A Comparative Study on Bayesian SVDD Models

Jongwu Kim

SKKUSTAT

06.Dec.2024

Motivation

Many sub-branches of Support Vector Machines (SVM; Hearst et al, 1998) have emerged.

Some of them specialize in handling imbalanced data.

- i.e. Support Vector Data Description (SVDD; Tax & Duin, 2004)

Recently, Bayesian modifications of SVDD such as those below were introduced.

- Bayesian Data Description (BDD; Ghasami et al., 2021)
- Bayesian SVDD (BSVDD; 오정민, 2023)
- Bayesian SVDD with Minor-class (BSVDD-M; 배희진, 2024)

Objective

The comparison of models below under various situations provides guidelines for using them.

- Support Vector Data Description (SVDD; Tax & Duin, 2004)
- Bayesian Data Description (BDD; Ghasami et al., 2021)
- Bayesian SVDD (BSVDD; 오정민, 2023)
- Bayesian SVDD with Minor-class (BSVDD-M; 배희진, 2024)

The Objective: **An intensive comparison of the 4 models related to SVDD across various data settings**

Imbalanced Data

The binary imbalanced dataset has the form below.

$$Y_i = \begin{cases} 1, & i = j \text{ for } j = 1, \cdots, n_1 \\ 0, & i = l \text{ for } l = 1, \cdots, n_0 \end{cases} \qquad \forall j \neq l$$

with $n = n_0 + n_1,\ \ n_0 >> n_1,\ \ \boldsymbol{x} = (\boldsymbol{x_0}^T, \boldsymbol{x_1}^T)^T \in R^{n \times p}$,
where $\boldsymbol{x_0} = (x_1^{(0)}, \cdots, x_{n_0}^{(0)})^T,\ \ \boldsymbol{x_1} = (x_1^{(1)}, \cdots, x_{n_1}^{(1)})^T$

- If n_1 is much smaller than n_0, $\boldsymbol{x_1} = (x_1^{(1)}, \cdots, x_{n_1}^{(1)})$ can be seen as anomalies.
- For the Anomaly Detection, SVDD(Tax & Duin, 2004) can be considered
  - SVDD is a one-class learning model and a sibling of the famous SVM(Hearst et al., 1998)

Support Vector Data Description (Tax & Duin, 2004)

SVDD has an objective function similar to that of SVM but uses only the $\boldsymbol{x_0} = (x_1^{(0)}, \cdots, x_{n_0}^{(0)})^T$, the major class.

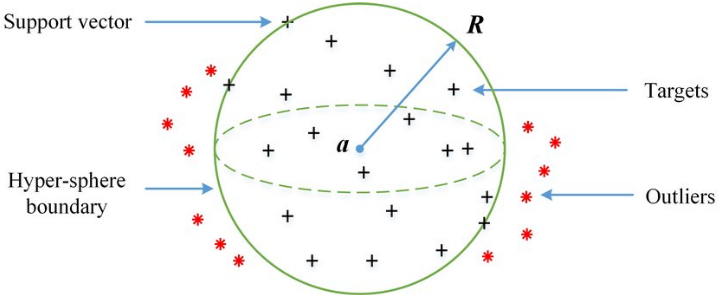
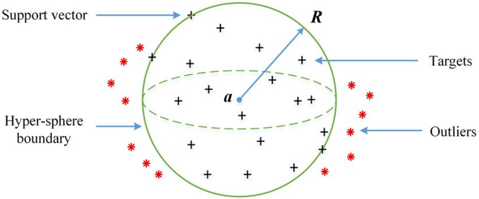


Figure: Liu et al., (2021)

Components of the Objective Function



The objective function

$$\min_{R,\mathbf{a},\xi_i^{(0)}} \quad R^2 + C\sum_{i=1}^{n_0}\xi_i^{(0)} \quad s.t.\ ||\phi(x_i^{(0)}) - \mathbf{a}||^2 \le R^2 + \xi_i^{(0)},\ \ \xi_i^{(0)} \ge 0,\ \ \forall i$$

$$\begin{cases} \mathbf{a} := \text{Center of the hypersphere} \\ R := \text{Radius of the hypersphere} \\ \phi(x_i^{(0)}) := i^{\text{th}} \text{ major class datapoint in kernel space} \\ \xi_i^{(0)} := \text{Amount of `slack' of the } i^{\text{th}} \text{ support vecter} \\ C := \text{Some constant} \end{cases}$$

Convert the Function into the Lagrangian Form

Lagrangian form of the objective function

$$\min_{R,\mathbf{a},\xi_i^{(0)}} \ (\max_{\alpha_i,\beta_i}) \ L,$$

where $L = R^2 + C\sum_{i=1}^{n_0}\xi_i^{(0)} - \sum_{i=1}^{n_0}\alpha_i(R^2+\xi_i^{(0)} - ||\phi(x_i^{(0)})-\mathbf{a}||^2) - \sum_{i=1}^{n_0}\beta_i\xi_i^{(0)}$

for $\alpha_i \geq 0, \ \ \beta_i \geq 0$

By the Karush-Kuhn-Tucker (KKT) condition,

$\min_{R,\mathbf{a},\xi_i^{(0)}} \ (\max_{\alpha_i,\beta_i}) \ L$ collapses into $\max_{\alpha_i} \ L, \ \forall i$. (See page 34∼36)

Forge Support Vectors

Some equations obtained by the KKT(See page 34) :

$$R^2 + \xi_i^{(0)} - ||\phi(x_i^{(0)}) - \mathbf{a}||^2 \geq 0 \;\; \text{(P1)}, \quad \alpha_i(R^2 + \xi_i^{(0)} - ||\phi(x_i^{(0)}) - \mathbf{a}||^2) = 0 \;\; \text{(C1)}$$

$$\beta_i \xi_i^{(0)} = 0 \quad \text{(C2)}, \qquad \mathbf{a} = \sum_{i=1}^{n_0} \alpha_i \phi(x_i^{(0)}) \quad \text{(D2)}, \qquad C - \alpha_i - \beta_i = 0 \quad \text{(D3)}$$

$$\xi_i^{(0)} \begin{cases} = 0 \text{ for } \phi(x_i^{(0)}) \text{ on the hypersphere} \\ > 0 \text{ for } \phi(x_i^{(0)}) \text{ out of the hypersphere} \end{cases} \quad \text{(def)}$$

