Selecting Stochastic Models, Especially for Networks

Cosma Shalizi
Statistics Department, Carnegie Mellon University
Santa Fe Institute

7 May 2013

Why Isn't Model Selection Easy?

"How can $R^2 = 0.9$ *(***) be wrong"?* Fit all the different model classes you'd consider, see which matches the data best, then use that

Why Isn't Model Selection Easy?

"How can $R^2 = 0.9$ *(***) be wrong"?* Fit all the different model classes you'd consider, see which matches the data best, then use that

This is a *very bad idea*

Why Isn't Model Selection Easy?

"How can $R^2 = 0.9$ (***) be wrong"? Fit all the different model classes you'd consider, see which matches the data best, then use that

This is a *very bad idea*

$$\text{(Error in-sample)} = \text{(Error out-of-sample)} + \text{(memorizing noise)}$$

Picking the best in-sample leads to over-fitting; model will work badly

Why Isn't Model Selection Easy?

"How can $R^2 = 0.9$ *(***) be wrong"?* Fit all the different model classes you'd consider, see which matches the data best, then use that
This is a *very bad idea*

(Error in-sample) = (Error out-of-sample) + (memorizing noise)

Picking the best in-sample leads to over-fitting; model will work badly
If a stochastic model is correct, it *shouldn't* fit the data perfectly

Bias-variance trade-off: more flexible model classes can fit more processes
but they are also more sensitive to noise

Bias-variance trade-off: more flexible model classes can fit more processes
but they are also more sensitive to noise
Picking the model which best fits the training data will over-fit
Over-fitting gets worse as we allow for more flexible models

Bias-variance trade-off: more flexible model classes can fit more processes
but they are also more sensitive to noise
Picking the model which best fits the training data will over-fit
Over-fitting gets worse as we allow for more flexible models
Sensible model selection methods all try to estimate and control over-fitting

Bias-variance trade-off: more flexible model classes can fit more processes
but they are also more sensitive to noise
Picking the model which best fits the training data will over-fit
Over-fitting gets worse as we allow for more flexible models
Sensible model selection methods all try to estimate and control over-fitting
We have *very little* understanding how to do this for networks

Model Selection

Given: candidate model classes $\Theta_1, \Theta_2, \ldots$
data $z_1, z_2, \ldots z_n = z_{1:n}$
Unknown: the best model class Θ_{k^*}
Return: a guess $\widehat{k}$ as to k^*

Model Selection

Given: candidate model classes $\Theta_1, \Theta_2, \ldots$
data $z_1, z_2, \ldots z_n = z_{1:n}$
Unknown: the best model class Θ_{k^*}
Return: a guess $\widehat{k}$ as to k^*
Consistency: A good method is one which reliably selects the best model class:

$$\lim_{n\to\infty} \Pr\left(\widehat{k} \neq k^*\right) = 0$$

"Best model class" can mean

1. The one which contains the (best approximation to the) truth, the process which generated the data
2. The one which will predict best on new data

These are distinct concepts!

The true model can be in a Θ_k with so many free parameters that estimation is hopeless, and we get better predictions *from limited data* with a systematically wrong but more tractable model

Log-Likelihood/Relative Entropy

Go back to information theory:

$$L_n(\theta) = -n^{-1} \log \Pr\left(Z_{1:n} = z_{1:n}; \theta\right)$$

$=$ How well does θ let us compress the data?

Log-Likelihood/Relative Entropy

Go back to information theory:

$$L_n(\theta) = -n^{-1} \log \Pr\left(Z_{1:n} = z_{1:n}; \theta\right)$$

$=$ How well does θ let us compress the data?
$\mathbb{E}\left[L(\theta)\right] = \lambda(\theta) =$ (true source entropy) $+$ (relative entropy of θ)
Best model in Θ_k ("pseudo-truth") is

$$\theta^*_k = \underset{\theta \in \Theta_k}{\operatorname{argmin}}\, \lambda(\theta)$$

Sampling Fluctuations

Fluctuations:

$$L_n(\theta) = \lambda(\theta) + G_n(\theta)$$

with $\mathbb{E}\left[G_n(\theta)\right] = 0$
We *fit* the model by minimizing L, so

$$\widehat{\theta}_k = \underset{\theta \in \Theta_k}{\operatorname{argmin}} \lambda(\theta) + G_n(\theta)$$

This means that

$$\mathbb{E}\left[G_n(\widehat{\theta}_k)\right] < 0$$

The in-sample fit is *optimistic* about how well the model will do

Bias-Variance

a.k.a. approximation vs. estimation, sensitivity vs. stability

Bias-Variance

a.k.a. approximation vs. estimation, sensitivity vs. stability

As Θ_k gets bigger, $\mathbb{E}\left[L(\theta^*_k)\right]$ goes down

smaller systematic approximation error, less bias, more sensitive to true process

Bias-Variance

a.k.a. approximation vs. estimation, sensitivity vs. stability

As Θ_k gets bigger, $\mathbb{E}\left[L(\theta^*_k)\right]$ goes down

smaller systematic approximation error, less bias, more sensitive to true process

As Θ_k gets bigger, $G_n(\widehat{\theta_k})$ gets more and more negative

bigger estimation error, more over-fitting, less stable against noise

Bias-Variance

a.k.a. approximation vs. estimation, sensitivity vs. stability

As Θ_k gets bigger, $\mathbb{E}\left[L(\theta^*_k)\right]$ goes down

smaller systematic approximation error, less bias, more sensitive to true process

As Θ_k gets bigger, $G_n(\widehat{\theta_k})$ gets more and more negative

bigger estimation error, more over-fitting, less stable against noise

$\therefore$ picking the smallest error *across models* is just going to select the model most sensitive to fluctuations

Comparing Models In-Sample

$$L_n(\widehat{\theta}_k) = \min_{\theta \in \Theta_k} \lambda(\theta) + G_n(\theta)$$

If we could just see $\lambda(\theta) = \mathbb{E}\left[L(\theta)\right]$ we'd be set

Comparing Models In-Sample

$$L_n(\widehat{\theta}_k) = \min_{\theta \in \Theta_k} \lambda(\theta) + G_n(\theta)$$

If we could just see $\lambda(\theta) = \mathbb{E}\left[L(\theta)\right]$ we'd be set
Penalties: try to *add* something to L_n to undo optimism
Now select the model with the best penalized log-likelihood

Information Criteria

Akaike information criterion [1]:

$$AIC(\Theta_k) = L_n(\widehat{\theta}_k) + \frac{\dim(\Theta_k)}{n}$$

RULE: $\widehat{k} = \operatorname{argmin}_k AIC(\Theta_k)$

Information Criteria

Akaike information criterion [1]:

$$AIC(\Theta_k) = L_n(\widehat{\theta}_k) + \frac{\dim(\Theta_k)}{n}$$

RULE: $\widehat{k} = \mathrm{argmin}_k \, AIC(\Theta_k)$
Schwarz's "Bayesian" Information Criterion (BIC) [34]:

$$BIC(\Theta_k) = L_n(\widehat{\theta}_k) + \frac{\dim(\Theta_k)}{2} \frac{\log n}{n}$$

RULE: $\widehat{k} = \mathrm{argmin}_k \, BIC(\Theta_k)$

Information Criteria

Akaike information criterion [1]:

$$AIC(\Theta_k) = L_n(\widehat{\theta}_k) + \frac{\dim(\Theta_k)}{n}$$

RULE: $\widehat{k} = \operatorname{argmin}_k AIC(\Theta_k)$
Schwarz's "Bayesian" Information Criterion (BIC) [34]:

$$BIC(\Theta_k) = L_n(\widehat{\theta}_k) + \frac{\dim(\Theta_k)}{2}\frac{\log n}{n}$$

RULE: $\widehat{k} = \operatorname{argmin}_k BIC(\Theta_k)$
Many, many others: all of the form

