Atypicality for Vector Gaussian Models

Elyas Sabeti, Anders Høst-Madsen

- Dept. of EE, University of Hawaii

Partly funded by NSF grant CCF 1434600

Motivation

Motivation

- BIG Data generates huge amounts of data
  - Medical sensors
  - Genetics
  - Surveillance: NSA
  - Environmental sensors

Motivation

- BIG Data generates huge amounts of data
  - Medical sensors
  - Genetics
  - Surveillance: NSA
  - Environmental sensors
- Often data is just stored, not being used

Motivation

- BIG Data generates huge amounts of data
  - Medical sensors
  - Genetics
  - Surveillance: NSA
  - Environmental sensors
- Often data is just stored, not being used
- What to use data for?
  - Statistics of “typical” data, “averages”
  - But perhaps what is interesting is the unique, rare event deviating from the norm, the atypical data
  - Art, scientific work (“genius”), entrepeneurship

Motivation

- BIG Data generates huge amounts of data
  - Medical sensors
  - Genetics
  - Surveillance: NSA
  - Environmental sensors
- Often data is just stored, not being used
- What to use data for?
  - Statistics of “typical” data, “averages”
  - But perhaps what is interesting is the unique, rare event deviating from the norm, the atypical data
  - Art, scientific work (“genius”), entrepeneurship

Motivation

- BIG Data generates huge amounts of data
  - Medical sensors
  - Genetics
  - Surveillance: NSA
  - Environmental s
- Often data is jus
- What to use dat
  - Statistics of "typi
  - But perhaps wha
    deviating from t
  - Art, scientific wo

Applications

- Medical
  - Most sensor data is indicative of normal
  - The rare event is indicative of decease
- Other
  - Gambling fraud or malfunction
  - Credit card fraud
  - Accounting, IRS
  - Computer network intrusion
  - Environmental monitoring
  - Electric power grids
  - Plant monitoring

⋮

Anomaly Detection with Universal Source Coding

Anomaly Detection with Universal Source Coding

- Atypical data can be thought of as anomalies
  - But more general application: data discovery

Anomaly Detection with Universal Source Coding

- Atypical data can be thought of as anomalies
  - But more general application: data discovery
- Looking for “unknown unknowns”
  - Need universal approach → information theory / universal source coding

Anomaly Detection with Universal Source Coding

- Atypical data can be thought of as anomalies
  - But more general application: data discovery
- Looking for "unknown unknowns"
  - Need universal approach → information theory / universal source coding
- Aim
  - Theoretically well-founded approach to anomaly detection with information theory

Is Information Theory Useful?

Is Information Theory Useful?

- Is information theory fundamental?

Is Information Theory Useful?

- Is information theory fundamental?
  - Entropy $H(X) \rightarrow$ Shortest codelength
  - Mutual Information $I(X;Y) \rightarrow$ Channel capacity

Is Information Theory Useful?

- Is i
  - E
  - M

A Mathematical Theory of Communication

By C. E. SHANNON

Introduction

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

Is Information Theory Useful?

- Is information theory fundamental?
  - Entropy $H(X) \rightarrow$ Shortest codelength
  - Mutual Information $I(X;Y) \rightarrow$ Channel capacity

Is Information Theory Useful?

- Is information theory fundamental?
  - Entropy $H(X) \rightarrow$ Shortest codelength
  - Mutual Information $I(X;Y) \rightarrow$ Channel capacity
- Minimum Descriptive Length (MDL)

Is Information Theory Useful?

- Is information theory fundamental?
  - Entropy $H(X) \rightarrow$ Shortest codelength
  - Mutual Information $I(X;Y) \rightarrow$ Channel capacity
- Minimum Descriptive Length (MDL)
  - Used to estimate model order in SP

Is Information Theory Useful?

- Is information theory fundamental?
  - Entropy $H(X) \rightarrow$ Shortest codelength
  - Mutual Information $I(X;Y) \rightarrow$ Channel capacity
- Minimum Descriptive Length (MDL)
  - Used to estimate model order in SP
  - But our thinking is that if the MDL of model A is shorter than the MDL of model B, model A describes the data better

Is Information Theory Useful?

