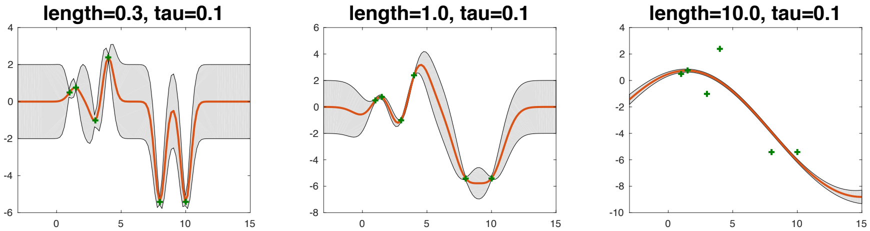
Data Analysis:
Statistical Modeling and Computation in Applications

Spatial and Environmental Data:
Model Selection and Long-range dependencies

Overview

- Gaussian Processes: Model selection
- Climate networks
- Degree distribution and connectivity
- Nonlinear relationships
- Dipole discovery

Which kernel function?



I could fit several models to my data – GPs with different kernels, kernels with different parameters (bandwidth, noise variance τ^2 etc) – which one is the most suitable?

Fitting covariance functions

many covariance functions have parameters θ, e.g. length scale ℓ

1. estimate generalization error: cross validation
   leave-one-out or k-fold

Fitting covariance functions

many covariance functions have parameters θ, e.g. length scale ℓ

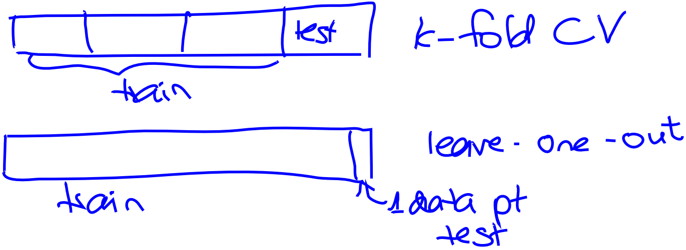
1. estimate generalization error: cross validation
   leave-one-out or k-fold

2. maximize log marginal likelihood of the data, $p(\mathbf{y}|\mathbf{X}, \theta)$ with respect to θ (e.g. $\theta = \ell$)

1. Leave-one-out cross validation

Given data $(x_1, y_1), \ldots, (x_N, y_N)$:

- For each i, remove (x_i, y_i) from the data and fit a GP to the rest $(\mathbf{X}_{-i}, \mathbf{y}_{-i})$.



1. Leave-one-out cross validation

Given data $(x_1, y_1), \ldots, (x_N, y_N)$:

- For each i, remove (x_i, y_i) from the data and fit a GP to the rest $(\mathbf{X}_{-i}, \mathbf{y}_{-i})$.
- Predict $\mu_{*|-i}$ and $\sigma_{*|-i}$. These depend on θ!

1. Leave-one-out cross validation

Given data $(x_1, y_1), \ldots, (x_N, y_N)$:

- For each i, remove (x_i, y_i) from the data and fit a GP to the rest $(\mathbf{X}_{-i}, \mathbf{y}_{-i})$.
- Predict $\mu_{*|-i}$ and $\sigma_{*|-i}$. These depend on θ!
- *What is the validation "loss"?*

1. Leave-one-out cross validation

Given data $(x_1, y_1), \ldots, (x_N, y_N)$:

- For each i, remove (x_i, y_i) from the data and fit a GP to the rest $(\mathbf{X}_{-i}, \mathbf{y}_{-i})$.
- Predict $\mu_{*|-i}$ and $\sigma_{*|-i}$. These depend on θ!
- *What is the validation "loss"?* compute predictive probability

$$\log p(y_i|\mathbf{X}, \mathbf{y}_{-i}, \theta) = -\frac{1}{2}\log\sigma^2_{*|-i} - \frac{(y_i - \mu_{*|-i})^2}{2\sigma^2_{*|-i}} - \frac{1}{2}\log 2\pi$$

$$= \log \frac{1}{\sqrt{2\pi\sigma_*^2}} \exp\left(-\frac{(y_i - \mu_{*|-i})^2}{2\sigma_*^2|-i}\right)$$

1. Leave-one-out cross validation

Given data $(x_1, y_1), \ldots, (x_N, y_N)$:

- For each i, remove (x_i, y_i) from the data and fit a GP to the rest $(\mathbf{X}_{-i}, \mathbf{y}_{-i})$.
- Predict $\mu_{*|-i}$ and $\sigma_{*|-i}$. These depend on θ!
- *What is the validation "loss"?* compute predictive probability

$$\log p(y_i|\mathbf{X}, \mathbf{y}_{-i}, \theta) = -\frac{1}{2}\log\sigma^2_{*|-i} - \frac{(y_i - \mu_{*|-i})^2}{2\sigma^2_{*|-i}} - \frac{1}{2}\log 2\pi$$

- choose parameters θ (e.g., bandwidth) that maximize log predictive probability

$$\max \quad \sum_{i=1}^{N} \log p(y_i|\mathbf{X}, \mathbf{y}_{-i}, \theta).$$

2. Maximize log marginal lielihood

- trades off data fit and complexity

2. Maximize log marginal lielihood

- trades off data fit and complexity
- maximize log marginal likelihood of the data, $p(\mathbf{y}|\mathbf{X}, \theta)$ with respect to θ (e.g. $\theta = \ell$)

$$\log p(\mathbf{y}|\mathbf{X}, \theta) = \underbrace{-\tfrac{1}{2}\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}} \quad \underbrace{-\tfrac{1}{2}\log|\mathbf{K}|} \quad - \tfrac{n}{2}\log 2\pi$$