The rule deciding support vectors

$$\begin{cases} \text{Case 1: } ||\phi(x_i^{(0)}) - \mathbf{a}||^2 < R^2 \overset{\text{P1}}{\to} R^2 + \xi_i^{(0)} - ||\phi(x_i^{(0)}) - \mathbf{a}||^2 > 0 \overset{\text{C1}}{\to} \alpha_i = 0 \\ \text{Case 2: } ||\phi(x_i^{(0)}) - \mathbf{a}||^2 = R^2 (\overset{\text{def}}{\leftrightarrow} \boldsymbol{\xi_i^{(0)} = 0}) \overset{\text{C1}}{\to} 0 \leq \alpha_i \leq C \overset{\text{D2}}{\to} \boldsymbol{0 < \alpha_i \leq C} \\ \text{Case 3: } ||\phi(x_i^{(0)}) - \mathbf{a}||^2 > R^2 (\overset{\text{def}}{\leftrightarrow} \boldsymbol{\xi_i^{(0)} > 0}) \overset{\text{C2}}{\to} \beta_i = 0 \overset{\text{D3}}{\to} \boldsymbol{\alpha_i = C} \end{cases}$$

Case 2 & 3 preserves α_i used in $\underset{\alpha_i}{max}\ L\ (\leftrightarrow \underset{R,\mathbf{a},\xi_i^{(0)}}{min}\ R^2)$

$\therefore x_i^{(0)}$ for Case 2 & 3 supports the hypersphere $\to$ Support Vector(S.V.)

Support Vectors support the Hypersphere

The rule deciding support vectors

$$\begin{cases} \text{Case 1: } ||\phi(x_i^{(0)}) - \mathbf{a}||^2 < R^2 \overset{\text{P1}}{\to} R^2 + \xi_i^{(0)} - ||\phi(x_i^{(0)}) - \mathbf{a}||^2 > 0 \overset{\text{C1}}{\to} \alpha_i = 0 \\ \text{Case 2: } ||\phi(x_i^{(0)}) - \mathbf{a}||^2 = R^2 (\overset{\text{def}}{\leftrightarrow} \boldsymbol{\xi_i^{(0)} = 0}) \overset{\text{C1}}{\to} 0 \le \alpha_i \le C \overset{\text{D2}}{\to} \boldsymbol{0 < \alpha_i \le C} \\ \text{Case 3: } ||\phi(x_i^{(0)}) - \mathbf{a}||^2 > R^2 (\overset{\text{def}}{\leftrightarrow} \boldsymbol{\xi_i^{(0)} > 0}) \overset{\text{C2}}{\to} \beta_i = 0 \overset{\text{D3}}{\to} \boldsymbol{\alpha_i = C} \end{cases}$$

After creating the hypersphere, S.V. $x_k^{(0)}$ in the Case 2 forms the radius.

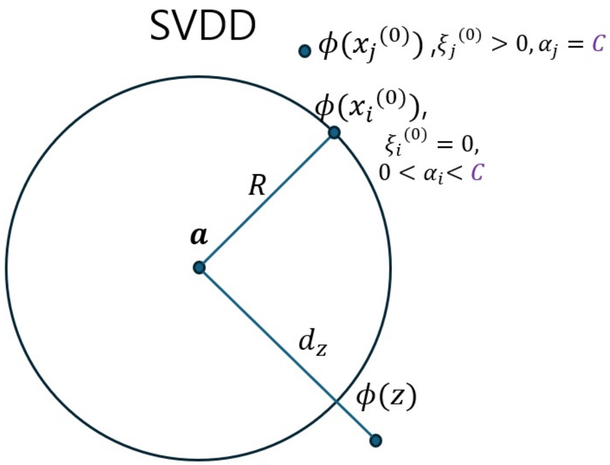
$$R^2 = ||\phi(x_k^{(0)}) - \mathbf{a}||^2$$

$$\downarrow$$

The decision rule of of SVDD

$$d_z^2 := ||\phi(z) - \mathbf{a}||^2 \begin{cases} \le R^2 & \to & z \text{ is normal (major class)} \\ > R^2 & \to & z \text{ is abnormal (minor class)} \end{cases}$$

Visualized Decision Rule



$$\min R^2 + C \sum_{i=1}^{n_0} \xi_i^{(0)}$$

Bayesian Data Description (Ghasami et al., 2021)

$\mathbf{a} = \sum_{i=1}^{n_0} \alpha_i \phi(x_i^{(0)})$: **Center** of the hypersphere in the SVDD.

Let $\boldsymbol{\alpha_0} := (\alpha_1 \cdots \alpha_{n_0})^T \in R^{n_0}$.

Bayes' theorem: $P(\boldsymbol{\alpha_0}|\phi(x_i^{(0)})) \propto P(\phi(x_i^{(0)})|\boldsymbol{\alpha_0})P(\boldsymbol{\alpha_0})$

Likelihood & Prior Distribution of $\boldsymbol{\alpha_0}$

$$\phi(x_i^{(0)})|\boldsymbol{\alpha_0} \sim MVN(\mathbf{a}, I), \quad \boldsymbol{\alpha_0} \sim MVN(\boldsymbol{m}, \boldsymbol{\Sigma})$$

Under the assumption: $\phi(x_i^{(0)})$ will be gathered around **a**, the center.

Spoiler! α_i has a limited range. We will see this later.

