A Statistical Framework for Dimensionality Reduction Implementation in FPGAs

Christos-S. Bouganis [#1], Iosifina Pournara [*2], Peter Y.K. Cheung [#3]

[#]*Department of Electrical and Electronic Engineering*
Imperial College London, University of London
South Kensington campus, London SW7 2AZ, UK
[1]ccb98@imperial.ac.uk
[3]p.cheung@imperial.ac.uk

[*]*Department of Crystallography, Birkbeck College, University of London*
Malet Street, London WC1E 7HX, UK
[2]i.pournara@cryst.bbk.ac.uk

***Abstract*— Dimensionality reduction or feature extraction has been widely used in applications that require a set of data to be represented by a small set of variables. A linear projection is often chosen due to its computational attractiveness. The calculation of the linear basis that best explains the data is usually addressed using the Karhunen-Loeve Transform (KLT). Moreover, for applications where real-time performance and flexibility to accommodate new data are required, the linear projection is implemented in FPGAs due to their fine-grain parallelism and reconfigurability properties. Currently, the optimization of such a design in terms of area usage is considered as a separate problem to the basis calculation. In this paper, we propose a novel approach that couples the calculation of the linear projection basis and the area optimization problems under a probabilistic Bayesian framework. The power of the proposed framework is based on the flexibility to insert information regarding the implementation requirements of the linear basis by assigning a proper prior distribution. Results using real-life examples demonstrate the effectiveness of our approach.**

I. Introduction

In many scientific fields, it is required to represent a set of data using a small number of variables. This problem is usually referred as *dimensionality reduction* or *feature extraction*. Examples can be found in the face recognition/detection problem [1] where images of people are mapped into a smaller space than the original one that captures the main characteristics of the faces, in image compression [2] and others.

The dimensionality reduction is often achieved by the linear projection of the new data to a smaller space than the original that is defined by a basis matrix. The basis of the new subspace is constructed using a set of data targeting a certain error in the data approximation. The data can be recovered from the subspace as in (1), where $x \in Z^P$ denotes the data vector with P elements, Λ denotes an orthogonal basis with dimensions $P \times K$, and $f \in Z^K$ denotes the factors vector. Note that K is usually much smaller than P.

$$x = \Lambda f \tag{1}$$

Many applications require the Λ matrix to be orthogonal, that is $\Lambda^T\Lambda = I$ and $\Lambda^{-1} = \Lambda^T$, where I denotes the identity matrix. The orthogonality of Λ implies that the factor vector f that corresponds to a data x can be calculated by using the same matrix Λ as shown in (2), reducing the amount of coefficients that need to be stored in the system.

$$f = \Lambda^T x \tag{2}$$

In many applications, the large dimensionality of matrix Λ, *i.e.* 500×40 for face detection applications, and the real-time performance requirements of the algorithm, lead to hardware based solutions. FPGAs are often used to achieve this goal due to their fine grain parallelism and reconfigurability. The current work addresses the case where maximum performance for the evaluation of (1) in terms of speed is required, thus only pipelined designs are considered.

The problem under consideration is the calculation of an orthogonal basis matrix Λ such that the required area for implementation of (1) is minimized, achieving at the same time a specific error in the approximation of the training data that is specified by the user.

Current techniques for mapping the above process into FPGAs treat the problem as a two steps process. Firstly, the appropriate subspace is calculated in the floating-point domain as the one that provides the best approximation of the data. Then, the elements of the Λ matrix are quantized for mapping into hardware. In order to optimize the design in terms of the required area, the elements of the basis are often encoded using Canonic Signed Digit recoding [3] or subexpression elimination [4]. The main drawback of the current approach is that the design process for the basis calculation does not take into account the hardware restrictions leading to suboptimal solutions.

