What have we learned until now: Part 1

- Direct simulation from probability densities
- Several different methods
  - inversion sampling
  - methods based on relationship between RV
  - rejection sampling (*)
  - importance sampling (*)
- We get samples that are independent of each other

What have we learned until now: Part 1 and 2

Bayesian paradigm

- Likelihood $\pi(y|x)$
- Prior $\pi(x)$
- Posterior $\pi(x|y) \propto \pi(y|x)\pi(x)$

What have we learned until now: Part 2

Hierarchical models are an extremely useful tool in Bayesian model building.

Three parts:

- Observation model $\boldsymbol{y}|\boldsymbol{x}$: Encodes information about observed data.
- The latent model $\boldsymbol{x}|\boldsymbol{\theta}$: The unobserved process.
- Hyperpriors for $\boldsymbol{\theta}$: Models for all of the parameters in the observation and latent processes.

Note: here we indicate the observed data by $\boldsymbol{y}$ while $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are parameters

What have we learned untill now: Part 2

MCMC algorithm:

- Problem: Sample from $\pi(x)$, $x \in S$.
- MCMC idea:
  - Construct Markov chain with $\pi(x)$ as limiting distribution.
  - Simulate the Markov chain for a long time so that it has time to converge.
  - Most MCMC samplers are based on reversible Markov chains
    $\Rightarrow$ Their convergence is proved by checking the detailed balance equation.
- Can be applied to virtually any bayesian model
- Convergence and slow mixing can be a big issue

What have we learned until now: Part 2

Integrated nested Laplace approximation:

- Can be applied only on a (large) class of models: Latent Gaussian models
- No sampling involved, based on numerical approximation
- The focus is on posterior marginals

TMA4300 - Part 3

TMA4300 - Part 3

*

Last part of this course

⇒ Not closely related to the two first parts
- no more MCMC
- mostly non-Bayesian perspective

⇒ Two topics (not closely related to each other):
- Bootstrapping
- Expectation-Maximization algorithm

*Slides are partially based on lecture notes kindly provided by Håkon Tjelmeland, Andrea Riebler, and Sara Martino.

Bootstrap

Bootstrap

http://tradingconsequences.blogs.edina.ac.uk/files/2013/10/Dr_Martens_black_old.jpg

An example for introduction

Group	Survival Time	Sample size	Mean	Estimated SE
Treatment	94,197,16,38 99,141,23	7	86.86	
Control	52,104,146,10,51,46 30,40,27,46	9	56.22	
		Differance:	30.63	

- Is the difference in mean significant?
-

An example for introduction

Group	Survival Time	Sample size	Mean	Estimated SE
Treatment	94,197,16,38 99,141,23	7	86.86	25.24
Control	52,104,146,10,51,46 30,40,27,46	9	56.22	14.14
		Differance:	30.63	30.93

- Is the difference in mean significant?
-

An example for introduction

Group	Survival Time	Sample size	Mean	Estimated SE
Treatment	94,197,16,38 99,141,23	7	86.86	25.24
Control	52,104,146,10,51,46 30,40,27,46	9	56.22	14.14
		Differance:	30.63	30.93

- Is the difference in mean significant?
- What if we want to compare the medians instead?
  Show code Bootstrap_intro.R

. . . pull oneself up by one's bootstraps

To begin an enterprise or recover from a setback without any outside help; to succeed only on one's own effort or abilities.

Wiktionary

The term is sometimes attributed to Rudolf Erich Raspe's story "The Surprising Adventures of Baron Munchausen", where the main character pulls himself (and his horse) out of a swamp by his hair

http://redstateeclectic.typepad.com/redstate_commentary/2010/11/sustainability-isnt-sustainable.html

The boostrap

- Bootstrap is a computer-based technique for doing statistical inference (usually with a minimum of assumptions)
- It is not Bayesian

Important concepts

- empirical distribution function
- plug in estimator
- bootstrap sample

Today's lecture

- Bootstrap
  - Non-parametric
  - Parametric
- Bootstrap estimate of SD
- Bootstrap estimate of bias

Empirical distribution function

Assume we have iid observations from an (unknown) distribution F:

$$F \to (x_1, \ldots, x_n)$$

The empirical distribution function $\hat{F}$ is the CDF that puts mass $1/n$ at each data point x_i:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1(x_i \leq x)$$

where $1(\cdot)$ denotes the indicator function.
For iid samples $\hat{F}$ is a sufficient estimator for F.

