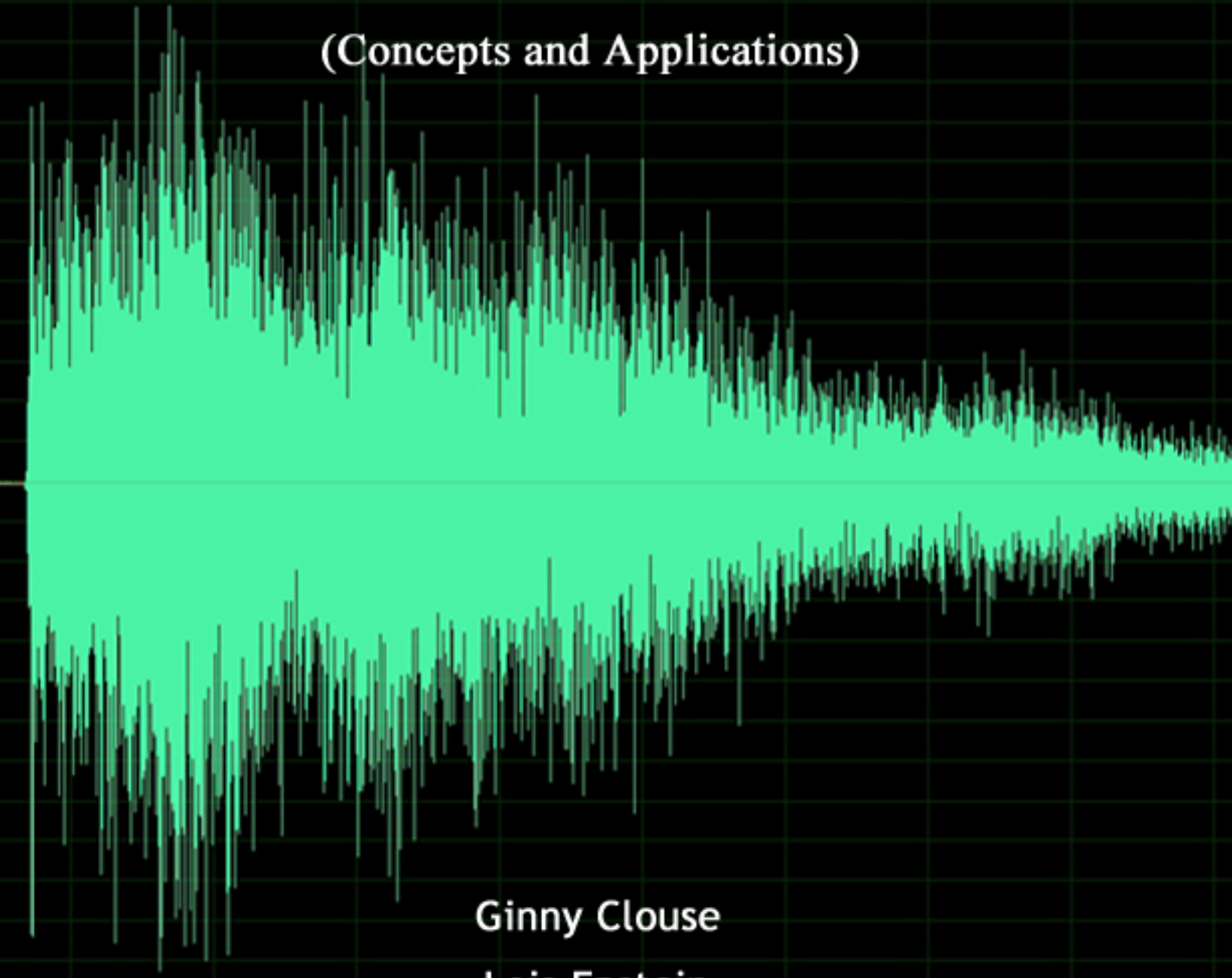
Telecommunication and Estimation Theories

(Concepts and Applications)



Ginny Clouse

Lois Epstein

First Edition, 2012

ISBN 978-81-323-0993-2

Published by:
Academic Studio
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter-1

Spectral Efficiency

Spectral efficiency, **spectrum efficiency** or **bandwidth efficiency** refers to the information rate that can be transmitted over a given bandwidth in a specific communication system. It is a measure of how efficiently a limited frequency spectrum is utilized by the physical layer protocol, and sometimes by the media access control (the channel access protocol).

Link spectral efficiency

The **link spectral efficiency** of a digital communication system is measured in *bit/s/Hz*, or, less frequently but unambiguously, in *(bit/s)/Hz*. It is the net bitrate (useful information rate excluding error-correcting codes) or maximum throughput divided by the bandwidth in hertz of a communication channel or a data link. Alternatively, the spectral efficiency may be measured in in *bit/symbol*, which is equivalent to *bits per channel use* (*bpcu*), implying that the net bit rate is divided by the symbol rate (modulation rate) or line code pulse rate.

Link spectral efficiency is typically used to analyse the efficiency of a digital modulation method or line code, sometimes in combination with a forward error correction (FEC) code and other physical layer overhead. In the latter case, a "bit" refers to a user data bit; FEC overhead is always excluded.

The **modulation efficiency** in bit/s is the gross bitrate (including any error-correcting code) divided by the bandwidth.

Example 1: A transmission technique using one kilohertz of bandwidth to transmit 1,000 bits per second has a modulation efficiency of 1 (bit/s)/Hz.
Example 2: A V.92 modem for the telephone network can transfer 56,000 bit/s downstream and 48,000 bit/s upstream over an analog telephone network. Due to filtering in the telephone exchange, the frequency range is limited to between 300 hertz and 3,400 hertz, corresponding to a bandwidth of 3,400 − 300 = 3,100 hertz. The spectral efficiency or modulation efficiency is 56,000/3,100 = 18.1 (bit/s)/Hz downstream, and 48,000/3,100 = 15.5 (bit/s)/Hz upstream.

An upper bound for the attainable modulation efficiency is given by the Nyquist rate or Hartley's law as follows: For a signaling alphabet with M alternative symbols, each symbol represents $N = \log_2 M$ bits. N is the modulation efficiency measured in *bit/symbol* or *bpcu*. In the case of baseband transmission (line coding or pulse-amplitude modulation) with a baseband bandwidth (or upper cut-off frequency) B, the symbol rate can not exceed $2B$ symbols/s in view to avoid intersymbol interference. Thus, the spectral efficiency can not exceed $2N$ (bit/s)/Hz in the baseband transmission case. In the passband transmission case, a signal with passband bandwidth W can be converted to an equivalent baseband signal (using undersampling or a superheterodyne receiver), with upper cut-off frequency $W/2$. If double-sideband modulation schemes such as QAM, ASK, PSK or OFDM are used, this results in a maximum symbol rate of W symbols/s, and in that the modulation efficiency can not exceed N (bit/s)/Hz. If digital single-sideband modulation is used, the passband signal with bandwidth W corresponds to a baseband message signal with baseband bandwidth W, resulting in a maximum symbol rate of $2W$ and an attainable modulation efficiency of $2N$ (bit/s)/Hz.

Example 3: An 16QAM modem has an alphabet size of $M = 16$ alternative symbols, with $N = 4$ bit/symbol or bpcu. Since QAM is a form of double sideband passband transmission, the spectral efficiency cannot exceed $N = 4$ (bit/s)/Hz.
Example 4: The 8VSB (8-level vestigial sideband) modulation scheme used in the ATSC digital television standard gives N=3 bit/symbol or bpcu. Since it can be described as nearly single-side band, the modulation efficiency is close to $2N = 6$ (bit/s)/Hz. In practice, ATSC transfers a gross bit rate of 32 Mbit/s over a 6 MHz wide channel, resulting in a modulation efficiency of 32/6 = 5.3 (bit/s)/Hz.
Example 5: The downlink of a V.92 modem uses a pulse-amplitude modulation with 128 signal levels, resulting in $N = 7$ bit/symbol. Since the transmitted signal before passband filtering can be considered as baseband transmission, the spectral efficiency cannot exceed $2N = 14$ (bit/s)/Hz over the full baseband channel (0 to 4 kHz). As seen above, a higher spectral efficiency is achieved if we consider the smaller passband bandwidth.

If a forward error correction code is used, the spectral efficiency is reduced from the uncoded modulation efficiency figure.

Example 6: If a forward error correction (FEC) code with code rate 1/2 is added, meaning that the encoder input bit rate is one half the encoder output rate, the spectral efficiency is 50% of the modulation efficiency. In exchange for this reduction in spectral efficiency, FEC usually reduces the bit-error rate, and typically enables operation at a lower signal to noise ratio (SNR).

An upper bound for the spectral efficiency possible without bit errors in a channel with a certain SNR, if ideal error coding and modulation is assumed, is given by the Shannon-Hartley theorem.

Example 7: If the SNR is 1 times expressed as a ratio, corresponding to 0 decibel, the link spectral efficiency can not exceed 1 (bit/s)/Hz for error-free detection (assuming an

ideal error-correcting code) according to Shannon-Hartley regardless of the modulation and coding.

Note that the goodput (the amount of application layer useful information) is normally lower than the maximum throughput used in the above calculations, because of packet retransmissions, higher protocol layer overhead, flow control, congestion avoidance, etc. On the other hand, a data compression scheme, such as the V.44 or V.42bis compression used in telephone modems, may however give higher goodput if the transferred data is not already efficiently compressed.

The link spectral efficiency of a wireless telephony link may also be expressed as the maximum number of simultaneous calls over 1 MHz frequency spectrum in erlangs per megahertz, or *E/MHz*. This measure is also affected by the source coding (data compression) scheme. It may be applied to analog as well as digital transmission.

In wireless networks, the *link spectral efficiency* can be somewhat misleading, as larger values are not necessarily more efficient in their overall use of radio spectrum. In a wireless network, high link spectral efficiency may result in high sensitivity to co-channel interference (crosstalk), which affects the capacity. For example, in a cellular telephone network with frequency reuse, spectrum spreading and forward error correction reduce the spectral efficiency in (bit/s)/Hz but substantially lower the required signal-to-noise ratio in comparison to non-spread spectrum techniques. This can allow for much denser geographical frequency reuse that compensates for the lower link spectral efficiency, resulting in approximately the same capacity (the same number of simultaneous phone calls) over the same bandwidth, using the same number of base station transmitters. As discussed below, a more relevant measure for wireless networks would be *system spectral efficiency* in bit/s/Hz per unit area. However, in closed communication links such as telephone lines and cable TV networks, and in noise-limited wireless communication system where co-channel interference is not a factor, the largest link spectral efficiency that can be supported by the available SNR is generally used.

System spectral efficiency or area spectral efficiency

In digital wireless networks, the *system spectral efficiency* or area spectral efficiency is typically measured in (bit/s)/Hz per unit area, (bit/s)/Hz per cell, or (bit/s)/Hz per site. It is a measure of the quantity of users or services that can be simultaneously supported by a limited radio frequency bandwidth in a defined geographic area. It may for example be defined as the maximum throughput or goodput, summed over all users in the system, divided by the channel bandwidth. This measure is affected not only by the single user transmission technique, but also by multiple access schemes and radio resource management techniques utilized. It can be substantially improved by dynamic radio resource management. If it is defined as a measure of the maximum goodput, retransmissions due to co-channel interference and collisions are excluded. Higher-layer protocol overhead (above the media access control sublayer) is normally neglected.

Example 8: In a cellular system based on frequency-division multiple access (FDMA) with a fixed channel allocation (FCA) cellplan using a frequency reuse factor of 4, each base station has access to 1/4 of the total available frequency spectrum. Thus, the maximum possible system spectral efficiency in *(bit/s)/Hz per site* is 1/4 of the link spectral efficiency. Each base station may be divided into 3 cells by means of 3 sector antennas, also known as a 4/12 reuse pattern. Then each cell has access to 1/12 of the available spectrum, and the system spectral efficiency in *(bit/s)/Hz per cell* or *(bit/s)/Hz per sector* is 1/12 of the link spectral efficiency.

The system spectral efficiency of a cellular network may also be expressed as the maximum number of simultaneous phone calls per area unit over 1 MHz frequency spectrum in E/MHz per cell, E/MHz per sector, E/MHz per site, or (E/MHz)/m^2. This measure is also affected by the source coding (data compression) scheme. It may be used in analog cellular networks as well.

Low link spectral efficiency in (bit/s)/Hz does not necessarily mean that an encoding scheme is inefficient from a system spectral efficiency point of view. As an example, consider Code Division Multiplexed Access (CDMA) spread spectrum, which is not a particularly spectral efficient encoding scheme when considering a single channel or single user. However, the fact that one can "layer" multiple channels on the same frequency band means that the system spectrum utilization for a multi-channel CDMA system can be very good.

Example 9: In the W-CDMA 3G cellular system, every phone call is compressed to a maximum of 8,500 bit/s (the useful bitrate), and spread out over a 5 MHz wide frequency channel. This corresponds to a link throughput of only 8,500/5,000,000 = 0.0017 *(bit/s)/Hz*. Let us assume that 100 simultaneous (non-silent) simultaneous calls are possible in the same cell. Spread spectrum makes it possible to have as low a frequency reuse factor as 1, if each base station is divided into 3 cells by means of 3 directional sector antennas. This corresponds to a system spectrum efficiency of over 1 × 100 × 0.0017 = 0.17 *(bit/s)/Hz per site*, and 0.17/3 = 0.06 *(bit/s)/Hz per cell or sector*.

The spectral efficiency can be improved by radio resource management techniques such as efficient fixed or dynamic channel allocation, power control, link adaptation and diversity schemes.

A combined fairness measure and system spectral efficiency measure is the fairly shared spectral efficiency.

Comparison table

Examples of numerical spectral efficiency values of some common communication systems can be found in the table below.

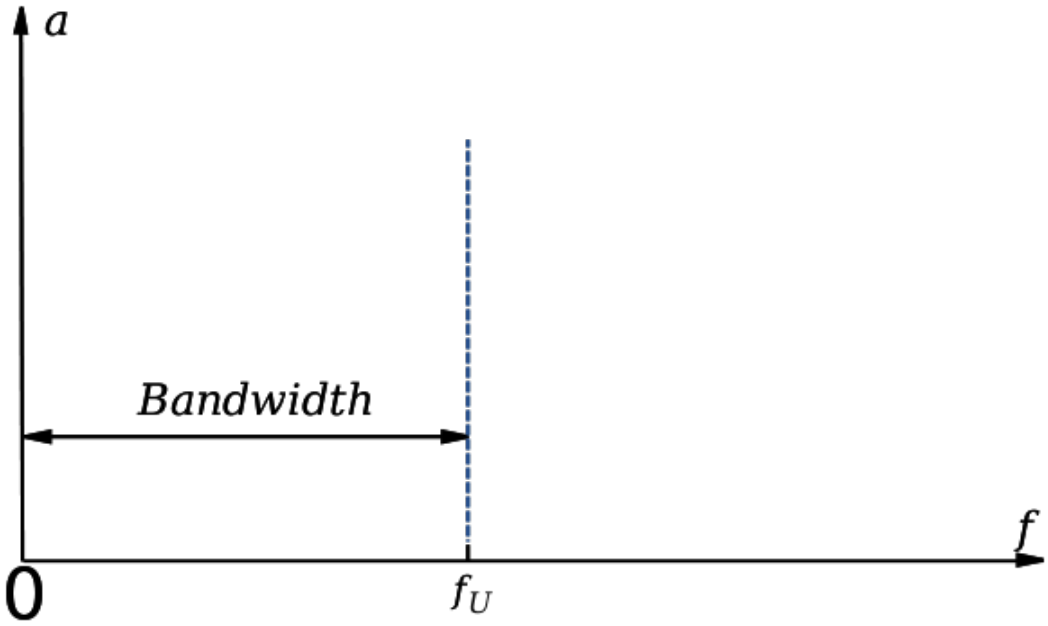
Spectral efficiency of common communication systems.

Service	Standard	Launched year	Net bitrate *R* per carrier (Mbit/s)	Bandwidth *B* per carrier (MHz)	Link spectral efficiency *R/B* ((bit/s)/Hz)	Typical reuse factor *1/K*	System spectral efficiency Approx. (*(R/B)/K*) ((bit/s)/Hz per site)
1G cellular	**NMT 450 modem**	1981	0.0012	0.025	0.45	1/7	0.064
1G cellular	**AMPS modem**	1983	0.0003	0.030	0.001	1/7	0.0015
2G cellular	**GSM**	1991	0.013 × 8 timeslots = 0.104	0.2	0.52	1/9 (1/3 in 1999)	0.17 (in 1999)
2G cellular	**D-AMPS**	1991	0.013 × 3 timeslots = 0.039	0.030	1.3	1/9 (1/3 in 1999)	0.45 (in 1999)
2.75G cellular	**CDMA2000 1× voice**	2000	Max. 0.0096 per phone call × typ 22 calls per carrier	1.2288	0.0078 per mobile × typ 22 calls per carrier	1	0.172 (fully loaded)
2.75G cellular	**GSM + EDGE**	2003	Max.: 0.384; Typ.: 0.20;	0.2	Max.: 1.92; Typ.: 1.00;	1/3	0.33
2.75G cellular	**IS-136HS + EDGE**		Max.: 0.384; Typ.: 0.27;	0.2	Max.: 1.92; Typ.: 1.35;	1/3	0.45
3G cellular	**WCDMA FDD**	2001	Max.: 0.384 per mobile;	5	Max.: 0.077 per mobile;	1	Max 0.51
3G cellular	**CDMA2000 1x PD**	2002	Max.: 0.153 per mobile;	1.2288	Max.: 0.125 per mobile;	1	Max 0.1720 (fully loaded)
3G cellular	**CDMA2000 1×EV-DO Rev.A**	2002	Max.: 3.072 per mobile;	1.2288	Max.: 2.5 per mobile;	1	Max 1.3 average loaded

							sector
Fixed WiMAX	**IEEE 802.16d**	2004	96	20 (1.75, 3.5, 7, ...)	4.8	¼	1.2
3.5G cellular	**HSDPA**	2007	Max.: 42.2 per mobile;	5	Max.: 8.44 per mobile;	1	Max 8.44
3.9G MBWA	**iBurst HC-SDMA**	2005	Max.: 3.9 per carrier;	0.625	Max.: 7.23 per carrier;	1	Max 7.23
3.9G cellular	**LTE**	2009	Max.: 326.4 per mobile;	20	Max.: 16.32 per mobile;	1	Max.: 16.32;
Wi-Fi	**IEEE 802.11a/g**	2003	Max.: 54;	20	Max.: 2.7;	⅓	0.9
Wi-Fi	**IEEE 802.11n Draft 2.0**	2007	Max.: 144.4;	20	Max.: 7.22;	⅓	Max 2.4
TETRA	**ETSI**	1998	4 timeslots = 0.036	0.025	1.44		
Digital radio	**DAB**	1995	0.576 to 1.152	1.712	0.34 to 0.67	⅕	0.08 to 0.17
Digital radio	**DAB with SFN**	1995	0.576 to 1.152	1.712	0.34 to 0.67	1	0.34 to 0.67
Digital TV	**DVB-T**	1997	Max.: 31.67; Typ.: 22.0;	8	Max.: 4.0; Typ.: 2.8;	⅕	0.55
Digital TV	**DVB-T with SFN**	1996	Max.: 31.67; Typ.: 22.0;	8	Max.: 4.0; Typ.: 2.8;	1	Max.: 4.0; Typ.: 2.8;
Digital TV	**DVB-H**	2007	5.5 to 11	8	0.68 to 1.4	⅕	0.14 to 0.28
Digital TV	**DVB-H with SFN**	2007	5.5 to 11	8	0.68 to 1.4	1	0.68 to 1.4
Digital cable TV	**DVB-C 256-QAM mode**		38	6	6.33	N/A	N/A
Broadband modem	**ADSL2 downlink**		12	0.962	12.47	N/A	N/A
Telephone modem	**V.92 downlink**	1999	0.056	0.004	14.0	N/A	N/A

Chapter-2

Bandwidth (signal processing)



Baseband bandwidth. Here the bandwidth equals the upper frequency.

Bandwidth is the difference between the upper and lower frequencies in a contiguous set of frequencies. It is typically measured in hertz, and may sometimes refer to *passband bandwidth*, sometimes to *baseband bandwidth*, depending on context. **Passband bandwidth** is the difference between the upper and lower cutoff frequencies of, for example, an electronic filter, a communication channel, or a signal spectrum. In case of a low-pass filter or baseband signal, the bandwidth is equal to its upper cutoff frequency. The term **baseband bandwidth** always refers to the upper cutoff frequency, regardless of whether the filter is bandpass or low-pass.

Bandwidth in hertz is a central concept in many fields, including electronics, information theory, radio communications, signal processing, and spectroscopy. A key characteristic of bandwidth is that a band of a given width can carry the same amount of information, regardless of where that band is located in the frequency spectrum (assuming equivalent noise level). For example, a 5 kHz band can carry a telephone conversation whether that band is at baseband (as in your POTS telephone line) or modulated to some higher (passband) frequency.

In computer networking and other digital fields, the term bandwidth often refers to a data rate measured in bits per second, for example network throughput, sometimes denoted *network bandwidth*, *data bandwidth* or *digital bandwidth*. The reason is that according to Hartley's law, the digital data rate limit (or channel capacity) of a physical communication link is proportional to its bandwidth in hertz, sometimes denoted **radio frequency (RF) bandwidth**, **signal bandwidth**, **frequency bandwidth**, **spectral bandwidth** or **analog bandwidth**. For *bandwidth* as a computing term, less ambiguous terms are bit rate, throughput, maximum throughput, goodput or channel capacity.

Overview

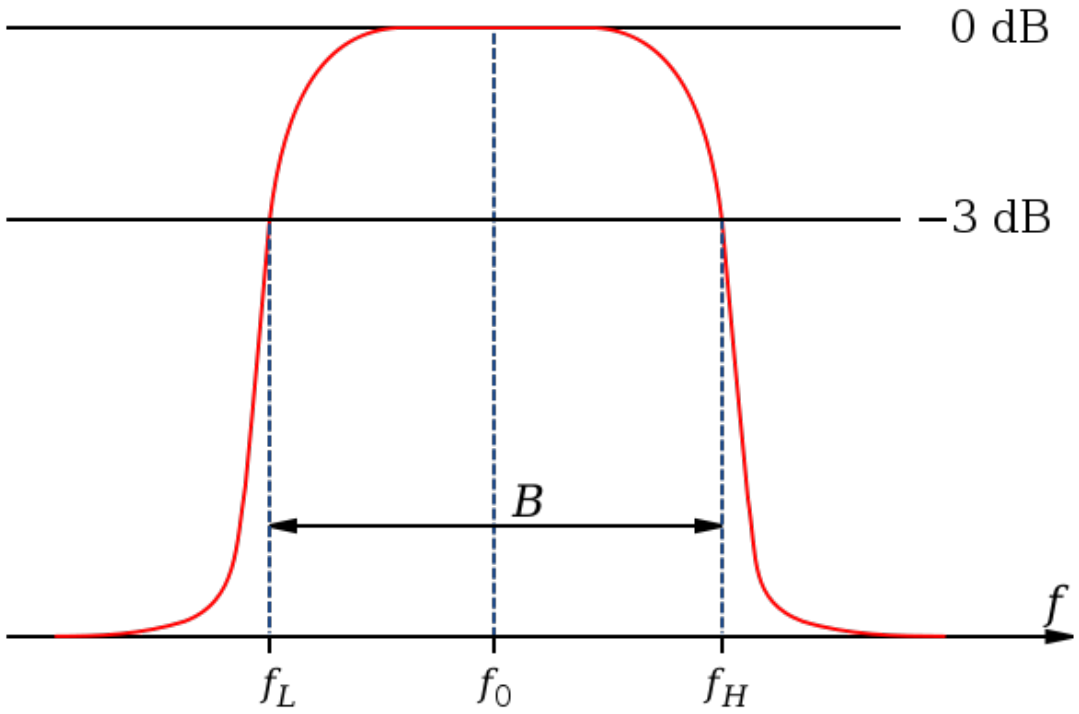
Bandwidth is a key concept in many telephony applications. In radio communications, for example, bandwidth is the frequency range occupied by a modulated carrier wave, whereas in optics it is the width of an individual spectral line or the entire spectral range.

In many signal processing contexts, bandwidth is a valuable and limited resource. For example, an FM radio receiver's tuner spans a limited range of frequencies. A government agency (such as the Federal Communications Commission in the United States) may apportion the regionally available bandwidth to licensed broadcasters so that their signals do not mutually interfere. Each transmitter owns a slice of bandwidth, a valuable (if intangible) commodity.

For different applications there are different precise definitions. For example, one definition of bandwidth could be the range of frequencies beyond which the frequency function is zero. This would correspond to the mathematical notion of the support of a function (i.e., the total "length" of values for which the function is nonzero). A less strict and more practically useful definition will refer to the frequencies where the frequency function is *small*. Small could mean less than 3 dB below (i.e., less than half of) the maximum value, or more rarely 10 dB below, or it could mean below a certain absolute value. As with any definition of the *width* of a function, many definitions are suitable for different purposes.

Bandwidth typically refers to baseband bandwidth in the context of for example sampling theorem and Nyquist sampling rate, while it refers to passband bandwidth in the context of Nyquist symbol rate or Shannon-Hartley channel capacity for communication systems.

X-dB bandwidth



A graph of a bandpass filter's gain magnitude, illustrating the concept of −3 dB bandwidth at a gain of 0.707. The frequency axis of this symbolic diagram can be linear or logarithmically scaled.

In some contexts, the signal bandwidth in hertz refers to the frequency range in which the signal's spectral density is nonzero or above a small threshold value. That definition is used in calculations of the lowest sampling rate that will satisfy the sampling theorem. Because this range of non-zero amplitude may be very broad or infinite, this definition is typically relaxed so that the bandwidth is defined as the range of frequencies in which the signal's spectral density is above a certain threshold relative to its maximum. Most commonly, bandwidth refers to the 3-dB bandwidth, that is, the frequency range within which the spectral density (in W/Hz or V^2/Hz) is above half its maximum value (or the spectral amplitude, in V or V/Hz, is more than 70.7% of its maximum); that is, above −3 dB relative to the peak.

The word bandwidth applies to signals as described above, but it could also apply to *systems*, for example filters or communication channels. To say that a system has a certain bandwidth means that the system can process signals of that bandwidth, or that the system reduces the bandwidth of a white noise input to that bandwidth.

The 3 dB bandwidth of an electronic filter or communication channel is the part of the system's frequency response that lies within 3 dB of the response at its peak, which in the passband filter case is typically at or near its center frequency, and in the lowpass filter is near 0 hertz. If the maximum gain is 0 dB, the 3 dB gain is the range where the gain is

more than -3dB, or the attenuation is less than + 3dB. This is also the range of frequencies where the amplitude gain is above 70.7% of the maximum amplitude gain, and above half the maximum power gain. This same "half power gain" convention is also used in spectral width, and more generally for extent of functions as full width at half maximum (FWHM).

In electronic filter design, a filter specification may require that within the filter passband, the gain is nominally 0 dB +/- a small number of dB , for example within the +/- 1 dB interval. In the stopband(s), the required attenuation in dB is above a certain level, for example >100 dB. In a transition band the gain is not specified. In this case, the filter bandwidth corresponds to the passband width, which in this example is the 1dB-bandwidth. If the filter shows amplitude ripple within the passband, the x dB point refers to the point where the gain is x dB below the nominal passband gain rather than x dB below the maximum gain.

A commonly used quantity is *fractional bandwidth*. This is the bandwidth of a device divided by its center frequency. E.g., a passband filter that has a bandwidth of 2 MHz with center frequency 10 MHz will have a fractional bandwidth of 2/10, or 20%.

In communication systems, in calculations of the Shannon–Hartley channel capacity, bandwidth refers to the 3dB-bandwidth. In calculations of the maximum symbol rate, the Nyquist sampling rate, and maximum bit rate according to the Hartley formula, the bandwidth refers to the frequency range within which the gain is non-zero, or the gain in dB is below a very large value.

The fact that in equivalent baseband models of communication systems, the signal spectrum consists of both negative and positive frequencies, can lead to confusion about bandwidth, since they are sometimes referred to only by the positive half, and one will occasionally see expressions such as $B = 2W$, where B is the total bandwidth (i.e. the maximum passband bandwidth of the carrier-modulated RF signal and the minimum passband bandwidth of the physical passband channel), and W is the positive bandwidth (the baseband bandwidth of the equivalent channel model). For instance, the baseband model of the signal would require a lowpass filter with cutoff frequency of at least W to stay intact, and the physical passband channel would require a passband filter of at least B to stay intact.

In signal processing and control theory the bandwidth is the frequency at which the closed-loop system gain drops 3 dB below peak.

In basic electric circuit theory, when studying band-pass and band-reject filters, the bandwidth represents the distance between the two points in the frequency domain where the signal is $\frac{1}{\sqrt{2}}$ of the maximum signal amplitude (half power).

Antenna systems

In the field of antennas, two different methods of expressing relative bandwidth are used for narrowband and wideband antennas. For either, a set of criteria is established to define the extents of the bandwidth, such as input impedance, pattern, or polarization.

Percent bandwidth, usually used for narrowband antennas, is used defined as $\%B = \frac{f_H - f_L}{f_c} = 2\frac{f_H - f_L}{f_H + f_L}$. The theoretical limit to fractional bandwidth is 200%, which occurs for $f_L = 0$.

Fractional bandwidth, usually used for wideband antennas, is defined as $B = f_H / f_L$, and is typically presented in the form of B:1. Fractional bandwidth is used for wideband antennas because of the compression of the percent bandwidth that occurs mathematically with percent bandwidths above 100%, which corresponds to a fractional bandwidth of 3:1.

