Prossinger, H., Schmehl, S. & Oberzaucher, E. (2019). Statistical Analysis of Gesture Encoding: How consistently can ethologists encode what they observe? Human Ethology, 34, 173-193. https://doi.org/10.22330/he/34/173-193

Human Ethology 34(2019): 173-193
Research Article

STATISTICAL ANALYSIS OF GESTURE ENCODING: HOW CONSISTENTLY CAN ETHOLOGISTS ENCODE WHAT THEY OBSERVE?

Hermann Prossinger, Susanne Schmehl, Elisabeth Oberzaucher

Department for Evolutionary Anthropology, University of Vienna, Austria

hermann.prossinger@univie.ac.at

ABSTRACT

Whenever persons describe localized pain, they include gestures along with their verbal descriptions. The "Fascial Distortion Model" (Typaldos, 2006) customizes its pain therapy by trying to classify these gestures. Practitioners claim that different Fascial Distortion Model classes necessitate different therapies. Here we present a statistical analysis method to assess whether the practitioners' claims are tenable if statistical rigor is assumed.

Five encoders observed 10 videos, one for each potential client describing his or her pain while including gestures. We took an inventory of gesture category loadings using an ethological toolkit derived from the "Gesture Action Coding System".

The outcomes of these observations are strings of category loadings for each client. All possible category loadings make up a dictionary of gestures, where each string of encodings is a (gesture) word in a generalized sense. If five encoders observe a client, the list of words should be statistically close, one hypothesizes. Our first approach was to look at how many different words occur overall and observe whether the (gesture) word lists for each client are consistent among encoders.

Our statistical methodology uses Bayesian probabilities and is to be seen in a much wider context: how to analyze the behaviors that ethologists observe, categorize, and classify. As ethologists' fieldwork quite often involves observing behaviors, be they gestures, facial expressions, etc., their classification challenges are analogous to Fascial Distortion Model ones, albeit with different dictionaries. We show how the statistical methodology we present here can be used to build gesture dictionaries and thereby enhance the statistical reliability of ethologists' encoding systems, including, for example, those for facial expressions.

Keywords: *Dirichlet distribution, Beta distribution, Bayesian probability, Gesture encoding, Fascial Distortion Model, Jeffreys prior*

* *Paper presented at XXIV Biennial Conference of Human Ethology 2018, in Santiago, Chile.*

submitted: 15.Apr. 2019; revised: 02. Aug. 2019; accepted: 20.Aug. 2019

INTRODUCTORY COMMENTS

It is common for humans to accompany their verbal descriptions with gestures. FDM ("Fascial Distortion Model", Typaldos, 2006) practitioners use the sequence of gestures accompanying verbal descriptions of potential clients' pain descriptions in order to diagnose pain and then design relief therapies.

In this paper, we describe how we develop statistical approaches to assess the FDM practitioners' claims that gestures accompanying pain descriptions are unique enough (in some sense) to differentiate among different possible pain therapies.

Our data set consists of 10 videos of 10 different potential FDM therapy clients while they are describing their pain both verbally and with gestures. Each video was observed by five different encoders, who had been trained to observe the 3–6 loadings of eight categories (Table 1). Thus, every video was observed 5 times and we initially presumed that the 5 lists of observations for each client would be consistent in some statistical sense.

The encoders were anonymous to us authors, as were the sexes and ages of the potential clients whose gestures had been videoed. The second author had access to the data files that included sex (but not age). The third author coordinated and monitored the data collection. The first author received only the partitioned lists of gesture category loadings.

This paper focuses on statistical methodology of categorical variables of encoding observations; it presents results that clarify the understanding and interpretability of statistical findings. It therefore intentionally restricts itself to using primarily one category of gestures in the calculations. Summary outcomes, such as whether any of the 10 potential clients described pain in ways that suggest to the FDM practitioners how many classes are present will not be exhaustively detailed here.

GESTURES, CATEGORIES, LOADINGS

Data collection

The FDM practitioner listened to the prospective client's description and watched the accompanying gestures; the latter were videoed during the session. The video was made available to us after the client had interacted with the FDM practitioner.

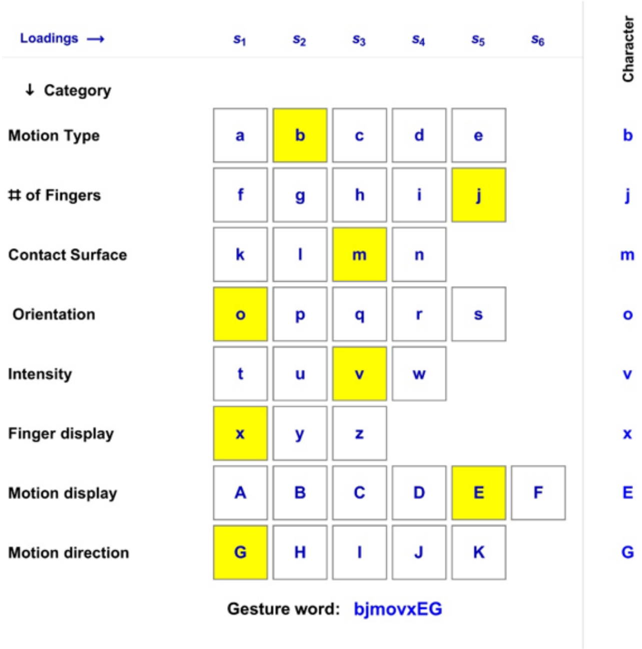
Categories and loadings

Table 1 shows the eight categories that the observers were to encode. Encoding took place by registering one loading per category. We note that the number of observed encodings can differ between encoders of the same video.

Because each of the five encoders observed each of the 10 videos once, all the observed loadings for one video will be distributed across the rows for each category (Fig. 1). We initially hypothesized that all, or at least most, of the loadings for each gesture of one category would occur in, one or at most two, cells of each category row. In Fig. 1, we show an example of one gesture that one encoder actually observed.

Table 1: The different loadings for each category. The loadings are categorical variables and have *no* numerical values associated with them.

