A CONTRARIO PATCH MATCHING, WITH AN APPLICATION TO KEYPOINT MATCHES VALIDATION

Rafael Grompone von Gioi[⋆] *Viorica Pătrăucean*[†]

[⋆] CMLA, ENS Cachan, France, grompone@cmla.ens-cachan.fr
[†] Department of Engineering, University of Cambridge, UK, vp344@cam.ac.uk

ABSTRACT

We describe a simple metric for image patches similarity, together with a robust criterion for unsupervised patch matching. The gradient orientations at corresponding positions in the two patches are compared and the normalized errors are accumulated. Based on the *a contrario* framework, the matching criterion validates a match between two patches when this cumulative error is too small to have occurred as the result of an accidental agreement. The method is illustrated in the validation of keypoint matches.

***Index Terms*—** patch matching, a contrario validation, false positives

1. INTRODUCTION

Image patches, as opposed to entire images [1], represent a good trade-off between informativeness and robustness to local deformations or occlusions, a highly desirable quality in various computer vision tasks. Discovered using specific detectors [2] or unsupervised learning [3], or simply sampled densely over the entire image [4], image patches appear under different representations in a wide range of applications: image retrieval [5], image classification [6], object recognition [7], (cross-domain) image matching [8], image editing [4], to name only a few.

Regardless of the application, appropriate metrics are needed to reason about patch similarity in a robust way. Simple metrics like L1 norm, L2 norm, or SSD (Sum of Squared Differences) over the intensity values of the patch pixels [9], despite their computational efficiency, have high sensitivity to small local deformations and noise, making them unsuitable for reliable estimation of patch similarity. Ardo and Astrom use a Bayesian formulation and learn from videos prior distributions of correlation coefficients to add robustness to the matching procedure [10].

To cope with metrics sensitivity, a large amount of research works focused on designing efficient descriptors of image patches that encode desired geometric and appearance invariances, using histogram representations, e.g. SIFT [11, 12], ASIFT [13], Shape Contexts [14], Self-Similarity descriptors [8], etc. Robust to predefined deformations, patch comparison based on these descriptors can then be performed more reliably using the simple metrics mentioned above, resulting in scalable recognition or indexing systems. For in-depth comparative studies on image patch detectors and descriptors we refer the reader to [15, 16, 17].

Despite their wide use in various setups due to their discriminative power, reasoning in the space defined by these descriptors is far from trivial in applications that require direct image matching, e.g. image registration for mosaicing [18], or stereo image matching [19]. Often, matches are found using nearest neighbour schemes [20], together with a hard-coded threshold, for a predefined distance function. A more robust criterion to match SIFT-like descriptors uses a threshold on the ratio between the nearest neighbour and the second nearest neighbour [11]. Hard-coded thresholds limit the flexibility of the methods, whilst the latter criterion fails in images containing repetitive patterns. To counteract these issues, Rabin et al. [21] proposed a framework based on the *a contrario* theory for robust parameterless SIFT descriptor matching. The method requires learning the distribution of the descriptor space and uses the earth mover distance to quantify the descriptor similarity.

In this paper we show that it is possible to obtain a simple unsupervised parameter-free matching criterion by reasoning directly in the image space, bypassing histogram representations. This is a natural extension of the work described in [22] for symmetry detection. Based on the *a contrario* theory [23], our method integrates the normalised gradient orientation errors between two patches and evaluates the probability of observing such an error as a result of an accidental agreement. If this probability is small, the match is considered as valid. In the following, we describe in detail the error computation procedure, together with the theoretical setup underlying the matching criterion (sect. 2). The applicability of the proposed approach is illustrated for SIFT keypoint matches validation, and is supported with qualitative results (sect. 3).

2. PATCH MATCHING

The proposed patch matching validation procedure is based on the *a contrario* theory, which relies mainly on the non-accidentalness principle [24, 25]; informally, this principle

states that there is no perception in noise. In the words of D. Lowe, "we need to determine the probability that each relation in the image could have arisen by accident, $P(a)$. Naturally, the smaller that this value is, the more likely the relation is to have a causal interpretation" [25, p. 39]. In our context, we need to assess the existence of a causal relation between two patches, based on an appropriate metric. If the distance is bigger than the expected distance for a pair of patches drawn from a random model, the match is rejected as there is not enough evidence to discard an accidental match.

