Analyse de shifts dans des données industrielles de capteurs par AutoEncodeur Variationnel parcimonieux

Brendan L'Ollivier*, Sonia Tabti*, Julien Budynek*

*Fieldbox, Quai Armand Lalande, 33300, Bordeaux, France
blollivier@fieldbox.ai, stabti@fieldbox.ai, jbudynek@fieldbox.ai

Résumé. Cet article explore l'utilisation d'AutoEncodeurs Variationnels (VAE) parcimonieux dans le cadre de l'analyse des perturbations affectant la distribution de données industrielles, aussi appelées shifts. À cette fin, plusieurs modèles sont comparés, en particulier, nous introduisons le LassoVAE, un VAE avec décodeur parcimonieux dont la procédure d'entraînement est efficace en termes de temps de calcul. La comparaison s'appuie sur un protocole expérimental que nous avons mis en place qui inclut un générateur de données synthétiques simulant à la fois l'interaction parcimonieuse entre les variables d'un processus industriel et des shifts subis par ces dernières. Deux nouvelles métriques sont introduites afin d'évaluer chaque modèle sur sa capacité à isoler la source des shifts. Les résultats des expériences montrent la supériorité des modèles parcimonieux pour cette tâche.

1 Introduction

Le domaine du monitoring statistique de processus industriels regroupe l'ensemble des modèles statistiques et d'apprentissage automatique visant à surveiller l'état de fonctionnement d'un processus industriel à partir des données de capteurs (Joe Qin, 2003). La prise en compte de la nature multivariée des données industrielles en est l'un des principaux enjeux, et constitue toujours un sujet de recherche ouvert. Dans ce contexte, les modèles non-supervisés à facteurs latents ont été largement utilisés pour dé-corréler les variables de processus industriels (Qin et al., 2020). Ces modèles ont pour but de représenter les données observables par des combinaisons d'un nombre réduit (relativement à la dimension de l'espace observable) de facteurs latents indépendants. Chaque variable latente encode l'information contenue dans un groupe de variables corrélées, potentiellement interprétable en termes de systèmes physiques.

En particulier, l'analyse en composantes principales (PCA), qui extrait séquentiellement les facteurs qui expliquent le plus de variabilité dans les données, a reçu beaucoup d'attention pour la détection d'anomalies dans les processus industriels (Teppola et al., 1998; Yin et al., 2012; Qin et Chiang, 2019). L'efficacité de cette approche pour l'isolation de la source du shift repose sur l'interprétabilité des facteurs latents. Dans le cas d'interactions linéaires, l'interprétabilité est généralement fournie par les poids de la matrice de projection (Cadima et Jolliffe, 1995). Cependant, comme le signale (Greco et Farcomeni, 2016), la non-parcimonie de la matrice de poids peut conduire à une interprétation erronée des interactions entre facteurs

latents et variables observables. Ce phénomène est connu sous le nom de "smearing-out-effect" ou "effet d'étalement" (Van den Kerkhof et al., 2013). Pour contourner ce problème d'interprétabilité, plusieurs approches ont été développées. Par exemple, la SparsePCA permet de reconstruire les données à partir de combinaisons parcimonieuses de composantes principales pour une interprétabilité explicite des facteurs latents (Luo et al., 2017; Theisen et al., 2021). La principale limite de la SparsePCA est qu'elle nécessite une estimation du nombre de facteurs latents à utiliser. Des critères de sélection tels que AIC ou BIC peuvent être appliqués, mais requièrent plusieurs entraînements pour l'optimisation de ce paramètre. Plus récemment, les AutoEncodeurs Variationnels (Kingma et Welling, 2014) ont aussi été utilisés dans des applications industrielles pour pallier les limites de la PCA en fournissant une architecture qui gère les non-linéarités et capable de désactiver automatiquement les facteurs latents superflus lors de l'entraînement (Zhu et al., 2022).

Un sujet étroitement lié à celui de la détection d'anomalies, est celui de la mesure et l'analyse des perturbations de la distribution des données d'un processus industriel donné. Ces perturbations peuvent être observées entre les données ayant servi à l'entraînement d'un modèle de machine learning et les données de production, ou entre les données d'équipements identiques mais opérant dans des conditions différentes. Détecter et interpréter les sources de ces perturbations, que l'on appellera dans la suite de l'article "shift", est un enjeu majeur pour la prévention d'impacts néfastes sur la performance des modèles de traitement de données mis en production. En effet, si les données de production changent de nature par rapport à l'entraînement, ou si un modèle est entraîné sur les données d'un équipement en particulier puis déployé sur une autre, disposer d'outils d'analyse de shifts pour prévenir une baisse de performances de ces modèles est crucial. Ce problème est référencé dans la littérature comme le "domain shift" (Lemberger et Panico, 2020).

Dans cet article, une méthode d'analyse du shift basée sur des VAE parcimonieux est proposée. La parcimonie imposée au décodeur, combinée à l'estimation du nombre de facteurs latents offre une analyse fine des sources de shifts entre datasets. Pour cela, nous introduisons le LassoVAE comme une option efficace en termes de temps d'entraînement pour induire de la parcimonie dans les poids du décodeur. Nous montrons que les décodeurs parcimonieux conduisent à une meilleure isolation des sources de shift dans l'espace latent ainsi qu'à une meilleure estimation des interactions entre facteurs latents et variables observables. Afin de quantifier ces deux aspects, deux métriques sont définies : le Mapping Recovery Score (MRS) et le Shift Dispersion Score (SDS). Le LassoVAE est comparé à d'autres modèles de type VAE ainsi qu'à des modèles de type PCA, avec des données générées par un protocole que nous avons mis en place puis avec les données du dataset Tennessee Eastman Process (Chen, 2019).

