Survivorship & Conditional Imputation

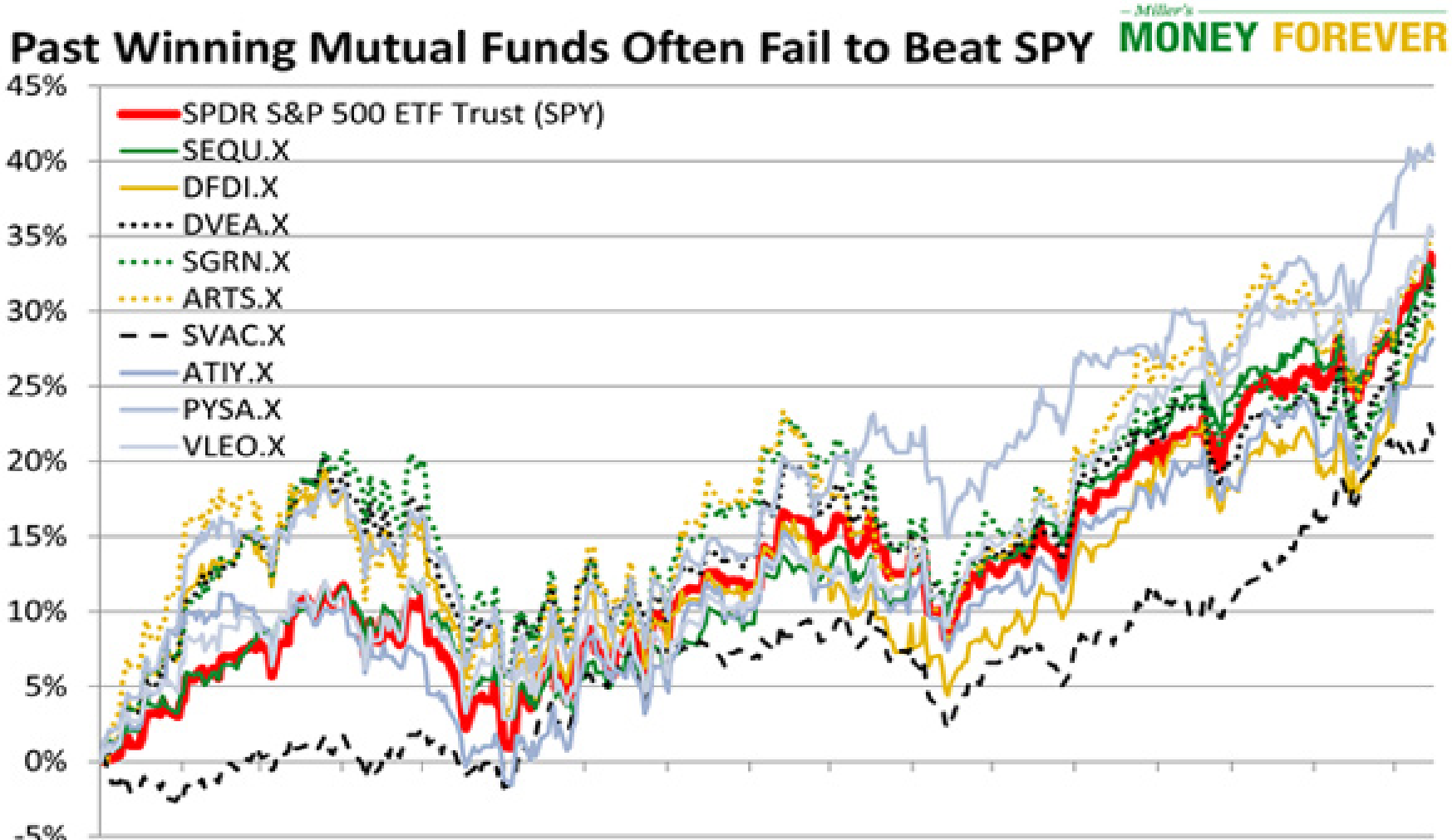
Don McLeish and Cyntha Struthers

University of Waterloo

Dec 5, 2015

Survivorship in Brownian Motion Model

Mutual Fund returns, CD4 counts, blood pressure...resemble Brownian motion.



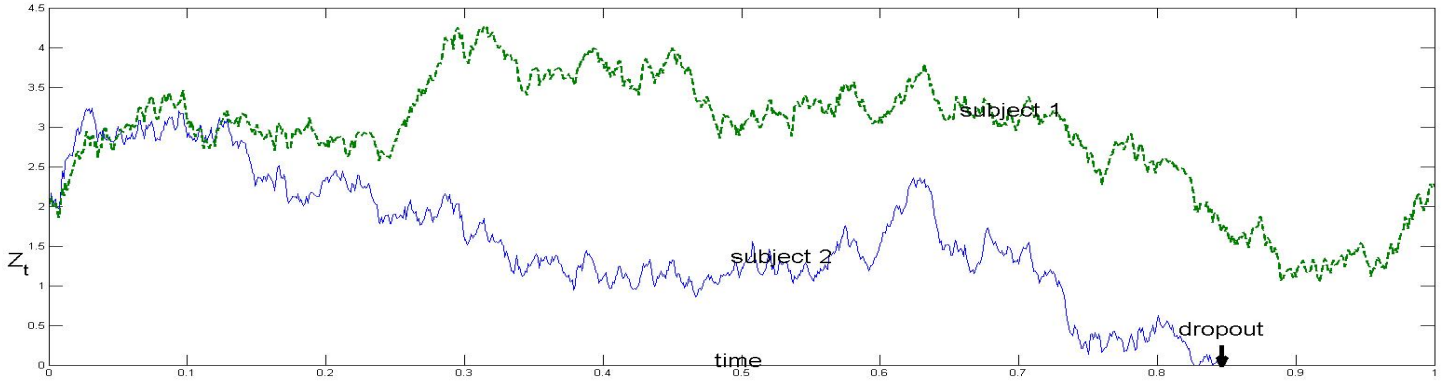
Survivorship for a Brownian Motion

- Suppose a random variable Z_t results from a continuous time process, $0 < t \leq T$.
- If Z_t ever drops below a barrier then the process disappears from view (e.g. dropout occurs).
- We are interested in the distribution of Z_T for the "survivors".

survivemovie.mp4

Survivorship Under a Brownian Motion:

- In a longitudinal study dropouts may occur either independently of the process Z_t or because Z_t falls below a threshold level b.
- The survivors' data Z_T obtains from the conditional distribution $f\left(z_T \middle| min_{t \leq T} Z_t > b\right)$.
- (When $R = 0$ we do not know whether $min(Z_t) > b$ or $\leq b$.



Survivorship/Post-Hoc Bias

Bias in Meta-analysis: Publication bias leads to test statistics Z_T that tend to be skewed towards significance[1].
Performance measures: Publication and survivorship lead to published portfolio investment returns or performance measures that are positively biased.
Group-Sequential Trials: We may sequentially monitor a trial and discontinue treatment (or end the trial) at time $t < T$ if the efficient score, say Z_t, (e.g. [2]) falls crosses a given threshold. In this case we observe the value of $R = 0$ if terminated. We observe values of R and Z_T $\sim f(z_T | min_{t \leq T} Z_t > b)$.

---

[1] Egger, et al. (1997)
[2] Whitehead (1997)

Models for survivorship Bias

Suppose a random variable Z is generated from a probability density function $f(z)$. Values of Z below a barrier b are deleted. Those with values close to a barrier survive with probability $G(z-b)$ where G is a c.d.f.

- We may not observe the random variable R. The probability density function **conditional on survival** $f(z|R=1)$ is proportional to $f(z)P(R=1|z)$or $f(z)G(z-b)$ for $z > b$.
- $G(z-b)$ "thins" the observations by selecting a fraction as survivors. This is an example of a **selection model**.
- Tractable choices for G include exponential, Normal c.d.f.[3], etc.

