DETERMINISTIC FEATURE DECOUPLING BY SURFING INVARIANCE MANIFOLDS

Eduardo Martínez-Enríquez and Javier Portilla

Instituto de Óptica, CSIC, Madrid

ABSTRACT

We introduce a formalism that justifies and extends a heuristic method for algebraically decoupling deterministic features that recently proved useful for improving feature-based classification. Our new formalism is based on defining transformations inside manifolds, by following trajectories along the features' gradients. Through these transformations we define a *normalization* that, we demonstrate, allows for decoupling differentiable features. By applying this to sampling moments, we obtain a quasi-analytic solution for the *orthokurtosis*, a modification of the kurtosis that is not just decoupled from mean and variance, but also from skewness. After theoretically motivating feature decoupling for random data distributions, we illustrate with a regression problem example how decoupled features may perform significantly better than coupled ones.

***Index Terms*—** nonlinear orthogonal features; sample statistics; regression; manifolds; decoupled features; orthokurtosis

1. INTRODUCTION

In many signal processing situations it is useful to operate with global features of discretized signals (e.g., audio, images, ECG, etc.). We may want to extract [1, 2, 3], impose [4, 5, 6, 7], transfer [7, 8, 9], or, generally, manipulate [10, 11], some global features without suffering the consequences of algebraic coupling between them [12, 13]. We say that some deterministic features are algebraically coupled (or just *coupled*) if their gradients are not mutually orthogonal everywhere. A consequence of coupling is that the set of possible (algebraically compatible) values of a feature depends on the other features' values. For instance, a high sample skewness value is mathematically incompatible with a low sample kurtosis value [14, 15]. Here we present a new formalism that builds upon a feature decoupling method (i.e., a method to make features' gradients mutually orthogonal) presented in [13]. The basic idea is that any set of global features defines a space foliation corresponding to the invariance manifolds obtained by following local linear combinations of the features' gradients, from any vector. We define a *feature translation* operation, consisting of finding the closest intersection of the vector's invariance manifold with the set of vectors having the desired feature values. We say that we *normalize* a vector when we translate its features to some reference values. Then we demonstrate how this normalization can be used to decouple features. By sequentially decoupling each global feature from the previous ones, we obtain a set of embedded mutually orthogonal global features. In this paper we apply these concepts to obtain a quasi-analytical expression for the *orthokurtosis*, a version of the kurtosis that is decoupled from the skewness [15, 13] (previous attempts to do this were [16, 17]). It is also important to note that, when extracting global features from signals, part of the joint structure of the vectors made by those features does not come from the statistical properties (joint pdf) of the signal being analyzed, but from the algebraic coupling of the features themselves. Here we analyze the effects of decoupling features on the statistics in the feature space and on discriminability, in the context of observing random data distributions. Finally, we apply feature decoupling to improve the estimation of the parameters of a distribution, in a regression problem. Using decoupled features we are able to reduce the mean square error (MSE) of the estimation by a factor around 2 in comparison to using the original coupled features.

Funded by the Spanish Government grant FIS2016-75891-P.

2. A NEW SIGNAL REPRESENTATION FRAMEWORK BASED ON MANIFOLDS

2.1. Feature translation by surfing the invariance manifolds

Definition 1: Invariance manifold. A set $\mathcal{S} = \{f_j(\vec{x}), j = 1 \dots M\}$ of M non-local deterministic differentiable scalar functions (or *global features*) defines a lossy transformation $\vec{f} : R^N \rightarrow R^M$, where typically $M << N$. Given a vector $\vec{x}_0 \in R^N$ representing an ordered set of N discrete samples we define the *invariance manifold* $\mathcal{I}(\vec{x}_0; \mathcal{S})$ as the set of all vectors that can be obtained by moving $\vec{x}_0$ along the direction of (any linear combination of) the features' gradients. This manifold, whose tangent space is spanned by the local features' gradients, expresses all possible transformations associated to the feature set $\mathcal{S}$ on $\vec{x}_0$.

Definition 2: Goal manifold. Given a vector $\vec{v}^{des} \in R^M$ containing the desired values for the features in $\mathcal{S}$, we define the *goal manifold*, $\mathcal{G}(\vec{v}^{des}; \mathcal{S}) = \vec{f}^{-1}(\vec{v}^{des})$ as the set of all vectors $\vec{x} \in R^N$ such that $\vec{f}(\vec{x}) = \vec{v}^{des}$.