Photonics

In photonics, the term *bandwidth* occurs in a variety of meanings:

- the bandwidth of the output of some light source, e.g., an ASE source or a laser; the bandwidth of ultrashort optical pulses can be particularly large
- the width of the frequency range that can be transmitted by some element, e.g. an optical fiber
- the gain bandwidth of an optical amplifier
- the width of the range of some other phenomenon (e.g., a reflection, the phase matching of a nonlinear process, or some resonance)
- the maximum modulation frequency (or range of modulation frequencies) of an optical modulator
- the range of frequencies in which some measurement apparatus (e.g., a powermeter) can operate
- the data rate (e.g., in Gbit/s) achieved in an optical communication system.

A related concept is the spectral linewidth of the radiation emitted by excited atoms.

Chapter-3

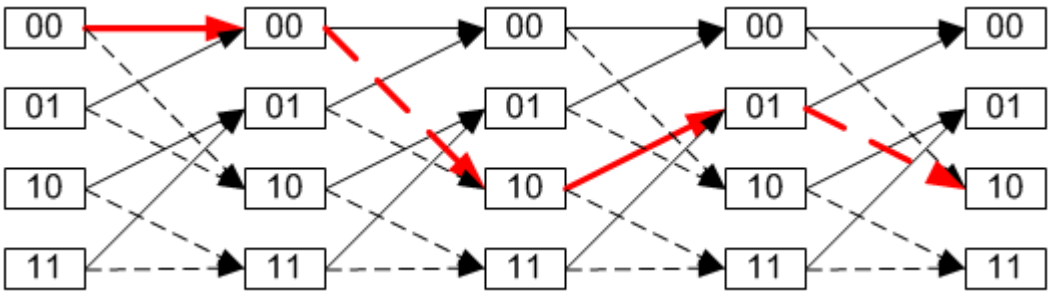
Trellis Modulation and Turbo Equalizer

Trellis modulation

In telecommunication, **trellis modulation** (also known as **trellis coded modulation**, or simply **TCM**) is a modulation scheme which allows highly efficient transmission of information over band-limited channels such as telephone lines. Trellis modulation was invented by Gottfried Ungerboeck working for IBM in the 1970s, and first described in a conference paper in 1976; but it went largely unnoticed until he published a new detailed exposition in 1982 which achieved sudden widespread recognition.

In the late 1980s, modems operating over plain old telephone service (*POTS*) typically achieved 9.6 kbit/s by employing 4 bits per symbol QAM modulation at 2,400 baud (symbols/second). This bit rate ceiling existed despite the best efforts of many researchers, and some engineers predicted that without a major upgrade of the public phone infrastructure, the maximum achievable rate for a POTS modem might be 14 kbit/s for two-way communication (3,429 baud × 4 bits/symbol, using QAM). However, 14 kbit/s is only 40% of the theoretical maximum bit rate predicted by Shannon's Theorem for POTS lines (approximately 35 kbit/s).

A new modulation method



Trellis diagram.

The name *trellis* was coined because a state diagram of the technique, when drawn on paper closely resembles the trellis lattice used in rose gardens. The scheme is basically a convolutional code of rates (r,r+1). Ungerboeck's unique contribution is to apply the parity check on a per symbol basis instead of the older technique of applying it to the bit stream then modulating the bits. The key idea he termed Mapping by Set Partitions. This idea was to group the symbols in a tree like fashion then separate them into two limbs of equal size. At each limb of the tree, the symbols were further apart. Although in multi-dimensions, it is hard to visualize, a simple one dimension example illustrates the basic procedure. Suppose the symbols are located at [1, 2, 3, 4, ...]. Then take all odd symbols and place them in one group, and the even symbols in the second group. This is not quite accurate because Ungerboeck was looking at the two dimensional problem, but the principle is the same, take every other one for each group and repeat the procedure for each tree limb. He next described a method of assigning the encoded bit stream onto the symbols in a very systematic procedure. Once this procedure was fully described, his next step was to program the algorithms into a computer and let the computer search for the best codes. The results were astonishing. Even the most simple code (4 state) produced error rates nearly 1,000 times lower than an equivalent uncoded system. For two years Ungerboeck kept these results private and only conveyed them to close colleagues. Finally, in 1982, Ungerboeck published a paper describing the principles of trellis modulation.

A flurry of research activity ensued, and by 1990 the International Telecommunication Union had published modem standards for the first trellis-modulated modem at 14.4 kbit/s (2,400 baud and 6 bits per symbol). Over the next several years further advances in encoding, plus a corresponding symbol rate increase from 2,400 to 3,429 baud, allowed modems to achieve rates up to 34.3 kbit/s (limited by maximum power regulations to 33.8 kbit/s). Today, the most common trellis-modulated V.34 modems use a 4-dimensional set partition which is achieved by treating two 2-dimensional symbols as a single lattice. This set uses 8, 16, or 32 state convolutional codes to squeeze the equivalent of 6 to 10 bits into each symbol sent by the modem (for example, 2,400 baud × 8 bits/symbol = 19,200 bit/s).

Once manufacturers introduced modems with trellis modulation, transmission rates increased to the point where interactive transfer of multimedia over the telephone became feasible (a 200 kilobyte image and a 5 megabyte song could be downloaded in less than 1 minute and 30 minutes, respectively). Sharing a floppy disk via a BBS could be done in just a few minutes, instead of an hour. Thus Ungerboeck's invention played a key role in the Information Age.

Turbo equalizer

In digital communications, a **turbo equalizer** is a type of receiver used to receive a message corrupted by a communication channel with intersymbol interference (ISI). It approaches the performance of a maximum a posteriori (MAP) receiver via iterative message passing between a soft-in soft-out (SISO) equalizer and a SISO decoder. It is

closely related to turbo codes, as a turbo equalizer may be considered a turbo decoder if the channel is viewed as a convolutional code.

History

In 1993, turbo codes were introduced by Berrou, Glavieux, and Thitimajshima. In 1995, the turbo principle, which was developed for turbo codes, was applied to an equalizer by Douillard, Jézéquel, and Berrou. They formulated the ISI receiver problem as a turbo code decoding problem, where the channel is thought of as a rate 1 convolutional code and the error correction coding is the second code. In 1997, Glavieux, Laot, and Labat demonstrated that a linear equalizer could be used in a turbo equalizer framework. This discovery made turbo equalization computationally efficient enough to be applied to a wide range of applications.

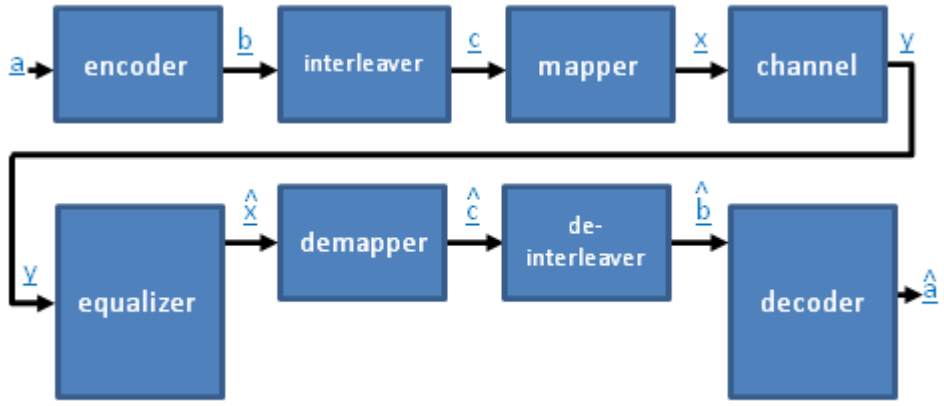
Overview

Standard Communication System Overview

Before discussing turbo equalizers, it is necessary to understand the basic receiver in the context of a communication system. At the transmitter, information bits a are encoded. Encoding adds redundancy by mapping the information bits *a* to a longer bit vector--the code bit vector *b*. The encoded bits *b* are then interleaved. Interleaving permutes the order of the code bits *b* resulting in bits *c*. The main reason for doing this is to insulate the information bits from bursty noise. Next, the symbol mapper maps the bits *c* into complex symbols *x*. These digital symbols are then converted into analog symbols with an D/A converter. Typically the signal is then up-converted to pass band frequencies by mixing it with a carrier signal. This is a necessary step for complex symbols. The signal is then ready to be transmitted through the channel.

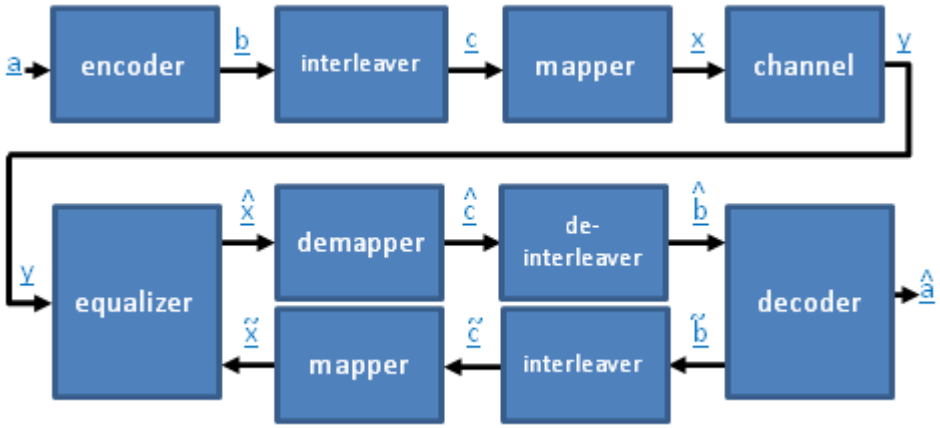
At the receiver, the operations performed by the transmitter are reversed to recover $\hat{a}$, an estimate of the information bits. The down-converter mixes the signal back down to baseband. The A/D converter then samples the analog signal, making it digital. At this point, *y* is recovered. The signal *y* is what would be received if *x* were transmitted through the digital baseband equivalent of the channel plus noise. The signal is then equalized. The equalizer attempts to unravel the ISI in the received signal to recover the transmitted symbols. It then outputs the bits $\hat{c}$associated with those symbols. The vector $\hat{c}$may represent hard decisions on the bits or soft decisions. If the equalizer makes soft decisions, it outputs information relating to the probability of the bit being a 0 or a 1. If the equalizer makes hard decisions on the bits, it quantizes the soft bit decisions and outputs either a 0 or a 1. Next, the signal is deinterleaved which is a simple permutation transformation that undoes the transformation the interleaver executed. Finally, the bits are decoded by the decoder. The decoder estimates $\hat{a}$from $\hat{b}$.

A diagram of the communication system is shown below. In this diagram, the channel is the equivalent baseband channel, meaning that it encompasses the A/D, the up converter, the channel, the down converter, and the D/A.



Turbo Equalizer Overview

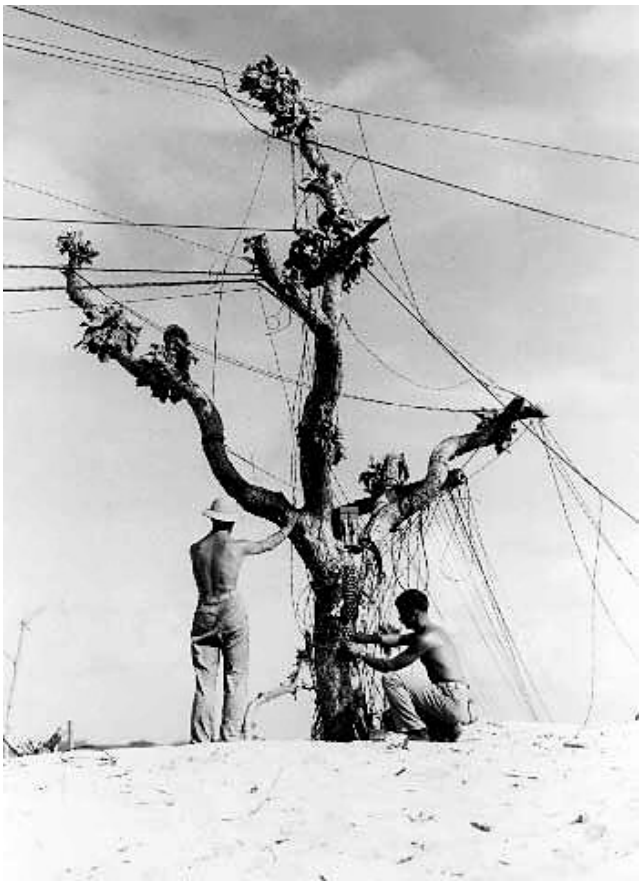
The block diagram of a communication system employing a turbo equalizer is shown below. The turbo equalizer encompasses the equalizer, the decoder, and the blocks in between.



The difference between a turbo equalizer and a standard equalizer is the feedback loop from the decoder to the equalizer. Due to the structure of the code, the decoder not only estimates the information bits *a*, but it also discovers new information about the coded bits *b*. The decoder is therefore able to output extrinsic information, $\tilde{b}$about the likelihood that a certain code bit stream was transmitted. Extrinsic information is new information that is not derived from information input to the block. This extrinsic information is then mapped back into information about the transmitted symbols *x* for use in the equalizer. These extrinsic symbol likelihoods, $\tilde{x}$, are fed into the equalizer as *a priori* symbol probabilities. The equalizer uses this *a priori* information as well as the input signal *y* to estimate extrinsic probability information about the transmitted symbols. The *a priori* information fed to the equalizer is initialized to 0, meaning that the initial estimate $\hat{a}$made by the turbo equalizer is identical to the estimate made by the standard receiver. The information $\hat{x}$is then mapped back into information about *b* for use by the decoder. The turbo equalizer repeats this iterative process until a stopping criterion is reached.

Chapter-4

Channel (communications)



Old telephone wires are a challenging communications channel for modern digital communications.

In telecommunications and computer networking, a **communication channel**, or **channel**, refers either to a physical transmission medium such as a wire, or to a logical connection over a multiplexed medium such as a radio channel. A channel is used to convey an information signal, for example a digital bit stream, from one or several *senders* (or transmitters) to one or several *receivers*. A channel has a certain capacity for

transmitting information, often measured by its bandwidth in Hz or its data rate in bits per second

In information theory, a channel refers to a theoretical *channel model* with certain error characteristics. In this more general view, a storage device is also a kind of channel, which can be sent to (written) and received from (read).

Examples

A channel can take many forms. Examples of communications channels include:

1. A connection between initiating and terminating nodes of a circuit.
2. A single path provided by a transmission medium via either
    - physical separation, such as by multipair cable or
    - electrical separation, such as by frequency-division or time-division multiplexing.
3. A path for conveying electrical or electromagnetic signals, usually distinguished from other parallel paths.
    - A storage which can communicate a message over time as well as space
    - The portion of a storage medium, such as a track or a band, that is accessible to a given reading or writing station or head.
    - A buffer from which messages can be 'put' and 'got'.
4. In a communications system, the physical or logical link that connects a data source to a data sink.
5. A specific radio frequency, pair or band of frequencies, usually named with a letter, number, or codeword, and often allocated by international agreement. Examples:
    - Marine VHF radio uses some 88 channels in the VHF band for two-way FM voice communication. Channel 16, for example, is 156.800 MHz. In the US, seven additional channels, WX1 - WX7, are allocated for weather broadcasts.
    - Television channels such as North American TV Channel 2 = 55.25 MHz, Channel 13 = 211.25 MHz. Each channel is 6 MHz wide. Besides these "physical channels", television also has "virtual channels".
    - Wi-Fi consists of unlicensed channels 1-13 from 2412 MHz to 2484 MHz in 5 MHz steps.
    - The radio channel between an amateur radio repeater and a ham uses two bands often 600 kHz (0.6 MHz) apart. For example, a repeater that transmits on 146.94 MHz typically listens for a ham transmitting on 146.34 MHz.
6. A room in the Internet Relay Chat (IRC) network, in which participants can communicate with each other.

All of these communications channels share the property that they transfer information. The information is carried through the channel by a signal.

Channel models

A channel can be modelled physically by trying to calculate the physical processes which modify the transmitted signal. For example in wireless communications the channel can be modelled by calculating the reflection off every object in the environment. A sequence of random numbers might also be added in to simulate external interference and/or electronic noise in the receiver.

Statistically a communication channel is usually modelled as a triple consisting of an input alphabet, an output alphabet, and for each pair *(i, o)* of input and output elements a transition probability *p(i, o)*. Semantically, the transition probability is the probability that the symbol *o* is received given that *i* was transmitted over the channel.

Statistical and physical modelling can be combined. For example in wireless communications the channel is often modelled by a random attenuation (known as fading) of the transmitted signal, followed by additive noise. The attenuation term is a simplification of the underlying physical processes and captures the change in signal power over the course of the transmission. The noise in the model captures external interference and/or electronic noise in the receiver. If the attenuation term is complex it also describes the relative time a signal takes to get through the channel. The statistics of the random attenuation are decided by previous measurements or physical simulations.

Channel models may be continuous channel models in that there is no limit to how precisely their values may be defined.

Communication channels are also studied in a discrete-alphabet setting. This corresponds to abstracting a real world communication system in which the analog->digital and digital->analog blocks are out of the control of the designer. The mathematical model consists of a transition probability that specifies an output distribution for each possible sequence of channel inputs. In information theory, it is common to start with memoryless channels in which the output probability distribution only depends on the current channel input.

A channel model may either be digital (quantified, e.g. binary) or analog.

Digital channel models

In a digital channel model, the transmitted message is modelled as a digital signal at a certain protocol layer. Underlying protocol layers, such as the physical layer transmission technique, is replaced by a simplified model. The model may reflect channel performance measures such as bit rate, bit errors, latency/delay, delay jitter, etc. Examples of digital channel models are:

- Binary symmetric channel (BSC), a discrete memoryless channel with a certain bit error probability
- Binary bursty bit error channel model, a channel "with memory"

- Binary erasure channel (BEC), a discrete channel with a certain bit error detection (erasure) probability
- Packet erasure channel, where packets are lost with a certain packet loss probability or packet error rate
- Arbitrarily varying channel (AVC), where the behavior and state of the channel can change randomly

Analog channel models

In an analog channel model, the transmitted message is modelled as an analog signal. The model can be a linear or non-linear, time-continuous or time-discrete (sampled), memoryless or dynamic (resulting in burst errors), time-invariant or time-variant (also resulting in burst errors), baseband, passband (RF signal model), real-valued or complex-valued signal model. The model may reflect the following channel impairments:

- Noise model, for example
  - Additive white Gaussian noise (AWGN) channel, a linear continuous memoryless model
  - Phase noise model
- Interference model, for example cross-talk (co-channel interference) and intersymbol interference (ISI)
- Distortion model, for example a non-linear channel model causing intermodulation distortion (IMD)
- Frequency response model, including attenuation and phase-shift
- Group delay model
- Modelling of underlying physical layer transmission techniques, for example a complex-valued equivalent baseband model of modulation and frequency response
- Radio frequency propagation model, for example
  - Log-distance path loss model
  - Fading model, for example Rayleigh fading, Ricean fading, log-normal shadow fading and frequency selective (dispersive) fading
  - Doppler shift model, which combined with fading results in a time-variant system
  - Ray tracing models, which attempt to model the signal propagation and distortions for specified transmitter-receiver geometries, terrain types, and antennas
  - Mobility models, which also causes a time-variant system

Types of communications channels

- Digital (discrete) or analog (continuous) channel
- Baseband and passband channel
- Transmission medium, for example a fibre channel
- Multiplexed channel
- Computer network virtual channel

- Simplex communication, duplex communication or half duplex communication channel
- Return channel
- Uplink or downlink (upstream or downstream channel)
- Broadcast channel, unicast channel or multicast channel

Multi-terminal channels, with application to cellular systems

In networks, as opposed to point-to-point communication, the communication media is shared between multiple nodes (terminals). Depending on the type of communication, different terminals can cooperate or interfere on each other. In general, any complex multi-terminal network can be considered as a combination of simplified multi-terminal channels. The following channels are the principal multi-terminal channels which was first introduced in the field of information theory:

- A point-to-multipoint channel, also known as broadcasting medium (not to be confused with broadcasting channel): In this channel, a single sender transmits multiple messages to different destination nodes. All wireless channels except radio links can be considered as broadcasting media, but may not always provide broadcasting service. The downlink of a cellular system can be considered as a point-to-multipoint channel, if only one cell is considered and inter-cell co-channel interference is neglected. However, the communication service of a phone call is unicasting.
- Multiple access channel: In this channel, multiple senders transmit multiple possible different messages over a shared physical medium to one or several destination nodes. This requires a channel access scheme, including a media access control (MAC) protocol combiend with a multiplexing scheme. This channel model has applications in the uplink of the cellular networks.
- Relay channel: In this channel, one or several intermediate nodes (called relay, repeater or gap filler nodes) cooperate with a sender to send the message to an ultimate destination node. Relay nodes are considered as a possible add-on in the upcoming cellular standards like 3GPP Long Term Evolution (LTE).
- Interference channel: In this channel, two different senders transmit their data to different destination nodes. Hence, the different senders can have a possible cross-talk or co-channel interference on the signal of each other. The inter-cell interference in the cellular wireless communications is an example of the interference channel. In spread spectrum systems like 3G, interference also occur inside the cell if non-orthogonal codes are used.
- A unicasting channel is a channel that provides a unicasting service, i.e. that sends data addressed to one specific user. An established phone call is an example.
- A broadcasting channel is a channel that provides a broadcasting service, i.e. that sends data addressed to all users in the network. Cellular network examples are the paging service as well as the Multimedia Broadcast Multicast Service.
- A multicasting channel is a channel where data is addressed to a group of subscribing users. LTE exampels are the Physical Multicast Channel (PMCH) and MBSFN (Multicast Broadcast Single Frequency Network).

From the above 4 basic multi-terminal channels, multiple access channel is the only one whose capacity region is known. Even for the special case of the Gaussian scenario, the capacity region of the other 3 channels except the broadcast channel is unknown in general.

Chapter-5

Detection Theory

Detection theory, or **signal detection theory**, is a means to quantify the ability to discern between signal and noise. According to the theory, there are a number of determiners of how a detecting system will detect a signal, and where its threshold levels will be. The theory can explain how changing the threshold will affect the ability to discern, often exposing how adapted the system is to the task, purpose or goal at which it is aimed.

When the detecting system is a human being, experience, expectations, physiological state (e.g. fatigue) and other factors can affect the threshold applied. For instance, a sentry in wartime will likely detect fainter stimuli than the same sentry in peacetime.

Much of the early work in detection theory was done by radar researchers. The psychological theory was first published by Wilson P. Tanner, David M. Green, and John A. Swets in 1954. Detection theory was used in 1966 by John A. Swets and David M. Green for psychophysics. Green and Swets criticized the traditional methods of psychophysics for their inability to discriminate between the real sensitivity of subjects and their (potential) response biases.

Detection theory has applications in many fields such as diagnostics of any kind, quality control, telecommunications, and psychology. The concept is similar to the signal to noise ratio used in the sciences and confusion matrices used in artificial intelligence. It is also usable in alarm management, where it is important to separate important events from background noise.

Psychology

Signal detection theory (SDT) is used when psychologists want to measure the way we make decisions under conditions of uncertainty, such as how we would perceive distances in foggy conditions. SDT assumes that the decision maker is not a passive receiver of information, but an active decision-maker who makes difficult perceptual judgements under conditions of uncertainty. In foggy circumstances, we are forced to decide how far away from us an object is, based solely upon visual stimulus which is

impaired by the fog. Since the brightness of the object, such as a traffic light, is used by the brain to discriminate the distance of an object, and the fog reduces the brightness of objects, we perceive the object to be much farther away than it actually is.

To apply signal detection theory to a data set where stimuli were either present or absent, and the observer categorized each trial as having the stimulus present or absent, the trials are sorted into one of four categories:

	Respond "Absent"	**Respond "Present"**
Stimulus Present	Miss	Hit
Stimulus Absent	Correct Rejection	False Alarm

Based on the proportions of these types of trials, numerical estimates of sensitivity can be obtained with statistics like the sensitivity index d' and A', and response bias can be estimated with statistics like β.

Signal detection theory can also be applied to memory experiments, where items are presented on a study list for later testing. A test list is created by combining these 'old' items with novel, 'new' items that did not apear on the study list. On each test trial the subject will respond 'yes, this was on the study list' or 'no, this was not on the study list'. Items presented on the study list are called Targets, and new items are called Distractors. Saying 'Yes' to a target constitutes a Hit, while saying 'Yes' to a distractor constitutes a False Alarm.

	Respond "No"	**Respond "Yes"**
Target	Miss	Hit
Distractor	Correct Rejection	False Alarm

Applications

Signal Detection Theory has wide application, both in humans and other animals. Topics include memory, stimulus characterists of schedules of reinforcement, etc.

Sensitivity or discriminability

Conceptually, sensitivity refers to how hard or easy it is to detect that a target stimulus is present from background events. For example, in a recognition memory paradigm, having longer to study to-be-remembered words makes it easier to recognize previously seen or heard words. In contrast, having to remember 30 words rather than 5 makes the discrimination harder. One of the most commonly used statistics for computing sensitivity is the so-called sensitivity index, or *d'*. There are also non-parametric measures.

Bias

Bias is the extent to which one response is more probable than another. That is, a receiver may be more likely to respond that a stimulus is present or more likely to respond that a stimulus is not present. Bias is independent of sensitivity. For example, if there is a penalty for either false alarms or misses, this may influence bias. If the stimulus is a bomber, then a miss (failing to detect the plane) may increase deaths, so a liberal bias is likely. In contrast, crying wolf (a false alarm) too often may make people less likely to respond, grounds for a conservative bias.

Mathematics

P(H1|y) > P(H2|y) / MAP Testing

In the case of making a decision between two hypotheses, *H1*, absent, and *H2*, present, in the event of a particular observation, *y*, a classical approach is to choose *H1* when *p(H1|y) > p(H2|y)* and *H2* in the reverse case. In the event that the two *a posteriori* probabilities are equal, one typically defaults to a single choice, say *H2*. One could also flip a coin although the expected number of errors would be the same.

When taking this approach, usually what one knows are the conditional probabilities, *p(y|H1)* and *p(y|H2)*, and the *a priori* probabilities $p(H1) = \pi_1$ and $p(H2) = \pi_2$. In this case,

$$p(H1|y) = \frac{p(y|H1) \cdot \pi_1}{p(y)},$$

$$p(H2|y) = \frac{p(y|H2) \cdot \pi_2}{p(y)}$$

where *p(y)* is the total probability of event *y*,

$$p(y|H1) \cdot \pi_1 + p(y|H2) \cdot \pi_2.$$

H2 is chosen in case

$$\frac{p(y|H2) \cdot \pi_2}{p(y|H1) \cdot \pi_1 + p(y|H2) \cdot \pi_2} \geq \frac{p(y|H1) \cdot \pi_1}{p(y|H1) \cdot \pi_1 + p(y|H2) \cdot \pi_2}$$

$$\Rightarrow \frac{p(y|H2)}{p(y|H1)} \geq \frac{\pi_1}{\pi_2}$$

and *H1* otherwise.

Often, the ratio $\frac{\pi_2}{\pi_1}$ is called τ_{MAP} and $\frac{p(y|H2)}{p(y|H1)}$ is called $L(y)$, the *likelihood ratio*.

Using this terminology, *H2* is chosen in case $L(y) \geq \tau_{MAP}$. This is called MAP testing, where MAP stands for "maximum *a posteriori*").

Taking this approach minimizes the expected number of errors one will make.

Bayes Criterion

In some cases, it is far more important to respond appropriately to *H1* than it is to respond appropriately to *H2*. For example, if one is trying to detect an incoming bomber known to be carrying a nuclear weapon, it is much more important to shoot down the bomber if it is there, than it is not to send a fighter squadron to inspect a false alarm (assuming a large supply of fighter squadrons). The Bayes criterion is an approach suitable for such cases.

Here a utility is associated with each of four situations:

- U_{11}: One responds with behavior appropriate to H1 and H1 is true: fighters destroy bomber, incurring fuel, maintenance, and weapons costs, take risk of some being shot down;
- U_{12}: One responds with behavior appropriate to H1 and H2 is true: fighters sent out, incurring fuel and maintenance costs, bomber location remains unknown;
- U_{21}: One responds with behavior appropriate to H2 and H1 is true: city destroyed;
- U_{22}: One responds with behavior appropriate to H2 and H2 is true: fighters stay home, bomber location remains unknown;

As is shown below, what is important are the differences, $U_{11} - U_{21}$ and $U_{22} - U_{12}$.

Similarly, there are four probabilities, P_{11}, P_{12}, etc., for each of the cases (which are dependent on one's decision strategy).

The Bayes criterion approach is to maximize the expected utility:

$$U = P_{11} \cdot U_{11} + P_{21} \cdot U_{21} + P_{12} \cdot U_{12} + P_{22} \cdot U_{22}$$

$$U = P_{11} \cdot U_{11} + (1 - P_{11}) \cdot U_{21} + P_{12} \cdot U_{21} + (1 - P_{12}) \cdot U_{22}$$

$$U = U_{12} + U_{21} + P_{11} \cdot (U_{11} - U_{21}) - P_{12} \cdot (U_{22} - U_{12})$$

Effectively, one may maximize the sum,

$$U' = P_{11} \cdot (U_{11} - U_{21}) - P_{12} \cdot (U_{22} - U_{12}),$$

and make the following substitutions:

$$P_{11} = \pi_1 \cdot \int_{R_1} p(y|H1)\, dy$$

$$P_{12} = \pi_2 \cdot \int_{R_1} p(y|H2)\, dy$$

where π_1 and π_2 are the *a priori* probabilities, $P(H1)$ and $P(H2)$, and R_1 is the region of observation events, y, that are responded to as though $H1$ is true.

$$\Rightarrow U' = \int_{R_1} \{\pi_1 \cdot (U_{11} - U_{21}) \cdot p(y|H1) - \pi_2 \cdot (U_{22} - U_{12}) \cdot p(y|H2)\}\, dy$$

U' and thus U are maximized by extending R_1 over the region where

$$\pi_1 \cdot (U_{11} - U_{21}) \cdot p(y|H1) - \pi_2 \cdot (U_{22} - U_{12}) \cdot p(y|H2) > 0$$

This is accomplished by deciding H2 in case

$$\pi_2 \cdot (U_{22} - U_{12}) \cdot p(y|H2) \geq \pi_1 \cdot (U_{11} - U_{21}) \cdot p(y|H1)$$

$$\Rightarrow L(y) \equiv \frac{p(y|H2)}{p(y|H1)} \geq \frac{\pi_1 \cdot (U_{11} - U_{21})}{\pi_2 \cdot (U_{22} - U_{12})} \equiv \tau_B$$

and H1 otherwise, where $L(y)$ is the so-defined *likelihood ratio*.