$\mathcal{N}(\mu, K)$

$\mu = 0$

2. Maximize log marginal lielihood

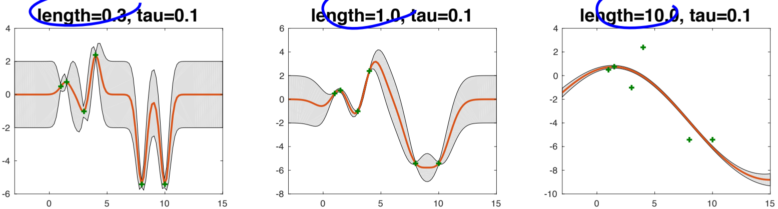
- trades off data fit and complexity
- maximize log marginal likelihood of the data, $p(\mathbf{y}|\mathbf{X}, \theta)$ with respect to θ (e.g. $\theta = \ell$)

$$\log p(\mathbf{y}|\mathbf{X}, \theta) = \underbrace{-\frac{1}{2}\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}_{\text{data fit}} \quad \underbrace{-\frac{1}{2}\log|\mathbf{K}|} \quad - \frac{n}{2}\log 2\pi$$

2. Maximize log marginal lielihood

- trades off data fit and complexity
- maximize log marginal likelihood of the data, $p(\mathbf{y}|\mathbf{X},\theta)$ with respect to θ (e.g. $\theta = \ell$)

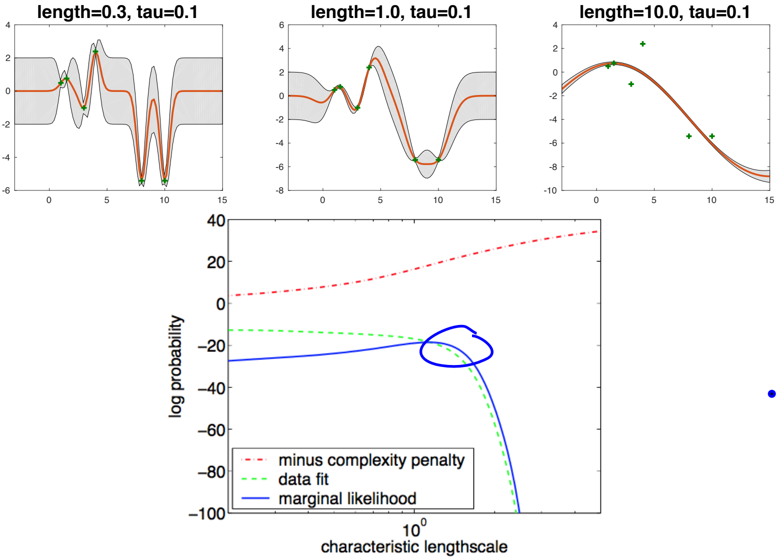
$$\log p(\mathbf{y}|\mathbf{X},\theta) = \underbrace{-\frac{1}{2}\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}_{\text{data fit}} \underbrace{-\frac{1}{2}\log|\mathbf{K}|}_{\text{complexity penalty}} - \frac{n}{2}\log 2\pi$$



RBF

lower plot: Rasmussen & Williams. Gaussian Processes for Machine Learning.

Data fit and complexity

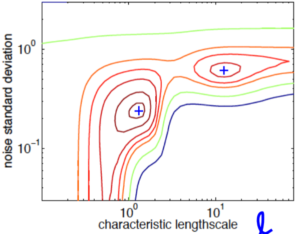


lower plot: Rasmussen & Williams. Gaussian Processes for Machine Learning.

Marginal likelihood

Maximum likelihood: $\log p(y|X,\ell,\tau) = -\frac{1}{2}y^\top K^{-1}y - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi$

τ



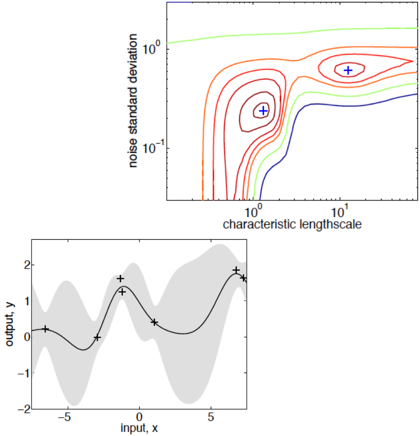
ℓ

Here, likelihood has 2 local maxima (2 "good" parameter settings (τ, ℓ)).

Source: Rasmussen & Williams. Gaussian Processes for Machine Learning.

Marginal likelihood

Maximum likelihood: $\log p(y|X, \ell, \tau) = -\frac{1}{2}y^\top K^{-1} y - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi$

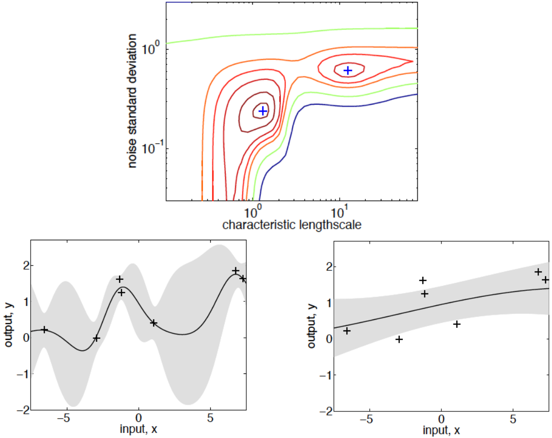


Here, likelihood has 2 local maxima (2 "good" parameter settings (τ, ℓ)).

Source: Rasmussen & Williams. Gaussian Processes for Machine Learning.

Marginal likelihood

Maximum likelihood: $\log p(y|X, \ell, \tau) = -\frac{1}{2} y^\top K^{-1} y - \frac{1}{2} \log |K| - \frac{n}{2} \log 2\pi$



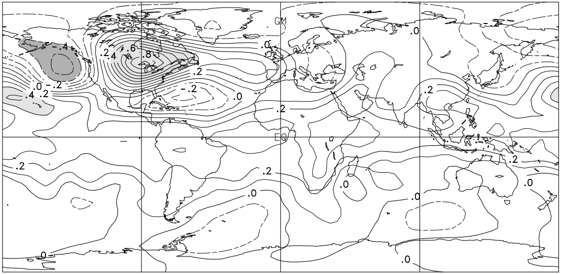
Here, likelihood has 2 local maxima (2 “good” parameter settings (τ, ℓ)).

