small, dynamics seems less chaotic than the imperfect models. Why?

Approach:
- Study the behavior of populations of oscillators as a function of the connection strength $\gamma$, (while ‘direction’ is fixed).

\[
\dot{x}_\mu = f(x_\mu, \theta_\mu) + \frac{\gamma}{M} \sum_{\nu \neq \mu} (x_\nu - x_\mu) \tag{1}
\]
From small to large connection strengths

Illustration: Lorenz 63, behaviour as function of $\gamma$

Time evolutions of the $x_{\mu}$-coordinates of a system of $M = 10$ connected of Lorenz 63 oscillators for different connection strengths $\gamma$. $\text{std}(\rho)/\rho = 0.1$
Small $\gamma$ approximation

Rewrite the dynamical equations for oscillator $\mu$ as

$$\dot{x}_\mu = f(x_\mu, \theta_\mu) + \frac{\gamma(M - 1)}{M}(a_\mu - x_\mu)$$

(2)

where $a_\mu$ is the average state of the other oscillators,

$$a_\mu = \frac{1}{M-1} \sum_{\nu \neq \mu} x_\nu.$$

Small $\gamma \rightarrow$ lowest order approximations

$\blacktriangleright$ $a_\mu \rightarrow$ ensemble mean state of $M - 1$ uncoupled oscillators.

$\blacktriangleright$ $\theta_\mu \rightarrow$ ensemble mean value $\theta$.

$$\dot{x} = f(x, \theta) + \frac{\gamma(M - 1)}{M}(a - x)$$

(3)

Coupling lead to a decay term!
Large \( \gamma \) approximation

Set

\[
\begin{align*}
\mathbf{x}_\mu &= \langle \mathbf{x}_\mu \rangle + \xi_\mu \equiv \mathbf{x} + \xi_\mu \quad (4) \\
\theta_\mu &= \langle \theta_\mu \rangle + \eta_\mu \equiv \theta + \eta_\mu \quad (5)
\end{align*}
\]

with the ensemble mean \( \langle \mathbf{x}_\mu \rangle = \frac{1}{M} \sum_\mu \mathbf{x}_\mu \).

Eqns for ensemble mean and covariances

\[
\begin{align*}
\Phi_{ai} &= \langle \eta_{a\mu} \xi_{i\mu} \rangle, \\
\Xi_{ij} &= \langle \xi_{i\mu} \xi_{j\mu} \rangle, \\
D_{ab} &= \langle \eta_{a\mu} \eta_{b\mu} \rangle
\end{align*}
\]

\[
\begin{align*}
\dot{\mathbf{x}} &= \mathbf{f} + \text{Tr}(\Phi \nabla_x \theta) \mathbf{f} + \frac{1}{2} \text{Tr}(\Xi \nabla \mathbf{x} \mathbf{x}) \mathbf{f} \\
\dot{\Phi} &= -\gamma \Phi + \Phi \nabla_x \mathbf{f} + D \nabla \theta \mathbf{f} \\
\dot{\Xi} &= -\gamma \Xi + \Xi \nabla_x \mathbf{f} + \Phi^T \nabla \theta \mathbf{f} + (\ldots)^T
\end{align*}
\]
From small to large connection strengths

Large $\gamma$ approximation versus $M = 5, 10, 50$

Lorenz 63. Mean and standard deviation taken over time of ensemble mean $x(t) \equiv \langle x_\mu(t) \rangle$. $t = 100$ time units, $t = 20$ time units transient. $h = 0.01$ Blue: $M = 5(\nabla)$, $M = 10(\triangle)$, $M = 50(\cdot)$. Red: large $\gamma$ approximation.
From small to large connection strengths

Large $\gamma$ approximation versus $M = 5, 10, 50$

Lorenz 63. Mean and standard deviation taken over time of variance $\xi^2(t) \equiv \langle x^2_\mu(t) \rangle - \langle x_\mu(t) \rangle^2$. $t = 100$ time units, $t = 20$ time units transient. $h = 0.01$ Blue: $M = 5(\bigtriangledown)$, $M = 10(\bigtriangleup)$, $M = 50(+)$ . Red: large $\gamma$ approximation.
Conclusion

- Regime transitions from independent chaotic behavior to synchronized chaos.
  - Uniformly connected Lorenz 63 oscillators.
    - Small $\gamma \rightarrow$ independent chaos.
    - Intermediate $\gamma \rightarrow$ strong damping.
    - Large $\gamma \rightarrow$ synchronization and average dynamics.
- Two approximations, one for small $\gamma$ and one for large $\gamma$.
  - Simulations showed the validity of these in the uniformly connected Lorenz 63 system.
The possibility of a non-smooth transition from independent chaotic behavior to synchronized chaos, should be taken into account when connected supermodels are optimized.