[3]leading to the skew normal distribution

A Distribution for Survivors

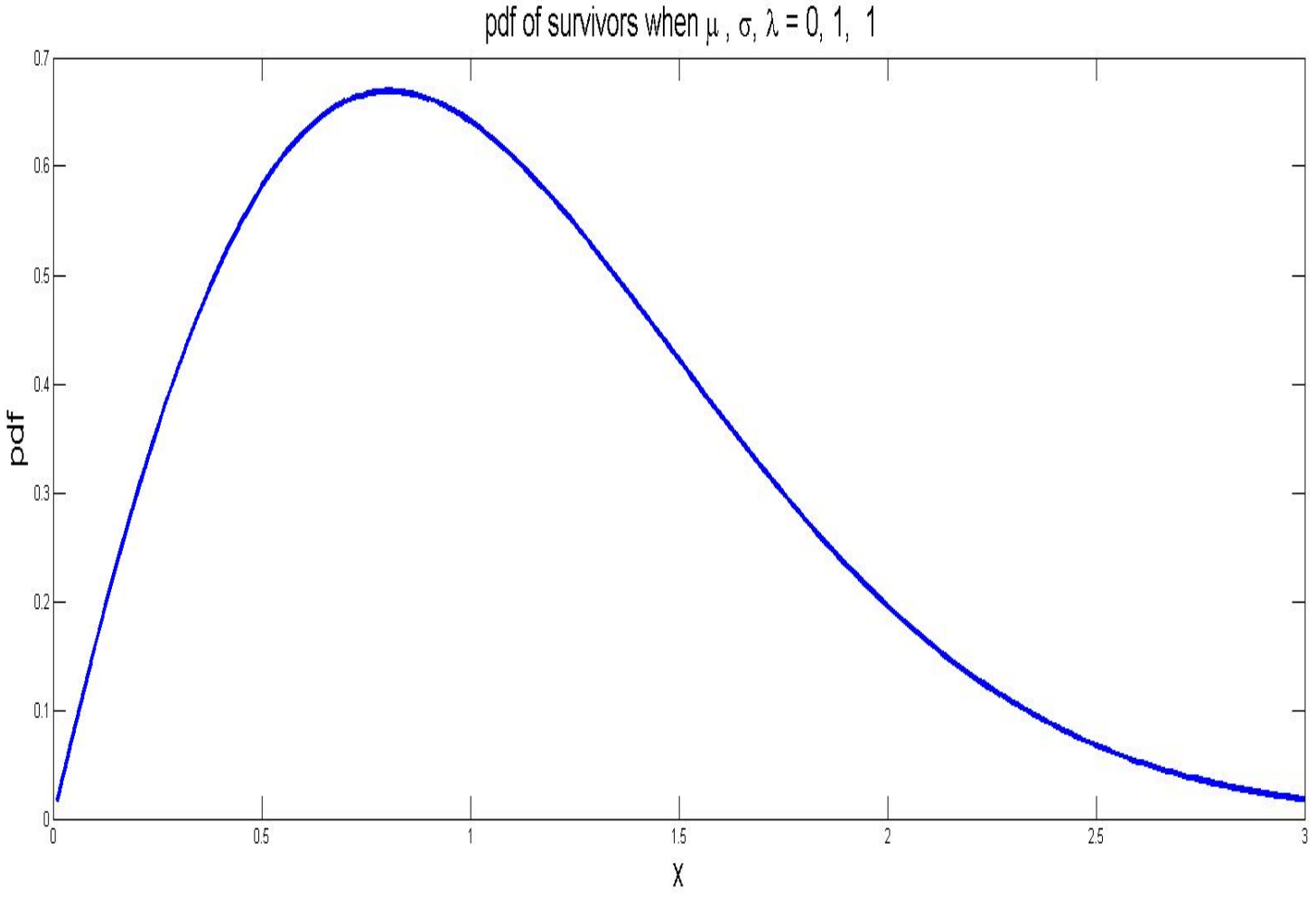
- Consider "exponential thinning"[4]. $G(z)$ is an exponential c.d.f. $G(z) = 1 - e^{-\lambda(z-b)}$, for $z > b$.
- If $f(z)$ is Normal we obtain $\frac{1}{k\sigma}\varphi\left(\frac{z-\mu}{\sigma}\right)\left(1 - e^{-\lambda(z-b)}\right)$, for $z > b$.
- Replacing Z by $Z - b$ obtain the family

$$h_\lambda(z;\mu,\sigma) = \frac{1}{k\sigma}\varphi\left(\frac{z-\mu}{\sigma}\right)\left(1 - e^{-\lambda z}\right), \text{ for } z > 0$$

with $\mu \in \Re, \sigma > 0, \lambda > 0$, $k = \Phi(\frac{\mu}{\sigma}) - p\Phi(\frac{\mu}{\sigma} - \lambda\sigma)$, and $p = e^{-\lambda\mu+\lambda^2\sigma^2/2}$.

[4]Thinning is exponential $G(z) = 1 - e^{-\lambda(z-b)}$ whenever the distribution of $Z_t|Z_0, Z_T$ for $0 < t < T$ is a Brownian bridge.

A p.d.f. for the survivors



The p.d.f. of the distribution of survivors h_λ

Survivorship Distribution: Moments

The density h_λ is exponential family so the maximum likelihood estimate of the parameters are obtained from the moments.

- First Moment:

$$E\left(\frac{Z-\mu}{\sigma}\right) = \frac{\lambda\sigma}{k} p\Phi\left(\frac{\mu}{\sigma} - \lambda\sigma\right)$$

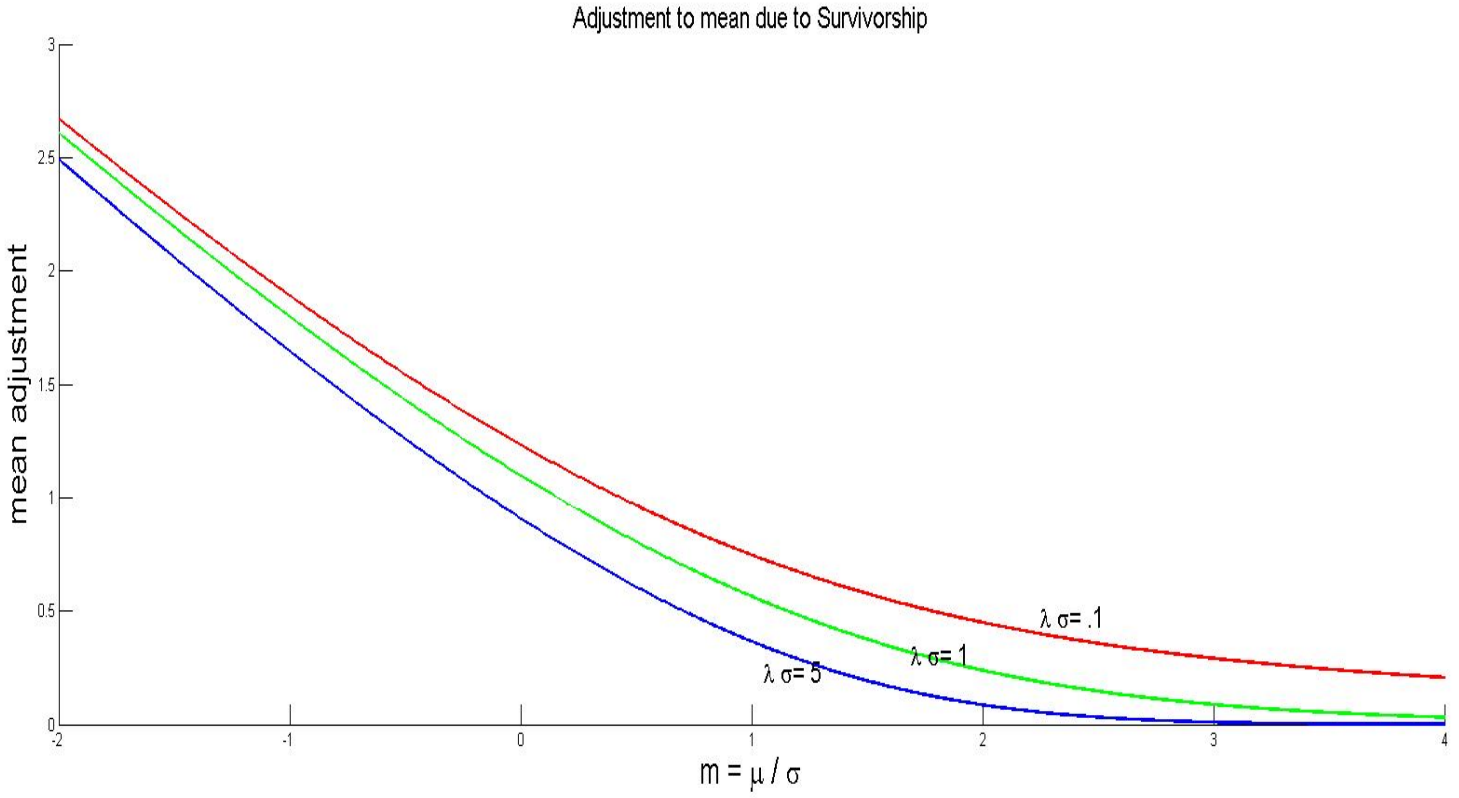
- Second Moment:

