Statistical analysis of chroma features in western music predicts human judgments of tonality*

Joan Serrà, Emilia Gómez, Perfecto Herrera, and Xavier Serra
Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
{joan.serraj,emilia.gomez,perfecto.herrera,xavier.serra}@upf.edu

Nov, 2008

Abstract

Motivated by evidence that image source statistics predict the response properties of several visual perception aspects, we provide an empirical analysis of the relation between chroma statistics and human judgments of tonality. To accomplish this, a statistical analysis method based on chroma feature covariance is proposed. It makes use of a large collection of western music to build a tonal profile. The obtained profile is compared to alternative tonal profiles proposed in the literature, either cognitively, perceptually, or theoretically inspired. The high degree of correlation we find between the covariance-based tonal profile proposed here and several ones proposed in the literature (reaching values higher than 0.9) is interpreted as evidence that human-derived profiles faithfully reflect the statistics of the musical input listeners have been exposed to. Furthermore, we show that very short time scales allow us to correctly predict these profiles, which brings us to discuss the role that local-scale implicit learning plays in building mental representations of tonality.

1 Introduction

Natural environments contain highly structured stimuli to which we are exposed in everyday life. The human brain internalizes these regularities by exposure, and the acquired implicit knowledge influences perception and performance. In human vision, plenty of research has been devoted to understand why the physical source of a retinal image cannot be directly derived from the stimulus information. This quandary is usually referred to as the *inverse optics* problem (Purves & Lotto, 2003). Nevertheless, evidence that image source statistics predict several aspects of visual perception exists (Simoncelli, 2003; Doi & Lewicki, 2005; Keil, 2008). This has motivated the study of statistical covariations in the spectral information that humans processed as typical visual stimuli. In fact, it has been demonstrated that cumulative density functions derived from these covariations accurately predict major colorimetric functions such as hue, saturation, and brightness (Long, Yang, & Purves, 2006). In audition, a similar *inverse acoustics* problem arises: the physical sources of sound stimuli are not uniquely specified by the sound pressure variations at the receptor surface. In this domain, statistical properties of naturally occurring sounds have been analyzed as well, providing valuable results. For instance, evidence that similarity of musical scales and consonance judgments across humans (with no explanation broadly accepted) arise from the statistical structure of naturally occurring periodic sound stimuli has been reported (Schwartz, Howe, & Purves, 2003).

Rogers and McClelland (2004) have studied, by means of specific neural network architectures, the role of covariance among stimulus dimensions on the process of learning new categories. According to the existing evidence, infants treat objects as being of the same kind because they are sensitive to patterns of experienced coherent covariation of properties across objects. This proposal goes beyond simple pairwise correlations and insists on the possibility that such a bias for the rapid learning of certain correlations over others can itself arise from higher-order patterns of coherent covariation among stimulus properties. This way, the properties that first become useful, informative, or salient to infants, and that are easiest to associate with one another, are just those that participate in the strongest patterns of coherent covariation across many different events and situations. The sensitivity of learning mechanisms to coherent covariation among stimulus properties provides a mechanism by which such domain-general perceptual learning can give rise to internal representations that capture similarity structure and modulate the acquisition of feature salience.

As it happens with the early acquisition of language categories, the early acquisition of music categories does not seem to involve explicit category labels, explicit response trials, or systematic explicit feedback provided by a teacher (Patel, 2008). Category information accompanies category instances during the acquisition process. Thus, it involves complex correlations (or covariations) between acoustic event sequences and various visual, auditory, olfactory, tactile, and other events occurring in the environment. Additionally, it also involves a complex pattern of co-occurrences between the acoustic and musical features themselves. This acquisition process is usually referred to as *statistical learning* or *implicit learning*. While it is far from clear how infants make use of this torrent of covariations, it is known that humans, infants and adults, are sensitive to statistical regularities at multiple levels (Saffran, Aslin, & Newport, 1996; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997; Saffran, Johnson, Aslin, & Newport, 1999), and that statistical learning of auditory stimulus seems to be stronger than visual or tactile learning, specially in sequential problems (Conway & Christiansen, 2005).

In tonality cognition, many studies support the acquisition of tonal structure through implicit learning (e.g. Tillmann, Bharucha, & Bigand, 2000; Tillmann, 2005). As musically naive listeners are in everyday life exposed to the tonal regularities underlying the music of their culture, they acquire implicit knowledge of them (Bharucha, 1984; Huron, 2006). On the other hand, some studies have suggested that the distribution of tones in music is related to the tonal hierarchy of a given musical tradition (Castellano, Bharucha, & Krumhansl, 1984; Krumhansl, 1990). Therefore, it can be rated by listeners provided that they are familiar with this musical tradition. These findings have been exploited for tonality estimation based on human ratings (the probe tone technique originally proposed in Krumhansl & Shepard, 1979; Krumhansl & Kessler, 1982) and further extended to the analysis of audio signals (e.g. Gómez, 2006).

It is the objective of the present work to contribute to the abovementioned flow of knowledge by providing experimental evidence that statistical covariations of pitch information strongly correlate with human ratings on tonal hierarchy. To automatically determine tonal regularities over massive music collections, one can use the so-called chroma features as a source of information. These are routinely employed for a wide variety of music information retrieval (MIR) tasks (e.g. Bartsch & Wakefield, 2001; Müller, 2007; Ellis, 2007), but a primary application for them is in automatic tonal analysis, including key determination and chord recognition (e.g. Fujishima, 1999; Sheh & Ellis, 2003; Paws, 2004; Gómez & Herrera, 2004; Purwins, Graepel, Blankertz, & Obermayer, 2004; Izmirli, 2006; Gómez, 2006). In general, a chroma profile is obtained by averaging within a long-term audio window (sometimes the entire song), and it is then correlated with theoretically or cognitively inspired profiles to derive an estimation of the main key. A similar procedure is usually followed for chord extraction, but using a different set of profiles. We want to emphasize that the present work is not focused on the usefulness of chroma features comparing to tonal profiles, which has been extensively addressed in the cited literature. Instead, we focus on covariances, and the role that these might play in tonality judgments. Notice that chroma feature covariance was also used by Ellis (2007) for the purpose of classifying music by artist, but no direct or implicit association was made with tonal profiles, nor with human tonality judgments. In contrast, we perform a chroma feature covariance-based analysis over a large western music collection to obtain a tonal profile, and subsequently compare the obtained profile with a set of well-established tonal profiles. This analysis yields high and statistically significant correlations, that assess the goodness of the

proposed procedure. Furthermore, we demonstrate that even covariance-based profiles derived from very small audio fragments outstandingly correlate with the well-established ones, which are usually derived from long audio fragments. This fact might indicate that tonality judgments could reflect listeners' exposure to tonal regularities. Our evidence, we discuss, supports the hypothesis that, by means of (computationally simple) implicit or statistical learning mechanisms, listeners' judgments reflect the statistics of the tonal information they have been exposed to.

