

Tour of VAE

Decoder:

- Recall PPCA: $x_i = Wu_i + \mu + \varepsilon_i$, where $x_i, \varepsilon_i \in \mathbb{R}^p$, $u_i \in \mathbb{R}^d$
- Use $f: \mathbb{R}^d \rightarrow \mathbb{R}^p$ to make it non-linear: $x_i = f(u_i) + \varepsilon_i$
- Bayesian jargon: $p(x|u) = f(u) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_p)$ and $u \sim \mathcal{N}(0, I)$
- or: $x|u \sim \mathcal{N}(f(u), \sigma_\varepsilon^2 I_p)$
- Importantly: fit f with a DNN, a *decoder*

Encoder:

- Posterior $p(u|x)$ will be approximated by $q_\phi(u|x) = \mathcal{N}(\mu(x), \sigma^2(x))$
- fit with a DNN, an *encoder*, outputting d means and (log) variances

Loss:

- Recall: $-ELBO(q) = -\mathbb{E}_q[\log p(x|u)] + D_{KL}[q(u|x)||p(u)]$
- Reconstruction loss: $-\mathbb{E}_q[\log p(x|u)] \propto \mathbb{E}_q[\|x - f(u)\|^2]$
- Regularization term: $D_{KL}[q(u|x)||p(u)] = -\frac{1}{2} \sum_{l=1}^d [1 + \gamma_l - e^{\gamma_l} - \mu_l^2]$, where $\gamma_l = \log \sigma_l^2$, $l = 1, \dots, d$

LMMVAE: Putting RE in VAE

- Problem: Data is not i.i.d.
- Dependencies: q locations, q subjects in longitudinal study, q levels of high-cardinality categorical feature, combinations
- Inspired by LMM: $x_{ij} = f(u_{ij}) + b_j + \varepsilon_{ij}$ ($j = 1, \dots, q; i = 1, \dots, n_j$)
- $b_j \in \mathbb{R}^p$ is a *random effect* (RE) vector, from $\mathcal{N}(0, D)$

In general: $X = f(U) + ZB + \mathcal{E}$, where:

- X of order $n \times p$, $U_{n \times d}$ "fixed" LV
- $f: \mathbb{R}^d \rightarrow \mathbb{R}^p$, "fixed" decoder
- $Z_{n \times q}$ "random" model matrix
- B of order $q \times p$ is a matrix of "random" effects and $B \sim \mathcal{MN}(0, \Psi, D)$
- \mathcal{E} of order $n \times p$, $\mathcal{N}(0, 1)$ noise
- $D = \text{diag}(\sigma_{b_1}^2, \dots, \sigma_{b_p}^2)$
- Z, Ψ depend on the model:

Model	Units	Z	Ψ
Categorical(s)	$Q = \sum q_k$	binary $Z_{n \times Q}$	I_Q
Longitudinal K polynomial terms	q subjects	$[Z_0; \dots; Z_{K-1}]_{n \times Kq}$	$\Phi_{K \times K} \otimes I_q$
Spatial	q locations	binary $Z_{n \times q}$	$K_{q \times q}$ RBF kernel

LMMVAE Loss

Assumption: $p(u, b|x) = p(u|x)p(b|x) \Rightarrow q_\phi(u, b|x) = q_\phi(u|x)q_\phi(b|x)$
 $\Rightarrow D_{KL}[q(u, b|x)||p(u, b)] = D_{KL}[q_\phi(u|x)||p(u)] + D_{KL}[q_\phi(b|x)||p(b)]$

$$-ELBO(q) = \frac{1}{m_{\text{batch}}} \sum_i (x_i - \hat{x}_i)^2 - \frac{\beta}{2} \sum_{l=1}^d [1 + \gamma_{ul} - e^{\gamma_{ul}} - \mu_{ul}^2] - \frac{\beta}{2} \sum_{k=1}^p [1 + \gamma_{bk} - \delta_{bk} - e^{\gamma_{bk} - \delta_{bk}} - \mu_{bk}^2 e^{-\delta_{bk}}]$$

where $\gamma_u = \log \tau_u^2$ and $\gamma_b = \log \tau_b^2$, $\delta_b = \log \sigma_b^2$, some prior, β is a h.p. (β -VAE)

