VARIABLES LATENTES, PROCESSUS GAUSSIENS ET RÉDUCTION DE LA DIMENSIONALITÉ

A. Arias

arias@irit.fr

1. INTRODUCTION

En reconnaissance de formes il est souvent nécessaire d'expliquer l'origine d'un ensemble de vecteurs. Si l'on considère ces vecteurs comme générés par un modèle de distribution de probabilité, la procédure à suivre est de choisir une forme spécifique de modèle et de calculer ses paramètres.

Si en même temps nous voulons expliquer les données observées avec un nombre plus réduit de variables (soit pour mieux comprendre leur processus de génération, soit pour pouvoir les visualiser), il est nécessaire de définir la fonction de probabilité conjointe entre les variables «génératrices» et les variables observées.

Finalement, il est possible que la relation conjointe des variables soit non linéaire, alors le modèle proposé doit aussi prendre en compte cette contrainte.

2. RÉDUCTION DE LA DIMENSIONALITÉ

Plusieurs tâches statistiques et d'apprentissage automatique sont sujettes à la «malédiction de la dimensionalité». En espaces de haute dimension, la représentativité d'un ensemble d'apprentissage est toujours compromise car le nombre d'échantillons nécessaires pour couvrir un hypervolume est en relation exponentielle avec la dimensionalité des variables concernées. La performance de plusieurs algorithmes (en termes de précision et de rapidité) est normalement améliorée si l'on travaille avec des représentations fiables des données de dimension réduite.

Par la réduction de la dimensionalité, l'estimation de la densité de probabilité à partir des données observées est simplifiée..

3. MODELE DE VARIABLES LATENTES ET PCA PROBABILISTE

3.1. Distribution normale

Nous allons considérer la modélisation de la distribution de N vecteurs $Y = [y_1 \dots y_N]^T$ de dimension D.

Le modèle le plus utilisé pour l'estimation d'une densité de probabilité est la distribution normale ou gaussienne, exprimé par :

$$p(y_n|\mu, \Sigma) = (2\pi)^{-D/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(y_n - \mu)^T\Sigma^{-1}(y_n - \mu)\right\} \tag{1}$$

où les paramètres sont la moyenne μ, la matrice de covariances Σ et le déterminant $|\Sigma|$. Une manière d'obtenir les valeurs de ces paramètres est avec la fonction de vraisemblance $\mathcal{L}$, qui considère le log de la probabilité des données observées selon le modèle construit avec ces paramètres.

$$\mathcal{L}(\mu, \Sigma) = \ln p(Y|\mu, \Sigma) = \sum_{n=1}^{N} \ln p(y_n|\mu, \Sigma) \tag{2}$$

On suppose que les vecteurs sont i.i.d. Si l'on voit $\mathcal{L}$ comme une fonction de μ et de Σ, la maximisation de cette fonction de vraisemblance est une procédure analytique qui obtient des valeurs pour ces paramètres les plus adaptées aux données observées (pour réviser la définition de ces paramètres dans le cadre d'inférence bayésienne voir [Bis95]).

Σ est constituée de $D(D+1)/2$ termes indépendants et μ de D termes, ce qui fait en total $D(D+3)/2$. Une manière de réduire cette $\approx D^2$ quantité de termes est de considérer la matrice Σ comme diagonale, mais c'est une affirmation trop optimiste car elle implique que les coordonnées des vecteurs sont indépendantes et un modèle comme celui là n'arrive pas à exprimer les corrélations entre les différents composants. En fait, les surfaces équiprobables générées par eq(1) sont des hyper-ellipsoïdes. Les axes principaux des ellipses sont les vecteurs propres de Σ dont les valeurs propres indiquent la variance sur les axes principaux.

Si l'on simplifie Σ comme étant diagonale, le nombre de termes est alors égal à $2D$ et les directions principales sont alignées avec les axes de coordonnées. En plus, les composantes des vecteurs sont considérées comme indépendantes et la distribution multivariable peut devenir le produit de D distributions monovariables.

3.2. Modèles de variables latentes

Une autre solution pour réduire le nombre de termes du modèle normal permet encore de faire ressortir certaines corrélations c'est le modèle des variables latentes

(LVM en anglais). Ce modèle est utilisé pour réduire la dimensionalité d'un ensemble en même temps que calcule l'estimation de sa densité de probabilité.