Get the estimator of $\boldsymbol{\alpha_0}$. (See page 37∼38)

MAP(Maximum a posteriori) estimator of $\boldsymbol{\alpha_0}$

$$\hat{\boldsymbol{\alpha_0}} = \underset{\boldsymbol{\alpha_0}}{argmin}\ \boldsymbol{\alpha_0}^T(n_0K_0 + \boldsymbol{\Sigma}^{-1})\boldsymbol{\alpha_0} - 2\boldsymbol{\alpha_0}^T(D_0\mathbf{1}_{\boldsymbol{n_0}} + \boldsymbol{\Sigma}^{-1}\boldsymbol{m})$$

Decision rule with the MAP estimator

MAP(Maximum a posteriori) estimator of $\boldsymbol{\alpha_0}$

$$\hat{\boldsymbol{\alpha}}_{\mathbf{0}} = \underset{\boldsymbol{\alpha_0}}{argmin}\ \boldsymbol{\alpha}_{\mathbf{0}}^T(n_0 K_0 + \boldsymbol{\Sigma}^{-1})\boldsymbol{\alpha_0} - 2\boldsymbol{\alpha}_{\mathbf{0}}^T(D_0 \mathbf{1}_{\boldsymbol{n_0}} + \boldsymbol{\Sigma}^{-1}\boldsymbol{m})$$

$$\rightarrow \quad \hat{\mathbf{a}} = \sum_{i=1}^{n_0} \hat{\alpha}_i \phi(x_i^{(0)}) \quad \rightarrow \quad ||\phi(x_k^{(0)}) - \hat{\mathbf{a}}||^2 =: d_k^2$$

Sort the distances d_k, $\forall k = 1, \cdots, n_0$ and set a **cutoff** $c (\leq n_0)$ then

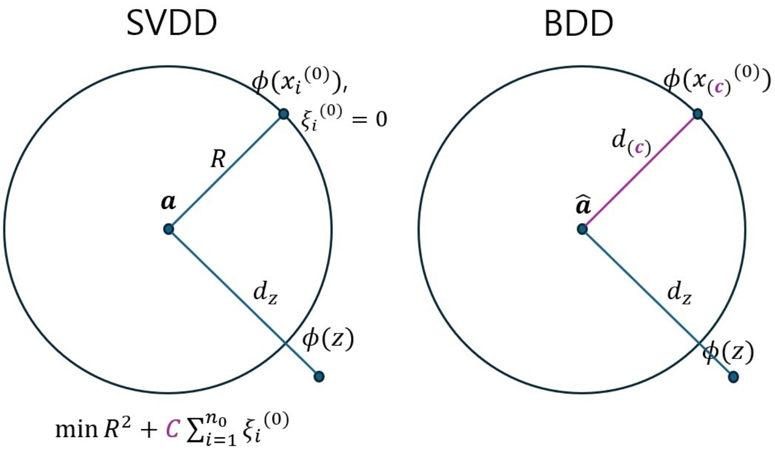
$$d_{(1)}{}^2 = ||\phi(x_{(1)}^{(0)}) - \hat{\mathbf{a}}||^2, \ \cdots, \ d_{(c)}{}^2 = ||\phi(x_{(c)}^{(0)}) - \hat{\mathbf{a}}||^2$$

$$\downarrow$$

The decision rule of of BDD

$$d_z^2 := ||\phi(z) - \hat{\mathbf{a}}||^2 \begin{cases} \leq d_{(c)}{}^2 & \rightarrow \quad z \text{ is normal (major class)} \\ > d_{(c)}{}^2 & \rightarrow \quad z \text{ is abnormal (minor class)} \end{cases}$$

Visualized Comparison



Prior Distribution of $\boldsymbol{\alpha_0}$ in the BDD

Prior Distribution of $\boldsymbol{\alpha_0}$

$$\boldsymbol{\alpha_0} \sim MVN(\boldsymbol{m},\ \boldsymbol{\Sigma})$$

Without prior knowledge, Ghasami et al.(2012) suggests

$$\begin{cases} m_i \propto -\sum_{j=1}^{n_0} K_{i,j} \ \rightarrow \ m_i = -(\sum_{j=1}^{n_0} K_{i,j})^v, \ \forall i = 1, \cdots, n_0, \ 0 < v < 1 \\ \qquad Q \text{ Why } -\sum_{j=1}^{n_0} K_{i,j}? \\ \qquad A \text{ To set a relatively high value to the data close to the hypersphere} \\ \qquad \left(\text{In the SVDD, } \xi_i \text{ depends on } \alpha_i \text{ in the center } \mathbf{a} = \sum_{i=1}^{n_0} \alpha_i \phi(x_i^{(0)})\right) \\ \boldsymbol{\Sigma} = I \end{cases}$$

"We limit α_i values to form a convex set, i.e.

$\sum_{i=1}^{n_0} \alpha_i = 1, \ 0 < \alpha_i < 1, \ \forall i$" (Ghasami et al., 2012)

Reparameterization

To match the prior dist^n $\boldsymbol{\alpha_0} \sim MVN(\boldsymbol{m}, \boldsymbol{\Sigma})$
and the constraints $0 < \alpha_i < 1, \ \forall i$,

Reparameterization of α_i

$$\alpha_i = \left\{ exp(\beta_i) \Big/ \sum_{i=1}^{n_0} exp(\beta_i) \right\}, \ \ \forall i = 1, \cdots, n_0$$

Need to set prior dist^n to the newly introduced $\boldsymbol{\beta}$.

Prior distribution of $\boldsymbol{\beta}$

$\boldsymbol{\beta} \sim MVN(\boldsymbol{m}, \boldsymbol{\Sigma})$, where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{n_0})^T$

With no prior knowledge, 오정민 (2023) suggests

$$m_i = -\sum_{j=1}^{n_0} K_{i,j}, \forall i, \ \ \boldsymbol{\Sigma} = I$$

Recall that Ghasemi et al.(2012) proposed $\alpha_i \sim N\big(-(\sum_{j=1}^{n_0} K_{i,j})^{\nu}, \ 1\big)$

Derive the Posterior distribution

$$
\begin{aligned}
log\ P(\boldsymbol{\beta}|\Phi(\boldsymbol{x_0})) &\propto log\ P(\Phi(\boldsymbol{x_0})|\boldsymbol{\beta})P(\boldsymbol{\beta}) \\
&\vdots \\
&\propto -\frac{1}{2}\Big[-2\boldsymbol{\alpha_0}^T D_0 \mathbf{1}_{\boldsymbol{n_0}} + n_0\boldsymbol{\alpha_0}^T K_0 \boldsymbol{\alpha_0} \\
&\qquad + (\boldsymbol{\beta}-\boldsymbol{m})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\Big]
\end{aligned}
$$

(See page 39∼40 for the notations & the calculation)

The above is not a well-known distribution.