Source: Rasmussen & Williams. Gaussian Processes for Machine Learning.

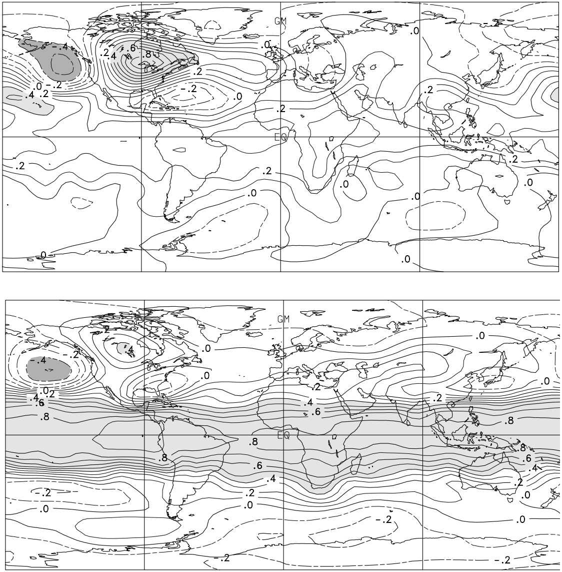
Overview

- Gaussian Processes: Model selection
- Climate networks
- Nonlinear relationships
- Dipole discovery

Long-range climate correlations



Long-range climate correlations



Climate networks

- Idea: build a graph $G = (V, E)$ to represent relationships!

Climate networks

- Idea: build a graph $G = (V, E)$ to represent relationships!
- **Construction:** based on e.g. sea surface temp. or pressure (hPa)

Climate networks

- Idea: build a graph $G = (V, E)$ to represent relationships!
- **Construction:** based on e.g. sea surface temp. or pressure (hPa)
  - nodes v_i: spatial grid points (2664), each has a time series.

Climate networks

- Idea: build a graph $G = (V, E)$ to represent relationships!
- **Construction:** based on e.g. sea surface temp. or pressure (hPa)
  - nodes v_i: spatial grid points (2664), each has a time series.
    $y'_{mt}(i)$: pressure in month m, year t at location i

Climate networks

- Idea: build a graph $G = (V, E)$ to represent relationships!
- **Construction:** based on e.g. sea surface temp. or pressure (hPa)
  - nodes v_i: spatial grid points (2664), each has a time series.
    $y'_{mt}(i)$: pressure in month m, year t at location i
  - edge (v_i, v_j) if time series $y(i)$, $y(j)$ are sufficiently correlated:
    $|r(y(i), y(j))| >$ threshold

- lag? 0

Climate networks

- Idea: build a graph $G = (V, E)$ to represent relationships!
- **Construction:** based on e.g. sea surface temp. or pressure (hPa)
  - nodes v_i: spatial grid points (2664), each has a time series.
    $y'_{mt}(i)$: pressure in month m, year t at location i
  - edge (v_i, v_j) if time series $y(i)$, $y(j)$ are sufficiently correlated:
    $|r(y(i), y(j))| >$ threshold
  - preprocessing: remove seasonal variation

$$y_{mt}(i) = \frac{y'_{mt}(i) - \bar{y'}_m(i)}{s(y'_m(i))}$$

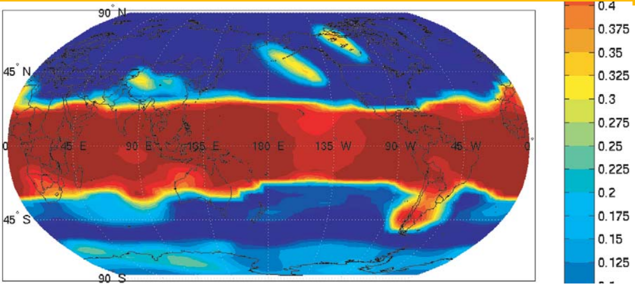
Climate networks

- Idea: build a graph $G = (V, E)$ to represent relationships!
- **Construction:** based on e.g. sea surface temp. or pressure (hPa)
  - nodes v_i: spatial grid points (2664), each has a time series.
    $y'_{mt}(i)$: pressure in month m, year t at location i
  - edge (v_i, v_j) if time series $y(i)$, $y(j)$ are sufficiently correlated:
    $|r(y(i), y(j))| > \text{threshold}$
  - preprocessing: remove seasonal variation

$$y_{mt}(i) = \frac{y'_{mt}(i) - \bar{y'}_m(i)}{s(y'_m(i))}$$

- **Network analysis**: degree distribution, clustering coefficient, centrality

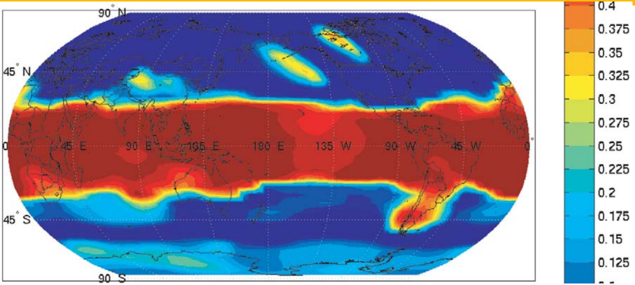
Climate network analysis *(Tsonis et al)*



fraction of total global area to which a geographical region is connected ("degree")

degree distributions: extratropical (30-65°N/S) and tropical (20°N-20°S) region

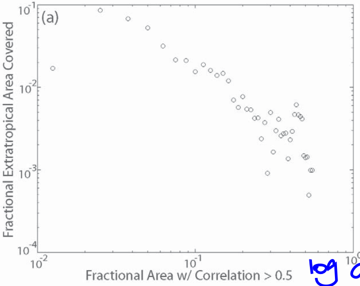
Climate network analysis *(Tsonis et al)*