Le but du LVM est d'exprimer la distribution $p(Y)$ des variables centrées $y_n \in \mathcal{R}^D$ en fonction de variables $x_n \in \mathcal{R}^q$, où $q < D$. On suppose donc une distribution conjointe $p(Y, X)$ qui est en fait le produit d'une distribution marginale $p(X)$ des variables latentes et de la distribution conditionnelle $p(Y|X)$. On suppose que la distribution conditionnelle est factorisable par rapport aux variables observées.

$$p(Y, X) = p(X)p(Y|X) = \prod_{n=1}^{N} p(x_n)p(y_n|x_n) \tag{3}$$

La distribution conditionnelle $p(y_n|x_n)$ représente la projection des variables latentes vers les variables observées :

$$y_n = t(W; x_n) + \eta_n \tag{4}$$

$t(W; X)$ est une fonction des variables latentes X avec paramètres $W \in \mathcal{R}^{D\times q}$ et η est un processus de bruit (si les composants de η ne sont pas corrélés, $p(y_n|x_n)$ se factorise toujours comme dans eq(3)). La fonction $t(W; X)$ définit une variété dans l'espace observé $\mathcal{S}$ à partir de l'espace latent.

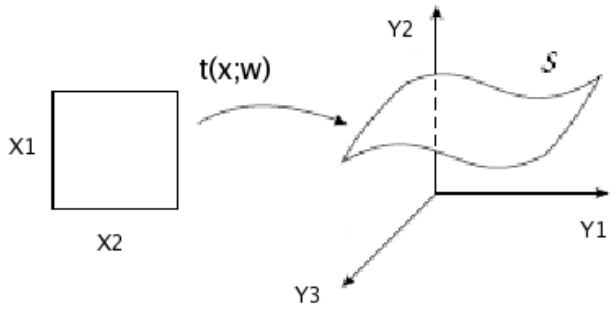

Fig. 1. La fonction $t(W; X)$ définit une variété à partir de la projection $x \rightarrow y$.

Pour définir un LVM on a donc besoin de spécifier : une distribution $p(\eta)$, une projection $t(W; X)$ et une distribution marginale $p(X)$.

3.3. PCA Probabiliste

Le LVM le plus simple est celui utilisé dans l'analyse factorielle. Il définit une projection linéaire $t(W; X)$ du type :

$$y_n = Wx_n + \eta_n \tag{5}$$

Si l'on considère cette projection plus un modèle de bruit dont la covariance est isotropique, $\eta_n \approx \mathcal{N}(0, \beta^{-1}I)$, la distribution de probabilité conditionnelle de y_n étant donné un ensemble x_n est :

$$p(y_n|x_n) = (2\pi\beta^{-1})^{-D/2} \exp\left\{-\frac{1}{2\beta^{-1}} \parallel y_n - Wx_n \parallel^2\right\} \tag{6}$$

Si les variables latentes $X = [x_1 \dots x_n]^T$ sont définies par une distribution gaussienne isotropique :

$$p(x_n) = (2\pi)^{-q/2} \exp\left\{-\frac{1}{2}x_n^T x_n\right\} \tag{7}$$

Le modèle cherché pour $p(y_n)$ est alors obtenu par la marginalisation des variables latentes :

$$\begin{aligned}
p(y_n) &= \int p(y_n|x_n)p(x_n)dx_n \\
&\approx \int \exp\left(-\frac{1}{2}(\beta y_n^T y_n - 2\beta y_n^T W x_n + x_n^T(\beta W^T W + I)x_n)\right) dx_n \\
&\approx \exp\left(-\frac{1}{2}(y_n^T(\beta I - \beta^2 W(\beta W^T W + I)^{-1}W^T)y_n)\right) \\
&= (2\pi)^{-D/2}|C|^{-1/2} \exp\left\{-\frac{1}{2}y_n^T C^{-1} y_n\right\} \\
p(Y) &= (2\pi)^{-ND/2}|C|^{-N/2} \exp\left\{-\frac{1}{2}tr(C^{-1}Y^T Y)\right\}
\end{aligned} \tag{8}$$

la covariance du modèle est : $C = \beta^{-1}I + WW^T$. Si l'on suppose l'indépendance entre les points, $p(Y) = \prod_{n=1}^{N} p(y_n)$. On obtient la probabilité $p(x_n|y_n)$ *a posteriori* avec le théorème de Bayes :