Plug in estimator

Let θ be an interesting feature of F, $\theta = t(F)$.

For example:

$$\theta = \mathrm{E}(X) = \int x f(x) dx$$

$$\theta = \mathrm{Var}(X) = \int (x - \mathrm{E}(X))^2 f(x) dx$$

The plug-in estimator for θ is defined by:

$$\hat{\theta} = t(\hat{F})$$

The plug-in principle is quite good, if the only information about F, comes from the sample x.

Examples

Thus

$$\theta = \mathrm{E}(X) \Rightarrow \hat{\theta} = \mathrm{E}_{\hat{F}}(X) = \sum_{i=1}^{n} x_i \frac{1}{n} = \bar{x}$$

Examples

Thus

$$\theta = \mathrm{E}(X) \Rightarrow \hat{\theta} = \mathrm{E}_{\hat{F}}(X) = \sum_{i=1}^{n} x_i \frac{1}{n} = \bar{x}$$

$$\theta = \mathrm{Var}(X) \Rightarrow \hat{\theta} = \mathrm{Var}_{\hat{F}}(X) = \mathrm{E}_{\hat{F}}[(X - \mathrm{E}_{\hat{F}}(X))^2]$$
$$= \sum_{i=1}^{n} (x_i - \mathrm{E}_{\hat{F}}(X))^2 \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Examples

Thus

$$\theta = \mathrm{E}(X) \Rightarrow \hat{\theta} = \mathrm{E}_{\hat{F}}(X) = \sum_{i=1}^{n} x_i \frac{1}{n} = \bar{x}$$

$$\theta = \mathrm{Var}(X) \Rightarrow \hat{\theta} = \mathrm{Var}_{\hat{F}}(X) = \mathrm{E}_{\hat{F}}[(X - \mathrm{E}_{\hat{F}}(X))^2]$$
$$= \sum_{i=1}^{n} (x_i - \mathrm{E}_{\hat{F}}(X))^2 \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\theta = \mathrm{SD}(X) \Rightarrow \hat{\theta} = \mathrm{SD}_{\hat{F}}(X) = \sqrt{\mathrm{Var}_{\hat{F}}(X)}$$
$$= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Example

$$E_F \left[\left(\frac{X - \mu}{\sigma} \right)^3 \right] = \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

See notes

Setting

Assume we have :

$$F \to (x_1, \ldots, x_n)$$

Thus $\hat{F}$ gives mass $\frac{1}{n}$ to each observed value.

A bootstrap sample is defined to be a random sample of size n from $\hat{F}$, say $x^\star = (x_1^\star, \ldots, x_n^\star)$

$$\hat{F} \to (x_1^\star, \ldots, x_n^\star)$$

Simple illustration

Suppose $n = 3$ univariate data points, namely

$$\{x_1, x_2, x_3\} = \{1, 2, 6\}$$

are observed as an iid sample from F that has mean θ . At each observed data value, $\hat{F}$ places mass $1/3$. Suppose the estimator to be bootstrapped is the sample mean $\hat{\theta}$.

There are $3^3 = 27$ possible outcomes for $\mathcal{X}^\star = \{X_1^\star, X_2^\star, X_3^\star\}$.

Simple illustration (II)

$\mathcal{X}^\star$	$\hat{\theta}^\star$	$P^\star(\hat{\theta}^\star)$	Observed frequency
1 1 1	3/3	1/27	36/1000
1 1 2	4/3	3/27	101/1000
1 2 2	5/3	3/27	123/1000
2 2 2	6/3	1/27	25/1000
1 1 6	8/3	3/27	104/1000
1 2 6	9/3	6/27	227/1000
2 2 6	10/3	3/27	131/1000
1 6 6	13/3	3/27	111/1000
2 6 6	14/3	3/27	102/1000
6 6 6	18/3	1/27	40/1000

Bootstrap estimate for standard error

- Parameter of interest: $\theta = t(F)$
- Our estimator for θ: $\hat{\theta} = s(x)$
- Want (to estimate) $\mathrm{SD}_F(\hat{\theta})$.

A bootstrap replication of $\hat{\theta}$ is

$$\hat{\theta}^\star = s(x^\star)$$

Use plug-in principle to estimate $\mathrm{SD}_F(\hat{\theta})$.
The bootstrap estimate of the standard error of $\hat{\theta} = s(x)$ is $\mathrm{SD}_{\hat{F}}(\hat{\theta}^\star)$.
This is called the ideal bootstrap estimate of standard error of $\hat{\theta}$.