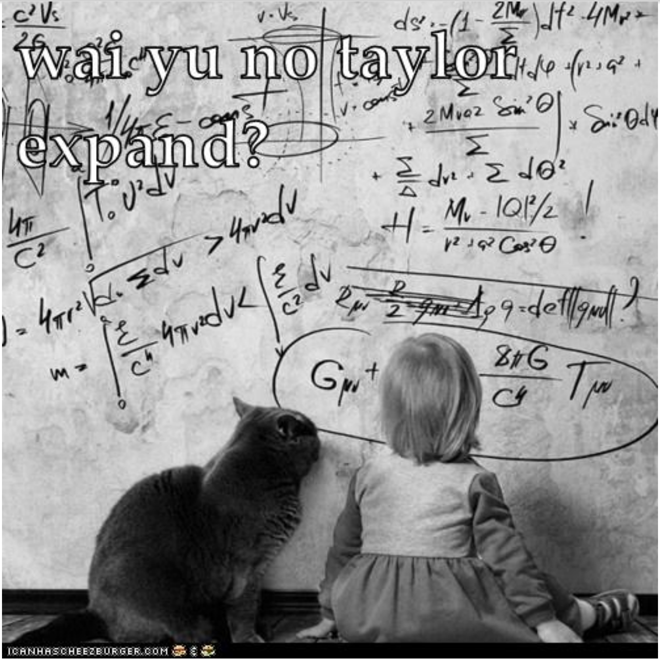
$$\widehat{k} = \operatorname*{argmin}_k L_n(\widehat{\theta}_k) + R_n(\Theta_k)$$

with $R_n = o_P(1)$

Origin Myths: AIC

$$AIC(\Theta_k) = L_n(\widehat{\theta_k}) + \frac{\dim(\Theta_k)}{n}$$

AIC is supposed to approximate $\mathbb{E}\left[\lambda(\widehat{\theta_k})\right]$
i.e., $n^{-1}\dim(\Theta_k)$ is supposed to approximate $G_n(\widehat{\theta_k})$
Why?


wai yu no taylor expand?
ICANHASCHEEZBURGER.COM

Classical Estimation Theory in One Slide

[20, 21, 23]

Classical Estimation Theory in One Slide

[20, 21, 23]

$$
\begin{aligned}
0 &= \nabla L_n(\widehat{\theta}_k) \\
&\approx \nabla L_n(\theta^*_k) + \nabla\nabla L_n(\theta^*_k)(\widehat{\theta}_k - \theta^*_k) \\
\widehat{\theta}_k &= \theta^*_k - \left(\nabla\nabla L_n(\theta^*_k)\right)^{-1} \nabla L_n(\theta^*_k)
\end{aligned}
$$

Classical Estimation Theory in One Slide

[20, 21, 23]

$$
\begin{aligned}
0 &= \nabla L_n(\widehat{\theta}_k) \\
&\approx \nabla L_n(\theta^*_k) + \nabla\nabla L_n(\theta^*_k)(\widehat{\theta}_k - \theta^*_k) \\
\widehat{\theta}_k &= \theta^*_k - (\nabla\nabla L_n(\theta^*_k))^{-1} \nabla L_n(\theta^*_k)
\end{aligned}
$$

$$
\begin{aligned}
L_n(\theta) &\rightarrow \lambda(\theta) \\
\nabla\nabla L_n(\theta^*_k) &\rightarrow \nabla\nabla\lambda(\theta^*_k) = \mathbf{i}_k \\
\nabla L_n(\theta^*_k) &\rightarrow 0 \\
\mathbb{V}\left[\nabla L_n(\theta^*_k)\right] &\rightarrow n^{-1}\mathbf{j}_k
\end{aligned}
$$

Classical Estimation Theory in One Slide

[20, 21, 23]

$$
\begin{aligned}
0 &= \nabla L_n(\widehat{\theta}_k) \\
&\approx \nabla L_n(\theta^*_k) + \nabla\nabla L_n(\theta^*_k)(\widehat{\theta}_k - \theta^*_k) \\
\widehat{\theta}_k &= \theta^*_k - \left(\nabla\nabla L_n(\theta^*_k)\right)^{-1} \nabla L_n(\theta^*_k)
\end{aligned}
$$

$$
\begin{aligned}
L_n(\theta) &\rightarrow \lambda(\theta) \\
\nabla\nabla L_n(\theta^*_k) &\rightarrow \nabla\nabla\lambda(\theta^*_k) = \mathbf{i}_k \\
\nabla L_n(\theta^*_k) &\rightarrow 0 \\
\mathbb{V}\left[\nabla L_n(\theta^*_k)\right] &\rightarrow n^{-1}\mathbf{j}_k
\end{aligned}
$$

$$
\begin{aligned}
\therefore \widehat{\theta}_k &\rightarrow \theta^*_k \\
\mathbb{V}\left[\widehat{\theta}_k\right] &\rightarrow n^{-1}\mathbf{i}_k^{-1}\mathbf{j}_k\mathbf{i}_k^{-1}
\end{aligned}
$$

How well will $\widehat{\theta}_k$ forecast? [1, 14] Taylor expand:

$$\begin{aligned}
\lambda(\widehat{\theta}_k) &\approx \lambda(\theta^*_k) + \frac{1}{2}\left\langle \widehat{\theta}_k - \theta^*_k | \mathbf{i}_k | \widehat{\theta}_k - \theta^*_k \right\rangle \\
\mathbb{E}\left[\lambda(\widehat{\theta}_k)\right] &\approx \lambda(\theta^*_k) + \frac{1}{2n}\mathrm{tr}\left(\mathbf{j}_k \mathbf{i}_k^{-1}\right)
\end{aligned}$$

How well will $\widehat{\theta}_k$ forecast? [1, 14] Taylor expand:

$$\begin{aligned}
\lambda(\widehat{\theta}_k) &\approx \lambda(\theta^*_k) + \frac{1}{2}\left\langle \widehat{\theta}_k - \theta^*_k | \mathbf{i}_k | \widehat{\theta}_k - \theta^*_k \right\rangle \\
\mathbb{E}\left[\lambda(\widehat{\theta}_k)\right] &\approx \lambda(\theta^*_k) + \frac{1}{2n}\operatorname{tr}\left(\mathbf{j}_k \mathbf{i}_k^{-1}\right)
\end{aligned}$$

Now Taylor-expand the other way:

$$\begin{aligned}
L_n(\theta^*_k) &\approx L_n(\widehat{\theta}_k) + \frac{1}{2}\left\langle \theta^*_k - \widehat{\theta}_k | \mathbf{i}_k | \theta^*_k - \widehat{\theta}_k \right\rangle \\
\lambda(\theta^*_k) &\approx \mathbb{E}\left[L_n(\widehat{\theta}_k)\right] + \frac{1}{2n}\operatorname{tr}\left(\mathbf{j}_k \mathbf{i}_k^{-1}\right)
\end{aligned}$$

How well will $\widehat{\theta}_k$ forecast? [1, 14] Taylor expand:

$$\begin{aligned}
\lambda(\widehat{\theta}_k) & \approx & \lambda(\theta^*_k) + \frac{1}{2}\left\langle \widehat{\theta}_k - \theta^*_k | \mathbf{i}_k | \widehat{\theta}_k - \theta^*_k \right\rangle \\
\mathbb{E}\left[\lambda(\widehat{\theta}_k)\right] & \approx & \lambda(\theta^*_k) + \frac{1}{2n}\operatorname{tr}\left(\mathbf{j}_k \mathbf{i}_k^{-1}\right)
\end{aligned}$$

Now Taylor-expand the other way:

$$\begin{aligned}
L_n(\theta^*_k) & \approx & L_n(\widehat{\theta}_k) + \frac{1}{2}\left\langle \theta^*_k - \widehat{\theta}_k | \mathbf{i}_k | \theta^*_k - \widehat{\theta}_k \right\rangle \\
\lambda(\theta^*_k) & \approx & \mathbb{E}\left[L_n(\widehat{\theta}_k)\right] + \frac{1}{2n}\operatorname{tr}\left(\mathbf{j}_k \mathbf{i}_k^{-1}\right)
\end{aligned}$$

$$\mathbb{E}\left[\lambda(\widehat{\theta}_k)\right] \approx \mathbb{E}\left[L_n(\widehat{\theta}_k)\right] + \frac{\operatorname{tr}\left(\mathbf{j}_k \mathbf{i}_k^{-1}\right)}{n}$$

