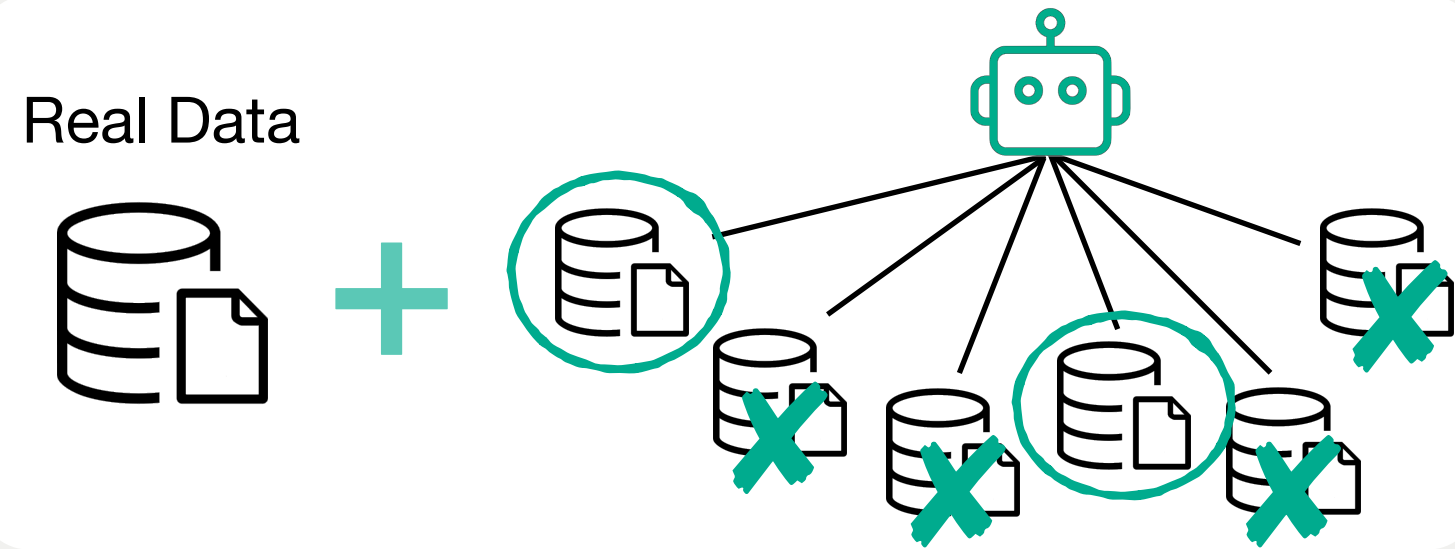


Problem

How to select synthetic data?



Setting

Data model



$$\begin{aligned}
 x_i &\sim \mathcal{N}(\mu_t, \Sigma_t) \\
 \varepsilon_i &\sim \mathcal{N}(0, \sigma^2) \\
 y_i &= \beta^\top x_i + \varepsilon_i \\
 \{y_i, x_i\}_{i=1}^{n_t}
 \end{aligned}$$

$$\begin{aligned}
 &\mathbb{R} \quad \mathbb{R}^p \quad \mathbb{R}^p \quad \mathbb{R} \\
 &y = \beta^\top x + \varepsilon \\
 &0 < \tau \lesssim \frac{n_t}{p}, \frac{n_s}{p} \lesssim \tau^{-1}
 \end{aligned}$$



$$\begin{aligned}
 x_j &\sim \mathcal{N}(\mu_s, \Sigma_s) \\
 \varepsilon^j &\sim \mathcal{N}(0, \sigma^2) \\
 y_j &= \beta^\top x_j + \varepsilon^j \\
 \{y_j, x_j\}_{j=1}^{n_s}
 \end{aligned}$$

Performance measure

$$R(\beta, \hat{\beta}) = \mathbb{E}_{x \sim \mathcal{N}(\mu_t, \Sigma_t)} |\hat{\beta}^\top x - \beta^\top x|^2$$

Estimator

$$\hat{\beta} := \{ \|b\|_2 : b \text{ minimizes } \sum_{i=1}^{n_t} |b^\top x_i - y_i|^2 + \sum_{j=1}^{n_s} |b^\top x_j - y_j|^2 \}$$

Main results

Theorem

There exists a deterministic value $\mathcal{R}(\Sigma_t, \Sigma_s)$ such that, with high probability, $\mathcal{R}(\Sigma_t, \Sigma_s) \approx R(\beta, \hat{\beta})$.