LMMVAE Scheme

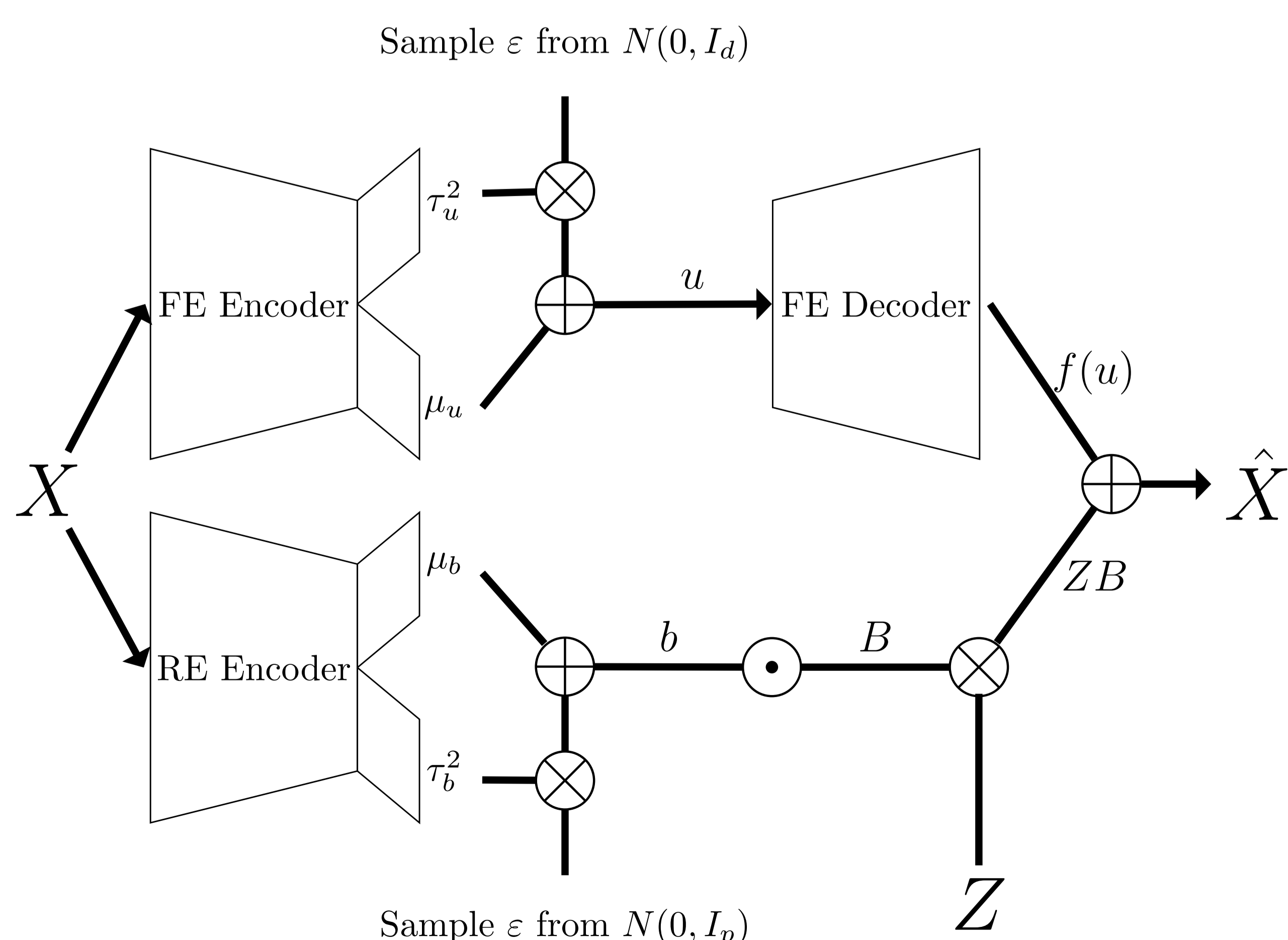


Figure 1. LMMVAE architecture: data X enters two separate FE and RE encoders, to produce the fixed LV u and RE b by the reparameterization trick. u goes through the FE decoder, its output $f(u)$ is added the RE term ZB after Z enters the model and multiplies the RE matrix B after it had been properly formed from the b RE vectors, as depicted by the \odot symbol (see the different correlation scenarios for more details). This produces the final reconstructions \hat{X} .