Category	Loading	Category	Loading
Motion Type	s_1 stroking with body contact	*# of fingers*	s_1 one finger
	s_2 pressing w/o movement		s_2 two fingers
	s_3 showing w/o contact		s_3 three fingers
	s_4 grabbing (palm & fingers)		s_4 four fingers
	s_5 pinching (palm & fingers)		s_5 five fingers or fist
Contact surface	s_1 finger tip	*Orientation*	s_1 palm → pain position
	s_2 finger edge		s_2 back of hand → pain position
	s_3 hand surface (palm or back)		s_3 inner edge of hand
	s_4 fist		s_4 outer edge of hand
			s_5 fist or clenched hand
Intensity	s_1 without	*Finger display*	s_1 outstretched
	s_2 light		s_2 curved
	s_3 medium		s_3 completely bent
	s_4 strong		
Motion display	s_1 linear, smooth	*Motion direction*	s_1 vertical
	s_2 linear, intermittent		s_2 horizontal
	s_3 circularly, smooth		s_3 diagonal
	s_4 circularly, intermittent		s_4 circular
	s_5 pointed; no motion		
	s_6 pointed, intermittent		



Loadings ⟶	s_1	s_2	s_3	s_4	s_5	s_6	Character
↓ Category							
Motion Type	a	b	c	d	e		b
⌗ of Fingers	f	g	h	i	j		j
Contact Surface	k	l	m	n			m
Orientation	o	p	q	r	s		o
Intensity	t	u	v	w			v
Finger display	x	y	z				x
Motion display	A	B	C	D	E	F	E
Motion direction	G	H	I	J	K		G

Gesture word: bjmovxEG

Figure 1: One raster supplied by one encoder during observation of one gesture in one video. Each raster square highlighted in yellow shows which loading has been marked during the encoding process. When encoding the description of pain from one video, each encoder supplies a set of such rasters with yellow squares that will then be analyzed. The raster shows the gesture word 'bjmovxEG'.

Developing a statistical procedure

Each loading of each category is encoded by a categorical variable, encoded conveniently by letters drawn from some alphabet (here: case-sensitive Latin). Fig. 1 shows the labels of the categorical variables for all possible loadings for all categories. Choosing to encode the loadings using an alphabet prevents the fallacy of calculating averages and standard deviations, which do not exist for categorical variables.

For one gesture encoded by one encoder of one video, the marked loadings in the raster constitute a word. For five encoding sequences of all gesture sequences in one video by five encoders, five *lists* of words are the outcomes of the encodings of that pain video. Because the loadings have been encoded as letters of some alphabet, the pain description (i.e. the FDM encoding) is encoded as a list of gesture words. The gesture words themselves must therefore also be categorical variables.

The statistical tasks are: (a) to determine how the population of gesture words is distributed, (b) to calculate the likelihoods of the observed sequence of gesture word loadings, (c) to define closeness of different gesture words, and, finally, (d) to derive a procedure to extract a signal from the five sequences of gesture words per video.

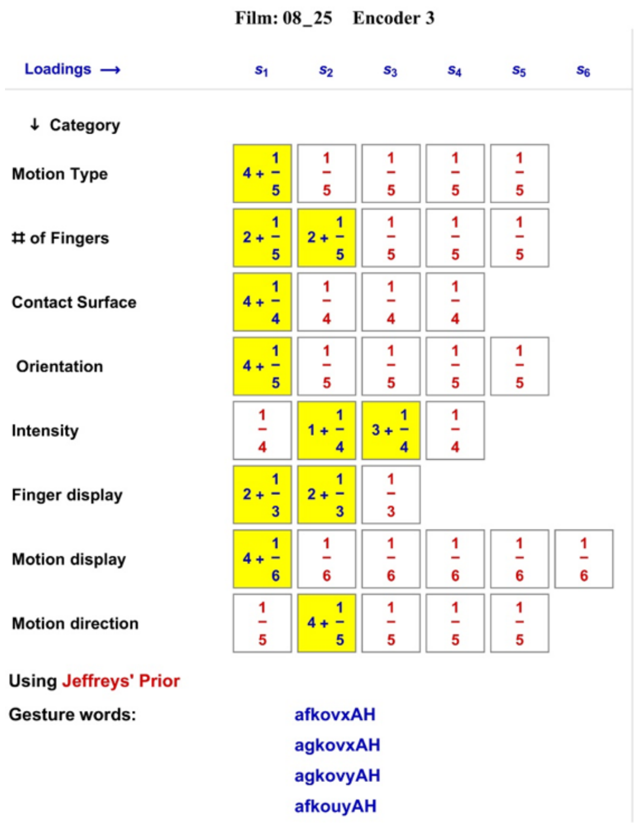
The gesture dictionary and observed gesture words

There are different possible encodings; therefore, potential clients could use 180000 different gesture words for their pain gestures. All these possible encodings make up a dictionary; the gesture dictionary therefore contains 180 thousand different words.

All possible words in the gesture dictionary could not have been observed by the encoders, because this would imply thousands of gestures per video.

Result #1: In this suite of 50 encoding sequences by five encoders encoding 10 videos, 271 gesture words have been encoded — roughly 5.4 gesture words per video. Result #2: In this suite of 50 encoding sequences, 124 different gesture words have been encoded.

All (different) gesture words encoded comprise the gesture dictionary for this suite of 10 pain descriptions. We emphasize that the analysis approach presented here is not restricted to the analyses of human gestures, nor is it restricted to visual gestures (acoustic gestures such as pitch and pronunciation can also be analyzed this way).



Film: 08_25 Encoder 3

Loadings ⟶ ↓ Category	s_1	s_2	s_3	s_4	s_5	s_6
Motion Type	$4+\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	
# of Fingers	$2+\frac{1}{5}$	$2+\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	
Contact Surface	$4+\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$		
Orientation	$4+\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	
Intensity	$\frac{1}{4}$	$1+\frac{1}{4}$	$3+\frac{1}{4}$	$\frac{1}{4}$		
Finger display	$2+\frac{1}{3}$	$2+\frac{1}{3}$	$\frac{1}{3}$			
Motion display	$4+\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
Motion direction	$\frac{1}{5}$	$4+\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	

Using Jeffreys' Prior

Gesture words:

afkovxAH
agkovxAH
agkovyAH
afkouyAH

Figure 2: The entries of all gestures observed by one encoder of one specific video including Jeffreys prior (Perks, 1947). For each category, the sum of observed loadings is a constant (namely, 4). For four categories the same loading was always observed. For others, such as '# of fingers', the encoder saw one loading twice and another loading twice. Using categorical variable encoding, four different gesture words were registered by this encoder for this client. The numerical entries for Jeffreys prior have been included and are explained in the text.

These 124 gesture words may or may not have been observed by all encoders or in all videos, but we do know from Result #1 and Result #2 that some gesture words have been observed more than once (Fig. 2).