2 Mesurer le shift avec un AutoEncodeur Variationnel

Pour la suite de l'article, nous introduisons les conventions de notation suivantes. Soit $\boldsymbol{x} = (x_1, ..., x_D) \in \mathbb{R}^D$ un vecteur aléatoire de dimension D décrivant les données de capteurs d'un processus industriel, et $\boldsymbol{X} = (X_1, ..., X_D) \in \mathbb{R}^{N \times D}$ un dataset composé de N échantillons tirés de $p(\boldsymbol{x})$, la distribution de probabilités de $\boldsymbol{x}$.

2.1 Introduction aux VAE

L'AutoEncodeur Variationnel (VAE) (Kingma et Welling, 2014), est une classe de modèles génératifs très efficaces dans l'approximation de distributions de grande dimension. Il apprend à générer les D variables observables $\boldsymbol{x} \in \mathbb{R}^D$ à partir de $K < D$ facteurs latents gaussiens $\boldsymbol{z} \in \mathbb{R}^K$, avec une architecture de type auto-encodeur. D'un point de vue probabiliste, l'objectif est d'approcher la distribution a posteriori $p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{x})}$. La spécificité du VAE est de contourner l'intractabilité du calcul de $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ par de l'inférence variationnelle. En introduisant $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, une distribution paramétrée par un réseau de neurones, le problème se ramène à la minimisation de la divergence Kullback–Leibler (KL) : $D_{KL}(p(\boldsymbol{z}|\boldsymbol{x})\|q_\phi(\boldsymbol{z}|\boldsymbol{x}))$, dont découle la fonction de perte du VAE :

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] + \beta D_{KL}\big(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})\big) \tag{1}$$

Le paramètre β permet de moduler la régularisation de l'espace latent.

2.2 Décomposition de la source de shift

Soit $\boldsymbol{X}^S \in \mathbb{R}^{N\times D}$, un dataset dit *source*, composé de D variables réelles et N échantillons, représentant les données d'un processus industriel collectées dans des conditions "normales". L'objectif est d'analyser les différents shifts potentiels entre le dataset source et un autre dataset $\boldsymbol{X}^T \in \mathbb{R}^{N\times D}$, dit *target* (cible en français), collecté dans des conditions inconnues.

Après entraînement d'un VAE sur les données sources, l'encodeur g_ϕ^S et le décodeur f_θ^S sont utilisés pour produire les espaces latents : $\boldsymbol{Z}^S = g_\phi(\boldsymbol{X}^S)$ et $\boldsymbol{Z}^T = g_\phi(\boldsymbol{X}^T) \in \mathbb{R}^{N\times K}$; ainsi que les espaces résiduels : $\boldsymbol{E}^S = f_\theta(\boldsymbol{Z}^S) - \boldsymbol{X}^S$ et $\boldsymbol{E}^T = f_\theta(\boldsymbol{Z}^T) - \boldsymbol{X}^T \in \mathbb{R}^{N\times D}$.

Soit $\Delta : \boldsymbol{X}, \boldsymbol{Y} \mapsto \Delta(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{R}$, une distance entre deux datasets $\boldsymbol{X}$ and $\boldsymbol{Y}$, définie comme suit : $\Delta(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{D} \delta(X_i, Y_i)$, avec, $\forall 1 \leq i \leq D$, $\delta(X_i, Y_i)$ représentant la contribution de la i-ème variable à la distance totale. La fonction δ est une distance entre distributions univariées arbitraire. On définit le profil de shift entre $\boldsymbol{X}$ et $\boldsymbol{Y}$ par le vecteur des contributions :

$$\delta(\boldsymbol{X}, \boldsymbol{Y}) = \big(\delta(X_i, Y_i)\big)_{1\leq i\leq D} \tag{2}$$

Δ est utilisé dans la mesure du shift dans l'espace latent et dans l'espace résiduel, produisant deux mesures de shift supplémentaires en plus de celle dans l'espace des variables observables, comme introduit dans (Lemberger et Panico, 2020) :

— Covariate Shift : $\Delta(\boldsymbol{Z}^S, \boldsymbol{Z}^T) \gg 0 \wedge \Delta(\boldsymbol{E}^S, \boldsymbol{E}^T) \simeq 0$,

— Concept Shift : $\Delta(\boldsymbol{E}^S, \boldsymbol{E}^T) \gg 0$,

où $\wedge$ est l'opérateur logique "ET". Cette décomposition du shift, permet de discriminer deux comportements. En effet, le covariate shift correspond aux situations où des fluctuations du processus industriel n'affectent pas les interactions entre les variables. Le concept shift est, quant à lui, relatif à l'usure ou une erreur de capteur, qui au contraire affecte les interactions entre les variables et se manifeste par une erreur de reconstruction par le VAE plus importante des variables concernées.