Definition 3: Feature translation. It is the action of shifting $\vec{x}_0$ along its invariance manifold until reaching the closest intersection with a desired goal manifold:

$$T(\vec{x}_0; \vec{v}^{des}; \mathcal{S}) = \arg \min_{\vec{x} \in \{\mathcal{G}(\vec{v}^{des}; \mathcal{S}) \bigcap \mathcal{I}(\vec{x}_0; \mathcal{S})\}} d_{\mathcal{I}(\vec{x}_0; \mathcal{S})}(\vec{x}, \vec{x}_0), \quad (1)$$

where $d_{\mathcal{I}(\vec{x}_0; \mathcal{S})}$ represents a manifold distance (i.e., the length of the shortest geodesic connecting two points within $\mathcal{I}(\vec{x}_0; \mathcal{S})$). Feature translation fulfils the properties:

- For a given feature set $\mathcal{S}$, all invariance manifolds (regardless of the choice of $\vec{x}_0$) are orthogonal to all goal manifolds (regardless of the choice of $\vec{v}^{des}$). This is because features' gradients are tangent to invariance manifolds and orthogonal to goal manifolds, by definition. As a consequence, their intersection is zero-dimensional (a discrete set of points, in general).
- The solution does not depend on the particular trajectory followed to reach the closest intersection point, as long as the trajectory lies inside the invariance manifold. Because of the intersection is made of isolated points (zero, one or more), there will be infinitely many possible trajectories inside the invariance manifold converging to the closest intersection point.
- $\mathcal{I}(\vec{x}_0;\mathcal{S}) = T^{-1}\Big(T(\vec{x}_0;\vec{v}^{des};\mathcal{S})\Big)$, that is, the invariance manifold passing by $\vec{x}_0$, corresponds to the set of vectors $\vec{x}$ such that when transformed by $T(\vec{x};\vec{v}^{des};\mathcal{S})$ yield the same transformed vector $T(\vec{x}_0;\vec{v}^{des};\mathcal{S})$ as $\vec{x}_0$. This property is what gives its name to the invariance manifold.

Definition 4: Normalization (and reference manifold). A set of reference values for the feature set, $\vec{v}^{ref}$, with the property that $T(\vec{x};\vec{v}^{ref};\mathcal{S})$ exists for all $\vec{x} \in R^N$ (except possibly for a zero-measurement set of singular vectors), defines a *normalization* $\hat{x}(\vec{x}) = T(\vec{x};\vec{v}^{ref};\mathcal{S})$, a special kind of feature translation. We name *reference manifold* the goal manifold used for normalization, $\mathcal{R}(\vec{v}^{ref};\mathcal{S}) = \check{f}^{-1}(\vec{v}^{ref})$.

2.2. Decoupling features through normalization

Proposition 2.1. Decoupling global features

Given a set $\mathcal{S}$ made of J differentiable and non-trivially redundant features, we want to transform the J-th feature such as its gradient becomes orthogonal to the gradients of the other $J-1$ features everywhere. Given a reference manifold $\mathcal{R}(\vec{v}^{ref};\mathcal{S}^*)$, being $\mathcal{S}^*$ the set obtained by removing the J-th feature from $\mathcal{S}$, the scalar function $\check{f}_J(\vec{x}) = f_J(T(\vec{x};\vec{v}^{ref};\mathcal{S}^*))$ fulfils that $\nabla\check{f}_J(\vec{x}) \cdot \nabla f_j(\vec{x}) = 0$, for all $\vec{x} \in R^N$, for $j = 1 \dots J-1$.