∴ Apply Metropolis-Hastings(MH).

Apply Metropolis

Recall that

$$\begin{aligned} P(\boldsymbol{\beta}|\Phi(\boldsymbol{x_0})) &\propto P(\Phi(\boldsymbol{x_0})|\boldsymbol{\beta}) \cdot P(\boldsymbol{\beta}) \\ &= N(\mathbf{a}, I) \cdot N(\boldsymbol{m}, \boldsymbol{\Sigma}) \end{aligned}$$

$\therefore$ Use Normal dist^n as the Jumping dist^n(J)

The J is symmetric so it can be canceled out from the acceptance ratio r

$$r = \frac{P(\theta^*|y)\cancel{J(\theta^*|\theta^{(t-1)})}}{P(\theta^{(t-1)}|y)\cancel{J(\theta^{(t-1)}|\theta^*)}}$$

$\rightarrow$ We can forget the J

$\rightarrow$ Not Metropolis-Hastings,
but just Metropolis.

Make distance samples

Now we get $\boldsymbol{\beta}$'s samples by MCMC. Recall the decision rule of SVDD:

The decision rule of of SVDD

$$d_z^2 := ||\phi(z) - \mathbf{a}||^2 \begin{cases} \leq R^2 & \rightarrow \quad z \text{ is normal (major class)} \\ > R^2 & \rightarrow \quad z \text{ is abnormal (minor class)} \end{cases}$$

For an obs. z, get the distance d_z 's posterior samples
from $\boldsymbol{\beta}$'s posterior samples

$$P(\boldsymbol{\beta}|\Phi(\boldsymbol{x_0})) \;\rightarrow\; \alpha_j = \frac{e^{\beta_j}}{\sum\limits_{i=1}^{n_0} e^{\beta_i}}, \;\; \forall j$$

$$\rightarrow \begin{cases} \mathbf{a} = \sum\limits_{j=1}^{n_0} \alpha_j \phi(x_j^{(0)}) \\ P(\mathbf{a}|\Phi(\boldsymbol{x_0})) \end{cases} \rightarrow \begin{cases} d_z := \sqrt{||\phi(z) - \mathbf{a}||^2} \\ P(d_z|\Phi(\boldsymbol{x_0})) \end{cases}$$

The posterior dist^n $P(d_z|\Phi(\boldsymbol{x_0}))$, can be utilized for a decision rule.

Decision rule with posterior dist^n of d_z

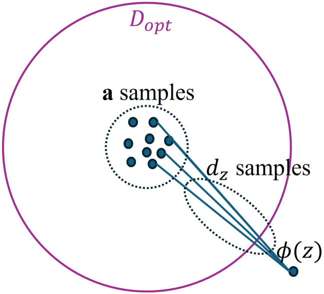


Figure: MCMC samples of $P(\mathbf{a}|\Phi(\boldsymbol{x_0}))$ and $P(d_z|\Phi(\boldsymbol{x_0}))$ for one z

Need to implement D_{opt} to use the posterior dist^n of d_z.

The decision rule of BSVDD

$$P(d_z \leq D_{opt}|\Phi(\boldsymbol{x_0})) \begin{cases} > 0.5 & \rightarrow \quad z \text{ is normal (major class)} \\ \leq 0.5 & \rightarrow \quad z \text{ is abnormal (minor class)} \end{cases}$$

E.g. Slice the distance samples with a Cutoff

Choose the cutoff D_{opt} by Cross-Validation

E.g., d_zs' distributions with $D_{opt} = 19.9$

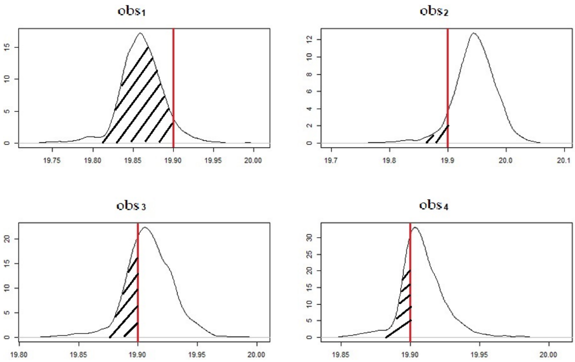
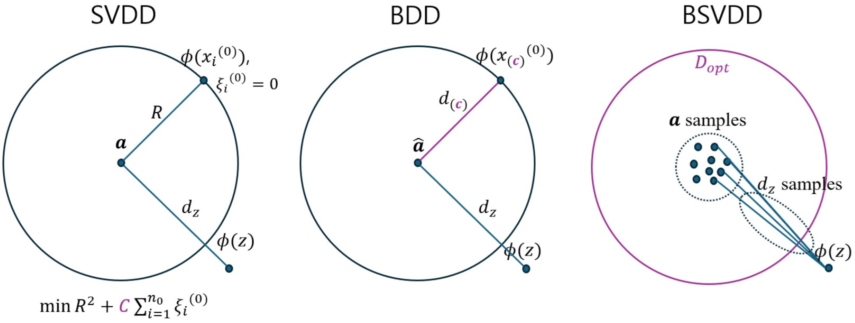


Figure: Example of $P(d_z \leq D_{opt}|\Phi(\boldsymbol{x_0}))$, (오정민, 2023)

- Only the 1st one is predicted as normal

Visualized Comparison



Exploit the Minor-class

Recall that **SVDD uses data from only the major class.**

If the objective function is modified a bit,
we can use the minor class to build the hypersphere. (Tax & Duin, 2004)

By applying the KKT condition to the Lagrangian form,

$$\mathbf{a} = \sum_{i=1}^{n_0} \alpha_i^{(0)} \phi(x_i^{(0)}) - \sum_{l=1}^{n_1} \alpha_l^{(1)} \phi(x_l^{(1)}).$$

(See page 41∼42 for the notations & the calculations)

Likelihood & Prior dist^n

$\mathbf{a} = \sum_{i=1}^{n_0} \alpha_i^{(0)} \phi(x_i^{(0)}) - \sum_{l=1}^{n_1} \alpha_l^{(1)} \phi(x_l^{(1)})$: Center of the hypersphere.