An unbiased estimate:

$$L_n(\widehat{\theta}_k) + \frac{(\operatorname{tr} \mathbf{j}_k \mathbf{i}_k^{-1})}{n}$$

If the model is well-specified, $\mathbf{i} = \mathbf{j}$ (Fisher information equality)
$\Rightarrow \mathrm{tr}\left(\mathbf{j}_k \mathbf{i}_k^{-1}\right) = \dim \Theta_k$

If the model is well-specified, $\mathbf{i} = \mathbf{j}$ (Fisher information equality)
$\Rightarrow \mathrm{tr}\left(\mathbf{j}_k \mathbf{i}_k^{-1}\right) = \dim \Theta_k$
$\therefore$ AIC is an unbiased estimate of how well the model class works

If the model is well-specified, $\mathbf{i} = \mathbf{j}$ (Fisher information equality)
$\Rightarrow \mathrm{tr}\left(\mathbf{j}_k \mathbf{i}_k^{-1}\right) = \dim \Theta_k$
$\therefore$ AIC is an unbiased estimate of how well the model class works

Assuming fixed dimension, unique interior quadratic minima for L_n and λ, $O(1/\sqrt{n})$ gradient noise, proper specification of the model...

If the model is well-specified, $\mathbf{i} = \mathbf{j}$ (Fisher information equality)

$\Rightarrow \mathrm{tr}\left(\mathbf{j}_k \mathbf{i}_k^{-1}\right) = \dim \Theta_k$

$\therefore$ AIC is an unbiased estimate of how well the model class works

Assuming fixed dimension, unique interior quadratic minima for L_n and λ, $O(1/\sqrt{n})$ gradient noise, proper specification of the model...

Doesn't control the *variance* of the estimate

Origin Myths: BIC

Introduce a prior distribution $\rho(\theta)$ over Θ

Origin Myths: BIC

Introduce a prior distribution $\rho(\theta)$ over Θ
Marginal/integrated log-likelihood, alias free energy:

$$\mathcal{L}(\Theta_k) = \log \Pr\left(z_{1:n} | \theta \in \Theta_k\right) = \log \int_{\Theta_k} \Pr\left(z_{1:n}; \theta\right) \rho(\theta | \theta \in \Theta_k) d\theta$$

($e^{\mathcal{L}(\Theta_k)}$ is also called "evidence" for Θ_k, ratios between them "Bayes factors" [27])
RULE: $\widehat{k} = \operatorname{argmax} \mathcal{L}(\Theta_k)$

Calculating $\mathcal{L}(\Theta_k)$ is generally intractable
Taylor expand in the exponent ("Laplace approximation") [27, 14]:

$$\begin{aligned}
\mathcal{L}(\Theta_k) &\approx \log \int e^{-n\left[L_n(\widehat{\theta}_k) + \frac{1}{2}\left\langle \theta - \widehat{\theta}_k | \nabla\nabla L_n(\widehat{\theta}_k) | \theta - \widehat{\theta}_k \right\rangle\right]} \rho(\theta | \theta \in \Theta_k) d\theta \\
&= -nL_n(\widehat{\theta}_k) + \log \int e^{-\frac{n}{2}\left\langle \theta - \widehat{\theta}_k | \nabla\nabla L_n(\widehat{\theta}_k) | \theta - \widehat{\theta}_k \right\rangle} \rho(\theta | \theta \in \Theta_k) d\theta \\
&\approx -nL_n(\widehat{\theta}_k) - \frac{\dim(\Theta_k)}{2} \log n \\
&\quad + \frac{\dim(\Theta_k)}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{i}_k| + \log \rho(\widehat{\theta}_k | \theta \in \Theta_k)
\end{aligned}$$

Divide by $-n$, discard the $o(1)$ terms (constant, Hessian, prior) $\Rightarrow$ BIC

RULE: $\widehat{k} = \operatorname{argmax} \mathcal{L}(\Theta_k)$

Justification 1 [34]: $\rho(\theta \in \Theta_k | Z_{1:n} = z_{1:n}) \propto e^{\mathcal{L}(\Theta_k)} \rho(\theta \in \Theta_k)$, so this is the Bayesian solution

Objection 1: *Real* Bayesians don't select models

Will come back to this

Objection 2: The prior term $\rho(\theta \in \Theta_k)$ really matters!

Miller-Harrison example [30]: standard ("Dirichlet process") prior for Gaussian clusters fed data from $\mathcal{N}(0, 1)$ converges on at least *two* clusters

RULE: $\widehat{k} = \operatorname{argmax} \mathcal{L}(\Theta_k)$
Justification 2 [25]: With $\rho(\theta|\theta \in \Theta_k)$ diffuse, as $\dim(\Theta_k)$ grows, more of the prior mass goes on large parameter vectors $\|\theta\| \gg 0$, most of which are bad

RULE: $\widehat{k} = \operatorname{argmax} \mathcal{L}(\Theta_k)$
Justification 2 [25]: With $\rho(\theta|\theta \in \Theta_k)$ diffuse, as $\dim(\Theta_k)$ grows, more of the prior mass goes on large parameter vectors $\|\theta\| \gg 0$, most of which are bad
$\therefore$ average gets pulled down from the high-likelihood θ by their crazy relatives

RULE: $\widehat{k} = \operatorname{argmax} \mathcal{L}(\Theta_k)$
Justification 2 [25]: With $\rho(\theta|\theta \in \Theta_k)$ diffuse, as $\dim(\Theta_k)$ grows, more of the prior mass goes on large parameter vectors $\|\theta\| \gg 0$, most of which are bad
$\therefore$ average gets pulled down from the high-likelihood θ by their crazy relatives
As $n \rightarrow \infty$ the prior gets swamped (hopefully)

RULE: $\widehat{k} = \operatorname{argmax} \mathcal{L}(\Theta_k)$
Justification 2 [25]: With $\rho(\theta|\theta \in \Theta_k)$ diffuse, as $\dim(\Theta_k)$ grows, more of the prior mass goes on large parameter vectors $\|\theta\| \gg 0$, most of which are bad
$\therefore$ average gets pulled down from the high-likelihood θ by their crazy relatives
As $n \rightarrow \infty$ the prior gets swamped (hopefully)

Most of the volume of a high-dimensional hypersphere is ϵ-close to the surface

$\therefore$ diffuse high-dimensional priors are weird

The Truth About Information Criteria

- If the true process is in some Θ_k of ours, and the data are IID/regression/Markov/etc., BIC is consistent [14, 15]
- AIC is *not* consistent and will tend to over-fit even as $n \to \infty$ (no control of variance) [14]
- AIC can give better generalization error than BIC when the truth is infinite-dimensional [14]
- Nothing magical about the AIC and BIC penalties
- Even for estimating risk, number of parameters is not really what's wanted, unless model is well-behaved *and* well-specified

Cross-Validation

Generalization performance = expected error on new data from the same source
Fake this by pretending that some of your data is really new

Cross-Validation

Generalization performance = expected error on new data from the same source
Fake this by pretending that some of your data is really new
Algorithm:

- For $j = 1 : m$
  - Randomly divide z into z_{train_j} and z_{test_j}
  - For each Θ_k, estimate $\widehat{\theta}_{k,j}$ using only z_{train_j}
  - Calculate $L(\widehat{\theta}_{k,j}; z_{\mathrm{test}_j})$
- $CV(\Theta_k) = m^{-1} \sum_j L(\widehat{\theta}_{k,j}; z_{\mathrm{test}_j})$

Cross-Validation

Generalization performance = expected error on new data from the same source
Fake this by pretending that some of your data is really new
Algorithm:

- For $j = 1 : m$
  - Randomly divide z into z_{train_j} and z_{test_j}
  - For each Θ_k, estimate $\widehat{\theta}_{k,j}$ using only z_{train_j}
  - Calculate $L(\widehat{\theta}_{k,j}; z_{\mathrm{test}_j})$
- $CV(\Theta_k) = m^{-1} \sum_j L(\widehat{\theta}_{k,j}; z_{\mathrm{test}_j})$

RULE: $\widehat{k} = \operatorname{argmin} CV(\Theta_k)$

Leave-One-Out vs. Multi-Fold

How big should z_{test_j} be? How big should m be?