Chapter-6

Filter (Signal Processing)

In signal processing, a **filter** is a device or process that removes from a signal some unwanted component or feature. Filtering is a class of signal processing, the defining feature of filters being the complete or partial suppression of some aspect of the signal. Most often, this means removing some frequencies and not others in order to suppress interfering signals and reduce background noise. However, filters do not exclusively act in the frequency domain; especially in the field of image processing many other targets for filtering exist.

There are many different bases of classifying filters and these overlap in many different ways; there is no simple hierarchical classification. Filters may be:

- analog or digital
- discrete-time (sampled) or continuous-time
- linear or non-linear
- time-invariant or time-variant, also known as shift invariance. If the filter operates in a spatial domain then the the characterization is space invariance.
- passive or active type of continuous-time filter
- infinite impulse response (IIR) or finite impulse response (FIR) type of discrete-time or digital filter.

Linear continuous-time filters

Linear continuous-time circuit is perhaps the most common meaning for filter in the signal processing world, and simply "filter" is often taken to be synonymous. These are filters that are designed to remove certain frequencies and allow others to pass. Such a filter is, of necessity, a linear filter. Any non-linearity will result in the output signal containing components of frequency which were not present in the input signal.

The modern design methodology for linear continuous-time filters is called network synthesis. Some important filter families designed in this way are:

- Chebyshev filter, has the best approximation to the ideal response of any filter for a specified order and ripple.
- Butterworth filter, has a maximally flat frequency response.
- Bessel filter, has a maximally flat phase delay.
- Elliptic filter, has the steepest cutoff of any filter for a specified order and ripple.

The difference between these filter families is that they all use a different polynomial function to approximate to the ideal filter response. This results in each having a different transfer function.

Another older, less-used methodology is the image parameter method. Filters designed by this methodology are archaically called "wave filters". Some important filters designed by this method are:

- Constant k filter, the original and simplest form of wave filter.
- m-derived filter, a modification of the constant k with improved cutoff steepness and impedance matching.

Terminology

Some terms used to describe and classify linear filters:

- The frequency response can be classified into a number of different bandforms describing which frequencies the filter passes (the passband) and which it rejects (the stopband):
    - Low-pass filter – low frequencies are passed, high frequencies are attenuated.
    - High-pass filter – high frequencies are passed, low frequencies are attenuated.
    - Band-pass filter – only frequencies in a frequency band are passed.
    - Band-stop filter or band-reject filter – only frequencies in a frequency band are attenuated.
    - Notch filter – rejects just one specific frequency - an extreme band-stop filter.
    - Comb filter – has multiple regularly spaced narrow passbands giving the bandform the appearance of a comb.
    - All-pass filter – all frequencies are passed, but the phase of the output is modified.

- Cutoff frequency is the frequency beyond which the filter will not pass signals. It is usually measured at a specific attenuation such as 3dB.
- Roll-off is the rate at which attenuation increases beyond the cut-off frequency.
- Transition band, the (usually narrow) band of frequencies between a passband and stopband.
- Ripple is the variation of the filters insertion loss in the passband.

- The order of a filter is the degree of the approximating polynomial and in passive filters corresponds to the number of elements required to build it. Increasing order increases roll-off and brings the filter closer to the ideal response.

Technologies

Filters can be built in a number of different technologies. The same transfer function can be realised in several different ways, that is the mathematical properties of the filter are the same but the physical properties are quite different. Often the components in different technologies are directly analogous to each other and fulfill the same role in their respective filters. For instance, the resistors, inductors and capacitors of electronics correspond respectively to dampers, masses and springs in mechanics. Likewise, there are corresponding components in distributed element filters.

- Electronic filters were originally entirely passive consisting of resistance, inductance and capacitance. Active technology makes design easier and opens up new possibilities in filter specifications.
- Digital filters operate on signals represented in digital form. The essence of a digital filter is that it directly implements a mathematical algorithm, corresponding to the desired filter transfer function, in its programming or microcode.
- Mechanical filters are built out of mechanical components. In the vast majority of cases they are used to process an electronic signal and transducers are provided to convert this to and from a mechanical vibration. However, examples do exist of filters that have been designed for operation entirely in the mechanical domain.
- Distributed element filters are constructed out of components made from small pieces of transmission line or other distributed elements. There are structures in distributed element filters that directly correspond to the lumped elements of electronic filters, and others that are unique to this class of technology.
- Waveguide filters consist of waveguide components or components inserted in the waveguide. Waveguides are a class of transmission line and many structures of distributed element filters, for instance the stub (electronics), can be implemented in waveguides also.
- Acoustic filters
- Optical filters were originally developed for purposes other than signal processing such as lighting and photography. With the rise of optical fiber technology, however, optical filters increasingly find signal processing applications and signal processing filter terminology, such as longpass and shortpass, are entering the field.

The transfer function

The transfer function $H(s)$ of a filter is the ratio of the output signal $Y(s)$ to that of the input signal $X(s)$ as a function of the complex frequency s:

$$H(s) = \frac{Y(s)}{X(s)}$$

with $s = \sigma + j\omega$.

The transfer function of all linear time-invariant filters generally share certain characteristics:

- For filters which are constructed of discrete components, their transfer function must be the ratio of two polynomials in s, i.e. a rational function of s. The order of the transfer function will be the highest power of $s$encountered in either the numerator or the denominator.
- The polynomials of the transfer function will all have real coefficients. Therefore, the poles and zeroes of the transfer function will either be real or occur in complex conjugate pairs.
- Since the filters are assumed to be stable, the real part of all poles (i.e. zeroes of the denominator) will be negative, i.e. they will lie in the left half-plane in complex frequency space.

Distributed element filters do not, in general, produce rational functions but can often approximate to them.

The proper construction of a transfer function involves the Laplace transform, and therefore it is needed to assume null initial conditions, because

$$\mathcal{L}\left\{\frac{df}{dt}\right\} = s \cdot \mathcal{L}\{f(t)\} - f(0),$$

And when f(0)=0 we can get rid of the constants and use the usual expression

$$\mathcal{L}\left\{\frac{df}{dt}\right\} = s \cdot \mathcal{L}\{f(t)\}$$

An alternative to transfer functions is to give the behavior of the filter as a convolution. The convolution theorem, which holds for Laplace transforms, guarantees equivalence with transfer functions.

Classification

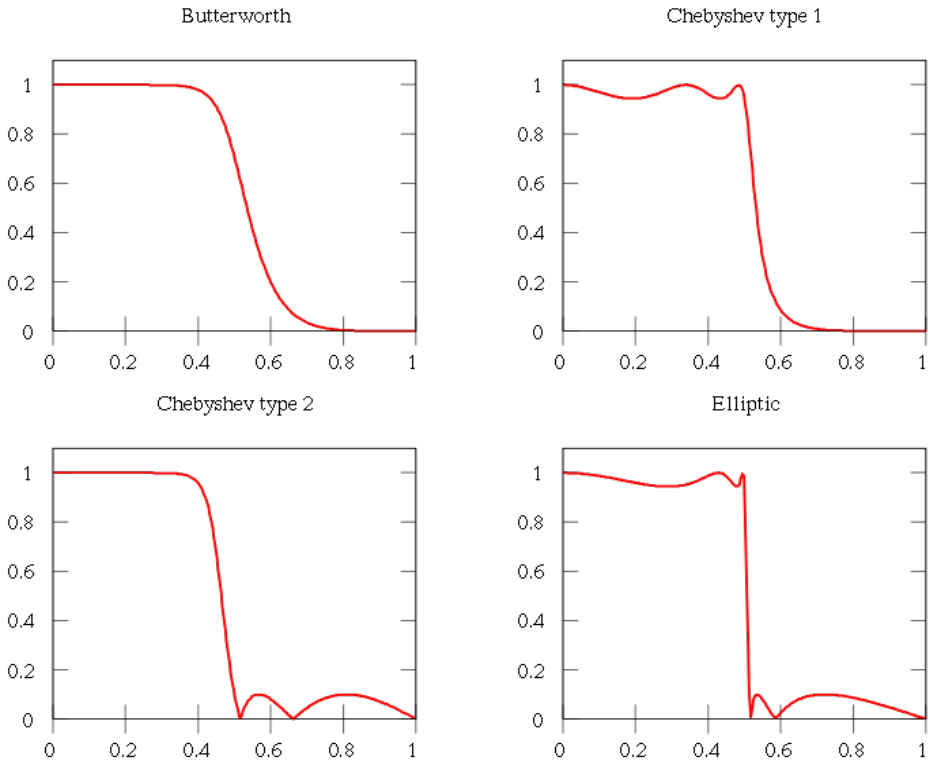
Filters may be specified by family and bandform. A filter's family is specified by the approximating polynomial used and each leads to certain characteristics of the transfer function of the filter. Some common filter families and their particular characteristics are:

- Butterworth filter – no gain ripple in pass band and stop band, slow cutoff

- Chebyshev filter (Type I) – no gain ripple in stop band, moderate cutoff
- Chebyshev filter (Type II) – no gain ripple in pass band, moderate cutoff
- Bessel filter – no group delay ripple, no gain ripple in both bands, slow gain cutoff
- Elliptic filter – gain ripple in pass and stop band, fast cutoff
- Optimum "L" filter
- Gaussian filter – no ripple in response to step function
- Hourglass filter
- Raised-cosine filter

Each family of filters can be specified to a particular order. The higher the order, the more the filter will approach the "ideal" filter; but also the longer the impulse response is and the longer the latency will be. An ideal filter has full transmission in the pass band, complete attenuation in the stop band, and an abrupt transition between the two bands, but this filter has infinite order (i.e., the response cannot be expressed as a linear differential equation with a finite sum) and infinite latency (i.e., its compact support in the Fourier transform forces its time response to be ever lasting).

Here is an image comparing Butterworth, Chebyshev, and elliptic filters. The filters in this illustration are all fifth-order low-pass filters. The particular implementation – analog or digital, passive or active – makes no difference; their output would be the same.



As is clear from the image, elliptic filters are sharper than all the others, but they show ripples on the whole bandwidth.

Any family can be used to implement a particular bandform of which frequencies are transmitted, and which, outside the passband, are more or less attenuated. The transfer function completely specifies the behavior of a linear filter, but not the particular technology used to implement it. In other words, there are a number of different ways of achieving a particular transfer function when designing a circuit. A particular bandform of filter can be obtained by transformation of a prototype filter of that family.

Impedance matching

Impedance matching structures invariably take on the form of a filter, that is, a network of non-dissipative elements. For instance, in a passive electronics implementation, this would likely take the form of a ladder topology of inductors and capacitors. The design of matching networks shares much in common with filters and the design invariably will have a filtering action as an incidental consequence. Although the prime purpose of a matching network is not to filter, it is often the case that both functions are combined in the same circuit. The need for impedance matching does not arise while signals are in the digital domain.

Some filters for specific purposes

- Audio filter
- Line filter
- Texture filtering

Filters for removing noise from data

- Wiener filter
- Kalman filter
- Savitzky–Golay smoothing filter

Chapter-7

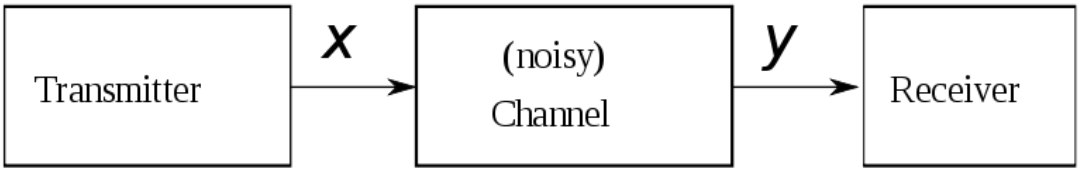
Channel Capacity and Frequency Mixer

Channel capacity

In electrical engineering, computer science and information theory, **channel capacity** is the tightest upper bound on the amount of information that can be reliably transmitted over a communications channel. By the noisy-channel coding theorem, the channel capacity of a given channel is the limiting information rate (in units of information per unit time) that can be achieved with arbitrarily small error probability.

Information theory, developed by Claude E. Shannon during World War II, defines the notion of channel capacity and provides a mathematical model by which one can compute it. The key result states that the capacity of the channel, as defined above, is given by the maximum of the mutual information between the input and output of the channel, where the maximization is with respect to the input distribution.

Formal definition



Let X represent the space of signals that can be transmitted, and Y the space of signals received, during a block of time over the channel. Let

$$p_{Y|X}(y|x)$$

be the conditional distribution function of Y given X. Treating the channel as a known statistic system, $p_{Y|X}(y \mid x)$ is an inherent fixed property of the communications channel (representing the nature of the noise in it). Then the joint distribution

$$p_{X,Y}(x,y)$$

of X and Y is completely determined by the channel and by the choice of

$$p_X(x) = \int_y p_{X,Y}(x,y)\,dy$$

the marginal distribution of signals we choose to send over the channel. The joint distribution can be recovered by using the identity

$$p_{X,Y}(x,y) = p_{Y|X}(y|x)\,p_X(x)$$

Under these constraints, next maximize the amount of information, or the message, that one can communicate over the channel. The appropriate measure for this is the mutual information $I(X;Y)$, and this maximum mutual information is called the **channel capacity** and is given by

$$C = \sup_{p_X} I(X;Y)$$

Noisy-channel coding theorem

The noisy-channel coding theorem states that for any $\varepsilon > 0$ and for any rate R less than the channel capacity C, there is an encoding and decoding scheme that can be used to ensure that the probability of block error is less than ε for a sufficiently long code. Also, for any rate greater than the channel capacity, the probability of block error at the receiver goes to one as the block length goes to infinity.

Example application

An application of the channel capacity concept to an additive white Gaussian noise (AWGN) channel with B Hz bandwidth and signal-to-noise ratio S/N is the Shannon–Hartley theorem:

$$C = B \log\left(1 + \frac{S}{N}\right)$$

C is measured in bits per second if the logarithm is taken in base 2, or nats per second if the natural logarithm is used, assuming B is in hertz; the signal and noise powers S and N are measured in watts or volts^2, so the signal-to-noise ratio here is expressed as a power ratio, *not* in decibels (dB); since figures are often cited in dB, a conversion may be needed. For example, 30 dB is a power ratio of $10^{30\,/\,10} = 10^3 = 1000$.

AWGN channel

If the average received power is $\bar{P}$[W] and the noise power spectral density is N_0 [W/Hz], the AWGN channel capacity is

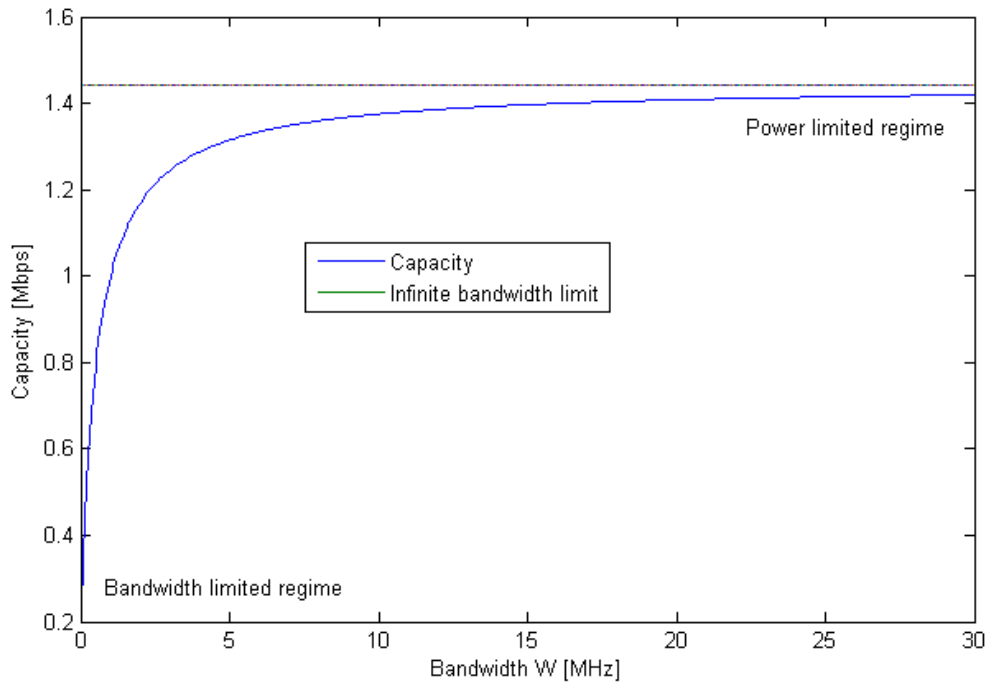
$$C_{awgn} = W \log_2 \left(1 + \frac{\bar{P}}{N_0 W}\right)_{\text{[bits/Hz]}},$$

where $\frac{\bar{P}}{N_0 W}$ is the received signal-to-noise ratio (SNR).

When the SNR is large (SNR >> 0 dB), the capacity $C \approx W \log_2 \frac{\bar{P}}{N_0 W}$ is logarithmic in power and approximately linear in bandwidth. This is called the *bandwidth-limited regime*.

When the SNR is small (SNR << 0 dB), the capacity $C \approx \frac{\bar{P}}{N_0} \log_2 e$ is linear in power but insensitive to bandwidth. This is called the *power-limited regime*.

The bandwidth-limited regime and power-limited regime are illustrated in the figure.



AWGN channel capacity with the power-limited regime and bandwidth-limited regime indicated. Here, $\frac{\bar{P}}{N_o} = 10^6$.

Frequency-selective channel

The capacity of the frequency-selective channel is given by so-called waterfilling power allocation,

$$C_{N_c} = \sum_{n=0}^{N_c-1} \log_2 \left(1 + \frac{P_n^* |\bar{h}_n|^2}{N_0}\right),$$

where $P_n^* = \max\left(\left(\frac{1}{\lambda} - \frac{N_0}{|\bar{h}_n|^2}\right), 0\right)$ and $|\bar{h}_n|^2$ is the gain of subchannel n, with λ chosen to meet the power constraint.

Slow-fading channel

In a slow-fading channel, where the coherence time is greater than the latency requirement, there is no definite capacity as the maximum rate of reliable communications supported by the channel, $\log_2(1 + |h|^2 SNR)$, depends on the random channel gain $|h|^2$. If the transmitter encodes data at rate R [bits/s/Hz], there is a certain probability that the decoding error probability cannot be made arbitrarily small,

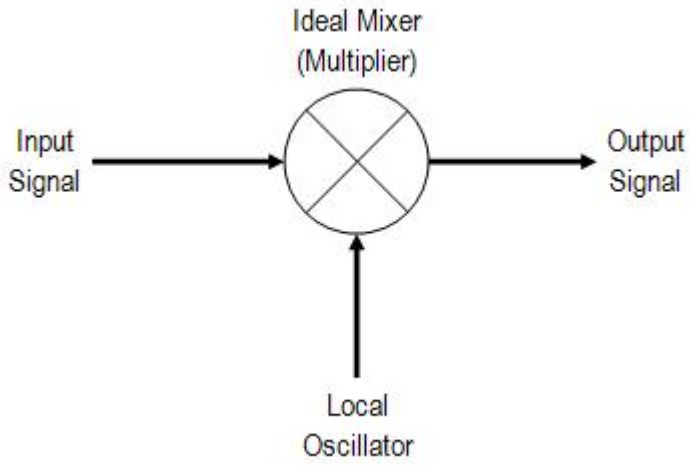
$$p_{out} = \mathbb{P}(\log(1 + |h|^2 SNR) < R),$$

in which case the system is said to be in outage. With a non-zero probability that the channel is in deep fade, the capacity of the slow-fading channel in strict sense is zero. However, it is possible to determine the largest value of R such that the outage probability p_{out} is less than ε. This value is known as the ε-outage capacity.

Fast-fading channel

In a fast-fading channel, where the latency requirement is greater than the coherence time and the codeword length spans many coherence periods, one can average over many independent channel fades by coding over a large number of coherence time intervals. Thus, it is possible to achieve a reliable rate of communication of $\mathbb{E}(\log_2(1 + |h|^2 SNR))$ [bits/s/Hz] and it is meaningful to speak of this value as the capacity of the fast-fading channel.

Frequency mixer



Frequency Mixer Symbol

In electronics a **mixer** or **frequency mixer** is a nonlinear electrical circuit that creates new frequencies from two signals applied to it. In its most common application, two signals at frequencies f_1 and f_2 are applied to a mixer, and it produces new signals at the sum $f_1 + f_2$ and difference $f_1 - f_2$ of the original frequencies. Other frequency components may also be produced in a practical frequency mixer.

Mixers are widely used to shift signals from one frequency range to another, a process known as heterodyning, for convenience in transmission or further signal processing. For example, a key component of a superheterodyne receiver is a mixer used to move received signals to a common intermediate frequency. Frequency mixers are also used to modulate a carrier frequency in radio transmitters.

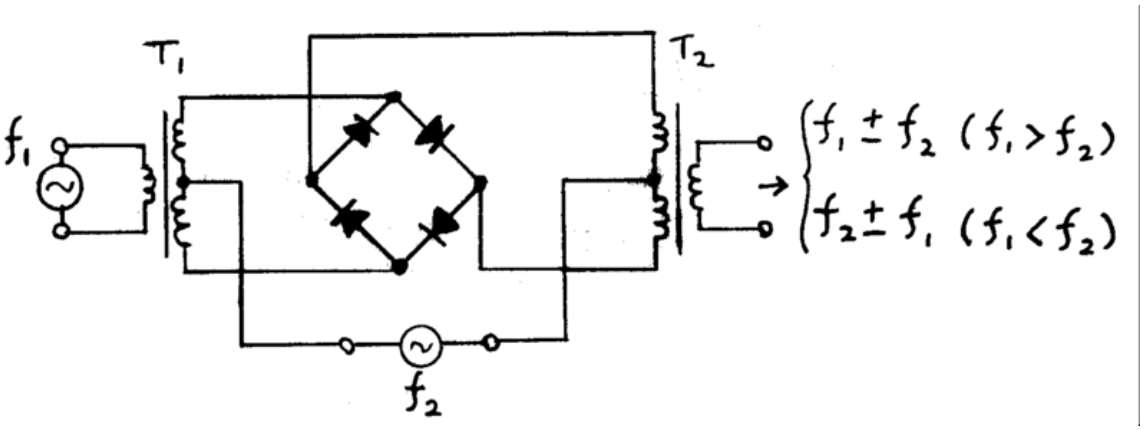
Types

Passive mixers use one or more diodes and rely on the non-linear relation between voltage and current to provide the multiplying element. In a passive mixer, the desired output signal is always of lower power than the input signals. Active mixers can increase the strength of the product signal. Active mixers improve isolation between the ports, but may have higher noise and more power consumption; an active mixer can be less tolerant of overload.

Mixers may be built of discrete components, may be part of integrated circuits, or can be delivered as hybrid modules.

Mixers may also be classified by their topology. Unbalanced mixers allow some of the input signal power to pass through to the output. A single-balanced mixer is arranged so

that the local oscillator, (or RF) signal port, cancels and cannot pass through to the output. A doubly-balanced mixer has symmetrical paths for both inputs, and will have no output if either input signal is not present.



Schematic diagram of a double-balanced passive diode mixer. There is no output unless both f1 and f2 inputs are present.

Selection of a mixer type is a trade off for a particular application. Mixer circuits are characterized by conversion gain, and noise figure. Balanced and double-balanced designs allow less of the input signals to feed through to the output.

Nonlinear electronic components that are used as mixers include diodes, transistors biased near cutoff, and at lower frequencies, analog multipliers. Ferromagnetic-core inductors driven into saturation have also been used. In nonlinear optics, crystals with nonlinear characteristics are used to mix two frequencies of laser light to create optical heterodynes.

Diode

A diode can be used to create a simple unbalanced mixer. This type of mixer produces the original frequencies as well as their sum and their difference. The importance of the diode is that it is non-linear (or non-Ohmic), which means its response (current) is not proportional to its input (voltage). The diode therefore does not reproduce the frequencies of its driving voltage in the current through it, which allows the desired frequency manipulation. Certain other non-linear devices such as tunnel diodes or Gunn diodes can be utilized similarly.

The current *I* through an ideal diode as a function of the voltage *V* across it is given by

$$I = I_S \left(e^{\frac{qV_D}{nkT}} - 1 \right)$$

where what is important is that *V* appears in *e*'s exponent. The exponential can be expanded as

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

and can be approximated for small *x* (that is, small voltages) by the first few terms of that series:

$$e^x - 1 \approx x + \frac{x^2}{2}$$

Suppose that the sum of the two input signals $v_1 + v_2$ is applied to a diode, and that an output voltage is generated that is proportional to the current through the diode (perhaps by providing the voltage that is present across a resistor in series with the diode). Then, disregarding the constants in the diode equation, the output voltage will have the form

$$v_o = (v_1 + v_2) + \frac{1}{2}(v_1 + v_2)^2 + \dots$$

The first term on the right is the original two signals, as expected, followed by the square of the sum, which can be rewritten as $(v_1 + v_2)^2 = v_1^2 + 2v_1v_2 + v_2^2$, where the multiplied signal is obvious. The ellipsis represents all the higher powers of the sum which we assume to be negligible for small signals.

Switching

Another form of mixer operates by switching, with the smaller input signal being passed inverted or uninverted according to the phase of the local oscillator (LO). This would be typical of the normal operating mode of a packaged double balanced mixer module such as an SBL-1, with the local oscillator drive considerably higher than the signal amplitude.

The aim of a switching mixer is to achieve linear operation over the signal level, and hard switching driven by the local oscillator. Mathematically the switching mixer is not much different from a multiplying mixer, just because instead of the LO sine wave term we would use the signum function. In the frequency domain the switching mixer operation leads to the usual sum and difference frequencies, but also to further terms e.g. +-3*fLO, +-5*fLO, etc. The advantage of a switching mixer is that it can achieve - with the same effort - a lower noise figure (NF) and larger conversion gain. This come because the switching diodes or transistors act either like a low resistor (switch closed) or large resistor (switch open) and in both cases only minimum noise is added. From the circuit perspective many multiplying mixers can be used as switching mixers, just by increasing the LO amplitude. So RF engineers simply talk about mixers, and mean switching mixers.

Applications

The mixer circuit can be used not only to shift the frequency of an input signal as in a receiver, but also as a product detector, modulator, phase detector or frequency multiplier. For example a communications receiver might contain two mixer stages for conversion of the input signal to an intermediate frequency, and another mixer employed as a detector for demodulation of the signal.

Chapter-8

Hilbert–Huang Transform

The **Hilbert–Huang transform** (**HHT**) is a way to decompose a signal into so-called intrinsic mode functions (IMF), and obtain instantaneous frequency data. It is designed to work well for data that are nonstationary and nonlinear. In contrast to other common transforms like the Fourier transform, the HHT is more like an algorithm (an empirical approach) that can be applied to a data set, rather than a theoretical tool.

Introduction

The **Hilbert–Huang transform** (**HHT**), a NASA designated name, was proposed by Huang et al. (1996, 1998, 1999, 2003). It is the result of the empirical mode decomposition (EMD) and the Hilbert spectral analysis (HSA). The HHT uses the EMD method to decompose a signal into so-called intrinsic mode function, and uses the HSA method to obtain instantaneous frequency data. The HHT provides a new method of analyzing nonstationary and nonlinear time series data.

Introduction to EMD and IMF

The fundamental part of the HHT is the **empirical mode decomposition** (**EMD**) method. Using the EMD method, any complicated data set can be decomposed into a finite and often small number of components, which is a collection of **intrinsic mode functions** (**IMF**). An IMF represents a generally simple oscillatory mode as a counterpart to the simple harmonic function. By definition, an IMF is any function with the same number of extrema and zero crossings, with its envelopes being symmetric with respect to zero. The definition of an IMF guarantees a well-behaved Hilbert transform of the IMF. This decomposition method operating in the time domain is adaptive and highly efficient. Since the decomposition is based on the local characteristic time scale of the data, it can be applied to nonlinear and nonstationary processes.

Introduction to HSA

The Hilbert spectral analysis (HSA) provides a method for examining the IMF's instantaneous frequency data as functions of time that give sharp identifications of

embedded structures. The final presentation of the results is an energy-frequency-time distribution, designated as the Hilbert spectrum.

Techniques

The empirical mode decomposition (EMD)

The EMD method is a necessary step to reduce any given data into a collection of intrinsic mode functions (IMF) to which the Hilbert spectral analysis can be applied. An IMF is defined as a function that satisfies the following requirements:

1. In the whole data set, the number of extrema and the number of zero-crossings must either be equal or differ at most by one.
2. At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

Therefore, an IMF represents a simple oscillatory mode as a counterpart to the simple harmonic function, but it is much more general: instead of constant amplitude and frequency in a simple harmonic component, an IMF can have variable amplitude and frequency along the time axis.
The procedure of extracting an IMF is called sifting. The sifting process is as follows:

1. Identify all the local extrema in the test data.
2. Connect all the local maxima by a cubic spline line as the upper envelope.
3. Repeat the procedure for the local minima to produce the lower envelope.