- Is information theory fundamental?
  - Entropy $H(X) \rightarrow$ Shortest codelength
  - Mutual Information $I(X;Y) \rightarrow$ Channel capacity
- Minimum Descriptive Length (MDL)
  - Used to estimate model order in SP
  - But our thinking is that if the MDL of model A is shorter than the MDL of model B, model A describes the data better
    - Model A is <u>fundamentally</u> more <u>meaningful</u>

Is Information Theory Useful?

- Is information theory fundamental?
  - Entropy $H(X) \rightarrow$ Shortest codelength
  - Mutual Information $I(X;Y) \rightarrow$ Channel capacity
- Minimum Descriptive Length (MDL)
  - Used to estimate model order in SP
  - But our thinking is that if the MDL of model A is shorter than the MDL of model B, model A describes the data better
    - Model A is <u>fundamentally</u> more <u>meaningful</u>
- This work is based on an assumption that information is fundamental

Is Information Theory Useful?

- Is information theory fundamental?
  - Entropy $H(X) \rightarrow$ Shortest codelength
  - Mutual Information $I(X;Y) \rightarrow$ Channel capacity
- Minimum Descriptive Length (MDL)
  - Used to estimate model order in SP
  - But our thinking is that if the MDL of model A is shorter than the MDL of model B, model A describes the data better
    - Model A is <u>fundamentally</u> more <u>meaningful</u>
- This work is based on an assumption that information is fundamental
  - Information measure is not <u>a</u> measure but <u>the</u> measure

Kolmogorov-Martin Löf Randomness

Kolmogorov-Martin Löf Randomness

- Infinite sequence of bits 10011011010100001…

Kolmogorov-Martin Löf Randomness

- Infinite sequence of bits 10011011010100001…
- When is the sequence truly (iid uniform) random?
  - 50 years of failed attempts
  - Solved by Martin-Löf in 1966

Kolmogorov-Martin Löf Randomness

- Infinite sequence of bits 10011011010100001…
- When is the sequence truly (iid uniform) random?
  - 50 years of failed attempts
  - Solved by Martin-Löf in 1966
- Kolmogorov
  - Typical sequences: truly random sequence
  - Special sequences: other sequences

Kolmogorov-Martin Löf Randomness

- Infinite sequence of bits 10011011010100001…
- When is the sequence truly (iid uniform) random?
  - 50 years of failed attempts
  - Solved by Martin-Löf in 1966
- Kolmogorov
  - <u>Typical</u> sequences: truly random sequence
  - <u>Special</u> sequences: other sequences
- Random Sequence

$$\exists c > 0 \forall n > 1 : K(x[1], \ldots, x[n]) \geq n - c$$

Kolmogorov-Martin Löf Randomness

Kolmogorov-Martin Löf Randomness

- Finite sequence of bits 10011011010100001

Kolmogorov-Martin Löf Randomness

- Finite sequence of bits 10011011010100001
- Algorithmically random if

$$K(x[1], \ldots, x[n] | n) \geq n$$

Kolmogorov-Martin Löf Randomness

- Finite sequence of bits 10011011010100001
- Algorithmically random if

$$K(x[1], \ldots, x[n]|n) \geq n$$

- Kolmogorov's terms
  - Typical $K(x[1], \ldots, x[n]|n) \geq n$
  - Special $K(x[1], \ldots, x[n]|n) < n \quad \rightarrow$ Atypical

Kolmogorov-Martin Löf Randomness

- Finite sequence of bits 10011011010100001
- Algorithmically random if

$$K(x[1],\ldots,x[n]|n) \geq n$$

- Kolmogorov's terms
  - Typical $K(x[1],\ldots,x[n]|n) \geq n$
  - Special $K(x[1],\ldots,x[n]|n) < n \quad \rightarrow$ Atypical
- Coding theory
  - If random, incompressible, identity coder optimum → Typical
  - If (universal) source coder can compress → Atypical