Proof. Because vectors in the invariance manifold for a given $\vec{x}_0$ have all the same transformed vector $\hat{x}(\vec{x}) = T(\vec{x};\vec{v}^{ref};\mathcal{S}^*)$, then all of them have the same value of the new feature $\check{f}_J$ (as it is just evaluating f_J on the same vector $\hat{x}(\vec{x})$). As a consequence, these manifolds are (or are included in) iso-level hyper-surfaces of $\check{f}_J$. Considering smoothness of all the involved functions, this implies that its gradient must exist and be orthogonal to the iso-level hypersurfaces, and to any iso-level manifolds inside them, which means to be orthogonal to the invariance manifolds' tangent spaces. As tangent spaces are the span of the features' gradients ∇f_j, from 1 to $J-1$, this finally implies the orthogonality of $\nabla\check{f}_J$ to each of these gradients. □

Normalization can be regarded as a transformation that removes the information (in an algebraic, not statistical, sense) about those features whose reference values are imposed. Intuitively, it makes sense that any measurement we make on a vector normalized in some of its features will be algebraically decoupled from those features. Figure 1 illustrates the method for decoupling a given second feature from a given first one.

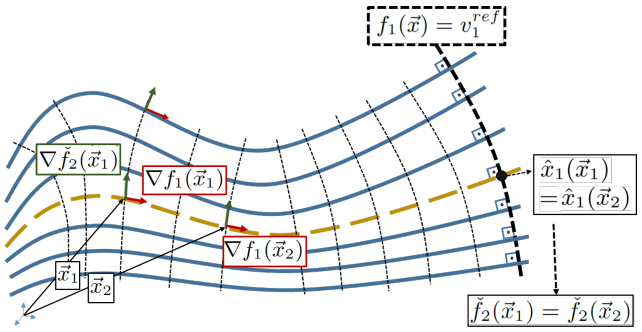

Fig. 1: Decoupling features through normalization. $\vec{x}_1$ and $\vec{x}_2$ belong to the same invariance manifold (yellow dashed curve), having the same normalized vector $\hat{x}_1$ and, thus, the same $\check{f}_2$. $\nabla\check{f}_2$ is orthogonal to iso-levels of $\check{f}_2$, and thus to ∇f_1. Normalizing is finding the intersection of the invariance manifold with $f_1(\vec{x}) = v_1^{ref}$.

In [13] we proposed the *Nested Normalizations* (NNs) method for decoupling features. NNs is based on sequentially adjusting each feature to its reference value, by following the projected feature's gradient on the orthogonal complement of the previous features gradients (so keeping unchanged the previously adjusted reference values). Mutually decoupled features are obtained by computing the original features on the corresponding normalized vectors: $\check{f}_j(\vec{x}) = f_j(\hat{x}_{j-1}(\vec{x}))$ (starting with $\hat{x}_0(\vec{x}) = \vec{x}$). Although in that paper we wrongly identified the orthogonalized gradients' directions with those of the orthofeatures gradients[1], that mistake does not affect the validness of the NNs algorithm itself, which is fully consistent with the theory explained in this section.

2.3. A quasi-analytic solution for the *orthokurtosis*

In [13] we attacked the problem of decoupling the marginal moments $f_j(\vec{x}) = \mu(\vec{x}^{\odot j})$, being $\mu()$ the sample mean operator and $\odot$ a pointwise vector operation indicator. However, our initial framework did not allow us to obtain an analytical solution for the adjustment of the skewness, a necessary step for obtaining an expression for the next decoupled moment, the *orthokurtosis*. We use as reference values the moments of a zero-mean univariate Gaussian. By solving the ODEs moving along the moments' gradient directions [13] we obtain the first normalizations:

$$\hat{x}_0(\vec{x}) = \vec{x} \quad (2)$$

$$\hat{x}_1(\vec{x}) = \hat{x}_0(\vec{x}) - \mu(\hat{x}_0(\vec{x})) \quad (3)$$

$$\hat{x}_2(\vec{x}) = \hat{x}_1(\vec{x}) / \sqrt{\mu(\hat{x}_1(\vec{x})^{\odot 2})}. \quad (4)$$

[1] These coincide in especial cases, like marginal moments up to order 3.