The upper and lower envelopes should cover all the data between them. Their mean is m_1. The difference between the data and m_1 is the first component h_1:

$$X(t) - m_1 = h_1 .$$

Ideally, h_1 should satisfy the definition of an IMF, for the construction of h_1 described above should have made it symmetric and having all maxima positive and all minima negative. After the first round of sifting, the crest may become a local maximum. New extrema generated in this way actually reveal the proper modes lost in the initial examination. In the subsequent sifting process, h_1 can only be treated as a proto-IMF. In the next step, it is treated as the data, then

$$h_1 - m_{11} = h_{11}.$$

After repeated sifting up to k times, h_1 becomes an IMF, that is

$$h_{1(k-1)} - m_{1k} = h_{1k}.$$

Then, it is designated as the first IMF component from the data:

$$c_1 = h_{1k}.$$

The stoppage criteria of the sifting process

The stoppage criterion determines the number of sifting steps to produce an IMF. Two different stoppage criteria have been used traditionally:

- 1. The first criterion is proposed by Huang et al. (1998). It similar to the Cauchy convergence test, and we define a sum of the difference, SD, as

  $$SD_k = \frac{\sum_{t=0}^{T} |h_{k-1}(t) - h_k(t)|^2}{\sum_{t=0}^{T} h_{k-1}^2(t)}.$$

  Then the sifting process is stop when SD is smaller than a pre-given value.

- 2. A second criterion is based on the number called the S-number, which is defined as the number of consecutive siftings when the numbers of zero-crossings and extrema are equal or at most differing by one. Specifically, an S-number is pre-selected. The sifting process will stop only if for S consecutive times the numbers of zero-crossings and extrema stay the same, and are equal or at most differ by one.

Once a stoppage criterion is selected, the first IMF, c_1, can be obtained. Overall, c_1 should contain the finest scale or the shortest period component of the signal. We can, then, separate c_1 from the rest of the data by $X(t) - c_1 = r_1$. Since the residue, r_1, still contains longer period variations in the data, it is treated as the new data and subjected to the same sifting process as described above.

This procedure can be repeated to all the subsequent r_j's, and the result is

$$r_{n-1} - c_n = r_n.$$

The sifting process stops finally when the residue, r_n, becomes a monotonic function from which no more IMF can be extracted. From the above equations, we can induce that

$$X(t) = \sum_{j=1}^{n} c_j + r_n.$$

Thus, a decomposition of the data into n-empirical modes is achieved. The components of the EMD are usually physically meaningful, for the characteristic scales are defined by the physical data. Flandrin et al. (2003) and Wu and Huang (2004) have shown that the EMD is equivalent to a dyadic filter bank.

Hilbert spectral analysis

Having obtained the intrinsic mode function components, the instantaneous frequency can be computed using the Hilbert Transform. After performing the Hilbert transform on each IMF component, the original data can be expressed as the real part, Real, in the following form:

$$X(t) = \text{Real}\sum_{j=1}^{n} a_j(t)e^{i\int \omega_j(t)dt}.$$

Current applications

- **Biomedical applications**: Huang et al. [1999b] analyzed the pulmonary arterial pressure on conscious and unrestrained rats.

- **Chemistry and chemical engineering**: Phillips et al. [2003] investigated a conformational change in Brownian dynamics(BD) and molecular dynamics(MD) simulations using a comparative analysis of HHT and wavelet methods. Wiley et al. [2004] used HHT to investigate the effect of reversible digitally filtered molecular dynamics(RDFMD) which can enhance or suppress specific frequencies of motion. Montesinos et al. [2002] applied HHT to signals obtained from BWR neuron stability.

- **Financial applications**: Huang et al. [2003b] applied HHT to nonstationary financial time series and used a weekly mortgage rate data.

- **Image processing**: Hariharan et al. [2006] applied EMD to image fusion and enhancement. Chang et al. [2009] applied an improved EMD to iris recognition, which reported a 100% faster in computational speed without losing accuracy than the original EMD.

- **Meteorological and atmospheric applications**: Salisbury and Wimbush [2002], using Southern Oscillation Index(SOI) data, applied the HHT technique to determine whether the SOI data are sufficiently noise free that useful predictions can be made and whether future El Nino southern oscillation(ENSO) events can be predicted from SOI data. Pan et al. [2002] used HHT to analyze satellite scatterometer wind data over the northwestern Pacific and compared the results to vector empirical orthogonal function(VEOF) results.

- **Ocean engineering**:Schlurmann [2002] introduced the application of HHT to characterize nonlinear water waves from two different perspectives, using laboratory experiments. Veltcheva [2002] applied HHT to wave data from nearshore sea. Larsen et al. [2004] used HHT to characterize the underwater electromagnetic environment and identify transient manmade electromagnetic disturbances.

- **Seismic studies**: Huang et al. [2001] used HHT to develop a spectral representation of earthquake data. Chen et al. [2002a] used HHT to determined the dispersion curves of seismic surface waves and compared their results to Fourier-based time-frequency analysis. Shen et al. [2003] applied HHT to ground motion and compared the HHT result with the Fourier spectrum.

- **Solar Physics**: Barnhart and Eichinger [2010] used HHT to extract the periodic components within sunspot data, including the 11-year Schwabe, 22-year Hale, and ~100-year Gleissberg cycles. They compared their results with traditional Fourier analysis.

- **Structural applications**: Quek et al. [2003] illustrate the feasibility of the HHT as a signal processing tool for locating an anomaly in the form of a crack, delamination, or stiffness loss in beams and plates based on physically acquired propagating wave signals. Using HHT, Li et al. [2003] analyzed the results of a pseudodynamic test of two rectangular reinforced concrete bridge columns.

- **Health monitoring**: Pines and Salvino [2002] applied HHT in structural health monitoring. Yang et al. [2004] used HHT for damage detection, applying EMD to extract damage spikes due to sudden changes in structural stiffness. Yu et al. [2003] used HHT for fault diagnosis of roller bearings.

- **System identification**: Chen and Xu [2002] explored the possibility of using HHT to identify the modal damping ratios of a structure with closely spaced modal frequencies and compared their results to FFT. Xu et al. [2003] compared the modal frequencies and damping ratios in various time increments and different winds for one of the tallest composite buildings in the world.

Limitations

Chen and Feng [undated][1] proposed a technique to improve the HHT procedure. The authors noted that the EMD is limited in distinguishing different components in narrow-band signals. The narrow band may contain either (a) components that have adjacent frequencies or (b) components that are not adjacent in frequency but for which one of the components has a much higher energy intensity than the other components. The improved technique is based on beating-phenomenon waves.

Datig and Schlurmann [2004] did the most comprehensive studies on the performance and limitations of HHT with particular applications to irregular waves. The authors did extensive investigation into the spline interpolation. The authors discussed using additional points, both forward and backward, to determine better envelopes. They also performed a parametric study on the proposed improvement and showed significant improvement in the overall EMD computations. The authors noted that HHT is capable of differentiating between time-variant components from any given data. Their study also showed that HHT was able to distinguish between riding and carrier waves.

Chapter-9

Intersymbol Interference and Pulse Shaping

Intersymbol interference

In telecommunication, **intersymbol interference** (**ISI**) is a form of distortion of a signal in which one symbol interferes with subsequent symbols. This is an unwanted phenomenon as the previous symbols have similar effect as noise, thus making the communication less reliable. ISI is usually caused by multipath propagation or the inherent non-linear frequency response of a channel causing successive symbols to "blur" together. The presence of ISI in the system introduces errors in the decision device at the receiver output. Therefore, in the design of the transmitting and receiving filters, the objective is to minimize the effects of ISI, and thereby deliver the digital data to its destination with the smallest error rate possible. Ways to fight intersymbol interference include adaptive equalization and error correcting codes.

Causes

Multipath propagation

One of the causes of intersymbol interference is what is known as multipath propagation in which a wireless signal from a transmitter reaches the receiver via many different paths. The causes of this include reflection (for instance, the signal may bounce off buildings), refraction (such as through the foliage of a tree) and atmospheric effects such as atmospheric ducting and ionospheric reflection. Since all of these paths are different lengths - plus some of these effects will also slow the signal down - this results in the different versions of the signal arriving at different times. This delay means that part or all of a given symbol will be spread into the subsequent symbols, thereby interfering with the correct detection of those symbols. Additionally, the various paths often distort the amplitude and/or phase of the signal thereby causing further interference with the received signal.

Bandlimited channels

Another cause of intersymbol interference is the transmission of a signal through a bandlimited channel, i.e., one where the frequency response is zero above a certain frequency (the cutoff frequency). Passing a signal through such a channel results in the removal of frequency components above this cutoff frequency; in addition, the amplitude of the frequency components below the cutoff frequency may also be attenuated by the channel.

This filtering of the transmitted signal affects the shape of the pulse that arrives at the receiver. The effects of filtering a rectangular pulse; not only change the shape of the pulse within the first symbol period, but it is also spread out over the subsequent symbol periods. When a message is transmitted through such a channel, the spread pulse of each individual symbol will interfere with following symbols.

As opposed to multipath propagation, bandlimited channels are present in both wired and wireless communications. The limitation is often imposed by the desire to operate multiple independent signals through the same area/cable; due to this, each system is typically allocated a piece of the total bandwidth available. For wireless systems, they may be allocated a slice of the electromagnetic spectrum to transmit in (for example, FM radio is often broadcast in the 87.5 MHz - 108 MHz range). This allocation is usually administered by a government agency; in the case of the United States this is the Federal Communications Commission (FCC). In a wired system, such as an optical fiber cable, the allocation will be decided by the owner of the cable.

The bandlimiting can also be due to the physical properties of the medium - for instance, the cable being used in a wired system may have a cutoff frequency above which practically none of the transmitted signal will propagate.

Communication systems that transmit data over bandlimited channels usually implement pulse shaping to avoid interference caused by the bandwidth limitation. If the channel frequency response is flat and the shaping filter has a finite bandwidh, it is possible to communicate with no ISI at all. Often the channel response is not known beforehand, and an adaptive equalizer is used to compensate the frequency response.

Effects on eye patterns

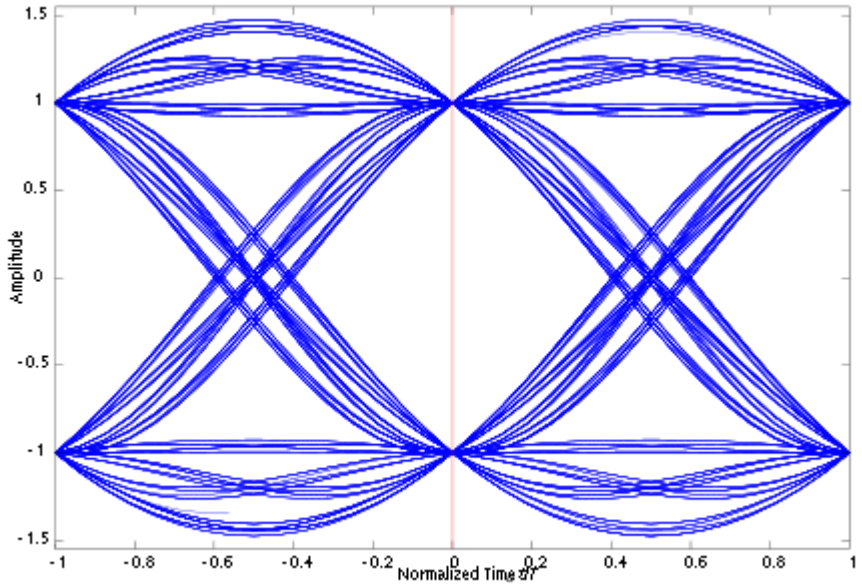
One way to study ISI in a PCM or data transmission system experimentally is to apply the received wave to the vertical deflection plates of an oscilloscope and to apply a sawtooth wave at the transmitted symbol rate R, 1/T to the horizontal deflection plates. The resulting display is called an eye pattern because of its resemblance to the human eye for binary waves. The interior region of the eye pattern is called the eye opening. An eye pattern provides a great deal of information about the performance of the pertinent system.

1. The width of the eye opening defines the time interval over which the received wave can be sampled without error from ISI. It is apparent that the preferred time for sampling is the instant of time at which the eye is open widest.
2. The sensitivity of the system to timing error is determined by the rate of closure of the eye as the sampling time is varied.
3. The height of the eye opening, at a specified sampling time, defines the margin over noise.

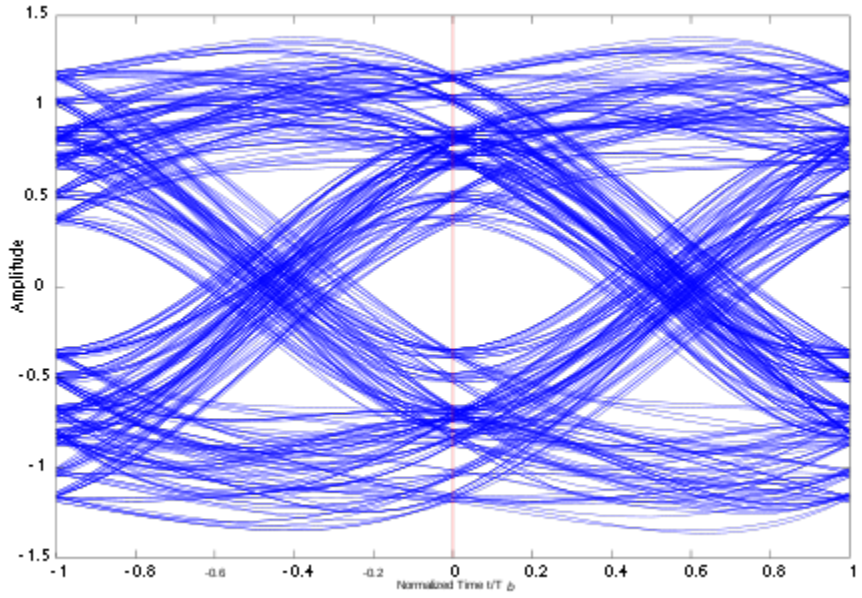
An eye pattern, which overlays many samples of a signal, can give a graphical representation of the signal characteristics. The first image below is the eye pattern for a binary phase-shift keying (PSK) system in which a one is represented by an amplitude of -1 and a zero by an amplitude of +1. The current sampling time is at the center of the image and the previous and next sampling times are at the edges of the image. The various transitions from one sampling time to another (such as one-to-zero, one-to-one and so forth) can clearly be seen on the diagram.

The noise margin - the amount of noise required to cause the receiver to get an error - is given by the distance between the signal and the zero amplitude point at the sampling time; in other words, the further from zero at the sampling time the signal is the better. For the signal to be correctly interpreted, it must be sampled somewhere between the two points where the zero-to-one and one-to-zero transitions cross. Again, the further apart these points are the better, as this means the signal will be less sensitive to errors in the timing of the samples at the receiver.

The effects of ISI are shown in the second image which is an eye pattern of the same system when operating over a multipath channel. The effects of receiving delayed and distorted versions of the signal can be seen in the loss of definition of the signal transitions. It also reduces both the noise margin and the window in which the signal can be sampled, which shows that the performance of the system will be worse (i.e. it will have a greater bit error ratio).



The eye diagram of a binary PSK system

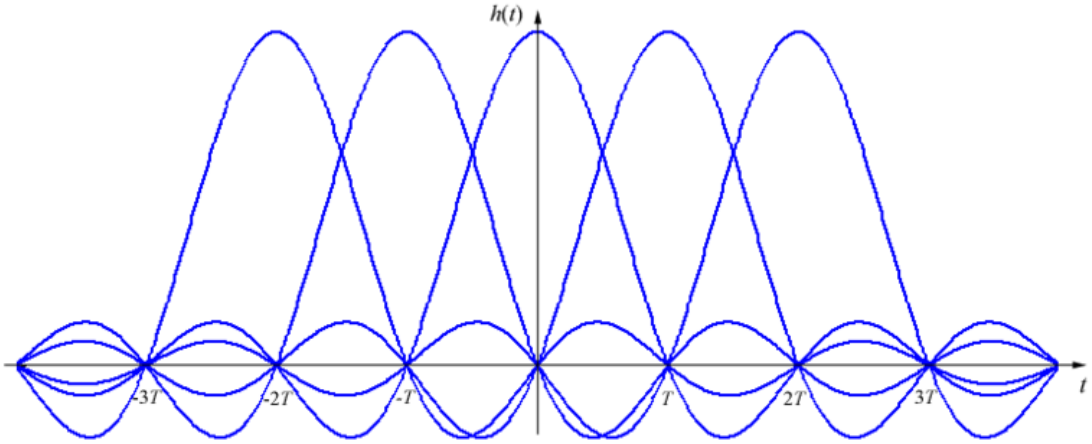


The eye diagram of the same system with multipath effects added

Countering ISI

There are several techniques in telecommunication and data storage that try to work around the problem of intersymbol interference.

- Design systems such that the impulse response is short enough that very little energy from one symbol smears into the next symbol.



Consecutive raised-cosine impulses, demonstrating zero-ISI property

- Separate symbols in time with guard periods.
- Apply an equalizer at the receiver, that, broadly speaking, attempts to undo the effect of the channel by applying an inverse filter.
- Apply a sequence detector at the receiver, that attempts to estimate the sequence of transmitted symbols using the Viterbi algorithm.

Pulse shaping

In digital telecommunication, **pulse shaping** is the process of changing the waveform of transmitted pulses. Its purpose is to make the transmitted signal better suited to the communication channel by limiting the effective bandwidth of the transmission. By filtering the transmitted pulses this way, the intersymbol interference caused by the channel can be kept in control. In RF communication, pulse shaping is essential for making the signal fit in its frequency band.

Typically pulse shaping occurs after line coding and before modulation.

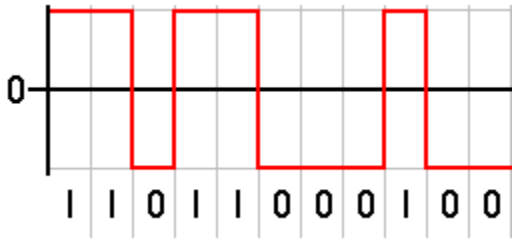
Need for pulse shaping

Transmitting a signal at high modulation rate through a band-limited channel can create intersymbol interference. As the modulation rate increases, the signal's bandwidth increases. When the signal's bandwidth becomes larger than the channel bandwidth, the channel starts to introduce distortion to the signal. This distortion is usually seen as intersymbol interference.

The signal's spectrum is determined by the pulse shaping filter used by the transmitter. Usually the transmitted symbols are represented as a time sequence of dirac delta pulses. This theoretical signal is then filtered with the pulse shaping filter, producing the transmitted signal. The spectrum of the transmission is thus determined by the filter.

In many base band communication systems the pulse shaping filter is implicitly a boxcar filter. Its spectrum is of the form *sin(x)/x*, and has significant signal power at frequencies higher than symbol rate. This is not a big problem when optical fibre or even twisted pair cable is used as the communication channel. However, in RF communications this would waste bandwidth, and only tightly specified frequency bands are used for single transmissions. In other words, the channel for the signal is band-limited. Therefore better filters have been developed, which attempt to minimise the bandwidth needed for a certain symbol rate.

Pulse filters



A typical NRZ coded signal is implicitly filtered with a boxcar filter.

Not all filters can be used as a pulse shaping filter. The filter itself must not introduce intersymbol interference — it needs to satisfy certain criteria. Nyquist ISI criterion is commonly used criterion for evaluation of filters, because it relates the frequency spectrum of the transmitter signal to intersymbol interference.

Examples of pulse-shaping filters that are commonly found in communication systems are:

- The trivial boxcar filter
- Sinc shaped filter
- Raised-cosine filter
- Gaussian filter

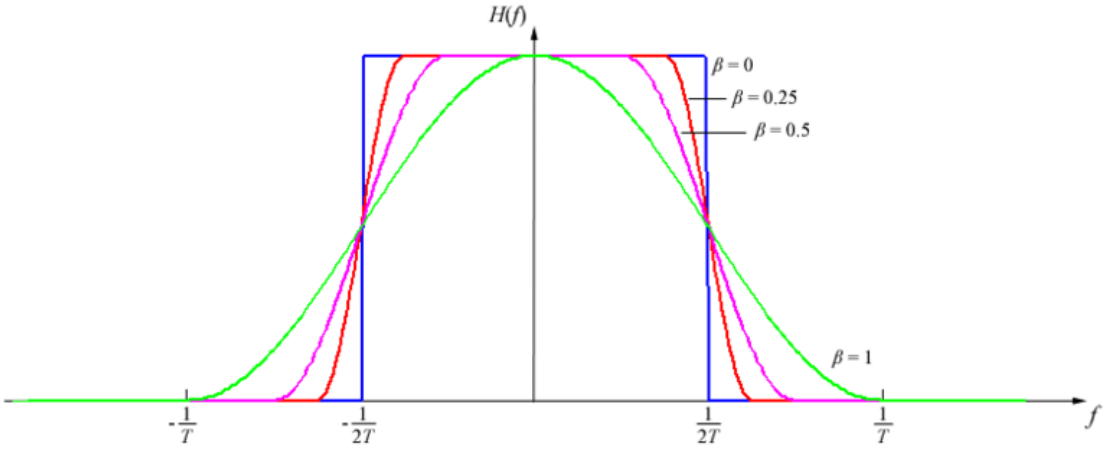
Sender side pulse shaping is often combined with a receiver side matched filter to achieve optimum tolerance for noise in the system. In this case the pulse shaping is equally distributed to the sender and receiver filters. The filters' amplitude responses are thus pointwise square-roots of the system filters.

Other approaches that eliminate complex pulse shaping filters have been invented. In OFDM, the carriers are modulated so slowly that each carrier is virtually unaffected by the bandwidth limitation of the channel.

Boxcar filter

The boxcar filter results in infinitely wide bandwidth for the signal. Thus its usefulness is limited, but it is used widely in wired baseband communications, where the channel has some extra bandwidth and the distortion created by the channel can be tolerated.

Sinc filter



Amplitude response of raised-cosine filter with various roll-off factors

Theoretically the best pulse shaping filter would be the sinc filter, but it cannot be implemented precisely. It is a non-causal filter with relatively slowly decaying tails. It is also problematic from a synchronisation point of view as any phase error results in steeply increasing intersymbol interference.

Raised-cosine filter

Raised-cosine filter is practical to implement and it is in wide use. It has a parametrisable excess bandwidth, so communication systems can choose a trade-off between a more complex filter and spectral efficiency.

Gaussian filter

This gives an output pulse shaped like a Gaussian function.

Chapter-10

Matched Filter

In telecommunications, a **matched filter** (originally known as a **North filter**) is obtained by correlating a known signal, or template, with an unknown signal to detect the presence of the template in the unknown signal. This is equivalent to convolving the unknown signal with a conjugated time-reversed version of the template. The matched filter is the optimal linear filter for maximizing the signal to noise ratio (SNR) in the presence of additive stochastic noise. Matched filters are commonly used in radar, in which a known signal is sent out, and the reflected signal is examined for common elements of the out-going signal. Pulse compression is an example of matched filtering. Two-dimensional matched filters are commonly used in image processing, e.g., to improve SNR for X-ray pictures.

Derivation of the matched filter

The following section derives the matched filter for a discrete-time system. The derivation for a continuous-time system is similar, with summations replaced with integrals.

The matched filter is the linear filter, *h*, that maximizes the output signal-to-noise ratio.

$$y[n] = \sum_{k=-\infty}^{\infty} h[n-k]x[k].$$

Though we most often express filters as the impulse response of convolution systems, as above, it is easiest to think of the matched filter in the context of the inner product, which we will see shortly.

We can derive the linear filter that maximizes output signal-to-noise ratio by invoking a geometric argument. The intuition behind the matched filter relies on correlating the received signal (a vector) with a filter (another vector) that is parallel with the signal, maximizing the inner product. This enhances the signal. When we consider the additive stochastic noise, we have the additional challenge of minimizing the output due to noise by choosing a filter that is orthogonal to the noise.

Let us formally define the problem. We seek a filter, *h*, such that we maximize the output signal-to-noise ratio, where the output is the inner product of the filter and the observed signal *x*.

Our observed signal consists of the desirable signal *s* and additive noise *v*:

$$x = s + v.$$

Let us define the covariance matrix of the noise, reminding ourselves that this matrix has Hermitian symmetry, a property that will become useful in the derivation:

$$R_v = E\{vv^H\}$$

where $.^H$ denotes Hermitian (conjugate) transpose, and *E* denotes expectation. Let us call our output, *y*, the inner product of our filter and the observed signal such that

$$y = \sum_{k=-\infty}^{\infty} h^*[k]x[k] = h^H x = h^H s + h^H v = y_s + y_v.$$

We now define the signal-to-noise ratio, which is our objective function, to be the ratio of the power of the output due to the desired signal to the power of the output due to the noise:

$$SNR = \frac{|y_s|^2}{E\{|y_v|^2\}}.$$

We rewrite the above:

$$SNR = \frac{|h^H s|^2}{E\{|h^H v|^2\}}.$$

We wish to maximize this quantity by choosing *h*. Expanding the denominator of our objective function, we have

$$E\{|h^H v|^2\} = E\{(h^H v)(h^H v)^H\} = h^H E\{vv^H\}h = h^H R_v h.$$

Now, our *SNR* becomes

$$SNR = \frac{|h^H s|^2}{h^H R_v h}.$$

We will rewrite this expression with some matrix manipulation. The reason for this seemingly counterproductive measure will become evident shortly. Exploiting the Hermitian symmetry of the covariance matrix R_v, we can write

$$SNR = \frac{|(R_v^{1/2}h)^H(R_v^{-1/2}s)|^2}{(R_v^{1/2}h)^H(R_v^{1/2}h)},$$

We would like to find an upper bound on this expression. To do so, we first recognize a form of the Cauchy-Schwarz inequality:

$$|a^H b|^2 \leq (a^H a)(b^H b),$$

which is to say that the square of the inner product of two vectors can only be as large as the product of the individual inner products of the vectors. This concept returns to the intuition behind the matched filter: this upper bound is achieved when the two vectors *a* and *b* are parallel. We resume our derivation by expressing the upper bound on our *SNR* in light of the geometric inequality above:

$$SNR = \frac{|(R_v^{1/2}h)^H(R_v^{-1/2}s)|^2}{(R_v^{1/2}h)^H(R_v^{1/2}h)} \leq \frac{\left[(R_v^{1/2}h)^H(R_v^{1/2}h)\right]\left[(R_v^{-1/2}s)^H(R_v^{-1/2}s)\right]}{(R_v^{1/2}h)^H(R_v^{1/2}h)}.$$

Our valiant matrix manipulation has now paid off. We see that the expression for our upper bound can be greatly simplified:

$$SNR = \frac{|(R_v^{1/2}h)^H(R_v^{-1/2}s)|^2}{(R_v^{1/2}h)^H(R_v^{1/2}h)} \leq s^H R_v^{-1} s.$$

We can achieve this upper bound if we choose,

$$R_v^{1/2}h = \alpha R_v^{-1/2}s$$

where α is an arbitrary real number. To verify this, we plug into our expression for the output *SNR*:

$$SNR = \frac{|(R_v^{1/2}h)^H(R_v^{-1/2}s)|^2}{(R_v^{1/2}h)^H(R_v^{1/2}h)} = \frac{\alpha^2|(R_v^{-1/2}s)^H(R_v^{-1/2}s)|^2}{\alpha^2(R_v^{-1/2}s)^H(R_v^{-1/2}s)} = \frac{|s^H R_v^{-1}s|^2}{s^H R_v^{-1}s} = s^H R_v^{-1} s.$$

Thus, our optimal matched filter is

$$h = \alpha R_v^{-1}s.$$

We often choose to normalize the expected value of the power of the filter output due to the noise to unity. That is, we constrain

$$E\{|y_v|^2\} = 1.$$

This constraint implies a value of α, for which we can solve:

$$E\{|y_v|^2\} = \alpha^2 s^H R_v^{-1} s = 1,$$

yielding

$$\alpha = \frac{1}{\sqrt{s^H R_v^{-1} s}},$$

giving us our normalized filter,

$$h = \frac{1}{\sqrt{s^H R_v^{-1} s}} R_v^{-1} s.$$

If we care to write the impulse response of the filter for the convolution system, it is simply the complex conjugate time reversal of *h*.

Though we have derived the matched filter in discrete time, we can extend the concept to continuous-time systems if we replace R_v with the continuous-time autocorrelation function of the noise, assuming a continuous signal *s*(*t*), continuous noise *v*(*t*), and a continuous filter *h*(*t*).

Alternative derivation of the matched filter

Alternatively, we may solve for the matched filter by solving our maximization problem with a Lagrangian. Again, the matched filter endeavors to maximize the output signal-to-noise ratio (*SNR*) of a filtered deterministic signal in stochastic additive noise. The observed sequence, again, is

$$x = s + v,$$

with the noise covariance matrix,

$$R_v = E\{vv^H\}.$$

The signal-to-noise ratio is

$$SNR = \frac{|y_s|^2}{E\{|y_v|^2\}}.$$

Evaluating the expression in the numerator, we have

$$|y_s|^2 = {y_s}^H y_s = h^H s s^H h.$$

and in the denominator,

$$E\{|y_v|^2\} = E\{{y_v}^H y_v\} = E\{h^H v v^H h\} = h^H R_v h.$$

The signal-to-noise ratio becomes

$$SNR = \frac{h^H s s^H h}{h^H R_v h}.$$

If we now constrain the denominator to be 1, the problem of maximizing *SNR* is reduced to maximizing the numerator. We can then formulate the problem using a Lagrange multiplier:

$$\begin{aligned}
&h^H R_v h = 1 \\
&\mathcal{L} = h^H s s^H h + \lambda(1 - h^H R_v h) \\
&\nabla_{h^*}\mathcal{L} = s s^H h - \lambda R_v h = 0 \\
&(s s^H) h = \lambda R_v h
\end{aligned}$$

which we recognize as an eigenvalue problem

$$h^H (s s^H) h = \lambda h^H R_v h = \lambda.$$

Since ss^H is of unit rank, it has only one nonzero eigenvalue. It can be shown that this eigenvalue equals

$$\lambda_{\max} = s^H R_v^{-1} s,$$

yielding the following optimal matched filter

$$h = \frac{1}{\sqrt{s^H R_v^{-1} s}} R_v^{-1} s.$$

This is the same result found in the previous section.