Simulated Data Results

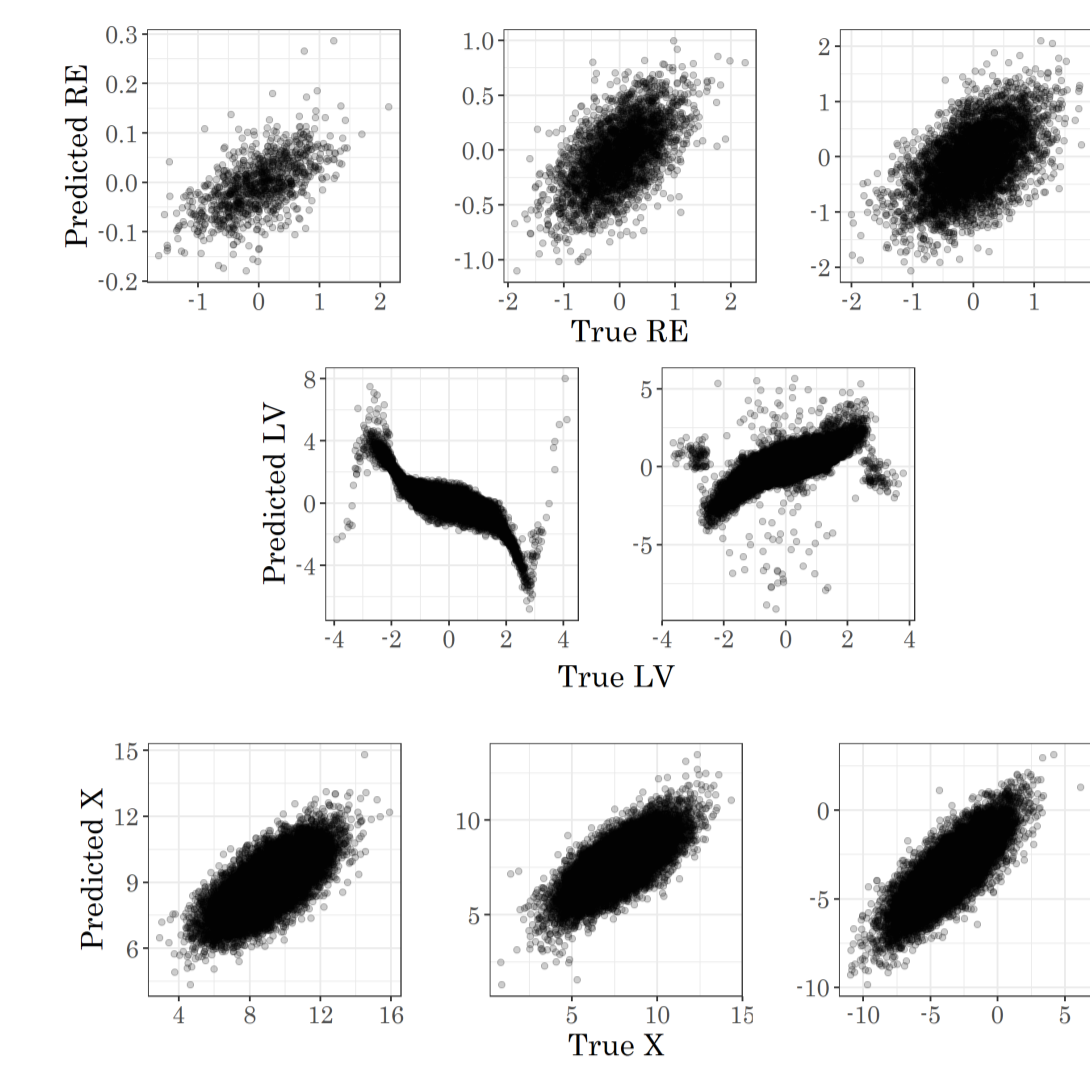


Figure 2. Simulated datasets with high-cardinality categorical features: predicted vs. true scatter plots for $n = 100000$ observations, $p = 100$ fixed features, 3 categorical features, with $q_1 = 1000, q_2 = 3000, q_3 = 5000$. First columns of B, U_{te}, X_{te} .

	$d = 1$		$d = 2$	
	Random mode	Future mode	Random mode	Future mode
PCA-Ignore	2.37 (.04)	2.54 (.03)	1.84 (.0)	1.98 (.0)
PCA-OHE	2.38 (.04)	2.56 (.03)	1.82 (.0)	1.98 (.0)
VAE-Ignore	2.19 (.02)	2.20 (.01)	2.27 (.0)	2.40 (.0)
VAE-OHE	2.24 (.01)	2.32 (.02)	2.22 (.0)	2.41 (.0)
VAE-Embed	2.18 (.02)	2.22 (.01)	2.06 (.0)	2.04 (.0)
SVGPVAE	2.10 (.0)	2.46 (.0)	2.49 (.0)	2.90 (.0)
VRAE	2.06 (.0)	2.54 (.0)	2.48 (.0)	2.94 (.1)
LMMVAE	1.41 (.05)	1.53 (.02)	1.14 (.0)	1.28 (.0)

Table 1. Mean test reconstruction errors for simulated model with 3 high-cardinality categorical features (left) and a longitudinal model with $q = 1000$ subjects, and $K = 3$ polynomial terms on t (right).