More formally, given a candidate pair of patches P and Q of equal size, a distance function $d(P, Q)$ will be defined, together with a stochastic model $\mathcal{H}_0$ for random patches used to evaluate accidentalness. We denote by $D_{\mathcal{H}_0}$ a random variable corresponding to the distance between two random patches drawn from $\mathcal{H}_0$. To assess the accidentalness of a match (P, Q), we need to evaluate the probability $\mathbb{P}[D_{\mathcal{H}_0} \leq d(P, Q)]$ of observing under $\mathcal{H}_0$ a distance $D_{\mathcal{H}_0}$ smaller than the observed one $d(P, Q)$. When this probability is small enough, there exists evidence to reject the null hypothesis and declare the candidate meaningful. However, one needs to consider that usually multiple patch pairs are tested. For example, if 100 tests are performed, it would not be surprising to observe an event that appears with probability 0.01 under random conditions. The number of tests N_T needs to be included as a correction term, as it is done in the statistical multiple hypothesis testing framework [26] (see [27, sect.4.4] for more details). Following the *a contrario* methodology [23], we define the *Number of False Alarms* (NFA) of a pair:

$$\mathrm{NFA}(P, Q) = N_T \cdot \mathbb{P}\Big[D_{\mathcal{H}_0} \leq d(P, Q)\Big].$$

Pairs with NFA $\leq \varepsilon$, for a predefined ε value, are accepted as matches. One can show [23, 27] that under $\mathcal{H}_0$ the expected number of pairs with NFA $\leq \varepsilon$, is bounded by ε:

$$\mathbb{E}_{\mathcal{H}_0}\left[\sum_{(P,Q)\in\mathcal{N}_T} \mathbb{1}_{\mathrm{NFA}(P,Q)\leq\varepsilon}\right] < \varepsilon,$$

where $\mathcal{N}_T$ is the set of N_T tests. As a result, ε corresponds to the mean number of false detections per random image pair. In most practical applications, the simple value $\varepsilon = 1$ is suitable; we will set it once and for all in our application as well. With this choice, the expected number of false positive patch matches per random image pair is guaranteed to be upper-bounded by 1.

Regarding the choice of $d(\cdot, \cdot)$ and $\mathcal{H}_0$, we suggest that a robust evaluation of patch similarity can be obtained by analysing the gradient orientation errors of the corresponding pixels in the candidate pair of patches. Let p_i and q_i be the corresponding i-th pixels of patches P and Q, extracted from images I_1 and I_2, respectively. The index i takes values in $\{1, \ldots, N_p\}$, where N_p is the number of pixels in the patches, which are of equal size. Then the orientation error of the pair of pixels is given by $\big|\mathrm{Angle}\big(\nabla I_1(p_i), \nabla I_2(q_i)\big)\big|$, where $\nabla I_1(p_i)$ is the image gradient at p_i and $\nabla I_2(q_i)$ is the image gradient at q_i. The metric[1] $d(P, Q)$ can now be defined as the additive normalised orientation error of the pairs of pixels:

$$d(P, Q) = \sum_{i=1}^{N_p} \frac{\big|\mathrm{Angle}\big(\nabla I_1(p_i), \nabla I_2(q_i)\big)\big|}{\pi}.$$

A perfect match has $d(P, Q) = 0$, whilst the worst has $d(P, Q) = N_p$.

With this choice, an appropriate (unstructured) null hypothesis $\mathcal{H}_0$ is an isotropic gradient field whose orientations are i.i.d. random variables, uniformly distributed over $[0, 2\pi]$. These properties hold in a Gaussian white noise model, under certain conditions of sub-sampling [23, p. 67].

Within this setup, $D_{\mathcal{H}_0}$ corresponds to the sum of N_p independent and uniformly distributed random variables taking values in $[0, 1]$. Using the Irwin-Hall distribution [28], for a given d, with $0 \leq d \leq N_p$, we obtain:

$$\mathbb{P}[D_{\mathcal{H}_0} \leq d] = \frac{1}{N_p!} \sum_{i=0}^{\lfloor d \rfloor} (-1)^i \binom{N_p}{i} (d - i)^{N_p},$$

where $\lfloor d \rfloor$ is the largest integer not bigger than d. Moreover, it can be observed that the first term of the sum gives an upper bound of this probability. For computational reasons, we keep only this first term, as it is a sufficient approximation[2] to evaluate the NFA test. Thence:

$$\mathbb{P}\Big[D_{\mathcal{H}_0} \leq d(P, Q)\Big] \leq \frac{\big[d(P, Q)\big]^{N_p}}{N_p!}.$$

Finally, to complete the reasoning, we need to compute the number of tests N_T. This term needs not be very accurate; it only has to reflect the order of magnitude of the number of tests to ensure that the proposed validation adapts to the image sizes while keeping under control the false positives for increasing image size (which implicitly leads to increased number of match candidates). The number of tests is determined by the number of patch pairs *potentially* evaluated, which is dependent on the problem being considered. For example, when comparing patches at a single scale and in the same image, the number of patches would be given by the number of pixels (each pixel of the image represents a patch centre) times the number of orientations; then, the number of tests would be equal to the number of pairs of patches, which is roughly the square of the number of patches. In multiscale

[1] It is simple to verify that $d(\cdot, \cdot)$ is a proper metric for orientation fields, satisfying the non-negativity, identity, symmetry, and subadditivity conditions.