$$E\left(\frac{Z-\mu}{\sigma}\right)^2 = 1 + \frac{\lambda\sigma}{k}\left[\varphi(\frac{\mu}{\sigma}) - \lambda^2\sigma^2 p\Phi(\frac{\mu}{\sigma} - \lambda\sigma)\right]$$

- Functions of 2 parameters, $\frac{\mu}{\sigma}$ and $\lambda\sigma$. $p = e^{-\lambda\mu+\lambda^2\sigma^2/2}$

How much difference to the mean does survivorship make?

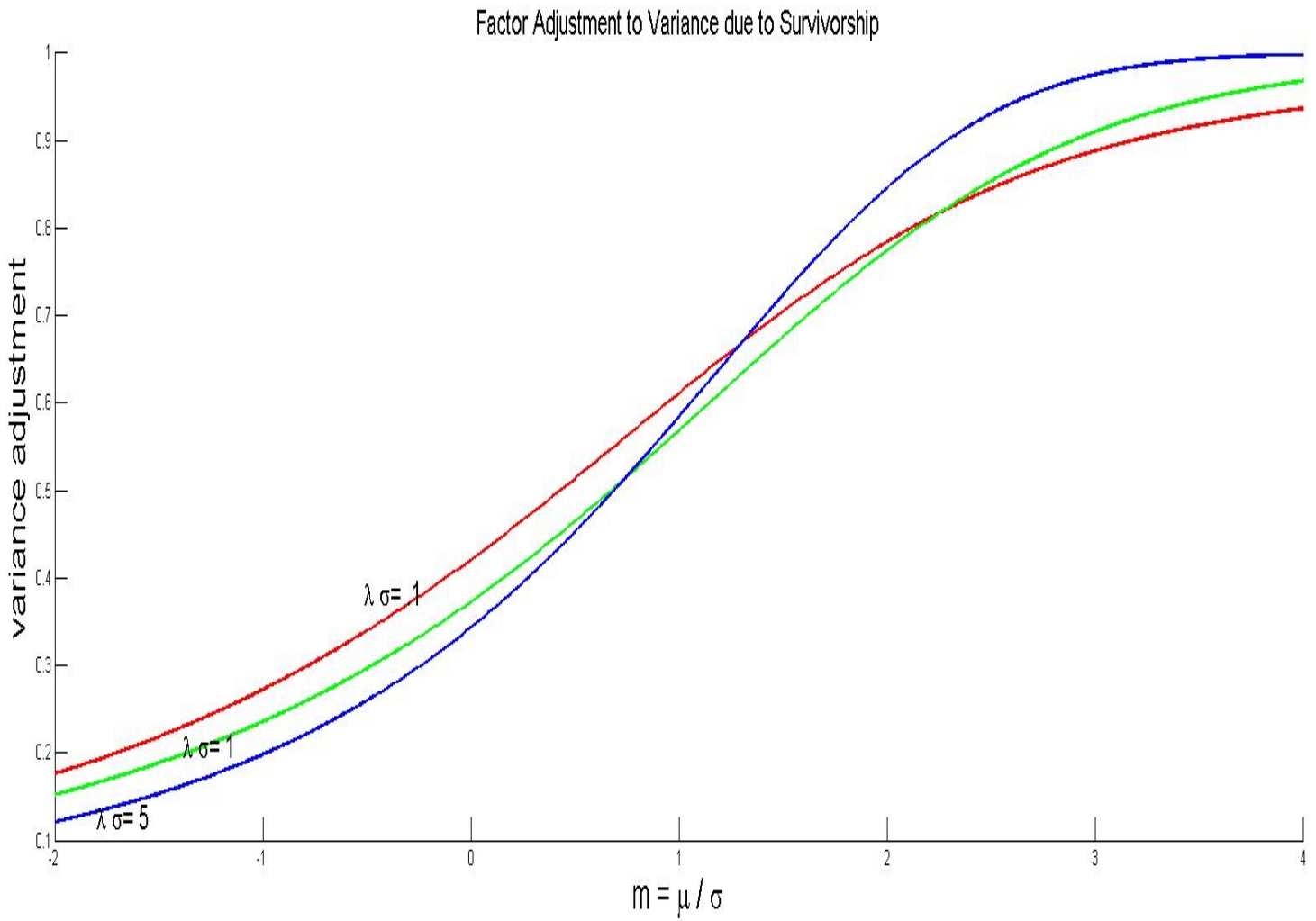
Assuming $\sigma = 1$, $E(Z) - \mu = \frac{\lambda}{k} p\Phi(\mu - \lambda)$



$E(Z) - \mu$ as a function of μ, and λ

How much difference to the variance does survivorship make?

$$Var(Z) = 1 + \frac{\lambda}{k}\left[\varphi(\mu) - \frac{\lambda p}{k}\Phi(\mu)\Phi(\mu - \lambda)\right]$$



When do I need to take survivorship into account?

- When we are close to the barrier (distance measured in units of σ) there is more "thinning" so
  - the mean of the survivors $>>$ unconditional mean
  - Variance of the survivors $<$ unconditional variance
  - The unconditional moments are biased.
- When there is selective dropout, treatments appear more effective, especially if the outcome is highly variable. (Risky investors have apparently better performance a posteriori)
- We should adjust estimators for survivorship using the corrected moments above.

