Sparse Covariance Coding

Aaron Courville, Dumitru Erhan, Pascal Vincent and Yoshua Bengio
Dept. IRO, Université de Montréal
$\{$*courvila,erhandum,vincentp,bengioy*$\}$*@iro.umontreal.ca*

Recently, there has been considerable interest in employing unsupervised learning methods as feature extractors for supervised learning tasks such as classification. The literature shows that methods based on this approach have proved to be competitive with established state-of-the-art machine learning strategies. One important recent advance was the discovery by (Hinton, Osindero, & Teh, 2006) of the role that unsupervised learning methods can play as an *initialization* for subsequent training for a supervised task. Unsupervised learning methods that extract sparse features of the data have received particular attention and has been shown by Raina, Battle, Lee, Packer, and Ng (2007) to significantly improve classification performance.

While there are some clear advantages of not including label information in learning the feature set or initialization — for instance, Raina et al. (2007) use a sparse coding scheme trained on a large quantity of "related" unlabeled data to augment the feature set training — the use of unsupervised feature extraction gives no safeguard against highly discriminative features being cast aside. In relatively uncomplicated tasks, such as classifying hand-written digits using the MNIST dataset (Lecun, Bottou, Bengio, & Haffner, 1998), the lack of supervisory information in determining the representation is not likely to present a problem as most of the salient features of the image are useful in the classification task. However in more complex tasks where a large number of salient features of the data can have nothing to do with the target task, unsupervised learning methods are not likely to efficiently generate discriminative features. On the other hand, the lesson of the utility of extracting features of the unsupervised input data pattern, demonstrated in Hinton et al. (2006) and in Raina et al. (2007) should not be ignored.

In this work, we focus on the problem of learning a sparse representation of data that stakes out a compromise position: explicitly taking into account information such as the labels in a classification task, while simultaneously attempting to capture descriptive features of the input data. Our method is based on a novel probabilistic interpretation of the canonical ridge analysis (Vinod, 1976), a regularized version of canonical correlation analysis.

Linear Latent Variable Models

Canonical correlation analysis (CCA) is a linear data projection method like principle components analysis (PCA), but where one can explicitly take label or other side information into account. CCA takes two or more datasets with corresponding entries (such as an input pattern and label) and determines projections that maximize the *correlation* between the datasets. Consider the case of two corresponding variables $X_1 \in \mathbb{R}^{m_1 \times n}$ and $x_2 \in \mathbb{R}^{m_2 \times n}$, each with mean zero, and projections $C_1^{\mathrm{CCA}} X_1$ and $C_2^{\mathrm{CCA}} X_2$ chosen to maximize $\mathrm{corr}^2(C_1 X_1, C_2 X_2)$. The directions chosen by CCA can be shown to correspond to the solution of the eigenvalue problem:

$$C_1 (X_1 X_1^T)^{-1} X_1 X_2^T (X_2 X_2^T)^{-1} X_2 X_1^T = \lambda C_1. \tag{1}$$

While CCA offers a means of incorporating target information into the learned representation of the data by constructing a projection of the input data that is maximally correlated with the target, it does so at the expense of other information about the input variable. This could well be detrimental to the goal of classification. After all, simple correlation is unlikely to capture all relevant information regarding the association between the input pattern and the corresponding label.

A compromise between the relative extremes of PCA (completely unsupervised) and the correlation seeking CCA, is the method of partial least squares (PLS). PLS describes a family of closely related algorithms that differ in details concerning the intended application, be it classification, regression or simply data summarization. However all PLS algorithms share the objective of finding projections for X_1 and X_2 that maximize *covariance*: i.e. find C_1^{PLS} and C_2^{PLS} to maximize $\mathrm{cov}(C_1 X_1, C_2 X_2) = \mathrm{var}(C_1 X_1) \mathrm{corr}^2(C_1 X_1, C_2 X_2) \mathrm{var}(C_2 X_2)$. Like CCA, PLS encodes correlation information, but PLS also weights the variance of X_1 and X_2 in the choice of the principle subspace.

In practice, the PLS directions are often dominated by the variance components of the covariance and one may wish to control the weighting of the variance contribution to the principle subspace. Canonical ridge analysis (CRA) was developed as a means of exploring a continuum of projections between PLS and CCA (Vinod, 1976; Rosipal & Krämer, 2006). CRA can be expressed directly as the solution, C_1^{CRA}, to the eigenvalue problem:

$$C_1(X_1X_1^T + k_1I_{m_1})^{-1}X_1X_2^T(X_2X_2^T + k_2I_{m_2})^{-1}X_2X_1^T = \lambda C_1 \tag{2}$$

where the parameters $k_1 \geq 0$ and $k_2 \geq 0$ control the weighting of the variance components of X_1 and X_2 respectively. With $k_1 = k_2 = 0$, CCA is recovered. As $k_1 = k_2 \to \infty$ the principle directions determined by CRA tend to those recovered by PLS. A comparable eigenvalue problem yeilds a solution for C_2^{CRA}.