[2] We performed exact (but slow) computations of this probability using arbitrary-precision arithmetic (GMP library `http://gmplib.org/`); the insignificant differences compared to the approximate computation confirm this choice for our problem.

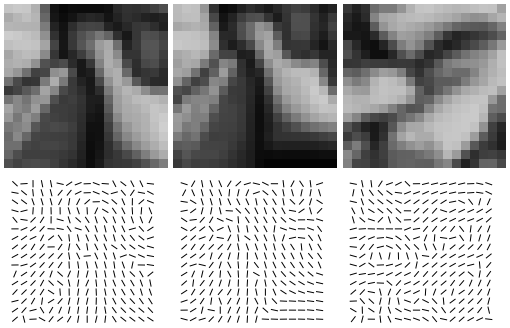

Fig. 1. Three 18x18 image patches and their corresponding 16x16 orientation fields used for matching; centered differences are used to compute gradient orientations. The first two validate as a match, while the third is rejected when compared with any of the former two.

comparisons, the number of scales multiplies the number of patches considered in the previous case. The next section illustrates the computation of the number of tests for the particular problem of keypoint matching.

To conclude, a pair of patches is accepted as valid match if its NFA satisfies the simple test:

$$\mathrm{NFA}(P,Q) = \frac{N_T}{N_p!}\left[\sum_{i=1}^{N_p}\frac{\left|\mathrm{Angle}\big(\nabla I_1(p_i),\nabla I_2(q_i)\big)\right|}{\pi}\right]^{N_p} \leq 1.$$

3. KEYPOINT MATCHES VALIDATION

To illustrate the applicability of the proposed method, we performed matches validation for SIFT keypoints. The classic method [11] to match SIFT keypoints basically creates a patch descriptor in the form of a histogram of the gradient orientations in the neighbourhood of the keypoint. Then, a match is accepted if the ratio between the distance to the nearest neighbour descriptor and the second nearest neighbour is below a predefined threshold.

Our method compares directly the gradient orientations of the patches. We use SIFT keypoints [11, 12] defined by location, scale, and orientation. For matching, we extract square image patches centered on the keypoints location, with the scale indicated by the keypoints scale, and rectify the patches to compensate for rotation. The patches are extracted by filtering with a Gaussian filter and then sampling using bilinear interpolation. Then the orientation fields are computed. Fig. 1 shows three examples of patches and the corresponding orientation fields. A threshold is applied on the gradient magnitude, in order to prevent comparing features that are not contrasted enough to be visible: if both corresponding pixels magnitudes are under a threshold, the pair is not counted in N_p; if only one of them is valid, a maximum normalised error of one is added to $d(P,Q)$; when both pixels are valid, the normalized error is computed as described in the previous section. Empirically, the threshold was fixed to 3.

Finally, we need to specify the number of tests for this particular application. When matching an image I_1 of size $m_1 \times n_1$ with an image I_2 of size $m_2 \times n_2$, the number of possible centres for patches in I_1 is about $m_1 n_1$; similarly, we have about $m_2 n_2$ patch centres in I_2. We consider $\sqrt{m_1 n_1}$ different patch orientations in I_1 and $\sqrt{m_2 n_2}$ in I_2. To account for multiple scales, we consider $\log_2\big(\max(m_1,n_1)\big)$ scales in I_1 and $\log_2\big(\max(m_2,n_2)\big)$ scales in I_2. All-in-all, the number of tests writes

$$N_T = (m_1 n_1)^{\frac{3}{2}} \cdot \log_2\big(\max(m_1,n_1)\big) \cdot (m_2 n_2)^{\frac{3}{2}} \cdot \log_2\big(\max(m_2,n_2)\big).$$

Our motivation here is to give a proof of concept by comparing the proposed validation method with the widely used SIFT second nearest neighbour criterion. (We used the implementation from [12].) Fig. 2 shows two examples that illustrate this comparative analysis. The first example is from the well-known VGG dataset (`http://www.robots.ox.ac.uk/~vgg/data/data-aff.html`); the second one is from a dataset with repetitive structures [21].