fraction of total global area to which a geographical region is connected (“degree”)

log fraction

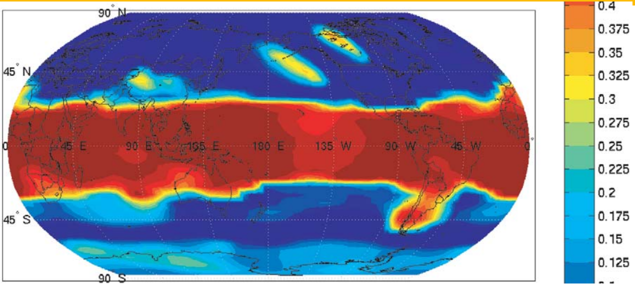
$p_k \approx k^{-8}$



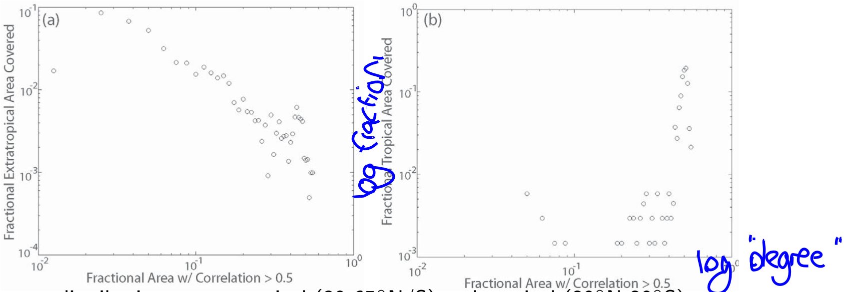
log degree

degree distributions: extratropical (30-65°N/S) and tropical (20°N-20°S) region

Climate network analysis *(Tsonis et al)*

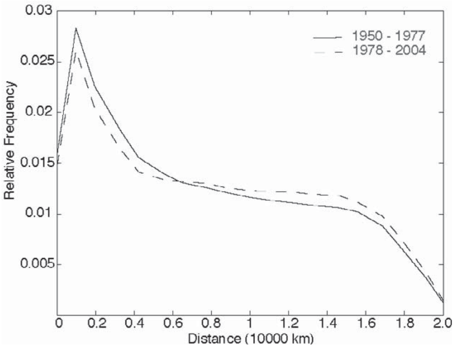


fraction of total global area to which a geographical region is connected ("degree")



degree distributions: extratropical (30-65°N/S) and tropical (20°N-20°S) region

Change of connectivity over time



Overview

- Gaussian Processes: Model selection
- Climate networks
- Nonlinear relationships
- Dipole discovery

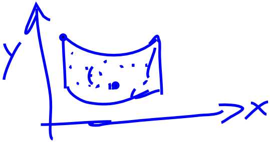
Climate networks: Nonlinear relationships *(Donges et al.)*

- correlation $\neq$ statistical dependence!

Climate networks: Nonlinear relationships *(Donges et al.)*

- correlation $\neq$ statistical dependence!
- *How could we capture nonlinear relations?*

Climate networks: Nonlinear relationships *(Donges et al.)*



- correlation $\neq$ statistical dependence!
- *How could we capture nonlinear relations?*
- alternative: *mutual information* $I(Y_1, Y_2) = H(Y_1) - H(Y_1|Y_2)$

Climate networks: Nonlinear relationships *(Donges et al.)*

- correlation $\neq$ statistical dependence!
- *How could we capture nonlinear relations?*
- alternative: *mutual information* $I(Y_1, Y_2) = H(Y_1) - H(Y_1|Y_2)$
- zero only if Y_1, Y_2 are independent

Climate networks: Nonlinear relationships *(Donges et al.)*

- correlation $\neq$ statistical dependence!
- *How could we capture nonlinear relations?*
- alternative: *mutual information* $I(Y_1, Y_2) = H(Y_1) - H(Y_1|Y_2)$
- zero only if Y_1, Y_2 are independent
- edge between v_i, v_j if $I(Y_i, Y_j) \geq \tau$

Quantifying uncertainty: (discrete) entropy

- discrete random variable Y, values in $\mathcal{Y} = \{1, \ldots m\}$:

Quantifying uncertainty: (discrete) entropy

- discrete random variable Y, values in $\mathcal{Y} = \{1, \ldots m\}$:

$$H(Y) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)}$$

Quantifying uncertainty: (discrete) entropy

- discrete random variable Y, values in $\mathcal{Y} = \{1, \dots m\}$:

$$H(Y) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)} = -\sum_{y \in \mathcal{Y}} p(y) \log p(y)$$

Quantifying uncertainty: (discrete) entropy

- discrete random variable Y, values in $\mathcal{Y} = \{1, \dots m\}$:

$$H(Y) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)} = -\sum_{y \in \mathcal{Y}} p(y) \log p(y)$$

- if $p(y) = \frac{1}{m}$ for all $y \in \mathcal{Y}$:

$$H(Y) = -\sum_{y} \frac{1}{m} \log \frac{1}{m}$$

$$= -\log \frac{1}{m} = \log m$$

Quantifying uncertainty: (discrete) entropy

- discrete random variable Y, values in $\mathcal{Y} = \{1, \dots m\}$:

$$H(Y) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)} = -\sum_{y \in \mathcal{Y}} p(y) \log p(y)$$

- if $p(y) = \frac{1}{m}$ for all $y \in \mathcal{Y}$: $H(Y) = \log m$.