Ideal bootstrap estimate of standard error

- For the sample mean it can be computed analytically
- For (very) small sample sizes it can be computed using all the possible bootstrap replicates. (Number of possible bootstrap sample: n^n.)
- In other cases it can be approximated via Monte Carlo techniques

Computational way of obtaining a good estimate

We can estimate $\mathrm{SD}_{\hat{F}}(\hat{\theta}^\star)$ by simulation:

1. Generate B bootstrap samples $x^{1\star}, \ldots, x^{B\star}$.
2. Evaluate the corresponding parameter estimates

$$\hat{\theta}^\star(b) = s(x^{b\star}), \quad b = 1, 2, \ldots, B$$

3. Estimate $\mathrm{SD}_{\hat{F}}(\hat{\theta}^\star)$ by

$$\widehat{\mathrm{SE}}_B = \sqrt{\frac{\sum_{b=1}^{B}(\hat{\theta}^\star(b) - \hat{\theta}^\star(\cdot))^2}{B-1}}$$

where

$$\hat{\theta}^\star(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^\star(b)$$

Note

$$\lim_{B\to\infty} \widehat{\mathrm{SE}}_B = \widehat{\mathrm{SE}}_\infty = \widehat{\mathrm{SD}}_{\hat{F}}(\hat{\theta}^\star)$$

Example

Setting

$$\theta = \mathrm{E}(X)$$

$$\hat{\theta} = s(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$\hat{\theta}^\star = s(x^\star) = \frac{1}{n}\sum_{i=1}^{n} x_i^\star = \bar{x^\star}$$

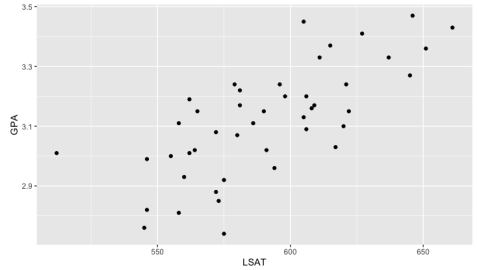
Here, the ideal bootstrap estimate exists

see blackboard

Example: The correlation coefficient

Scores for 15 law schools in the USA

$$y_i = (LSAT_i, GPA_i), \; t = i \ldots, 15$$



The correlation between the two scores is estimated to be 0.78, but what is its standard error?

Example: The correlation coefficient

- 1000 bootstrap replicates

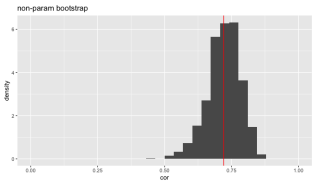
$$y^{1\star}, \ldots, y^{1000\star}$$

- For each replicates compute

$$\hat{\theta}^{i\star} = s(y^{i\star})$$

- Estimate bootstrap SE

$$\hat{SD}_{\hat{F}}(\theta) = 0.121$$



How large do we need B?

Intuitively we understand that the $\widehat{\mathrm{SE}}_B$ has larger standard deviation than $\widehat{\mathrm{SE}}_\infty$.

Theory, not to be discussed here, gives the following rules of thumb:

1. Even a small B is informative, say $B = 25$ or $B = 50$ is often enough to get a good estimate of $\mathrm{SE}_F(\hat{\theta})$.
2. Very seldomly more than $B = 200$ is necessary to estimate $\mathrm{SE}_F(\hat{\theta})$.

The parametric bootstrap

When data are modeled to originate from a parametric distribution, so

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x, \xi),$$

another estimate of F may be employed.

Suppose that the observed data are used to estimate ξ by $\hat{\xi}$. Then each parametric bootstrap pseudo-dataset $\mathcal{X}^\star$ can be generated by drawing $X_1^\star, \ldots, X_n^\star \overset{\text{iid}}{\sim} F(x, \hat{\xi}) = \hat{F}_{\text{par}}$.

Again ...