Real Data Results

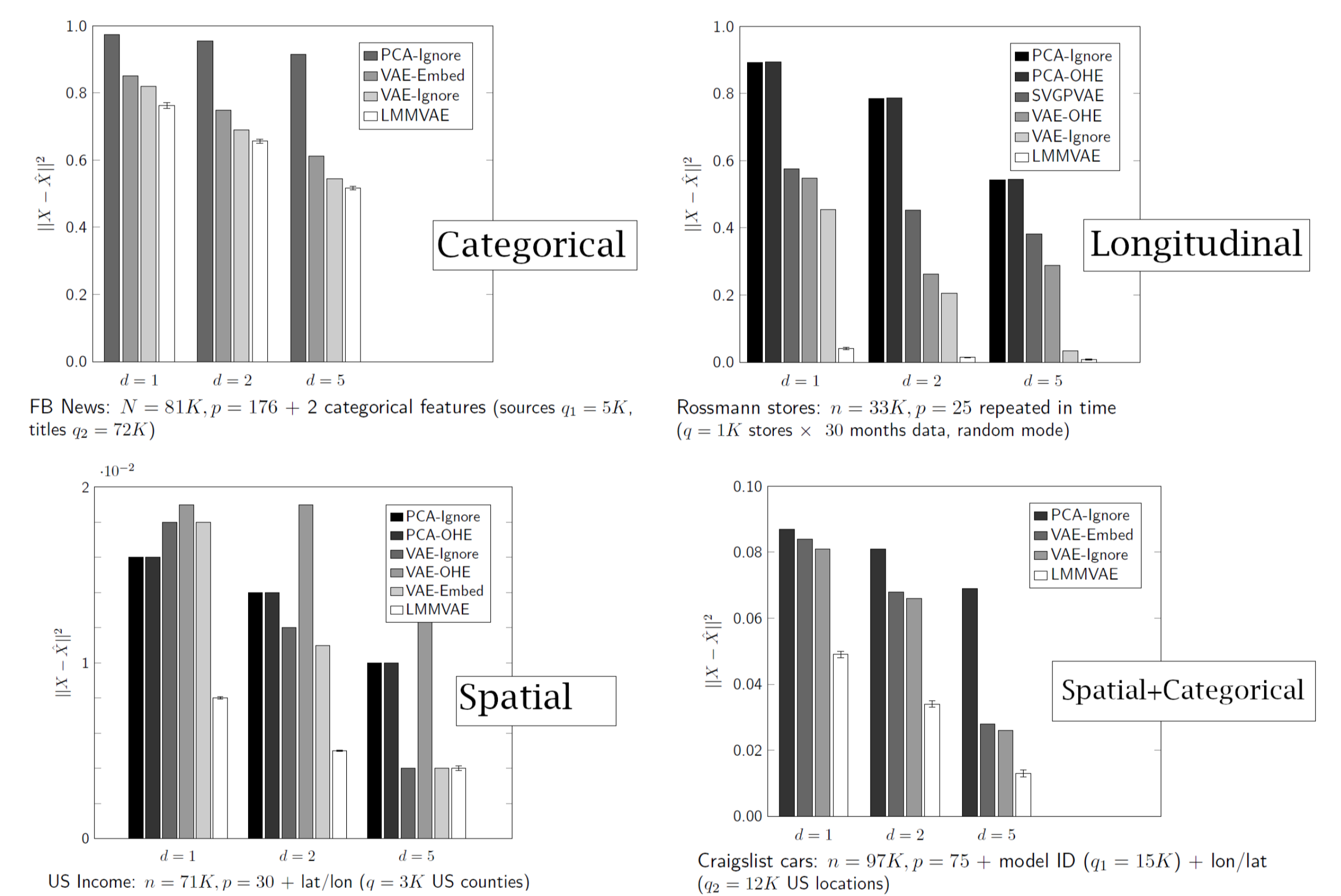


Figure 3. Real datasets with various covariance structures: mean reconstruction loss on 20% unseen data with 5-CV

Additional Viz

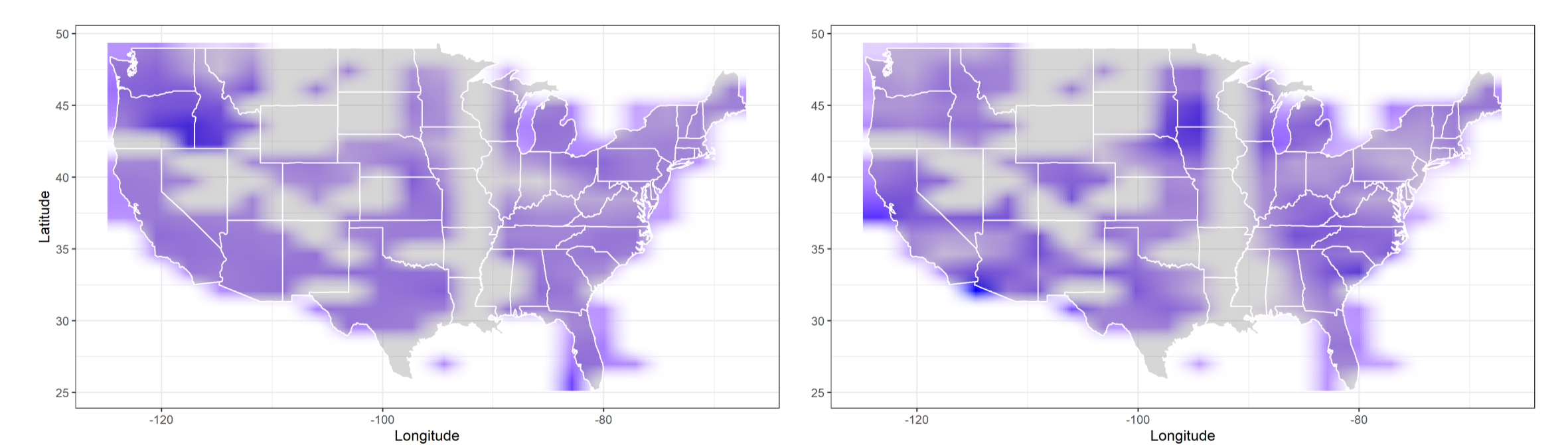


Figure 4. Exploring the \hat{B} RE matrix from the Cars dataset containing spatial features. Left: distribution across the US of the \hat{B} column corresponding to the price feature; Right: the \hat{B} column of the odometer feature.