2.3 Généralisation du décodeur : prérequis à la détection du covariate shift

Par définition, le covariate shift combine les conditions $\Delta(\boldsymbol{Z}^S, \boldsymbol{Z}^T) \gg 0$ et $\Delta(\boldsymbol{E}^S, \boldsymbol{E}^T) \simeq 0$, ce qui exige que le décodeur généralise à des régions non explorées de l'espace latent. Cette propriété n'est généralement pas garantie par un décodeur non-linéaire à couches profondes. En supposant que les interactions entre facteurs latents et variables observables sont linéaires, un décodeur affine permet de généraliser à toute nouvelle région de l'espace latent. Cette hypothèse est raisonnable dans un contexte industriel pour modéliser la plupart des interactions entre signaux de capteurs, tant que le point de fonctionnement du processus sous-jacent reste fixe, assurant la stationnarité de l'interaction entre variables. Dans le cas de points de fonctionnement multiples, chacun peut être considéré comme un nouveau domaine, et le problème est ramené à l'analyse du shift entre les différents points de fonctionnement.

3 Interprétabilité des facteurs latents

Mesurer le shift dans l'espace latent n'est utile qu'à condition de pouvoir l'expliquer avec les variables observables et lui donner une interprétation physique relative au processus industriel. Une façon d'y parvenir est de rendre explicite les interactions entre variables latentes $\boldsymbol{z} = (z_1, ..., z_K) \in \mathbb{R}^K$ et variables observables $\boldsymbol{x} = (x_1, ..., x_D) \in \mathbb{R}^D$. La notion d'interaction est ici à prendre au sens de : la variation d'un facteur latent est-t-elle liée à la variation d'une variable observable ?

3.1 Matrice d'interactions

Les interactions entre facteurs latents et variables observables sont modélisées par la matrice d'interactions $W \in \mathbb{R}^{K \times D}$, tel que $\forall\ 1 \leq k \leq K$ and $1 \leq d \leq D$:

$$w_{k,d} = \begin{cases} 1 \text{ si } z_k \text{ interagit avec } x_d \\ 0 \text{ sinon} \end{cases}$$

Dans un processus industriel composé de plusieurs sous-systèmes, représentés par différents groupes de capteurs, il est cohérent de supposer que la matrice d'interactions est parcimonieuse, limitant ainsi l'effet "smearing-out" évoqué en Section 1.

3.2 Apprendre la matrice d'interactions avec un décodeur parcimonieux

Dans cette section, deux méthodes d'estimation de la matrice d'interactions entre facteurs latents et variables observables sont présentées, dont une que nous proposons : le LassoVAE.

3.2.1 LassoVAE

Nous proposons une nouvelle architecture, le LassoVAE : un VAE avec une régularisation L1 sur les poids du décodeur affine, qui, combinée à la désactivation automatique des facteurs latents superflus, induite par la divergence KL, apprend des interactions parcimonieuses entre les facteurs latents et les variables observables tout en estimant la dimension de l'espace latent.

Sa fonction de perte en équation (3) est la somme de la fonction de perte du VAE (rappelée en équation (1)) et d'un terme de pénalité égal à la somme des valeurs absolues sur l'ensemble Θ des poids et des biais du décodeur linéaire :

$$\mathcal{L}_{\text{LassoVAE}} = \mathcal{L}_{\text{VAE}} + \alpha \sum_{\theta \in \Theta} |\theta| \tag{3}$$

Le paramètre α contrôle l'intensité de la parcimonie. La valeur optimale correspond à un compromis entre l'erreur de reconstruction et la régularisation L1. Avec le LassoVAE, l'estimation de la matrice d'interaction n'est pas explicite mais il a l'avantage d'avoir un entraînement économe en puissance de calcul, par opposition au SparseVAE présenté ci-dessous.

3.2.2 SparseVAE

Le LassoVAE est comparé à une seconde méthode d'induction de la parcimonie dans le décodeur : le SparseVAE (Moran et al., 2022). La matrice d'interactions W est estimée de manière explicite pendant la phase d'apprentissage en imposant un a priori dit de Spike and Slab Lasso (Ročková et George, 2016) sur les poids d'une matrice $W \in \mathbb{R}^{K \times D}$. Cette matrice W est utilisée comme masque de sélection des facteurs latents lors de la reconstruction de chaque variable. L'entraînement du SparseVAE est plus coûteux que celui du LassoVAE du fait de l'architecture du décodeur : il y a, en parallèle, autant de décodeurs que de variables observables.

4 Génération d'un dataset synthétique

Afin de comparer plusieurs méthodes de détection de shifts, nous définissons un protocole permettant de générer un jeu de données synthétiques en accord avec l'hypothèse d'interactions parcimonieuses dans des données industrielles mentionnée en section 3.1. Ce jeu de données est composé d'une matrice d'interactions parcimonieuses, de données source respectant ces interactions, et de données target obtenues par application de shifts dans l'espace latent des données sources. Ce procédé permet d'avoir le contrôle sur les interactions entre variables et sur les shifts, et ainsi d'évaluer les modèles selon le Mapping Recovery Score et le Shift Dispersion Score définis respectivement en sections 5.1 et 5.2.

4.1 Génération de la matrice d'interactions

Ce paragraphe décrit comment générer les coefficients $w_{k,d}$ d'une matrice d'interactions binaire parcimonieuse avec une distribution Beta-Bernouilli : $\forall k \in [\![1 : K]\!]$ and $d \in [\![1 : D]\!]$:

$$\eta_k \sim \text{Beta}(a, b), \quad w_{k,d} \sim \text{Bernoulli}(\eta_k) \tag{4}$$

où η_k contrôle la proportion de variables observables qui interagissent avec le $k^{\text{ième}}$ facteur. La distribution $w_{k,d}$ contrôle si la $d^{\text{ième}}$ variable x_d, interagit avec le $k^{\text{ième}}$ facteur z_k. Les paramètres de la distribution Beta sont $a, b > 0$. Quand a vaut 1, $b \in [1, N = dim(\boldsymbol{x})]$ contrôle l'intensité de la parcimonie. Un exemple de matrice W est donné en annexe sur la figure 6.