This encoding analysis is reminiscent of the "crossword puzzle problem": which words must be included in a dictionary to successfully construct a crossword puzzle?

CONJUGATE DISTRIBUTIONS

Comparing frequentist (orthodox) statistics with Bayesian statistics

In frequentist statistics, the probability p is a number (a 'point estimator') achieved by observing the ratio of looked for ('favorable') outcomes versus all outcomes, as the number of observations $N \rightarrow \infty$.

$$p = \frac{N_{\text{favorable}}}{N}$$

This probability is sometimes called a Laplace probability, or more specifically, the Laplace paradigm for probability.

In Bayesian statistics, the probability is a random variable s with a probability density function conveniently called its likelihood function (Krushke, 2015), which is defined over the domain $\mathbb{D} = [0,1] \subset \mathbb{R}$. For any number of observations, the mode of s will be the *most likely* (ML) *probability* (s_{ML}) for the event to be observed. The Bayesian probability is sometimes called 'a belief' (MacKay, 2015) and the ML probability would then be the most likely belief (conveniently called the *conviction*; MacKay, 2015). We note that these words, common in psychology research, are thereby statistically quantifiable.

Frequentist statistics using the Laplace probability paradigm become impossible when (a) the number of observations is small and (b) the underlying process that generates the outcome is not constant during the observations.

If, for example, a weather forecast predicts "78% probability of rain at 21:18 local time," the predicted probability is not the result of repeated observations. Weather phenomena are chaotic; therefore, observation *repetitions* (in contrast to: many observations) can *never* exist. The predicted probability of a weather phenomenon is therefore a most likely probability, calculated using Bayesian statistics. The predicted weather probability is never a Laplace probability.

In biological and psychological systems, observation repetitions with a constant generating process are also rare, even when they are not the outcome of a chaotic process. There could — and indeed oftentimes does — exist an underlying drift or a discontinuity during a sequence of observations. Hence, frequentist statistics is rarely appropriate for biological or psychological phenomena. An example of a situation in which the use of frequentist statistics is appropriate is the measurement of the mass and/or diameter of the millions of eggs released by a female oyster at one event.

Bayesian statistics for Bernoulli trials

Assuming we have a sequence of events in which a list of binary outcomes is registered, these outcomes are called Bernoulli trial outcomes (Table 2).

Table 2: Eight examples of Bernoulli trial outcomes. Each row is an example of a Bernoulli trial. The rows may not be combined. The trial in the second row is for mammals, not birds. Last row: testing for the distribution of (human) twins in a population. The heading shows the parameters of the Beta distribution; they are to be estimated using the data and Bayesian methods.

$\mathcal{B}e(\alpha,\beta)$	
□ yes	□ no
□ ♂ (XY)	□ ♀ (XX)
□ guilty	□ innocent
□ tooth present	□ tooth absent
□ dead	□ alive
□ pass	□ fail
□ score a goal	□ not score a goal
□ heterozygous	□ homozygous

Because, in Bayesian statistics, the probability of success *s* is a random variable, the probability of failure (1-*s*) is likewise a random variable.

The (conjugate) distribution of a (statistical) population that meets the above conditions (Bayesian statistics and Bernoulli trial) is the Beta Distribution (Bishop 2006).

Consider the distribution of sexes (not genders, as gender attribution is not the outcome of a Bernoulli trial; Table 4) of clients visiting an FDM practitioner. There are eight female and two male clients who show the same gesture category loading. What is the ML probability that a female showed this gesture category loading?

Applying Bayes' Theorem for a 1st observation of a female (we assume the probability is *s* for observing a female and 1 – *s* for observing a male)

$$\mathcal{L}_{post} = data \times \mathcal{L}_{prior}$$

leads to

$$\mathcal{L}_{\text{post}} = \text{constant} \times s^{1} \times \mathcal{L}_{\text{prior}}$$

with some to-be-determined constant. The sequence of 10 = 8 + 2 observations (perhaps in random order) in this example results in

$$\mathcal{L}_{\text{post}} = \text{constant} \times s^{8} \times (1 - s)^{2} \times \mathcal{L}_{\text{prior}}$$

In general, the posterior likelihood $\mathcal{L}_{\text{post}}$ is

$$\mathcal{L}_{\text{post}} = pdf(\mathcal{B}e(\alpha,\beta), s) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} s^{\alpha-1}(1-s)^{\beta-1}\mathcal{L}_{\text{prior}}$$

where $\Gamma(k)$ is the Gamma function, a generalization of the factorial $k!$; where $\Gamma(k) = (k - 1)!$ and k is not restricted to integer k (Abramowitz & Stegun, 1972).

The often-used Bayesian prior is $\mathcal{L}_{\text{prior}} = 1$. As we clarify below, this prior is not statistically 'clean' for our example, because a Bayesian prior infers prior information.

In almost all statistical analyses, one may not assume that the prior outcome of a trial will be observed. Phrased more rigorously: we need a prior likelihood $\mathcal{L}_{\text{prior}}$ that does not include prior information. Jeffreys (1932) derived this prior, using Fisher's information matrix. In this example, it could be that no client of either sex show the pain gesture loading. The *Jeffreys prior* ensures that no prior information is available prior to the first observation.

The Jeffreys prior for a Bernoulli trial is

$$\mathcal{L}_{\text{prior}} = \frac{\Gamma(\frac{1}{2}+\frac{1}{2})}{\Gamma(\frac{1}{2})\cdot\Gamma(\frac{1}{2})} s^{\frac{1}{2}-1}(1-s)^{\frac{1}{2}-1} = \frac{1}{\pi\sqrt{s}\sqrt{(1-s)}}$$

Using Jeffreys prior, the inferred posterior prior $\mathcal{L}_{\text{post}}$ of the 10 clients (eight females and two males) showing this gesture category loading is

$$\mathcal{L} = pdf(\mathcal{Be}(8+1+\tfrac{1}{2}, 2+1+\tfrac{1}{2}), s) = \frac{\Gamma(13)}{\Gamma(9\frac{1}{2})\Gamma(3\frac{1}{2})} s^{8\frac{1}{2}}(1-s)^{2\frac{1}{2}}$$

Fig. 3 shows the graph of this posterior likelihood (i.e. including Jeffreys prior) along with the most likely probability s_{ML}. The exponents for Jeffreys prior for each category for four gesture words (Fig. 2) are the fractions in that graph (Perks, 1947).