Kolmogorov-Martin Löf Randomness

- Finite sequence of bits 10011011010100001
- Algorithmically random if
$$K(x[1],\ldots,x[n]|n) \geq n$$
- Kolmogorov's terms
  - Typical $K(x[1],\ldots,x[n]|n) \geq n$
  - Special $K(x[1],\ldots,x[n]|n) < n \quad \rightarrow$ Atypical
- Coding theory
  - If random, incompressible, identity coder optimum → Typical
  - If (universal) source coder can compress → Atypical

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

Atypicality

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

Atypicality

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

Atypicality

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

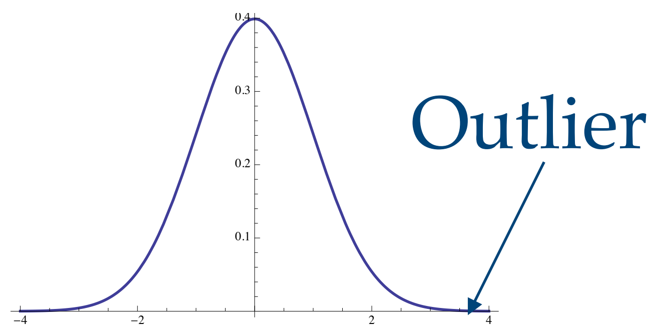
$$C_t(x) - C_a(x) > 0$$

Atypicality

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

$$C_t(x) - C_a(x) > 0$$

- Outlier detection
  - Low likelihood, rarity: $C_t(x)$ large

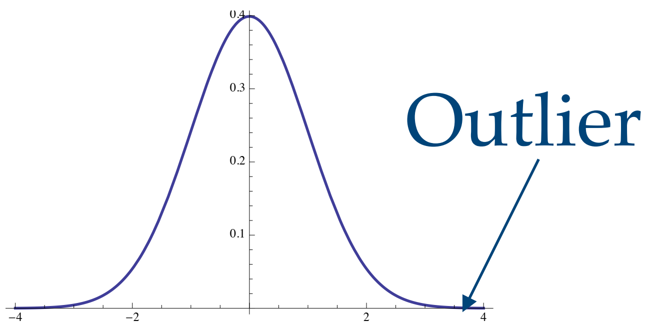


Atypicality

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

$$C_t(x) - C_a(x) > 0$$

- Outlier detection
  - Low likelihood, rarity: $C_t(x)$ large
- Iid random case



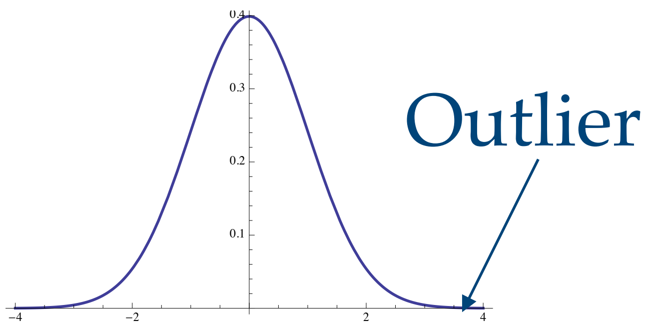

```
10011011010100001
11111111111111111
```

Atypicality

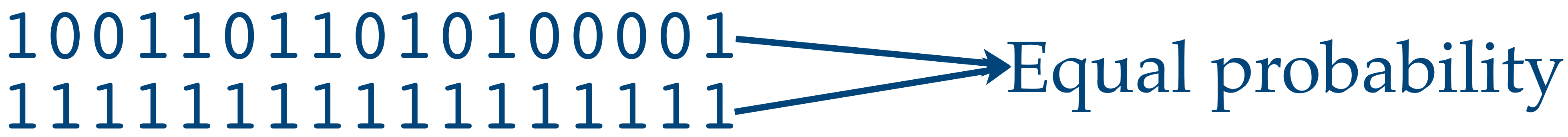
A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