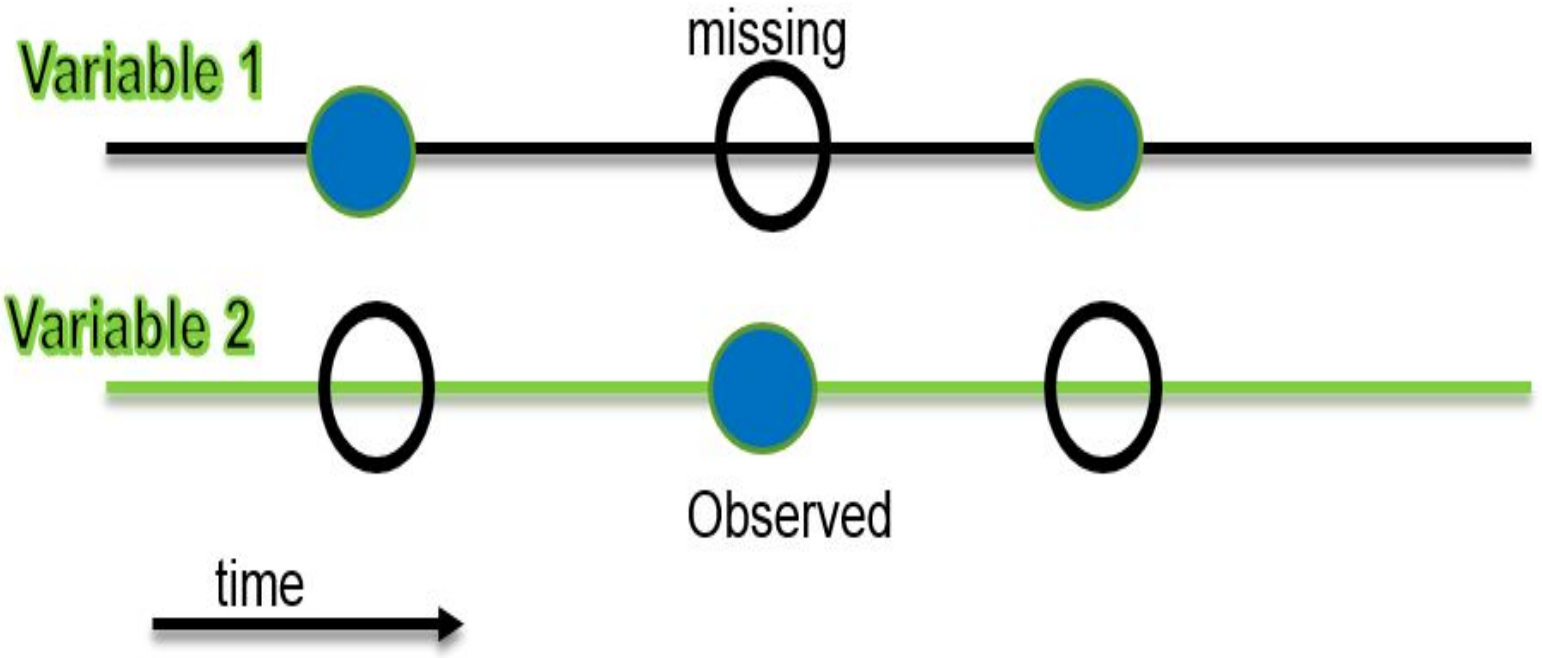
Why multivariate time series data is often incomplete.

- There are often related time series which are disregarded. We may
  - Analyze stock without corresponding index,
  - analyze blood pressure without diet, alcohol, physical activity, etc.
- Does the world stop turning when I sleep? There are gaps in data collection nights, holidays, week-ends, sick-days,
- Asynchronous reporting. Weight, lab test results at discrete non-synchronous times.

Example of asynchronous Observations

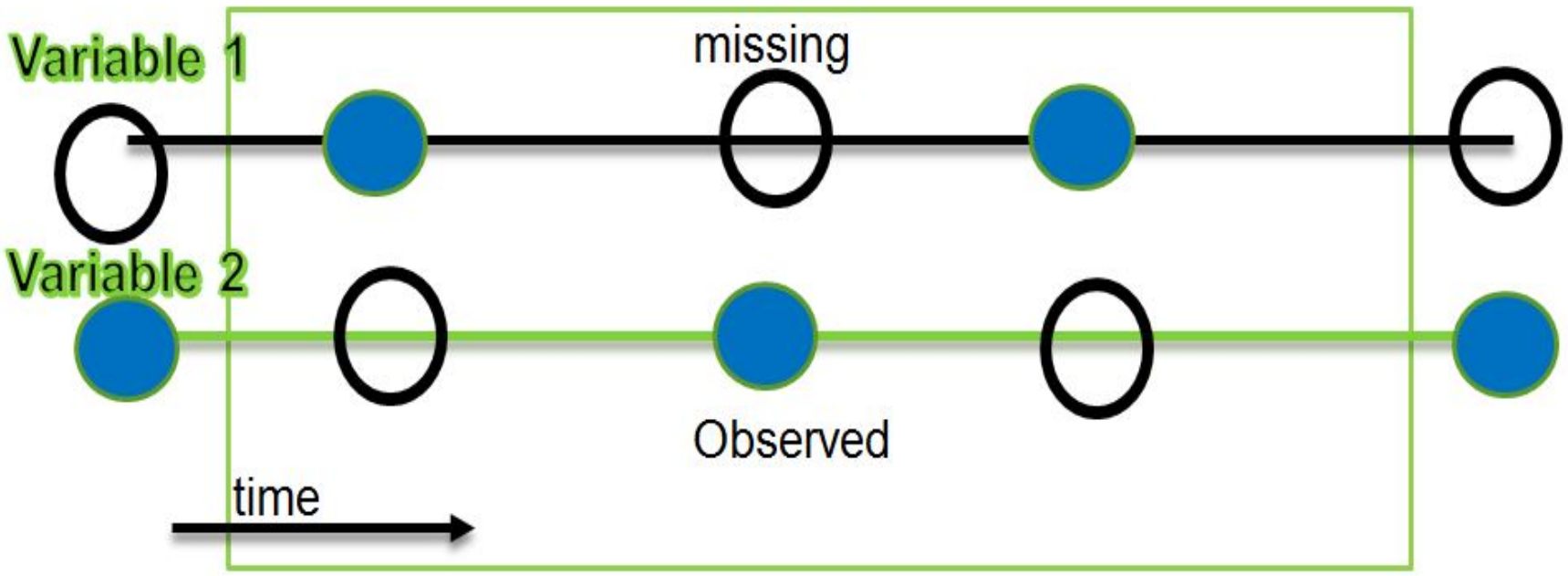
- Suppose I wish to study the joint behaviour of stock market indices, e.g. DJIA, Hang Seng, etc. Usually use LOCF then treat as a synchronous multivariate observation. Harmless if small volatility and reasonably frequent observations. However, 12 hours separate indices in Hong Kong from New York. Holidays and market closures are additional complications.

Example: asynchronous measurements



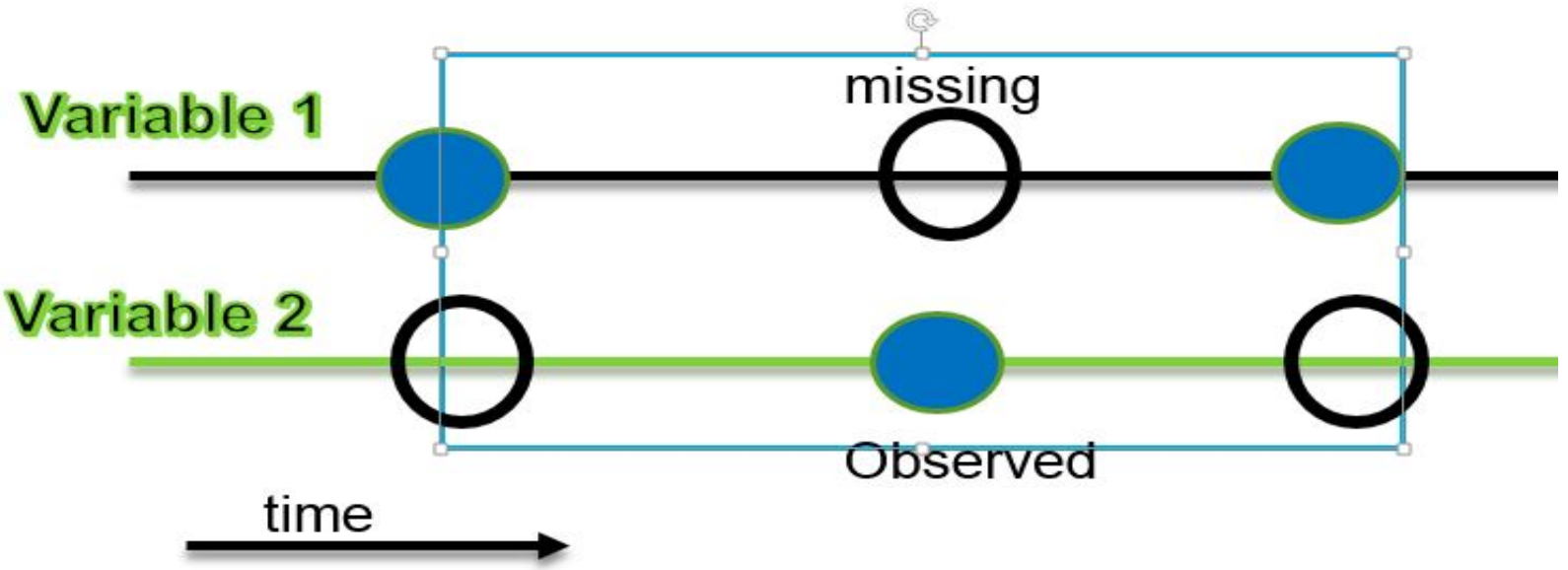
- Two correlated processes, measurements at different times. Blue=observed.