One superiority of Bayesian statistics (in addition to the possibility of defining probabilities for very small sample sizes) is the possibility of calculating probability uncertainties. These can be defined by quantile intervals about the mode. The statistician and the researcher using the results are required to make a choice about how the quantiles are defined, because a standard deviation is not meaningfully defined for a Beta function. Below, we show one way of estimating the uncertainty interval. We (again) stress that the probability is along the horizontal axis in Fig. 3, while the probability density function (*pdf*) is the likelihood function of *s*.

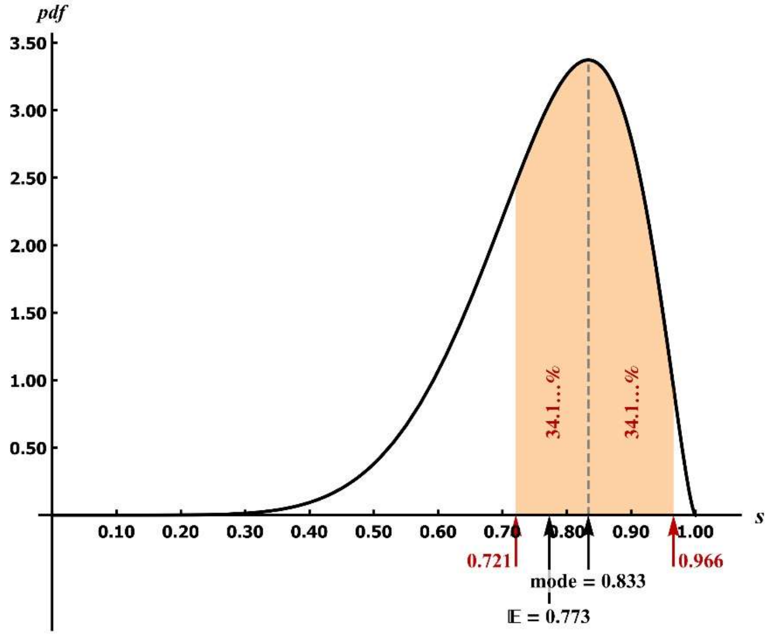


Figure 3: The likelihood function (*pdf*) and the ±34.1...% uncertainties with the resulting confidence interval for the probability *s* of observing a further female showing the pain gesture, using Jeffreys prior. The mode is s_{ML} = 0.833... and the ±34.1% confidence interval is $\begin{matrix} +0.721 \\ +0.966 \end{matrix}$ or $0.721 \leq s \leq 0.966$.

We observe that the estimation using the Laplace probability (alas often used, even though a Laplace estimation is not feasible for such a small sample) is $p_{\mathrm{Laplace}} = \dfrac{8}{8+2} = 0.800\ldots$ This Laplace probability is to be rejected. Note that s_{ML} is the mode, not the expectation value $\mathbb{E}$.

Bayesian statistics for multinoulli trials

In many situations the response or observed outcome is not that of a Bernoulli trial. In such situations, the name we use is 'multinoulli' trial (MacKay, 2015). Table 3 lists seven examples of multinoulli trials with three categorical outcomes and Table 4 lists four examples for multinoulli trials with five categorical outcomes.

Table 3: Some examples of multinoulli trials with three outcomes. If one of the outcomes is "other" — as in row 5 —, then this is an example of a marginalization, as described in the text. The heading shows the parameters of the Dirichlet distribution; they are to be estimated using the data and Bayesian methods.

$\mathcal{D}ir\,(\alpha_1,\alpha_2,\alpha_3)$		
□ win	□ loss	□ tie
□ yes	□ no	□ abstain
□ bent	□ curved	□ outstretched
□ healthy	□ damaged	□ missing
□ heterosexual	□ homosexual	□ other
□ married	□ divorced	□ widowed
□ healthy	□ sick	□ dead

Table 4: Some examples of multinoulli trials with five outcomes. The first example can be considered typical in some gender research polls. The second example is a typical multinoulli trial for a questionnaire concerning customer feedback. The third are possible responses to questions in examinations or on tests (including a marginalization). The fourth example lists the loadings of the category "# of fingers" in the gesture word inventory (Table 1). The heading shows the parameters of the Dirichlet distribution; they are to be estimated using the data and Bayesian methods.

$\mathcal{D}ir\,(\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5)$				
□ L(esbian)	□ G(ay)	□ B(isexual)	□ T(ransgender)	□ Q(ueer)
□ very dissatisfied	□ dissatisfied	□ neutral	□ satisfied	□ very satisfied
□ A correct	□ B correct	□ C correct	□ D correct	□ none are correct
□ one finger	□ two fingers	□ three fingers	□ four fingers	□ five fingers or fist

The (conjugate) distribution of a (statistical) population that meets the two conditions — Bayesian statistics and multinoulli trial — is the Dirichlet Distribution (Bishop, 2006). Dirichlet distributions can be used for an arbitrary number of categorical variables. The Dirichlet distribution is a generalization of the Beta distribution for a multinoulli trial with n_1, n_2, n_3, n_K different outcomes. We note that the n_K occurrences for the K^{th} outcome are defined via the other $(n - (n_1 + n_2 + \ldots + n_{K-1}))$ ones. More stringently: the (Bayesian) probabilities s_1, s_2, s_3, … , $s_{K\text{-}1}$ define the probability s_K because $s_K = (1 - (s_1 + s_2 + s_3 + . + s_{K-1}))$…

Bayes' Theorem for a multinoulli trial with K outcomes is

$$\mathcal{L}_{\text{post}} = \mathit{pdf}\ (\mathcal{D}ir(\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_K), \mathrm{s}_1, \mathrm{s}_2, \mathrm{s}_3, \ldots, \mathrm{s}_K)\ \mathcal{L}_{\text{prior}}$$

where we have used the symbol $\mathcal{D}ir\ldots$ for a Dirichlet distribution.

If there are three possible outcomes

$$\mathcal{L}_{\text{post}} = \mathit{pdf}\ (\mathcal{D}ir(\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_K) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2) \cdot \Gamma(\alpha_3)} s_1^{\alpha_1 - 1} \cdot s_2^{\alpha_2 - 1} \cdot s_3^{\alpha_3 - 1}$$

Because these three outcomes are not independent, we can graph the 2D-surface of $\mathcal{L}_{\text{post}}$ for s_1 and s_2; noting that the domain is the triangle $s_2 = 1 - s_1 \wedge s_1 \in [0,1] \subset \mathbb{E}$ (Fig. 4).