Frequency-domain interpretation

When viewed in the frequency domain, it is evident that the matched filter applies the greatest weighting to spectral components that have the greatest signal-to-noise ratio. Although in general this requires a non-flat frequency response, the associated distortion is not significant in situations such as radar and digital communications, where the original waveform is known and the objective is to detect the presence of this signal against the background noise.

Example of matched filter in radar and sonar

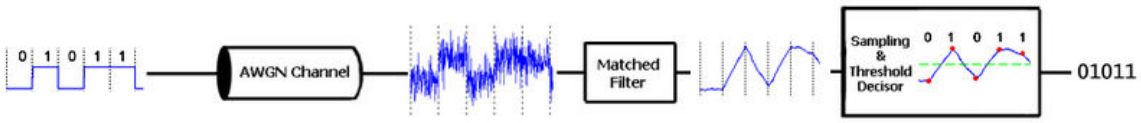
Matched filters are often used in signal detection. As an example, suppose that we wish to judge the distance of an object by reflecting a signal off it. We may choose to transmit a pure-tone sinusoid at 1 Hz. We assume that our received signal is an attenuated and phase-shifted form of the transmitted signal with added noise.

To judge the distance of the object, we correlate the received signal with a matched filter, which, in the case of white (uncorrelated) noise, is another pure-tone 1-Hz sinusoid. When the output of the matched filter system exceeds a certain threshold, we conclude with high probability that the received signal has been reflected off the object. Using the speed of propagation and the time that we first observe the reflected signal, we can estimate the distance of the object. If we change the shape of the pulse in a specially-designed way, the signal-to-noise ratio and the distance resolution can be even improved after matched filtering: this is a technique known as pulse compression.

Additionally, matched filters can be used in parameter estimation problems. To return to our previous example, we may desire to estimate the speed of the object, in addition to its position. To exploit the Doppler effect, we would like to estimate the frequency of the received signal. To do so, we may correlate the received signal with several matched filters of sinusoids at varying frequencies. The matched filter with the highest output will reveal, with high probability, the frequency of the reflected signal and help us determine the speed of the object. This method is, in fact, a simple version of the discrete Fourier transform (DFT). The DFT takes an N-valued complex input and correlates it with N matched filters, corresponding to complex exponentials at N different frequencies, to yield N complex-valued numbers corresponding to the relative amplitudes and phases of the sinusoidal components.

Example of matched filter in digital communications

The matched filter is also used in communications. In the context of a communication system that sends binary messages from the transmitter to the receiver across a noisy channel, a matched filter can be used to detect the transmitted pulses in the noisy received signal.



Imagine we want to send the sequence "0101100100" coded in non polar Non-return-to-zero (NRZ) through a certain channel.

Mathematically, a sequence in NRZ code can be described as a sequence of unit pulses or shifted rect functions, each pulse being weighted by +1 if the bit is "1" and by 0 if the bit is "0". Formally, the scaling factor for the k^{th} bit is,

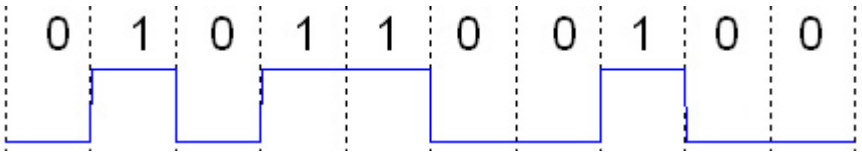
$$a_k = \begin{cases} 1, & \text{if bit } k \text{ is } 1, \\ 0, & \text{if bit } k \text{ is } 0. \end{cases}$$

We can represent our message, $M(t)$, as the sum of shifted unit pulses:

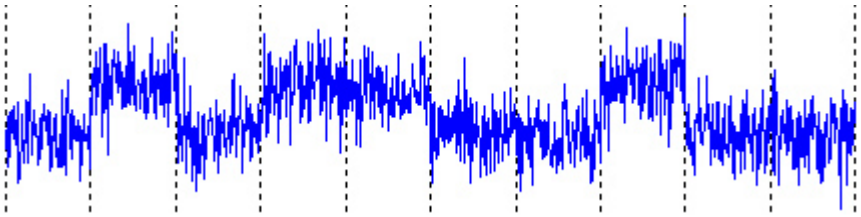
$$M(t) = \sum_{k=-\infty}^{\infty} a_k \times \Pi\left(\frac{t - kT}{T}\right).$$

where T is the time length of one bit.

Thus, the signal to be sent by the transmitter is



If we model our noisy channel as an AWGN channel, white Gaussian noise is added to the signal. At the receiver end, for a Signal-to-noise ratio of 3dB, this may look like:



A first glance will not reveal the original transmitted sequence. There is a high power of noise relative to the power of the desired signal (i.e., there is a low signal-to-noise ratio). If the receiver were to sample this signal at the correct moments, the resulting binary message would possibly belie the original transmitted one.

To increase our signal-to-noise ratio, we pass the received signal through a matched filter. In this case, the filter should be matched to an NRZ pulse (equivalent to a "1" coded in NRZ code). Precisely, the impulse response of the ideal matched filter, assuming white (uncorrelated) noise should be a time-reversed complex-conjugated scaled version of the signal that we are seeking. We choose

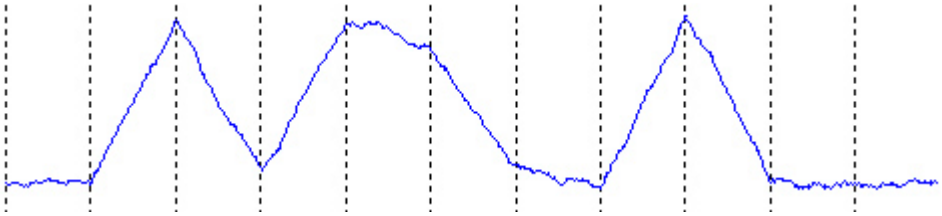
$$h(t) = \Pi\left(\frac{t}{T}\right).$$

In this case, due to symmetry, the time-reversed complex conjugate of $h(t)$ is in fact $h(t)$, allowing us to call $h(t)$ the impulse response of our matched filter convolution system.

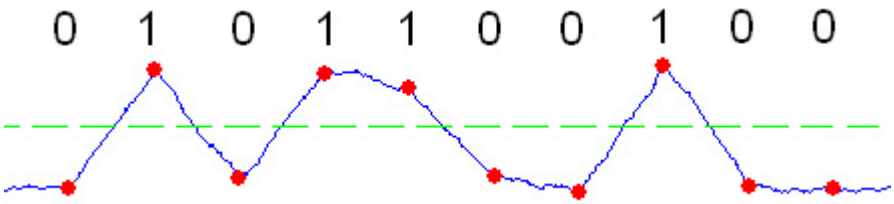
After convolving with the correct matched filter, the resulting signal, $M_{\text{filtered}}(t)$ is,

$$M_{\text{filtered}}(t) = M(t) * h(t)$$

where * denotes convolution.



Which can now be safely sampled by the receiver at the correct sampling instants, and compared to an appropriate threshold, resulting in a correct interpretation of the binary message.



Chapter-11

Estimation Theory

Estimation theory is a branch of statistics and signal processing that deals with estimating the values of parameters based on measured/empirical data that has a random component. The parameters describe an underlying physical setting in such a way that the value of the parameters affects the distribution of the measured data. An estimator attempts to approximate the unknown parameters using the measurements.

For example, it is desired to estimate the proportion of a population of voters who will vote for a particular candidate. That proportion is the unobservable parameter; the estimate is based on a small random sample of voters.

Or, for example, in radar the goal is to estimate the range of objects (airplanes, boats, etc.) by analyzing the two-way transit timing of received echoes of transmitted pulses. Since the reflected pulses are unavoidably embedded in electrical noise, their measured values are randomly distributed, so that the transit time must be estimated.

In estimation theory, it is assumed the measured data is random with probability distribution dependent on the parameters of interest. For example, in electrical communication theory, the measurements which contain information regarding the parameters of interest are often associated with a noisy signal. Without randomness, or noise, the problem would be deterministic and estimation would not be needed.

Estimation process

The entire purpose of estimation theory is to arrive at an estimator, and preferably an implementable one that could actually be used. The estimator takes the measured data as input and produces an estimate of the parameters.

It is also preferable to derive an estimator that exhibits optimality. Estimator optimality usually refers to achieving minimum average error over some class of estimators, for example, a minimum variance unbiased estimator. In this case, the class is the set of unbiased estimators, and the average error measure is variance (average squared error

between the value of the estimate and the parameter). However, optimal estimators do not always exist.

These are the general steps to arrive at an estimator:

- In order to arrive at a desired estimator, it is first necessary to determine a probability distribution for the measured data, and the distribution's dependence on the unknown parameters of interest. Often, the probability distribution may be derived from physical models that explicitly show how the measured data depends on the parameters to be estimated, and how the data is corrupted by random errors or noise. In other cases, the probability distribution for the measured data is simply "assumed", for example, based on familiarity with the measured data and/or for analytical convenience.
- After deciding upon a probabilistic model, it is helpful to find the limitations placed upon an estimator. This limitation, for example, can be found through the Cramér–Rao bound.
- Next, an estimator needs to be developed or applied if an already known estimator is valid for the model. The estimator needs to be tested against the limitations to determine if it is an optimal estimator (if so, then no other estimator will perform better).
- Finally, experiments or simulations can be run using the estimator to test its performance.

After arriving at an estimator, real data might show that the model used to derive the estimator is incorrect, which may require repeating these steps to find a new estimator. A non-implementable or infeasible estimator may need to be scrapped and the process started anew.

In summary, the estimator estimates the parameters of a physical model based on measured data.

Teori Estimasi

Basics

To build a model, several statistical "ingredients" need to be known. These are needed to ensure the estimator has some mathematical tractability instead of being based on "good feel".

The first is a set of statistical samples taken from a random vector (RV) of size *N*. Put into a vector,

$$\mathbf{x} = \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}.$$

Secondly, we have the corresponding *M* parameters

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_M \end{bmatrix},$$

which need to be established with their probability density function (pdf) or probability mass function (pmf)

$$p(\mathbf{x}|\theta).$$

It is also possible for the parameters themselves to have a probability distribution (e.g., Bayesian statistics). It is then necessary to define the Bayesian probability

$$\pi(\theta).$$

After the model is formed, the goal is to estimate the parameters, commonly denoted $\hat{\theta}$, where the "hat" indicates the estimate.

One common estimator is the minimum mean squared error (MMSE) estimator, which utilizes the error between the estimated parameters and the actual value of the parameters

$$\mathbf{e} = \hat{\theta} - \theta$$

as the basis for optimality. This error term is then squared and minimized for the MMSE estimator.

Estimators

Commonly-used estimators and estimation methods, and topics related to them:

- Maximum likelihood estimators
- Bayes estimators
- Method of moments estimators
- Cramér–Rao bound

- Minimum mean squared error (MMSE), also known as Bayes least squared error (BLSE)
- Maximum a posteriori (MAP)
- Minimum variance unbiased estimator (MVUE)
- Best linear unbiased estimator (BLUE)
- Unbiased estimators.
- Particle filter
- Markov chain Monte Carlo (MCMC)
- Kalman filter
- Ensemble Kalman filter (EnKF)
- Wiener filter

Examples

Unknown constant in additive white Gaussian noise

Consider a received discrete signal, $x[n]$, of N independent samples that consists of an unknown constant A with additive white Gaussian noise (AWGN) $w[n]$ with known variance σ^2 (*i.e.*, $\mathcal{N}(0, \sigma^2)$). Since the variance is known then the only unknown parameter is A.

The model for the signal is then

$$x[n] = A + w[n] \quad n = 0, 1, \ldots, N-1$$

Two possible (of many) estimators are:

- $\hat{A}_1 = x[0]$
- $\hat{A}_2 = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$ which is the sample mean

Both of these estimators have a mean of A, which can be shown through taking the expected value of each estimator

$$\mathrm{E}\left[\hat{A}_1\right] = \mathrm{E}\left[x[0]\right] = A$$

and

$$\mathrm{E}\left[\hat{A}_2\right] = \mathrm{E}\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = \frac{1}{N}\left[\sum_{n=0}^{N-1} \mathrm{E}\left[x[n]\right]\right] = \frac{1}{N}\left[NA\right] = A$$

At this point, these two estimators would appear to perform the same. However, the difference between them becomes apparent when comparing the variances.

$$\operatorname{var}\left(\hat{A}_1\right) = \operatorname{var}\left(x[0]\right) = \sigma^2$$

and

$$\operatorname{var}\left(\hat{A}_2\right) = \operatorname{var}\left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right) \overset{independence}{=} \frac{1}{N^2}\left[\sum_{n=0}^{N-1} \operatorname{var}(x[n])\right] = \frac{1}{N^2}\left[N\sigma^2\right] = \frac{\sigma^2}{N}$$

It would seem that the sample mean is a better estimator since, as $N \to \infty$, the variance goes to zero.

Maximum likelihood

Continuing the example using the maximum likelihood estimator, the probability density function (pdf) of the noise for one sample *w*[*n*] is

$$p(w[n]) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} w[n]^2\right)$$

and the probability of *x*[*n*] becomes (*x*[*n*] can be thought of a $\mathcal{N}(A, \sigma^2)$)

$$p(x[n]; A) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x[n] - A)^2\right)$$

By independence, the probability of $\mathbf{x}$becomes

$$p(\mathbf{x}; A) = \prod_{n=0}^{N-1} p(x[n]; A) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^N} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2\right)$$

Taking the natural logarithm of the pdf

$$\ln p(\mathbf{x}; A) = -N \ln\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2$$

and the maximum likelihood estimator is

$$\hat{A} = \arg\max \ln p(\mathbf{x}; A)$$

Taking the first derivative of the log-likelihood function

$$\frac{\partial}{\partial A}\ln p(\mathbf{x};A) = \frac{1}{\sigma^2}\left[\sum_{n=0}^{N-1}(x[n]-A)\right] = \frac{1}{\sigma^2}\left[\sum_{n=0}^{N-1}x[n]-NA\right]$$

and setting it to zero

$$0 = \frac{1}{\sigma^2}\left[\sum_{n=0}^{N-1}x[n]-NA\right] = \sum_{n=0}^{N-1}x[n]-NA$$

This results in the maximum likelihood estimator

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1}x[n]$$

which is simply the sample mean. From this example, it was found that the sample mean is the maximum likelihood estimator for *N* samples of a fixed, unknown parameter corrupted by AWGN.

Cramér–Rao lower bound

To find the Cramér–Rao lower bound (CRLB) of the sample mean estimator, it is first necessary to find the Fisher information number

$$\mathcal{I}(A) = \mathrm{E}\left(\left[\frac{\partial}{\partial\theta}\ln p(\mathbf{x};A)\right]^2\right) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln p(\mathbf{x};A)\right]$$

and copying from above

$$\frac{\partial}{\partial A}\ln p(\mathbf{x};A) = \frac{1}{\sigma^2}\left[\sum_{n=0}^{N-1}x[n]-NA\right]$$

Taking the second derivative

$$\frac{\partial^2}{\partial A^2}\ln p(\mathbf{x};A) = \frac{1}{\sigma^2}(-N) = \frac{-N}{\sigma^2}$$

and finding the negative expected value is trivial since it is now a deterministic constant

$$-\mathrm{E}\left[\frac{\partial^2}{\partial A^2}\ln p(\mathbf{x};A)\right] = \frac{N}{\sigma^2}$$

Finally, putting the Fisher information into

$$\operatorname{var}\left(\hat{A}\right) \geq \frac{1}{\mathcal{I}}$$

results in

$$\operatorname{var}\left(\hat{A}\right) \geq \frac{\sigma^2}{N}$$

Comparing this to the variance of the sample mean (determined previously) shows that the sample mean is *equal to* the Cramér–Rao lower bound for all values of *N* and *A*. In other words, the sample mean is the (necessarily unique) efficient estimator, and thus also the minimum variance unbiased estimator (MVUE), in addition to being the maximum likelihood estimator.

Maximum of a uniform distribution

One of the simplest non-trivial examples of estimation is the estimation of the maximum of a uniform distribution. It is used as a hands-on classroom exercise and to illustrate basic principles of estimation theory. Further, in the case of estimation based on a single sample, it demonstrates philosophical issues and possible misunderstandings in the use of maximum likelihood estimators and likelihood functions.

Given a discrete uniform distribution $1, 2, \ldots, N$ with unknown maximum, the UMVU estimator for the maximum is given by

$$\frac{k+1}{k}m - 1 = m + \frac{m}{k} - 1$$

where *m* is the sample maximum and *k* is the sample size, sampling without replacement. This problem is commonly known as the German tank problem, due to application of maximum estimation to estimates of German tank production during World War II.

The formula may be understood intuitively as:

> "The sample maximum plus the average gap between observations in the sample",

the gap being added to compensate for the negative bias of the sample maximum as an estimator for the population maximum.

This has a variance of

$$\frac{1}{k}\frac{(N-k)(N+1)}{(k+2)} \approx \frac{N^2}{k^2} \text{ for small samples } k \ll N$$

so a standard deviation of approximately N / k, the (population) average size of a gap between samples; compare $\frac{m}{k}$ above. This can be seen as a very simple case of maximum spacing estimation.

The sample maximum is the maximum likelihood estimator for the population maximum, but, as discussed above, it is biased.

Applications

Numerous fields require the use of estimation theory. Some of these fields include (but are by no means limited to):

- Interpretation of scientific experiments
- Signal processing
- Clinical trials
- Opinion polls
- Quality control
- Telecommunications
- Project management
- Software engineering
- Control theory
- Network intrusion detection system
- Orbit determination

Measured data are likely to be subject to noise or uncertainty and it is through statistical probability that optimal solutions are sought to extract as much information from the data as possible.

Chapter-12

Bayes Estimator

In estimation theory and decision theory, a **Bayes estimator** or a **Bayes action** is an estimator or decision rule that minimizes the posterior expected value of a loss function (i.e., the **posterior expected loss**). Equivalently, it maximizes the posterior expectation of a utility function. An alternative way of formulating an estimator within Bayesian statistics is Maximum a posteriori estimation.

Definition

Suppose an unknown parameter θ is known to have a prior distribution π. Let δ = δ(x) be an estimator of θ (based on some measurements x), and let L(θ,δ) be a loss function, such as squared error. The **Bayes risk** of δ is defined as $E_\pi\{L(\theta,\delta)\}$, where the expectation is taken over the probability distribution of θ: this defines the risk function as a function of δ. An estimator δ is said to be a *Bayes estimator* if it minimizes the Bayes risk among all estimators. Equivalently, the estimator which minimizes the posterior expected loss $E\{L(\theta,\delta) \mid x\}$ *for each* x also minimizes the Bayes risk and therefore is a Bayes estimator.

If the prior is improper then an estimator which minimizes the posterior expected loss *for each* x is called a **generalized Bayes estimator**.

Examples

Minimum mean square error estimation

The most common risk function used for Bayesian estimation is the mean square error (MSE), also called squared error risk. The MSE is defined by

$$\mathrm{MSE} = E\left[(\widehat{\theta}(x) - \theta)^2\right],$$

where the expectation is taken over the joint distribution of θ and x.

Using the MSE as risk, the Bayes estimate of the unknown parameter is simply the mean of the posterior distribution,

$$\widehat{\theta}(x) = E[\theta|x] = \int \theta f(\theta|x)\, d\theta.$$

This is known as the *minimum mean square error* (MMSE) estimator. The Bayes risk, in this case, is the posterior variance.

Bayes estimators for conjugate priors

If there is no inherent reason to prefer one prior probability distribution over another, a conjugate prior is sometimes chosen for simplicity. A conjugate prior is defined as a prior distribution belonging to some parametric family, for which the resulting posterior distribution also belongs to the same family. This is an important property, since the Bayes estimator, as well as its statistical properties (variance, confidence interval, etc.), can all be derived from the posterior distribution.

Conjugate priors are especially useful for sequential estimation, where the posterior of the current measurement is used as the prior in the next measurement. In sequential estimation, unless a conjugate prior is used, the posterior distribution typically becomes more complex with each added measurement, and the Bayes estimator cannot usually be calculated without resorting to numerical methods.

Following are some examples of conjugate priors.

- If $x|\theta$ is normal, $x|\theta \sim N(\theta,\sigma^2)$, and the prior is normal, $\theta \sim N(\mu,\tau^2)$, then the posterior is also normal and the Bayes estimator under MSE is given by

$$\widehat{\theta}(x) = \frac{\sigma^2}{\sigma^2 + \tau^2}\mu + \frac{\tau^2}{\sigma^2 + \tau^2}x.$$

- If $x_1,...,x_n$ are iid Poisson random variables $x_i|\theta \sim P(\theta)$, and if the prior is Gamma distributed $\theta \sim G(a,b)$, then the posterior is also Gamma distributed, and the Bayes estimator under MSE is given by

$$\widehat{\theta}(X) = \frac{n\overline{X} + a}{n + \frac{1}{b}}.$$

- If $x_1,...,x_n$ are iid uniformly distributed $x_i|\theta \sim U(0,\theta)$, and if the prior is Pareto distributed $\theta \sim Pa(\theta_0,a)$, then the posterior is also Pareto distributed, and the Bayes estimator under MSE is given by

$$\widehat{\theta}(X) = \frac{(a+n)\max(\theta_0, x_1, ..., x_n)}{a+n-1}.$$

Alternative risk functions

Risk functions are chosen depending on how one measures the distance between the estimate and the unknown parameter. The MSE is the most common risk function in use, primarily due to its simplicity. However, alternative risk functions are also occasionally used. The following are several examples of such alternatives. We denote the posterior generalized distribution function by *F*.

- A "linear" loss function, with $a > 0$, which yields the posterior median as the Bayes' estimate:

$$L(\theta, \widehat{\theta}) = a|\theta - \widehat{\theta}|$$
$$F(\widehat{\theta}(x)|X) = \tfrac{1}{2}.$$

- Another "linear" loss function, which assigns different "weights" $a,b > 0$ to over or sub estimation. It yields a quantile from the posterior distribution, and is a generalization of the previous loss function:

$$L(\theta, \widehat{\theta}) = \begin{cases} a|\theta - \widehat{\theta}|, & \text{for } \theta - \widehat{\theta} \geq 0 \\ b|\theta - \widehat{\theta}|, & \text{for } \theta - \widehat{\theta} < 0 \end{cases}$$
$$F(\widehat{\theta}(x)|X) = \frac{a}{a+b}.$$

- The following loss function is trickier: it yields either the posterior mode, or a point close to it depending on the curvature and properties of the posterior distribution. Small values of the parameter $K > 0$ are recommended, in order to use the mode as an approximation ($L > 0$):

$$L(\theta, \widehat{\theta}) = \begin{cases} 0, & \text{for } |\theta - \widehat{\theta}| < K \\ L, & \text{for } |\theta - \widehat{\theta}| \geq K. \end{cases}$$

Other loss functions can be conceived, although the mean squared error is the most widely used and validated.

Generalized Bayes estimators

The prior distribution π has thus far been assumed to be a true probability distribution, in that

$$\int \pi(\theta) d\theta = 1.$$

However, occasionally this can be a restrictive requirement. For example, there is no distribution (covering the set, **R**, of all real numbers) for which every real number is equally likely. Yet, in some sense, such a "distribution" seems like a natural choice for a non-informative prior, i.e., a prior distribution which does not imply a preference for any particular value of the unknown parameter. One can still define a function $\pi(\theta) = 1$, but this would not be a proper probability distribution since it has infinite mass,

$$\int \pi(\theta)d\theta = \infty.$$

Such measures $\pi(\theta)$, which are not probability distributions, are referred to as improper priors.

The use of an improper prior means that the Bayes risk is undefined (since the prior is not a probability distribution and we cannot take an expectation under it). As a consequence, it is no longer meaningful to speak of a Bayes estimator that minimizes the Bayes risk. Nevertheless, in many cases, one can define the posterior distribution

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)d\theta}.$$

This is a definition, and not an application of Bayes' theorem, since Bayes' theorem can only be applied when all distributions are proper. However, it is not uncommon for the resulting "posterior" to be a valid probability distribution. In this case, the posterior expected loss

$$\int L(\theta, a)\pi(\theta|x)d\theta$$

is typically well-defined and finite. Recall that, for a proper prior, the Bayes estimator minimizes the posterior expected loss. When the prior is improper, an estimator which minimizes the posterior expected loss is referred to as a **generalized Bayes estimator**.

Example

A typical example concerns the estimation of a location parameter with a loss function of the type $L(a - \theta)$. Here θ is a location parameter, i.e., $p(x \mid \theta) = f(x - \theta)$.

It is common to use the improper prior $\pi(\theta) = 1$ in this case, especially when no other more subjective information is available. This yields

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} = \frac{f(x-\theta)}{p(x)}$$

so the posterior expected loss equals

$$E[L(a-\theta)] = \int L(a-\theta)\pi(\theta|x)d\theta = \frac{1}{p(x)}\int L(a-\theta)f(x-\theta)d\theta.$$

The generalized Bayes estimator is the value $a(x)$ which minimizes this expression for all x. This is equivalent to minimizing

$$\int L(a-\theta)f(x-\theta)d\theta \quad \text{for all } x. \qquad (1)$$

It can be shown that, in this case, the generalized Bayes estimator has the form $x + a_0$, for some constant a_0. To see this, let a_0 be the value minimizing (1) when $x = 0$. Then, given a different value x_1, we must minimize

$$\int L(a-\theta)f(x_1-\theta)d\theta = \int L(a-x_1-\theta')f(-\theta')d\theta'. \qquad (2)$$

This is identical to (1), except that a has been replaced by $a - x_1$. Thus, the expression minimizing is given by $a - x_1 = a_0$, so that the optimal estimator has the form

$$a(x) = a_0 + x.$$

Empirical Bayes estimators

A Bayes estimator derived through the empirical Bayes method is called an *empirical Bayes estimator*. Empirical Bayes methods enable the use of auxiliary empirical data, from observations of related parameters, in the development of a Bayes estimator. This is done under the assumption that the estimated parameters are obtained from a common prior. For example, if independent observations of different parameters are performed, then the estimation performance of a particular parameter can sometimes be improved by using data from other observations.

There are parametric and non-parametric approaches to empirical Bayes estimation. Parametric empirical Bayes is usually preferable since it is more applicable and more accurate on small amounts of data.

Example

The following is a simple example of parametric empirical Bayes estimation. Given past observations $x_1, \ldots, x_n$ having conditional distribution $f(x_i \mid \theta_i)$, one is interested in estimating θ_{n+1} based on x_{n+1}. Assume that the θ_i's have a common prior π which depends on unknown parameters. For example, suppose that π is normal with unknown mean μ_π and variance σ_π. We can then use the past observations to determine the mean and variance of π in the following way.

First, we estimate the mean μ_m and variance σ_m of the marginal distribution of $x_1, \ldots, x_n$ using the maximum likelihood approach:

$$\begin{aligned}\widehat{\mu}_m &= \frac{1}{n}\sum x_i, \\ \widehat{\sigma}_m^2 &= \frac{1}{n}\sum (x_i - \widehat{\mu}_m)^2.\end{aligned}$$

Next, we use the relation

$$\begin{aligned}\mu_m &= E_\pi[\mu_f(\theta)], \\ \sigma_m^2 &= E_\pi[\sigma_f^2(\theta)] + E_\pi[\mu_f(\theta) - \mu_m],\end{aligned}$$

where $\mu_f(\theta)$ and $\sigma_f(\theta)$ are the moments of the conditional distribution $f(x_i \mid \theta_i)$, which are assumed to be known. In particular, suppose that $\mu_f(\theta) = \theta$ and that $\sigma_f^2(\theta) = K$; we then have

$$\begin{aligned}\mu_\pi &= \mu_m, \\ \sigma_\pi^2 &= \sigma_m^2 - \sigma_f^2 = \sigma_m^2 - K.\end{aligned}$$

Finally, we obtain the estimated moments of the prior,

$$\begin{aligned}\widehat{\mu}_\pi &= \widehat{\mu}_m, \\ \widehat{\sigma}_\pi^2 &= \widehat{\sigma}_m^2 - K.\end{aligned}$$

For example, if $x_i \mid \theta_i \tilde{} N(\theta_i,1)$, and if we assume a normal prior (which is a conjugate prior in this case), we conclude that $\theta_{n+1} \sim N(\widehat{\mu}_\pi, \widehat{\sigma}_\pi^2)$, from which the Bayes estimator of θ_{n+1} based on x_{n+1} can be calculated.

Properties

Admissibility

Bayes rules having finite Bayes risk are typically admissible. The following are some specific examples of admissibility theorems.

- If a Bayes rule is unique then it is admissible. For example, as stated above, under mean squared error (MSE) the Bayes rule is unique and therefore admissible.
- If θ belongs to a discrete set, then all Bayes rules are admissible.
- If θ belongs to a continuous (non-discrete set), and if the risk function $R(\theta,\delta)$ is continuous in θ for every δ, then all Bayes rules are admissible.

By contrast, generalized Bayes rules often have undefined Bayes risk in the case of improper priors. These rules are often inadmissible and the verification of their admissibility can be difficult. For example, the generalized Bayes estimator of a location parameter θ based on Gaussian samples (described in the "Generalized Bayes estimator" section above) is inadmissible for $p > 2$; this is known as Stein's phenomenon.

Asymptotic efficiency

Let θ be an unknown random variable, and suppose that $x_1, x_2, \cdots$ are iid samples with density $f(x_i \mid \theta)$. Let $\delta_n = \delta_n(x_1, \ldots, x_n)$ be a sequence of Bayes estimators of θ based on an increasing number of measurements. We are interested in analyzing the asymptotic performance of this sequence of estimators, i.e., the performance of δ_n for large n.