2 Experimental data

2.1 Chroma features

Chroma features are derived from the energy found within a given frequency range (usually from 50 to 5000 Hz) in short-time spectral representations (typically 100 msec) of audio signals extracted on a frame-by-frame basis. This energy is usually collapsed into a 12-bin octave-independent histogram representing the relative intensity of each of the 12 semitones of an equal-tempered chromatic scale (figure 1). Reliable chroma features should, ideally (Gómez, 2006), (a) represent the pitch class distribution of both monophonic and polyphonic signals, (b) consider the presence of harmonic frequencies, (c) be robust to noise and non-tonal sounds, (d) be independent of timbre and played instrument, (e) be independent of loudness and dynamics and (f) be independent of tuning, so that the reference frequency can be different from the standard A 440 Hz.

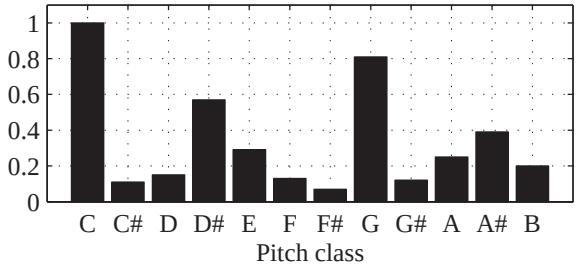

Figure 1: Example of a chroma feature vector extracted from an audio frame.

There are several algorithms and variants to compute this set of features from audio signals. In this study, we compare three approaches for chroma feature computation which we consider representative of the state-of-the-art: harmonic pitch class profiles (HPCP, Gómez, 2006, from now on denoted as C_{Gomez}), MIRToolbox[1] chroma features (Lartillot, Toivianen, & Eerola, 2008, denoted as C_{MIRTbx}), and the ones provided by Ellis in his web page[2] (Ellis, 2007, denoted as C_{Ellis}). All the chroma features were extracted in a frame-by-frame basis with the default parameters provided in their respective references except framelengths and overlapping, which we set manually to 93 msec and 50%, respectively[3] (figure 2).

2.2 Tonal profiles

Different major and minor tonal profiles have been documented in the literature. Some of them are obtained through human listening tests, as the probe tones (Krumhansl & Kessler, 1982). In this study, an unfinished C major scale (without the final tonic) was played in either its ascending or descending form, in order to establish a C major key context. After establishing this context, listeners were presented with one of the 12 chromatic scale tones in the next octave (called the probe tones), and they rated how well each tone completed the scale. Notice that Krumhansl and

[1] `http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox`

[2] `http://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/`

[3] As we are using the STFT for extracting chroma feature representations, we prefer the frame length to be a power of two. Therefore, using 2^{12} samples per frame and an audio sampling rate of 44.1 KHz yields a final frame length of $4096/44100 \approx 0.093$ seconds.

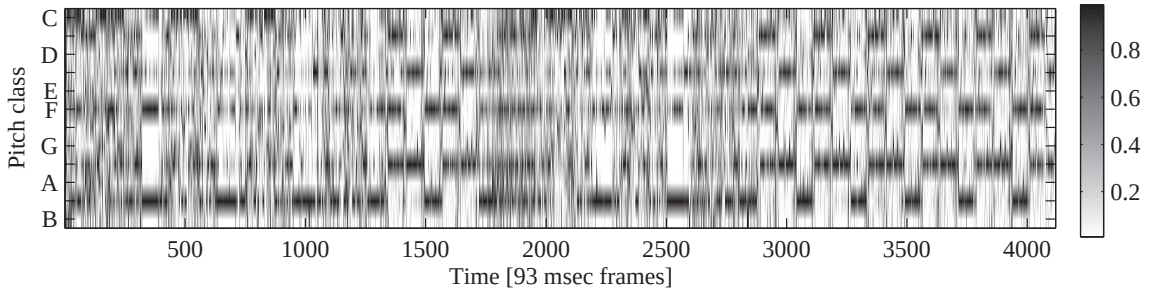

Figure 2: Example of a chroma feature sequence corresponding to an excerpt of the song "Roxanne" by The Police. Time (in frames) is represented in the horizontal axis and pitches in the vertical axis.

Kessler (1982) tested musicians only, but subsequent research also showed very similar ratings for trained and untrained listeners (Hébert, Peretz, & Gagnon, 1995). This method was also extended to a variety of ways to establish the key, including chord cadences and both major and minor scales (Krumhansl, 1990).

Another possibility to consider is that tone judgments reflect musical experience. In this sense, some papers introduce the idea of learning these profiles by analyzing symbolic representations of musical pieces which are manually labelled in terms of key. Krumhansl computed the statistical distributions of pitch classes on the melodic lines of entire classical pieces (Krumhansl, 1990) and found that these distributions were strongly correlated with the probe tone profiles obtained through human listening tests. Finally, music theory knowledge can be also used to set these tonal profiles, as Temperley (1999) has proposed.