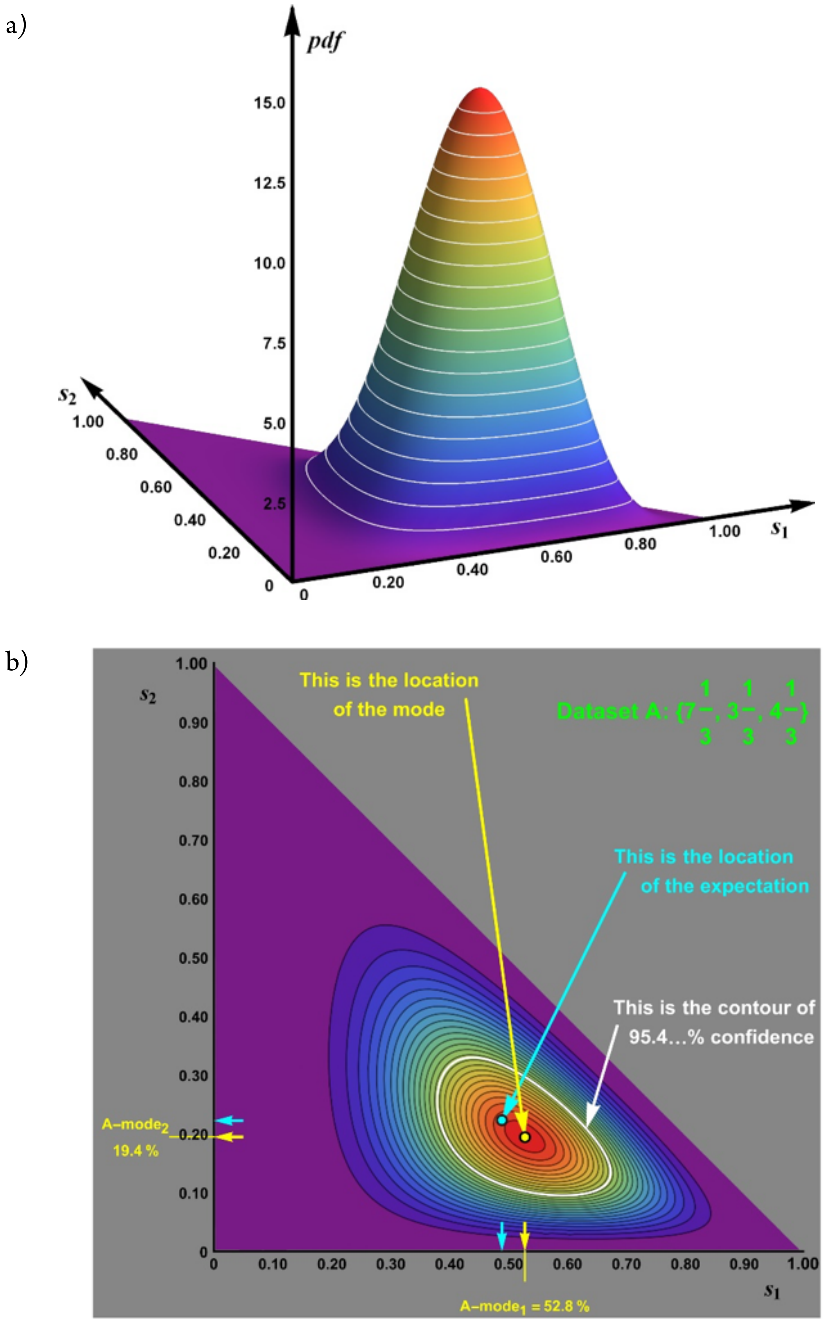


Figure 4: The *likelihood function* of a Dirichlet distribution using a Jeffreys prior for an encoding of a gesture category with loadings {7,3,4} and a Jeffreys prior $\{\frac{1}{3},\frac{1}{3},\frac{1}{3}\}$ (Perks, 1947). **(a)** The likelihood surface shown in 3D; the light gray contours show the 95%, 90%, … fractions of the maximum likelihood. **(b)** The projection of the likelihood surface onto the s_1-s_2 plane. The 95%, 90%, … fractions of the maximum likelihood are contours rendered in black; the 95.4…% confidence contour is rendered as a white contour. The values of the modes of the probability are rendered in yellow.

The curve of the *pdf*-surface on the s_1-*pdf* plane is not a projection, but rather the integral

$$\int_0^1 pdf(\mathcal{D}ir(a_1, a_2), s_1, s_2)ds_2$$

likewise for s_1. Each of these integrations is called a marginalization. Marginalization reduces the number of parameters for any multi-parameter distribution. If, for a Dirichlet distribution, all parameters except two are absorbed through marginalization, then the resultant distribution is a Beta distribution; a Beta Distribution is the special case of a Dirichlet distribution with only two parameters.

AN OVERVIEW OF DISTRIBUTIONS OF CATEGORY LOADINGS

Histograms

The histogram (Fig. 5) shows how many loadings different encoders observed of the same category for one client's pain description. It is conspicuous (and is also reason for the statistical challenges that are dealt with in this paper) that there is no obvious consistency of category loadings observed. The challenge includes, of course, whether the differences in gesture words are *significantly* different.

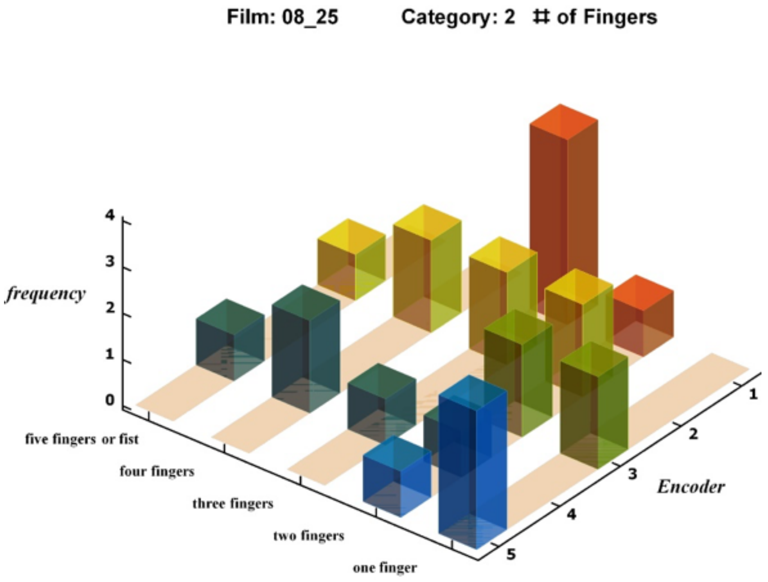


Figure 5: The histogram of loadings by the five different encoders of category$_2$ for one specific video. We note that the number of fingers observed by different encoders varies considerably. The statistical challenge is to determine whether the differences in observation outcome(s) are significantly different.