Risk only depends on the covariance matrices, and not on the means.

Theorem

For a fixed Σ_t it holds that $\operatorname{argmin}_{\Sigma} \mathcal{R}(\Sigma_t, \Sigma) \sim \Sigma_t$.

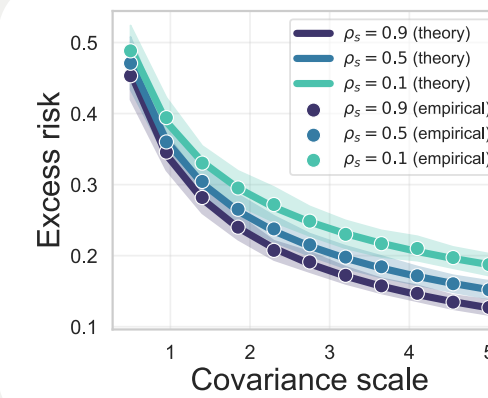
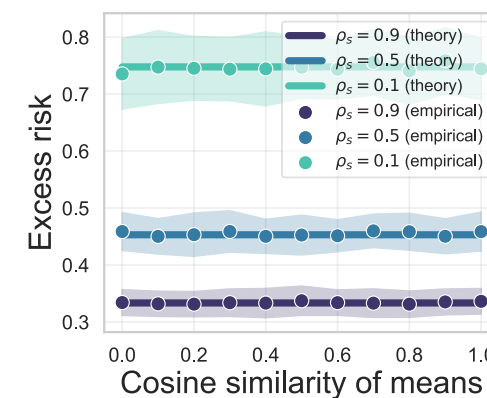
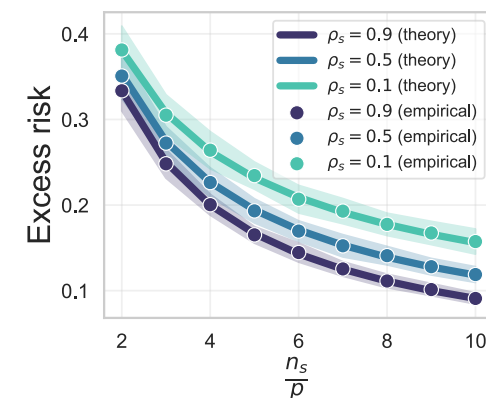
Given a training covariance Σ_t , optimal to choose synthetic data with $\Sigma_s \propto \Sigma_t$, i.e., **match the covariances**.

Formally stated

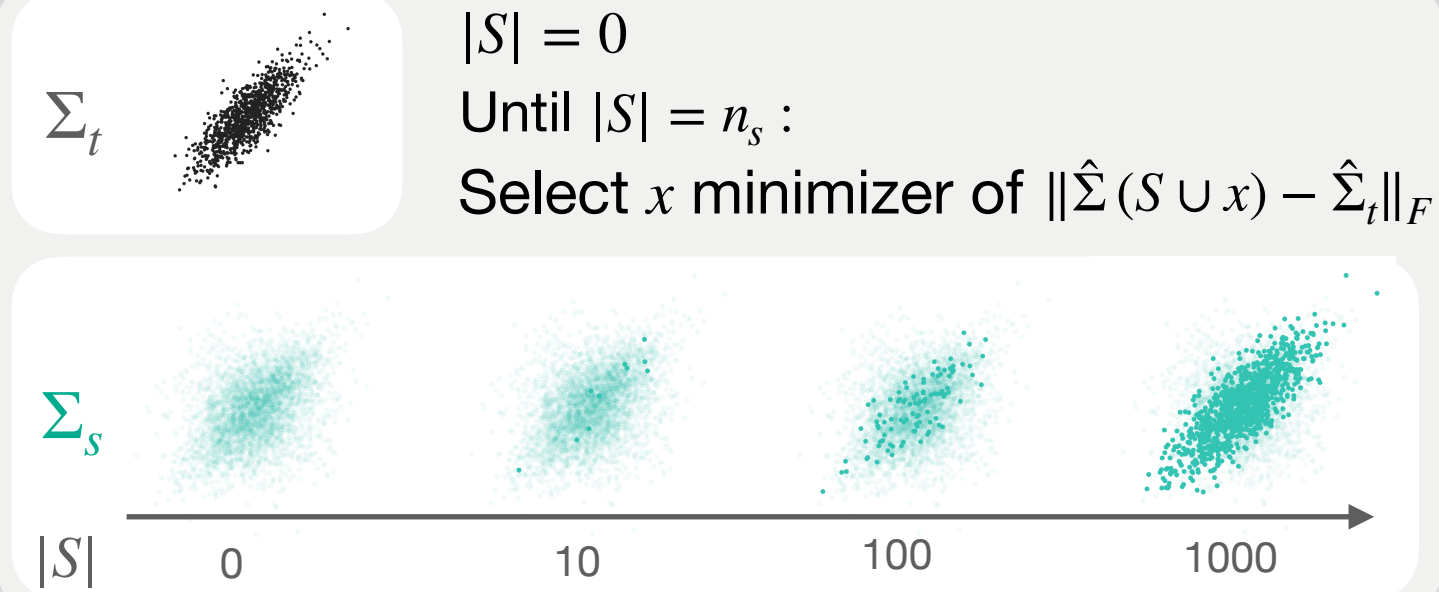
It holds that $\lim_{p \rightarrow \infty} |R(\beta, \hat{\beta}) - \mathcal{R}(\Sigma_t, \Sigma_s)| = 0$, where for $n = n_s + n_t$