Need prior dist^n for $\alpha_i^{(0)}, \forall i; \quad \alpha_l^{(1)}, \forall l$.

Turkoz & Kim (2022) implemented BDD with the minor class.

$$P(\boldsymbol{\alpha_0}, \boldsymbol{\alpha_1} | \phi(x_i^{(0)}), \phi(x_l^{(1)})) \propto P(\phi(x_i^{(0)}), \phi(x_l^{(1)}) | \boldsymbol{\alpha_0}, \boldsymbol{\alpha_1}) P(\boldsymbol{\alpha_0} | \boldsymbol{\alpha_1}) P(\boldsymbol{\alpha_1})$$

where $\boldsymbol{\alpha_0} = (\alpha_1^{(0)} \cdots \alpha_{n_0}^{(0)})^T$ & $\boldsymbol{\alpha_1} = (\alpha_1^{(1)} \cdots \alpha_{n_1}^{(1)})^T$

Likelihood & Prior Dist^n of $\boldsymbol{\alpha_0}, \boldsymbol{\alpha_1}$ (Turkoz & Kim, 2022)

$$\phi(x_i^{(0)}) | \phi(x_l^{(1)}), \boldsymbol{\alpha_0}, \boldsymbol{\alpha_1} \sim MVN(\mathbf{a}, I),$$
$$\boldsymbol{\alpha_0} | \boldsymbol{\alpha_1} \sim MVN(\boldsymbol{m_1}, \ \boldsymbol{\Sigma}_1), \quad \boldsymbol{\alpha_1} \sim MVN(\boldsymbol{m_2}, \ \boldsymbol{\Sigma}_2)$$

with $\sum_{j=1}^{n_0} \alpha_j^{(0)} - \sum_{l=1}^{n_1} \alpha_l^{(1)} = 1, \ 0 \leq \alpha_l^{(1)} \leq 1$ (Turkoz & Kim, 2022)

Reparametrization

To meet the constraints, apply reparametrization to the priors.

Reparameterization of $\boldsymbol{\alpha_0}, \boldsymbol{\alpha_1}$

$$\alpha_l^{(1)} \sim Beta(a,b),\ \ \alpha_i^{(0)}(\boldsymbol{\alpha_1}, \boldsymbol{\tau}) = \left\{ e^{\tau_i} \left(1 + \sum_{l=1}^{n_1} \alpha_l^{(1)} \right) \Big/ \sum_{i=1}^{n_0} e^{\tau_i} \right\}$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{n_0}),\ \ \forall i = 1, \ldots, n_0,\ \ \forall l = 1, \ldots, n_1$

Set $a = b = 1$ if there is no prior knowledge, then we get

$$\alpha_l^{(1)} \sim Beta(1,1) \overset{d}{=} U(0,1)$$

A prior distribution is required for the newly adopted $\boldsymbol{\tau}$

Prior distribution of $\boldsymbol{\tau}$

$$\boldsymbol{\tau} \sim MVN(\boldsymbol{m}, \boldsymbol{\Sigma})$$

With no prior knowledge, 배희진 (2024) suggests

$$m_i = -\sum_{j=1}^{n_0} K_{0,(i,j)}, \forall i,\ \ \boldsymbol{\Sigma} = I,\ \text{where } K_{0,(i,j)} = <\phi(x_i^{(0)}), \phi(x_j^{(0)})>$$

Recall that Ghasemi et al.(2012) proposed $\alpha_i^{(0)} \sim N\big(-(\sum_{j=1}^{n_0} K_{0,(i,j)})^{\nu},\ 1\big)$

Derive the Posterior distribution

$$
\begin{aligned}
log\ P(\boldsymbol{\tau}, \boldsymbol{\alpha_1} | \Phi(\boldsymbol{x})) &\propto log\ P(\Phi(\boldsymbol{x}) | \boldsymbol{\alpha_0}(\boldsymbol{\alpha_1}, \boldsymbol{\tau})) P(\boldsymbol{\tau} | \boldsymbol{\alpha_1}) P(\boldsymbol{\alpha_1}) \\
&\vdots \\
&\propto -\frac{1}{2}\{-2\boldsymbol{\alpha_0}^T D_0 \mathbf{1}_{\boldsymbol{n_0}} + n_0 \boldsymbol{\alpha_0}^T K_0 \boldsymbol{\alpha_0} + n_0 \boldsymbol{\alpha_1}^T K_1 \boldsymbol{\alpha_1} \\
&\quad + 2\boldsymbol{\alpha_1}^T D_{01} \mathbf{1}_{\boldsymbol{n_1}} - 2n_0 \boldsymbol{\alpha_1}^T K_{01}^T \boldsymbol{\alpha_0}\} \\
&\quad - \frac{1}{2}(\boldsymbol{\tau} - \boldsymbol{m})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\tau} - \boldsymbol{m})
\end{aligned}
$$

(See page 43∼46 for the notations & the calculation)

The above is not a well-known distribution.

∴ Apply Metropolis-Hastings(MH).