Some specific issues evident in Fig. 5:

(a) Encoder$_5$, encoder$_3$ and encoder$_1$ observed 'only' two different loadings, while encoder$_2$ and encoder$_4$ observed four different loadings, with different frequencies.

(b) Encoder$_5$ observed four loadings (as did encoder$_3$), while encoder$_1$ observed five loadings, as did encoder$_4$.

(c) Encoder$_2$ observed seven loadings; and four different ones.

The issues listed imply that the loadings observed vary considerably. In this paper, we do not address the issue of reliability of the encoders' skills (an issue often addressed so as to discard some encodings). Rather, we consider all five encoders equally qualified and we take their observations (encodings) equally seriously. We use no weightings to model preferences of one encoder (or a select few encoders) over the others. Our approach, therefore, implies one statistical challenge: whether the different frequencies of encodings are significantly different. As we demonstrate below, some of them are, but not all.

Marginalization leading to a Beta Distribution

Fig. 5 shows that only the loadings s_2 and s_3 were registered by encoder$_1$. We can therefore integrate

$$\int_0^1 \int_0^1 \int_0^1 pdf(\mathcal{D}ir(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5), s_1, s_2, s_3, s_4, s_5) ds_1 ds_4 ds_5$$

to obtain the likelihood function of a Beta Distribution $pdf(\mathcal{B}e(\alpha, \beta), s)$. Fig. 6 shows the likelihood function of this distribution, along with the expected value ($\mathbb{E} \approx 0.700$), the mode ($s_{\mathrm{ML}} \approx 0.833$) and the confidence interval ($0.601 \leqq s \leqq 1$).

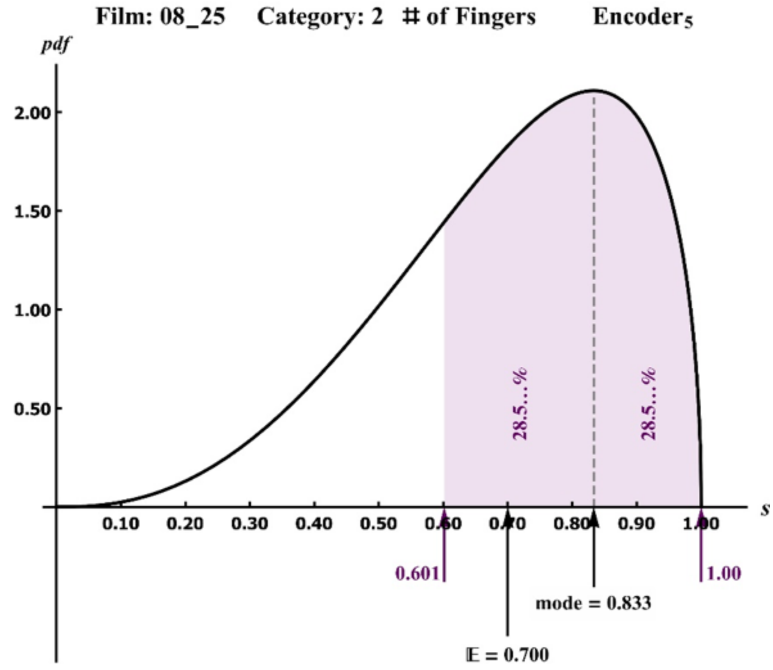


Figure 6: The likelihood function of the probability of observing three fingers (loading$_3$) by encoder$_1$, including Jeffreys prior. The likelihood function is a Beta distribution, and the Bernoulli trial is either loading$_3$ (three fingers) or loading$_2$ (two fingers) (Fig. 5). The areas are the ±28.5…% probability quantiles (the maximum permissible areas, because the upper confidence limit cannot exceed 1). The uncertainty interval about the mode is asymmetric. Observe that the most likely probability (s_{ML} … the mode) is considerably different from the expectation value.

Marginalization leading to a multinoulli trial

Fig. 5 shows that encoder$_2$ and encoder$_4$ have comparable histograms. If we assume the loadings differ due to chance, we need to determine whether the differences are significant. They are not if the uncertainty contours overlap at a predefined significance level. We define 95.4 % confidence contours (Fig. 4b). Furthermore, we marginalize out loading$_1$ (no observation) and loading$_5$ (the smaller of the equal counts). The result is shown in Fig. 7. We observe that the contours overlap, so the differences in the two ML probabilities are insignificant. Not all information is discarded via marginalization: the contours remain considerably extended.

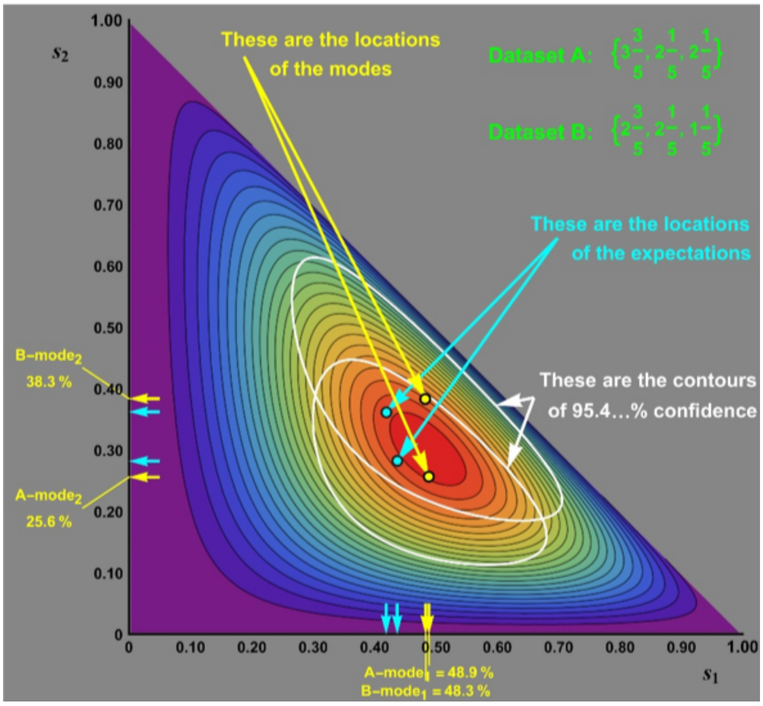


Figure 7: The 95.4…% confidence contours for the loadings observed by encoder$_2$ (Dataset A) and by encoder$_4$ (Dataset B) of category$_2$ ('# of fingers') including Jeffreys prior. The overlap of the contours shows that the observed differences in loading frequencies are not significantly different. The contours of the sum of the likelihood functions are rendered in dark blue. The ML probabilities for mode$_1$ hardly differ for the two observers — a fact that is not evident when assessing the histogram (Fig. 5). ML probabilities for mode$_2$ differ considerably (difference ~12.7 %), but not significantly. The projections of the contour extremes onto each of the axes are not the uncertainty intervals of the respective probabilities; the intervals on each axis must be obtained by marginalization (i.e. integration over all other loading probabilities as described in the text).