Quantifying uncertainty: (discrete) entropy

- discrete random variable Y, values in $\mathcal{Y} = \{1, \ldots m\}$:

$$H(Y) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)} = -\sum_{y \in \mathcal{Y}} p(y) \log p(y)$$

- if $p(y) = \frac{1}{m}$ for all $y \in \mathcal{Y}$: $H(Y) = \log m$.
- if $p(1) = 1$ and $p(y) = 0$ for all $y \neq 1$: $H(Y) = 0 + 0 = 0$

Quantifying uncertainty: (discrete) entropy

- discrete random variable Y, values in $\mathcal{Y} = \{1, \ldots m\}$:

$$H(Y) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)} = -\sum_{y \in \mathcal{Y}} p(y) \log p(y)$$

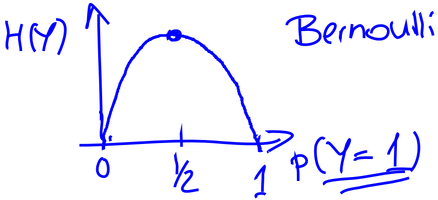
- if $p(y) = \frac{1}{m}$ for all $y \in \mathcal{Y}$: $H(Y) = \log m$.
- if $p(1) = 1$ and $p(y) = 0$ for all $y \neq 1$: $H(Y) = 0$.

Quantifying uncertainty: (discrete) entropy

- discrete random variable Y, values in $\mathcal{Y} = \{1, \dots m\}$:

$$H(Y) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)} = -\sum_{y \in \mathcal{Y}} p(y) \log p(y)$$

- if $p(y) = \frac{1}{m}$ for all $y \in \mathcal{Y}$: $H(Y) = \log m$.
- if $p(1) = 1$ and $p(y) = 0$ for all $y \neq 1$: $H(Y) = 0$.
- in general: $0 \leq H(Y) \leq \log m$



Quantifying uncertainty: (discrete) entropy

- discrete random variable Y, values in $\mathcal{Y} = \{1, \dots m\}$:

$$H(Y) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)} = -\sum_{y \in \mathcal{Y}} p(y) \log p(y)$$

- if $p(y) = \frac{1}{m}$ for all $y \in \mathcal{Y}$: $H(Y) = \log m$.
- if $p(1) = 1$ and $p(y) = 0$ for all $y \neq 1$: $H(Y) = 0$.
- in general: $0 \leq H(Y) \leq \log m$
- **Joint entropy:** 2 random variables Y, Z:

$$H(Y, Z) = -\sum_{\substack{y \in \mathcal{Y}, \\ z \in \mathcal{Z}}} p(y, z) \log p(y, z)$$

Quantifying uncertainty: (discrete) entropy

- discrete random variable Y, values in $\mathcal{Y} = \{1, \dots m\}$:

$$H(Y) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)} = -\sum_{y \in \mathcal{Y}} p(y) \log p(y)$$

- if $p(y) = \frac{1}{m}$ for all $y \in \mathcal{Y}$: $H(Y) = \log m$.
- if $p(1) = 1$ and $p(y) = 0$ for all $y \neq 1$: $H(Y) = 0$.
- in general: $0 \leq H(Y) \leq \log m$
- **Joint entropy:** 2 random variables Y, Z:

$$H(Y, Z) = -\sum_{\substack{y \in \mathcal{Y}, \\ z \in \mathcal{Z}}} p(y, z) \log p(y, z) \;\leq\; H(Y) + H(Z)$$

Quantifying uncertainty: (discrete) entropy

- discrete random variable Y, values in $\mathcal{Y} = \{1, \dots m\}$:

$$H(Y) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)} = -\sum_{y \in \mathcal{Y}} p(y) \log p(y)$$

- if $p(y) = \frac{1}{m}$ for all $y \in \mathcal{Y}$: $H(Y) = \log m$.
- if $p(1) = 1$ and $p(y) = 0$ for all $y \neq 1$: $H(Y) = 0$.
- in general: $0 \leq H(Y) \leq \log m$
- **Joint entropy:** 2 random variables Y, Z:

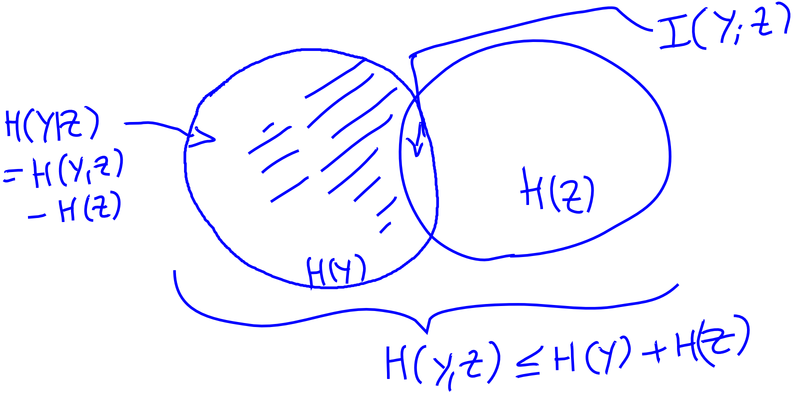
$$H(Y, Z) = -\sum_{\substack{y \in \mathcal{Y}, \\ z \in \mathcal{Z}}} p(y, z) \log p(y, z) \ \leq \ H(Y) + H(Z)$$

- **Conditional entropy:**

$$H(Y|Z) = H(Y, Z) - H(Z)$$

- reduction in uncertainty: how much does Z tell about Y?

$$I(Y;Z) = H(Y) - H(Y|Z)$$



Mutual Information

- reduction in uncertainty: how much does Z tell about Y?

$$I(Y;Z) = H(Y) - H(Y|Z) = H(Z) - H(Z|Y)$$

Mutual Information

- reduction in uncertainty: how much does Z tell about Y?

$$I(Y;Z) = H(Y) - H(Y|Z) = H(Z) - H(Z|Y)$$

- also captures *nonlinear* relationships
  $I(Y;Z) = 0$ if and only if Y, Z are independent.