$$C_t(x) - C_a(x) > 0$$

- Outlier detection
  - Low likelihood, rarity: $C_t(x)$ large



- Iid random case

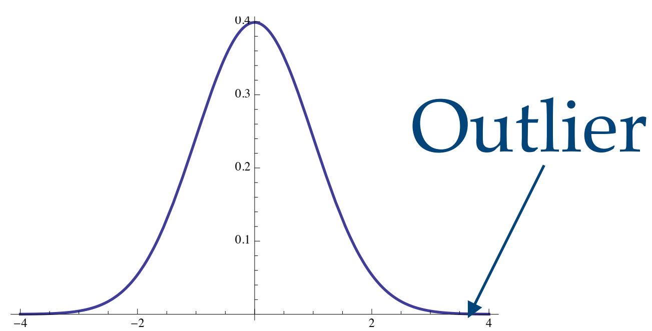


Atypicality

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

$$C_t(x) - C_a(x) > 0$$

- Outlier detection
  - Low likelihood, rarity: $C_t(x)$ large



- Iid random case

```
1001101101010001
1111111111111111
```

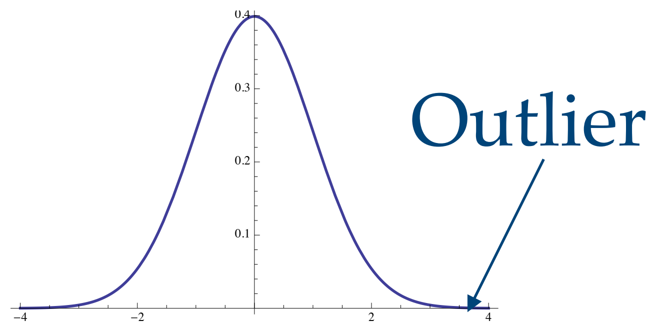
Equal probability

  - $C_t(x)$ same, but $C_a(x)$ different

Atypicality

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

$$C_t(x) - C_a(x) > 0$$



- Outlier detection
  - Low likelihood, rarity: $C_t(x)$ large
- Iid random case

```
10011011010100001
11111111111111111
```

→ Equal probability

  - $C_t(x)$ same, but $C_a(x)$ different
- Also prioritizes these cases
  - The larger $C_t(x) - C_a(x)$ the more atypical

Binary IID sequences

1000011010011111111111100001010101111110001

Binary IID sequences

- Default law: $P(0)=1-p$, $P(1)=p$, p known

10000110100111111111111000101010111110001

Binary IID sequences

- Default law: $P(0)=1-p$, $P(1)=p$, p known
  - Codelength $L(l) = l\left(\hat{p}\log\frac{1}{p} + (1-\hat{p})\log\frac{1}{1-p}\right), \hat{p} = \frac{1}{l}\sum X_i$

1000011010011111111111100010101011110001

Binary IID sequences

- Default law: $P(0)=1-p$, $P(1)=p$, p known
  - Codelength $L(l) = l\left(\hat{p}\log\frac{1}{p} + (1-\hat{p})\log\frac{1}{1-p}\right), \hat{p} = \frac{1}{l}\sum X_i$
- Alternative law: $P(1)\neq p$

1000011010011111111111000101010111110001

Binary IID sequences

- Default law: $P(0)=1-p$, $P(1)=p$, p known
  - Codelength $L(l) = l\left(\hat{p}\log\frac{1}{p} + (1-\hat{p})\log\frac{1}{1-p}\right), \hat{p} = \frac{1}{l}\sum X_i$
- Alternative law: $P(1)\neq p$
  - Universal source code from Cover's book

1000011010011111111111110001010101111110001

Binary IID sequences

- Default law: $P(0)=1-p$, $P(1)=p$, p known
  - Codelength $L(l) = l\left(\hat{p}\log\frac{1}{p} + (1-\hat{p})\log\frac{1}{1-p}\right), \hat{p} = \frac{1}{l}\sum X_i$
- Alternative law: $P(1)\neq p$
  - Universal source code from Cover's book
  - Codelength $L_{\hat{p}}(l) = lH(\hat{p}) + \frac{1}{2}\log l$

1 0 0 0 0 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 0 1 0 1 0 1 1 1 1 0 0 0 1

Binary IID sequences

- Default law: $P(0)=1-p$, $P(1)=p$, p known
  - Codelength $L(l) = l\left(\hat{p}\log\frac{1}{p} + (1-\hat{p})\log\frac{1}{1-p}\right), \hat{p} = \frac{1}{l}\sum X_i$
- Alternative law: $P(1)\neq p$
  - Universal source code from Cover's book
  - Codelength $L_{\hat{p}}(l) = lH(\hat{p}) + \frac{1}{2}\log l$
- Need to tell beginning and end