To this end, it is customary to regard θ as a deterministic parameter whose true value is θ_0. Under specific conditions, for large samples (large values of n), the posterior density of θ is approximately normal. In other words, for large n, the effect of the prior probability on the posterior is negligible. Moreover, if δ is the Bayes estimator under MSE risk, then it is asymptotically unbiased and it converges in distribution to the normal distribution:

$$\sqrt{n}(\delta_n - \theta_0) \to N\left(0, \frac{1}{I(\theta_0)}\right),$$

where $I(\theta_0)$ is the fisher information of θ_0. It follows that the Bayes estimator δ_n under MSE is asymptotically efficient.

Another estimator which is asymptotically normal and efficient is the maximum likelihood estimator (MLE). The relations between the maximum likelihood and Bayes estimators can be shown in the following simple example.

Consider the estimator of θ based on binomial sample $x \sim b(\theta, n)$ where θ denotes the probability for success. Assuming θ is distributed according to the conjugate prior, which in this case is the Beta distribution $B(a,b)$, the posterior distribution is known to be B(a+x,b+n-x). Thus, the Bayes estimator under MSE is

$$\delta_n(x) = E[\theta|x] = \frac{a+x}{a+b+n}.$$

The MLE in this case is x/n and so we get,

$$\delta_n(x) = \frac{a+b}{a+b+n}E[\theta] + \frac{n}{a+b+n}\delta_{MLE}.$$

The last equation implies that, for $n \to \infty$, the Bayes estimator (in the described problem) is close to the MLE. On the other hand, when n is small, the prior information is still relevant to the decision problem and affects the estimate.

Practical example

The Internet Movie Database uses a formula for calculating the Top Rated 250 Titles which gives a true Bayesian estimate:

$$W = \frac{Rv + Cm}{v + m}$$

where:

W= weighted rating
R= average for the movie as a number from 0 to 10 (mean) = (Rating)
v= number of votes for the movie = (votes)
m= minimum votes required to be listed in the Top 250 (currently 3000)
C= the mean vote across the whole report (currently 6.9)

for the Top 250, only votes from regular voters are considered.

Chapter-13

Cramer–Rao Bound

In estimation theory and statistics, the **Cramér–Rao bound (CRB)** or **Cramér–Rao lower bound (CRLB)**, named in honor of Harald Cramér and Calyampudi Radhakrishna Rao who were among the first to derive it, expresses a lower bound on the variance of estimators of a deterministic parameter. The bound is also known as the **Cramér–Rao inequality** or the **information inequality**.

In its simplest form, the bound states that the variance of any unbiased estimator is at least as high as the inverse of the Fisher information. An unbiased estimator which achieves this lower bound is said to be efficient. Such a solution achieves the lowest possible mean squared error among all unbiased methods, and is therefore the minimum variance unbiased (MVU) estimator. However, in some cases, no unbiased technique exists which achieves the bound. This may occur even when an MVU estimator exists.

The Cramér–Rao bound can also be used to bound the variance of *biased* estimators of given bias. In some cases, a biased approach can result in both a variance and a mean squared error that are *below* the unbiased Cramér–Rao lower bound.

Statement

The Cramér–Rao bound is stated in this section for several increasingly general cases, beginning with the case in which the parameter is a scalar and its estimator is unbiased. All versions of the bound require certain regularity conditions, which hold for most well-behaved distributions.

Scalar unbiased case

Suppose θ is an unknown deterministic parameter which is to be estimated from measurements x, distributed according to some probability density function $f(x;\theta)$. The variance of any *unbiased* estimator $\hat{\theta}$ of θ is then bounded by the inverse of the Fisher information $I(\theta)$:

$$\operatorname{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

where the Fisher information $I(\theta)$ is defined by

$$I(\theta) = \mathrm{E}\left[\left(\frac{\partial \ell(x;\theta)}{\partial \theta}\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2 \ell(x;\theta)}{\partial \theta^2}\right]$$

and $\ell(x;\theta) = \log f(x;\theta)$ is the natural logarithm of the likelihood function and E denotes the expected value.

The efficiency of an unbiased estimator $\hat{\theta}$ measures how close this estimator's variance comes to this lower bound; estimator efficiency is defined as

$$e(\hat{\theta}) = \frac{I(\theta)^{-1}}{\operatorname{var}(\hat{\theta})}$$

or the minimum possible variance for an unbiased estimator divided by its actual variance. The Cramér–Rao lower bound thus gives

$$e(\hat{\theta}) \leq 1.$$

General scalar case

A more general form of the bound can be obtained by considering an unbiased estimator $T(X)$ of a function $\psi(\theta)$ of the parameter θ. Here, unbiasedness is understood as stating that $E\{T(X)\} = \psi(\theta)$. In this case, the bound is given by

$$\operatorname{var}(T) \geq \frac{[\psi'(\theta)]^2}{I(\theta)}$$

where $\psi'(\theta)$ is the derivative of $\psi(\theta)$, and $I(\theta)$ is the Fisher information defined above.

Apart from being a bound on estimators of functions of the parameter, this approach can be used to derive a bound on the variance of biased estimators with a given bias, as follows. Consider an estimator $\hat{\theta}$ with bias $b(\theta) = E\{\hat{\theta}\} - \theta$, and let $\psi(\theta) = b(\theta) + \theta$. By the result above, any unbiased estimator whose expectation is $\psi(\theta)$ has variance greater than or equal to $(\psi'(\theta))^2 / I(\theta)$. Thus, any estimator $\hat{\theta}$ whose bias is given by a function $b(\theta)$ satisfies

$$\operatorname{var}\left(\hat{\theta}\right) \geq \frac{[1+b'(\theta)]^2}{I(\theta)}.$$

Clearly, the unbiased version of the bound is a special case of this result, with $b(\theta) = 0$.

Of course, it's trivial to have a small variance − an "estimator" that is constant has a variance of zero. But from the above equation we find that the mean squared error of a biased estimator is bounded by

$$\mathrm{E}\left((\hat{\theta}-\theta)^2\right) \geq \frac{[1+b'(\theta)]^2}{I(\theta)} + b(\theta)^2,$$

and this can be less than the unbiased Cramér-Rao bound $1/I(\theta)$.

Multivariate case

Extending the Cramér–Rao bound to multiple parameters, define a parameter column vector

$$\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_d]^T \in \mathbb{R}^d$$

with probability density function $f(x;\boldsymbol{\theta})$ which satisfies the two regularity conditions below.

The Fisher information matrix is a $d \times d$ matrix with element $I_{m,k}$ defined as

$$I_{m,k} = \mathrm{E}\left[\frac{d}{d\theta_m}\log f\left(x;\boldsymbol{\theta}\right)\frac{d}{d\theta_k}\log f\left(x;\boldsymbol{\theta}\right)\right].$$

Let $\boldsymbol{T}(X)$ be an estimator of any vector function of parameters, $\boldsymbol{T}(X) = (T_1(X), \dots, T_n(X))^T$, and denote its expectation vector $\mathrm{E}[\boldsymbol{T}(X)]$ by $\boldsymbol{\psi}(\boldsymbol{\theta})$. The Cramér–Rao bound then states that the covariance matrix of $\boldsymbol{T}(X)$ satisfies

$$\operatorname{cov}_{\boldsymbol{\theta}}\left(\boldsymbol{T}(X)\right) \geq \frac{\partial \boldsymbol{\psi}\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\left[I\left(\boldsymbol{\theta}\right)\right]^{-1}\left(\frac{\partial \boldsymbol{\psi}\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right)^T$$

where

- The matrix inequality $A \geq B$ is understood to mean that the matrix $A - B$ is positive semidefinite, and
- $\partial\boldsymbol{\psi}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is a matrix whose ijth element is given by $\partial\psi_i(\boldsymbol{\theta})/\partial\theta_j$.

If $\boldsymbol{T}(X)$ is an unbiased estimator of $\boldsymbol{\theta}$ (i.e., $\boldsymbol{\psi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$), then the Cramér–Rao bound reduces to

$$\operatorname{cov}_{\boldsymbol{\theta}}(\boldsymbol{T}(X)) \geq I(\boldsymbol{\theta})^{-1}.$$

Regularity conditions

The bound relies on two weak regularity conditions on the probability density function, $f(x;\theta)$, and the estimator $T(X)$:

- The Fisher information is always defined; equivalently, for all x such that $f(x;\theta) > 0$,

  $$\frac{\partial}{\partial\theta} \ln f(x;\theta)$$

  exists, and is finite.

- The operations of integration with respect to x and differentiation with respect to θ can be interchanged in the expectation of T; that is,

  $$\frac{\partial}{\partial\theta}\left[\int T(x) f(x;\theta)\,dx\right] = \int T(x)\left[\frac{\partial}{\partial\theta} f(x;\theta)\right] dx$$

  whenever the right-hand side is finite.
  This condition can often be confirmed by using the fact that integration and differentiation can be swapped when either of the following cases hold:

  1. The function $f(x;\theta)$ has bounded support in x, and the bounds do not depend on θ;
  2. The function $f(x;\theta)$ has infinite support, is continuously differentiable, and the integral converges uniformly for all θ.

Simplified form of the Fisher information

Suppose, in addition, that the operations of integration and differentiation can be swapped for the second derivative of $f(x;\theta)$ as well, i.e.,

$$\frac{\partial^2}{\partial\theta^2}\left[\int T(x) f(x;\theta)\,dx\right] = \int T(x)\left[\frac{\partial^2}{\partial\theta^2} f(x;\theta)\right] dx.$$

In this case, it can be shown that the Fisher information equals

$$I(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right].$$

The Cramèr–Rao bound can then be written as

$$\operatorname{var}\left(\widehat{\theta}\right) \geq \frac{1}{I(\theta)} = \frac{1}{-\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]}.$$

In some cases, this formula gives a more convenient technique for evaluating the bound.

Single-parameter proof

The following is a proof of the general scalar case of the Cramér–Rao bound, which was described above; namely, that if the expectation of T is denoted by $\psi(\theta)$, then, for all θ,

$$\operatorname{var}(t(X)) \geq \frac{[\psi'(\theta)]^2}{I(\theta)}.$$

Let X be a random variable with probability density function $f(x;\theta)$. Here $T = t(X)$ is a statistic, which is used as an estimator for $\psi(\theta)$. If V is the score, i.e.

$$V = \frac{\partial}{\partial\theta}\ln f(X;\theta)$$

then the expectation of V, written $\mathrm{E}(V)$, is zero. If we consider the covariance $\operatorname{cov}(V,T)$ of V and T, we have $\operatorname{cov}(V,T) = \mathrm{E}(VT)$, because $\mathrm{E}(V) = 0$. Expanding this expression we have

$$\operatorname{cov}(V,T) = \mathrm{E}\left(T\cdot\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)$$

This may be expanded using the chain rule

$$\frac{\partial}{\partial\theta}\ln Q = \frac{1}{Q}\frac{\partial Q}{\partial\theta}$$

and the definition of expectation gives, after cancelling $f(x;\theta)$,

$$\mathrm{E}\left(T\cdot\frac{\partial}{\partial\theta}\ln f(X;\theta)\right) = \int t(x)\left[\frac{\partial}{\partial\theta}f(x;\theta)\right]dx = \frac{\partial}{\partial\theta}\left[\int t(x)f(x;\theta)\,dx\right] = \psi'(\theta)$$

because the integration and differentiation operations commute (second condition).

The Cauchy–Schwarz inequality shows that

$$\sqrt{\operatorname{var}(T)\operatorname{var}(V)} \geq |\operatorname{cov}(V,T)| = |\psi'(\theta)|$$

therefore

$$\operatorname{var}\, T \geq \frac{[\psi'(\theta)]^2}{\operatorname{var}(V)} = \frac{[\psi'(\theta)]^2}{I(\theta)} = \left[\frac{\partial}{\partial\theta}\mathrm{E}(T)\right]^2 \frac{1}{I(\theta)}$$

which proves the proposition.

Examples

Multivariate normal distribution

For the case of a *d*-variate normal distribution

$$\boldsymbol{x} \sim N_d\left(\boldsymbol{\mu}\left(\boldsymbol{\theta}\right), \boldsymbol{C}\left(\boldsymbol{\theta}\right)\right)$$

the Fisher information matrix has elements

$$I_{m,k} = \frac{\partial \boldsymbol{\mu}^T}{\partial \theta_m} \boldsymbol{C}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \theta_k} + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{C}}{\partial \theta_m}\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{C}}{\partial \theta_k}\right)$$

where "tr" is the trace.

For example, let *w*[*n*] be a sample of *N* independent observations) with unknown mean θ and known variance σ^2

$$w[n] \sim \mathbb{N}_N\left(\theta\mathbf{1}, \sigma^2\boldsymbol{I}\right).$$

Then the Fisher information is a scalar given by

$$I(\theta) = \left(\frac{\partial \boldsymbol{\mu}(\theta)}{\partial \theta}\right)^T \boldsymbol{C}^{-1}\left(\frac{\partial \boldsymbol{\mu}(\theta)}{\partial \theta}\right) = \sum_{i=1}^{N} \frac{1}{\sigma^2} = \frac{N}{\sigma^2},$$

and so the Cramér–Rao bound is

$$\operatorname{var}\left(\hat{\theta}\right) \geq \frac{\sigma^2}{N}.$$

Normal variance with known mean

Suppose X is a normally distributed random variable with known mean μ and unknown variance σ^2. Consider the following statistic:

$$T = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}.$$

Then T is unbiased for σ^2, as $E(T) = \sigma^2$. What is the variance of T?

$$\operatorname{var}(T) = \frac{\operatorname{var}(X-\mu)^2}{n} = \frac{1}{n}\left[E\left\{(X-\mu)^4\right\} - \left(E\left\{(X-\mu)^2\right\}\right)^2\right]$$

(the second equality follows directly from the definition of variance). The first term is the fourth moment about the mean and has value $3(\sigma^2)^2$; the second is the square of the variance, or $(\sigma^2)^2$. Thus

$$\operatorname{var}(T) = \frac{2(\sigma^2)^2}{n}.$$

Now, what is the Fisher information in the sample? Recall that the score V is defined as

$$V = \frac{\partial}{\partial \sigma^2} \log L(\sigma^2, X)$$

where L is the likelihood function. Thus in this case,

$$V = \frac{\partial}{\partial \sigma^2} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X-\mu)^2/2\sigma^2}\right] = \frac{(X-\mu)^2}{2(\sigma^2)^2} - \frac{1}{2\sigma^2}$$

where the second equality is from elementary calculus. Thus, the information in a single observation is just minus the expectation of the derivative of V, or

$$I = -E\left(\frac{\partial V}{\partial \sigma^2}\right) = -E\left(-\frac{(X-\mu)^2}{(\sigma^2)^3} + \frac{1}{2(\sigma^2)^2}\right) = \frac{\sigma^2}{(\sigma^2)^3} - \frac{1}{2(\sigma^2)^2} = \frac{1}{2(\sigma^2)^2}.$$

Thus the information in a sample of n independent observations is just n times this, or $\frac{n}{2(\sigma^2)^2}$.

The Cramer Rao bound states that

$$\operatorname{var}(T) \geq \frac{1}{I}.$$

In this case, the inequality is saturated (equality is achieved), showing that the estimator is efficient.

However, we can achieve a lower mean squared error using a biased estimator. The estimator

$$T = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n+2}.$$

obviously has a smaller variance, which is in fact

$$\operatorname{var}(T) = \frac{2n(\sigma^2)^2}{(n+2)^2}.$$

Its bias is

$$\left(1 - \frac{n}{n+2}\right)\sigma^2 = \frac{2\sigma^2}{n+2}$$

so its mean squared error is

$$\operatorname{MSE}(T) = \left(\frac{2n}{(n+2)^2} + \frac{4}{(n+2)^2}\right)(\sigma^2)^2 = \frac{2(\sigma^2)^2}{n+2}$$

which is clearly less than the Cramér-Rao bound found above.

When the mean is not known, the minimum mean squared error estimate of the variance of a sample from Gaussian distribution is achieved by dividing by *n*+1, rather than *n*-1 or *n*+2.

Chapter-14

Extended Kalman Filter

In estimation theory, the **extended Kalman filter** (EKF) is the nonlinear version of the Kalman filter which linearizes about the current mean and covariance. The EKF has been considered the *de facto* standard in the theory of nonlinear state estimation, navigation systems and GPS.

Formulation

In the extended Kalman filter, the state transition and observation models need not be linear functions of the state but may instead be differentiable functions.

$$\begin{aligned}\mathbf{x}_k &= f(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}) + \mathbf{w}_{k-1} \\ \mathbf{z}_k &= h(\mathbf{x}_k) + \mathbf{v}_k\end{aligned}$$

Where $\mathbf{w}_k$ and $\mathbf{v}_k$ are the process and observation noises which are both assumed to be zero mean multivariate Gaussian noises with covariance $\mathbf{Q}_k$ and $\mathbf{R}_k$ respectively.

The function *f* can be used to compute the predicted state from the previous estimate and similarly the function *h* can be used to compute the predicted measurement from the predicted state. However, *f* and *h* cannot be applied to the covariance directly. Instead a matrix of partial derivatives (the Jacobian) is computed.

At each timestep the Jacobian is evaluated with current predicted states. These matrices can be used in the Kalman filter equations. This process essentially linearizes the non-linear function around the current estimate.

Predict and update equations

Predict

Predicted state $\hat{\mathbf{x}}_{k|k-1} = f(\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{u}_{k-1})$

Predicted estimate covariance $\mathbf{P}_{k|k-1} = \mathbf{F}_{k-1}\mathbf{P}_{k-1|k-1}\mathbf{F}_{k-1}^{\top} + \mathbf{Q}_{k-1}$

Update

Innovation or measurement residual	$\tilde{\mathbf{y}}_k = \mathbf{z}_k - h(\hat{\mathbf{x}}_{k\|k-1})$
Innovation (or residual) covariance	$\mathbf{S}_k = \mathbf{H}_k \mathbf{P}_{k\|k-1} \mathbf{H}_k^\top + \mathbf{R}_k$
Near-Optimal Kalman gain	$\mathbf{K}_k = \mathbf{P}_{k\|k-1} \mathbf{H}_k^\top \mathbf{S}_k^{-1}$
Updated state estimate	$\hat{\mathbf{x}}_{k\|k} = \hat{\mathbf{x}}_{k\|k-1} + \mathbf{K}_k \tilde{\mathbf{y}}_k$
Updated estimate covariance	$\mathbf{P}_{k\|k} = (I - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_{k\|k-1}$

where the state transition and observation matrices are defined to be the following Jacobians

$$\mathbf{F}_{k-1} = \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{u}_{k-1}}$$

$$\mathbf{H}_k = \left.\frac{\partial h}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}_{k|k-1}}$$

Continuous-time extended Kalman filter

Model

$$\dot{\mathbf{x}}(t) = f\big(\mathbf{x}(t), \mathbf{u}(t)\big) + \mathbf{w}(t), \quad \mathbf{w}(t) \sim N\big(\mathbf{0}, \mathbf{Q}(t)\big)$$

$$\mathbf{z}(t) = h\big(\mathbf{x}(t)\big) + \mathbf{v}(t), \quad \mathbf{v}(t) \sim N\big(\mathbf{0}, \mathbf{R}(t)\big)$$

Initialize

$$\hat{\mathbf{x}}(t_0) = E\big[\mathbf{x}(t_0)\big], \; \mathbf{P}(t_0) = Var\big[\mathbf{x}(t_0)\big]$$

Predict-Update

$$\dot{\hat{\mathbf{x}}}(t) = f\big(\hat{\mathbf{x}}(t), \mathbf{u}(t)\big) + \mathbf{K}(t)\Big(\mathbf{z}(t) - h\big(\hat{\mathbf{x}}(t)\big)\Big)$$

$$\dot{\mathbf{P}}(t) = \mathbf{F}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}(t)^\top - \mathbf{K}(t)\mathbf{H}(t)\mathbf{P}(t) + \mathbf{Q}(t)$$

$$\mathbf{K}(t) = \mathbf{P}(t)\mathbf{H}(t)^\top \mathbf{R}(t)^{-1}$$

$$\mathbf{F}(t) = \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}(t), \mathbf{u}(t)}$$

$$\mathbf{H}(t) = \left.\frac{\partial h}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}(t)}$$

Unlike discrete-time extended Kalman filter, the prediction and update steps are coupled in continuous-time extended Kalman filter.

Continuous-discrete extended Kalman filter

Most physical systems are represented as continuous-time models while discrete-time measurements are frequently taken for state estimation via a digital processor. Therefore, the system model and measurement model are given by

$$\begin{aligned}\dot{\mathbf{x}}(t) &= f\big(\mathbf{x}(t), \mathbf{u}(t)\big) + \mathbf{w}(t), \quad \mathbf{w}(t) \sim N\big(\mathbf{0}, \mathbf{Q}(t)\big) \\ \mathbf{z}_k &= h(\mathbf{x}_k) + \mathbf{v}_k, \qquad\qquad\quad \mathbf{v}_k \sim N(\mathbf{0}, \mathbf{R}_k)\end{aligned}$$

where $\mathbf{x}_k = \mathbf{x}(t_k)$.

Initialize

$$\hat{\mathbf{x}}_{0|0} = E\big[\mathbf{x}(t_0)\big], \mathbf{P}_{0|0} = Var\big[\mathbf{x}(t_0)\big]$$

Predict

$$\begin{cases}\dot{\hat{\mathbf{x}}}(t) = f\big(\hat{\mathbf{x}}(t), \mathbf{u}(t)\big) \\ \dot{\mathbf{P}}(t) = \mathbf{F}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}(t)^\top + \mathbf{Q}(t)\end{cases}, \text{ with } \begin{cases}\hat{\mathbf{x}}(t_{k-1}) = \hat{\mathbf{x}}_{k-1|k-1} \\ \mathbf{P}(t_{k-1}) = \mathbf{P}_{k-1|k-1}\end{cases}$$

$$\Rightarrow \begin{cases}\hat{\mathbf{x}}_{k|k-1} = \hat{\mathbf{x}}(t_k) \\ \mathbf{P}_{k|k-1} = \mathbf{P}(t_k)\end{cases}$$

where

$$\mathbf{F}(t) = \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}(t), \mathbf{u}(t)}$$

Update

$$\begin{aligned}\mathbf{K}_k &= \mathbf{P}_{k|k-1}\mathbf{H}_k^\top \big(\mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}_k^\top + \mathbf{R}_k\big)^{-1} \\ \hat{\mathbf{x}}_{k|k} &= \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k\big(\mathbf{z}_k - h(\hat{\mathbf{x}}_{k|k-1})\big) \\ \mathbf{P}_{k|k} &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k|k-1}\end{aligned}$$

where

$$\mathbf{H}_k = \left.\frac{\partial h}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}_{k|k-1}}$$

The update equations are identical to those of discrete-time extended Kalman filter.

Disadvantages of the extended Kalman filter

Unlike its linear counterpart, the extended Kalman filter in general is *not* an optimal estimator (of course it is optimal if the measurement and the state transition model are both linear, as in that case the extended Kalman filter is identical to the regular one). In addition, if the initial estimate of the state is wrong, or if the process is modeled incorrectly, the filter may quickly diverge, owing to its linearization. Another problem with the extended Kalman filter is that the estimated covariance matrix tends to underestimate the true covariance matrix and therefore risks becoming inconsistent in the statistical sense without the addition of "stabilising noise".

Having stated this, the extended Kalman filter can give reasonable performance, and is arguably the *de facto* standard in navigation systems and GPS.

Unscented Kalman filters

An improvement to the extended Kalman filter led to the development of the Unscented Kalman filter (UKF), also a nonlinear filter. In the UKF, the probability density is approximated by the nonlinear transformation of a random variable, which returns much more accurate results than the first-order Taylor expansion of the nonlinear functions in the EKF. The approximation utilizes a set of sample points, which guarantees accuracy with the posterior mean and covariance to the second order for any nonlinearity. The UKF tends to be more robust and more accurate than the EKF in its estimation of error.

"The extended Kalman filter (EKF) is probably the most widely used estimation algorithm for nonlinear systems. However, more than 35 years of experience in the estimation community has shown that is difficult to implement, difficult to tune, and only reliable for systems that are almost linear on the time scale of the updates. Many of these difficulties arise from its use of linearization."

Invariant extended Kalman filter

The invariant extended Kalman filter (IEKF) is a modified version of the EKF for nonlinear systems possessing symmetries (or *invariances*). It combines the advantages of both the EKF and the recently introduced symmetry-preserving filters. Indeed, instead of using a linear correction term based on a linear output error, it uses a geometrically adapted correction term based on an invariant output error; in the same way the gain matrix is not updated from a linear state error, but from an invariant state error. The main benefit is that the gain and covariance equations converge to constant values on a much

bigger set of trajectories than equilibrium points as it is the case for the EKF, which results in a better convergence of the estimation.

Chapter-15

Fisher Information

In mathematical statistics and information theory, the **Fisher information** (sometimes simply called **information**) is the variance of the score. In Bayesian statistics, the asymptotic distribution of the posterior mode depends on the Fisher information and not on the prior (according to the Bernstein–von Mises Theorem, which was anticipated by Laplace for exponential families). The role of the Fisher information in the asymptotic theory of maximum-likelihood estimation was emphasized by the statistician R.A. Fisher (following some initial results by F. Y. Edgeworth). The Fisher information is also used in the calculation of the Jeffreys prior, which is used in Bayesian statistics.

The Fisher-information matrix is used to calculate the covariance matrices associated with maximum-likelihood estimates. It can also be used in the formulation of test statistics, such as the Wald test.

History

The Fisher information was discussed by several early statisticians, notably F. Y. Edgeworth. For example, Savage says: "In it [Fisher information], he [Fisher] was to some extent anticipated (Edgeworth 1908–9 esp. 502, 507–8, 662, 677–8, 82–5 and references he [Edgeworth] cites including Pearson and Filon 1898 [. . .])." There are a number of early historical sources and a number of reviews of this early work.

Definition

The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ upon which the probability of X depends. The probability function for X, which is also the likelihood function for θ, is a function $f(X; \theta)$; it is the probability mass (or probability density) of the random variable X conditional on the value of θ. The partial derivative with respect to θ of the log of the likelihood function is called the score. Under certain regularity conditions, it can be shown that the first moment of the score is 0. The second moment is called the Fisher information:

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2 \middle| \theta\right],$$

where, for any given value of θ, the expression E[…|θ] denotes the conditional expectation over values for X with respect to the probability function $f(x; \theta)$ given θ. Note that $0 \leq \mathcal{I}(\theta) < \infty$. A random variable carrying high Fisher information implies that the absolute value of the score is often high.The Fisher information is not a function of a particular observation, as the random variable X has been averaged out.

Since the expectation of the score is zero, the Fisher information is also the variance of the score.

If ln $f(x; \theta)$ is twice differentiable with respect to θ, and under certain regularity conditions, then the Fisher information may also be written as

$$\mathcal{I}(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta) \middle| \theta\right] \ .$$

Thus, the Fisher information is the negative of the expectation of the second derivative with respect to θ of the log of f. Information may be seen to be a measure of the "curvature" of the support curve near the maximum likelihood estimate of θ. A "blunt" support curve (one with a shallow maximum) would have a low negative expected second derivative, and thus low information; while a sharp one would have a high negative expected second derivative and thus high information.

Information is additive, in that the information yielded by two independent experiments is the sum of the information from each experiment separately:

$$\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta).$$

This result follows from the elementary fact that if random variables are independent, the variance of their sum is the sum of their variances. Hence the information in a random sample of size n is n times that in a sample of size 1 (if observations are independent).

The information provided by a sufficient statistic is the same as that of the sample X. This may be seen by using Neyman's factorization criterion for a sufficient statistic. If $T(X)$ is sufficient for θ, then

$$f(X;\theta) = g(T(X),\theta)h(X)$$

for some functions g and h. The equality of information then follows from the following fact:

$$\frac{\partial}{\partial\theta}\ln\left[f(X;\theta)\right] = \frac{\partial}{\partial\theta}\ln\left[g(T(X);\theta)\right]$$

which follows from the definition of Fisher information, and the independence of $h(X)$ from θ. More generally, if $T = t(X)$ is a statistic, then

$$\mathcal{I}_T(\theta) \leq \mathcal{I}_X(\theta)$$

with equality if and only if T is a sufficient statistic.

Informal derivation of the Cramér–Rao bound

The Cramér–Rao bound states that the inverse of the Fisher information is a lower bound on the variance of any unbiased estimator of θ. Van Trees (1968) and Frieden (2004) provide the following method of deriving the Cramér–Rao bound, a result which describes use of the Fisher information, informally:

Consider an unbiased estimator $\hat{\theta}(X)$. Mathematically, we write

$$\mathrm{E}\left[\hat{\theta}(X) - \theta \middle| \theta\right] = \int \left[\hat{\theta}(X) - \theta\right] \cdot f(X;\theta)\,dx = 0.$$

The likelihood function $f(X; \theta)$ describes the probability that we observe a given sample x *given* a known value of θ. If f is sharply peaked with respect to changes in θ, it is easy to intuit the "correct" value of θ given the data, and hence the data contains a lot of information about the parameter. If the likelihood f is flat and spread-out, then it would take many, many samples of X to estimate the actual "true" value of θ. Therefore, we would intuit that the data contain much less information about the parameter.

