On Noisy Source Vector Quantization via a Subspace Constrained Mean Shift Algorithm

Youness Aliyari Ghassabeh, Tamás Linder, Glen Takahara

Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada
Email: {aliyari,linder,takahara}@mast.queensu.ca

***Abstract*—The subspace constrained mean shift (SCMS) algorithm is an iterative method for finding an underlying manifold associated with an intrinsically low dimensional data set embedded in a high dimensional space. We investigate the application of the SCMS algorithm to the problem of noisy source vector quantization where the clean source needs to be estimated from its noisy observation before quantizing with an optimal vector quantizer. We demonstrate that an SCMS-based preprocessing step can be effective for sources that have intrinsically low dimensionality in situations where clean source samples are unavailable and the system design relies only on noisy source samples for training.**

***Index Terms*—Noisy sources, vector quantization, subspace constrained mean shift algorithm, principal curves and surfaces.**

I. Introduction

Vector quantization is an important building block used in lossy data compression. A vector quantizer encodes (maps) vectors from a multidimensional input space into a finite subset of the space, called the codebook. The design of quantizers has been extensively studied. A classical result shows that an optimal quantizer of a given codebook size has to satisfy the Lloyd-Max conditions [1], [2]. This gives rise to the Lloyd-Max algorithm, an iterative method for scalar quantizer design that alternates between optimizing the the codebook and the partition induced by the codebook. The generalized version of the Lloyd-Max algorithm, known as the LBG algorithm, is used to design (locally) optimal vector quantizers [3] [4]. The classical problem of optimal vector quantization assumes that the source is available noise free to the quantizer. However, in some situations the source output may be corrupted by noise due to, e.g., measurement errors [5]. In this case, only a noisy version of the source is available for the quantization, and the quantizer's goal is then to minimize the expected distortion between the clean (unobserved) source and the output of the quantizer. Some practical examples where this model may apply are pilot's speech in the presence of aircraft noise, digital signal processing at transmitter or receiver that introduce quantization and round-off errors, satellite images affected by measurement error, or speech signal for a mobile phone in a noisy environment.

The theory of noisy source coding was first investigated by Dobrushin and Tsybakov [6] who analyzed the optimal rate-distortion performance. The structure of the optimal noisy source quantizer under the mean square distortion was studied by Fine [7], Sakrison [8], and Wolf and Ziv [9]. It has been shown that for the mean square distortion an optimal noisy source quantization system can be decomposed into an optimum estimator followed by an optimum source coder operating on the estimator output [9]. Some properties of an optimum noisy source quantizer, and its relations with the optimal estimator for the general problem, are derived by Ayanoglu [10]. By appropriately modifying the given distortion measure, Ephraim and Gray [11] showed the noisy source quantization problem becomes a standard quantization problem for the noisy source using the modified distortion measure. The problem of empirical vector quantizer design for noisy sources has been investigated by Linder, Lugosi, and Zeger [12].

The classical results imply that in order to minimize the mean square distortion with respect to the clean data, one needs to quantize the conditional expectation of the clean data given the noisy data. Thus, we need to find a good approximation, in the minimum mean square error (MMSE) sense, of the clean data from the observed noisy data. In practical situations where the statistics of the data and noise are unknown, the clean data can be estimated by applying nonparametric techniques, such as kernel regression [13], based on training data. However, in practice training data from the clean source may not be available and the designer of the quantizer only has access to the noisy observations.

In this paper, we consider sources that, with high probability, take values in a lower dimensional manifold. In addition, we assume that the noisy source quantizer is to be designed on the basis of noisy observations only. To obtain an estimate of the clean data in this situation, we use a non-parametric, iterative method, recently introduced by Ozertem and Erdogmos [14] for estimating principal curves and surfaces, called the subspace constrained mean shift (SCMS) algorithm.