In this study, we consider the following set of tonal profiles (figure 4):

- **Diatonic (P_D)**: Flat diatonic profiles. These are based on music analysis and assign the same weight to all the major and minor scale degrees. Although we can consider different minor scales (e.g. harmonic or melodic), we have selected the harmonic minor scale (minor sixth and major seventh) as proposed in Temperley, 1999. Minor sixth has been shown to be rated higher by humans in Krumhansl, 1990, and Temperley (1999) increased the weight of the raised seventh in minor mode.
- **Krumhansl (P_K)**: Krumhansl's profiles obtained from human ratings as presented in Krumhansl, 1990.
- **Temperley theoretic (P_{T_T})**: Temperley (1999) profiles derived from the ones by Krumhansl. The author increased the weight of the seventh scale degree in major and the raised seventh in minor mode.
- **Temperley empiric (P_{T_E})**: Major and minor key profiles are empirically derived in Temperley, 2005 using a corpus of excerpts taken from the Kotska and Payne (1995) music theory textbook, in which keys are explicitly marked.
- **Chai (P_{CH})**: Chai (2005) obtains a tonal profile by considering 7673 folk music scores[4]. The profile was generated as follows: get the key of each piece, count the number of times that each note appears, average the vectors over all the considered pieces and finally normalize it.

2.3 Music collection

We included 1953 songs for the analysis, amounting to more than 120 hours of music data. This collection was intended to be representative of "western music" and songs were distributed according to different instrumentation, genres, and styles, including live performances, remixes, *a capella* songs

[4]Chai, 2005, personal communication.

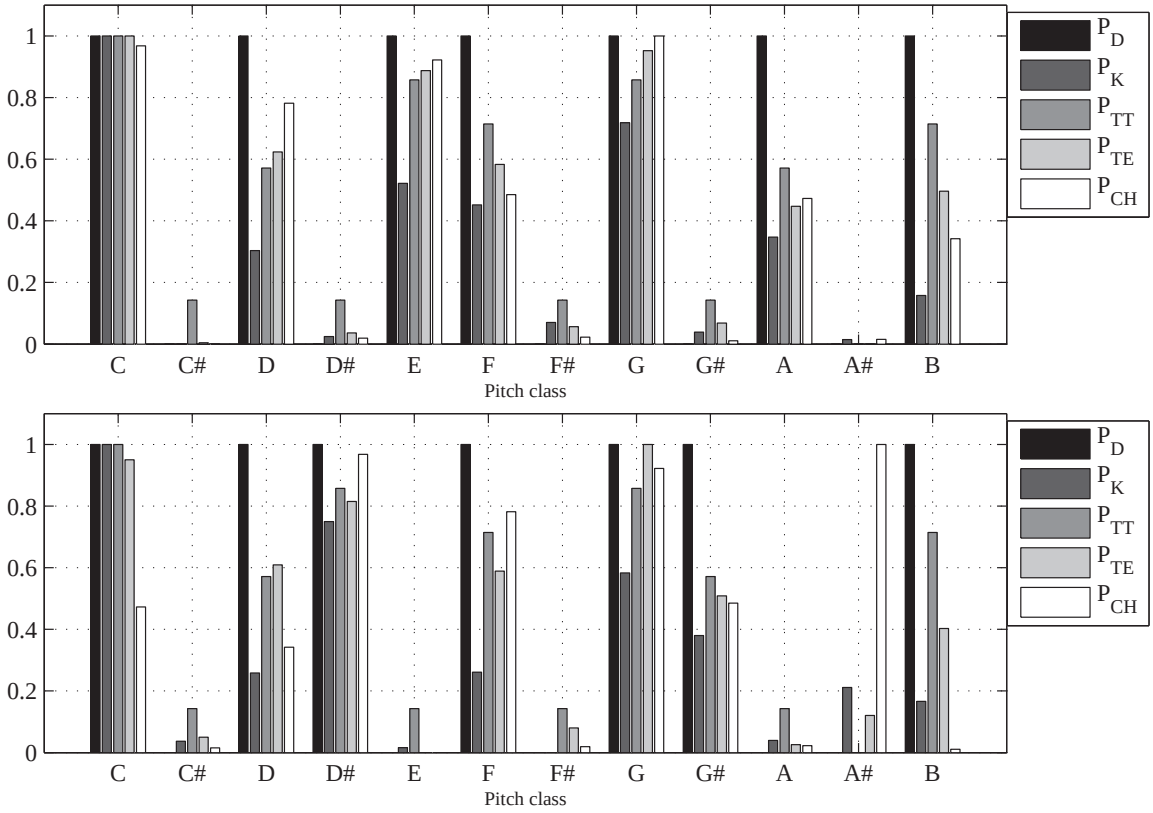

Figure 3:

Figure 4: Major (top) and minor (bottom) profiles for a C context. For visualization purposes, these have been normalized by subtracting the minimum value and then dividing by the maximum.

and instrumental pieces. The distribution of musical genres for the collection is shown in table 1. We have to note that the category *World* (corresponding to world music) did not include any non-western piece (in the sense that the used tonal hierarchy corresponds to the western tonal music tradition).

Table 1: Distribution of genres for the employed music collection.

Genre	Number of songs
Pop	668
Rock	501
Electronic	203
World	166
Classical	157
Jazz/Blues	152
Miscellaneous	106
Total	1953

3 Covariance-based statistical analysis of chroma features

As we mentioned in the introduction, the availability and the exploitation of implicit knowledge after intensive exposure to complex environments exhibiting regularities has been demonstrated in several studies. Existing literature has also shown that humans are highly sensitive to statistical

regularities (section 1). We here focus on tonality, but follow the same strategy found in previous research on visual perception when applying statistical techniques to massive data of co-occurring pitches (Simoncelli, 2003; Doi & Lewicki, 2005; Long et al., 2006; Keil, 2008). We hypothesize that human tonality judgments and pitch relations among a well-established tonal context are strongly influenced by co-occurrences of pitches to which one is being exposed in everyday listening experience.

Covariance is a measure of how much two random variables vary together (Wasserman, 2004). If two variables tend to vary together (that is, when one variable is above its expected value and the other one tends to be also above its expected value), then the covariance between them will be positive. On the other hand (if one of the variables is above its expected value and the other tends to be below its expected value), then the covariance between them will be negative. Thus, covariance is a useful tool to analyze co-occurrences of pitches.