In addition, the gradients of the resulting orthogonal features (mean, variance and skewness) are:

$$\nabla \check{f}_1(\vec{x}) \propto \vec{1} \quad (5)$$

$$\nabla \check{f}_2(\vec{x}) \propto \hat{x}_1(\vec{x}) \quad (6)$$

$$\nabla \check{f}_3(\vec{x}) \propto \hat{x}_2(\vec{x})^{\odot 2} - \check{f}_3(\vec{x})\hat{x}_2(\vec{x}) - 1. \quad (7)$$

In this case we see that the linear span of these gradients (and, thus, the corresponding invariance manifold) is the same as that of the original moments. Thus, we can use for the normalization *any* concatenation of solutions of ODE equations using the *original* features' gradients that impose the reference values. In particular, we can adjust the third order moment along its gradient, by solving the Ricatti equation $d\vec{x}(t)/dt = \vec{x}(t)^{\odot 2}$, whose solution is

$$\vec{x}(t) = \vec{x}(0) \odot /(1 - t\vec{x}(0)). \quad (8)$$

Now we just normalize the inferior moments, which do not affect the skewness: $\hat{x}_3(\vec{x}) = \hat{x}_2(\vec{x} \odot /(1 - t_0(\vec{x})\vec{x}))$, with $t_0(\vec{x})$ a function forcing the skewness to achieve its reference value (0, in this case), $t_0(\vec{x}) = \arg_t\{\check{f}_3(\vec{x} \odot /(1 - t\vec{x})) = 0\}$ [2]. Finally, the *orthokurtosis* is: $\check{f}_4(\vec{x}) = f_4(\hat{x}_3(\vec{x}))$ [3].

3. DECOUPLING FEATURES FOR ANALYSIS

Here we show how features' decoupling removes local covariance in the feature space, and how this improves discrimination.

3.1. Covariance-free "balls" in the feature space

Let $\vec{x} \in \mathbb{R}^N$ be a random vector made of N i.i.d. samples obeying a probability distribution $p(x)$. Let us assume a feature set $\mathcal{S}$ of M global features $\{f_j\}$. Define a vector $\vec{c} \in R^M$ containing the expected value of the features $f_j(\vec{x})$ for different realizations of $\vec{x}$, i.e., $c_j = \mathbb{E}\{f_j(\vec{x})\}$. Define a goal manifold $\mathcal{G}(\vec{c}; \mathcal{S}) = \vec{f}^{-1}(\vec{c})$, i.e., the manifold containing the set of all vectors $\vec{x}$ having the same $\vec{c}$. Describe vector samples as $\vec{x}_i = \vec{x}_{0i} + \vec{d}_i$, where $\vec{x}_{0i} \in \mathcal{G}$, and $\vec{d}_i$ is a (relatively small) sampling fluctuation, with $\mathbb{E}\{d_k d_l\} = 0, k, l \in [1, \ldots, N]$, for all k-th and l-th components of vector $\vec{d}_i$.

Proposition 3.1. Decoupled features are locally uncorrelated *Under previous assumptions, for N large, decoupled features will have uncorrelated deviations from their expected values, i.e.,* $\mathbb{E}\{(\check{f}_n(\vec{x}) - c_n)(\check{f}_m(\vec{x}) - c_m)\} = 0, n, m \in [1, \ldots, M]$.

Proof. For N large, features' values will not deviate much from their expected values and thus vector samples $\{\vec{x}_i\}$ will be located in the vicinity of $\mathcal{G}$. A first order local approximation yields

$$f_j(\vec{x}_i) \approx f_j(\vec{x}_{0i}) + \nabla f_j(\vec{x}_{0i}) \cdot \vec{d}_i = c_j + \nabla f_j(\vec{x}_{0i}) \cdot \vec{d}_i. \quad (9)$$

Therefore, $\mathbb{E}\{(f_n(\vec{x}) - c_n)(f_m(\vec{x}) - c_m)\} \approx \mathbb{E}\{(\nabla f_n(\vec{x}_0) \cdot \vec{d}(\vec{x}_0))(\nabla f_m(\vec{x}_0) \cdot \vec{d}(\vec{x}_0))\}$, yielding the covariance:

$$cov_{n,m} = \sigma_d^2 \mathbb{E}\{\nabla f_n(\vec{x}_0) \cdot \nabla f_m(\vec{x}_0)\}, \quad (10)$$

where σ_d^2 is the expected quadratic dispersion of the features fluctuations. In the decoupled features case gradients are mutually orthogonal, and thus vector differences for the different features will be uncorrelated. In contrast, when using coupled features, $\vec{d}_i$ is projected onto non-orthogonal directions, leading to correlated sampling fluctuations in the feature space. □

Figure 2 shows an example (also used in next section) of Gaussian distributions, of two features before (skewness and kurtosis) and after (skewness and *orthokurtosis*) being decoupled.