Uncertainties

In order to evaluate the consistency of the encodings by the five different observers, it is insufficient to compare modes; rather, uncertainties must also be compared. Because of marginalization it is not possible to use uncertainty intervals, unless marginalization results in a Beta distribution (as in Fig. 6). The (white) contours in Fig. 7 are the uncertainties for a case of two modes. Another possibility when comparing uncertainties is to use the square root of the variance $\sqrt{var}$ after marginalization. However, it is fallacious to calculate the uncertainty about the mode using $\sqrt{var}$, even in the case of a Beta distribution, because the uncertainty quantiles are not due to equal integrals, as demonstrated in Fig. 6.

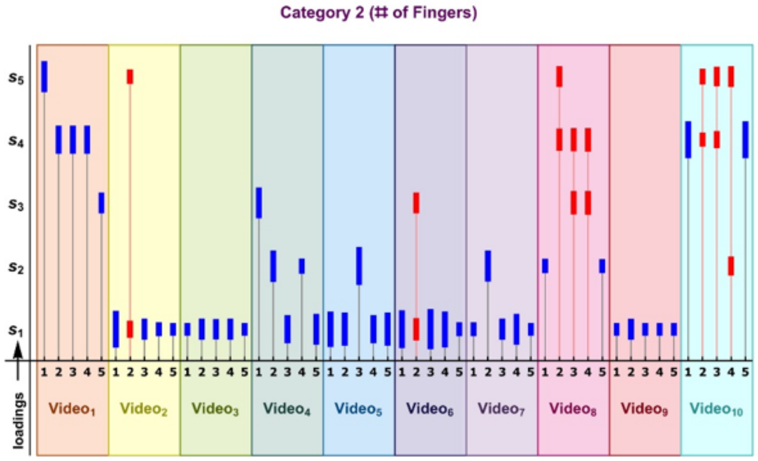


Figure 8: The square roots of the variances $\sqrt{var}$ (which are linear measures) of the observations of the five encoders of the ten clients. The blue bars are those $\sqrt{var}$ for which marginalization resulted in one dominant mode; the red bars for those where marginalization resulted in two modes. The vertical axis does not have a numerical scale. Comparison of the lengths of the bars ($\sqrt{var}$) is only permitted for a specific loading; for example, in Video$_3$, all encoders observed only one mode (for loading s_1) after marginalization, but the $\sqrt{var}$ differed among different encoders. The graph (indirectly) also shows for which loading the mode was most likely — if there was only one mode — or for which loadings two modes were most likely, albeit with different likelihoods. The graph shows that no marginalization resulted in three modes. Only the encodings of category$_2$ for Video$_3$ and for Video$_9$ are remarkably consistent.

Table 5: Patterns in the modes and uncertainties (estimated by $\sqrt{var}$), extracted from Fig. 8. The most likely mode for all 10 clients is pointing with one finger, but there are very many exceptions. The uncertainties quantify the reliability (or lack thereof) of the encoders.

Video No.	After Marginalization			
	1 mode (majority)	1 mode (others)	2 modes	$\sqrt{var}$ comparable?
1	$3 \times s_4$	$1 \times s_3$ $5 \times s_5$		all; except $\sqrt{var_3}$
2	$4 \times s_1$		s_1 s_5	all; except $\sqrt{var_1}$
3	$5 \times s_1$			all
4	$2 \times s_1$ $2 \times s_3$	s_3		all; except $1 \times \sqrt{var_2}$
5	$4 \times s_1$	s_2		all
6	$5 \times s_1$		s_1 s_3	$3 \times \sqrt{var_1}$ $2 \times \sqrt{var_1}$ $1 \times \sqrt{var_3}$
7	$4 \times s_1$	$1 \times s_2$		$3 \times \sqrt{var_1}$ $1 \times \sqrt{var_1}$ $1 \times \sqrt{var_3}$
8		$2 \times s_2$	$2 \times s_{3,} 2 \times s_4$ $1 \times s_{4,} 1 \times s_5$	all; except $2 \times \sqrt{var_2}$
9	$5 \times s_1$			all
10	$2 \times s_4$		$1 \times s_{2,} 1 \times s_4$ $2 \times s_{5,} 2 \times s_5$	all; except $2 \times \sqrt{var_4}$

The patterns in Table 5 constitute one part of a conclusion. They summarize the results for one category. Here, we do not present the results for the other seven categories. Clearly, it is to be expected that the distribution of loading modes and $\sqrt{var}$ for each of the observed loading modes will sometimes differ wildly (as for Video$_8$ and Video$_{10}$) and sometimes be highly consistent (as for Video$_3$ and Video$_9$).

ML loading probabilities

In the previous section, we addressed the question whether the modes for various loadings roughly agreed or strongly disagreed. We also need to assess how large (in magnitude) the ML loadings are: the likelihoods of the respective modes.

In Fig. 9, we present such an analysis and outcome for Video$_{10}$ — one of the two statistically most challenging cases (Fig. 8 and Table 7).