Now, we differentiate the unbiased-ness condition above to get

$$\frac{\partial}{\partial\theta}\int \left[\hat{\theta}(X) - \theta\right] \cdot f(X;\theta)\,dx = \int \left(\hat{\theta} - \theta\right)\frac{\partial f}{\partial\theta}\,dx - \int f\,dx = 0.$$

We now make use of two facts. The first is that the likelihood f is just the probability of the data given the parameter. Since it is a probability, it must be normalized, implying that

$$\int f\,dx = 1.$$

Second, we know from basic calculus that

$$\frac{\partial f}{\partial \theta} = f \frac{\partial \ln f}{\partial \theta}.$$

Using these two facts in the above let us write

$$\int \left(\hat{\theta} - \theta\right) f \frac{\partial \ln f}{\partial \theta} \, dx = 1.$$

Factoring the integrand gives

$$\int \left(\left(\hat{\theta} - \theta\right) \sqrt{f}\right) \left(\sqrt{f} \frac{\partial \ln f}{\partial \theta}\right) \, dx = 1.$$

If we square the equation, the Cauchy–Schwarz inequality lets us write

$$\left[\int \left(\hat{\theta} - \theta\right)^2 f \, dx\right] \cdot \left[\int \left(\frac{\partial \ln f}{\partial \theta}\right)^2 f \, dx\right] \geq 1.$$

The right-most factor is defined to be the Fisher Information

$$\mathcal{I}(\theta) = \int \left(\frac{\partial \ln f}{\partial \theta}\right)^2 f \, dx.$$

The left-most factor is the expected mean-squared error of the estimator θ, since

$$\mathrm{E}\left[\left(\hat{\theta}(X) - \theta\right)^2 \middle| \theta\right] = \int \left(\hat{\theta} - \theta\right)^2 f \, dx.$$

Notice that the inequality tells us that, fundamentally,

$$\mathrm{Var}\left[\hat{\theta}\right] \geq \frac{1}{\mathcal{I}(\theta)}.$$

In other words, the precision to which we can estimate θ is fundamentally limited by the Fisher Information of the likelihood function.

Single-parameter Bernoulli experiment

A Bernoulli trial is a random variable with two possible outcomes, "success" and "failure", with "success" having a probability of θ. The outcome can be thought of as

determined by a coin toss, with the probability of obtaining a "head" being θ and the probability of obtaining a "tail" being $1 - \theta$.

The Fisher information contained in n independent Bernoulli trials may be calculated as follows. In the following, A represents the number of successes, B the number of failures, and $n = A + B$ is the total number of trials.

$$\begin{aligned}
\mathcal{I}(\theta) &= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln(f(A;\theta))\middle|\theta\right] && (1)\\
&= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln\left(\theta^A(1-\theta)^B\frac{(A+B)!}{A!B!}\right)\middle|\theta\right] && (2)\\
&= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\left(A\ln(\theta)+B\ln(1-\theta)\right)\middle|\theta\right] && (3)\\
&= -\mathrm{E}\left[\frac{\partial}{\partial\theta}\left(\frac{A}{\theta}-\frac{B}{1-\theta}\right)\middle|\theta\right] && \text{(on differentiating ln } x\text{)} \quad (4)\\
&= +\mathrm{E}\left[\frac{A}{\theta^2}+\frac{B}{(1-\theta)^2}\middle|\theta\right] && (5)\\
&= \frac{n\theta}{\theta^2}+\frac{n(1-\theta)}{(1-\theta)^2} && \text{(as the expected value of } A \text{ given } \theta \text{ is } n\theta\text{, etc.)} \quad (6)\\
&= \frac{n}{\theta(1-\theta)} && (7)
\end{aligned}$$

(1) defines Fisher information. (2) invokes the fact that the information in a sufficient statistic is the same as that of the sample itself. (3) expands the log term and drops a constant. (4) and (5) differentiate with respect to θ. (6) replaces A and B with their expectations. (7) is algebra.

The end result, namely,

$$\mathcal{I}(\theta) = \frac{n}{\theta(1-\theta)},$$

is the reciprocal of the variance of the mean number of successes in n Bernoulli trials, as expected.

Matrix form

When there are N parameters, so that θ is a Nx1 vector $\theta = \left[\theta_1, \theta_2, \ldots, \theta_N\right]^{\mathrm{T}}$, then the Fisher information takes the form of an NxN matrix, the Fisher Information Matrix (FIM), with typical element:

$$(\mathcal{I}(\theta))_{i,j} = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\ln f(X;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\ln f(X;\theta)\right)\middle|\,\theta\right].$$

The FIM is a *N*x*N* positive semidefinite symmetric matrix, defining a Riemannian metric on the *N*-dimensional parameter space, thus connecting Fisher information to differential geometry. In that context, this metric is known as the Fisher information metric, and the topic is called information geometry.

Under certain regularity conditions, the Fisher Information Matrix may also be written as:

$$(\mathcal{I}(\theta))_{i,j} = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\ln f(X;\theta)\middle|\,\theta\right].$$

Orthogonal parameters

We say that two parameters θ_i and θ_j are orthogonal if the element of the *i*th row and *j*th column of the Fisher information matrix is zero. Orthogonal parameters are easy to deal with in the sense that their maximum likelihood estimates are independent and can be calculated separately. When dealing with research problems, it is very common for the researcher to invest some time searching for an orthogonal parametrization of the densities involved in the problem.

Multivariate normal distribution

The FIM for a *N*-variate multivariate normal distribution has a special form. Let $\mu(\theta) = \left[\mu_1(\theta), \mu_2(\theta), \ldots, \mu_N(\theta)\right]^{\mathrm{T}}$, and let $\Sigma(\theta)$ be the covariance matrix. Then the typical element $\mathcal{I}_{m,n}$, $0 \le m, n < M$, of the FIM for $X \sim N(\mu(\theta), \Sigma(\theta))$ is:

$$\mathcal{I}_{m,n} = \frac{\partial\mu^{\mathrm{T}}}{\partial\theta_m}\Sigma^{-1}\frac{\partial\mu}{\partial\theta_n} + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_m}\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_n}\right),$$

where $(..)^{\mathrm{T}}$ denotes the transpose of a vector, tr(..) denotes the trace of a square matrix, and:

- $\frac{\partial\mu}{\partial\theta_m} = \left[\frac{\partial\mu_1}{\partial\theta_m} \quad \frac{\partial\mu_2}{\partial\theta_m} \quad \cdots \quad \frac{\partial\mu_N}{\partial\theta_m}\right]^{\mathrm{T}}$;

$$\frac{\partial \Sigma}{\partial \theta_m} = \begin{bmatrix} \frac{\partial \Sigma_{1,1}}{\partial \theta_m} & \frac{\partial \Sigma_{1,2}}{\partial \theta_m} & \cdots & \frac{\partial \Sigma_{1,N}}{\partial \theta_m} \\ \frac{\partial \Sigma_{2,1}}{\partial \theta_m} & \frac{\partial \Sigma_{2,2}}{\partial \theta_m} & \cdots & \frac{\partial \Sigma_{2,N}}{\partial \theta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \Sigma_{N,1}}{\partial \theta_m} & \frac{\partial \Sigma_{N,2}}{\partial \theta_m} & \cdots & \frac{\partial \Sigma_{N,N}}{\partial \theta_m} \end{bmatrix}.$$

Note that a special, but very common, case is the one where $\Sigma(\theta) = \Sigma$, a constant. Then

$$\mathcal{I}_{m,n} = \frac{\partial \mu^{\mathrm{T}}}{\partial \theta_m} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_n}.$$

In this case the Fisher information matrix may be identified with the coefficient matrix of the normal equations of least squares estimation theory.

Properties

Reparametrization

The Fisher information depends on the parametrization of the problem. If θ and η are two scalar parametrizations of an estimation problem, and θ is a continuously differentiable function of η, then

$$\mathcal{I}_\eta(\eta) = \mathcal{I}_\theta(\theta(\eta)) \left(\frac{\mathrm{d}\theta}{\mathrm{d}\eta} \right)^2$$

where $\mathcal{I}_\eta$ and $\mathcal{I}_\theta$ are the Fisher information measures of η and θ, respectively.

In the vector case, suppose $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are *k*-vectors which parametrize an estimation problem, and suppose that $\boldsymbol{\theta}$ is a continuously differentiable function of $\boldsymbol{\eta}$, then,

$$\mathcal{I}_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = \boldsymbol{J}^{\mathrm{T}} \mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{\eta})) \boldsymbol{J}$$

where the (*i*, *j*)th element of the *k* × *k* Jacobian matrix $\boldsymbol{J}$ is defined by

$$J_{ij} = \frac{\partial \theta_i}{\partial \eta_j},$$

and where $\boldsymbol{J}^{\mathrm{T}}$ is the matrix transpose of $\boldsymbol{J}$.

In information geometry, this is seen as a change of coordinates on a Riemannian manifold, and the intrinsic properties of curvature are unchanged under different parametrization.

Applications

Optimal design of experiments

Fisher information is widely used in optimal experimental design. Because of the reciprocity of estimator-variance and Fisher information, ***minimizing* the *variance*** corresponds to ***maximizing* the *information***.

When the linear (or linearized) statistical model has several parameters, the mean of the parameter-estimator is a vector and its variance is a matrix. The inverse matrix of the variance-matrix is called the "information matrix". Because the variance of the estimator of a parameter vector is a matrix, the problem of "minimizing the variance" is complicated. Using statistical theory, statisticians compress the information-matrix using real-valued summary statistics; being real-valued functions, these "information criteria" can be maximized.

Traditionally, statisticians have evaluated estimators and designs by considering some summary statistic of the covariance matrix (of a mean-unbiased estimator), usually with positive real values (like the determinant or matrix trace). Working with positive real-numbers brings several advantages: If the estimator of a single parameter has a positive variance, then the variance and the Fisher information are both positive real numbers; hence they are members of the convex cone of nonnegative real numbers (whose nonzero members have reciprocals in this same cone). For several parameters, the covariance-matrices and information-matrices are elements of the convex cone of nonnegative-definite symmetric matrices in a partially ordered vector space, under the Loewner (Löwner) order. This cone is closed under matrix-matrix addition, under matrix-inversion, and under the multiplication of positive real-numbers and matrices. An exposition of matrix theory and the Loewner-order appears in Pukelsheim.

The traditional optimality-criteria are the information-matrix's invariants; algebraically, the traditional optimality-criteria are functionals of the eigenvalues of the (Fisher) information matrix.

Jeffreys prior in Bayesian statistics

In Bayesian statistics, the Fisher information is used to calculate the Jeffreys prior, which is a standard, non-informative prior for continuous distribution parameters.

Relation to KL-divergence

Fisher information is the curvature of the Kullback–Leibler information.

Distinction from the Hessian of the entropy

In general, the Fisher Information

$$\mathcal{I}(\theta) = \int f(X;\theta) \left(-\frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2} \right) dx = \int f(X;\theta) \left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 dx$$

is not the same as the negative of the second derivative of the entropy

$$-\frac{\partial^2 H}{\partial \theta^2} = \frac{\partial^2}{\partial \theta^2} \int f(X;\theta) \ln f(X;\theta)\, dx\,.$$

For instance, with $f(X;\theta) = \frac{e^{-(x-\theta)^2/2}}{\sqrt{2\pi}}$, the entropy H is independent of the distribution mean θ. Thus, in this case, the second derivative of the entropy is zero. However, for the Fisher information, we have $\mathcal{I}(\theta) = 1$.

Chapter-16

Generalized Method of Moments

In econometrics, **generalized method of moments** (**GMM**) is a generic method for estimating parameters in statistical models. Usually it is applied in the context of semiparametric models, where the parameter of interest is finite-dimensional, whereas the full shape of the distribution function of the data may not be known, and therefore the maximum likelihood estimation is not applicable.

The method requires that a certain number of *moment conditions* were specified for the model. These moment conditions are functions of the model parameters and the data, such that their expectation is zero at the true values of the parameters. The GMM method then minimizes a certain norm of the sample averages of the moment conditions.

The GMM estimators are known to be consistent, asymptotically normal, and efficient in the class of all estimators that don't use any extra information aside from that contained in the moment conditions.

GMM was developed by Lars Peter Hansen in 1982 as a generalization of the method of moments.

Description

Suppose the available data consists of T of iid observations $\{Y_t\}_{t=1,\ldots,T}$, where each observation Y_t is an n-dimensional multivariate random variable. The data comes from a certain statistical model, defined up to an unknown parameter $\theta \in \Theta$. The goal of the estimation problem is to find the "true" value of this parameter, θ_0, or at least a reasonably close estimate.

In order to apply GMM there should exist a vector-valued function $g(Y,\theta)$ such that

$$m(\theta_0) \equiv \mathrm{E}[\, g(Y_t, \theta_0)\,] = 0,$$

where E denotes expectation, and Y_t is a generic observation, which are all assumed to be iid. Moreover, function $m(\theta)$ must not be equal to zero for $\theta \neq \theta_0$, or otherwise parameter θ will not be identified.

The basic idea behind GMM is to replace the theoretical expected value $\mathrm{E}[\cdot]$ with its empirical analog — sample average:

$$\hat{m}(\theta) = \hat{\mathrm{E}}\big[\, g(Y_t, \theta) \,\big] \equiv \frac{1}{T} \sum_{t=1}^{T} g(Y_t, \theta)$$

and then to minimize the norm of this expression with respect to θ.

By the law of large numbers, $\hat{m}(\theta) \approx \mathrm{E}[g(Y_t, \theta)] = m(\theta)$ for large values of T, and thus we expect that $\hat{m}(\theta_0) \approx m(\theta_0) = 0$. The generalized method of moments looks for a number $\hat{\theta}$ which would make $\hat{m}(\hat{\theta})$ as close to zero as possible. Mathematically, this is equivalent to minimizing a certain norm of $\hat{m}(\theta)$ (norm of m, denoted as $\|m\|$, measures the distance between m and zero). The properties of the resulting estimator will depend on the particular choice of the norm function, and therefore the theory of GMM considers an entire family of norms, defined as

$$\|\hat{m}(\theta)\|_W^2 = \hat{m}(\theta)' \, W \hat{m}(\theta),$$

where W is a positive-definite weighting matrix, and m' denotes transposition. In practice, the weighting matrix W is computed based on the available data set, which will be denoted as $\hat{W}$. Thus, the GMM estimator can be written as

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \left(\frac{1}{T} \sum_{t=1}^{T} g(Y_t, \theta) \right)' \hat{W} \left(\frac{1}{T} \sum_{t=1}^{T} g(Y_t, \theta) \right)$$

Under suitable conditions this estimator is consistent, asymptotically normal, and with right choice of weighting matrix $\hat{W}$ asymptotically efficient.

Properties

Consistency

Consistency is the most important property of an estimator. It means that having sufficient number of observations, the estimator will get arbitrarily close to the true value of parameter:

$$\hat{\theta} \xrightarrow{p} \theta_0 \text{ as } T \to \infty$$

Necessary and sufficient conditions for a GMM estimator to be consistent are as follows:

1. $\hat{W}_T \xrightarrow{p} W$, where W is a positive semi-definite matrix,
2. $W\mathrm{E}[\, g(Y_t, \theta)\,] = 0$ only for $\theta = \theta_0$,
3. $\theta_0 \in \Theta$, which is compact,
4. $g(Y, \theta)$ is continuous at each θ with probability one,
5. $\mathrm{E}[\sup_{\theta \in \Theta} \|g(Y, \theta)\|] < \infty$.

The second condition here (so-called **Global identification** condition) is often particularly hard to verify. There exist simpler necessary but not sufficient conditions, which may be used to detect non-identification problem:

- **Order condition**. The dimension of moment function *m(θ)* should be at least as large as the dimension of parameter vector *θ*.
- **Local identification**. If *g(Y,θ)* is continuously differentiable in a neighborhood of θ_0, then matrix $W\mathrm{E}[\nabla_\theta g(Y_t, \theta_0)]$ must have full column rank.

In practice applied econometricians often simply *assume* that global identification holds, without actually proving it.

Asymptotic normality

Asymptotic normality is a useful property, as it allows us to construct confidence bands for the estimator, and conduct different tests. Before we can make a statement about the asymptotic distribution of the GMM estimator, we need to define two auxiliary matrices:

$$G = \mathrm{E}[\, \nabla_\theta\, g(Y_t, \theta_0)\,], \qquad \Omega = \mathrm{E}[\, g(Y_t, \theta_0) g(Y_t, \theta_0)'\,]$$

Then under conditions 1–6 listed below, the GMM estimator will be asymptotically normal with limiting distribution

$$\sqrt{T}\big(\hat{\theta} - \theta_0\big) \xrightarrow{d} \mathcal{N}\big[0, (G'WG)^{-1} G'W\Omega WG(G'WG)^{-1}\big]$$

Conditions:

1. $\hat{\theta}$ is consistent,
2. θ_0 lies in the interior of set Θ,
3. $g(Y, \theta)$ is continuously differentiable in some neighborhood N of θ_0 with probability one,
4. $\mathrm{E}[\, \|g(Y_t, \theta)\|^2\,] < \infty$,
5. $\mathrm{E}[\sup_{\theta \in N} \|\nabla_\theta g(Y_t, \theta)\|\,] < \infty$,
6. matrix *G'WG* is nonsingular.

Efficiency

So far we have said nothing about the choice of matrix W, except that it must be positive semi-definite. In fact any such matrix will produce a consistent and asymptotically normal GMM estimator, the only difference will be in the asymptotic variance of that estimator. It can be shown that taking

$$W \propto \Omega^{-1}$$

will result in the most efficient estimator in the class of all asymptotically normal estimators. Efficiency in this case means that such an estimator will have the smallest possible variance (we say that matrix A is smaller than matrix B if $B–A$ is positive semi-definite).

In this case the formula for the asymptotic distribution of the GMM estimator simplifies to

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left[0, (G'\,\Omega^{-1}G)^{-1}\right]$$

The proof that such a choice of weighting matrix is indeed optimal is quite elegant, and is often adopted with slight modifications when establishing efficiency of other estimators. As a rule of thumb, a weighting matrix is optimal whenever it makes the "sandwich formula" for variance collapse into a simpler expression.

Proof. We will consider the difference between asymptotic variance with arbitrary W and asymptotic variance with $W = \Omega^{-1}$. If we can factor this difference into a symmetric product of the form CC' for some matrix C, then it will guarantee that this difference is nonnegative-definite, and thus $W = \Omega^{-1}$ will be optimal by definition.

$$\begin{aligned} V(W) - V(\Omega^{-1}) &= (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1} - (G'\Omega^{-1}G)^{-1} \\ &= (G'WG)^{-1}\Big(G'W\Omega WG - G'WG(G'\Omega^{-1}G)^{-1}G'WG\Big)(G'WG)^{-1} \\ &= (G'WG)^{-1}G'W\Omega^{1/2}\Big(I - \Omega^{-1/2}G(G'\Omega^{-1}G)^{-1}G'\Omega^{-1/2}\Big)\Omega^{1/2}WG(G'WG)^{-1} \\ &= A(I-B)A', \end{aligned}$$

where we introduced matrices A and B in order to slightly simplify notation; I is an identity matrix. We can see that matrix B here is symmetric and idempotent: $B^2 = B$. This means $I–B$ is symmetric and idempotent as well: $I - B = (I - B)(I - B)'$. Thus we can continue to factor the previous expression as

$$= A(I-B)(I-B)'A' = \Big(A(I-B)\Big)\Big(A(I-B)\Big)' \geq 0$$

Implementation

One difficulty with implementing the outlined method is that we cannot take $W = \Omega^{-1}$ because, by the definition of matrix Ω, we need to know the value of θ_0 in order to compute this matrix, and θ_0 is precisely the quantity we don't know and are trying to estimate in the first place.

Several approaches exist to deal with this issue, the first one being the most popular:

- **Two-step feasible GMM**:
  - *Step 1*: Take $W = I$ (the identity matrix), and compute preliminary GMM estimate $\hat{\theta}_{(1)}$. This estimator is consistent for θ_0, although not efficient.
  - *Step 2*: Take

$$\hat{W} = \left(\frac{1}{T} \sum_{t=1}^{T} g(Y_t, \hat{\theta}_{(1)}) g(Y_t, \hat{\theta}_{(1)})' \right)^{-1},$$

    where we have plugged in our first-step preliminary estimate $\hat{\theta}_{(1)}$. This matrix converges in probability to Ω^{-1} and therefore if we compute $\hat{\theta}$with this weighting matrix, the estimator will be asymptotically efficient.

- **Iterated GMM**. Essentially the same procedure as 2-step GMM, except that the matrix $\hat{W}_T$is recalculated several times. That is, the estimate obtained in step 2 is used to calculate the weighting matrix for step 3, and so on. Such estimator, denoted $\hat{\theta}_{(i)}$, is equivalent to solving the following system of equations:

$$\left(\frac{1}{T} \sum_{t=1}^{T} \frac{\partial g}{\partial \theta'}(Y_t, \hat{\theta}_{(i)}) \right)' \left(\frac{1}{T} \sum_{t=1}^{T} g(Y_t, \hat{\theta}_{(i)}) g(Y_t, \hat{\theta}_{(i)})' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} g(Y_t, \hat{\theta}_{(i)}) \right) = 0$$

  Asymptotically no improvement can be achieved through such iterations, although certain Monte-Carlo experiments suggest that finite-sample properties of this estimator are slightly better.

- **Continuously Updating GMM** (CUGMM, or CUE). Estimates $\hat{\theta}$simultaneously with estimating the weighting matrix W:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \left(\frac{1}{T} \sum_{t=1}^{T} g(Y_t, \theta) \right)' \left(\frac{1}{T} \sum_{t=1}^{T} g(Y_t, \theta) g(Y_t, \theta)' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} g(Y_t, \theta) \right)$$

  In Monte-Carlo experiments this method demonstrated a better performance than

the traditional two-step GMM: the estimator has smaller median bias (although fatter tails), and the J-test for overidentifying restrictions in many cases was more reliable.

Another important issue is implementation of minimization procedure is that the function is supposed to search through (possibly high-dimensional) parameter space Θ and find the value of θ which minimizes the objective function. No generic recommendation for such procedure exists, it is a subject of its own field, numerical optimization.

J-test

When the number of moment conditions is greater than the dimension of the parameter vector θ, the model is said to be *over-identified*. Over-identification allows us to check whether the model's moment conditions match the data well or not.

Conceptually we can check whether $\hat{m}(\hat{\theta})$ is sufficiently close to zero to suggest that the model fits the data well. The GMM method has then replaced the problem of solving the equation $\hat{m}(\theta) = 0$, which chooses θ to match the restrictions exactly, by a minimization calculation. The minimization can always be conducted even when no θ_0 exists such that $m(\theta_0) = 0$. This is what J-test does. The J-test is also called a *test for over-identifying restrictions*.

Formally we consider two hypotheses:

- $H_0 : \ m(\theta_0) = 0$ (the null hypothesis that the model is "valid"), and
- $H_1 : \ m(\theta) \neq 0, \ \forall \theta \in \Theta$ (the alternative hypothesis that model is "invalid"; the data do not come close to meeting the restrictions)

Under hypothesis H_0, the following so-called J-statistic is asymptotically *chi-squared* with $k–l$ degrees of freedom. Define J to be:

$$J \equiv T \cdot \left(\frac{1}{T} \sum_{t=1}^{T} g(Y_t, \hat{\theta}) \right)' \hat{W}_T \left(\frac{1}{T} \sum_{t=1}^{T} g(Y_t, \hat{\theta}) \right) \xrightarrow{d} \chi^2_{k-\ell}$$ under H_0,

where $\hat{\theta}$ is the GMM estimator of the parameter θ_0, k is the number of moment conditions (dimension of vector g), and l is the number of estimated parameters (dimension of vector θ). Matrix $\hat{W}_T$ must converge in probability to Ω^{-1}, the efficient weighting matrix (note that previously we only required that W be proportional to Ω^{-1} for estimator to be efficient; however in order to conduct the J-test W must be exactly equal to Ω^{-1}, not simply proportional).

Under the alternative hypothesis H_1, the J-statistic is asymptotically unbounded:

$J \xrightarrow{p} \infty$ under H_1

To conduct the test we compute the value of J from the data. It is a nonnegative number. We compare it with (say) the 0.95 quantile of the $\chi^2_{k-\ell}$ distribution:

- H_0 is *rejected* at 95% confidence level if $J > q^{\chi^2_{k-\ell}}_{0.95}$
- H_0 cannot be rejected at 95% confidence level if $J < q^{\chi^2_{k-\ell}}_{0.95}$

Scope

Many other popular estimation techniques can be cast in terms of GMM optimization:

- Ordinary Least Squares (OLS) is equivalent to GMM with moment conditions:

$$\mathrm{E}[\, x_t(y_t - x_t'\beta)\,] = 0$$

- Generalized Least Squares (GLS)

$$\mathrm{E}[\, x_t(y_t - x_t'\beta)/\sigma^2(x_t)\,] = 0$$

- Instrumental variables regression (IV)

$$\mathrm{E}[\, z_t(y_t - x_t'\beta)\,] = 0$$

- Non-linear Least Squares (NLLS):

$$\mathrm{E}[\, \nabla_\beta\, g(x_t, \beta) \cdot (y_t - g(x_t, \beta))\,] = 0$$

- Maximum likelihood estimation (MLE):

$$\mathrm{E}[\, \nabla_\theta \ln f(x_t, \theta)\,] = 0$$

Chapter-17

Estimator

In statistics, an **estimator** is a rule for calculating an estimate of a given quantity based on observed data: thus the rule and its result (the estimate) are distinguished. Here we discusses estimators that are point estimators; that is, they yield single-valued results, although this includes the possibility of single vector-valued results and results that can be expressed as a single function. This is in contrast to an interval estimator, where the result would be a range of plausible values (or vectors or functions).

Statistical theory is concerned with the properties of estimators; that is, with defining properties that can be used to compare different estimators (different rules for creating estimates) for the same quantity, based on the same data. Such properties can be used to determine the best rules to use under given circumstances. However, in robust statistics, statistical theory goes on to consider the balance between having good properties, if tightly defined assumptions hold, and having less good properties that hold under wider conditions.

Background

An "estimator" or "point estimate" is a statistic (that is, a measurable function of the data) that is used to infer the value of an unknown parameter in a statistical model. The parameter being estimated is sometimes called the *estimand*. It can be either finite-dimensional (in parametric and semi-parametric models), or infinite-dimensional (semi-nonparametric and non-parametric models). If the parameter is denoted θ then the estimator is typically written by adding a “hat” over the symbol: $\hat{\theta}$. Being a function of the data, the estimator is itself a random variable; a particular realization of this random variable is called the "estimate". Sometimes the words “estimator” and “estimate” are used interchangeably.

The definition places virtually no restrictions on which functions of the data can be called the “estimators”. The attractiveness of different estimators can be judged by looking at their properties, such as unbiasedness, mean square error, consistency, asymptotic distribution, etc.. The construction and comparison of estimators are the subjects of the

estimation theory. In the context of decision theory, an estimator is a type of decision rule, and its performance may be evaluated through the use of loss functions.

When the word “estimator” is used without a qualifier, it usually refers to point estimation. The estimate in this case is a single point in the parameter space. Other types of estimators also exist: interval estimators, where the estimates are subsets of the parameter space.

The problem of density estimation arises in two applications. Firstly, in estimating the probability density functions of random variables and secondly in estimating the spectral density function of a time series. In these problems the estimates are functions that can be thought of as point estimates in an infinite dimensional space, and there are corresponding interval estimation problems.

Definition

Suppose there is a fixed *parameter* θ that needs to be estimated. Then an "estimator" is a function that maps the sample space to a set of *sample estimates*. An estimator of θ is usually denoted by the symbol $\widehat{\theta}$. It is often convenient to express the theory using the algebra of random variables: thus if *X* is used to denote a random variable corresponding to the observed data, the estimator (itself treated as a random variable) is symbolised as a function of that random variable, $\widehat{\theta}(X)$. The estimate for a particular observed dataset (i.e. for *X*=*x*) is then $\widehat{\theta}(x)$, which is a fixed value. Often an abbreviated notation is used in which $\widehat{\theta}$ is interpreted directly as a random variable, but this can cause confusion.

Quantified properties

The following definitions and attributes apply:

Error

For a given sample x, the "error" of the estimator $\widehat{\theta}$ is defined as

$$e(x) = \widehat{\theta}(x) - \theta,$$

where θ is the parameter being estimated. Note that the error, *e*, depends not only on the estimator (the estimation formula or procedure), but on the sample.

Mean squared error

The *mean squared error* of $\widehat{\theta}$ is defined as the expected value (probability-weighted average, over all samples) of the squared errors; that is,

$$\mathrm{MSE}(\widehat{\theta}) = \mathrm{E}[(\widehat{\theta}(X) - \theta)^2].$$

It is used to indicate how far, on average, the collection of estimates are from the single parameter being estimated. Consider the following analogy. Suppose the parameter is the bull's-eye of a target, the estimator is the process of shooting arrows at the target, and the individual arrows are estimates (samples). Then high MSE means the average distance of the arrows from the bull's-eye is high, and low MSE means the average distance from the bull's-eye is low. The arrows may or may not be clustered. For example, even if all arrows hit the same point, yet grossly miss the target, the MSE is still relatively large. Note, however, that if the MSE is relatively low, then the arrows are likely more highly clustered (than highly dispersed).

Sampling deviation

For a given sample x, the *sampling deviation* of the estimator $\widehat{\theta}$ is defined as

$$d(x) = \widehat{\theta}(x) - \mathrm{E}(\widehat{\theta}(X)) = \widehat{\theta}(x) - \mathrm{E}(\widehat{\theta}),$$

where $\mathrm{E}(\widehat{\theta}(X))$ is the expected value of the estimator. Note that the sampling deviation, *d*, depends not only on the estimator, but on the sample.

Variance

The *variance* of $\widehat{\theta}$ is simply the expected value of the squared sampling deviations; that is, $\mathrm{var}(\widehat{\theta}) = \mathrm{E}[(\widehat{\theta} - \mathrm{E}(\widehat{\theta}))^2]$. It is used to indicate how far, on average, the collection of estimates are from the *expected value* of the estimates. Note the difference between MSE and variance. If the parameter is the bull's-eye of a target, and the arrows are estimates, then a relatively high variance means the arrows are dispersed, and a relatively low variance means the arrows are clustered. Some things to note: even if the variance is low, the cluster of arrows may still be far off-target, and even if the variance is high, the diffuse collection of arrows may still be unbiased. Finally, note that even if all arrows grossly miss the target, if they nevertheless all hit the same point, the variance is zero.