3.2. Features' covariance and discriminability

Let us assume now that our pdf depends on a parameter θ, $p(x; \theta)$. How well can we discriminate samples coming from similar values of θ, based on some global features? Consider a mapping $s_{(\theta,\theta_0)} : \mathbb{R} \to \mathbb{R}$ such that $x(\theta) = s_{(\theta,\theta_0)}(x_0)$, with $x(\theta) \sim p(x; \theta)$ and $x_0 \sim p(x; \theta_0)$, and a vector of global features $\vec{f}(\vec{x})$. This allows us to study the dependency of the feature vector $\vec{f}$ on θ, by expressing:

$$\frac{d\mathbb{E}\{\vec{f}(\vec{x}(\theta))\}}{d\theta} = \mathbb{E}\left\{\mathbf{J}_{\vec{f}}(\vec{x}(\theta))\frac{d\vec{x}(\theta)}{d\theta}\right\}, \quad (11)$$

where $\mathbf{J}_{\vec{f}}$ is the Jacobian matrix of $\vec{f}(\vec{x})$. On the other hand, from Eq.(10) we can write the whole covariance matrix $\mathbf{C}(\theta)$ of the features fluctuations, as:

$$\mathbf{C}(\theta) = \sigma_d^2 \mathbb{E}\{\mathbf{J}_{\vec{f}}(\vec{x}(\theta))\mathbf{J}_{\vec{f}}^T(\vec{x}(\theta))\}. \quad (12)$$

The connection between Eqs.(11) and (12) is now clear: if we choose a set of global features highly sensitive to changes in parameter θ, then the first right singular vector of $\mathbf{J}_{\vec{f}}(\vec{x})$ (the one causing the largest change in $\vec{f}$) will be roughly aligned with $d\vec{x}(\theta)/d\theta$, on average. As a consequence, $d\mathbb{E}\{\vec{f}(\vec{x}(\theta))\}/d\theta$ will also be roughly aligned with the first *left* singular vector of $\mathbf{J}_{\vec{f}}(\vec{x})$, which, given Eq. (12), is just the first eigenvector (the dominant covariance direction) of $\mathbf{C}(\theta)$. Such an alignment between the feature vector local covariance and the expected feature vector curve, implies that heavily coupled features will have strongly overlapping pdf's along the curve of expected values for different parameter's values in the feature space. Fig. 2(left) illustrates this phenomenon. Fig. 2(right) shows the effect of decoupling the kurtosis from the skewness (*orthokurtosis*). We used 128 random vectors of 1024 i.i.d. samples each, from $x(\theta) = x_0^\theta$, being $x_0 \sim U(0, 1)$. Ellipses correspond to a Mahalanobis radius of 2, and $\theta = 5$ (black), 6 (blue), and 7 (red). Error probabilities are 12.4% (coupled) vs. 4.5% (decoupled).

4. EXPERIMENTS AND DISCUSSION

Different approaches have been proposed in the literature to estimate the parameters of a distribution that best describe a dataset, such as the classical maximum likelihood estimation (MLE) or the method of moments (MoMs). Here we have used, in addition to the above mentioned, Support Vector Regression (SVR)

[2] Using Eq. 8 we can adjust the 3rd order moment to any real value with $t \in (1/\min(\vec{x}), 1/\max(\vec{x}))$, whenever $\max(\vec{x}) > 0, \min(\vec{x}) < 0$.

[3] Matlab code available in https://www.researchgate.net/project/Nested-Normalizations-for-Decoupling-Global-Features

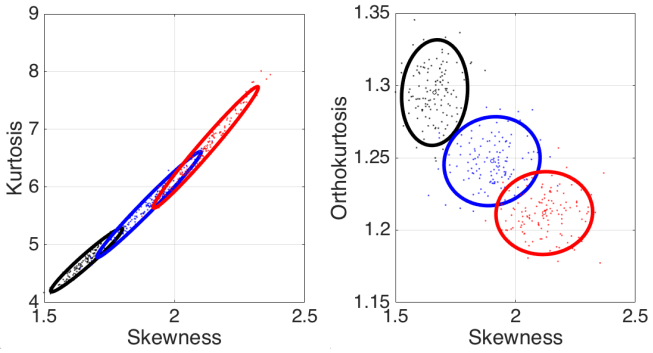

Fig. 2: Comparing classes in the original feature space (left) and in the decoupled feature space (right).