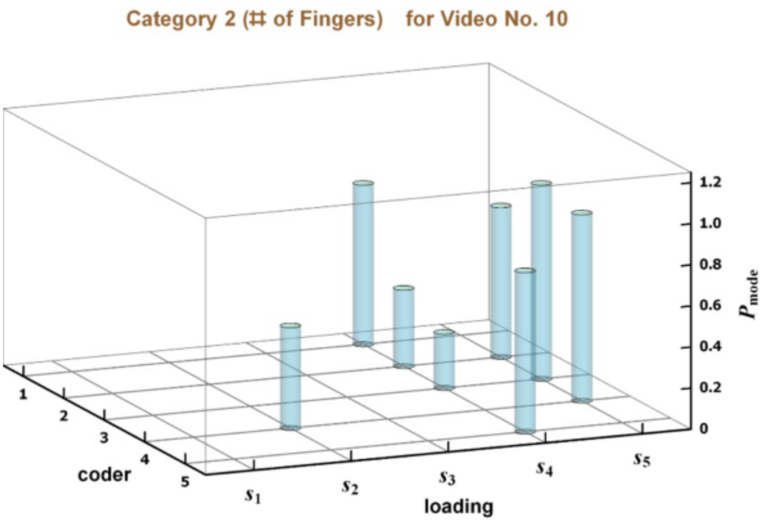


Figure 8: The modes of the loadings (ML probabilities) graphed for the five encoders for one category$_2$ observed in Video$_{10}$. We note that two encoders observed only one mode, three encoders only two modes, and none more than two. However, only two encoders observed the same loadings, albeit with different $pdf(s_{\text{ML}})$. In summary, none of the encoders observed rigorously consistent observation statistics for this video.

We note that the *likelihoods* of five modes are almost equal, while the likelihoods of three (much smaller) modes differ considerably.

In detail: we observe that (a) only two encoders observed a single mode for one loading, with a likelihood $P_{\text{mode}} \approx 80\%$ (consequently, these two encoders did not observe only one single loading in this video, but both predominantly observed the loading s_4) and (b) encoder$_2$, encoder$_3$ and encoder$_4$ observed so many loadings that these were Dirichlet-distributed (furthermore, the many ML P_{mode} were not the same for these encoders).

SUMMARY

Conclusions about statistical methodology

The number of gesture words observed by different encoders in the same video varied. Small variations in encoding frequencies resulted in well-defined modes of one loading and one small variance — whenever the encoders observed such a situation. Quite often, small variations in loadings did not occur; then the marginalization resulted in more than one mode. We needed to make inventories of modes and variances, because observing and encoding some clients' gestures sequences were statistically very noisy (either because of the difficulties of observing the gestures or their ambiguities).

The longer the video, the greater were the variances of the loadings. Because this paper is about statistical methods needed for categorical variables, we did not include the durations of the videos of clients describing their pain. However, none of the videos were short. Some clients used only a few gestures, some very many, and not all encoders agreed about this fact.

When comparing encodings of gestures by different encoders, we do not observe any consensus of gestures occurring. However, this statement is too broad — it cannot, as stated, be considered a conclusion. To sharpen it, we require the significance of any differences between frequencies of gesture word occurrences to be calculated. Assessing the necessary significance levels employed requires a justification for the percentages used to calculate quantiles of (marginalized) Beta distributions or contours of (marginalized) 3-parameter Dirichlet distributions. We always succeeded in marginalizing to these two options. Nonetheless, we cannot expect this to be possible for other behavior studies.

Although we have only demonstrated marginalization for the encodings for category$_2$ ('# of fingers'), it is generally true (at least for this data set and for these encoders) that marginalization of the Dirichlet distributions is necessary (because of the high variability of loadings) to interpret the ML loadings for each category. We can conclude that either the encoders had great difficulty encoding consistently or the gestures of some clients are not adequately evident for the encoder to observe them clearly. For example, if one encoder observes the loading s_3 ('three fingers'), another encoder might observe s_4 ('four fingers'), because it may seem to another encoder that the fourth finger is actually being used during the pointing. This issue of ambiguity is omnipresent, yet does not appear to be adequately addressed in the ethology literature. It may have often been overlooked in published presentations, perhaps because careful statistical analysis (a) using Dirichlet distributions of categorical variables, (b) using maximum likelihood methods, (c) including Jeffreys' prior, and (d) marginalizing very infrequent registrations had not been employed.

Other methods of analyzing these noisy data sets are available in a statistician's toolkit: Singular Value Decomposition (SVD) and Correspondence Analysis (CA), but both are not applicable in this case. SVD of the frequencies, as in Fig. 5, will hardly be useful for such low frequencies (of which many are zero). Likewise, using CA to determine associations of frequencies (as in Fig. 5) will also be unsuccessful for the same reason. Neither SVD nor CA can be used across all categories, because the number of category loadings vary (Fig. 1) and no matrix can be constructed.

Inferences for statistical analysis of behaviors

The statistics of observed behavior published by ethologists must be considered tenuous, unless the methods presented here have been rigorously applied. To do so, more than one encoder is imperative. If Bayesian statistics are used, small numbers of encoders can be employed, where the uncertainties will depend on the number of encoders.

Consider the case of expert ethologists who 'train' future observers until 'all' trainees observe what the trained ethologist observed: this approach must be considered flawed. For one, statisticians cannot assume that the observations made by the trainer have the optimal number of modes of loadings and minimal variances (the 'golden standard'). Even if one were to assume that the observations made by the trainer are 'better' (in some sense), then the trainer would have to perform the statistical analyses presented here in order to quantify the trainees' observational skills. It is better, we argue, to require initial training and thereafter analyze the statistical outcomes made by all encoders. Marginalization would presumably incorporate a large fraction of differing observations. If, after marginalization, significant differences remain: so be it! We reject the view that there exists a 'golden standard' that is to be considered a successful sequence of observations. Perhaps observations and encodings by humans are simply always very noisy. We insist that more than one observer encode. With video and other recording equipment being ubiquitous, repeated viewings can be employed.

The large variances and the necessary marginalization of loadings imply that AI (Artificial Intelligence) methods are much more promising than the analytical statistics presented here. In the case of AI algorithms, training with a subset of the observations and 'bagging' with the remainder sounds promising. But, of course, this necessitates very large data sets (in the order of ~1000). Because ethologists would presumably have difficulties setting up such an observation program, the methods introduced here will most likely (pun intended!) not become obsolete in the near future.

ACKNOWLEDGEMENTS

We thank the encoders for their painstaking observations.

REFERENCES

Abramowitz, M. & Stegun, I. A. (Eds) (1972). "Gamma (Factorial) Function" §6.1 in *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing*. New York: Dover, pp 255–258.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer Science + Business Media, New York, NY, USA. DOI

Jeffreys, H. (1932). On the Theory of Errors and Least Squares. *Proc. Roy. Soc. 138*, 48–55.

Krushke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS and Stan*. Academic Press & Elsevier, London, UK. DOI

MacKay, D.J.C (2015). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK.

Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries* 73, 285–334. DOI

Typaldos, S. (2006). *Orthopädische Medizin. Die Verbindung von Orthopdädie und Osteopathie durch das Fasziendistorsionmodell*. European FDM Association.