Section II gives a brief review of the noisy source vector quantization problem. In Section III, we review the SCMS algorithm and state a convergence result. Section IV is devoted to simulation results which demonstrate the effectiveness of the SCMS approach to noisy source quantization for some special source distributions.

II. Noisy source vector quantization

A fixed rate N-point vector quantizer $Q : \mathbb{R}^k \to \mathcal{C}$ is a mapping from the k-dimensional Euclidean space $\mathbb{R}^k$ into a

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

finite set $\mathcal{C} \subset \mathbb{R}^k$ of cardinality N, called the codebook [15]. The elements of $\mathcal{C}$ are called the codevectors. The performance of a fixed rate quantizer in approximating the input vector is measured using a nonnegative function $d : \mathbb{R}^k \times \mathbb{R}^k \to [0, \infty)$ called the distortion measure. For a k-dimensional random vector $\mathbf{X}$ the overall distortion D of a quantizer Q is the expected value of the reconstruction error $D = Ed(\mathbf{X}, Q(\mathbf{X}))$. The most common and tractable distortion measure is the mean square error (MSE) distortion, i.e., $D = E\|\mathbf{X} - Q(\mathbf{X})\|^2$. Let $\mathbf{X}$ and $\mathbf{Y}$ be k-dimensional random vectors with $\mathbf{X}$ representing the clean source and $\mathbf{Y}$ the noisy version of $\mathbf{X}$. The problem of noisy source vector quantization is to approximate the clean data $\mathbf{X}$ with the lowest distortion based on quantizing its noisy version $\mathbf{Y}$ at a given fixed rate. Formally, our encoder is a member of the set of all N-level quantizers $\mathcal{Q}_N$ on $\mathbb{R}^k$. Assuming that $E\|\mathbf{X}\|^2$ is finite, the noisy source quantization problem is to find $Q^* \in \mathcal{Q}_N$ with minimum distortion

$$E\|\mathbf{X} - Q^*(\mathbf{Y})\|^2 = \min_{Q \in \mathcal{Q}_N} E\|\mathbf{X} - Q(\mathbf{Y})\|^2. \qquad (1)$$

It has been shown that the structure of an optimal N-level quantizer Q^* can be obtained via a useful decomposition [8] [9]. The following summarizes these results.

Proposition 1 ([8], [9]). *Let $m(\mathbf{y}) = E[\mathbf{X}|\mathbf{Y} = \mathbf{y}]$. Then an optimal quantizer Q^* is given by $Q^*(\mathbf{y}) = \hat{Q}(m(\mathbf{y}))$, where $\hat{Q} \in \mathcal{Q}_N$ is an MSE optimum N-level quantizer for $m(\mathbf{Y})$, i.e., $\hat{Q} = \arg\min_{Q \in \mathcal{Q}_N} E\|m(\mathbf{Y}) - Q(m(\mathbf{Y}))\|^2$. Furthermore, $\min_{Q \in \mathcal{Q}_N} E\|\mathbf{X} - Q(\mathbf{Y})\|^2 = E\|\mathbf{X} - m(\mathbf{Y})\|^2 + \min_{Q \in \mathcal{Q}_N} E\|m(\mathbf{Y}) - Q(m(\mathbf{Y}))\|^2$.*

Thus, in order to minimize the distortion, one needs to find a good approximation of the clean data $\mathbf{X}$ based on the observed noisy data $\mathbf{Y}$. In practical situations where the statistics of the data and noise are unknown, the clean data can be estimated using nonparametric techniques, such as kernel regression, based on training data. If a set of training data $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1\dots,n}$ is available in advance, the conditional expectation $m(\mathbf{y}) = E[\mathbf{X}|\mathbf{Y} = \mathbf{y}]$ of $\mathbf{X}$ given $\mathbf{Y}$ can be estimated using the kernel regression method as

$$\hat{m}(\mathbf{y}) = \frac{\sum_{i=1}^n \mathbf{x}_i K_h(\mathbf{y} - \mathbf{y}_i)}{\sum_{i=1}^n K_h(\mathbf{y} - \mathbf{y}_i)}, \qquad (2)$$

where $K_h : \mathbb{R}^k \to [0, \infty)$ is an integrable kernel function with bandwidth h. In this paper we assume that the designer of the quantizer only has access to the noisy observations and training data from the clean source are not available.