Example: asynchronous measurements



- If this is Markovian, in order to impute the value A, can I restrict to measurements in green window? Not even!

Example: Asynchronous measurements



- If we had observed values everywhere, Markov property allows us to restrict to the smaller (blue) window containing nearest neighbours.
- Imputed observations allows us to reduce the size of the (immediate) problem. So use imputed neighbouring missing values ("o") in blue rectangle to simulate value at point A . Move along and repeat.

Gibb's Sampling..... A Dummy's guide

Begin with completed data set: missing values imputed arbitrarily (e.g. mean, or purely random)

Gibbs Sampling[5]: Repeat many times:

Move about (systematically? at random?)

1. Draw parameter values from their posterior distribution given the complete data (or estimate parameters from completed data set as in stochastic EM algorithm)
2. Impute a missing value using the conditional distribution given the neighbours and the parameters. Repeat for all missing values.

[5] Gilks et al. (1996)

Example: Gibbs Sampling

Consider a random walk $Y_t, t = 1, 2, ...10$ with N(μ,σ^2) increments. The observation at $t = 5$ is missing. The conditional distribution of Y_5 given Y_4, Y_6 is

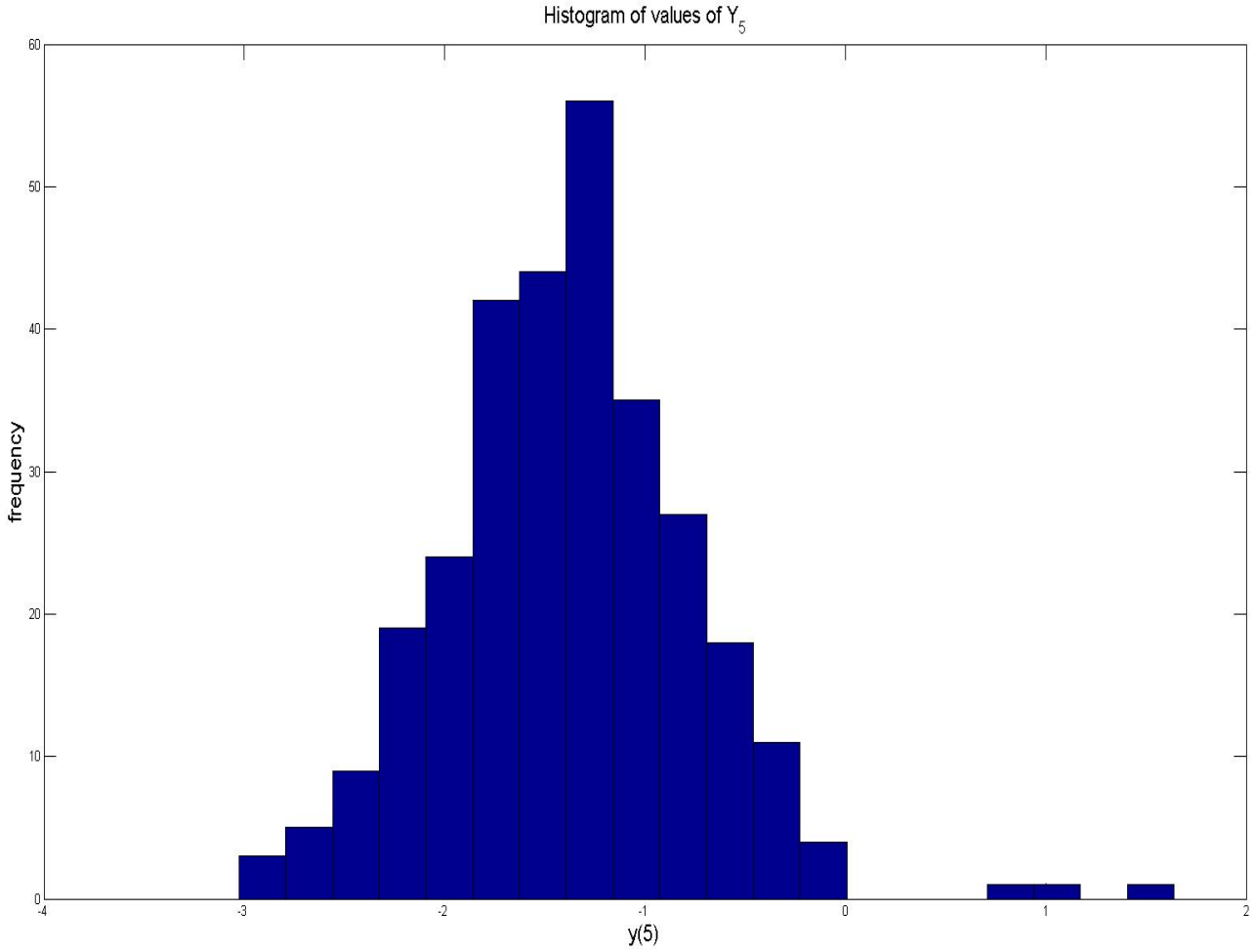
$$N\left(\frac{1}{2}(Y_6 + Y_4), \frac{\sigma^2}{2}\right)$$

so repeatedly:

1. Impute the value of $Y_5 \sim N\left(\frac{1}{2}(Y_6 + Y_4), \frac{\hat{\sigma}^2}{2}\right)$
2. Generate σ^2 from its posterior distribution. $\frac{1}{\sigma^2} \backsim Gamma\left(\frac{n-1}{2}, \frac{2}{(n-1)S^2}\right)$ where S^2 = sample variance of $X_i = Y_{i+1} - Y_i$.