In this paper we propose a novel framework where the two steps of subspace estimation and hardware implementation are considered simultaneously, allowing us to target area optimized designs. This is achieved by formulating the problem of the subspace calculation in a Bayesian framework, where the cost of the implementation of the necessary components in Λ matrix is inserted to the system as a prior information. Experiments using real data from computer vision applications demonstrate the effectiveness of the proposed framework.

In summary, the original contributions of this paper are:

- the combination of the subspace estimation problem and the area optimization problem for hardware realization under a uniform Bayesian framework,
- the exploration of a family of functions for incorporating the implementation cost of the system into a Bayesian framework.

The paper is structured as follows. Section II gives a description of current work in the field. Section III describes the proposed Bayesian factor analysis framework, where Section IV discusses the requirements of the functions that map the area implementation information to a probability distribution. Section V presents the evaluation results of the proposed algorithm, where Section VI concludes the paper.

II. Related work

The problem of dimensionality reduction can be formulated as follows. Given a set of N data $x^i \in R^P$, with $X = [x^1, x^2, \cdots, x^N]$, the goal is to find a subspace Λ with dimensions $P \times K$ such that the original data can be expressed as in (3), such that $\mathcal{E}\{E^2\}$ is minimized, where E denotes the error in the approximation and $\mathcal{E}\{\cdot\}$ denotes the mean operator.

$$X = \Lambda F + E \tag{3}$$

The matrix Λ is called *basis matrix* or *factor loadings*, where the matrix F is called *factor matrix* and has dimensions $K \times N$. Out of all possible decompositions, we are seeking the most compact orthogonal matrix Λ, in terms of the number of columns, for a given target mean square reconstruction error.

Current methodology applies the Karhunen-Loeve Transform (KLT) [5] for the calculation of the Λ matrix. It has been shown that the KLT transform produces the most compact representation of the original data assuming that the error has the same power in all dimensions.

The motivation behind this work is demonstrated graphically in Figure 1. The two dimensional data can be expressed using a one-dimension space achieving small error in their approximation. The current methodology that is based on the KLT algorithm, finds that the best basis to describe the data is $W = [0.52; 0.49]$, without taking into account the implementation cost associated with such a basis. However, the basis $W = [0.5; 0.5]$ requires considerable less area and at the same time approximates the original data achieving less error than the required.

In this paper we propose a novel framework that is based on a Bayesian formulation of a factor analysis model for dimensionality reduction where the cost of the hardware implementation of the elements in the Λ matrix is minimized. The proposed approach addresses the problem of dimensionality reduction in FPGAs in a unified framework allowing a better exploration of the design space regarding the approximation of the data versus the cost of the implementation. Moreover, the error in each dimension of the original space is assumed to be independent from the errors in the other dimensions. This provides a larger flexibility than the KLT transform, where the power of the error is assumed to be the same in all dimensions.

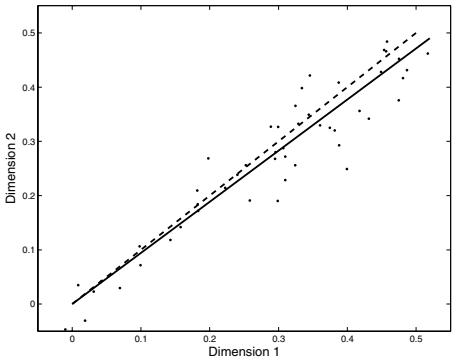

Fig. 1. Projection of 2D data into 1D space. The KLT algorithm finds a basis $W = [0.52, 0.49]$ that corresponds to the solid line and best explains the data in the floating point domain. However, in a fixed-point domain, the basis $W = [0.5, 0.5]$, which corresponds to the dashed line, leads to a design that requires less area than the KLT basis.