Apply Metropolis

Recall that

$$
\begin{aligned}
P(\boldsymbol{\tau}, \boldsymbol{\alpha_1} | \Phi(\boldsymbol{x})) &\propto P(\Phi(\boldsymbol{x}) | \boldsymbol{\alpha_0}(\boldsymbol{\alpha_1}, \boldsymbol{\tau})) \cdot P(\boldsymbol{\tau} | \boldsymbol{\alpha_1}) \cdot P(\boldsymbol{\alpha_1}) \\
&= N(\mathbf{a}, I) \cdot N(\boldsymbol{m}, \boldsymbol{\Sigma}) \cdot U(0, 1)
\end{aligned}
$$

$\therefore$ Use Normal dist^n as the Jumping dist^n(J)

The J is symmetric so it can be canceled out from the acceptance ratio r

$$
r = \frac{P(\theta^* | y) \cancel{J(\theta^* | \theta^{(t-1)})}}{P(\theta^{(t-1)} | y) \cancel{J(\theta^{(t-1)} | \theta^*)}} \rightarrow \text{We can forget the } J
$$

$\rightarrow$ Not Metropolis-Hastings,

but just Metropolis.

Make distance samples

Now we get $\boldsymbol{\alpha_1}$'s samples by MCMC. Recall the decision rule of SVDD:

The decision rule of of SVDD

$$d_z^2 := ||\phi(z) - \mathbf{a}||^2 \begin{cases} \leq R^2 & \rightarrow \quad z \text{ is normal (major class)} \\ > R^2 & \rightarrow \quad z \text{ is abnormal (minor class)} \end{cases}$$

For an obs. z, get the distance d_z 's posterior samples
from $\boldsymbol{\alpha_1}$'s posterior samples

$$\begin{aligned} P(\boldsymbol{\tau}, \boldsymbol{\alpha_1} | \Phi(\boldsymbol{x})) \quad &\rightarrow \alpha_i^{(0)}(\boldsymbol{\alpha_1}, \boldsymbol{\tau}) = \frac{e^{\tau_i}\left(1 + \sum\limits_{l=1}^{n_1} \alpha_l^{(1)}\right)}{\sum\limits_{i=1}^{n_0} e^{\tau_i}}, \ \forall i \\ &\rightarrow \mathbf{a} = \sum_{i=1}^{n_0} \alpha_i^{(0)} \phi(x_i^{(0)}) - \sum_{l=1}^{n_1} \alpha_l^{(1)} \phi(x_l^{(1)}), \ \ P(\mathbf{a} | \Phi(\boldsymbol{x})) \\ &\rightarrow d := \sqrt{||\phi(z) - \mathbf{a}||^2}, \ \ P(d_z | \Phi(\boldsymbol{x})) \end{aligned}$$

The posterior dist^n $P(d_z | \Phi(\boldsymbol{x}))$ can be utilized for a decision rule.

Decision rule with posterior dist^n of d_z

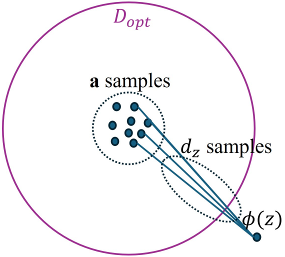


Figure: MCMC samples of $P(\mathbf{a}|\Phi(\boldsymbol{x}))$ and $P(\boldsymbol{d_z}|\Phi(\boldsymbol{x}))$ for one z

Need to implement D_{opt} to use the posterior dist^n of d_z.

The decision rule of BSVDD-M

$$P(d_z \leq D_{opt}|\Phi(\boldsymbol{x_0}), \Phi(\boldsymbol{x_1})) \begin{cases} > 0.5 & \rightarrow \quad z \text{ is normal (major class)} \\ \leq 0.5 & \rightarrow \quad z \text{ is abnormal (minor class)} \end{cases}$$

Unlike the decision rule of BSVDD, the above uses both $\boldsymbol{x_0}, \boldsymbol{x_1}$

Toy example settings

$$\text{Train data size} = 1{,}000, \text{ Test data size} = 10{,}000$$

$$\text{Imbalance ratio} := \frac{\text{Size of the abnormal data}}{\text{Size of the all data}} \approx \frac{1}{10}$$

$$\eta = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots \beta_{50} x_{i,50}$$

v: the tuning parameter for meaningful β s variance.

- 10 pairs of train and test datasets were generated for each $v = \mathbf{5,10,20}$

(See page 47 for the detailed explanation)

All models (SVDD, BDD, BSVDD, BSVDD-M) were fitted and predicted on all pairs of datasets.

Performance chart: SVDD, BDD

SVDD						
v	5		10		20	
Accuracy	0.5631	(0.0109)	0.5502	(0.0069)	0.5443	(0.0069)
F1-score	**0.167**	(0.0335)	**0.1896**	(0.0194)	**0.1863**	(0.0184)
FPR	0.0664	(0.0354)	0.0871	(0.0152)	0.089	(0.0140)
FNR	0.9003	(0.0261)	0.8841	(0.0142)	0.8864	(0.0132)
PPV	0.5618	(0.0617)	0.5279	(0.0309)	0.5221	(0.0306)
G-mean	0.3022	(0.0333)	0.3246	(0.0181)	0.3212	(0.0172)
BDD						
v	5		10		20	
Accuracy	0.528	(0.0145)	0.5249	(0.0226)	0.5045	(0.0284)
F1-score	**0.6101**	(0.0385)	**0.6078**	(0.0494)	**0.5864**	(0.0448)
FPR	0.6521	(0.0980)	0.6672	(0.0959)	0.6742	(0.0576)
FNR	0.3283	(0.0885)	0.3146	(0.1043)	0.3437	(0.0845)
PPV	0.5637	(0.0131)	0.551	(0.0147)	0.5322	(0.0196)
G-mean	0.4749	(0.0284)	0.4674	(0.0354)	0.4579	(0.0242)

Performance chart: BSVDD, BSVDD-M

BSVDD						
v	5		10		20	
Accuracy	0.447	(0.0071)	0.4576	(0.0067)	0.4617	(0.0070)
F1-score	**0.6161**	(0.0065)	**0.6262**	(0.0063)	**0.6305**	(0.0065)
FPR	0.994	(0.0029)	0.9942	(0.0026)	0.9958	(0.0018)
FNR	0.0016	(0.0008)	0.0017	(0.0008)	0.0012	(0.0005)
PPV	0.4455	(0.0067)	0.4562	(0.0066)	0.4607	(0.0069)
G-mean	0.0751	(0.0186)	0.0742	(0.0171)	0.0631	(0.0143)
BSVDD-M						
v	5		10		20	
Accuracy	0.447	(0.0102)	0.4578	(0.0096)	0.4621	(0.0091)
F1-score	**0.614**	(0.0076)	**0.6242**	(0.0077)	**0.6285**	(0.0082)
FPR	0.9872	(0.0390)	0.9868	(0.0396)	0.9872	(0.0385)
FNR	0.0101	(0.0312)	0.0103	(0.0312)	0.0104	(0.0317)
PPV	0.4452	(0.0069)	0.4561	(0.0068)	0.4606	(0.0070)
G-mean	0.0509	(0.0998)	0.0562	(0.0991)	0.0531	(0.0987)

Outcome

We can see the following outcomes in the toy example.