1000011010011111111111000101010111110001

Binary IID sequences

- Default law: $P(0)=1-p$, $P(1)=p$, p known
  - Codelength $L(l) = l\left(\hat{p}\log\frac{1}{p} + (1-\hat{p})\log\frac{1}{1-p}\right), \hat{p} = \frac{1}{l}\sum X_i$
- Alternative law: $P(1)\neq p$
  - Universal source code from Cover's book
  - Codelength $L_{\hat{p}}(l) = lH(\hat{p}) + \frac{1}{2}\log l$
- Need to tell beginning and end

10000110100.11111111111000101010111110001

$\xrightarrow{\quad l \quad}$

Binary IID sequences

- Default law: $P(0)=1-p$, $P(1)=p$, p known
  - Codelength $L(l) = l\left(\hat{p}\log\frac{1}{p} + (1-\hat{p})\log\frac{1}{1-p}\right), \hat{p} = \frac{1}{l}\sum X_i$
- Alternative law: $P(1)\neq p$
  - Universal source code from Cover's book
  - Codelength $L_{\hat{p}}(l) = lH(\hat{p}) + \frac{1}{2}\log l$
- Need to tell beginning and end
  - Cost of encoding '.': $\tau = \log\frac{1}{P('.')}$

10000110100.111111111111000101010111110001

$\xrightarrow{\quad l \quad}$

Binary IID sequences

- Default law: $P(0)$=1-p, $P(1)$=p, p known
  - Codelength $L(l) = l\left(\hat{p}\log\frac{1}{p} + (1-\hat{p})\log\frac{1}{1-p}\right), \hat{p} = \frac{1}{l}\sum X_i$
- Alternative law: $P(1)\neq p$
  - Universal source code from Cover's book
  - Codelength $L_{\hat{p}}(l) = lH(\hat{p}) + \frac{1}{2}\log l$
- Need to tell beginning and end
  - Cost of encoding '.': $\tau = \log\frac{1}{P('.')}$
  - Cost of encoding length (Rissanen, Elias):

$$\log^*(l) = \log l + \log\log l + \log\log\log l + \cdots$$

10000110100.111111111111000101010111110001

$\xrightarrow{\quad\quad}$ l

Binary IID sequences

- Default law: $P(0)=1-p$, $P(1)=p$, p known
  - Codelength $L(l) = l\left(\hat{p}\log\frac{1}{p} + (1-\hat{p})\log\frac{1}{1-p}\right), \hat{p} = \frac{1}{l}\sum X_i$
- Alternative law: $P(1)\neq p$
  - Universal source code from Cover's book
  - Codelength $L_{\hat{p}}(l) = lH(\hat{p}) + \frac{1}{2}\log l$
- Need to tell beginning and end
  - Cost of encoding '.': $\tau = \log\frac{1}{P('.')}$
  - Cost of encoding length (Rissanen, Elias):
    $$\log^*(l) = \log l + \log\log l + \log\log\log l + \cdots$$
- Total codelength

$$L_{\hat{p}}(l) = lH(\hat{p}) + \frac{3}{2}\log l + \tau$$

Binary IID sequences

- Default law: P(0)=1-p, P(1)=p, p known
  - Codelength $L(l) = l\left(\hat{p}\log\frac{1}{p} + (1-\hat{p})\log\frac{1}{1-p}\right), \hat{p} = \frac{1}{l}\sum X_i$
- Alternative law: P(1)≠p
  - Codelength: $L_{\hat{p}}(l) = lH(\hat{p}) + \dfrac{3}{2}\log l + \tau$