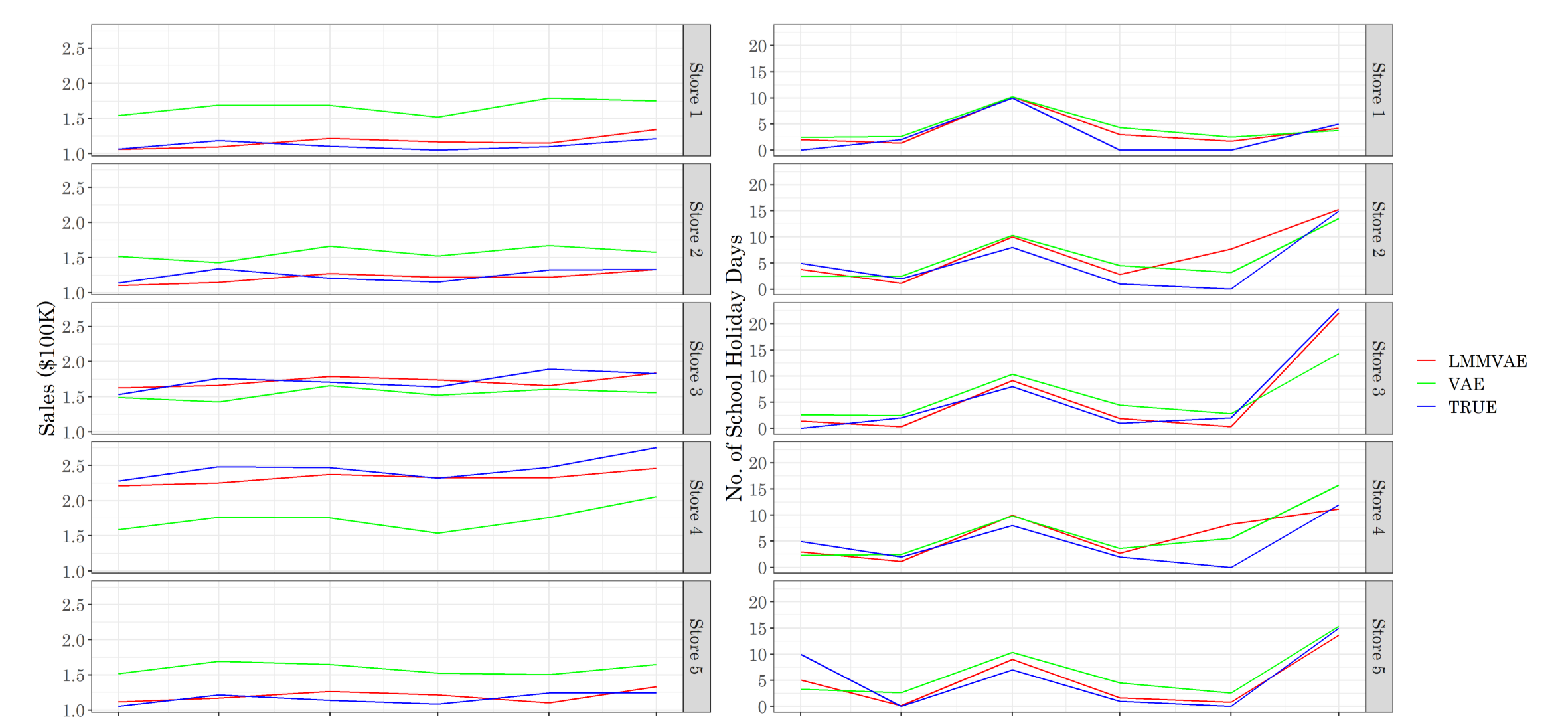


Figure 5. Comparing true vs. reconstructed X_{te} for the Rossmann dataset. The model is trained on the first 25 months of dataset to reconstruct the last 6 months. Left: sales feature; Right: school days feature.