The first example illustrates the typical behaviour of the two methods: SIFT shows favourable results, the proposed method found less matches; however, qualitatively, the results are comparable. The second example shows clearly the advantage of the proposed method when repetitive structures are present. It is well known that this is a problematic case for the SIFT criterion, which results in a reduced number of matches. The proposed method handles naturally repetitive structures as it evaluates how good a match is independently of other matches, allowing to produce multiple matches for a single patch, as shown in the figure.

These preliminary results are encouraging, and future work will explore the applicability of the proposed method in retrieval applications, using directly the orientation field as keypoint descriptor, and our method for matching validation.

4. CONCLUSION

We presented a metric for image patches and an unsupervised method to validate patch matches. Its use was illustrated in validating matches between SIFT keypoints. Our method works out of the box to compare patches surrounding the keypoints; the results are comparable to SIFT's second nearest neighbour criterion. The proposed method is able to handle naturally repetitive structures and is able to produce multiple matches per patch. The control of false matches, based on the *a contrario* framework, results in reliable matches. Future work will concentrate on providing a solid foundation to the thresholding of the gradient, as well as improving the robustness to affine transformations.

Acknowledgments: We thank Julie Delon for kindly providing test images and Ives Rey-Otero for numerous suggestions.

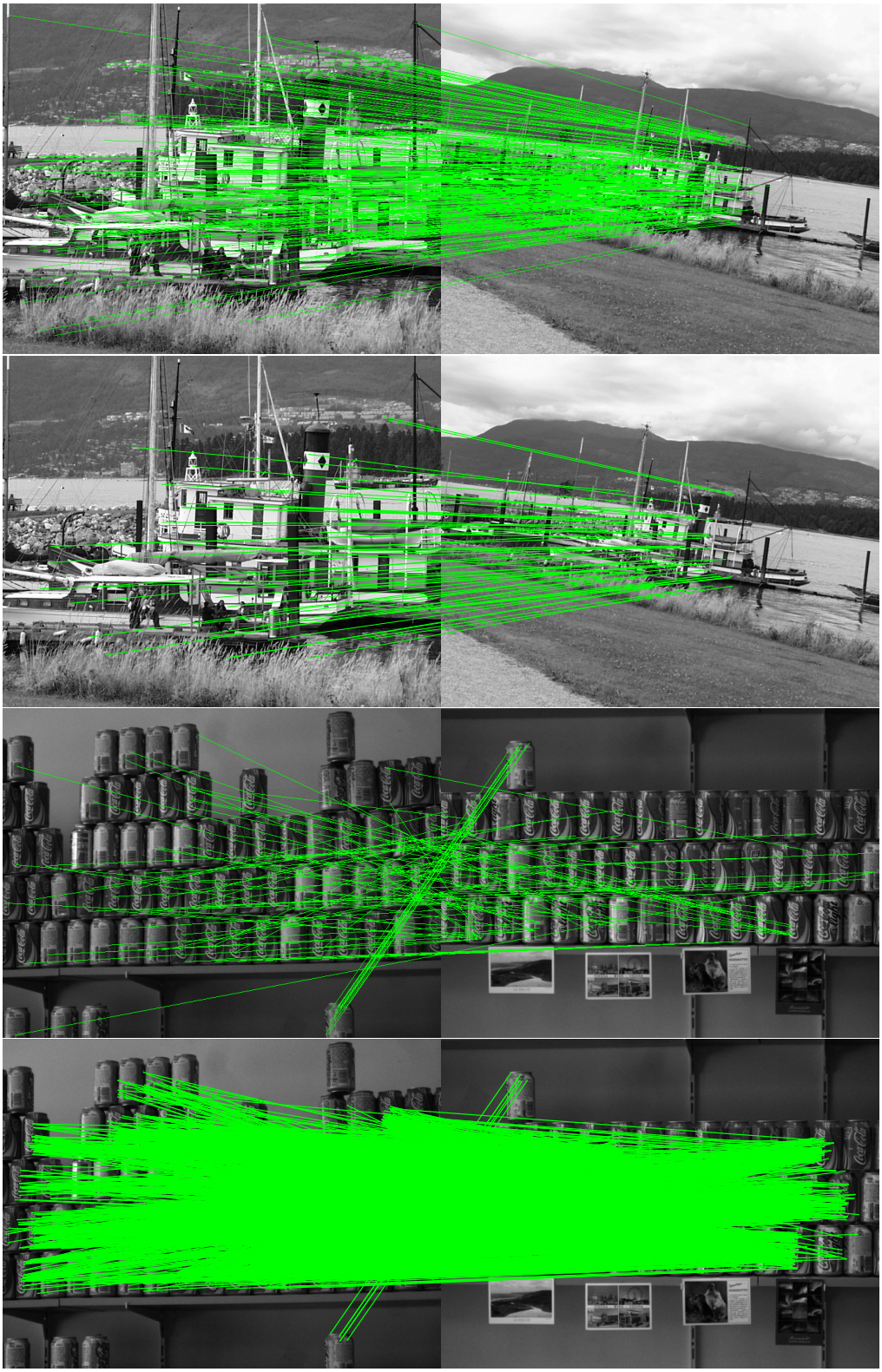

Fig. 2. Comparison of the proposed method with SIFT second nearest neighbour criterion. **1st row:** SIFT. **2nd row:** proposed method. **3rd row:** SIFT. **4th row:** proposed method.

5. REFERENCES

[1] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[2] T. Lindeberg, "Scale selection properties of generalized scale-space interest point detectors," *JMIV*, vol. 46, no. 2, pp. 177–210, 2013.

[3] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *ECCV*, 2012.

[4] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 24:1–24:11, 2009.

[5] A. Mikulík, M. Perdoch, O. Chum, and J. Matas, "Learning a fine vocabulary," in *ECCV*, 2010, pp. 1–14.

[6] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[7] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints," *IJCV*, vol. 66, no. 3, pp. 231–259, 2006.

[8] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *CVPR*, 2007, pp. 1–8.