Binary IID sequences

- Default law: P(0)=1-p, P(1)=p, p known
  - Codelength $L(l) = l\left(\hat{p}\log\frac{1}{p} + (1-\hat{p})\log\frac{1}{1-p}\right), \hat{p} = \frac{1}{l}\sum X_i$
- Alternative law: P(0)≠p
  - Codelength: $L_{\hat{p}}(l) = lH(\hat{p}) + \dfrac{3}{2}\log l + \tau$
- Atypicality criterion

$$D(\hat{p}\|p) > \frac{\tau + \frac{3}{2}\log l}{l}$$

Theoretical Analysis

- The probability P_A that a sequence of length l is classified as atypical is bounded by

$$P_A \leq 2^{-\tau+1}\frac{1}{l^{3/2}}K(l,\tau), \quad \forall\tau : \lim_{l\to\infty} K(l,\tau) = 1$$

Theoretical Analysis

- The probability P_A that a sequence of length l is classified as atypical is bounded by

$$P_A \leq 2^{-\tau+1} \frac{1}{l^{3/2}} K(l,\tau), \quad \forall \tau : \lim_{l\to\infty} K(l,\tau) = 1$$

- Consider the case $p=\frac{1}{2}$. The probability $P_A(X_n)$ that a given sample X_n is part of an atypical subsequence of any length is upper bounded by

$$P_A(X_n) \leq (K_1\sqrt{\tau} + K_2)2^{-\tau}$$

for some constants K_1, K_2

Real-Valued Data

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

Real-Valued Data

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

Real-Valued Data

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

- Generalization to real valued data
  - Definition based on exact encoding, not rate-distortion

Real-Valued Data

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

- Generalization to real valued data
  - Definition based on exact encoding, not rate-distortion
- Exact encoding of real-valued data
  - Lossless audio coding (MPEG-4 ALS, Apple Lossless)

Real-Valued Data

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

- Generalization to real valued data
  - Definition based on exact encoding, not rate-distortion
- Exact encoding of real-valued data
  - Lossless audio coding (MPEG-4 ALS, Apple Lossless)
- Abstract encoding
  - Fixed point, r bits after ., unlimited bits prior
  - Codelength (Rissanen)

$$L(x) = -\log \int_x^{x+2^{-r}} f(t)dt \approx -\log(f(x)) + r$$

Real-Valued Data

$$L(x) = -\log \int_{x}^{x+2^{-r}} f(t)dt \approx -\log(f(x)) + r$$

Real-Valued Data

- Abstract encoding
  - Fixed point, r bits after ., unlimited bits prior
  - Codelength (Rissanen)

$$L(x) = -\log \int_{x}^{x+2^{-r}} f(t)dt \approx -\log(f(x)) + r$$

Real-Valued Data

- Abstract encoding
  - Fixed point, r bits after ., unlimited bits prior
  - Codelength (Rissanen)

$$L(x) = -\log \int_x^{x+2^{-r}} f(t)dt \approx -\log(f(x)) + r$$

- Only need comparison of codelengths
  - r cancels out
  - Can let $r \rightarrow \infty$, L(x)=-log(f(x))

Real-Valued Data

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

$$L(x) = -\log \int_x^{x+2^{-r}} f(t)dt \approx -\log(f(x)) + r$$

- Only need comparison of codelengths
  - r cancels out
  - Can let $r \to \infty$, $L(x) = -\log(f(x))$

Real-Valued Data

A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code designed for typical sequences.

$$L(x) = -\log \int_x^{x+2^{-r}} f(t)dt \approx -\log(f(x)) + r$$

- Only need comparison of codelengths
  - r cancels out
  - Can let $r \rightarrow \infty$, L(x)=-log(f(x))
- Parametric model $f(\mathbf{x}|\boldsymbol{\theta})$
  - Need to encode data and parameters
  - Rissanen's MDL: $L = -\log f(\mathbf{x}|\hat{\boldsymbol{\theta}}_{\text{ML}}) + \frac{k}{2}\log l$

Vector Gaussian case

- Model
$$\mathbf{x}[n] = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{w}[n]$$
where $\mathbf{w}[n] \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \mathbf{s}(\boldsymbol{\theta}) k$-parameter
- Used to find atypical *relationships* between data streams
- **Theorem**: Probability of intrinsically atypical sequence
$$\limsup_{l\to\infty} \frac{\ln P_A(l)}{\frac{k+2}{2}\ln l} \leq 1$$
- Or
$$P_A(l) \lesssim l^{\frac{k+2}{2}}$$