$$p(x_n|y_n) = (2\pi)^{-q/2}|\beta^{-1}M|^{-1/2} \exp\left\{-\frac{1}{2}(x_n - \tilde{x_n})^T(\beta^{-1}M)^{-1}(x_n - \tilde{x_n})\right\} \tag{9}$$

dans ce cas la covariance est : $\beta^{-1}M = \beta^{-1}(\beta^{-1}I + W^T W)^{-1}$, et la moyenne de la distribution est : $\tilde{x_n} = M^{-1}W^T y_n$. M a une dimensionalité $q \times q$ pendant que C est de dimension $D \times D$.

La vraisemblance des données observées sur ce modèle LVM est :

$$\mathcal{L} = \sum_{n=1}^{N} \ln\{p(y_n)\} = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|C| - \frac{N}{2}Tr\left\{C^{-1}S\right\} \tag{10}$$

S est la matrice de covariance $N^{-1}Y^TY$. La vraisemblance est maximisée si :

$$W = U_q(\Lambda_q - \beta I)^{1/2}R \tag{11}$$

où $U_q \in \mathcal{R}^{D\times q}$ sont les q principaux vecteurs propres de S, Λ_q est une matrice diagonale avec les valeurs propres correspondants et R est une matrice $(q \times q)$ orthogonale de rotation arbitraire (normalement $R = I$). C'est la solution du PCA probabiliste.

Pour visualiser les variables latentes, nous faisons appel à la moyenne de la distribution *a posteriori* $\tilde{X}$.

Le nombre des paramètres indépendants de ce modèle LVM est $(D + 1)(q + 1) - q(q + 1)/2$. Désormais le nombre de paramètres augmente linéairement par rapport a D.

4. PROCESSUS GAUSSIENS ET VARIABLES LATENTES

Nous avons considéré les variables latentes comme définies par une distribution gaussienne sphérique. Si l'on considère les paramètres w_d aussi comme des variables aléatoires avec une distribution similaire,

$$p(w_d) = (2\pi)^{-q/2}\exp\left\{-\frac{1}{2}w_d^Tw_d\right\}; \quad p(W) = \prod_{d=1}^{D} p(w_d) \tag{12}$$

Précédemment nous avons marginalisé $X \in R^{N\times q}$ car ce facteur était d'une dimensionalité plus grande que $W \in R^{D\times q}$, mais on pourrait décider de marginaliser W. Les deux approches sont équivalentes [Law05].

Si l'on marginalise W, avec la probabilité conditionnelle $p(y_n|x_n)$ de (6) et le modèle de bruit isotropique η_n,

$$p(Y) = \int \prod_{n=1}^{N} p(y_n|x_n)p(W)dW \tag{13}$$

et le modèle $p(Y)$ obtenu est :

$$p(Y) = \prod_{d=1}^{D}(2\pi)^{-N/2}|K|^{-1/2}\exp\left\{-\frac{1}{2}y_{:,d}^TK^{-1}y_{:,d}\right\} \tag{14}$$

$$p(Y) = (2\pi)^{-ND/2}|K|^{-D/2}\exp\left\{-\frac{1}{2}tr(K^{-1}YY^T)\right\} \tag{15}$$

$K = \beta^{-1}I + XX^T$. La distribution marginale P(Y) se factorise sur le nombre D de lignes en W, qui est en relation avec chaque colonne de Y ($Y \in \mathcal{R}^{N\times D}$). La vraisemblance est :

$$\mathcal{L} = -\frac{Nd}{2}\ln(2\pi) - \frac{d}{2}\ln|K| - \frac{1}{2}Tr\left(K^{-1}YY^T\right) \tag{16}$$

Les gradients de (16) par rapport à K et à X sont :

$$\frac{\partial \mathcal{L}}{\partial K} = K^{-1}YY^TK^{-1} - dK^{-1} \tag{17}$$

Dans le cas où $K = \beta^{-1}I + XX^T$, $\frac{\partial K}{\partial X} = X$, alors,

$$\frac{\partial \mathcal{L}}{\partial X} = K^{-1}YY^TK^{-1}X - dK^{-1}X \tag{18}$$

Les valeurs de X qui minimisent la vraisemblance sont [Law05] :

$$X = U_qLV^T \tag{19}$$

où U_q est une matrice $N \times q$ dont les colonnes sont les vecteurs propres de YY^T. L est une matrice $q \times q$ diagonale avec éléments $l_j = (\lambda_j - \beta^{-1})^{-1/2}$, λ_j est la valeur propre associée au l_j vecteur propre de YY^T/d et V est une matrice $q \times q$ de rotation arbitraire.