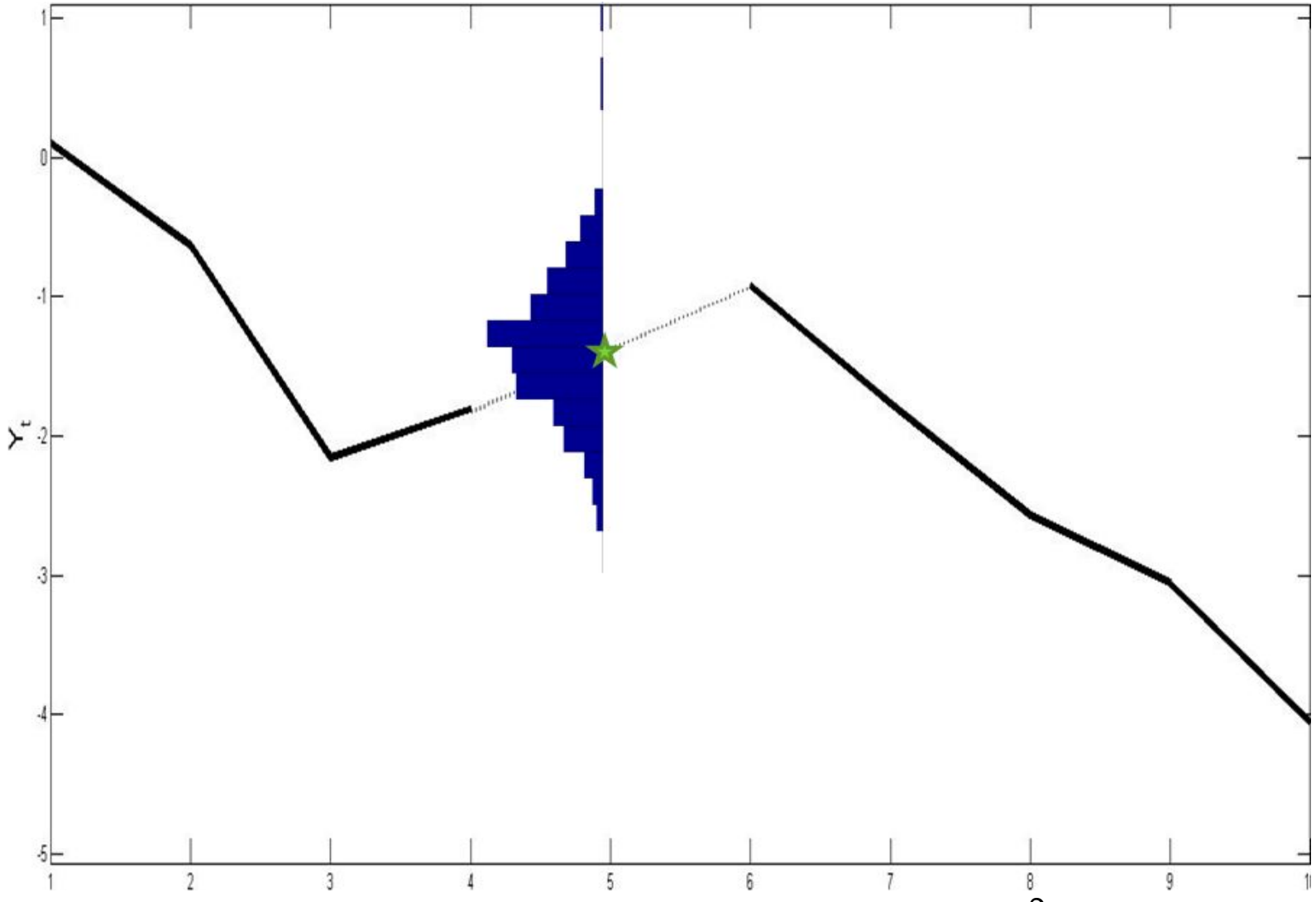
Movie: Gibbsdemo

Example: Gibbs Sampling - imputed values of Y



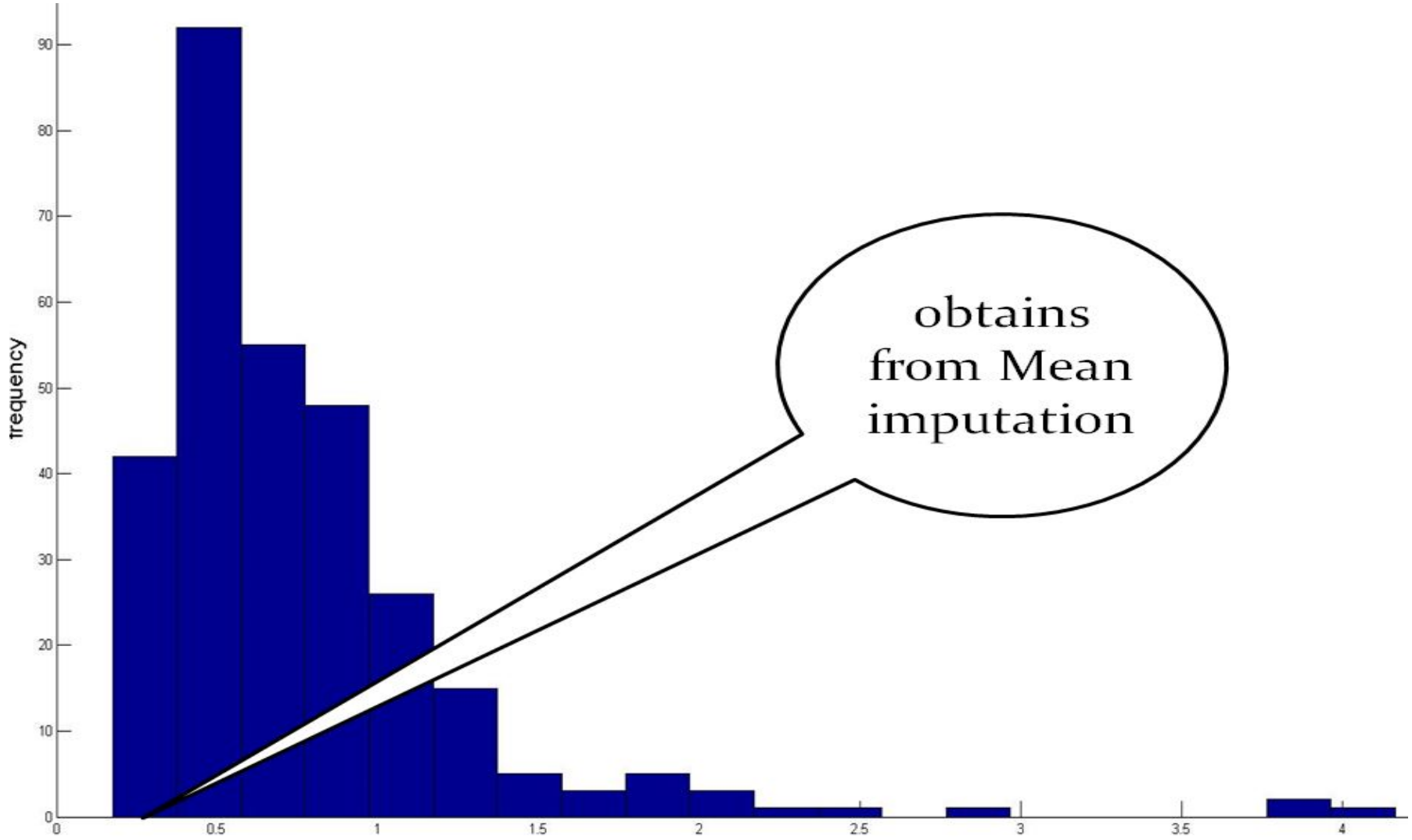
The imputed values of $Y_5 \sim N(\frac{1}{2}(Y_6 + Y_4), \frac{\sigma^2}{2})$

Example: Gibbs Sampling - imputed values of Y



The imputed values of $Y_5 \sim N(\frac{1}{2}(Y_6 + Y_4), \frac{\sigma^2}{2})$

Example: Gibbs Sampling



The values of σ^2 sampled from posterior

Example: Gibbs Sampling for a 2d Random Walk

Consider a 2 dimensional random walk $(\mathbf{Y}_t), t = 1, 2, ...10$ with bivariate N(μ,$\mathbf{\Sigma}$) increments.