Given the n-th song of a music collection, the analysis starts by computing chroma features for each frame m (section 2.1). This yields a set of chroma sequences $S_n = \{s_{n_m}\}$ for $n = 1, \ldots, N$ and $m = 1, \ldots, M$, where each s_{n_m} represents a chroma feature vector like the one depicted in figure 1. Here $N = 1953$ corresponds to the total number of songs and M to the total number of frames per song. To avoid expensive computations[5], we randomly select $\hat{M} = 20$ chroma feature vectors for each song and compute an averaging between the previous ones according to a predefined temporal scope. This yields a random sample of chroma feature vector sets $\hat{S}_n = \{\hat{s}_{n_m}\}$ for $n = 1, \ldots, N$ and $m = \tau + 1, \ldots, M$, where

$$\hat{s}_{n_m} = \frac{\sum_{k=m-\tau}^{m} s_{n_k}}{\max\left\{\sum_{k=m-\tau}^{m} s_{n_k}\right\}}, \tag{1}$$

and τ corresponds to the length (in frames) of this temporal scope. Short-time chroma feature representations (e.g., 93 msec) are usually averaged over larger segments of consecutive frames depending on the temporal scope of the desired description (equation 1). Traditionally, a short segment represents the pitch class distribution in a given chord and a large segment represents the pitch class distribution for a given key (Leman, 1995; Sapp, 2005). We are aware that, with long segments, some modulations might be introduced in the averaging, but we assume that these are not statistically significant for our analysis and, therefore, we do not give any special treatment for them. Furthermore, in the music genres we use for our analysis, the amount of modulations is expected to be low.

Once a chroma feature sample $\hat{S}_n$ is obtained for each song, all samples from all songs in the entire music collection are aggregated to form a chroma sample matrix $\hat{S}$ with $N \times \hat{M}$ chroma vectors (rows) of 12 pitch classes each (columns). Let p_i, p_j represent the i, j-th pitch class from sample $\hat{S}$. Then, the covariance matrix $\Phi = \{\phi_{i,j}\}$ for $i, j = 1, \ldots, 12$ is computed, where

$$\phi_{i,j} = \mathrm{E}\left[(p_i - \mathrm{E}[p_i])^T (p_j - \mathrm{E}[p_j])\right], \tag{2}$$

and T represents vector transposition. Notice that, with chroma feature covariance, each pitch is considered as a real-valued random variable between 0 and 1 which has been evaluated $\hat{M}$ times for each song, and that covariance will quantify the dependencies between these random variables. Accordingly, each element $\phi_{i,j}$ quantifies if variable p_j jointly evolves with variable p_i. Therefore, the i-th row of Φ will contain values quantifying how all 12 pitch classes vary according to reference pitch class p_i. Notice also that covariance will treat the chroma vectors in $\hat{S}$ independently, therefore disregarding any representativity for a given musical piece, specially in the case of small temporal window averagings.

[5]Notice that assuming an average 4 min song length ($M = 5217$) we would have had a total sample of $N \times M > 10^7$ chroma feature vectors.

To counteract the side-effects that a non-uniform distribution of main keys might introduce into Φ, a row-wise normalized matrix $\Psi = \{\psi_{i,j}\}$ for $i, j = 1, \ldots, 12$ is computed, where

$$\psi_{i,j} = \frac{\phi_{i,j}}{\max\{\phi_i\}}, \tag{3}$$

and ϕ_i denotes the i-th row vector of Φ.

Finally, to obtain a tonal profile C, one can circularly shift each row i by $i-1$ positions and sum column-wise. Therefore, each pitch class c_j of C can be calculated as:

$$c_j = \sum_{i=1}^{12} \psi_{i,l} \tag{4}$$

for $j = 1, \ldots, 12$, where l is the remainder of $\frac{j+i-1}{12}$.

The above process is repeated 10 times for each of the 3 different approaches for chroma feature computation outlined in section 2.1, and considering 3 temporal averaging windows of 0.5, 15 and 30 sec ($\tau = 11$, 326, and 652 frames, respectively). The profiles obtained from each of the 10 runs are averaged to form a unique profile. This yields to 9 different covariance-based profiles (C_{Gomez} with 0.5 sec averaging, C_{Gomez} with 15 sec, C_{Gomez} with 30 sec, C_{MIRTbx} with 0.5 sec, C_{MIRTbx} with 15 sec, and so forth).

4 Results and discussion

To assess tonal profile similarity, we compute Pearson's linear correlation coefficient between the covariance-based profiles and the different major and minor profiles found in the literature (table 2).

Since tonal profiles are given for major and minor modes, it would be desirable to obtain a major and a minor covariance-based profile. This would require having the songs labeled with the correct mode at each (at least) 0.5 sec. Unfortunately, reliable information of this kind is very difficult to obtain (labeling a big music corpus as the one used in this study would require a huge amount of time and effort). In addition, automatic solutions to this problem might introduce estimation errors that we want to avoid. If we were planning to extend the proposed method to distinguish between modes, we should create two sample sets (major and minor), correlate each chroma feature with a major and a minor profile (e.g., any of the ones presented in section 2.2), and aggregate it to the corresponding sample set. However, we cannot use this variant of the method here because, as we want to compare the obtained representation against tonal profiles, it would obviously bias the results of the analysis. Therefore, to provide further comparison, we consider tonal profiles obtained by combining major and minor modes found in the literature according to (a) an equal distribution of major and minor pieces, and (b) a distribution of 61% major against 39% minor pieces (table 3). The latter is an attempt to imitate the uneven distribution of major and minor pieces found in our music collection. These percentages are obtained after analyzing the mode distribution over a different audio file collection of around 1400 musical pieces of different styles and historical periods, which have been manually labeled in terms of key (Gómez, 2006). We simply obtain these averaged tone profiles by summing the correspondingly weighted major and minor ones.

Tables 2 and 3 show several statistically significant correlation values. For 12 observations, and considering a non-directional probability distribution, statistical significance at 1% (denoted by '$*$') corresponds to a correlation value higher than 0.708, and statistical significance at 0.1% (denoted by '$**$') corresponds to correlation values higher than 0.823. To compute an additional baseline for the results shown in the forementioned tables, we also computed correlations for randomly generated profiles, which yielded values lower than ± 0.001 against all tonal profiles tested.

Above tables also give us clear insights on which tonal profiles correlate better with the covariance-based ones. This information can be very useful for key estimation algorithms, as it

Table 2: Pearson's linear correlation coefficient between major (left) and minor (right) profiles. '$*$' and '$**$' denote statistical significance at $p < 0.01$ and $p < 0.001$, respectively. The second column corresponds to the time spanned by the chroma features (in seconds).