$$\mathcal{R}(M) = \frac{\sigma^2}{n} \operatorname{Tr} \left[\left(\alpha_1 \Sigma_t^{-1/2} \Sigma_s \Sigma_t^{-1/2} + \alpha_2 I_p \right)^{-1} \right],$$

such that $\alpha_1 + \alpha_2 = 1 - \frac{p}{n}$, $\alpha_1 + \frac{1}{n} \sum_{i=1}^p \frac{\lambda_i \alpha_1}{\lambda_i \alpha_1 + \alpha_2} = \frac{n_s}{n}$, $\lambda_i = \lambda_i(\Sigma_t^{-1/2} \Sigma_s \Sigma_t^{-1/2})$.



Covariance Matching



Experiments

CIFAR 10

Method	Scratch	Distillation	Pretrained
No synthetic	44.36 $\pm$ 1.51	47.33 $\pm$ 0.57	63.40 $\pm$ 1.33
Center matching	50.04 $\pm$ 2.84	53.83 $\pm$ 0.59	67.01 $\pm$ 0.89
Center sampling	50.48 $\pm$ 2.03	54.91 $\pm$ 1.07	67.71 $\pm$ 0.90
DS3	52.83 $\pm$ 2.19	58.32 $\pm$ 0.43	68.21 $\pm$ 0.66
K-means	50.74 $\pm$ 1.77	56.06 $\pm$ 0.68	66.50 $\pm$ 1.11
Random	49.38 $\pm$ 2.43	54.89 $\pm$ 0.91	67.65 $\pm$ 0.77
Text matching	50.94 $\pm$ 1.40	55.17 $\pm$ 0.57	67.81 $\pm$ 0.76
Text sampling	50.28 $\pm$ 1.18	54.82 $\pm$ 0.72	67.45 $\pm$ 1.02
Covariance matching (ours)	54.00 $\pm$ 1.89	59.77 $\pm$ 0.61	69.20 $\pm$ 0.56
Real upper bound	61.08 $\pm$ 2.54	65.38 $\pm$ 0.51	74.35 $\pm$ 0.56

ImageNet 100

Method	Truncated models	T2I models
No synthetic	40.78 $\pm$ 1.29	
Center matching	53.39 $\pm$ 0.37	53.96 $\pm$ 1.06
DS3	57.47 $\pm$ 0.87	53.51 $\pm$ 0.31
Random	54.14 $\pm$ 0.82	49.84 $\pm$ 1.32
Text matching	53.39 $\pm$ 0.99	53.37 $\pm$ 0.72
Covariance matching (ours)	57.52 $\pm$ 0.36	53.07 $\pm$ 0.89
Real upper bound	62.67 $\pm$ 0.65	

RxRx1

Method	MorphGen
No synthetic	86.83 $\pm$ 2.44
Center matching	88.17 $\pm$ 2.35
Random	87.33 $\pm$ 2.03
K-means	89.00 $\pm$ 1.70
DS3	89.67 $\pm$ 1.45
Center sampling	88.75 $\pm$ 2.27
Covariance matching (ours)	90.00 $\pm$ 1.86