III. Subspace constrained mean shift (SCMS) algorithm

Principal curves and surfaces are the nonlinear generalizations of principal components. Assuming that a high dimensional observed data set is located on a low dimensional manifold, principal curves and surfaces have been proposed to estimate the structure of that low dimensional manifold [16] [17] [18]. An interesting recent definition of a d-dimensional principal surface in $\mathbb{R}^D$ ($d < D$) is given by Ozertem and Erogmus [14]. According to this definition, a given point is in a d-dimensional principal manifold associated with a probability distribution having a probability density function (pdf) if and only if the gradient of the pdf is orthogonal to at least $D-d$ eigenvectors of the Hessian of the pdf at that point, and the eigenvalues corresponding to these $D-d$ orthogonal eigenvectors are negative. It was shown in [14] that points in a d-dimensional principal manifold are local maxima of the pdf in a local orthogonal $D-d$-dimensional subspace.

An iterative algorithm, called the SCMS algorithm, was proposed to find points that satisfy this definition. The SCMS algorithm can be considered as a generalization of the mean shift (MS) algorithm [19] [20] [21] to estimate higher order principal curves and surfaces ($d \geq 1$). The MS algorithm is a non-parametric, iterative technique for locating modes of a pdf obtained via a kernel density estimate (see, e.g., [22]) from a given data set. The collection of these modes can be viewed as a zero-dimensional principal manifold. The MS algorithm is initialized to one of the observed data points, then it iteratively relocates this point to a weighted average of the neighboring data points to find stationary points of the estimated pdf [21].

Similar to the MS algorithm, the SCMS algorithm starts from a data set sampled from the probability distribution. The algorithm first forms a kernel density estimate $\hat{f}$ based on the data, then in each iteration it evaluates the MS vector for every data point. Then each MS vector is projected to the local subspace spanned by the $D-d$ eigenvectors corresponding to the $D-d$ largest eigenvalues of the estimated local covariance matrix at that point. The estimated local covariance matrix at $\mathbf{x}$ is defined by [14]

$$\hat{\mathbf{\Sigma}}^{-1}(\mathbf{x}) = -\mathbf{H}(\mathbf{x})\hat{f}(\mathbf{x})^{-1} + \mathbf{g}(\mathbf{x})\mathbf{g}(\mathbf{x})^t\hat{f}(\mathbf{x})^{-2},$$

where $\hat{f}(\mathbf{x})$ is the pdf estimate at $\mathbf{x}$, and $\mathbf{H}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ are the Hessian and gradient of the pdf estimate at $\mathbf{x}$, respectively. When the underlying pdf is multivariate Gaussian, the authors in [14] showed that projection of the MS vector to the $D-d$ largest eigenvectors of the local covariance matrix leads to the principal components.

We will slightly modify this projection step and instead we project the MS vector to the $D-d$ eigenvectors corresponding to the $D-d$ smallest eigenvalues of the Hessian matrix of the estimated pdf. This change is motivated by the observation that a point $\mathbf{x}$ is located on a d-dimensional ridge of the pdf if the gradient of the pdf is orthogonal to the $D-d$ smallest eigenvectors of the Hessian matrix of the pdf at $\mathbf{x}$ and the corresponding eigenvalues are negative [23]. The projection and MS steps are iterated until the norm of the difference between two consecutive projections becomes less than a predefined threshold.

The following result shows that this is a valid stopping criterion for the algorithm (such a result is missing in [14]). The proof of the theorem is omitted due to space constraints.