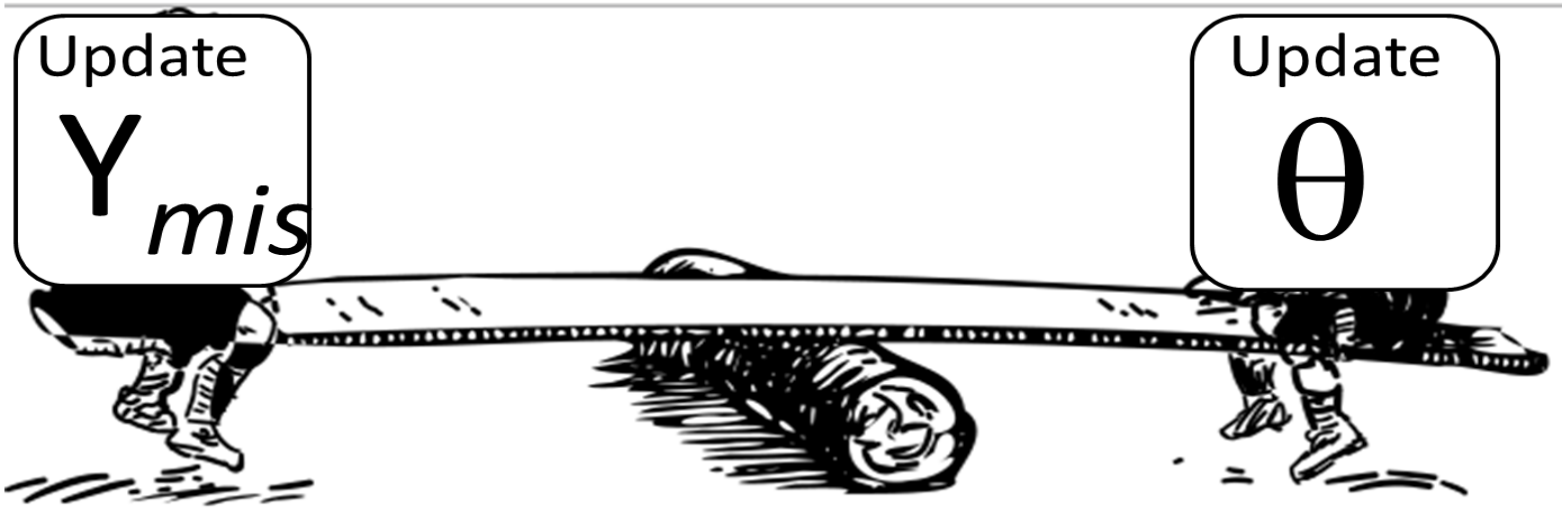
- The observation at $t = 5$ is missing.
- The conditional distribution of Y_5 given $\mathbf{Y}_4, \mathbf{Y}_6$ is

$$\mathbf{Y}_5 | \mathbf{Y}_4, \mathbf{Y}_6 \sim \mathbf{MVN}\left(\frac{1}{2}(\mathbf{Y}_6 + \mathbf{Y}_4), \frac{1}{2}\mathbf{\Sigma}\right)$$

- Repeatedly draw μ,Σ them from their posterior distribution , the Normal-inverse-Wishart (or re-estimate μ,Σ from completed data). and then impute $\mathbf{Y}_5$

Movie: Gibbsdemo2

What Did We Just Do?? The Gibbs Two-step.



1. Generated (imputed) $Y_{mis}|\mu, \Sigma$ using $P(Y_{mis}|\mu, \Sigma, Y_{obs})$.
2. **(Bayes)** Generate $\mu, \mathbf{\Sigma}$ from $p(\mu, \mathbf{\Sigma}|Y)$. MCMC theory shows (under some conditions) the process converges to its stationary distribution for which $\mu, \mathbf{\Sigma}$ are draws from the posterior distribution given the data Y_{obs}, and Y_{mis} are draws from the predictive distribution $P(Y_{mis}|\mu, \Sigma, Y_{obs})$.

The Gibbs Two-step.

1. Generate (impute) $Y_{mis}|\mu, \Sigma$ using $P(Y_{mis}|\mu, \Sigma, Y_{obs})$.
   As alternative to 2 above we may

2* Obtain the maximum likelihood estimate of
   $\mu, \Sigma|Y_{mis}, Y_{obs}$

This 2-step algorithm (Gibbs Sampling) is a special case of Markov Chain Monte Carlo. It lies at the heart of most of missing data imputation. The Bayes version is often called Data Augmentation[6].

[6] Tanner and Wong, 1987

Gibbs alternative to mean imputation: simulate parameters, impute unobserved data

- Imputation requires some model for the data, observed and missing.
- If our interest is in a specific parameter we could use EM algorithm (e.g. mean imputation). Using simulated values rather than the mean this is the "stochastic EM" algorithm.
- Simulating both parameters and missing values honours both the random characteristics of the data and the uncertainty concerning the parameter value. Under the correct model, we obtain consistent estimators of all parameters.

Generating from a difficult joint distribution: The Hammersley-Clifford Theorem

- A joint distribution is completely determined by its conditional distributions.
- We can generate from a joint distribution in small bits: i.e. generate from each of the conditional distributions of Y_i given the others. i.e.
  - generate $Y_1 | Y_2, Y_2, ..., Y_d$, then
  - generate $Y_2 | Y_1, Y_3, ..., Y_d$,
  - generate $Y_3 | Y_1, Y_2, Y_4...., Y_d$, etc.

Question: www.socrative.com (calgary)

Suppose we begin with arbitrary values of $(Y_1 = 1, Y_2 = 0)$ and then repeatedly generate:

1. $Y_1 = \frac{1}{2}Y_2 + \varepsilon$ where ε is an independent N(0,1) random variables.
2. $Y_2 = \frac{1}{2}Y_1 + \varepsilon$ where $\varepsilon \sim$ independent N(0,1).

Then after 1000 iterations of steps 1&2 the pair of values generated (Y_1, Y_2)

A. Will explode to infinity
B. Will converge to a point.
C. Will resemble independent normal random variables
D. Will resemble correlated normal random variables with correlation $\frac{1}{2}$.

Variations on Gibbs in higher dimension

- Suppose process X is d-dimensional, $d > 2$.
- In general we adopt an order in which the d variables are drawn and updated. The original order $1, 2, 3, \ldots d$ called systematic scan.
- If we draw a random component from $\{1, 2, 3, \ldots d\}$ to determine next variable sampled called Random-scan (this chain is reversible)
- Updating highly correlated components jointly appears to result in faster convergence to stationary distribution.
- Often updating occurs in blocks of components

Poor (Wo)man's Algorithm for dealing with incomplete data revisited

Begin with a model and a preliminary estimate of the parameters

When you wish to estimate a parameter in the presence of incomplete data, write down the complete-data estimator.

1. Simulate any missing values required for the estimate using the conditional distribution given the observed data and the current parameter value.
2. Calculate the value of the estimate treating the simulated values as if they were observed. Alternatively draw the parameters from the posterior distribution.

Repeat steps 1-2 until convergence.

You may use as a parameter point estimator the average of the all values obtained on step 2 (perhaps leaving off break-in period)

Numerical Shortcuts for Gaussian models

See Schafer (1997). The conditional distributions are all regression models. There are numerical shortcuts to obtaining the conditional parameters.

Sweeping: Now not just for curlers!

Sweep Operator:[7] A computational efficient way to generate parameters of conditional distribution for multivariate normal. Operates on one row of symmetric matrix at a time.

$$A = SWP[k]G :$$

$$A_{jl} \leftarrow \begin{cases} G_{jl} - \frac{G_{jl}G_{kl}}{G_{kk}} & \text{for} \quad j \neq k, l \neq k \\ \frac{G_{jl}}{G_{kk}} & \text{for} \quad j \neq k, l = k \text{ or } j = k, l \neq k \\ -\frac{1}{G_{kk}} & \text{for} \quad j = l = k \end{cases}$$

SWP[1,2,...,k] indicates sweeping on rows 1 to k successively.

[7] see Schafer (1997) P. 159

Sweeping up the regression coefficients

Suppose $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is MVN$\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$.

where Σ_{11} is $p_1 \times p_1$. Then

$$SWP[2,\ldots 1+p_1]\begin{bmatrix} -1 & \mu_1^T & \mu_2^T \\ \mu_1 & \Sigma_{11} & \Sigma_{12} \\ \mu_2 & \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} -1-\mu_1^T\Sigma_{11}^{-1}\mu_1 & * & * \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} & * \\ \alpha_{2\bullet 1} & \beta_{2\bullet 1} & \Sigma_{22\bullet 1} \end{bmatrix}$$

provides the parameters of the distribution $X_2 \mid X_1$. Here

$X_2 \mid X_1$ has distribution of $\alpha_{2\bullet 1} + \beta_{2\bullet 1} X_1 + N(0, \Sigma_{22\bullet 1})$

Conditional mean

conditional covariance

Read these from the swept matrix

Poor (Wo)man's Algorithm for dealing with incomplete data revisited yet again

Begin with a model and a preliminary estimate of the parameters
When you wish to estimate a parameter in the presence of incomplete data, write down the complete-data estimator.