- SVDD's F1-score is relatively much lower than the other methods.
- In the BDD, F1-score $\downarrow$ as $v \uparrow$
- In the BSVDD and BSVDD-M, F1-score $\uparrow$ as $v \uparrow$

Statistical reinforcement of the BSVDD and BSVDD-M may enhance prediction robustness in the high variance of meaningful β s compared to the BDD.

The next steps

- Set $p > 50$
- Make the imbalance ratio larger than the current one$(= 1/10)$
- Implement non-linear decision boundaries
- etc

SVDD: Apply KKT to Lagrangean - 1

Lagrangian form of the objective function

$$\min_{R,\mathbf{a},\xi_i^{(0)}} \ \left(\max_{\alpha_i,\beta_i}\right) \ L,$$

where $L = R^2 + C\sum_{i=1}^{n_0}\xi_i^{(0)} - \sum_{i=1}^{n_0}\alpha_i(R^2 + \xi_i^{(0)} - ||\phi(x_i^{(0)}) - \mathbf{a}||^2) - \sum_{i=1}^{n_0}\beta_i\xi_i^{(0)}$

for $\alpha_i \geq 0, \ \ \beta_i \geq 0$

By the Karush-Kuhn-Tucker(KKT) condition,

Primal feasibility: $\alpha_i(R^2 + \xi_i^{(0)} - ||\phi(x_i^{(0)}) - \mathbf{a}||^2) \geq 0$ (P1)

$$\frac{\partial L}{\partial R} = 2R - 2R\sum_{i=1}^{n_0}\alpha_i = 0 \ \rightarrow \ \sum_{i=1}^{n_0}\alpha_i = 1$$

$$\frac{\partial L}{\partial \mathbf{a}} = 2\sum_{i=1}^{n_0}\alpha_i\phi(x_i^{(0)}) - 2\mathbf{a}\sum_{i=1}^{n_0}\alpha_i = 0 \ \rightarrow \ \mathbf{a} = \sum_{i=1}^{n_0}\alpha_i\phi(x_i^{(0)}) : \ \textbf{Center} \quad \text{(D2)}$$

$$\frac{\partial L}{\partial \xi_i^{(0)}} = C - \alpha_i - \beta_i = 0, \ \ \forall i = 0, \cdots, n_0 \quad \text{(D3)}$$

Complementary Slackness:

$$\alpha_i(R^2 + \xi_i^{(0)} - ||\phi(x_i^{(0)}) - \mathbf{a}||^2) = 0 \quad \text{(C1)}, \qquad \beta_i\xi_i^{(0)} = 0 \quad \text{(C2)}$$

SVDD: Apply KKT to Lagrangean - 2

From the KKT, $\sum_{i=1}^{n_0} \alpha_i = 1$, $\mathbf{a} = \sum_{i=1}^{n_0} \alpha_i \phi(x_i^{(0)})$, $C - \alpha_i - \beta_i = 0$, $\forall i$

$$
\begin{aligned}
L =& R^2 + C \sum_{i=1}^{n_0} \xi_i^{(0)} - \sum_{i=1}^{n_0} \alpha_i \left(R^2 + \xi_i^{(0)} - \phi(x_i^{(0)}) \cdot \phi(x_i^{(0)}) + 2\mathbf{a} \cdot \phi(x_i^{(0)}) - \mathbf{a} \cdot \mathbf{a}\right) \\
& - \sum_{i=1}^{n_0} \beta_i \xi_i^{(0)} \\
=& \cancelto{0}{R^2 - R^2 \sum_{i=1}^{n_0} \alpha_i} + \sum_{i=1}^{n_0} \xi_i^{(0)} \cancelto{0}{(C - \alpha_i - \beta_i)} \\
& + \sum_{i=1}^{n_0} \alpha_i \phi(x_i^{(0)}) \cdot \phi(x_i^{(0)}) - 2\mathbf{a} \cdot \sum_{i=1}^{n_0} \alpha_i \phi(x_i^{(0)}) + \cancelto{1}{\sum_{i=1}^{n_0} \alpha_i} (\mathbf{a} \cdot \mathbf{a}) \qquad \because \text{KKT} \\
=& \sum_{i=1}^{n_0} \alpha_i \phi(x_i^{(0)}) \cdot \phi(x_i^{(0)}) - 2\Big(\sum_{j=1}^{n_0} \alpha_j \phi(x_j^{(0)})\Big) \cdot \sum_{i=1}^{n_0} \alpha_i \phi(x_i^{(0)}) \\
& + \Big(\sum_{i=1}^{n_0} \alpha_i \phi(x_i^{(0)})\Big) \cdot \Big(\sum_{j=1}^{n_0} \alpha_j \phi(x_j^{(0)})\Big) \qquad \because \mathbf{a} = \sum_{i=1}^{n_0} \alpha_i \phi(x_i^{(0)})
\end{aligned}
$$

SVDD: Apply KKT to Lagrangean - 3

(cont'd)

$$= \sum_{i=1}^{n_0} \alpha_i K_{i,i} - \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} \alpha_i \alpha_j K_{i,j} \text{ for } 0 \le \alpha_i \le C$$

where C is some constant, $K_{i,j} = < \phi(x_i^{(0)}), \phi(x_j^{(0)}) >$

has only α related terms

$\therefore \underset{R,\mathbf{a},\xi_i^{(0)}}{min} \; \underset{\alpha_i,\beta_i}{(max)} \; L$ collapses into $\underset{\alpha_i}{max} \; L, \;\; \forall i = 0, \cdots, n_0$