III. Bayesian Factor Analysis Model

Let's assume that we have a random observed vector x with dimensions $P \times 1$. An instance of this vector is denoted as x^i, and we assume that we have N such instances x^i where $i = 1, \ldots, N$. We denote as $f = (f_1, \ldots, f_K)'$ a vector of K variables, known as *factors*, where f^i denotes an instance of this vector. Note that the number K of factors is always smaller or equal to the number P of observed variables. The factor analysis model states that the observed variables are a linear combination of the factors plus a mean and an error term. For an instance i, that is

$$x^i = \mu + \Lambda f^i + \epsilon^i \tag{4}$$

where $\mu = (\mu_1, \ldots, \mu_P)'$ and $\epsilon^i = (\epsilon^i_1, \ldots, \epsilon^i_P)'$ are both column vectors of P dimensions with each element corresponding to the mean and the error term of each observed dimension, respectively. The vector μ is the same for all cases i. In our case, we centralize the data, which implies that the mean vector is zero, and it will be discarded in the rest of the paper. Λ is the unobserved basis matrix or more often referred to as the *factor loadings matrix*. The factor loadings matrix has $P \times K$ dimensions. That is, each column corresponds to a factor and each row corresponds to an observed variable. The entries of the factor loadings matrix indicate the strength of the dependence of each observed variable on each factor. For example, if λ_{pk} is zero, then variable x_p is independent of factor f_k.

Factor analysis models assume that the error terms ϵ^i are independent, and multivariate normally distributed with mean zero and covariance matrix Ψ. Thus the probability distribution of x for each observed case i has a multivariate normal density given by

$$\begin{aligned} p(x^i|f^i, \Lambda, \Psi) &= \mathcal{N}(x^i|\Lambda f_i, \Psi) \\ &= (2\pi)^{-P/2}|\Psi|^{-1/2} \times \\ &\quad \exp\left(-\frac{1}{2}(x^i - \Lambda f^i)'\Psi^{-1}(x^i - \Lambda f^i)\right) \end{aligned} \tag{5}$$

In the following subsections, we discuss the prior and posterior probabilities of the parameters F, Λ and Ψ.

A. Factors

The factors are assumed to be normally distributed with mean zero and covariance matrix Σ_F. That is,

$$f^i \sim \mathcal{N}(0, \Sigma_F)$$

The posterior probability of the factors is now derived as

$$p(f^i|x^i, \Lambda, \Psi) \propto p(f^i)p(x^i|f^i, \Lambda, \Psi) = \mathcal{N}(f^i|m_F^*, \Sigma_F^*)$$

where the posterior mean and variance are given by

$$\begin{aligned} \Sigma_F^* &= (\Sigma_F + \Lambda'\Psi^{-1}\Lambda)^{-1} \\ m_F^* &= \Sigma_F^*\Lambda'\Psi^{-1}x^i \end{aligned}$$

By expressing (5) in matrix notation and integrating over F, we get the complete density of the data as

$$\begin{aligned} p(X|\Lambda, \Psi) &= \mathcal{N}(X|\Lambda\Sigma_F\Lambda' + \Psi) \\ &= (2\pi)^{-N/2}|\Lambda\Sigma_F\Lambda' + \Psi|^{-1/2} \times \\ &\quad \exp\left(-\frac{1}{2}\text{tr}[X'(\Lambda\Sigma_F\Lambda' + \Psi)^{-1}X]\right) \end{aligned} \tag{6}$$

As shown in (6), the complete density of the data is given by a normal distribution with covariance matrix $\Lambda\Sigma_F\Lambda' + \Psi$. There is a scale identifiability problem associated with Λ and Σ_F. In order to avoid this problem, we can either restrict the columns of Λ to unit vectors or set Σ_F to the identity matrix. The second approach is often used in factor analysis, and which is also adopted here.

B. Factor loadings matrix Λ

The main advantage of the proposed framework lies on the flexibility in selecting the prior distribution of the factor loadings matrix Λ.