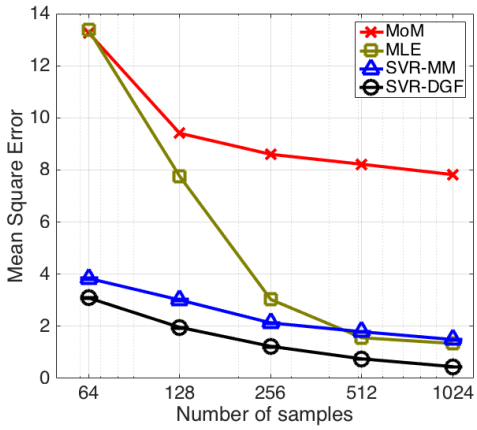

Fig. 3: Mean square error in the estimation of the shape parameter (SP) as function of the sample size.

[18]. In these experiments we used the skew normal (SN) distribution [19], which is a generalization of the normal distribution that, by means of its *shape parameter* (SP), has non-zero skewness (when SP$\neq$0) [19]. Let $\vec{x}$ represent an N-D vector of i.i.d. samples drawn from a SN distribution (we generated the samples following [20]) with location, scale, and SP parameters sampled from uniform distributions in $[-2, 2]$, $[1, 5]$ and $[-10, 10]$, respectively. Marginal standardized moments (MM) and their corresponding decoupled features (DGF) were obtained from the mean to kurtosis/*orthokurtosis*, leading to two sets of predictors $\{f_j(\vec{x}), j = 1, \ldots, 4\}$ and $\{\check{f}_j(\vec{x}), j = 1, \ldots, 4\}$ for each parameter triplet {location, scale, SP}. We compared the SP prediction accuracy of these two sets. We generated $d = 2048$ vectors $\vec{x}$ thus having 2048 pairs of predictors ($\{f_j\}$ or $\{\check{f}_j(\vec{x})\}$) and targets (known SP values). We used a Gaussian kernel for SVR, and averaged 100 5-fold cross-validation runs to measure the MSE of the estimated SPs in both cases. For comparison purposes, the SP was also estimated with the MoMs, using the sample skewness [19], and with the MLE. Vector realizations with absolute sample skewness greater than 0.99 (for which the MoMs may not have solution [19]) were discarded from all the methods.

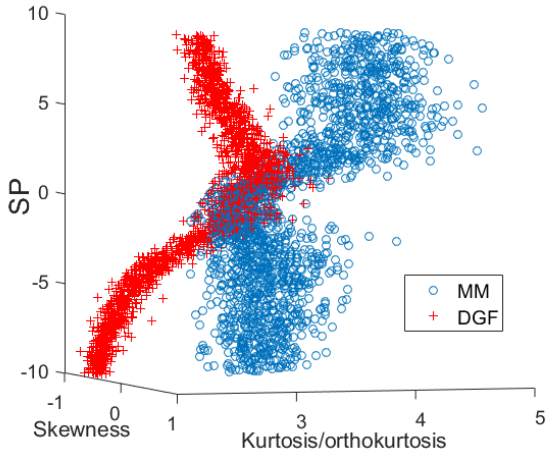

Fig. 4: Scatter plot of extracted features in the joint space with the shape parameter (SP). We show 2048 dots, each obtained from a vector of 1024 i.i.d. skew normal samples.

Figure 3 shows the MSE for the compared methods and sample sizes N. SVR-based estimations clearly outperformed both MLE and MoMs in this case. MLE estimation worked better than MoMs for $N > 64$. Nevertheless, given the difficulty of the optimization, the MLE result strongly depended on the initialization values and restrictions in the allowed solutions. We used the *mle.m* ®Matlab function for the optimization, using MoMs estimate as initialization, and restricting the solutions to $[-10.1, 10.1]$. Specialized algorithms, as those based on genetic algorithms proposed in [21], can be used instead. The SVR approach using decoupled features clearly outperformed all compared methods; e.g., for $N = 256$, with MSE reduction by factors of 1.84, 2.52 and 7.80, with respect to the original features (SVR-MM), MLE and MoMs, respectively. Specificially, the MSE was consistently lower using decoupled features (*orthomoments*) than using the classical standardized moments. Figure 4 shows a set of feature vectors (skewness and kurtosis/*orthokurtosis*) for random SP values. The lower dispersion observed in the decoupled features in all the SP range explains the reduction in MSE.