Leave-one-out CV: each testing set is 1 data point, $m = n$, use each point once

Multi-fold CV: fix m to set 5 or 10, use n/m points in each testing set, each point used once

Many more variants

On each “fold” j,

$$L(\widehat{\theta}_{k,j}; z_{\mathrm{test}_j}) \approx \mathbb{E}\left[L(\theta^*_k)\right]$$

so averaging them together seems reasonable, but it’s a correlated average

On each "fold" j,

$$L(\widehat{\theta}_{k,j}; z_{\mathrm{test}_j}) \approx \mathbb{E}\left[L(\theta_k^*)\right]$$

so averaging them together seems reasonable, but it's a correlated average

Can also write CV as a penalty method; penalty is random and data-dependent [3]

On each "fold" j,

$$L(\widehat{\theta}_{k,j}; z_{\mathrm{test}_j}) \approx \mathbb{E}\left[L(\theta^*_k)\right]$$

so averaging them together seems reasonable, but it's a correlated average

Can also write CV as a penalty method; penalty is random and data-dependent [3]

Leave-one-out CV penalty $\rightarrow n^{-1}\,\mathrm{tr}\left(\mathbf{j}_k \mathbf{i}_k^{-1}\right)$

$\therefore$ AIC is an asymptotic approximation to leave-one-out [14]

On each "fold" j,

$$L(\widehat{\theta}_{k,j}; z_{\mathrm{test}_j}) \approx \mathbb{E}\left[L(\theta^*_k)\right]$$

so averaging them together seems reasonable, but it's a correlated average
Can also write CV as a penalty method; penalty is random and data-dependent [3]
Leave-one-out CV penalty $\rightarrow n^{-1} \operatorname{tr}\left(\mathbf{j}_k \mathbf{i}_k^{-1}\right)$
$\therefore$ AIC is an asymptotic approximation to leave-one-out [14]
Multi-fold CV has a stronger penalty

On each "fold" j,

$$L(\widehat{\theta}_{k,j}; z_{\mathrm{test}_j}) \approx \mathbb{E}\left[L(\theta_k^*)\right]$$

so averaging them together seems reasonable, but it's a correlated average
Can also write CV as a penalty method; penalty is random and data-dependent [3]
Leave-one-out CV penalty $\rightarrow n^{-1} \operatorname{tr}(\mathbf{j}_k \mathbf{i}_k^{-1})$
$\therefore$ AIC is an asymptotic approximation to leave-one-out [14]
Multi-fold CV has a stronger penalty
Extremely reliable, robust, and practical; 5- or 10- fold CV is basically "industry standard"

On each "fold" j,

$$L(\widehat{\theta}_{k,j}; z_{\mathrm{test}_j}) \approx \mathbb{E}\left[L(\theta^*_k)\right]$$

so averaging them together seems reasonable, but it's a correlated average
Can also write CV as a penalty method; penalty is random and data-dependent [3]
Leave-one-out CV penalty $\rightarrow n^{-1}\,\mathrm{tr}\,(\mathbf{j}_k \mathbf{i}_k^{-1})$
$\therefore$ AIC is an asymptotic approximation to leave-one-out [14]
Multi-fold CV has a stronger penalty
Extremely reliable, robust, and practical; 5- or 10- fold CV is basically "industry standard"
Relies on dividing data into *independent* training/testing sets

Bootstrapping

Draw a bootstrap sample $\tilde{Z}$ of size n
Set

$$\tilde{\theta}_k = \operatorname*{argmin}_{\theta \in \Theta_k} \tilde{L}_n(\theta)$$

Bootstrap estimate of the optimism:

$$\tilde{\mathbb{E}}\left[L_n(\tilde{\theta}_k) - \tilde{L}_n(\tilde{\theta}_k)\right]$$

= penalty to apply to Θ_k
Closely related to CV: directly looking at generalizing from a sample to a whole ensemble

Capacity Control

If over-fitting is the problem, *control over-fitting*

$$G_n(\Theta_k) \equiv \sup_{\theta \in \Theta_k} |G_n(\theta)|$$

Capacity Control

If over-fitting is the problem, *control over-fitting*

$$G_n(\Theta_k) \equiv \sup_{\theta \in \Theta_k} |G_n(\theta)|$$

$G_n(\Theta_k)$ will vary with

- n

Capacity Control

If over-fitting is the problem, *control over-fitting*

$$G_n(\Theta_k) \equiv \sup_{\theta \in \Theta_k} |G_n(\theta)|$$

$G_n(\Theta_k)$ will vary with

- n
- Pointwise rate at which $|G_n(\theta)| \to 0$
  Large deviations / measure concentration (usually) says:
  $\Pr\left(|G_n(\theta)| > \epsilon\right) < c_1 e^{-c_2 n \epsilon^2}$

Capacity Control

If over-fitting is the problem, *control over-fitting*

$$G_n(\Theta_k) \equiv \sup_{\theta \in \Theta_k} |G_n(\theta)|$$

$G_n(\Theta_k)$ will vary with

- n
- Pointwise rate at which $|G_n(\theta)| \to 0$
  Large deviations / measure concentration (usually) says:
  $\Pr\left(|G_n(\theta)| > \epsilon\right) < c_1 e^{-c_2 n \epsilon^2}$
- Size (in some sense) of Θ_k

Capacity Control

If over-fitting is the problem, *control over-fitting*

$$G_n(\Theta_k) \equiv \sup_{\theta \in \Theta_k} |G_n(\theta)|$$

$G_n(\Theta_k)$ will vary with

- n
- Pointwise rate at which $|G_n(\theta)| \rightarrow 0$
  Large deviations / measure concentration (usually) says:
  $\Pr\left(|G_n(\theta)| > \epsilon\right) < c_1 e^{-c_2 n \epsilon^2}$
- Size (in some sense) of Θ_k

The number of *effectively* distinct θ in Θ_k grows with n

With $n = 5$ there are at most 32 distinguishable classifiers

Similarly, but scale-dependently, for regression, etc.

How rapidly does the number of distinguishable models grow with the amount of data?

- Exponentially-small error probabilities $\times$ polynomial number of models $\Rightarrow$ consistency
- Exponentially-small error probabilities $\times$ exponentially-large number of models $\Rightarrow$ trouble

How rapidly does the number of distinguishable models grow with the amount of data?

- Exponentially-small error probabilities $\times$ polynomial number of models $\Rightarrow$ consistency
- Exponentially-small error probabilities $\times$ exponentially-large number of models $\Rightarrow$ trouble

Capacity = growth rate in number of distinguishable models can be quantified in various ways

covering numbers, bracketing numbers, Vapnik-Chervonenkis dimension, Pollard pseudo-dimension, fat-shattering dimension, Rademacher complexity, ...

How rapidly does the number of distinguishable models grow with the amount of data?

- Exponentially-small error probabilities $\times$ polynomial number of models $\Rightarrow$ consistency
- Exponentially-small error probabilities $\times$ exponentially-large number of models $\Rightarrow$ trouble

Capacity = growth rate in number of distinguishable models can be quantified in various ways

covering numbers, bracketing numbers, Vapnik-Chervonenkis dimension, Pollard pseudo-dimension, fat-shattering dimension, Rademacher complexity, ...

Complexities of *model classes*, not of *process* modeled *or* of fitting

Complexity $\neq$ number of parameters

How rapidly does the number of distinguishable models grow with the amount of data?

- Exponentially-small error probabilities $\times$ polynomial number of models $\Rightarrow$ consistency
- Exponentially-small error probabilities $\times$ exponentially-large number of models $\Rightarrow$ trouble

Capacity = growth rate in number of distinguishable models can be quantified in various ways

covering numbers, bracketing numbers, Vapnik-Chervonenkis dimension, Pollard pseudo-dimension, fat-shattering dimension, Rademacher complexity, ...