... we can/must estimate $\mathrm{SD}_{\hat{F}_{\text{par}}}(\hat{\theta}^\star)$ by simulation:

1. Generate B bootstrap samples $x^{1\star}, \ldots, x^{B\star}$, where
$$x^{b\star} = (x_1^{b\star}, \ldots, x_n^{b\star})$$
with $x_1^{b\star}, \ldots, x_n^{b\star} \overset{\text{iid}}{\sim} \hat{F}_{\text{par}}$.
2. Evaluate the corresponding parameter estimates
$$\hat{\theta}^\star(b) = s(x^{b\star}), \quad b = 1, 2, \ldots, B$$
3. Estimate $\mathrm{SD}_{\hat{F}_{\text{par}}}(\hat{\theta}^\star)$ by
$$\widehat{\mathrm{SE}}_B = \sqrt{\frac{\sum_{b=1}^B (\hat{\theta}^\star(b) - \hat{\theta}^\star(\cdot))^2}{B-1}}$$
where
$$\hat{\theta}^\star(\cdot) = \frac{1}{B}\sum_{b=1}^B \hat{\theta}^\star(b)$$

Example: Correlation coefficients

We assume now that

$$y_i = (LSAT_i, GPA_i) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ i.i.d}$$

where $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and obtain:

$$\hat{F}_{(\hat{\mu}, \hat{\Sigma})}$$

Example: The correlation coefficient

- 1000 bootstrap replicates

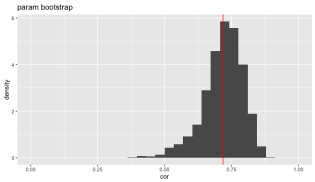
$$y^{1\star}, \ldots, y^{1000\star} \sim \hat{F}_{(\hat{\mu},\hat{\Sigma})}$$

- For each replicates compute

$$\hat{\theta}^{i\star} = s(y^{i\star})$$

- Estimate bootstrap SE

$$\hat{SD}_{\hat{F}}(\theta)$$



Bootstrapping regression

Consider the ordinary multiple regression model

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \text{for } i = 1, \ldots, n,$$

where ϵ_i are iid mean zero random variables with constant variance.

- Parameters of interest $\boldsymbol{\beta}$
- Want to estimate $SD(\hat{\beta})$

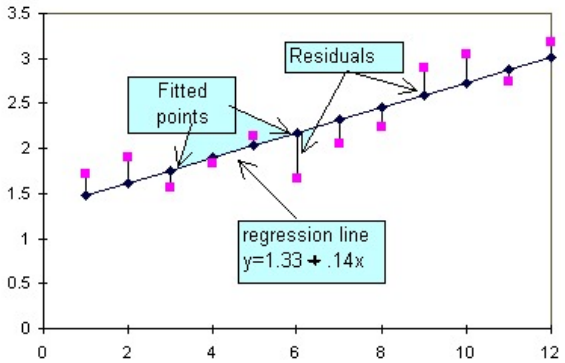
Review: Linear Regression

- Least square estimate of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \text{argmin}\{\sum (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2\} \Rightarrow \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$

- Residuals

$$e_i = Y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$$



Bootstrap regression

Alternative 1: Bootstrap the residuals $e_i = Y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$

Alternative 2: Bootstrap the pairs $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$

Bootstrap the residuals

1. Fit the regression model to the observed data and obtain the fitted responses $\hat{y}_i$ and residuals $\hat{\epsilon}_i$.
2. Sample a bootstrap set of residuals $\hat{\epsilon}_1^\star, \ldots, \hat{\epsilon}_n^\star$ from the set of fitted residuals completely at random and with replacement.
3. Generate a bootstrap set of pseudo responses

$$Y_i^\star = \hat{y}_i + \hat{\epsilon}_i^\star, \quad \text{for } i = 1, \ldots, n.$$

4. Regress $Y^\star$ on $\boldsymbol{x}$ to obtain a bootstrap estimate $\hat{\boldsymbol{\beta}}^\star$.

Repeat this process to get an empirical distribution of $\hat{\boldsymbol{\beta}}^\star$.

Bootstrapping residuals: Remarks

This approach is also used for autoregressive models, for example.

Note: Bootstrapping the residuals is reliant on

- The model provides an appropriate fit
- The residuals have a constant variance

Otherwise, a different scheme is recommended.

Comment: No need to bootstrap for linear regression model with least squares estimation, as analytical results are then available.

Bootstrap the pair $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$

Suppose response and predictors are measured from a collection of individuals selected at random

$\Rightarrow$ Data pairs $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$ can be regarded as iid realisation from $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$ drawn from a joint response-predictor distribution.