Mutual Information

- reduction in uncertainty: how much does Z tell about Y?

$$I(Y;Z) = H(Y) - H(Y|Z) = H(Z) - H(Z|Y)$$

- also captures *nonlinear* relationships
  $I(Y;Z) = 0$ if and only if Y, Z are independent.
- $I(Z;Y) = I(Y;Z) \geq 0$ because
  "information never hurts": $H(Y) \geq H(Y|Z)$

Mutual Information

- reduction in uncertainty: how much does Z tell about Y?

$$I(Y;Z) = H(Y) - H(Y|Z) = H(Z) - H(Z|Y)$$

- also captures *nonlinear* relationships
  $I(Y;Z) = 0$ if and only if Y, Z are independent.
- $I(Z;Y) = I(Y;Z) \geq 0$ because
  "information never hurts": $H(Y) \geq H(Y|Z)$
- Computation:

Mutual Information

- reduction in uncertainty: how much does Z tell about Y?

$$
\begin{aligned}
I(Y;Z) = &\ H(Y) - H(Y|Z) \ = \ H(Z) - H(Z|Y) \\
= &\ H(Y) + H(Z) - H(Y,Z)
\end{aligned}
$$

- also captures *nonlinear* relationships
  $I(Y;Z) = 0$ if and only if Y, Z are independent.
- $I(Z;Y) = I(Y;Z) \geq 0$ because
  "information never hurts": $H(Y) \geq H(Y|Z)$

- Computation:

Mutual Information

- reduction in uncertainty: how much does Z tell about Y?

$$
\begin{aligned}
I(Y;Z) &= H(Y) - H(Y|Z) = H(Z) - H(Z|Y) \\
&= H(Y) + H(Z) - H(Y,Z)
\end{aligned}
$$

- also captures *nonlinear* relationships
  $I(Y;Z) = 0$ if and only if Y, Z are independent.
- $I(Z;Y) = I(Y;Z) \geq 0$ because
  "information never hurts": $H(Y) \geq H(Y|Z)$

- Computation:

$$
I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}
$$

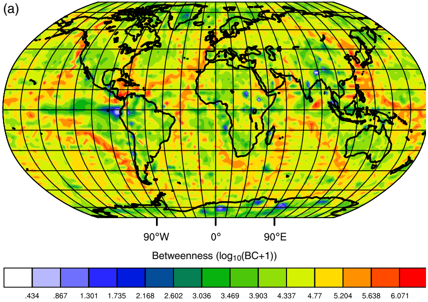
Distance from being independent

Climate networks: Nonlinear relationships *(Donges et al.)*

- correlation $\neq$ statistical dependence!
- alternative: *mutual information* $I(Y_1, Y_2) = H(Y_1) - H(Y_1|Y_2)$
- edge between v_i, v_j if $I(Y_i, Y_j) \geq \tau$

Climate networks: Nonlinear relationships *(Donges et al.)*

- correlation $\neq$ statistical dependence!
- alternative: *mutual information* $I(Y_1, Y_2) = H(Y_1) - H(Y_1 | Y_2)$
- edge between v_i, v_j if $I(Y_i, Y_j) \geq \tau$

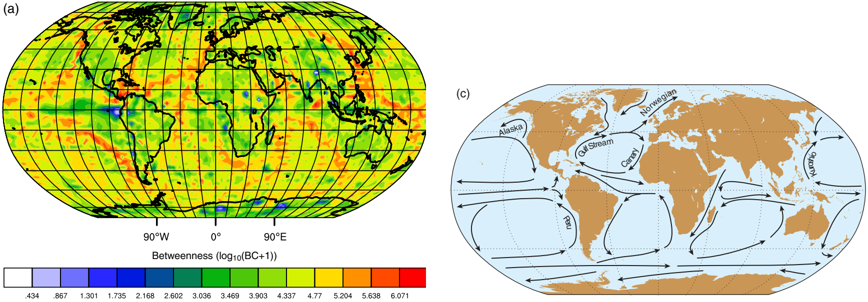


Betw. centrality (v) =

$\sum_{u,w}$ fraction of shortest paths between $u \rightsquigarrow w$ through (v)

Climate networks: Nonlinear relationships *(Donges et al.)*

- correlation $\neq$ statistical dependence!
- alternative: *mutual information* $I(Y_1, Y_2) = H(Y_1) - H(Y_1|Y_2)$
- edge between v_i, v_j if $I(Y_i, Y_j) \geq \tau$



Betweenness centrality "backbone" follows ocean surface currents!

Comments

- Mutual information vs correlation: nonlinear, but harder to estimate

Comments

- Mutual information vs correlation: nonlinear, but harder to estimate
- Check effect of your modeling choices (e.g., threshold for edges)

Comments

- Mutual information vs correlation: nonlinear, but harder to estimate
- Check effect of your modeling choices (e.g., threshold for edges)
- Significance testing

  - ...

Comments

- Mutual information vs correlation: nonlinear, but harder to estimate
- Check effect of your modeling choices (e.g., threshold for edges)
- Significance testing
  - null hypothesis: network is a random graph with given degree sequence
  - ...

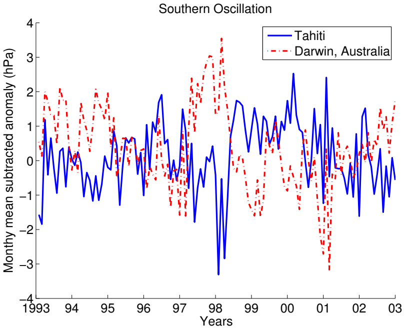
Comments

- Mutual information vs correlation: nonlinear, but harder to estimate
- Check effect of your modeling choices (e.g., threshold for edges)
- Significance testing
  - null hypothesis: network is a random graph with given degree sequence
  - scramble time series
  - ...