Ce modèle d'estimation de $p(Y)$ à partir de $p(X)$ est nommé GPLVM (gaussian process latent variable model).

Désormais on va considérer que K n'est pas une matrice linéaire de covariance.

5. LVM NON LINÉAIRES

Les LVM non linéaires sont capables de modéliser des relations non linéaires entre les variables latentes et les variables observées. Il y a trois algorithmes représentatifs de ces méthodes : les density networks [Mac95], le generative topographic mapping (GTM)[BSW98] et les GPLVM [Law05] mentionnés auparavant et que l'on va traiter avec plus de détails en leur extension non linéaire.

Les density networks sont une extension des réseaux de neurones bayésiens dédiés à l'estimation non supervisée d'une densité. La fonction $t(W; X)$ est modélisée avec un réseau de neurones pondérés avec les poids W. La probabilité de y_n est donc :

$$p(y_n) = \int p(y_n|Wx_n)p(x_n)dx_n \tag{20}$$

Et :

$$p(Y) = \prod_{n=1}^{N} p(y_n). \tag{21}$$

Pour une distribution sphérique de $p(X)$ et un modèle de bruit gaussien, l'intégrale (20) peut être évaluée par une approximation de Monte Carlo. La valeur de W est estimée par la minimisation du log négatif du (21). Après la définition de $p(Y)$ et $p(y_n|Wx_n)$, la distribution *a posteriori* de $p(X)$ se calcule avec la formule de Bayes. Des valeurs spécifiques de x_n sont obtenues avec la moyenne de $p(x_n)$.

Le modèle des density networks est très coûteux en temps de calcul et non déterministe dû à l'intégration de Monte Carlo.

Le GTM est un cas spécial de la procédure de density networks, avec une distribution *a priori* $p(X)$ discrète et uniforme et une fonction $t(W;X)$ approchée par une réseau de neurones RBF. Soit c_m un ensemble des coordonnées fixes en l'espace latente :

$$p(x_n) = \frac{1}{m}\sum_{m=1}^{M}\delta(x_n - c_m) \tag{22}$$

Cette distribution $p(x_n)$ plus un modèle de bruit isotropique donne une mélange de gaussiennes dans l'espace des observations y_n :

$$\begin{aligned} p(y_n) &= \frac{1}{m}\sum_{m=1}^{M} p(y_n|Wx_n = c_m) \\ p(y_n) &= \frac{1}{m2\pi\beta^{-1}}\sum_{m=1}^{M}\exp\left\{-\frac{1}{2\beta}\parallel W\phi(x_m) - y_n \parallel^2\right\} \end{aligned} \tag{23}$$

L'estimation des paramètres du modèle de mélanges (23) se fait avec l'algorithme d'expectation-maximization.

La moyenne *a posteriori* permettre de trouver x_n pour représenter y_n :

$$< x_n >= \sum_{m=1}^{M} p(x_m = c_m|y_n, W)c_m \tag{24}$$

Les density networks et le GTM donnent des modèles explicites $\mathcal{R}^D \rightarrow \mathcal{R}^q$. Le problème des density networks c'est la difficile et déficiente approximation de Monte Carlo et celui de GTM est la distribution uniforme en l'espace latente, qui donne des visualisations un peu bizarres.

6. REFERENCES

[Bis95] C. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[Bis99] C. Bishop. Latent variable models. *Learning in Graphical Models*, pages 371–403, 1999.

[BSW98] C. Bishop, M. Svenson, and C. Williams. Gtm : Generative topographic mapping. *Neural Computation*, 1(10) :215–234, 1998.

[Law05] N. Lawrence. Probabilistic non-linear principal components analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6 :1783–1816, 2005.

[Mac95] D. MacKay. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research*, 354(1) :73–80, 1995.

[TB99] M. E. Tipping and C. M. Bishop. Probabilistic principal components analysis. *Journal of the Royal Statistical Society*, 3(6) :611–622, 1999.