A Probabilistic Interpretation of Canonical Ridge Analysis

Building on the earlier work of Tipping and Bishop (1999) and their development of a probabilistic model of PCA, Bach and Jordan (2005) proposed a probabilistic interpretation of canonical correlation analysis. We further build on Bach and Jordan's (2005) model to construct a probabilistic model of canonical ridge analysis.

Taking the data to have mean zero as before, Bach and Jordan (2005) showed that the maximum likelihood estimates of the parameters W_1, W_2, Ψ_1, Ψ_2 for the model over the data, $x_1 \in \mathbb{R}^{m_1}$ and $x_2 \in \mathbb{R}^{m_2}$:

$$z \sim \mathcal{N}(0, I_d), \qquad x_1 \sim \mathcal{N}(W_1z, \Psi_1), \qquad x_2 \sim \mathcal{N}(W_2z, \Psi_2), \tag{3}$$

result in the model recovering the subspace spanned by the first d canonical directions. Let $W = [W_1, W_2]^T$ and Ψ be a block diagonal matrix defined as $\Psi = \mathrm{Diag}(\Psi_1, \Psi_2)$, the *marginal* joint covariance of $[x_1, x_2]$ is $\Sigma = WW^T + \Psi$. We now introduce an inverse Wishart prior over the marginal joint covariance:

$$\Sigma \sim \mathcal{W}^{-1}(b, \Phi) = \frac{|\Phi|^{b/2}|\Sigma|^{(b+p+1)/2}\exp\{-\mathrm{trace}(\Phi\Sigma^{-1})/2\}}{2^{bp/2}\Gamma_p(b/2)} \tag{4}$$

with parameters $b \in \mathbb{R}$, $p = m_1 + m_2$, and the matrix Φ defined as: $\Phi = \begin{bmatrix} k_1I_{m_1} & 0 \\ 0 & k_2I_{m_2} \end{bmatrix}$.

We show that by taking $b = n - p - 1$ (where n is the number of data points), the maximum *a posteriori* (MAP) estimates of the model parameters result in subspace projections of x_1 and x_2, implicit in the posterior expectations $E[z \mid x_1]$ and $E[z \mid x_2]$, that correspond to the subspaces spanned by the canonical ridge analysis projections, C_1^{CRA} and C_2^{CRA} respectively as defined in eq. 2. Thus we now have a probabilistic interpretation of canonical ridge analysis that affords a latent variable model and explores the space of data representations between CCA and PLS.

Sparse Covariance Codes

The goal of this research project is to use our new latent variable interpretation of CRA to derive a novel sparse coding scheme that captures covariance information between input patterns and the corresponding label or other side information. By replacing the Gaussian prior on the latent variable z in eq. **??** with a Laplace prior, we define a joint sparse coding model of x_1 and x_2 with the goal of extracting correlation and/or covariance information from (x_1, x_2) data pairs. The parameters are estimated in a MAP gradient-descent framework. At each learning iteration, the latent representation (or bases coefficients) z is optimized, for each paired input pattern and label, via conjugate gradient in a convex setting. Then the model parameters W and Ψ are updated in the direction of the gradient of the log posterior, incorporating the Wishart prior.

We present results exploring the properties of the sparse covariance coding scheme. We show the learned bases for various values of the prior hyperparameters (k_1 and k_2), and evaluate the representation as the input to a standard classifier. The sparse covariance coding scheme is compared to more established representation learning schemes (such as PCA, PLS, CCA and unsupervised sparse codes (Olshausen & Field, 1996)) and across a range of experiments chosen to highlight the need for supervisory information in the feature learning/selecting process.

References

Bach, F. R., & Jordan, M. I. (2005). A probabilistic interpretation of canonical correlation analysis. Tech. rep. 688, University of California, Berkeley.

Hinton, G. E., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527–1554.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Raina, R., Battle, A., Lee, H., Packer, B., & Ng, A. Y. (2007). Self-taught learning: transfer learning from unlabeled data. In *ICML*, pp. 759–766.

Rosipal, R., & Krämer, N. (2006). *Subspace, Latent Structure and Feature Selection*, Vol. 3940, chap. Overview and Recent Advances in Partial Least Squares, pp. 34–51. Springer Berlin / Heidelberg.

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal components analysis. *Journal of the Royal Statistical Society B*, *61*(3), 611–622.

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, *4*, 147–166.