We aim to identify a factor loadings matrix that can represent faithfully the data in the high dimension space, but at the same time provides an optimized hardware design in terms of area usage. The suggested prior is a function of the area that is required for implementing a LUT based multiplier in an FPGA. In order to reduce the computational complexity of the algorithm, it is assumed that the variables in the Λ matrix are independent. This assumption allows us to express the probability distribution of the Λ matrix as the product of the probabilities of the individual elements as in (7).

$$p(\Lambda) = \prod_{p=1}^{P}\prod_{k=1}^{K} p(\lambda_{pk}) \tag{7}$$

In the current work, we select the prior probability distribution $p(\lambda_{pk})$ to be inverse proportional to the area that is required for the realization using LUTs of a multiplication by λ_{pk}. The function $g(\cdot)$ relates the area, $A(\lambda_{pk})$, required by the constant coefficient multiplier to the prior probability distribution as $p(\lambda_{pk}) = g(A(\lambda_{pk}))$. The selection of the function g is discussed below. The posterior probability of each element λ_{pk} of Λ is given by

$$p(\lambda_{pk}|X, F, \Psi) = p(X|F, \Lambda, \Psi)\prod_{p=1}^{P}\prod_{k=1}^{K} p(\lambda_{pk}).$$

The above distribution does not have a known form, thus we have to calculate it for all possible values of λ_{pk} and then use the uniform probability distribution to sample the new value of λ_{pk}.

C. Noise covariance matrix Ψ

A convenient conjugate prior is assigned to the inverse of the noise covariance matrix Ψ so that its posterior distribution has a known form. The Gamma distribution is selected as the prior distribution for each ψ_p^{-2}.

D. Orthogonality

The above statistical framework does not necessarily produces an orthogonal basis of the new space. However, in computer vision field, which is our target domain, this condition is often required. Under the proposed framework, this requirement is enforced by finding firstly the direction that mostly explains the data, that is the direction with the maximum variance, and then by projecting the data to the obtained space and retrieving them back. This process is repeated in order to calculate each vector in the new space. Moreover, the above technique minimizes the error due to quantization of the factor loadings Λ and factors F. A similar methodology has been proposed by Bouganis *et al.* [6], [7] for 2D filter design exploration.

IV. Mapping implementation cost to prior distribution

The prior distribution for the Λ matrix has to be a valid distribution, that is:

$$p(\lambda_{pk}) \geq 0 \ ,\forall\lambda_{pk} \tag{8}$$

and

$$\sum_{\lambda_{pk}} p(\lambda_{pk}) = 1 \tag{9}$$

Thus, we are seeking for functions that map the space of the area cost to the space of valid distributions. These functions should be (a) monotonically decreasing, (b) no negative and (c) sum to one. In the rest of the paper we use the family of functions to map the area required by a constant coefficient multiplier to a valid distribution as shown in (10), where c is a constant and ensures that $\sum_{\lambda_{pk}} g(A(\lambda_{pk})) = 1$.

$$g(A(\lambda_{pk})) = c\,(A(\lambda_{pk}))^{-a} \quad , \quad a, c > 0 \tag{10}$$

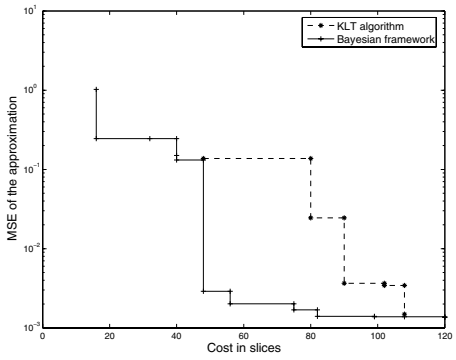

Fig. 2. Mean square error of the data approximation versus the required area for mapping data from R^3 space to Z^2 space.

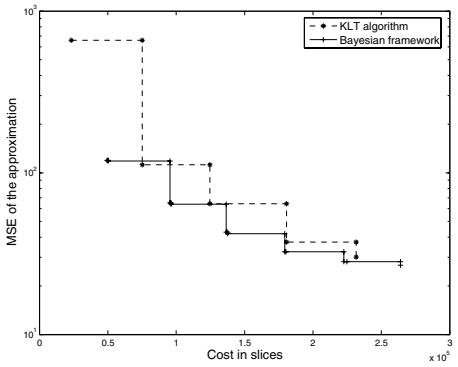

Fig. 3. Mean square error of the data approximation versus the required area for a face recognition application. The algorithm maps the data from Z^{500} space to Z^{40} space.