Chroma features		Major profiles					Minor profiles				
		P_D	P_K	P_{T_T}	P_{T_E}	P_{CH}	P_D	P_K	P_{T_T}	P_{T_E}	P_{CH}
C_{Gomez}	*0.5*	0.478	0.833**	0.596	0.651	0.635	0.478	0.758*	0.614	0.701	0.516
	15	0.486	0.792*	0.565	0.632	0.629	0.485	0.723*	0.590	0.703	0.592
	30	0.486	0.779*	0.555	0.625	0.625	0.485	0.713*	0.582	0.700	0.609
C_{MIRTbx}	*0.5*	0.444	0.812*	0.606	0.623	0.586	0.441	0.769*	0.613	0.635	0.309
	15	0.482	0.809*	0.597	0.638	0.620	0.478	0.765*	0.614	0.681	0.458
	30	0.488	0.800*	0.589	0.636	0.623	0.484	0.758*	0.609	0.687	0.494
C_{Ellis}	*0.5*	0.405	0.824**	0.565	0.604	0.578	0.406	0.779*	0.578	0.626	0.340
	15	0.456	0.816*	0.580	0.633	0.619	0.456	0.783*	0.598	0.678	0.472
	30	0.468	0.811*	0.571	0.631	0.624	0.466	0.764*	0.594	0.687	0.521

Table 3: Pearson's linear correlation coefficient between two different averaged profiles. '$*$' and '$**$' denote statistical significance at $p < 0.01$ and $p < 0.001$, respectively. The second column corresponds to the time spanned by the chroma features (in seconds).

Chroma features		0.5×Major + 0.5×Minor					0.61×Major + 0.39×Minor				
		P_D	P_K	P_{T_T}	P_{T_E}	P_{CH}	P_D	P_K	P_{T_T}	P_{T_E}	P_{CH}
C_{Gomez}	*0.5*	0.590	0.916^{**}	0.701	0.798^{*}	0.823^{*}	0.583	0.916^{**}	0.692	0.782^{*}	0.820^{*}
	15	0.599	0.874^{**}	0.669	0.787^{*}	0.872^{**}	0.591	0.872^{**}	0.660	0.769^{*}	0.857^{**}
	30	0.599	0.859^{**}	0.659	0.781^{*}	0.882^{**}	0.591	0.858^{**}	0.650	0.763^{*}	0.863^{**}
C_{MIRTbx}	*0.5*	0.546	0.910^{**}	0.705	0.742^{*}	0.638	0.540	0.905^{**}	0.699	0.734^{*}	0.665
	15	0.592	0.906^{**}	0.701	0.778^{*}	0.770^{*}	0.585	0.901^{**}	0.693	0.764^{*}	0.776^{*}
	30	0.599	0.897^{**}	0.694	0.780^{*}	0.798^{*}	0.592	0.893^{**}	0.685	0.766^{*}	0.798^{*}
C_{Ellis}	*0.5*	0.501	0.920^{**}	0.662	0.725^{*}	0.655	0.494	0.914^{**}	0.655	0.715^{*}	0.676
	15	0.562	0.919^{**}	0.683	0.773^{*}	0.780^{*}	0.555	0.918^{**}	0.675	0.759^{*}	0.784^{*}
	30	0.575	0.907^{**}	0.675	0.778^{*}	0.818^{*}	0.569	0.903^{**}	0.666	0.762^{*}	0.814^{*}

Table 4: Pearson's linear correlation coefficient between empirically-derived tonal profiles from different temporal windows. All values are statistically significant at $p < 0.001$.

Chroma features		C_{Gomez}			C_{MIRTbx}			C_{Ellis}		
		0.5	*15*	*30*	*0.5*	*15*	*30*	*0.5*	*15*	*30*
C_{Gomez}	*0.5*	1.000	0.988	0.982	0.935	0.981	0.985	0.950	0.993	0.996
	15	0.988	1.000	0.999	0.882	0.962	0.975	0.899	0.971	0.987
	30	0.982	0.999	1.000	0.866	0.954	0.969	0.883	0.962	0.982
C_{MIRTbx}	*0.5*	0.935	0.882	0.866	1.000	0.973	0.957	0.983	0.958	0.936
	15	0.981	0.962	0.956	0.973	1.000	0.998	0.965	0.987	0.986
	30	0.985	0.975	0.969	0.957	0.998	1.000	0.951	0.987	0.990
C_{Ellis}	*0.5*	0.950	0.899	0.883	0.983	0.965	0.951	1.000	0.977	0.955
	15	0.993	0.971	0.962	0.958	0.987	0.987	0.977	1.000	0.996
	30	0.996	0.987	0.982	0.936	0.986	0.990	0.955	0.996	1.000

has been widely acknowledged in the literature (Gómez, 2006). Tonal profiles obtained through probe-tone experiments using western music listeners (P_K) are better correlated with covariance-based profiles, while profiles derived from music theory are less correlated (P_D and P_{T_T}). Empirically derived ones (from symbolic music annotations, P_{T_E} and P_{CH}) also yield high correlation values. Note that, in addition, when comparing covariance-based profiles with a 'mixed' P_K tonal profile, we get a substantial increase in correlation. This leads up to some values higher than 0.9 (which corresponds to a probability $p < 0.0001$ of occurring by chance).

It is shown in tables 2 and 3 that, in many of the cases, covariance-based profiles extracted from very short musical excerpts (0.5 sec) are the best correlated with the tonal profiles considered. This association is specially strong with P_K (in fact it is the only association that is statistically significant, whereas the association between any other profiles and any of the chroma features used does not reach that level). To further explore the relation between different chroma feature extraction methods and temporal windows, we computed the correlation coefficients between all covariance-based profiles (table 4).

We first should note the high correlation values obtained and the big amount of significance achieved. Accordingly, no relevant differences were found when using different approaches for chroma feature extraction (figure 5). This fact indicates that any could be useful for future research in this area. In addition, the combination of the obtained correlation values (tables 2 and 3), together with the absence of relevant differences between chroma feature extraction methods (table 4), proves the suitability of our covariance-based tonal profile computation to estimate a tonal profile.