Theorem 1. *Let $\{\mathbf{y}_j\}_{j=1,2,\dots}$ be the sequence generated by the SCMS algorithm. Assume that the kernel K used for the kernel density estimate has a differentiable, convex, and monotonically decreasing profile $k : [0, \infty) \to [0, \infty)$ ($K(\mathbf{x}) \propto k(\|\mathbf{x}\|^2)$). Let $\{\hat{f}_{h,k}(\mathbf{x})\}$ denote the estimated pdf at point $\mathbf{x}$ with profile k and bandwidth h. Then the sequence*

$\{\hat{f}_{h,k}(\mathbf{y}_j)\}_{j=1,2,\ldots}$ *is monotonically increasing and convergent and* $\lim_{j\to\infty} \|\mathbf{y}_{j+1} - \mathbf{y}_j\| = 0$.

In the next section we will apply the SCMS algorithm as an estimate for the conditional expectation of the clean data given the noisy data. A heuristic explanation as to why this should work is as follows: If the clean source has a distribution that is supported on a lower-dimensional (smooth) manifold and the noisy source is obtained by adding independent noise having low variance, one expects that samples from the noisy source will be concentrated around the manifold. The estimate of the clean source sample based on a noisy observation is then obtained by running the SCMS algorithm, initialized at the noisy observation, until convergence. The output, which will lie on or very near to the manifold, will serve as the estimate of the clean source sample. Note that the SCMS algorithm only uses the noisy data and has no need for observations from the clean source, as opposed to the (theoretically optimal) estimate obtained from the kernel regression method (2).

IV. Simulation results

Since very little is known theoretically about the performance of the SCMS algorithm, we will use numerical examples to assess how well the SCMS algorithm approximates the clean data for the purposes of quantization. We compare the performance of the obtained system with that of a system using the kernel regression method trained on a data set consisting of pairs of clean and noisy data samples. In particular, in two different scenarios we compare the mean square distortion that results from quantizing the output of the SCMS algorithm with the near-optimal distortion resulting from quantizing the estimated clean data using the kernel regression method. We note that the kernel regression method is asymptotically optimal in the limit of large training set sizes.

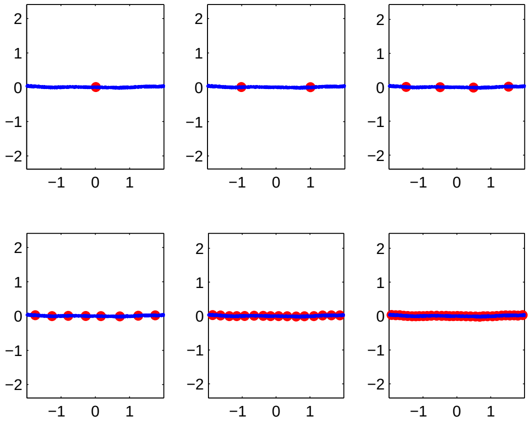

Fig. 1: *Quantization of a noisy line.* The blue points represent the output of the SCMS algorithm applied to the points from the noisy line and the red points are the codewords generated by the LBG vector quantization algorithm.

A. Quantization of a noisy line

We examine the performance of the SCMS algorithm as a preprocessing step for noisy vector quantization on a straight line in $\mathbb{R}^2$. In the design stage, we uniformly select 500 samples from the straight line of length 4 and perturb them by additive, independent zero-mean bivariate Gaussian noise with per component variance 0.4. These points are fed the SCMS algorithm and the resulting 500 output points are then used as a training set to design a vector quantizer using the LBG algorithm. For testing, we select another 500 samples from the straight line, perturb them by noise, the noisy data is fed to the SCMS algorithm, and the output of the algorithm is quantized using the designed vector quantizer. We vary the number of codewords and run the simulations for quantizers of codebook size 1, 2, 4, 8, 16, and 32. Fig. 1 shows the output points of the SCMS algorithm and the computed codewords for each choice of the codebook size. The blue points in Fig. 1 represent the output points of the SCMS algorithm and the red points represent the codewords computed by the LBG algorithm.