1. Simulate any missing values required for the estimate using the conditional distribution given the observed data and the current parameter value.
2. Calculate the value of the estimate treating the simulated values as if they were observed. Alternatively draw the parameters from the posterior distribution.

Repeat steps 1- until the convergence.
You may use as a parameter point estimator the average of the all values obtained on step 2 (perhaps leaving off break-in period)

Fully Conditional Specification versus joint modelling

Gibbs sampling or data augmentation assumed that we begin with a joint distribution and generate from conditionals derived from it.

- What if our "model' consists **only** of specified conditional distributions for the missing variables given the others.
- Like specifying regression relationships among a set of variables in place of a multivariate normal assumption.
- Are these consistent with any multivariate distribution?

Imputation by Chained Equations

Use fully conditional specification to describe your "model'

1. Impute values of any unknowns in the estimator under the assumed conditional distribution given the observed data and the current parameter value.
2. Calculate the value of the parameter estimator treating the imputed values as if they were observed, or draw parameters from posterior distribution given completed data.

Fully Conditional Specification: Pros

- A suitable joint model often hard to specify.
  Instead, we specify generating mechanisms (regression relationships) e.g.
  $P(Y_1|Y_3, Y_4), P(Y_2|Y_1, Y_3, Y_4), P(Y_3|Y_2, Y_4)$ **not** obtained from a joint distribution and run Gibbs using these to impute values.
- No laborious calculation of conditional distributions.
- In a model for Y_j **you** control which other variables are important to be included and how.
- Allows some variables to be continuous (e.g. normal) and others discrete/dichotomous. (e.g. use logistic regression). For example you might assume $Y_1|Y_2, Y_3$ is $N(\alpha + \beta Y_2, \sigma^2)$ and
  $logit(P(Y_3 = 1)) = a + b_1 Y_1 + b_2 Y_2$.

Fully Conditional Specification: Cons

1. Your "model' may **not** be a model. The conditional distributions may not be compatible with any joint distribution.
2. May be subject to "feedback loops", converge slowly, fail to converge, loop, explode.
3. Will this converge to a stationary distribution? Hard to assess convergence.
4. If it converges, the joint distribution is described "numerically". Harder to understand.

An underdeveloped alternative: copula-based imputation models.

Convergence. A nice place if you can get there....

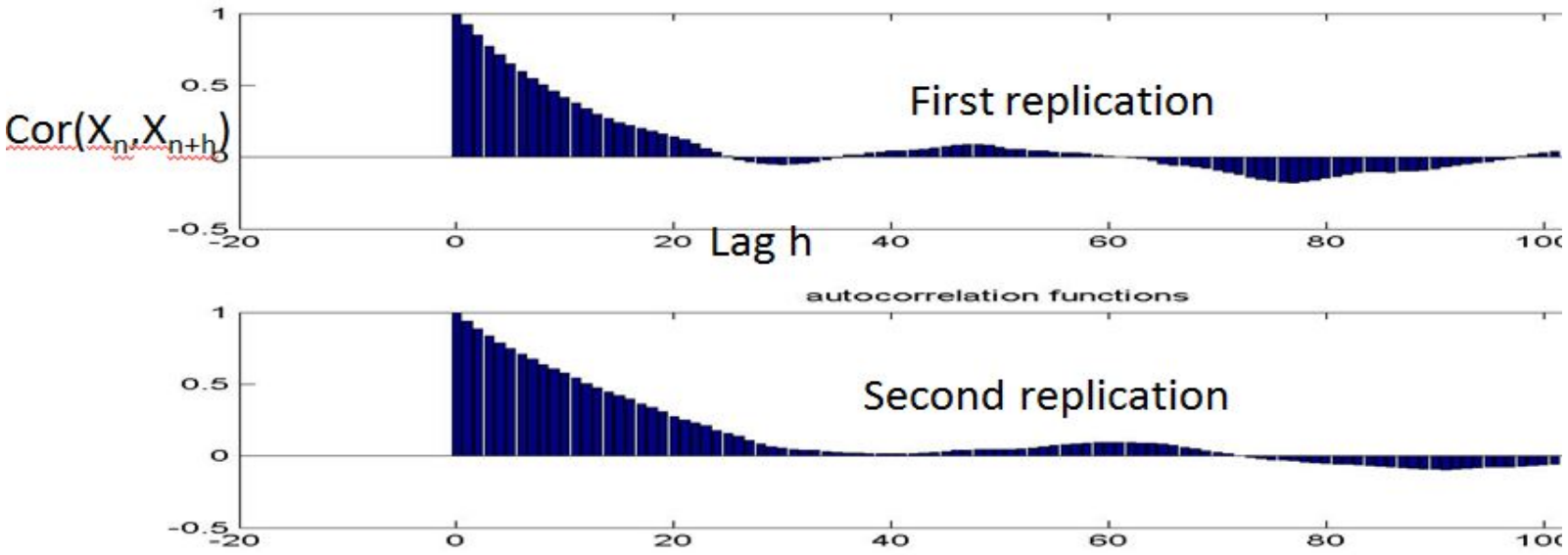
- Convergence in distribution is very difficult to assess.
- The early iterations reflect (retain a memory of) the starting value, and these are called the burn-in.
- After the burn-in, we say the chain has converged (poor language). We omit the burn-in period from averages.
- Methods for determining an appropriate length of the burn-in period called convergence diagnostics.

But how do I know I have converged?....

- One way of determining whether the chain is close to stationarity is to run parallel chains with different starting values. How many?
- Several long runs (Gelman and Rubin,1992) helps to indicate convergence and provides an estimator of standard error.
- One very long run (Geyer, 1992) reaches regions in the sample space a shorter run might not reach.

Autocorrelation Function?

- We only have a draw from the stationary distribution $f(y)$ once the Markov chain has converged to its stationary distribution.
- Can use visual inspection of autocorrelation function to see whether the effect of the initial condition on the chain have died out.



Different Starting Values?

- Can use visual inspection of process with different starting values to see whether the effect of the initial condition on the chain have died out.
- For example, two chains, one starting with 3 the other with -3.

In general, for multiple chains....

Delete burn-in period, then:

- Compare quantities "between" chains and "within" chains.
- e.g. Compare means, variances, kernel density estimates.
- Hard to formally compare (since formal exact tests not available)

For a single (long chain)

Divide the chain into a number of long blocks and pretend they are independent chains.

- The Geweke statistic[8] compares the mean of the first 10% of the chain with that of the last 50%.
- Could also use the variance or another moment.
- Can also plot; the cumulative median, and upper and lower x% intervals, moving averages, empirical characteristic functions?

[8] Geweke, J. 1992.

Gelman-Rubin Statistics-I

- Conduct multiple (n) MCMC chains each with different starting values.
- Select a set of quantities of interest, exclude the "burn-in" period and thin the chain (select every $m'th$ value).
- Compares the mean of the empirical variance within chains to the variance of the mean across the chains.

see:
http://faculty.washington.edu/rayh/558_2004/ppt_files_lectures/Diag.PPT

One Long Run or Many Short Runs?

- Many short runs allow a fairly direct check on whether convergence has occurred. However this depends on good coverage for the set of initial parameter vectors.
- Many (too) short runs may never reach regions in the sample space.
- Try to conduct many (5-10?) runs.

Other Statistics:

- Heidelberger-Welsh: tests for stationarity of the chain.
- Raftery-Lewis: based on how many iterations are necessary to estimate the posterior for a given quantity.

see:

http://faculty.washington.edu/rayh/558_2004/ppt_files_lectures/Diag.PPT

Convergence and Faith

- *Conventional wisdom* says not to worry about convergence[9], if you can avoid periodicity and "getting stuck" in a portion of the space. Worry just enough.
- Convergence of a FCS. Half Theorem, Half Faith.

[9] technically this requires irredicible, aperiodic and recurrence.