Une fois que la matrice W est générée, le modèle génératif global s'exprime par :

$$z_k \sim \mathcal{N}(0,1), k = 1, ..., K \quad \text{et} \quad x_d \sim \mathcal{N}\big(f(w_d \odot z_k), \sigma_d^2\big), d = 1, ..., D \tag{5}$$

Avec σ_d^2 la variance du bruit appliqué à la d^{ieme} variable et $\odot$ la multiplication terme à terme. Les résultats d'échantillonnages sont alors notés $\boldsymbol{Z} = (Z_1, ..., Z_K)$ et $\boldsymbol{X} = (X_1, ..., X_D)$.

4.2 Génération des données sources

L'équation (5) permet de générer la matrice $W^{\text{true}} \in \mathbb{R}^{K^{\text{true}} \times D}$, et les datasets $\boldsymbol{X}^{\text{true}} \in \mathbb{R}^{N \times D}$ et $\boldsymbol{Z}^{\text{true}} \in \mathbb{R}^{N \times K^{\text{true}}}$. Sous l'hypothèse d'interactions linéaires, la fonction f du décodeur peut être simplifiée en un produit matriciel entre les variables latentes et une la matrice de poids $C^{\text{true}} \in \mathbb{R}^{K^{\text{true}} \times D}$ définie comme suit : $\boldsymbol{X}^{\text{true}} = \boldsymbol{Z}^{\text{true}} C^{\text{true}}$. La position des coefficients non nuls de C^{true} est indiquée par la présence de 1 dans la matrice d'interactions associée W^{true}. Pour rendre les données plus réalistes, un bruit gaussien de faible intensité (relativement aux variances des variables observables), d'écart type σ^{noise}, est ajouté à la vraie distribution :

$$\boldsymbol{X}^{\text{obs}} = \boldsymbol{X}^{\text{true}} + \mathcal{N}(\mathbf{0}, \sigma^{\text{noise}} I_D) \tag{6}$$

Où $\mathbf{0}$ est le vecteur nul.

4.3 Génération des données cible

Le dataset target, ou cible en français, est généré de la même manière que le dataset source à la différence près que des shifts unidirectionnels sont appliqués dans l'espace latent, avant la génération des variables observables. Pour une direction donnée k, le shift unidirectionnel appliqué à un batch de N_{samples} échantillons est défini de la façon suivante :

$$\boldsymbol{Z}^{\text{true}}_{\text{batch}} \sim \mathcal{N}(\mathbf{0}, \sigma^{\text{noise}}_{\text{batch}} I_K), \quad \boldsymbol{Z}^{\text{cov}}_{\text{batch}} = u_k^{\text{cov}}(\boldsymbol{Z}^{\text{true}}_{\text{batch}})$$
$$\text{avec } u_k^{\text{cov}} : \boldsymbol{z} = (z_1, ..., z_K) \mapsto (z_1, ..., z_k + \sigma_{z_k}, ..., z_K)$$

L'amplitude de chaque translation est fixée à l'écart-type σ_{z_k} de la variable z_k affectée par le shift. Faire varier le paramètre N_{samples} permet de tester la mesure du shift dans des contextes différents. On s'attend à observer une meilleure mesure du shift sur des batchs de plus grande taille. En pratique, en milieu industriel, la taille du batch a un impact sur le délai de collecte d'échantillons, rendant significative la mesure du shift.

5 Métriques d'évaluation de la mesure du shift

A notre connaissance, aucune métrique de la littérature ne permet de comparer des modèles à facteurs latents sur leurs capacités d'identification des interactions entre facteurs latents et variables observables et d'isolation de la direction du shift dans l'espace latent. Dans cette section, deux métriques sont proposées dans ce but.

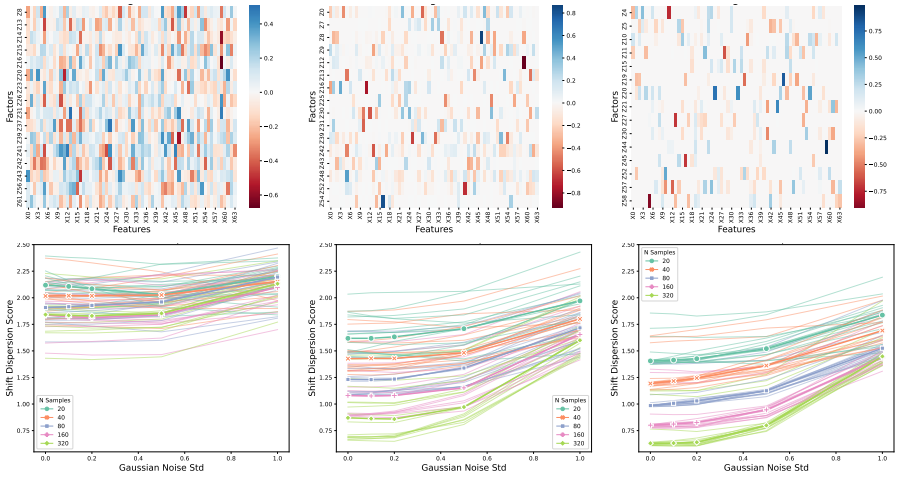


FIG. 1 – Résultats des expériences menées respectivement avec le LinearVAE (première colonne), le LassoVAE (deuxième colonne) et le SparseVAE (dernière colonne) sur le dataset synthétique. Pour chaque modèle : la première ligne montre la matrice de coefficients C calculée pendant l'entraînement. La deuxième ligne montre la relation entre le score de dispersion des profils de shift (SDS), la taille du batch et le niveau de bruit. Différentes courbes d'une même couleur correspondent aux différentes directions de shift. La courbe en gras montre la moyenne sur les 16 directions. De façon générale, les modèles parcimonieux permettent une meilleure isolation de la source du shift, avec un score de dispersion qui tend à augmenter lorsque le bruit augmente et lorsque la taille de l'échantillon diminue.