Complexities of *model classes*, not of *process* modeled *or* of fitting

Complexity $\neq$ number of parameters

Different size measures lead to different bounds on $G_n(\Theta_k)$ [31]

Many bounds are distribution-free (though worst case) [38]

Stabilization

Capacity looks at the size of the whole model class, but we're not using all of that

Stabilization

Capacity looks at the size of the whole model class, but we're not using all of that

Stability of learning: how much does $\widehat{\theta}_k$ change if we perturb the data $z_1, \ldots z_n$ a little?

Stabilization

Capacity looks at the size of the whole model class, but we're not using all of that

Stability of learning: how much does $\widehat{\theta}_k$ change if we perturb the data $z_1, \ldots z_n$ a little?

Really, how much does $\lambda(\widehat{\theta}_k)$ change if we perturb the data a little?

Stabilization

Capacity looks at the size of the whole model class, but we're not using all of that

Stability of learning: how much does $\widehat{\theta}_k$ change if we perturb the data $z_1, \ldots z_n$ a little?

Really, how much does $\lambda(\widehat{\theta}_k)$ change if we perturb the data a little?

Uniform stability + law of large numbers for data $\Rightarrow$ bounds on over-fitting [10]

Stabilization

Capacity looks at the size of the whole model class, but we're not using all of that

Stability of learning: how much does $\widehat{\theta}_k$ change if we perturb the data $z_1, \ldots z_n$ a little?

Really, how much does $\lambda(\widehat{\theta}_k)$ change if we perturb the data a little?

Uniform stability + law of large numbers for data $\Rightarrow$ bounds on over-fitting [10]

Cross-validation or bootstrap $\approx$ stability control without math (or guarantees)

Structural Risk Minimization

[38, 29]
$B(\Theta_k, n)$ = your favorite learning-theory bound on over-fitting
RULE: $\widehat{k} = \operatorname{argmin} L(\widehat{\theta_k}) + B(\Theta_k, n)$

Structural Risk Minimization

[38, 29]
$B(\Theta_k, n)$ = your favorite learning-theory bound on over-fitting
RULE: $\widehat{k} = \operatorname{argmin} L(\widehat{\theta_k}) + B(\Theta_k, n)$

- Consistent, because model classes are penalized *directly* by how badly they could be over-fitting
- Tends to work well when it can be applied; major difficulty is getting suitable bounds

Method of Sieves

[24, 22, 37]
$\Theta_k \subset \Theta_{k+1}$
With n samples, estimate in $\Theta_{k(n)}$
Let $k(n) \rightarrow n$ as $n \rightarrow \infty$, but slowly
Handles the bias/variance trade-off as well
Examples: non-parametric smoothing methods for density estimation and regression

Other Model Selection Ideas

Selection tests (χ^2, Cox, Vuong): Test the hypothesis that $\lambda(\widehat{\theta}_k) > \lambda(\widehat{\theta}_j)$ [32, 39]

Other Model Selection Ideas

Selection tests (χ^2, Cox, Vuong): Test the hypothesis that $\lambda(\widehat{\theta}_k) > \lambda(\widehat{\theta}_j)$ [32, 39]

Encompassing: True model should predict pseudo-truth for other models, but not vice versa

Specification checking: Look for systematic errors, accept everything without them, hope that this confidence set shrinks

Bayesian Model Averaging

Posterior distribution over parameters, including models:

$$\rho(\theta|z_{1:n}) = \frac{\rho(\theta)\Pr\left(z_{1:n};\theta\right)}{\int_{\Theta}\rho(\theta')\Pr\left(z_{1:n};\theta'\right)d\theta'}$$

Bayesian Model Averaging

Posterior distribution over parameters, including models:

$$\rho(\theta|z_{1:n}) = \frac{\rho(\theta)\Pr\left(z_{1:n};\theta\right)}{\int_{\Theta}\rho(\theta')\Pr\left(z_{1:n};\theta'\right)d\theta'}$$

Posterior predictive distribution:

$$\Pr\left(z_{n+1}|z_{1:n};\rho\right) = \int_{\Theta}\Pr\left(z_{n+1}|z_{1:n};\theta\right)\rho(\theta|z_{1:n})d\theta$$

Bayesian Model Averaging

Posterior distribution over parameters, including models:

$$\rho(\theta|z_{1:n}) = \frac{\rho(\theta)\Pr\left(z_{1:n};\theta\right)}{\int_{\Theta}\rho(\theta')\Pr\left(z_{1:n};\theta'\right)d\theta'}$$

Posterior predictive distribution:

$$\Pr\left(z_{n+1}|z_{1:n};\rho\right) = \int_{\Theta}\Pr\left(z_{n+1}|z_{1:n};\theta\right)\rho(\theta|z_{1:n})d\theta$$

Never select a model, *always* keep a distribution

Bayesian Model Averaging

Posterior distribution over parameters, including models:

$$\rho(\theta|z_{1:n}) = \frac{\rho(\theta)\Pr\left(z_{1:n};\theta\right)}{\int_{\Theta}\rho(\theta')\Pr\left(z_{1:n};\theta'\right)d\theta'}$$

Posterior predictive distribution:

$$\Pr\left(z_{n+1}|z_{1:n};\rho\right) = \int_{\Theta}\Pr\left(z_{n+1}|z_{1:n};\theta\right)\rho(\theta|z_{1:n})d\theta$$

Never select a model, *always* keep a distribution

- Prior ρ biases towards certain θ, but reduces variance
- Smoothing effect: error of the posterior = (weighted average error of each θ) - (weighted diversity of θs' predictions)

Consistency of Bayesian Model Averaging

Bayesian learning $\equiv$ replicator dynamics for θ (fitness = likelihood) [35]

Consistency of Bayesian Model Averaging

Bayesian learning $\equiv$ replicator dynamics for θ (fitness = likelihood) [35]

Bayesian learning is *not* generally consistent [17, 13] or even convergent [5, 6]

Need to build constraints like a sieve into the prior [4, 35]

Consistency of Bayesian Model Averaging

Bayesian learning $\equiv$ replicator dynamics for θ (fitness = likelihood) [35]
Bayesian learning is *not* generally consistent [17, 13] or even convergent [5, 6]
Need to build constraints like a sieve into the prior [4, 35]
Then there's a large deviations principle [35]

$$\log \rho(\theta \in A|z_{1:n}) \approx -n \left[\inf_{\theta \in A} \lambda(\theta) - \inf_{\theta' \in \Theta} \lambda(\theta') \right] + O_p(n^{1/2})$$

Consistency of Bayesian Model Averaging

Bayesian learning $\equiv$ replicator dynamics for θ (fitness = likelihood) [35]
Bayesian learning is *not* generally consistent [17, 13] or even convergent [5, 6]
Need to build constraints like a sieve into the prior [4, 35]
Then there's a large deviations principle [35]

$$\log \rho(\theta \in A | z_{1:n}) \approx -n \left[\inf_{\theta \in A} \lambda(\theta) - \inf_{\theta' \in \Theta} \lambda(\theta') \right] + O_p(n^{1/2})$$

Predictive consistency if truth is in the prior:

$$\Pr\left(z_{n+1} | z_{1:n}; \rho\right) \rightarrow \Pr\left(z_{n+1} | z_{1:n}\right)$$

ρ need *not* concentrate on the true model even when that's available

Other Forms of Model Averaging

Mixture-of-experts: average predictions of all $\widehat{\theta}_k$ with weights $\propto e^{-nL_n(\widehat{\theta}_k)}$

Can give strong bounds on regret [12]

Other Forms of Model Averaging

Mixture-of-experts: average predictions of all $\widehat{\theta}_k$ with weights $\propto e^{-nL_n(\widehat{\theta}_k)}$

Can give strong bounds on regret [12]

Bagging: draw many bootstrap samples, fit different model to each, average their predictions [11]

Other Forms of Model Averaging

Mixture-of-experts: average predictions of all $\widehat{\theta}_k$ with weights $\propto e^{-nL_n(\widehat{\theta}_k)}$

Can give strong bounds on regret [12]

Bagging: draw many bootstrap samples, fit different model to each, average their predictions [11]

Boosting: fit a model, then do a weighted bootstrap with more weight on ill-fit points, repeat; average all models [33]

Other Forms of Model Averaging

Mixture-of-experts: average predictions of all $\widehat{\theta}_k$ with weights $\propto e^{-nL_n(\widehat{\theta}_k)}$

Can give strong bounds on regret [12]

Bagging: draw many bootstrap samples, fit different model to each, average their predictions [11]

Boosting: fit a model, then do a weighted bootstrap with more weight on ill-fit points, repeat; average all models [33]

etc., etc.