Bootstrap:

- Sample $\boldsymbol{Z}_1^\star, \ldots, \boldsymbol{Z}_n^\star$ completely at random with replacement from $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$.
- Apply regression model on pseudo dataset to get $\hat{\boldsymbol{\beta}}^\star$.

Repeat this approach many times.

Note: Paired bootstrap is less sensitive to violation of assumptions, e.g. adequacy of regression model, than bootstrapping the residuals.

Copper-nickel alloy

Data: 13 measurements of corrosion loss (y_i) in copper-nickel alloys, each with a specific iron content (x_i).

Question: Change in corrosion loss in the alloys as the iron content increases, relative to corrosion loss where there is no iron, i.e. $\theta = \beta_1/\beta_0$.

x_i	0.01	0.48	0.71	0.95	1.19	0.01	0.48
y_i	127.6	124.0	110.8	103.9	101.5	130.1	122.0
x_i	1.44	0.71	1.96	0.01	1.44	1.96	
y_i	92.3	113.1	83.7	128.0	91.4	86.2	

The observed data yield $\hat{\theta} = \hat{\beta}_1/\hat{\beta}_0 = -0.185$.

Bias of an estimator

- We observe $X_1, X_2, \ldots, X_n \sim F$ iid
- Parameter of interest $\theta = t(F)$
- Estimator $\hat{\theta} = s(X)$
  (may or may not be based on the plug-in principle)
- Bias definition

$$\text{bias}_F(\hat{\theta}, \theta) = \mathsf{E}_F[\hat{\theta}] - \theta = \mathsf{E}_F[s(\boldsymbol{x})] - t(F)$$

Bootstrap estimate of bias

We want to estimate

$$\text{bias}_F(\hat{\theta}, \theta) = \mathsf{E}_F[s(\boldsymbol{x})] - t(F)$$

Idea: Apply the plug-in principle and define the bootstrap estimate of bias as:

$$\text{bias}_{\hat{F}} = \mathsf{E}_{\hat{F}}[s(\boldsymbol{x}^\star)] - t(\hat{F})$$

where $\hat{F}$ is an estimate of F (for example the empirical distribution)

Bias estimate of the bias

1. Generate B bootstrap samples $x^{1\star}, \ldots, x^{B\star}$.
2. Evaluate the corresponding parameter estimates

$$\hat{\theta}^\star(b) = s(x^{b\star}), \quad b = 1, 2, \ldots, B$$

3. Approximate the bootstrap expectation $\mathsf{E}_{\hat{F}}[s(\boldsymbol{x}^\star)]$ as:

$$\hat{\theta}^\star(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^\star(b)$$

4. Approximate the ideal bootstrap estimate for bias as

$$\widehat{\text{bias}}_B = \hat{\theta}^\star(\cdot) - t(\hat{F})$$

Bias corrected estimate

One we have estimated the bias we can compute the bias-corrected estimator

$$\hat{\theta}_c = \hat{\theta} - \widehat{\text{bias}}_B = \hat{\theta} - [\hat{\theta}^\star(\cdot) - t(\hat{F})]$$

Bias corrected estimate

One we have estimated the bias we can compute the bias-corrected estimator

$$\hat{\theta}_c = \hat{\theta} - \widehat{\text{bias}}_B = \hat{\theta} - [\hat{\theta}^\star(\cdot) - t(\hat{F})]$$

Note: Bias correction will not always give an improved estimator. We have that $\text{Var}(\hat{\theta}_c) \geq \text{Var}(\hat{\theta})$ so if the bias is small is better not to do bias correction.

Confidence intervals (percentile method)

A "simple-minded" two-sided confidence interval with coverage $(1-\alpha)$ for a parameter α is given by

$$[q^\star_{\alpha/2}, q^\star_{1-\alpha/2}]$$

where $q^\star_\alpha$ is the α-bootstrap quantile in the distribution of $\hat{\theta}^\star$.

Experience: Often good, but often too low coverage, i.e the true α for the interval is lower than the specified value.
Note: Better bootstrap confidence intervals exist and often have better coverage accuracy — at the price of being somewhat more difficult to implement
Show R-code bootstrap_regression.R

Bootstrap bias correction

Copper-nickel alloy example

The mean value of

$$\hat{\theta}^\star - \hat{\theta}$$

among the pseudo datasets is about -0.00125.

The bias-corrected bootstrap estimate of β_1/β_0 is $-0.18507 - (-0.00125) = -0.184$.