Theoretical Analysis

- The probability P_A that a sequence of length l is classified as atypical is bounded by

$$P_A \leq 2^{-\tau+1} \frac{1}{l^{3/2}} K(l,\tau), \quad \forall \tau : \lim_{l \to \infty} K(l,\tau) = 1$$

Vector Gaussian case

- Model

$$\mathbf{x}[n] = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{w}[n]$$

  where $\mathbf{w}[n] \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \mathbf{s}(\boldsymbol{\theta})$ k-parameter
- Used to find atypical *relationships* between data streams
- **Theorem**: Probability of intrinsically atypical sequence

$$\limsup_{l \to \infty} \frac{\ln P_A(l)}{\frac{k+2}{2} \ln l} \leq 1$$

- Or

$$P_A(l) \lesssim l^{\frac{k+2}{2}}$$

Proof

- Atypicality criterion

$$r(\mathbf{x}) = -\log \frac{f(\mathbf{x}|\hat{\boldsymbol{\theta}})}{f(\mathbf{x}|\boldsymbol{\theta})} \geq \tau + \frac{k+2}{2}\log l$$

- Chernoff bound

$$P\left(r(\mathbf{x}) \geq \tau + \frac{k+2}{2}\log l\right) \leq \exp(-s(\tau + \frac{k+2}{2}\log l))M_r(s)$$

- Need to prove $M_r(s) = E[e^{sr}] \leq K < \infty$ independent of l for $s < \ln 2$

Proof

- Need to prove $M_r(s) = E[e^{sr}] \le K < \infty$ independent of l for $s < \ln 2$

$$
\begin{aligned}
-\ln \frac{p(\mathbf{x}|\hat{\boldsymbol{\theta}})}{p(\mathbf{x}|\boldsymbol{\theta})} &= \frac{1}{2}\sum_{n=1}^{l} \mathbf{x}[n]^T \boldsymbol{\Sigma}^{-1} \mathbf{x}[n] \\
&\quad - \frac{1}{2}\sum_{n=1}^{l} \left(\mathbf{x}[n] - \mathbf{s}(\hat{\boldsymbol{\theta}})\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}[n] - \mathbf{s}(\hat{\boldsymbol{\theta}})\right) \\
&\le \frac{1}{2l}\left(\sum_{n=1}^{l} \mathbf{x}[n]\right)^T \boldsymbol{\Sigma}^{-1} \left(\sum_{n=1}^{l} \mathbf{x}[n]\right)
\end{aligned}
$$

- Here $t = \sum_{n=1}^{l} \mathbf{x}[n]$ is sufficient statistic

Proof

- Need to prove $M_r(s) = E[e^{sr}] \leq K < \infty$ independent of l for $s < \ln 2$

$$E[e^{sr}] \leq \frac{1}{(2\pi)^{l/2}\sqrt{l\det\mathbf{\Sigma}}} \int \exp\left(\frac{s}{2l\ln 2}\mathbf{t}^T\mathbf{\Sigma}^{-1}\mathbf{t}\right)$$
$$\times \exp\left(-\frac{1}{2l}\mathbf{t}^T\mathbf{\Sigma}^{-1}\mathbf{t}\right) d\mathbf{t}$$
$$\leq K$$