Model Averaging and Over-Fitting

Model ensembles are effectively *huge* models

Model Averaging and Over-Fitting

Model ensembles are effectively *huge* models
Gain in performance because of smoothing/diversity

Model Averaging and Over-Fitting

Model ensembles are effectively *huge* models
Gain in performance because of smoothing/diversity
Avoids over-fitting because the ensemble is very stable [19]

Prediction

$\lambda(\theta) = \mathbb{E}\left[L(\theta)\right]$ is how well, on average, θ will fit new data

Prediction

$\lambda(\theta) = \mathbb{E}[L(\theta)]$ is how well, on average, θ will fit new data
Question: Expectation *over what?* What's "new data"?

Prediction

$\lambda(\theta) = \mathbb{E}[L(\theta)]$ is how well, on average, θ will fit new data
Question: Expectation *over what?* What's "new data"?
Answer: depends on the kind of process producing the data

IID Processes

$Z_1, Z_2, \ldots Z_n, \ldots$ all independent
$\Pr(Z_{n+1}|Z_{1:n}) = \Pr(Z_{n+1})$

IID Processes

$Z_1, Z_2, \ldots Z_n, \ldots$ all independent
$\Pr(Z_{n+1} | Z_{1:n}) = \Pr(Z_{n+1})$
$\therefore$ expectation over Z_{n+1} = expectation over Z_1
and expectation over Z_{n+1} is constant

IID Processes

$Z_1, Z_2, \ldots Z_n, \ldots$ all independent
$\Pr(Z_{n+1}|Z_{1:n}) = \Pr(Z_{n+1})$
$\therefore$ expectation over Z_{n+1} = expectation over Z_1
and expectation over Z_{n+1} is constant

$$\begin{aligned} \mathbb{E}[L_n(\theta)] &= -\mathbb{E}\left[n^{-1}\sum_{i=1}^{n}\log \Pr(Z_i;\theta)\right] \\ &= -\mathbb{E}[\log \Pr(Z_1;\theta)] = \lambda(\theta) \end{aligned}$$

IID Processes

$Z_1, Z_2, \ldots Z_n, \ldots$ all independent
$\Pr(Z_{n+1}|Z_{1:n}) = \Pr(Z_{n+1})$
$\therefore$ expectation over Z_{n+1} = expectation over Z_1
and expectation over Z_{n+1} is constant

$$\begin{aligned} \mathbb{E}[L_n(\theta)] &= -\mathbb{E}\left[n^{-1}\sum_{i=1}^{n}\log\Pr(Z_i;\theta)\right] \\ &= -\mathbb{E}[\log\Pr(Z_1;\theta)] = \lambda(\theta) \end{aligned}$$

"New data" = new sample from the unchanging distribution

Markov Processes

Time series: $Z_1, Z_2, \ldots Z_n, \ldots$ sequential, dependent on last k steps

$$\Pr\left(Z_{n+1} | Z_1, \ldots Z_n\right) = \Pr\left(Z_{n+1} | Z_{n-(k-1):n}\right) \neq \Pr\left(Z_{n+1}\right)$$

Markov Processes

Time series: $Z_1, Z_2, \ldots Z_n, \ldots$ sequential, dependent on last k steps

$$\Pr\left(Z_{n+1}|Z_1, \ldots Z_n\right) = \Pr\left(Z_{n+1}|Z_{n-(k-1):n}\right) \neq \Pr\left(Z_{n+1}\right)$$

$\therefore$ expectation over Z_{n+1} is *not* expectation over Z_1
and expectation over Z_{n+1} fluctuates

Markov Processes

Time series: $Z_1, Z_2, \ldots Z_n, \ldots$ sequential, dependent on last k steps

$$\Pr\left(Z_{n+1}|Z_1, \ldots Z_n\right) = \Pr\left(Z_{n+1}|Z_{n-(k-1):n}\right) \neq \Pr\left(Z_{n+1}\right)$$

$\therefore$ expectation over Z_{n+1} is *not* expectation over Z_1
and expectation over Z_{n+1} fluctuates
Ergodicity:

$$\begin{aligned} \mathbb{E}\left[L_n(\theta)\right] &= -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log\Pr\left(Z_i|Z_{i-k:i-1};\theta\right)\right] \\ \rightarrow & \quad -\mathbb{E}\left[\frac{1}{n-k}\sum_{i=k+1}^{n}\log\Pr\left(Z_i|Z_{i-k:i-1};\theta\right)\right] \rightarrow \lambda(\theta) \end{aligned}$$

Markov Processes

Time series: $Z_1, Z_2, \ldots Z_n, \ldots$ sequential, dependent on last k steps

$$\Pr\left(Z_{n+1} | Z_1, \ldots Z_n\right) = \Pr\left(Z_{n+1} | Z_{n-(k-1):n}\right) \neq \Pr\left(Z_{n+1}\right)$$

$\therefore$ expectation over Z_{n+1} is *not* expectation over Z_1
and expectation over Z_{n+1} fluctuates
Ergodicity:

$$\begin{aligned} \mathbb{E}\left[L_n(\theta)\right] &= -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log \Pr\left(Z_i | Z_{i-k:i-1};\theta\right)\right] \\ \to & \quad -\mathbb{E}\left[\frac{1}{n-k}\sum_{i=k+1}^{n}\log \Pr\left(Z_i | Z_{i-k:i-1};\theta\right)\right] \to \lambda(\theta) \end{aligned}$$

"New data" = future of the series, averaging over blocks of length $k+1$

More General Time Series

Dependence goes all the way back to the beginning

More General Time Series

Dependence goes all the way back to the beginning
Can still hope for ergodicity:

$$\mathbb{E}\left[L_n(\theta)\right] = -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log\Pr\left(Z_i | Z_{1:i-1};\theta\right)\right] \rightarrow \lambda(\theta)$$

More General Time Series

Dependence goes all the way back to the beginning
Can still hope for ergodicity:

$$\mathbb{E}\left[L_n(\theta)\right] = -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log \Pr\left(Z_i|Z_{1:i-1};\theta\right)\right] \rightarrow \lambda(\theta)$$

This requires averaging over the *whole* future of the process

More General Time Series

Dependence goes all the way back to the beginning
Can still hope for ergodicity:

$$\mathbb{E}\left[L_n(\theta)\right] = -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log \Pr\left(Z_i | Z_{1:i-1};\theta\right)\right] \rightarrow \lambda(\theta)$$

This requires averaging over the *whole* future of the process
Having a limit here relies on weak dependence: the past matters less and less further and further into the future

Spatial Processes

No longer a sequence

Spatial Processes

No longer a sequence
Look at larger and larger spatial domains, say size of a is $|a|$

Spatial Processes

No longer a sequence
Look at larger and larger spatial domains, say size of a is $|a|$
Hope for spatial ergodicity: as $|a| \rightarrow \infty$,

$$-\mathbb{E}\left[|a|^{-1} \log \Pr\left(Z_a; \theta\right)\right] \rightarrow \lambda(\theta)$$

Spatial Processes

No longer a sequence
Look at larger and larger spatial domains, say size of a is $|a|$
Hope for spatial ergodicity: as $|a| \to \infty$,

$$-\mathbb{E}\left[|a|^{-1}\log \Pr\left(Z_a;\theta\right)\right] \to \lambda(\theta)$$

"New data" = parts of space not included in the old domain
Again, need weak dependence

The Role of Weak Dependence

"Blocking": divide $Z_1, Z_2, \ldots Z_n$ into μ blocks of size a

The Role of Weak Dependence

"Blocking": divide $Z_1, Z_2, \ldots Z_n$ into μ blocks of size a
Weak dependence $\Rightarrow$ if a is big, blocks are nearly independent

The Role of Weak Dependence

"Blocking": divide $Z_1, Z_2, \ldots Z_n$ into μ blocks of size a
Weak dependence $\Rightarrow$ if a is big, blocks are nearly independent
$\therefore$ Approximate distribution of $Z_{1:n}$ with μ IID copies of $Z_{1:a}$

The Role of Weak Dependence

"Blocking": divide $Z_1, Z_2, \ldots Z_n$ into μ blocks of size a
Weak dependence $\Rightarrow$ if a is big, blocks are nearly independent
$\therefore$ Approximate distribution of $Z_{1:n}$ with μ IID copies of $Z_{1:a}$
Want $a \rightarrow \infty$ to make approximation tight
Want $\mu \rightarrow \infty$ to use asymptotics

The Role of Weak Dependence

"Blocking": divide $Z_1, Z_2, \ldots Z_n$ into μ blocks of size a
Weak dependence $\Rightarrow$ if a is big, blocks are nearly independent
$\therefore$ Approximate distribution of $Z_{1:n}$ with μ IID copies of $Z_{1:a}$
Want $a \to \infty$ to make approximation tight
Want $\mu \to \infty$ to use asymptotics
Effective sample size $\mu = O(n)$, rate varying with dependence range
[40, 28, 16]

What About Networks?