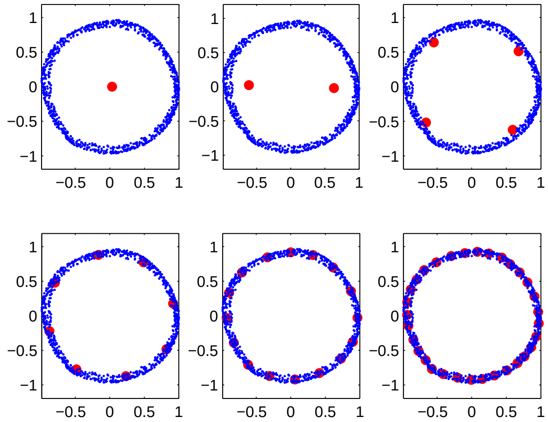

Fig. 2: *Quantization of a noisy circle.* The blue points represent the output of the SCMS algorithm applied to the points from the noisy circle and the red points are the codewords generated by the LBG vector quantization algorithm.

To compare the performance of the SCMS approach with the theoretical optimum, we generate 500 pairs of clean and noisy data point to train a kernel regression function in order to estimate the conditional expectation of the clean data given the noisy version. Another 500 noisy data points are then generated and fed to the kernel regression method and the output is used to train a vector quantizer with the LBG method. In the testing phase, another 500 noisy data points are generated, fed to the kernel regression estimator, and the output is quantized using the vector quantizer obtained in the training phase. Table I compares the mean square distortions resulting from the quantization of the estimated clean data using the kernel regression method and the output of the SCMS algorithm, respectively, as a function of the number of the codevectors (ranging from 2 to 128). Although the SCMS algorithm does not have access to the clean data, the simulation results indicate that the resulting mean square distortion is close to that achieved by the near-optimal scheme where the clean data estimates are obtained using the kernel regression method.

TABLE I: *Quantization of a noisy line.* The mean square distortion resulting from quantization of the output of the SCMS algorithm and the (near) optimal mean square distortion for different number of codebook sizes ranging from 2 to 128 for the noisy line.

Number of the codevectors	2	4	8	16	32	64	128
Optimal distortion [1]	0.4986	0.2507	0.1287	0.0639	0.0330	0.0172	0.0081
SCMS algorithm	0.5001	0.2683	0.1477	0.0731	0.0415	0.0271	0.0135

TABLE II: *Quantization of a noisy circle.* The mean square distortion resulting from quantization of the output of the SCMS algorithm and the (near) optimal mean square distortion for different number of codebook sizes ranging from 2 to 128 for the noisy circle.

Number of the codevectors	2	4	8	16	32	64	128
Optimal distortion [1]	0.7271	0.3858	0.2016	0.1064	0.0595	0.0367	0.0294
SCMS algorithm	0.7274	0.3945	0.2120	0.1220	0.071	0.0498	0.0379

B. Quantization of a noisy circle

The simulation setup is similar to the previous one, but now we consider the uniform distribution on the unit circle as the clean source and the additive bivariate zero-mean Gaussian noise has per sample variance 0.3. For training and testing two sets of 1024 noisy data points are generated for the SCMS approach, and 1024 pairs of clean and noisy data points are generated for the kernel regression approach. Fig. 2 shows the output of the SCMS algorithm and the computed codewords for simulations with quantizer codebook sizes 1, 2, 4, 8, 16, and 32. Table II compares the mean square distortions for quantization of the SCMS estimates and that of the kernel regression method, respectively, as the number of the codevectors ranges from 2 to 128. The measurements indicate that the mean square distortion achieved by quantization of the output of the SCMS algorithm is close to the near-optimum mean square distortion obtained by quantization of the estimates using the kernel regression method.