V. Performance Evaluation

The proposed framework is evaluated for its performance when the number of dimensions of the target space is fixed and is specified by the user. This scenario can arise in applications where the number of dimensions of the factors has to be restricted *i.e.* image compression [2]. From the hardware perspective, this can also be enforced due to the available memory bandwidth in the system where the factors F are stored.

We have compared our proposed framework with the current available approach which is based on the KLT transform and subsequent quantization of the Λ and F matrices. In addition, the reference algorithm has been extended to search the space of possible basis by varying the woldlength of the elements in the Λ matrix and imposing a common wordlength for all the elements. The factors F are quantized to 8 bits, in both the proposed framework and the reference algorithm. In all the cases, it is assumed that the input data have 8 bits wordlength, which is common for image processing applications. The error in the final approximation is calculated by projecting the data back to the original space after having calculated and quantized the factors F.

Figure 2 illustrates the performance of the proposed Bayesian framework and the reference algorithm for mapping data from the R^3 space to Z^2 subspace. The figure plots the achieved mean square error approximation of the data as a function of the number of slices that are required for implementation of the system. The proposed framework can reduce the required area up to a factor of two, achieving at the same time the same mean square error in the data approximation with the reference algorithm. It should be noted that the reduction in the area that is achieved by applying the proposed framework depends on the data to be approximated. Finally, Figure 3 illustrates the results obtained by the proposed framework and the reference algorithm for a face recognition application. The original space is Z^{500} and is mapped to a Z^{40} space. In all the cases, the proposed framework outperforms the reference algorithm achieving designs that require less area and at the same time have the same error in the data approximation.

VI. Conclusions

This paper proposes a novel Bayesian Factor Analysis framework for dimensionality reduction implementation in FPGAs. The proposed approach couples the problem of data approximation using a small set of variables, and the problem of area design optimization under a unified framework. It has been demonstrated that by injecting information to the system regarding the area requirements for the implementation of the constant coefficient multipliers using a prior distribution, we are able to target designs that have a significant reduction in the area requirement when they are compared against current techniques, achieving at the same time the same error in the approximation of the data. Future work will involve the extension of the framework to exploit the heterogeneous components of modern FPGAs for the problem of dimensionality reduction.

Acknowledgment

The authors wish to acknowledge the financial support of the EPSRC under the platform grant EP/C549481/1, the UK Research Council (Basic Technology Research Programme "Reverse Engineering Human Visual Processes" GR/R87642/02), and the BBSRC.

References

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 72–86, 1991.

[2] J. Taur and C. Tao, "Medical image compression using principal component analysis," in *International Conference on Image Processing*, vol. 2, 1996, pp. 903–906.

[3] I. Koren, *Computer Arithmetic Algorithms*, 2nd ed. New Jersey: Prentice-Hall Inc., 2002.

[4] A. Dempster and M. D. Macleod, "Use of minimum-adder multiplier blocks in FIR digital filters," *IEEE Trans. Circuits Systems II*, vol. 42, pp. 569 – 577, September 1995.

[5] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 1933.

[6] C.-S. Bouganis, G. A. Constantinides, and P. Y. K. Cheung, "A Novel 2D Filter Design Methodology For Heterogeneous Devices," in *Proc. Field-Programmable Custom Computing Machines*, 2005, pp. 1–10.

[7] C.-S. Bouganis, P. Y. K. Cheung, and G. A. Constantinides, "Heterogeneity Exploration for Multiple 2D Filter Designs," in *Proc. Field Programmable Logic and Applications*, 2005, pp. 263–268.