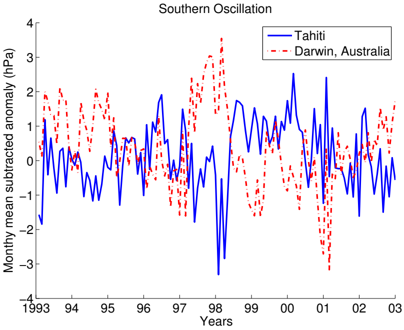
Overview

- Gaussian Processes: Model selection
- Climate networks
- Nonlinear relationships
- Dipole discovery

Dipoles, Oscillations, Teleconnections



Dipoles, Oscillations, Teleconnections



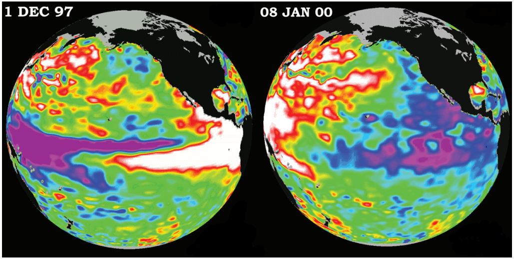
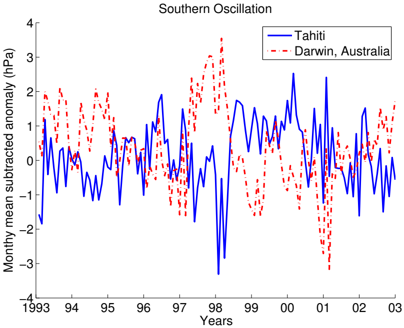


Figure 16.1. *SST anomalies (°C) observed under El Niño conditions (December, 1997; left) and La Niña conditions (January, 2000; right). Reprinted courtesy of NASA/JPL-Caltech.*



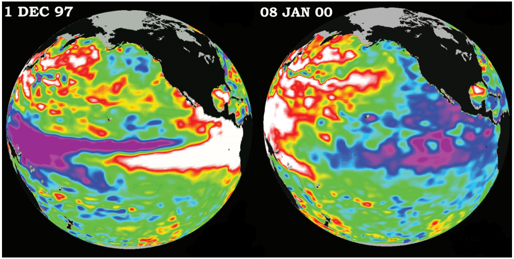


Figure 16.1. *SST anomalies (°C) observed under El Niño conditions (December, 1997; left) and La Niña conditions (January, 2000; right). Reprinted courtesy of NASA/JPL-Caltech.*

- global effects on temperature, rain
- El Niño: warm water in east Pacific, floods in Peru, dry in Australia/Indonesia
- La Niña: opposite
- Impact on economy, health, world events

Image sources: Kaper, H. Engler. Mathematics & Climate. Kawale et al, SDM 2011.

Dipole discovery *(Kawale et al.)*

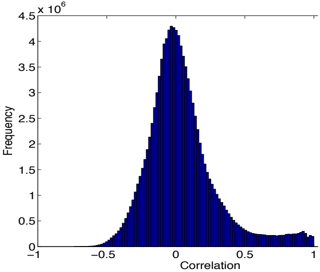
- Can we discover dipoles automatically?

Dipole discovery *(Kawale et al.)*

- Can we discover dipoles automatically?
- First: explore pairwise correlations

Dipole discovery *(Kawale et al.)*

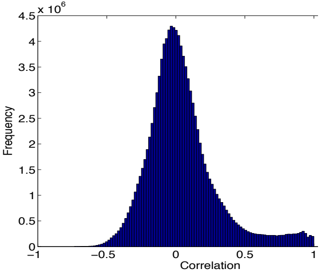
- Can we discover dipoles automatically?
- First: explore pairwise correlations



correlations

Dipole discovery *(Kawale et al.)*

- Can we discover dipoles automatically?
- First: explore pairwise correlations

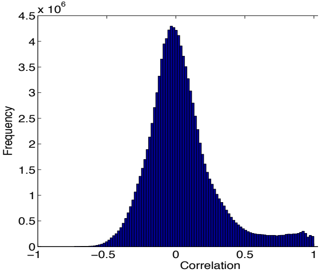


correlations

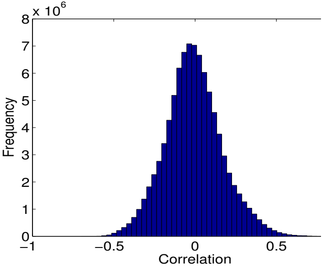
- treat positive & negative correlations differently

Dipole discovery *(Kawale et al.)*

- Can we discover dipoles automatically?
- First: explore pairwise correlations



correlations

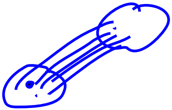


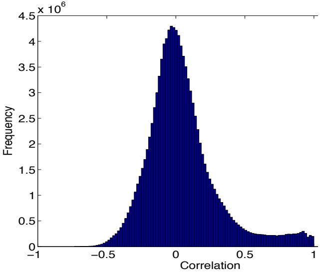
correlations for distance > 5000

- treat positive & negative correlations differently

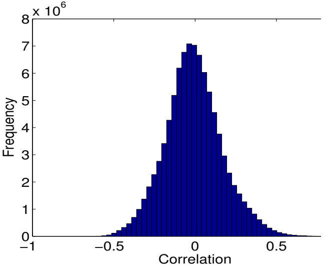
Dipole discovery *(Kawale et al.)*

- Can we discover dipoles automatically?
- First: explore pairwise correlations





correlations



correlations for distance > 5000

- treat positive & negative correlations differently
- reduction of search space: consistency in space

Procedure (roughly): (1)

Given: time series of pressure data (monthly mean) for 1948-2011.

- Preprocessing: smooth each time series by averaging over 3-month-windows; remove seasonality (as for climate networks)

Procedure (roughly): (1)

Given: time series of pressure data (monthly mean) for 1948-2011.

- Preprocessing: smooth each time series by averaging over 3-month-windows; remove seasonality (as for climate networks)
- Divide data into 9 periods of 20 years each, shifted by 5 years

Procedure (roughly): (1)

Given: time series of pressure data (monthly mean) for 1948-2011.