V. Conclusion

The simulation results demonstrate the effectiveness of the SCMS approach to noisy source quantization for the special source distributions we have investigated. In general, one can expect similarly good results if the source distribution is supported on a lower dimensional manifold that the SCMS algorithm can effectively reconstruct from noisy data. A theoretical analysis and performance guarantees for the SCMS approach to noisy source quantization are the subject of future research.

References

[1] S. P. Lloyd, "Least squared quantization in PCM," unpublished memorandum, Bell labs., 1957; reprinted in *IEEE Trans. Inform. Theory*, vol. 28, no. 2, pp. 129-137, Mar. 1982.

[2] J. Max, "Quantizing for minimum distortion," *IEEE Trans. Inform. Theory*, vol. 6, pp. 7-12, Mar. 1960.

[3] Y. Linde, A. Buzo, R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communications*, vol. 28, pp. 702-710, Jan. 1980.

[4] R. M. Gray, J. C. Kieffer, Y. Linde, "Locally optimum block quantizer design," *Inform. and Contr.*, vol. 45, pp. 178-198, May 1980.

[5] D. Rebollo-Monedero, S.Rane, B. Girod, "Wyner-Ziv quantization and transform coding of noisy sources at high rates," *Eighth Asilomar Conf. on Signals, Systems, and Computers*, pp 2084 -2088, Mar. 2005.

[6] R. L. Dorbushin, B. S. Tsybakov, "Information transmission with additional noise," *IEEE Trans. Inform. Theory*, vol. 18, pp. 293-304, Sep. 1962.

[7] T. Fine, "Optimum mean-square quantization of a noisy input," *IEEE Trans. Inform. Theory*, vol. 11, pp. 293-294, Apr. 1965.

[8] D. J. Sakrison, "Source encoding quantization of a noisy input," *IEEE Trans. Inform. Theory*, vol. 14, pp. 165-167, Jan. 1968.

[9] J. K. Wolf, J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. on Inform. Theory*, vol. 16, pp. 406-411, Jul. 1970.

[10] E. Ayanoglu,"On optimal quantization of noisy sources," *IEEE Trans. on Inform. Theory*, vol. 36, no. 6, pp. 1450-1452, Nov. 1990.

[11] Y. Ephraim, R. M. Gray, "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization," *IEEE Trans. on Inform. Theory*, vol. 34, no. 4, pp. 826-834, Jul. 1988.

[12] T. Linder, G. Lugosi, K. Zeger, "Empirical quantizer design in the presence of source noise or channel noise," *IEEE Trans. on Inform. Theory*, vol. 43, no. 2, pp. 612-623, Mar. 1997.

[13] E. A. Nadaraya,"On estimating regression,"*Theory of probability and Its Applications*, vol. 9, no. 1, pp. 141-142, 1964.

[14] U. Ozertem, D. Erdogmus, "Locally defined principal curves and surfaces," *Journal of Machine Learning Research*, vol. 12, pp. 1249-1286, Apr. 2011.

[15] A. Gersho, R. M. Gray, *Vector Quantization and Signal Compression*. Springer, 1991.

[16] T. Hastie, W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502-516, 1989.

[17] B. Kegl, A. Krzyzak, T. Linder, K. Zeger, "Learning and design of principal curves," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 281-197, Mar. 2000.

[18] K. Chang and J. Ghosh, "A unified model for probabilistic principal surfaces," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 1, pp. 22-41, Aug. 2001.

[19] K. Fukunaga, L. D. Hostetler, "Estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. on Inform. Theory*, vol. 21, pp. 32-40, Jan. 1975.

[20] Y. Cheng, "Mean shift, mode seeking and clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 790-799, Aug. 1995.

[21] D. Comanicio, P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603-619, May 2002.

[22] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[23] D. Eberly, *Ridges in image and data analysis*. Kluwer, 1996.

[1] Strictly speaking, this distortion is only near the theoretical optimum since the kernel estimate converges to the desired conditional expectation only in the limit of large training set sizes. Also the LBG algorithm is not guaranteed to produce globally optimal quantizers.