[9] W. Pratt, "Correlation techniques of image registration," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. AES-10, no. 3, pp. 353–358, 1974.

[10] H. Ardo and K. Astrom, "Bayesian formulation of image patch matching using cross-correlation," in *ICDSC*, 2009, pp. 1–8.

[11] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[12] I. Rey-Otero and M. Delbracio, "Anatomy of the sift method," *Image Processing On Line*, vol. 4, pp. 370–396, 2014.

[13] J. M. Morel and G.Yu, "Asift, a new framework for fully affine invariant image comparison," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 438–469, 2009.

[14] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *TPAMI*, vol. 24, no. 4, pp. 509–522, 2002.

[15] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *IJCV*, vol. 60, no. 1, pp. 63–86, 2004.

[16] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *TPAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[17] P. Moreels and P. Perona, "Evaluation of features detectors and descriptors based on 3d objects," in *ICCV*, 2005, pp. 800–807.

[18] L. Moisan, P. Moulon, and P. Monasse, "Automatic Homographic Registration of a Pair of Images, with A Contrario Elimination of Outliers," *Image Processing On Line*, 2012.

[19] M.Z. Brown, D. Burschka, and G.D. Hager, "Advances in computational stereo," *TPAMI*, vol. 25, no. 8, pp. 993–1008, 2003.

[20] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, and A.Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *J. ACM*, vol. 45, no. 6, pp. 891–923, 1998.

[21] J. Rabin, J. Delon, and Y. Gousseau, "A statistical approach to the matching of local features," *SIAM J. Img. Sci.*, vol. 2, no. 3, pp. 931–958, 2009.

[22] V. Pătrăucean, R. Grompone von Gioi, and M. Ovsjanikov, "Detection of mirror-symmetric image patches," in *CVPRW*, 2013, pp. 211–216.

[23] A. Desolneux, L. Moisan, and J.-M. Morel, *From Gestalt Theory to Image Analysis*, Springer, 2008.

[24] A. P. Witkin and J. M. Tenenbaum, "On the role of structure in vision," in *Human and Machine Vision*, J. Beck, B. Hope, and A. Rosenfeld, Eds., pp. 481–543. Academic Press, 1983.

[25] D. Lowe, *Perceptual Organization and Visual Recognition*, Kluwer Academic Publishers, 1985.

[26] A. Gordon, G. Glazko, X. Qiu, and A. Yakovlev, "Control of the mean number of false discoveries, bonferroni and stability of multiple testing," *Ann. Appl. Stat.*, vol. 1, pp. 179–190, 2007.

[27] V. Pătrăucean, *Detection and Identification of Elliptical Structure Arrangements in Images: Theory and Algorithms*, Ph.D. thesis, Institut National Polytechnique de Toulouse, France, 2012.

[28] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous univariate distributions*, Distributions in statistics. Wiley, New York, NY [u.a.], 2. ed. edition, 1995.