5.1 Mapping Recovery Score

En supposant que la matrice d'interactions théorique $W^{\text{true}} \in \mathbb{R}^{K^{\text{true}} \times D}$ soit connue, nous proposons une métrique quantifiant son écart à la matrice d'interactions prédite $W^{\text{pred}} \in \mathbb{R}^{K^{\text{pred}} \times D}$: le Mapping Recovery Score (MRS) défini comme suit :

$$MRS(W^{\text{true}}, W^{\text{pred}}) = \frac{1}{K^{\text{true}}} \sum_{k_t=1}^{K^{\text{true}}} \max_{1 \leq k_p \leq K^{\text{pred}}} J(w_{k_t}^{\text{true}}, w_{k_p}^{\text{pred}}) \tag{7}$$

Où $J(w_{k_t}^{\text{true}}, w_{k_p}^{\text{pred}})$ est l'indice de Jaccard entre la $k_t^{\text{ème}}$ ligne de W^{true} et la $k_p^{\text{ème}}$ ligne de W^{pred} :

$$J(w_{k_t}^{\text{true}}, w_{k_p}^{\text{pred}}) = \frac{w_{k_t}^{\text{true}} \cap w_{k_p}^{\text{pred}}}{w_{k_t}^{\text{true}} \cup w_{k_p}^{\text{pred}}}$$

Puisque l'indice de Jaccard varie entre 0 (aucune similarité) et 1 (récupération parfaite de la matrice d'interactions), le MRS varie également entre 0 et 1.

5.2 Score de Dispersion du Shift

Les modèles sont classés en fonction de leur capacité à isoler la direction du shift dans chaque batch du dataset target. Cette propriété est mesurée en comparant les profils de shift (définis dans la section 2.2) dans l'espace latent, après entraînement du modèle sur les données sources, aux profils de shift théoriques.
Le profil de shift estimé $\hat{\delta} = (\hat{\delta}_1, ..., \hat{\delta}_D)$, pouvant être vu comme une distribution de probabilités sur les directions possibles du shift, nous introduisons le Score de Dispersion de Shift ou Shift Dispersion Score (SDS), comme l'entropie croisée entre le profil de shift classé dans l'ordre décroissant des valeurs de shift : $(\max_{1\leq d\leq D} \hat{\delta}_d, ..., \min_{1\leq d\leq D} \hat{\delta}_d)$, et le profil de shift théorique associé, noté $\delta^{\text{true}} = (1, 0, ..., 0)$:

$$SDS\big(\hat{\delta}, \delta^{\text{true}}\big) = -\log\big(\max_{1\leq d\leq D} \hat{\delta}_d\big) \tag{8}$$

Notez que l'équation (8) a été simplifiée en raison de la nature binaire de δ^{true}. Des valeurs faibles du SDS suggèrent une bonne isolation de la source du shift.

6 Expériences et Résultats

Cette section présente les résultats d'expériences sur deux jeux de données différents : un jeu de données synthétiques constitué grâce au protocole de la section 4 et un jeu de données industrielles.

6.1 Expériences sur données synthétiques

6.1.1 Préparation des datasets source et target

La distribution décrite dans l'équation (4) est utilisée pour générer une matrice d'interactions parcimonieuses entre 16 facteurs latents et 64 variables observables. Un dataset source de 5000 échantillons, bruités avec un bruit gaussien d'écart type 0.2 est généré en suivant le processus décrit par l'équation (6). Le dataset target est composé de différents batchs de données shiftées. Chaque batch est obtenu par application de l'équation (7) pour une direction de shift k donnée, une taille de batch N_{samples} donnée et un niveau de bruit $\sigma^{\text{noise}}_{\text{batch}}$ donné. L'opération est répétée pour les $1 \leq k \leq K$ dimensions de l'espace latent, ainsi que pour $N_{\text{samples}} \in \{20, 40, 80, 160, 320\}$ et $\sigma^{\text{noise}}_{\text{batch}} \in \{0, 0.1, 0.2, 0.5, 1\}$. Au total, les données target sont constituées de 400 batchs de données shiftées, avec 20 à 320 échantillons par batch.

6.1.2 Modèles comparés

Plusieurs modèles sont comparés dans cette étude, notamment des modèles basés sur des VAE : le LassoVAE et le SparseVAE, décrits en section 3.2, ainsi que le LinearVAE (un VAE constitué d'un décodeur affine mais non régularisé). Le nombre de facteurs latents est initialement défini comme égal à la dimension de l'espace observable. La propriété d'auto-désactivation des facteurs latents du VAE lors de l'entraînement (Dai et al., 2019), permet une estimation de la dimension intrinsèque du dataset. La figure 2 illustre ce phénomène pour les modèles LassoVAE et SparseVAE.