- Preprocessing: smooth each time series by averaging over 3-month-windows; remove seasonality (as for climate networks)
- Divide data into 9 periods of 20 years each, shifted by 5 years
- Build a network via correlations, but different thresholds for positive (0.85) and negative (0.4) correlations

Procedure (roughly): (2)

Criteria for dipole:

Dipole search:

Procedure (roughly): (2)

Criteria for dipole:

- two coherent regions in space

Dipole search:

Procedure (roughly): (2)

Criteria for dipole:

- two coherent regions in space
- negatively correlated

Dipole search:

Procedure (roughly): (2)

Criteria for dipole:

- two coherent regions in space
- negatively correlated
- far apart (long-range)

Dipole search:

Procedure (roughly): (2)

Criteria for dipole:

- two coherent regions in space
- negatively correlated
- far apart (long-range)

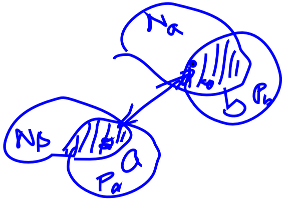
Dipole search:

1. Pick strongest negative edge (a, b).

Procedure (roughly): (2)

Criteria for dipole:

- two coherent regions in space
- negatively correlated
- far apart (long-range)



Dipole search:

1. Pick strongest negative edge (a, b).
2. Pick k most positive local neighbors of a and b each (gives neighborhoods P_a, P_b)

Procedure (roughly): (2)

Criteria for dipole:

- two coherent regions in space
- negatively correlated
- far apart (long-range)

Dipole search:

1. Pick strongest negative edge (a, b).
2. Pick k most positive local neighbors of a and b each (gives neighborhoods P_a, P_b)
3. Pick k farthest negative neighbors of a and b each (gives neighborhoods N_a, N_b)

Procedure (roughly): (2)

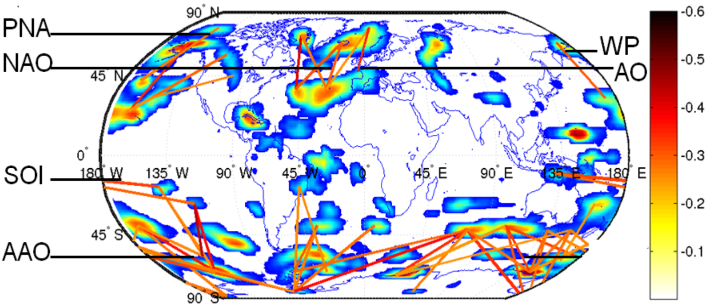
Criteria for dipole:

- two coherent regions in space
- negatively correlated
- far apart (long-range)

Dipole search:

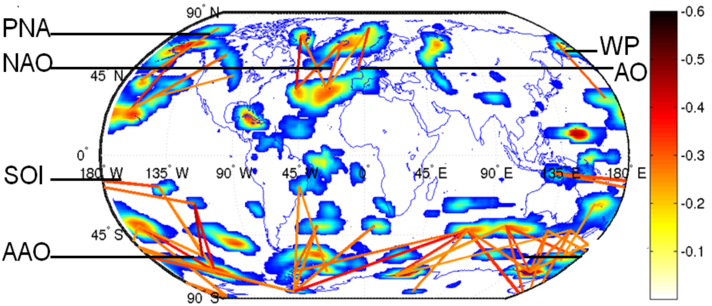
1. Pick strongest negative edge (a, b).
2. Pick k most positive local neighbors of a and b each (gives neighborhoods P_a, P_b)
3. Pick k farthest negative neighbors of a and b each (gives neighborhoods N_a, N_b)
4. Use $P_a \cap N_b$ and $P_b \cap N_a$ as dipole pair, if large enough.

Automatic Dipole Discovery: Results

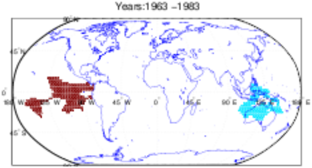


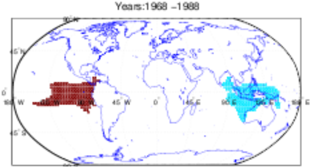
Credit: V. Kumar

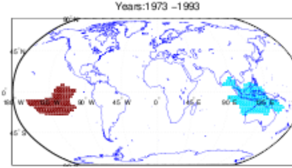
Automatic Dipole Discovery: Results



Credit: V. Kumar







Different phases of the Southern Oscillation (SOI), from pressure data

Summary

- Modeling short-range interactions: Gaussian Processes
  - Interpretation, prediction, experimental design
- Modeling & studying long-range interactions: network analysis (climate networks) on spatial grid
  - Edges via correlation or mutual information
  - Apply network analysis and time series tools from previous modules

References and further reading

- **Entropy, mutual information**: T.M. Cover, J.A. Thomas. Elements of Information Theory. Chapter 2.1–2.4.
- **Climate networks:** A.A. Tsonis, K. L. Swanson, and P. J. Roebber. *What do networks have to do with climate?*. Bulletin of the American Meteorological Society 87.5 (2006): 585–595.
- **Dipole Discovery:** J. Kawale, M. Steinbach, V. Kumar. *Discovering dynamic dipoles in climate data*. In SDM, 2011.

Image sources

- C. E. Rasmussen & C. K. I. Williams. Gaussian Processes for Machine Learning, 2006.
- A.A. Tsonis, K. L. Swanson, and P. J. Roebber. *What do networks have to do with climate?*. Bulletin of the American Meteorological Society 87.5 (2006): 585–595.
- J. F. Donges, Y. Zou, N. Marwan, J. Kurths. *The backbone of the climate network*. EPL (Europhysics Letters), 87(4), 48007, 2009
- H. Von Storch and F. Zwiers. *Statistical analysis in climate research*. Cambridge Univ Pr, 2002
- J. Kawale, M. Steinbach, V. Kumar. *Discovering dynamic dipoles in climate data*. In SDM, 2011.