- Possibly one or more intermediate damping regimes.

The connection strengths of optimized supermodels should be sufficiently large so that the system is in the synchronized chaos regime.
Motivation and approach

Motivation:
▶ Real ground truth models are more complex than the real imperfect models
  - Unresolved processes.
▶ Perfect supermodeling cannot be assumed
▶ Fit on vectorfield does not imply fit on attractor
  - vectorfield = tendency
  - weather versus climate
▶ Attractor measures
▶ Attractor learning
Assumed ground truth: driven Lorenz 63

\[
\begin{align*}
\dot{x}_v &= \sigma(y_v - x_v) + \epsilon z_h, \\
\dot{y}_v &= x_v(\rho - z_v) - y_v, \\
\dot{z}_v &= x_v y_v - \beta z_v + \delta(x_h + \eta), \\
\dot{x}_h &= \sigma(y_h - x_h), \\
\dot{y}_h &= x_h(\rho - z_h) - y_h, \\
\dot{z}_h &= x_h y_h - \beta z_h.
\end{align*}
\]

(9)

with

\[
\sigma = 10, \ \rho = 28, \ \beta = 8/3, \ \epsilon = 1, \ \delta = 5, \ \eta = 2
\]

(10)

▶ v: visible states / observable variables
▶ h: hidden states / latent variables / unresolved scales
Imperfect models: Lorenz 63 with constant forcings

\[ \begin{align*}
\dot{x}_\mu &= \sigma_\mu(y - x) + \alpha_\mu \\
\dot{y}_\mu &= x(\rho_\mu - z) - y \\
\dot{z}_\mu &= xy - \beta_\mu z + \gamma_\mu
\end{align*} \]  (11, 12, 13)

The assumed parameter settings

<table>
<thead>
<tr>
<th></th>
<th>( \mu = 1 )</th>
<th>( \mu = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_\mu )</td>
<td>10</td>
<td>6.5</td>
</tr>
<tr>
<td>( \rho_\mu )</td>
<td>28</td>
<td>38</td>
</tr>
<tr>
<td>( \beta_\mu )</td>
<td>8/3</td>
<td>1.6</td>
</tr>
<tr>
<td>( \alpha_\mu )</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>( \gamma_\mu )</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>
Supermodeling

- Weighted supermodels
- Weighted combination of imperfect model vector fields
- Optimized on the basis of ground truth vector field.

Result:

<table>
<thead>
<tr>
<th>( \mu )</th>
<th>( \text{RMSE}_{vf-\text{Test}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu = 1 )</td>
<td>0.24 (0.01)</td>
</tr>
<tr>
<td>( \mu = 2 )</td>
<td>0.54 (0.02)</td>
</tr>
<tr>
<td>\text{sm}_vf</td>
<td>0.23 (0.01)</td>
</tr>
</tbody>
</table>
Climate: long free runs

Supermodeling deteriorates climate.
Long term statistics error: Wasserstein error

Wasserstein distance between two Gaussian distributions $\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_1, \Sigma_1)$:

$$W^2 = (\mu_1 - \mu_0)^2 + \text{Tr} \left( \Sigma_0 + \Sigma_1 - 2(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2} \right). \quad (14)$$

- Metric for two sets on the basis of their means and covariances
- Quality metric for models: use model to generate data sets and compare with observations
Wasserstein learning

Optimize $\mathbf{w}$ w.r.t. Wasserstein distance to training set of ground truth.

- Rough landscape $\rightarrow$ global optimization methods

<table>
<thead>
<tr>
<th></th>
<th>$RMSE_{vf-Test}$</th>
<th>$W_{Test}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu = 1$</td>
<td>0.24 (0.01)</td>
<td>17.4 (0.8)</td>
</tr>
<tr>
<td>$\mu = 2$</td>
<td>0.54 (0.02)</td>
<td>10.0 (0.5)</td>
</tr>
<tr>
<td>sm_vf</td>
<td>0.23 (0.01)</td>
<td>16.8 (0.8)</td>
</tr>
<tr>
<td>sm_W</td>
<td>0.27 (0.02)</td>
<td>3.1 (1.3)</td>
</tr>
</tbody>
</table>
| truth          | -                | 1.5 (0.6) }
Wasserstein learning

test model 1

test model 2

test sumo_vf

test sumo_ws
Conclusions

- With more complex models, sumo improves weather prediction.
- However, improved weather prediction (vectorfield) does not improve on climate (attractor).
  - Driven Lorenz 63
- Wasserstein metric as attractor measure.
  - based on mean and covariance.
  - other metrics can be used as well
- Attractor learning
  - attractor measure as objective
  - regularized by supermodel architecture!
  - fit on vectorfield as by product
  - requires global methods
- How important in practice?
  - Criteria?