(a) LassoVAE (b) SparseVAE

FIG. 2 – Évolution de l'écart type des 64 facteurs latents au cours de l'entraînement. Les facteurs latents dont l'écart type est proche de zéro (au sens d'un seuil prédéfini) à la fin de l'entraînement, sont considérés comme inactifs, et donc ignorés dans l'analyse du shift.

Les algorithmes Probabilistic PCA (PPCA) et SparsePCA de la bibliothèque *scikit-learn* sont également entraînés et testés comme modèles de référence. Notons que ces deux modèles n'éliminent pas automatiquement les facteurs latents superflus. Le choix du nombre de facteurs est donc déterminant. Les deux modèles PPCA et SparsePCA sont entraînés une première fois avec le nombre optimal de facteurs : 16, et une seconde fois avec un nombre non optimal arbitraire : 32. Ces modèles ont respectivement un suffixe "16" ou "32" ajouté à leur nom de base (par exemple, "SparsePCA16"). Les paramètres utilisés pour les différents modèles peuvent être consultés en annexe A.3.

6.1.3 Entraînement et mesure du covariate shift

Chacun des modèles est entraîné sur le dataset "source". Pour améliorer la robustesse statistique de nos résultats, l'entraînement est répété pour 5 états aléatoires initiaux distincts. Après cette étape, les poids de la matrice du décodeur sont utilisés pour calculer le Mapping Recovery Score défini par l'équation (7). Chaque modèle est ensuite utilisé pour projeter le dataset target dans l'espace latent. Pour chaque batch du dataset target, on calcule le profil de shift dans l'espace latent, conformément à l'équation (2). La distance statistique δ utilisée pour le calcul des shifts univariés est la distance de Wasserstein 1D. Les profils de shift sont alors agrégés en un SDS (cf. éq.8), servant à classer les modèles selon leurs capacités à isoler la source du shift. Un SDS bas signifie que le profil de shift est très concentré sur une direction latente, correspondant à la direction du shift théorique. Au contraire, un SDS élevé reflète une difficulté à discerner une direction de shift privilégié dans le profil de shift mesuré.

6.1.4 Résultats

La figure 1 illustre le lien entre la parcimonie du décodeur et le score d'isolation du shift. On y voit que même dans les pires conditions de mesure de shift (petite taille de batch et haut niveau de bruit), les modèles avec parcimonie induite obtiennent un meilleur SDS que le LinearVAE pris dans les conditions les plus propices (grande taille de batch et faible niveau de bruit). Les figures 4b et 4a montrent que les modèles non régularisés (LinearVAE et PPCA) ne permettent pas d'estimer correctement les vraies interactions parcimonieuses entre facteurs latents et variables observables (au sens du Mapping Recovery Score), confirmant ainsi le lien

entre mécanismes induisant la parcimonie, estimation des interactions théoriques et isolation de la source du shift dans l'espace latent. Il est aussi à noter que le SparseVAE et le SparsePCA ont des MRS et SDS comparables alors que le SparseVAE doit de lui-même estimer la dimension intrinsèque des données. La figure 3 montre l'évolution des métriques Mean Square Error (MSE) et la norme L1 des poids du décodeur (L1 Loss) au cours des phases d'apprentissage des différents modèles. Bien qu'étant le plus performant au sens de la norme L1, le SparseVAE est le plus lent à converger. Sans compter que chaque epoch est aussi plus coûteuse en temps de calcul que le LinearVAE et le LassoVAE car, par construction, le calcul de la sortie du décodeur nécessite un passage par feature (cf. annexe A.4). Le LassoVAE possède la même dynamique de convergence que le LinearVAE, tout en assurant une faible norme L1.

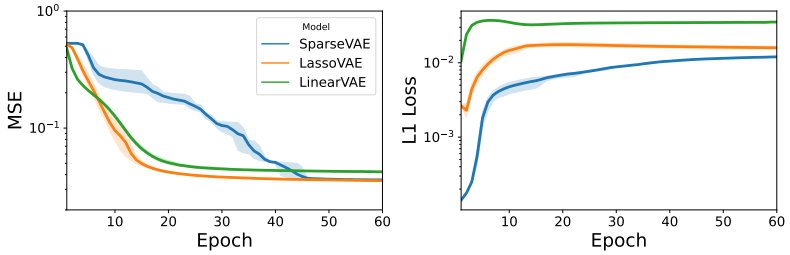
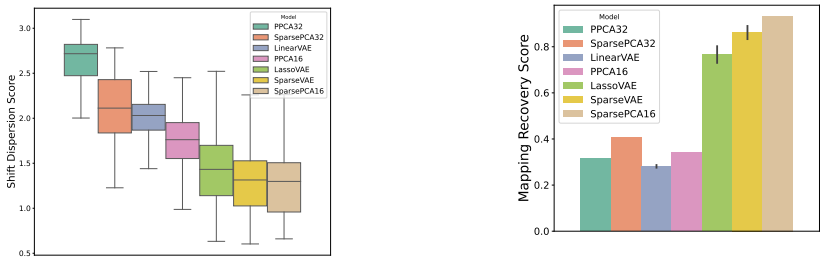


FIG. 3 – Historique d'apprentissage des trois modèles VAE. La principale observation est la différence de vitesse de convergence entre les deux modèles parcimonieux LassoVAE et SparseVAE. Le LassoVAE converge plus vite mais obtient une norme L1 finale plus élevée que celle du SparseVAE.