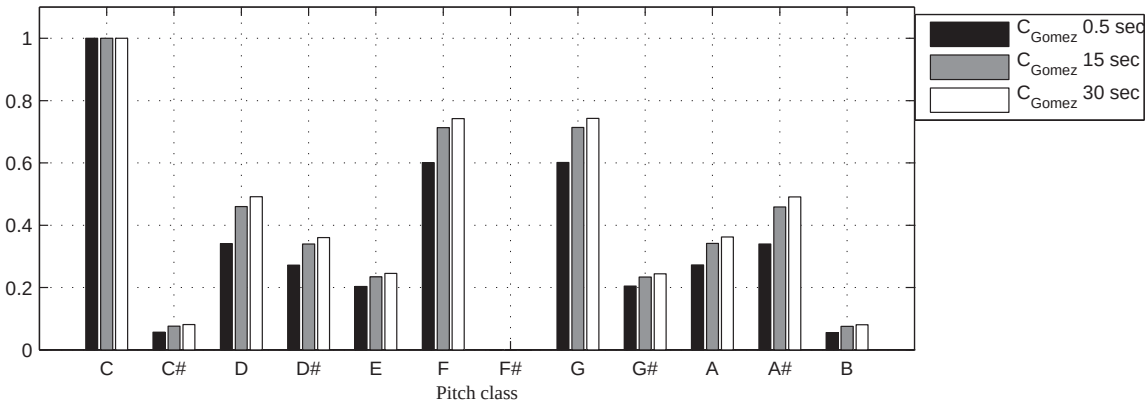

Figure 5: Covariance-based profiles obtained for 0.5, 15 and 30 second averaging (first, second, and third bar in each pitch class, respectively).

It is even more outstanding that, in any case, the difference in correlation between short, mid,

and long-term empirical profiles is very small. In general, the majority of the empirically, cognitively, or music-theoretically inspired profiles are grounded into "well-established" tonal contexts (section 2.2). Therefore, obtaining similar (as well as statistically significant, table 4) correlations between profiles for very small (0.5 sec) and very large (30 sec) temporal windows, is a surprising and relevant outcome. This leads us to reason that tonal stability among a predefined key and the hierarchy of notes derived from human ratings could be strongly influenced by our short-term music listening experience. In other words, one could interpret that, in order to derive tonal representations for a piece, it is enough with intensively exploiting short-term information. To this respect, Tillmann and Bigand (2004) speculated that "numerous experiments highlight the strength of local structure for music perception. In short temporal windows, listeners understand the local function of cadences, perceive changes in tonality, and go somewhat beyond the musical surface. However, their perception fails to integrate these structural markers into more global structures. For short time-spans, simulations with the connectionist models suggest that the influence of local structures on musical processing can emerge from the simple accumulation of activation patterns over time (weighted by temporal decay), without necessarily extracting a hierarchical organization".

From the point of view of the neurophysiology of learning, it seems plausible to consider that the neural mechanisms involved in covariation detection operate optimally at a temporal scale of a very few hundred milliseconds (Chen, Haykin, Eggermont, & Becker, 2007), and it is also the case with those related with chord and syntactic processing in the brain (Koelsch, 2005; Koelsch & Siebel, 2005). Therefore, our findings seem to be supporting the hypothesis that, computationally, local-scale implicit learning could be involved in building long-term mental representations of tonality by means of the mechanisms advanced in Bigand, Poulin, Tillmann, & D'Adamo, 2003. Further research on the statistical properties of chroma feature sequences (contrasting to the current one, which is only considering their simultaneous occurrences) could give support to their distinction between psychoacoustic and syntactic components in music processing. To this respect, Huron (2006), discusses the pivotal role that statistical learning has in his ITRPA theory in order to account for musical expectation, understanding and appraisal. Surprisingly, his analyses do not take into account co-occurrences, but first and higher order temporal dependencies.

The perceptual ordering of chromatic scale tone combinations (i.e., the so-called *consonance theory*) also has important ties with tonality judgments and the discussion made here. Perceived differences among possible combinations of tones have been studied from different perspectives and terms, including psychophysical studies (e.g. Malmberg, 1918; Krumhansl, 1990; Huron, 1994), and mathematical or geometrical relationships (e.g. Helmholtz, 1954; Kameoka & Kuriyagawa, 1969). More interestingly, it has been shown that consonance ordering is also predicted by the statistical structure of human speech sounds (Schwartz et al., 2003).

For completeness of the analysis, we plot an exemplary covariance-based profile for a short term averaging window (0.5 sec) versus the averaged P_D, P_K, P_{T_T}, P_{T_E} and P_{CH} profiles (figure 6). We observe that, in the covariance-based one, the minor third's magnitude is higher than the major third, whereas the opposite relation obtains in the tonal profile curves. Similarly, the major seventh's magnitude is higher in the tonal hierarchy than the minor seventh, and the opposite holds for the covariance-based profile. As noted in Krumhansl, 1990, these discrepancies reflect scale membership, and they are more evident with theoretic profiles (P_D and P_{T_T}) than with others.

Despite the differences highlighted in the above paragraph, the pitch class rankings mostly agree. We can further assess it with a pitch class ranking plot (figure 7). This plot is obtained by considering the ranks of each pitch class within each tonal profile. That is, if the tonic (T) has the highest value in the tonal profile, then $rank(T) = 1$, if the fifth (5) has the second highest value, then $rank(5) = 2$, and so on. Kendall τ rank correlation coefficients for the three different averagings of C_{Gomez} are also shown (table 5).