5. CONCLUSIONS

We have proposed a conceptual framework based on normalizing features using invariance manifolds, which extends and provides more formal rigor to our previous work on algebraically decoupling global features. We have obtained a quasi-analytic expression for the *orthokurtosis*, a new feature obtained from the classical kurtosis that is decoupled from the skewness. We have also theoretically analyzed the implications of using decoupled features in data analysis, and obtained significant improvements in a regression experiment, in comparison to using classical (non-decoupled) features. These new results reinforce our previous results in texture classification, and, we believe, they jointly provide a solid evidence of the positive practical impact of feature decoupling for signal analysis.

6. REFERENCES

[1] B. Julesz, "Visual pattern discrimination," *IRE Transactions on Information Theory*, vol. 8(2), pp. 84–92, 1962.

[2] R. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, no. 6, 1973.

[3] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, pp. 51–59, 1996.

[4] J. Portilla, R. Navarro, O. Nestares, and A. Tabernero, "Texture synthesis-by-analysis based on a multiscale early-vision model," *Optical Engineering*, vol. 35, no. 8, pp. 2403–2417, 1996.

[5] D. J. Heeger and J. R. Bergen, "Pyramid-based texture analysis/synthesis," in *Proceedings., International Conference on Image Processing*, Oct 1995, vol. 3, pp. 648–651 vol.3.

[6] S. C. Zhu, Y. N. Wu, and D. Mumford, "Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling," *International Journal of Computer Vision*, vol. 27, no. 2, pp. 107–126, Mar 1998.

[7] J. Portilla and E.P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, vol. 40(1), pp. 49–71, 2000, ®Matlab code publicly available at http://www.cns.nyu.edu/ lcv/texture/.

[8] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2414–2423.

[9] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 3320–3328. Curran Associates, Inc., 2014.

[10] R. Hummel, "Image enhancement by histogram transformation," *Computer Graphics and Image Processing*, vol. 6, no. 2, pp. 184 – 195, 1977.

[11] J. Portilla and E. P. Simoncelli, "Image denoising via adjustment of wavelet coefficient magnitude correlation," in *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, Sep. 2000, vol. 3, pp. 277–280 vol.3.

[12] J. Balas, "Texture synthesis and perception: Using computational models to study texture representations in the human visual system," *Vision Research*, vol. 46, no. 3, pp. 299 – 309, 2006.

[13] J. Portilla and E. Martinez-Enriquez, "Nested normalizations for decoupling global features," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 2112–2116.

[14] K. Pearson, "IX. Mathematical contributions to the theory of evolution.—XIX. Second supplement to a memoir on skew variation," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 216, no. 538-548, pp. 429–457, 1916.

[15] R. Sharma and R. Bhandari, "Skewness, kurtosis and Newton's inequality," *ArXiv e-prints*, Sept. 2013.

[16] D. C. Blest, "A new measure of kurtosis adjusted for skewness," *Australian and New Zealand Journal of Statistics*, vol. 45, no. 2, pp. 175–179, 2003.

[17] J.F. Rosco, A.Pewsey, and M. C. Jones, "On Blest's measure of kurtosis adjusted for skewness," *Communications in Statistics: Theory and Methods*, vol. 44, no. 17, pp. 3628–3638, 2015.

[18] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug 2004.

[19] A. Azzalini, "A class of distributions which includes the normal ones," *Scandinavian Journal of Statistics*, vol. 12, pp. 171–178, 1985.

[20] D. Ghorbanzadeh, L. Jaupi, and P. Durand, "A method to simulate the skew normal distribution," *Applied Mathematics*, vol. 5, pp. 2073–2076, 2014.

[21] A. Yalçinkaya, B. Şenoğlu, and U. Yolcu, "Maximum likelihood estimation for the parameters of skew normal distribution using genetic algorithm," *Swarm and Evolutionary Computation*, vol. 38, pp. 127 – 138, 2018.