(a) Shift Dispersion Scores (SDS)

(b) Mapping Recovery Scores (MPS)

FIG. 4 – Figure 4a : Shift Dispersion Scores agrégés. Une valeur plus faible signifie une meilleure isolation de la véritable source de shift. Chaque boxplot montre la distribution du SDS pour différentes valeurs de N_{samples} et de $\sigma^{\text{noise}}_{\text{batch}}$ et pour différents . Figure 4b : Mapping Recovery Scores de tous les modèles testés dans cette étude. Nous constatons la supériorité des modèles parcimonieux LassoVAE, SparseVAE et SparcePCA16 pour cette tâche.

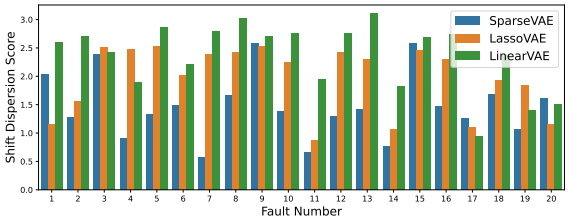


FIG. 5 – Shift Dispersion Scores (SDS) calculés à partir du profil de shift de chacune des vingt anomalies de fonctionnement du dataset Tennesse Eastman Process. Les valeurs de SDS les plus basses montrent la meilleur performance des modèles parcimonieux pour l'isolation de la source du shift dans l'espace latent.

6.2 Expériences sur dataset industriel : le Tennessee Eastman Process

Le jeu de données Tennessee Eastman Process (Chen, 2019) provient de la simulation numérique d'un processus industriel chimique. Ce jeu de données est régulièrement utilisé pour comparer les algorithmes de détection d'anomalies. Le jeu de données est divisé en deux parties, la première contient des runs de simulation "sans défaut", et la seconde contient 20 types différents de runs "défectueux". Nous avons comparé les performances de chacun des trois modèles basés sur des VAE pour isoler les sources de shift dans les données "défectueuses", après un entraînement sur les données "sans défaut". Pour chaque type de défaut, un profil de shift est obtenu en calculant les distances de Wasserstein sur tous les facteurs latents entre les données "sans défaut" et "avec défaut". Les profils de shift résultants sont comparés en fonction de leur Shift Dispersion Score. La figure 5 montre les résultats de cette expérience. Les modèles parcimonieux améliorent l'isolation de la source du shift dans l'espace latent pour la plupart des types de défaut.

7 Conclusion et perspectives

Cet article aborde le problème de l'identification de la source du shift mesuré dans l'espace latent de modèles du type AutoEncodeur Variationnel dans le contexte du monitoring de processus industriels. Nous montrons que la modélisation parcimonieuse de la corrélation entre les variables d'un processus industriel est une solution prometteuse. Nous avons introduit le LassoVAE, un VAE avec un décodeur linéaire parcimonieux capable d'apprendre des interactions parcimonieuses tout en estimant la dimension intrinsèque des données. Notre étude montre que la parcimonie améliore l'isolation des sources de shifts dans l'espace latent par rapport au VAE traditionnel et les méthodes de types PCA. Aussi, deux nouvelles mesures ont été définies : le Mapping Recovery Score et le Shift Dispersion Score, servant à mesurer respectivement la qualité de l'estimation de la matrice d'interaction et l'isolation de la source du shift. Une possible contribution future consisterait à généraliser le cadre d'analyse du shift présenté dans cet article aux interactions spatio-temporelles. De la même manière que ce qui a été fait avec la PCA dynamique, le LassoVAE peut être étendu à l'apprentissage des interactions temporelles parcimonieuses dans les données de capteurs.

Références

Cadima, J. et I. T. Jolliffe (1995). Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics 22*(2), 203–214. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/757584614.

Chen, X. (2019). Tennessee Eastman simulation dataset. Publisher : IEEE Type : dataset.

Dai, B., Y. Wang, J. Aston, G. Hua, et D. Wipf (2019). Hidden Talents of the Variational Autoencoder. *arXiv :1706.05148 [cs]*. arXiv : 1706.05148.

Greco, L. et A. Farcomeni (2016). A plug-in approach to sparse and robust principal component analysis. *TEST 25*(3), 449–481.

Joe Qin, S. (2003). Statistical process monitoring : basics and beyond. *Journal of Chemometrics : A Journal of the Chemometrics Society 17*(8-9), 480–502.

Kingma, D. P. et M. Welling (2014). Auto-Encoding Variational Bayes. *arXiv :1312.6114 [cs, stat]*. arXiv : 1312.6114.

Lemberger, P. et I. Panico (2020). A Primer on Domain Adaptation. *arXiv :2001.09994 [cs, stat]*. arXiv : 2001.09994.

Luo, L., S. Bao, J. Mao, et D. Tang (2017). Fault Detection and Diagnosis Based on Sparse PCA and Two-Level Contribution Plots. *Industrial & Engineering Chemistry Research 56*(1), 225–240. Publisher : American Chemical Society.

Moran, G. E., D. Sridhar, Y. Wang, et D. M. Blei (2022). Identifiable Deep Generative Models via Sparse Decoding. *arXiv :2110.10804 [cs, stat]*. arXiv : 2110.10804.

Qin, S. J. et L. H. Chiang (2019). Advances and opportunities in machine learning for process data analytics. *Computers & Chemical Engineering 126*, 465–473.

Qin, S. J., Y. Dong, Q. Zhu, J. Wang, et Q. Liu (2020). Bridging systems theory and data science : A unifying review of dynamic latent variable analytics and process monitoring. *Annual Reviews in Control 50*, 29–48.