5 Conclusions

Motivated by evidence that image source statistics predict the response properties of several aspects of visual perception, we investigate the relation between chroma statistics and human judgments of

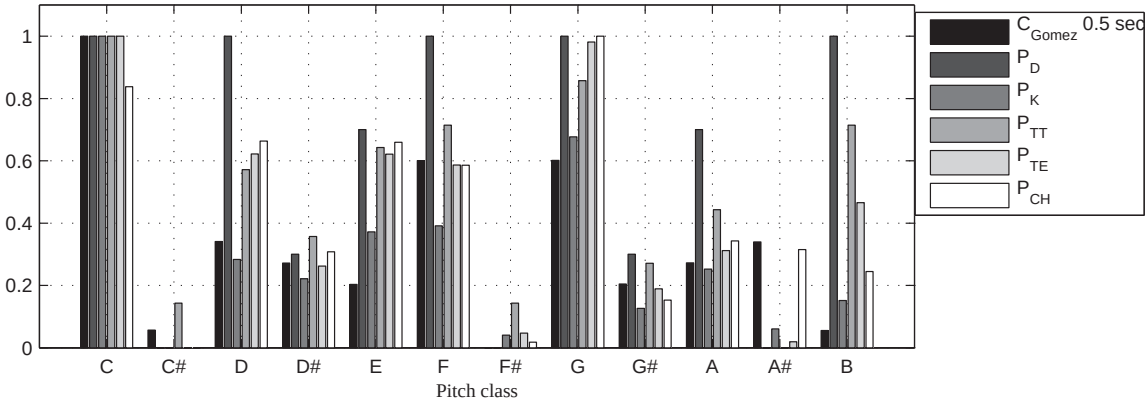

Figure 6: Plot of an example short-term covariance-based profile (first bar in each pitch) versus the other profiles considered (0.61×Major + 0.39×Minor).

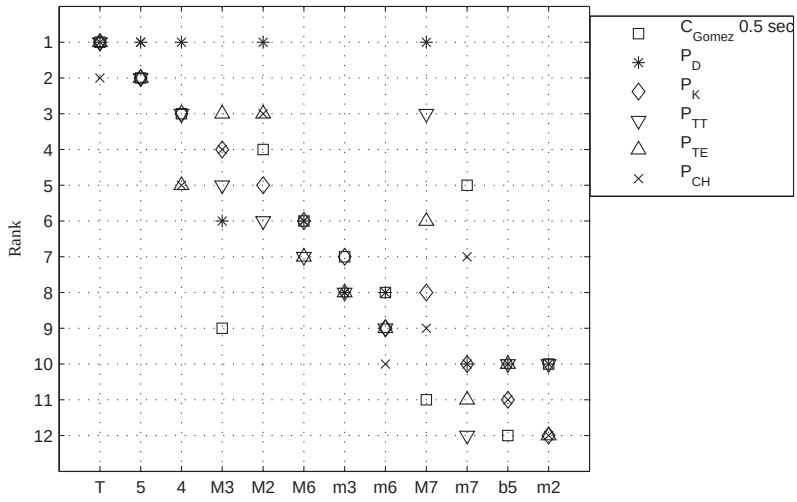

Figure 7: Pitch class ranking of the obtained covariance-based profile (red squares) versus the other profiles considered (0.61×Major + 0.39×Minor).

Table 5: Kendall's τ for covariance-based profiles versus the other considered profiles.

Chroma features		0.61×Major + 0.39×Minor				
		P_D	P_K	P_{T_T}	P_{T_E}	P_{CH}
C_{Gomez}	0.5 sec	0.465	0.606	0.431	0.515	0.697
C_{Gomez}	15 sec	0.499	0.636	0.461	0.545	0.727
C_{Gomez}	30 sec	0.499	0.636	0.461	0.545	0.727

tonality. Following the existing literature in close research areas, we hypothesize that judgments of tonality are strongly influenced by co-occurrences of chroma values to which one is being exposed in everyday listening experience. Accordingly, we propose a novel statistical technique based on covariance analysis of chroma features computed from musical recordings to extract tonal profiles from massive music collections.

The goodness of the proposed technique is assessed by exhaustively comparing empirical tonal profiles extracted using the covariation of 3 different chroma feature implementations versus 5 well-established tonal profiles. Therefore, the present article explicitly addresses the link between common MIR tools and existing cognition and perception measures. To the authors' knowledge, there are few works comparing different tonal profiles and even fewer extending the analysis to more than

one chroma feature extraction method. Moreover, these comparisons are usually made on the basis of single songs and long-term averagings. In contrast, we consider covariance analysis over massive short, mid, and long-term chroma feature averagings across a 1953 song collection of western tonal music.

The obtained empirical profiles yield high and statistically significant correlations with existing tonal profiles obtained from human judgments, music theory, and empirical studies performed on symbolic music. Statistical significance has also been achieved among all the different approaches for chroma feature extraction considered. This observation partially supports the goodness of our covariance-based tonal profile computation.

In addition, we empirically demonstrate that our initial hypothesis could be true: human listeners could build tonal representations with the help of very simple statistical mechanisms. We present strong evidence that covariance-based profiles obtained from very short real musical excerpts predict tonal profiles from different studies. Results are partially supported by existing research and hold to literature on the implicit learning of tonality. Specifically, very high correlations are achieved with cognitively inspired tonal profiles derived by means of probe tone ratings (Krumhansl, 1990). Note that, in the case of tonal learning, these profiles corresponded to the best 'cognitive representation' that could be derived from this kind of statistical analysis.

As future work, we will further push this approach by considering data and listeners from other cultures. We will focus on non-western music and try to exploit this and other techniques to empirically extract tonal information from them. We hope that this article will encourage the MIR community to study the learning of tonality and their cognitive aspects from an audio content processing perspective. In addition, we believe that MIR-based techniques and resources like the chroma features can be of great benefit for other researchers studying music and tonality from a different perspective.

6 Acknowledgments

The authors are extremely grateful to Barbara Tillmann, who provided helpful comments and suggestions for improving the quality of the manuscript. They also want to thank their colleagues and staff at the Music Technology Group (UPF) for their support and work, specially Graham Coleman. Furthermore, the authors wish to thank the anonymous reviewers for very helpful comments. This research has been partially funded by the EU-IP project PHAROS[6] (IST-2006-045035), the e-Content Plus project VARIAZIONI[7] (ECP-2005-CULT-038264) and the European FET project EmCAP[8] (FP6-IST-013123).

References

Bartsch, N. A., & Wakefield, G. H. (2001). To catch a chorus: using chroma-based representations for audio thumbnailing. *IEEE Workshop on Apps. of Signal Processing to Audio and Acoustics (WASPAA)*, 15-18.

Bharucha, J. J. (1984). Anchoring effects in music: the resolution of dissonance. *Cognitive Psychology*, *16*, 485-518.

Bigand, E., Poulin, B., Tillmann, B., & D'Adamo, D. (2003). Cognitive versus sensory components in harmonic priming effects. *Journal of Experimental Psychology: Human perception and performance*, *29*(1), 159-171.