Bias

The *bias* of $\widehat{\theta}$ is defined as $B(\widehat{\theta}) = \mathrm{E}(\widehat{\theta}) - \theta$. It is the distance between the average of the collection of estimates, and the single parameter being estimated. It also is the expected value of the error, since $\mathrm{E}(\widehat{\theta}) - \theta = \mathrm{E}(\widehat{\theta} - \theta)$. If the parameter is the bull's-eye of a target, and the arrows are estimates, then a relatively high absolute value for the bias means the average position of the arrows is off-target, and a relatively low absolute bias means the average position of the arrows is on target. They may be

dispersed, or may be clustered. The relationship between bias and variance is analogous to the relationship between accuracy and precision.

Unbiased

The estimator $\widehat{\theta}$ is an *unbiased estimator* of θ if and only if $B(\widehat{\theta}) = 0$. Note that bias is a property of the estimator, not of the estimate. Often, people refer to a "biased estimate" or an "unbiased estimate," but they really are talking about an "estimate from a biased estimator," or an "estimate from an unbiased estimator." Also, people often confuse the "error" of a single estimate with the "bias" of an estimator. Just because the error for one estimate is large, does not mean the estimator is biased. In fact, even if all estimates have astronomical absolute values for their errors, if the expected value of the error is zero, the estimator is unbiased. Also, just because an estimator is biased, does not preclude the error of an estimate from being zero (we may have gotten lucky). The ideal situation, of course, is to have an unbiased estimator with low variance, and also try to limit the number of samples where the error is extreme (that is, have few outliers). Yet unbiasedness is not essential. Often, if just a little bias is permitted, then an estimator can be found with lower MSE and/or fewer outlier sample estimates.

An alternative to the version of "unbiased" above, is "median-unbiased", where the median of the distribution of estimates agrees with the true value; thus, in the long run half the estimates will be too low and half too high. While this applies immediately only to scalar-valued estimators, it can be extended to any measure of central tendency of a distribution.

Relationships

- The MSE, variance, and bias, are related: $\mathrm{MSE}(\widehat{\theta}) = \mathrm{var}(\widehat{\theta}) + (B(\widehat{\theta}))^2$, i.e. mean squared error = variance + square of bias. In particular, for an unbiased estimator, the variance equals the MSE.
- The standard deviation of an estimator of θ (the square root of the variance), or an estimate of the standard deviation of an estimator of θ, is called the *standard error* of θ.

Behavioural properties

Consistency

A consistent sequence of estimators is a sequence of estimators that converge in probability to the quantity being estimated as the index (usually the sample size) grows without bound. In other words, increasing the sample size increases the probability of the estimator being close to the population parameter.

Mathematically, a sequence of estimators $\{t_n; n \geq 0\}$ is a consistent estimator for parameter θ if and only if, for all $\epsilon > 0$, no matter how small, we have

$$\lim_{n\to\infty} \Pr\left\{|t_n - \theta| < \epsilon\right\} = 1.$$

The consistency defined above may be called weak consistency. The sequence is *strongly consistent*, if it converges almost surely to the true value.

An estimator that converges to a *multiple* of a parameter can be made into a consistent estimator by multiplying the estimator by a scale factor, namely the true value divided by the asymptotic value of the estimator. This occurs frequently in estimation of scale parameters by measures of statistical dispersion.

Asymptotic normality

An asymptotically normal estimator is a consistent estimator whose distribution around the true parameter θ approaches a normal distribution with standard deviation shrinking in proportion to $1/\sqrt{n}$ as the sample size n grows. Using $\xrightarrow{D}$ to denote convergence in distribution, t_n is asymptotically normal if

$$\sqrt{n}(t_n - \theta) \xrightarrow{D} N(0, V),$$

for some V, which is called the *asymptotic variance* of the estimator.

The central limit theorem implies asymptotic normality of the sample mean $\bar{x}$ as an estimator of the true mean. More generally, maximum likelihood estimators are asymptotically normal under fairly weak regularity conditions section of the maximum likelihood article. However, not all estimators are asymptotically normal, the simplest examples being case where the true value of a parameter lies in the boundary of the allowable parameter region.

Efficiency

Two naturally desirable properties of estimators are for them to be unbiased and have minimal mean squared error (MSE). These cannot in general both be satisfied simultaneously: a biased estimator may have lower mean squared error (MSE) than any unbiased estimator: despite having bias, the estimator variance may be sufficiently smaller than that of any unbiased estimator, and it may be preferable to use, despite the bias.

Among unbiased estimators, there often exists one with the lowest variance, called the minimum variance unbiased estimator (MVUE). In some cases an unbiased efficient estimator exists, which, in addition to having the lowest variance among unbiased estimators, satisfies the Cramér–Rao bound, which is an absolute lower bound on variance for statistics of a variable.

Chapter-18

Invariant Estimator

In statistics, the concept of being an **invariant estimator** is a criterion that can be used to compare the properties of different estimators for the same quantity. It is a way of formalising the idea that an estimator should have certain intuitively appealing qualities. Strictly speaking, "invariant" would mean that the estimates themselves are unchanged when both the measurements and the parameters are transformed in a compatible way, but the meaning has been extended to allow the estimates to change in appropriate ways with such transformations. The term **equivariant estimator** is used in formal mathematical contexts that include a precise description of the relation of the way the estimator changes in response to changes to the dataset and parameterisation: this corresponds to the use of "equivariance" in more general mathematics.

General setting

Background

In statistical inference, there are several approaches to estimation theory that can be used to decide immediately what estimators should be used according to those approaches. For example, ideas from Bayesian inference would lead directly to Bayesian estimators. Similarly, the theory of classical statistical inference can sometimes lead to strong conclusions about what estimator should be used. However, the usefulness of these theories depends on having a fully prescribed statistical model and may also depend on having a relevant loss function to determine the estimator. Thus a Bayesian analysis might be undertaken, leading to a posterior distribution for relevant parameters, but the use of a specific utility or loss function may be unclear. Ideas of invariance can then be applied to the task of summarising the posterior distribution. In other cases, statistical analyses are undertaken without a fully defined statistical model or the classical theory of statistical inference cannot be readily applied because the family of models being considered are not amenable to such treatment. In addition to these cases where general theory does not prescribe an estimator, the concept of invariance of an estimator can be applied when seeking estimators of alternative forms, either for the sake of simplicity of application of the estimator or so that the estimator is robust.

The concept of invariance is sometimes used on its own as a way of choosing between estimators, but this is not necessarily definitive. For example, a requirement of invariance may be incompatible with the requirement that the estimator be mean-unbiased; on the other hand, the criterion of median-unbiasedness is defined in terms of the estimator's sampling distribution and so is invariant under many transformations.

One use of the concept of invariance is where a class or family of estimators is proposed and a particular formulation must be selected amongst these. One procedure is to impose relevant invariance properties and then to find the formulation within this class that has the best properties, leading to what is called the optimal invariant estimator.

Some classes of invariant estimators

There are several types of transformations that are usefully considered when dealing with invariant estimators. Each gives rise to a class of estimators which are invariant to those particular types of transformation.

- Shift invariance: Notionally, estimates of a location parameter should be invariant to simple shifts of the data values. If all data values are increased by a given amount, the estimate should change by the same amount. When considering estimation using a weighted average, this invariance requirement immediately implies that the weights should sum to one. While the same result is often derived from a requirement for unbiasedness, the use of "invariance" does not require that a mean value exists and makes no use of any probability distribution at all.
- Scale invariance: Note that this is a topic not directly covered in scale invariance.
- Parameter-transformation invariance: Here, the transformation applies to the parameters alone. The concept here is that essentially the same inference should be made from data and a model involving a parameter θ as would be made from the same data if the model used a parameter φ, where φ is a one-to-one transformation of θ, $\varphi=h(\theta)$. According to this type of invariance, results from transformation-invariant estimators should also be related by $\varphi=h(\theta)$. Maximum likelihood estimators have this property.
- Permutation invariance: Where a set of data values can be represented by a statistical model that they are outcomes from independent and identically distributed random variables, it is reasonable to impose the requirement that any estimator of any property of the common distribution should be permutation-invariant: specifically that the estimator, considered as a function of the set of data-values, should not change if items of data are swapped within the dataset.

The combination of permutation invariance and location invariance for estimating a location parameter from an independent and identically distributed dataset using a weighted average implies that the weights should be identical and sum to one. Of course, estimators other than a weighted average may be preferable.

Optimal invariant estimators

Under this setting, we are given a set of measurements x which contains information about an unknown parameter θ. The measurements x are modelled as a vector random variable having a probability density function $f(x \mid \theta)$ which depends on a parameter vector θ.

The problem is to estimate θ given x. The estimate, denoted by a, is a function of the measurements and belongs to a set A. The quality of the result is defined by a loss function $L = L(a,\theta)$ which determines a risk function $R = R(a,\theta) = E[L(a,\theta) \mid \theta]$. The sets of possible values of x, θ, and a are denoted by X, Θ, and A, respectively.

Mathematical setting

Definition

An invariant estimator is an estimator which obeys the following two rules:

1. Principle of Rational Invariance: The action taken in a decision problem should not depend on transformation on the measurement used
2. Invariance Principle: If two decision problems have the same formal structure (in terms of X, Θ, $f(x \mid \theta)$ and L), then the same decision rule should be used in each problem.

To define an invariant or equivariant estimator formally, some definitions related to groups of transformations are needed first. Let X denote the set of possible data-samples. A group of transformations of X, to be denoted by G, is a set of (measurable) 1:1 and onto transformations of X into itself, which satisfies the following conditions:

1. If $g_1 \in G$ and $g_2 \in G$ then $g_1 g_2 \in G$
2. If $g \in G$ then $g^{-1} \in G$, where $g^{-1}(g(x)) = x$. (That is, each transformation has an inverse within the group.)
3. $e \in G$ (i.e there is an identity transformation $e(x) = x$)

Datasets x_1 and x_2 in X are equivalent if $x_1 = g(x_2)$ for some $g \in G$. All the equivalent points form an equivalence class. Such an equivalence class is called an orbit (in X). The x_0 orbit, $X(x_0)$, is the set $X(x_0) = \{g(x_0) : g \in G\}$. If X consists of a single orbit then g is said to be transitive.

A family of densities F is said to be invariant under the group G if, for every $g \in G$ and $\theta \in \Theta$ there exists a unique $\theta^* \in \Theta$ such that $Y = g(x)$ has density $f(y \mid \theta^*)$. θ^* will be denoted $\bar{g}(\theta)$.

If F is invariant under the group G then the loss function $L(\theta,a)$ is said to be invariant under G if for every $g \in G$ and $a \in A$ there exists an $a^* \in A$ such that $L(\theta, a) = L(\bar{g}(\theta), a^*)$ for all $\theta \in \Theta$. The transformed value a^* will be denoted by $\tilde{g}(a)$.

In the above, $\bar{G} = \{\bar{g} : g \in G\}$ is a group of transformations from Θ to itself and $\tilde{G} = \{\tilde{g} : g \in G\}$ is a group of transformations from A to itself.

An estimation problem is invariant(equivariant) under G if there exist three groups $G, \bar{G}, \tilde{G}$ as defined above.

For an estimation problem that is invariant under G, estimator $\delta(x)$ is an invariant estimator under G if, for all $x \in X$ and $g \in G$,

$$\delta(g(x)) = \tilde{g}(\delta(x)).$$

Properties

1. The risk function of an invariant estimator, δ, is constant on orbits of Θ. Equivalently $R(\theta, \delta) = R(\bar{g}(\theta), \delta)$ for all $\theta \in \Theta$ and $\bar{g} \in \bar{G}$.
2. The risk function of an invariant estimator with transitive $\bar{g}$ is constant.

For a given problem, the invariant estimator with the lowest risk is termed the "best invariant estimator". Best invariant estimator cannot always be achieved. A special case for which it can be achieved is the case when $\bar{g}$ is transitive.

Example: Location parameter

Suppose θ is a location parameter if the density of X is of the form $f(x - \theta)$. For $\Theta = A = \mathbb{R}^1$ and $L = L(a - \theta)$, the problem is invariant under $g = \bar{g} = \tilde{g} = \{g_c : g_c(x) = x + c, c \in \mathbb{R}\}$. The invariant estimator in this case must satisfy

$$\delta(x + c) = \delta(x) + c, \text{ for all } c \in \mathbb{R},$$

thus it is of the form $\delta(x) = x + K$ ($K \in \mathbb{R}$). $\bar{g}$ is transitive on Θ so the risk does not vary with θ: that is, $R(\theta, \delta) = R(0, \delta) = \mathrm{E}[L(X + K)|\theta = 0]$. The best invariant estimator is the one that brings the risk $R(\theta,\delta)$ to minimum.

In the case that L is the squared error $\delta(x) = x - \mathrm{E}[X|\theta = 0]$.

Pitman estimator

The estimation problem is that $X = (X_1, \ldots, X_n)$ has density $f(x_1 - \theta, \ldots, x_n - \theta)$, where θ is a parameter to be estimated, and where the loss function is $L(\mid a - \theta \mid)$. This problem is invariant with the following (additive) transformation groups:

$$\begin{aligned} G &= \{g_c : g_c(x) = (x_1 + c, \ldots, x_n + c), c \in \mathbb{R}^1\}, \\ \bar{G} &= \{g_c : g_c(\theta) = \theta + c, c \in \mathbb{R}^1\}, \\ \tilde{G} &= \{g_c : g_c(a) = a + c, c \in \mathbb{R}^1\}. \end{aligned}$$

The best invariant estimator $\delta(x)$ is the one that minimizes

$$\frac{\int_{-\infty}^{\infty} L(\delta(x) - \theta) f(x_1 - \theta, \ldots, x_n - \theta) d\theta}{\int_{-\infty}^{\infty} f(x_1 - \theta, \ldots, x_n - \theta) d\theta},$$

and this is Pitman's estimator (1939).

For the squared error loss case, the result is

$$\delta(x) = \frac{\int_{-\infty}^{\infty} \theta f(x_1 - \theta, \ldots, x_n - \theta) d\theta}{\int_{-\infty}^{\infty} f(x_1 - \theta, \ldots, x_n - \theta) d\theta}.$$

If $x \sim N(\theta 1_n, I)$ (i.e. a multivariate normal distribution with independent, unit-variance components) then

$$\delta_{pitman} = \delta_{ML} = \frac{\sum x_i}{n}.$$

If $x \sim C(\theta 1_n, I\sigma^2)$ (independent components having a Cauchy distribution with scale parameter σ) then $\delta_{pitman} \neq \delta_{ML}$,. However the result is

$$\delta_{pitman} = \sum_{k=1}^{n} x_k \left[\frac{Re\{w_k\}}{\sum_{m=1}^{n} Re\{w_k\}} \right], \qquad n > 1,$$

with

$$w_k = \prod_{j \neq k} \left[\frac{1}{(x_k - x_j)^2 + 4\sigma^2} \right] \left[1 - \frac{2\sigma}{(x_k - x_j)} i \right].$$

Chapter-19

James–Stein Estimator & Kaplan–Meier Estimator

James–Stein Estimator

The **James–Stein estimator** is a nonlinear estimator which can be shown to dominate, or outperform, the "ordinary" (least squares) technique. As such, it is the best-known example of Stein's phenomenon.

An earlier version of the estimator was developed by Charles Stein in 1956, and is sometimes referred to as **Stein's estimator**. The result was improved by Willard James and Charles Stein in 1961.

Setting

Suppose $\boldsymbol{\theta}$ is an unknown parameter vector of length m, and let $\mathbf{y}$ be a vector of observations of $\boldsymbol{\theta}$ of length m, such that the observations are normally distributed:

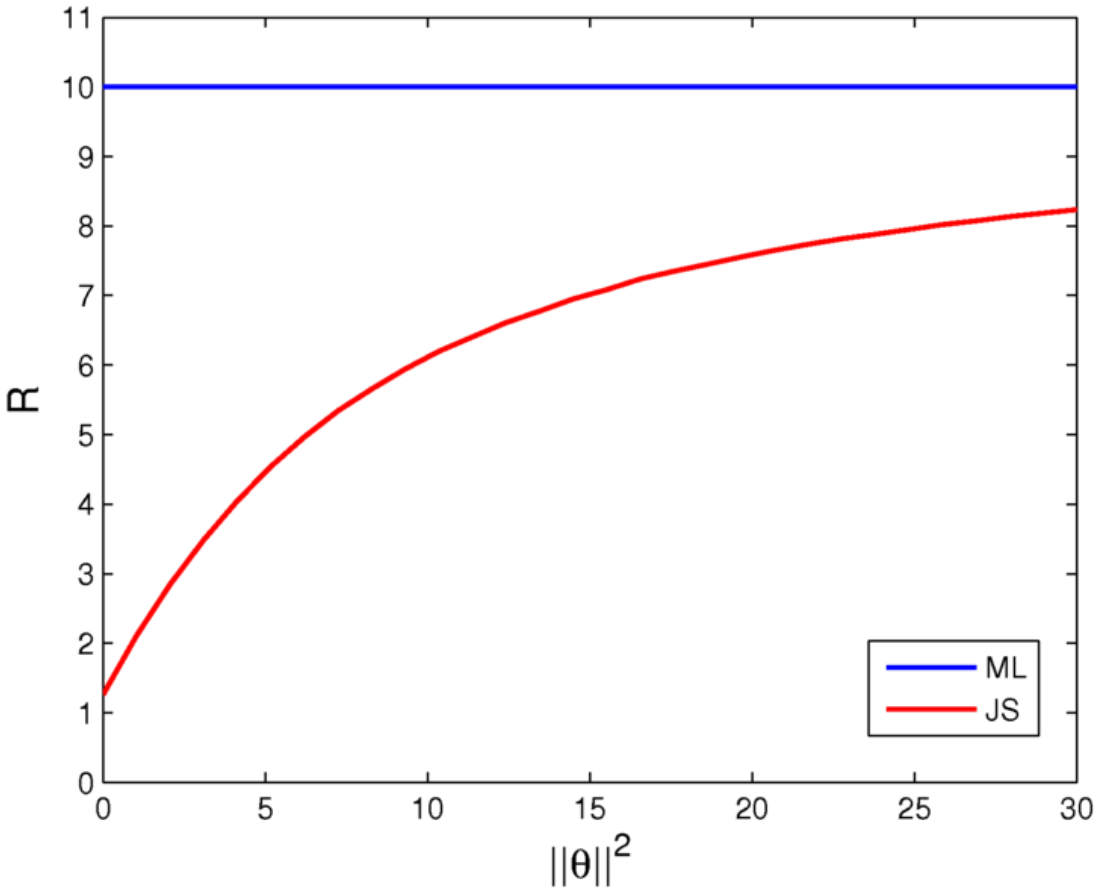
$$\mathbf{y} \sim N(\boldsymbol{\theta}, \sigma^2 I).$$

We are interested in obtaining an estimate $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{y})$ of $\boldsymbol{\theta}$, based on the observations $\mathbf{y}$.

This is an everyday situation in which a set of parameters is measured, and the measurements are corrupted by independent Gaussian noise. Since the noise has zero mean, it is very reasonable to use the measurements themselves as an estimate of the parameters. This is the approach of the least squares estimator, which is $\widehat{\boldsymbol{\theta}}_{LS} = \overline{\mathbf{y}}$, where $\overline{\mathbf{y}}$ is the average of the observations. If there is only a single observation y, then $\widehat{\boldsymbol{\theta}}_{LS} = y$.

As a result, there was considerable shock and disbelief when Stein demonstrated that, in terms of mean squared error $E\{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2\}$, this approach is suboptimal. The result became known as Stein's phenomenon.

The James–Stein estimator



MSE (R) of least squares estimator (ML) vs. James–Stein estimator (JS). The James–Stein estimator gives its best estimate when the norm of the actual parameter vector θ is near zero.

If σ is known, the James–Stein estimator is given by

$$\widehat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\|\overline{\mathbf{y}}\|^2}\right)\overline{\mathbf{y}}.$$

James and Stein showed that the above estimator dominates $\widehat{\boldsymbol{\theta}}_{LS}$ for any $m \geq 3$, meaning that the James–Stein estimator always achieves lower MSE than the least squares estimator.

Notice that if $(m-2)\sigma^2 < \|\overline{\mathbf{y}}\|^2$ then this estimator simply takes the natural estimator $\overline{\mathbf{y}}$ and shrinks it towards the origin $\mathbf{0}$. In fact this is not the only direction of shrinkage that works. Let $\boldsymbol{v}$ be an arbitrary fixed vector of length m. Then there exists a James–Stein estimator that shrinks toward $\boldsymbol{v}$, namely

$$\widehat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\|\overline{\mathbf{y}} - \boldsymbol{\nu}\|^2}\right)(\overline{\mathbf{y}} - \boldsymbol{\nu}) + \boldsymbol{\nu}.$$

It is interesting to note that the James–Stein estimator dominates the usual estimator for any $\boldsymbol{\nu}$. A natural question to ask is whether the improvement over the usual estimator is independent of the choice of $\boldsymbol{\nu}$. The answer is no. The improvement is small if $\|\boldsymbol{\theta} - \boldsymbol{\nu}\|$ is large. Thus to get a very great improvement some knowledge of the location of $\boldsymbol{\theta}$ is necessary. Of course this is the quantity we are trying to estimate so we don't have this knowledge a priori. But we may have some guess as to what the mean vector is. This can be considered a disadvantage of the estimator because the choice is not objective, it depends on the beliefs of the researcher.

Stein has shown that, for $m \leq 2$, the least squares estimator is admissible, meaning that no estimator dominates it.

Interpretation

A consequence of the above discussion is the following counterintuitive result: When three or more unrelated parameters are measured, their total MSE can be reduced by using a combined estimator such as the James–Stein estimator; whereas when each parameter is estimated separately, the least squares (LS) estimator is admissible. This quirk has caused some to sarcastically ask whether, in order to estimate the speed of light, one should jointly estimate tea consumption in Taiwan and hog weight in Montana. The response is that the James–Stein estimator always improves upon the *total* MSE, i.e., the sum of the expected errors of each component. Therefore, the total MSE in measuring light speed, tea consumption and hog weight would improve by using the James–Stein estimator. However, any particular component (such as the speed of light) would improve for some parameter values, and deteriorate for others. Thus, although the James–Stein estimator dominates the LS estimator when three or more parameters are estimated, any single component does not dominate the respective component of the LS estimator.

The conclusion from this hypothetical example is that measurements should be combined if one is interested in minimizing their total MSE. For example, in a telecommunication setting, it is reasonable to combine channel tap measurements in a channel estimation scenario, as the goal is to minimize the total channel estimation error. Conversely, it is probably not reasonable to combine channel estimates of different users, since no user would want their channel estimate to deteriorate in order to improve the average network performance.

Improvements

The basic James–Stein estimator has the peculiar property that for small values of $\|\overline{\mathbf{y}} - \boldsymbol{\nu}\|$, the multiplier on $\overline{\mathbf{y}} - \boldsymbol{\nu}$ is actually negative. This can be easily remedied by replacing this multiplier by zero when it is negative. The resulting estimator is called the *positive-part James–Stein estimator* and is given by

$$\widehat{\boldsymbol{\theta}}_{JS+} = \left(1 - \frac{(m-2)\sigma^2}{\|\overline{\mathbf{y}} - \boldsymbol{\nu}\|^2}\right)^+ (\overline{\mathbf{y}} - \boldsymbol{\nu}) + \boldsymbol{\nu}.$$

This estimator has a smaller risk than the basic James–Stein estimator. It follows that the basic James–Stein estimator is itself inadmissible.

It turns out, however, that the positive-part estimator is also inadmissible. This follows from a general result which requires admissible estimators to be smooth.

Extensions

The James–Stein estimator may seem at first sight to be a result of some peculiarity of the problem setting. In fact, the estimator exemplifies a very wide-ranging effect, namely, the fact that the "ordinary" or least squares estimator is often inadmissible for simultaneous estimation of several parameters. This effect has been called Stein's phenomenon, and has been demonstrated for several different problem settings, some of which are briefly outlined below.

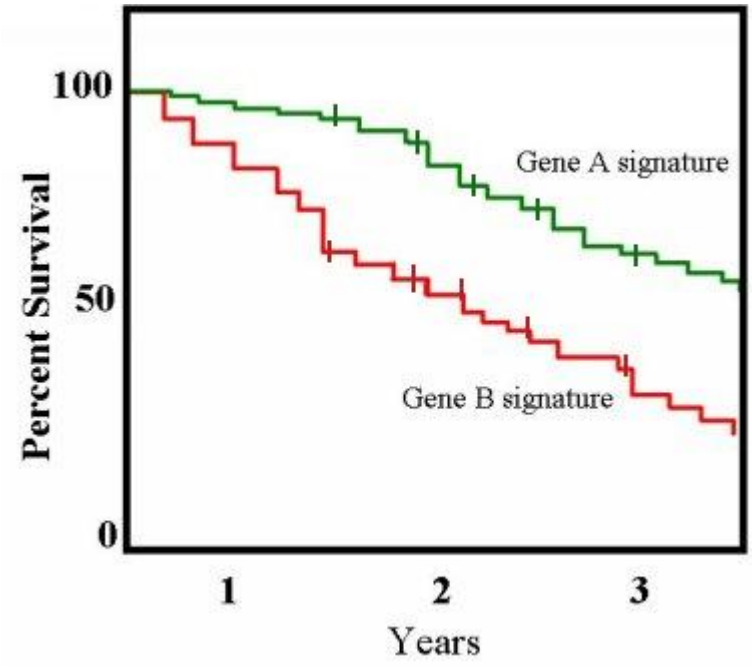
- James and Stein demonstrated that the estimator presented above can still be used when the variance σ^2 is unknown, by replacing it with the standard estimator of the variance, $\widehat{\sigma}^2 = \frac{1}{n}\sum(y_i - \overline{y})^2$. The dominance result still holds under the same condition, namely, $m > 2$.
- Bock extended the work of James and Stein to the case of a general measurement covariance matrix, i.e., where measurements may be statistically dependent and may have differing variances. A similar dominating estimator can be constructed, with a suitably generalized dominance condition. This can be used to construct a linear regression technique which outperforms the standard application of the LS estimator.
- Stein's result was substantially extended by Lawrence D. Brown to a wide class of distributions and loss functions. However, his theorem is an existence result only, in that explicit dominating estimators were not actually exhibited. It is quite difficult to obtain explicit estimators improving upon the usual estimator without specific restrictions on the underlying distributions.

Kaplan–Meier Estimator

The **Kaplan–Meier estimator** (named after Edward L. Kaplan and Paul Meier), also known as the **product limit estimator**, estimates the survival function from life-time data. In medical research, it might be used to measure the fraction of patients living for a certain amount of time after treatment. An economist might measure the length of time people remain unemployed after a job loss. An engineer might measure the time until

failure of machine parts. An ecologist may use it to estimate how long fleshy fruits remain on plants before they are removed by frugivores.

A plot of the Kaplan–Meier estimate of the survival function is a series of horizontal steps of declining magnitude which, when a large enough sample is taken, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations ("clicks") is assumed to be constant.



An example of a Kaplan–Meier plot for two conditions associated with patient survival

An important advantage of the Kaplan–Meier curve is that the method can take into account some types of censored data, particularly *right-censoring*, which occurs if a patient withdraws from a study, i.e. is lost from the sample before the final outcome is observed. On the plot, small vertical tick-marks indicate losses, where a patient's survival time has been right-censored. When no truncation or censoring occurs, the Kaplan–Meier curve is equivalent to the empirical distribution.

In medical statistics, a typical application might involve grouping patients into categories, for instance, those with Gene A profile and those with Gene B profile. In the graph, patients with Gene B die much more quickly than those with gene A. After two years about 80% of the Gene A patients still survive, but less than half of patients with Gene B.

Formulation

Let $S(t)$ be the probability that an item from a given population will have a lifetime exceeding t. For a sample from this population of size N let the observed times until death of N sample members be

$$t_1 \le t_2 \le t_3 \le \cdots \le t_N.$$

Corresponding to each t_i is n_i, the number "at risk" just prior to time t_i, and d_i, the number of deaths at time t_i.

Note that the intervals between each time typically will not be uniform. For example, a small data set might begin with 10 cases, have a death at Day 3, a loss (censored case) at Day 9, and another death at Day 11. Then we have ($t_1 = 3$, $t_2 = 11$), ($n_1 = 10$, $n_2 = 8$), and ($d_1 = 1$, $d_2 = 1$).

The Kaplan–Meier estimator is the nonparametric maximum likelihood estimate of $S(t)$. It is a product of the form

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}.$$

When there is no censoring, n_i is just the number of survivors just prior to time t_i. With censoring, n_i is the number of survivors less the number of losses (censored cases). It is only those surviving cases that are still being observed (have not yet been censored) that are "at risk" of an (observed) death.

There is an alternative definition that is sometimes used, namely

$$\hat{S}(t) = \prod_{t_i\le t} \frac{n_i - d_i}{n_i}.$$

The two definitions differ only at the observed event times. The latter definition is right-continuous whereas the former definition is left-continuous.

Let T be the random variable that measures the time of failure and let $F(t)$ be its cumulative distribution function. Note that

$$S(t) = P[T > t] = 1 - P[T \le t] = 1 - F(t).$$

Consequently, the right-continuous definition of $\hat{S}(t)$ may be preferred in order to make the estimate compatible with a right-continuous estimate of $F(t)$.

Statistical considerations

The Kaplan–Meier estimator is a statistic, and several estimators are used to approximate its variance. One of the most common such estimators is Greenwood's formula:

$$\widehat{\operatorname{Var}}(\widehat{S}(t)) = \widehat{S}(t)^2 \sum_{t_i<t} \frac{d_i}{n_i(n_i - d_i)}.$$

In some cases, one may wish to compare different Kaplan–Meier curves. This may be done by several methods including:

- The log rank test
- The Cox proportional hazards test