Ročková, V. et E. George (2016). The Spike-and-Slab LASSO. *Journal of the American Statistical Association 113*(521), 431–444.

Teppola, P., S.-P. Mujunen, P. Minkkinen, T. Puijola, et P. Pursiheimo (1998). Principal component analysis, contribution plots and feature weights in the monitoring of sequential process data from a paper machine's wet end. *Chemometrics and Intelligent Laboratory Systems 44*(1), 307–317.

Theisen, M., G. Dörgő, J. Abonyi, et A. Palazoglu (2021). Sparse PCA Support Exploration of Process Structures for Decentralized Fault Detection. *Industrial & Engineering Chemistry Research*. Publisher : American Chemical Society.

Van den Kerkhof, P., J. Vanlaer, G. Gins, et J. F. M. Van Impe (2013). Analysis of smearing-out in contribution plot based fault isolation for Statistical Process Control. *Chemical Engineering Science 104*, 285–293.

Yin, S., S. X. Ding, A. Haghani, H. Hao, et P. Zhang (2012). A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of Process Control 22*(9), 1567–1581.

Zhu, J., M. Jiang, et Z. Liu (2022). Fault Detection and Diagnosis in Industrial Processes

with Variational Autoencoder : A Comprehensive Study. *Sensors 22*(1), 227. Number : 1 Publisher : Multidisciplinary Digital Publishing Institute.

A Annexe

A.1 Exemple de matrice d'interaction synthétique

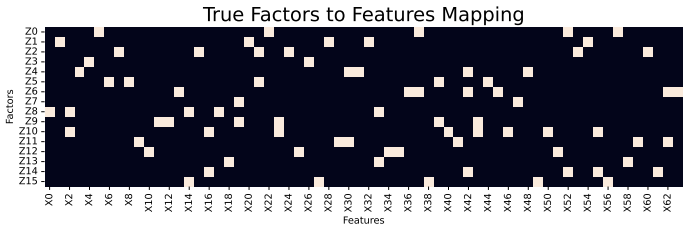


FIG. 6 – Matrice d'interaction synthétique générée avec l'équation (4). Les cellules en blanc représentent une valeur de $w_{k,d}$ égale à 1. Ce qui signifie que le k^{ieme} facteur latent interagie avec la d^{ieme} variable observable.

A.2 Architecture des modèles

A.2.1 Encodeurs

Les trois modèles VAE possèdent la même architecture pour l'encodeur : deux réseaux parallèles d'une seule couche cachée de 128 neurones qui calcule respectivement μ_z et $\sigma_z \in \mathbb{R}^K$:

```
dense(input_dim, 128) -> relu -> dense(128, latent_dim)
```

A.2.2 Décodeurs

Le décodeur calcule $\mu_{\boldsymbol{x}} \in \mathbb{R}^D$ à partir de $\boldsymbol{z}$. Le vecteur des écarts type $\sigma_{\boldsymbol{x}} \in \mathbb{R}^D$ est estimé de façon globale : les écarts type de chaque variable $\sigma_{x_1}, ..., \sigma_{x_D}$ sont des paramètres du modèle, au même titre que les poids des neurones, estimés lors de l'apprentissage.

Le décodeur est constitué d'une seule couche linéaire de dimensions `(latent_dim, input_dim)` régularisé par une norme L1 sur les poids dans le cas du LassoVAE, et sans régularisation dans le cas du LinearVAE.

Le décodeur du SparseVAE est composé de `input_dim` couche linéaire de dimensions `(latent_dim, 1)`, prenant en entrée une version masquée de $\boldsymbol{z}$, calculée à à partir de la matrice W.

A.3 Paramètres des modèles

Les paramètres listés ci-dessous pour chaque modèle ont été définis empiriquement de façon à optimiser les scores MRS et SDS.

1. **SparsePCA** : Le paramètre alpha, contrôlant l'intensité de la régularisation L1 est fixé à 1.
2. **LassoVAE** : L'intensité de la régularisation L1 sur les poids du décodeur est fixé à 0.07.
3. **SparseVAE** : Les paramètres a et b de la distribution beta sont fixés respectivement à 1 et 32. Les paramètres λ_0 et λ_1 sont eux fixés respectivement à 20 et 0.01.

Dans les trois cas, la valeur de β (intensité de la régularisation de l'espace latent par la divergence Kullback-Leibler) est fixé à 0.3.

A.4 Entraînement des modèles VAE

Les trois modèles ont été entraînés avec le même taux d'apprentissage : $1e^{-3}$ et la même taille de batch : 50. En plus des courbes d'historique d'apprentissage présentées en figure 3, voici les durées moyennes de calcul d'un epoch (en seconde) pour chaque modèle obtenues avec le processeur Intel® Core™ i7-8850H CPU @ 2.60GHz × 12 :

LinearVAE	LassoVAE	SparseVAE
0.25	0.26	2.22

Summary

This paper explores the use of sparse Variational Autoencoders (VAE) for the analysis of distribution shifts in industrial datasets. To this end, several models are compared, in particular, we introduce the LassoVAE, a sparse model with a computationally efficient training. Comparisons are obtained thanks to an experimental protocol we designed that allows to generate synthetic data and different types of shifts with various parameters. New metrics are also introduced to evaluate the models' ability to retrieve the sources of shifts. Results show that sparse models are highly more efficient at recovering the true interactions between variables than a VAE with a dense decoder.