Castellano, M. A., Bharucha, J. J., & Krumhansl, C. L. (1984). Tonal hierarchies in the music of north india. *Journal of Experimental Psychology: General*, *113*(3), 394-412.

Chai, W. (2005). *Automated analysis of musical structure*. Unpublished doctoral dissertation, Massachussets Institute of Technology, USA.

[6] `http://www.pharos-audiovisual-search.eu`
[7] `http://www.variazioniproject.org`
[8] `http://emcap.iua.upf.es`

Chen, Z., Haykin, S., Eggermont, J. J., & Becker, S. (2007). *Correlative learning: a basis for brain and adaptive systems*. John Wiley and Sons.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 24-39.

Doi, E., & Lewicki, M. (2005, July). Relations between the statistical regularities of natural images and the response properties of the early visual system. *Japanese Cognitive Science Society, SIG Pattern Recognition and Perception Models*. (Preprint.)

Ellis, D. P. W. (2007, October). Classifying music audio with timbral and chroma features. *Int. Symp. on Music Information Retrieval (ISMIR)*, 339-340.

Fujishima, T. (1999). Realtime chord recognition of musical sound: a system using common lisp music. *Int. Computer Music Conference (ICMC)*, 464-467.

Gómez, E. (2006). *Tonal description of music audio signals*. Unpublished doctoral dissertation, Universitat Pompeu Fabra, Barcelona, Spain.

Gómez, E., & Herrera, P. (2004). Estimating the tonality of polyphonic audio files: cognitive versus machine learning modelling strategies. *Int. Symp. on Music Information Retrieval (ISMIR)*, 92-95.

Hébert, S., Peretz, I., & Gagnon, L. (1995). Perceiving the tonal ending of tune excerpts: the roles of pre-existing representation and musical expertise. *Canadian Journal of Experimental Psychology*, *49*, 193-209.

Helmholtz, H. L. F. (1954). *On the sensations of tone as a physiological basis for the theory of music*. Dover.

Huron, D. (1994). Interval-class content in equally-tempered pitch-class sets: common scales exhibit optimum tonal consonance. *Music Perception*, *11*, 289-305.

Huron, D. (2006). *Sweet anticipation: music and the psychology of expectation*. MIT Press.

Izmirli, Ö. (2006, October). Audio key finding using low-dimensional spaces. *Int. Symp. on Music Information Retrieval (ISMIR)*, 8-12.

Kameoka, A., & Kuriyagawa, M. (1969). Consonance theory part ii: consonance of complex tones and its calculation. *Journal of the Acoustical Society of America*, *45*, 1460-1469.

Keil, M. (2008). Does face image statistics predict a preferred spatial frequency for human face processing? *Proc. of the Royal Society B*, *275*(1647), 2095-2100.

Koelsch, S. (2005). Neural substrates of processing syntax and semantics in music. *Current Opinion in Neurobiology*, *15*, 1-6.

Koelsch, S., & Siebel, S. A. (2005, December). Towards a neural basis of music perception. *Trends in Cognitive Sciences*, *9*(12), 578-584.

Kotska, S., & Payne, D. (1995). *Tonal harmony*. McGraw-Hill.

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*. Oxford University Press.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, *89*(4), 334-368.

Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human perception and performance*, *5*, 579-594.

Lartillot, O., Toivianen, P., & Eerola, T. (2008). *A matlab toolbox for music information retrieval*. Springer.

Leman, M. (1995). Music and schema theory: cognitive foundations of systematic musicology. *Information Science*(31).

Long, F., Yang, Z., & Purves, D. (2006). Spectral statistics in natural scenes predict hue, saturation, and brightness. *Proc. of the Nat. Academy of Sciences*, *103*(15), 6013-6018.

Malmberg, C. F. (1918). The perception of consonance and dissonance. *Psychological Monographs*, *25*(2), 93-133.

Müller, M. (2007). *Information retrieval for music and motion*. Springer.

Patel, A. (2008). *Music, language, and the brain*. Oxford University Press.

Paws, S. (2004). Musical key extraction from audio. *Int. Symp. on Music Information Retrieval (ISMIR)*, 96-99.

Purves, D., & Lotto, R. B. (2003). *Why we see what we do: evidence for an empirical theory of vision*. Sinauer.

Purwins, H., Graepel, T., Blankertz, B., & Obermayer, K. (2004). *Correspondence analysis for*

visualizing interplay of pitch class, key, and composer (G. Mazzola, T. Noll, & E. Luis-Puebla, Eds.). Osnabrück series on Music and Computation.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: a parallel distributed processing approach*. MIT Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27-52.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: listening (and learning) out of the corner of your ear. *Psychological Science*, *8*, 101-105.

Sapp, C. (2005). Visual hierarchical key analysis. *Computers in Entertainment*, *3*(4).

Schwartz, D., Howe, C. Q., & Purves, D. (2003, August). The statistical structure of human speech sounds predicts musical universals. *Journal of Neuroscience*, *23*(18), 7160-7168.

Sheh, A., & Ellis, D. P. W. (2003). Chord segmentation and recognition using em-trained hidden markov models. *Int. Symp. on Music Information Retrieval (ISMIR)*, 183-189.

Simoncelli, E. (2003). Vision and the statistics of visual environment. *Current opinion in Neurobiology*, *13*(2), 144-149.

Temperley, D. (1999). What's key for key? the krumhansl-schmuckler key finding algorithm reconsidered. *Music Perception*, *17*(1), 65-100.

Temperley, D. (2005). A bayesian key-finding model. *MIREX extended abstract*.

Tillmann, B. (2005). Implicit investigations of tonal knowledge in nonmusician listeners. *Annals of the New York Academy of Sciences*, *1060*(1), 100-110.

Tillmann, B., Bharucha, J. J., & Bigand, E. (2000). Implicit learning of tonality: a self-organizing approach. *Psychological Review*, *107*(4), 885-913.

Tillmann, B., & Bigand, E. (2004). The relative importance of local and global structures in music perception. *Journal of Aesthetics and Art Criticism*, *62*(2), 211-222.

Wasserman, L. (2004). *All of statistics*. Springer.