Distinguish between prediction *on* network and prediction *of* network

"On" is much easier: basically, space

- Graph defines geometry
- "New data" = values on parts of graph not included in old domain
- Weak dependence across the graph leads to ergodicity

Prediction *of* the Graph

See some of the graph, try to predict the rest

Prediction *of* the Graph

See some of the graph, try to predict the rest

Link prediction/correction $\approx$ leave-v-out CV

Prediction *of* the Graph

See some of the graph, try to predict the rest
Link prediction/correction $\approx$ leave-v-out CV
Estimation: See subgraph, find θ, predict super-graph

Prediction *of* the Graph

See some of the graph, try to predict the rest
Link prediction/correction ≈ leave-v-out CV
Estimation: See subgraph, find θ, predict super-graph
Problems:

Prediction *of* the Graph

See some of the graph, try to predict the rest
Link prediction/correction $\approx$ leave-v-out CV
Estimation: See subgraph, find θ, predict super-graph
Problems:

- node-specific parameters + sparse data = high-dimensional estimation

Prediction *of* the Graph

See some of the graph, try to predict the rest
Link prediction/correction $\approx$ leave-v-out CV
Estimation: See subgraph, find θ, predict super-graph
Problems:

- node-specific parameters + sparse data = high-dimensional estimation
- graph geometry is random; it's what we want to predict!

Prediction *of* the Graph

See some of the graph, try to predict the rest
Link prediction/correction $\approx$ leave-v-out CV
Estimation: See subgraph, find θ, predict super-graph
Problems:

- node-specific parameters + sparse data = high-dimensional estimation
- graph geometry is random; it's what we want to predict!
- global, long-range dependence

Prediction *of* the Graph

See some of the graph, try to predict the rest
Link prediction/correction $\approx$ leave-v-out CV
Estimation: See subgraph, find θ, predict super-graph
Problems:

- node-specific parameters + sparse data = high-dimensional estimation
- graph geometry is random; it's what we want to predict!
- global, long-range dependence
- hard to find an "inside" and "outside" for blocking

A Very Rough Stat-Mech Analogy

Embedding the graph in $\mathbb{R}^p$ needs huge p

A Very Rough Stat-Mech Analogy

Embedding the graph in $\mathbb{R}^p$ needs huge p
and most of the volume of a high-dimensional set is ϵ-close to its surface

A Very Rough Stat-Mech Analogy

Embedding the graph in $\mathbb{R}^p$ needs huge p
and most of the volume of a high-dimensional set is ϵ-close to its surface
so even with short-range interactions, surface-energy terms $\approx$ volume terms
$\therefore$ thermodynamic limit gets weird

A Very Rough Stat-Mech Analogy

Embedding the graph in $\mathbb{R}^p$ needs huge p
and most of the volume of a high-dimensional set is ϵ-close to its surface
so even with short-range interactions, surface-energy terms $\approx$ volume terms
$\therefore$ thermodynamic limit gets weird
Completely messes up most exponential-family random graph models [36]

Ways Out

1. Turn dependence off: Erdos-Renyi

Ways Out

1. Turn dependence off: Erdos-Renyi
2. Dyadic independence: $\Pr\left(Z_{ij}, Z_{kl}\right) = \Pr\left(Z_{ij}\right)\Pr\left(Z_{kl}\right)$

Ways Out

1. Turn dependence off: Erdos-Renyi
2. Dyadic independence: $\Pr\left(Z_{ij}, Z_{kl}\right) = \Pr\left(Z_{ij}\right)\Pr\left(Z_{kl}\right)$
3. Conditional dyadic independence: $\Pr\left(Z_{ij}, Z_{kl} | U_i, U_j, U_k, U_l\right) = \Pr\left(Z_{ij} | U_i, U_j\right)\Pr\left(Z_{kl} | U_k, U_l\right)$

Graph Nonparametrics, Maybe

Aldous, Hoover, Kallenberg Infinite unlabeled graph distributions are always mixtures of C.D.I. processes [26]
C.D.I. processes are limits of block models

Graph Nonparametrics, Maybe

Aldous, Hoover, Kallenberg Infinite unlabeled graph distributions are always mixtures of C.D.I. processes [26]
C.D.I. processes are limits of block models

Borgs, Chayes, Lovasz Dense graph sequence converges if all motif densities converge
Dense graph sequences converge to C.D.I. processes [9]

Graph Nonparametrics, Maybe

Aldous, Hoover, Kallenberg Infinite unlabeled graph distributions are always mixtures of C.D.I. processes [26]
C.D.I. processes are limits of block models

Borgs, Chayes, Lovasz Dense graph sequence converges if all motif densities converge
Dense graph sequences converge to C.D.I. processes [9]

Diaconis, Jansson C.D.I. processes have strong independence-across-subgraph properties, and large-deviations bounds for motif densities [18]

Graph Nonparametrics, Maybe

Aldous, Hoover, Kallenberg Infinite unlabeled graph distributions are always mixtures of C.D.I. processes [26]
C.D.I. processes are limits of block models

Borgs, Chayes, Lovasz Dense graph sequence converges if all motif densities converge
Dense graph sequences converge to C.D.I. processes [9]

Diaconis, Jansson C.D.I. processes have strong independence-across-subgraph properties, and large-deviations bounds for motif densities [18]

Bickel, Chen, Levina Possible route to nonparametrics for graphs [7, 8]

What We Need

- Good notions of weak dependence for graphs
- Good notions of *sparse* graph convergence
- Something like v-fold cross-validation for graphs
  Omit n^2/v edges? Omit n/v nodes? What?
- Bootstrap/resampling for graphs
- Smoothing for graphs

Reliability is fundamental; specific criteria are not

Reliability is fundamental; specific criteria are not
Appropriate model selection depends on the purpose of the model

Realistic representation or predictive instrument?

Reliability is fundamental; specific criteria are not

Appropriate model selection depends on the purpose of the model

Realistic representation or predictive instrument?

Control of capacity and background assumptions

Reliability is fundamental; specific criteria are not
Appropriate model selection depends on the purpose of the model

Realistic representation or predictive instrument?

Control of capacity and background assumptions
Consistency of specific techniques for relational data are (largely) open questions

Reliability is fundamental; specific criteria are not

Appropriate model selection depends on the purpose of the model

Realistic representation or predictive instrument?

Control of capacity and background assumptions

Consistency of specific techniques for relational data are (largely) open questions

Someone needs to figure our network cross-validation and bootstrapping

[1] Akaike, Hirotugu (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In *Proceedings of the Scond International Symposium on Information Theory* (B. N. Petrov and F. Caski, eds.), pp. 267–281. Budapest: Akademiai Kiado. Reprinted in [2, pp. 199–213].

[2] — (1998). *Selected Papers of Hirotugu Akaike*. Berlin: Springer-Verlag. Edited by Emanuel Parzen, Kunio Tanabe and Genshiro Kitagawa.

[3] Arlot, Sylvain (2008). "V-fold cross-validation improved: V-fold penalization." Electronic preprint, arxiv.org. URL `http://arxiv.org/abs/0802.0566`.