- Here $t = \sum_{n=1}^{l} \mathbf{x}[n]$ is sufficient statistic

Example: S & P 500

Example: S & P 500

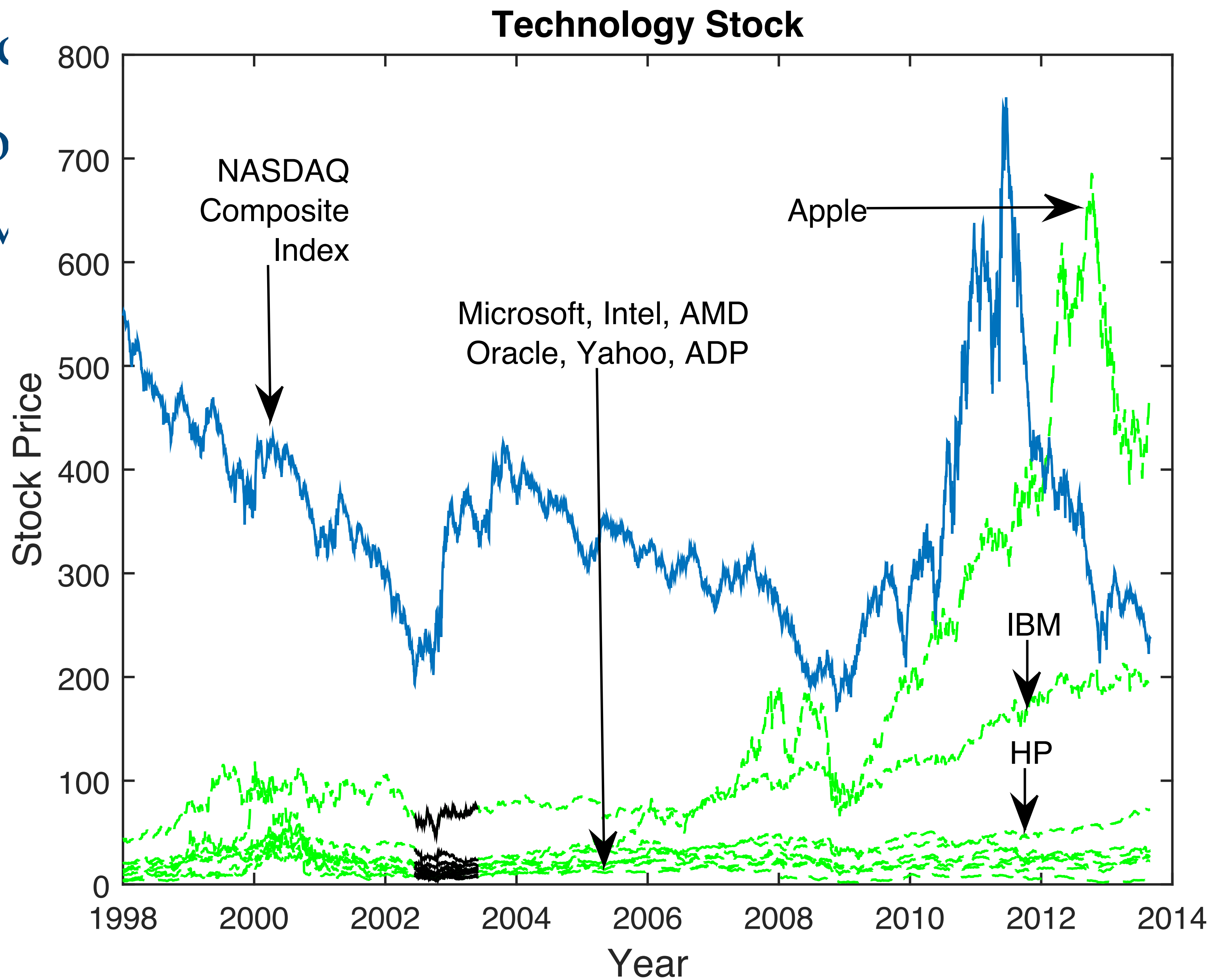
- Daily trading prices 1998-2013

Example: S & P 500

- Daily trading prices 1998-2013
- 9 tech stocks
  - ADP, AMD, HP, IBM, Intel, Microsoft, Oracle, Yahoo

Example: S & P 500

- Daily tra(
- 9 tech sto
  - ADP, AM



Example: S & P 500

- Daily trading prices 1998-2013
- 9 tech stocks
  - ADP, AMD, HP, IBM, Intel, Microsoft, Oracle, Yahoo

Example: S & P 500

- Daily trading prices 1998-2013
- 9 tech stocks
  - ADP, AMD, HP, IBM, Intel, Microsoft, Oracle, Yahoo
- Atypical segment in 2003
  - Not clear from stocks themselves
  - Low point of Nasdaq after bubble
    - Perhaps stocks move more in sync?

Conclusion

- We have developed an information theory criterion of atypicality
  - Fundamental
- Works for
  - Discrete valued data
  - Real valued data
- Upper bounded probability of intrinsically atypical data
  - Same for real and discrete case
- Experimental results for stock market data