[4] Barron, Andrew, Mark J. Schervish and Larry Wasserman (1999). "The Consistency of Posterior Distributions in Nonparametric Problems." *Annals of Statistics*, **27**: 536–561. URL `http://projecteuclid.org/euclid.aos/1018031206`.

[5] Berk, Robert H. (1966). "Limiting Behavior of Posterior Distributions when the Model is Incorrect." *Annals of Mathematical Statistics*, **37**: 51–58. URL `http://projecteuclid.org/euclid.aoms/1177699597`. doi:10.1214/aoms/1177699597. See also correction, volume 37 (1966), pp. 745–746.

[6] — (1970). "Consistency a Posteriori." *Annals of Mathematical Statistics*, **41**: 894–906. URL `http://projecteuclid.org/euclid.aoms/1177696967`. doi:10.1214/aoms/1177696967.

[7] Bickel, Peter J. and Aiyou Chen (2009). "A Nonparametric View of Network Models and Newman-Girvan and Other Modularities." *Proceedings of the National Academy of Sciences (USA)*, **106**: 21068–21073. doi:10.1073/pnas.0907096106.

[8] Bickel, Peter J., Aiyou Chen and Elizaveta Levina (2011). "The method of moments and degree distributions for

network models." *Annals of Statistics*, **39**: 38–59. URL http://arxiv.org/abs/1202.5101.

[9] Borgs, Christian, Jennifer T. Chayes, László Lovász, Vera T. Sós, Balázs Szegedy and Katalin Vesztergombi (2006). "Graph Limits and Parameter Testing." In *Proceedings of the 38th Annual ACM Symposium on the Theory of Computing [STOC 2006]*, pp. 261–270. New York: ACM. URL http://research.microsoft.com/en-us/um/people/jchayes/Papers/TestStoc.pdf.

[10] Bousquet, Olivier and André Elisseeff (2002). "Stability and Generalization." *Journal of Machine Learning Research*, **2**: 499–526. URL http://jmlr.csail.mit.edu/papers/v2/bousquet02a.html.

[11] Breiman, Leo (1996). "Bagging Predictors." *Machine Learning*, **24**: 123–140.

[12] Cesa-Bianchi, Nicolò and Gábor Lugosi (1999). "Prediction of Individual Sequences." *Annals of Statistics*,

27: 1865–1895. URL http://projecteuclid.org/euclid.aos/1017939242.

[13] Christensen, Ronald (2009). "Inconsistent Bayesian Estimation." *Bayesian Analysis*, **4**: 759–762. doi:10.1214/09-BA428.

[14] Claeskens, Gerda and Nils Lid Hjort (2008). *Model Selection and Model Averaging*. Cambridge, England: Cambridge University Press.

[15] Csiszár, Imre and Paul C. Shields (2000). "The Consistency of the BIC Markov order estimator." *Annals of Statistics*, **28**: 1601–1619. URL http://projecteuclid.org/euclid.aos/1015957472.

[16] Dedecker, Jérôme, Paul Doukhan, Gabriel Lang, José Rafael León R., Sana Louhichi and Clémentine Prieur (2007). *Weak Dependence: With Examples and Applications*. New York: Springer.

[17] Diaconis, Persi and David Freedman (1986). "On the Consistency of Bayes Estimates." *Annals of Statistics*, **14**:

1–26. URL http://projecteuclid.org/euclid.aos/1176349830.

[18] Diaconis, Persi and Svante Janson (2008). "Graph Limits and Exchangeable Random Graphs." *Rendiconti di Matematica e delle sue Applicazioni*, **28**: 33–61. URL http://arxiv.org/abs/0712.2749.

[19] Domingos, Pedro (1999). "The Role of Occam's Razor in Knowledge Discovery." *Data Mining and Knowledge Discovery*, **3**: 409–425. URL http://www.cs.washington.edu/homes/pedrod/papers/dmkd99.pdf.

[20] Fisher, R. A. (1922). "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society A*, **222**: 309–368. URL http://digital.library.adelaide.edu.au/dspace/handle/2440/15172.

[21] — (1934). "Two New Properties of Mathematical Likelihood." *Proceedings of the Royal Society of London A*,

144: 285–307. URL
`http://digital.library.adelaide.edu.au/coll/special//fisher/108.pdf`.

[22] Geman, Stuart and Chii-Ruey Hwang (1982). "Nonparametric Maximum Likelihood Estimation by the Method of Sieves." *Annals of Statistics*, **10**: 401–414. URL
`http://projecteuclid.org/euclid.aos/1176345782`.

[23] Geyer, Charles J. (2005). *Le Cam Made Simple: Asymptotics of Maximum Likelihood without the LLN or CLT or Sample Size Going to Infinity*. Tech. Rep. 643, School of Statistics, University of Minnesota. URL
`http://arxiv.org/abs/1206.4762`.

[24] Grenander, Ulf (1981). *Abstract Inference*. New York: Wiley.

[25] Grünwald, Peter D. (2007). *The Minimum Description Length Principle*. Cambridge, Massachusetts: MIT Press.

[26] Kallenberg, Olav (2005). *Probabilistic Symmetries and Invariance Principles*. New York: Springer-Verlag.

[27] Kass, Robert E. and Adrian E. Raftery (1995). "Bayes Factors." *Journal of the American Statistical Association*, **90**: 773–795. URL http://www.stat.cmu.edu/~kass/papers/bayesfactors.pdf.

[28] Keane, Michael and Karl Petersen (2006). "Easy and nearly simultaneous proofs of the Ergodic Theorem and Maximal Ergodic Theorem." In *Dynamics and Stochastics: Festschrift in honor of M.S. Keane* (Dee Denteneer and Frank Den Hollander and Evgeny Verbitskiy, eds.), vol. 48 of *IMS Lecture Notes-Monographs Series*, pp. 248–251. Hayward, California: Institute of Mathematical Statistics. URL http://arxiv.org/abs/math.DS/0608251.

[29] Massart, Pascal (2007). *Concentration Inequalities and Model Selection*. Berlin: Springer-Verlag. URL http://eprints.pascal-network.org/archive/00002827/.

[30] Miller, Jeffrey W. and Matthew T. Harrison (2013). "A simple example of Dirichlet process mixture inconsistency

for the number of components." arxiv:1301.2708. URL http://arxiv.org/abs/1301.2708.

[31] Mohri, Mehryar, Afshin Rostamizadeh and Ameet Talwalkar (2012). *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, Massachusetts: MIT Press.

[32] Rivers, Douglas and Quang H. Vuong (2002). "Model selection tests for nonlinear dynamic models." *The Econometrics Journal*, **5**: 1–39. doi:10.1111/1368-423X.t01-1-00071.

[33] Schapire, Robert E. and Yoav Freund (2012). *Boosting: Foundations and Algorithms*. Cambridge, Massachusetss: MIT Press.

[34] Schwarz, Gideon (1978). "Estimating the Dimension of a Model." *Annals of Statistics*, **6**: 461–464. URL http://projecteuclid.org/euclid.aos/1176344136.

[35] Shalizi, Cosma Rohilla (2009). "Dynamics of Bayesian Updating with Dependent Data and Misspecified

Models." *Electronic Journal of Statistics*, **3**: 1039–1074. URL
`http://arxiv.org/abs/0901.1342`.
doi:10.1214/09-EJS485.

[36] Shalizi, Cosma Rohilla and Alessandro Rinaldo (2013). "Consistency Under Sampling of Exponential Random Graph Models." *Annals of Statistics*, **41**: 508–535. URL `http://arxiv.org/abs/1111.3054`.

[37] van de Geer, Sara A. (2000). *Empirical Processes in M-Estimation*. Cambridge, England: Cambridge University Press.

[38] Vapnik, Vladimir N. (1995). *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag, 1st edn.

[39] Vuong, Quang H. (1989). "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica*, **57**: 307–333. URL `http://www.jstor.org/pss/1912557`.

[40] Yu, Bin (1994). "Rates of Convergence for Empirical Processes of Stationary Mixing Sequences." *Annals of*

Probability, **22**: 94–116. URL http://projecteuclid.org/euclid.aop/1176988849.