BDD: Posterior calculation - 1

$$logP(\boldsymbol{\alpha_0}|\Phi(\boldsymbol{x_0})) \propto logP(\Phi(\boldsymbol{x_0})|\boldsymbol{\alpha_0})P(\boldsymbol{\alpha_0})$$

where $\boldsymbol{\alpha_0} = (\alpha_1 \cdots \alpha_{n_0})^T, \ \Phi(\boldsymbol{x_0}) = (\phi(x_1^{(0)}) \ \cdots \ \phi(x_{n_0}^{(0)}))^T$

$$= -\frac{1}{2}\left[\sum_{i=1}^{n_0} \cancel{K_{i,i}} - 2\sum_{j=1}^{n_0}\alpha_j K_{i,j} + \sum_{i=1}^{n_0}\sum_{j=1}^{n_0}\alpha_i\alpha_j K_{i,j}\} \right.$$

$$\left. + \boldsymbol{\alpha_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha_0} - 2\boldsymbol{\alpha_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{m} + \cancel{\boldsymbol{m}^T\boldsymbol{\Sigma}^{-1}}\boldsymbol{m}\right]$$

$$\propto \sum_{i=1}^{n_0}\sum_{j=1}^{n_0}\alpha_j K_{i,j} - \frac{n_0}{2}\sum_{i=1}^{n_0}\sum_{j=1}^{n_0}\alpha_i\alpha_j K_{i,j}$$

$$-\frac{1}{2}\boldsymbol{\alpha_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha_0} + \boldsymbol{\alpha_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{m}$$

$$= \boldsymbol{\alpha_0}^T D_0 \boldsymbol{1_{n_0}} - \frac{n_0}{2}\boldsymbol{\alpha_0}^T K_0 \boldsymbol{\alpha_0} - \frac{1}{2}\boldsymbol{\alpha_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha_0} + \boldsymbol{\alpha_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{m}$$

where $D_{0,(i,i)} = \sum_{j=1}^{n_0} K_{i,j}, \ \boldsymbol{1_{n_0}} = (1 \ \cdots \ 1)^T \in R^{n_0},$

$$K_{0,(i,j)} = K_{i,j}, \ \forall i,j = 1,\cdots,n_0$$

BDD: Posterior calculation - 2

(cont'd)

$$
\begin{aligned}
logP(\boldsymbol{\alpha_0}|\Phi(\boldsymbol{x_0})) &\propto 2\boldsymbol{\alpha_0}^T D_0 \mathbf{1}_{\boldsymbol{n_0}} - n_0 \boldsymbol{\alpha_0}^T K_0 \boldsymbol{\alpha_0} - \boldsymbol{\alpha_0}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha_0} + 2\boldsymbol{\alpha_0}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{m} \\
&= -\boldsymbol{\alpha_0}^T (n_0 K_0 + \boldsymbol{\Sigma}^{-1}) \boldsymbol{\alpha_0} + 2\boldsymbol{\alpha_0}^T (D_0 \mathbf{1}_{\boldsymbol{n_0}} + \boldsymbol{\Sigma}^{-1} \boldsymbol{m})
\end{aligned}
$$

where $D_{0,(i,i)} = \sum_{j=1}^{n_0} K_{i,j}, \;\; K_{0,(i,j)} =< \phi(x_i^{(0)}), \phi(x_j^{(0)}) >$

&

$$
\begin{aligned}
\underset{\boldsymbol{\alpha_0}}{argmax}\; P(\boldsymbol{\alpha_0}|\Phi(\boldsymbol{x_0})) &= \underset{\boldsymbol{\alpha_0}}{argmax}\; log\; P(\boldsymbol{\alpha_0}|\Phi(\boldsymbol{x_0})) \\
&= \underset{\boldsymbol{\alpha_0}}{argmin}\; -log\; P(\boldsymbol{\alpha_0}|\Phi(\boldsymbol{x_0}))
\end{aligned}
$$

$$\downarrow$$

MAP(Maximum a posteriori) estimator of $\boldsymbol{\alpha_0}$

$$
\hat{\boldsymbol{\alpha_0}} = \underset{\boldsymbol{\alpha_0}}{argmin}\; \boldsymbol{\alpha_0}^T (n_0 K_0 + \boldsymbol{\Sigma}^{-1}) \boldsymbol{\alpha_0} - 2\boldsymbol{\alpha_0}^T (D_0 \mathbf{1}_{\boldsymbol{n_0}} + \boldsymbol{\Sigma}^{-1} \boldsymbol{m})
$$

BSVDD: Posterior Calculation

$$
\begin{aligned}
log\ P(\boldsymbol{\beta}|\Phi(\boldsymbol{x_0})) &\propto log\ P(\Phi(\boldsymbol{x_0})|\boldsymbol{\beta})P(\boldsymbol{\beta}) \\
&= -\frac{1}{2}\Bigg[\sum_{i=1}^{n_0}\{\cancel{K_{i,i}} - 2\sum_{j=1}^{n_0}\alpha_j K_{i,j} + \sum_{i=1}^{n_0}\sum_{j=1}^{n_0}\alpha_i\alpha_j K_{i,j}\} \\
&\quad + (\boldsymbol{\beta}-\boldsymbol{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\Bigg] \\
&\propto -\frac{1}{2}\Bigg[-2\boldsymbol{\alpha_0}^T D_0 \mathbf{1}_{\boldsymbol{n_0}} + n_0\boldsymbol{\alpha_0}^T K_0\boldsymbol{\alpha_0} \\
&\quad + (\boldsymbol{\beta}-\boldsymbol{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\Bigg] \\
&\quad \text{where} \quad \boldsymbol{\alpha_0} = (\alpha_1,\ldots,\alpha_{n_0})^T = \left(\frac{e^{\beta_1}}{\sum_{i=1}^{n} e^{\beta_i}},\ldots,\frac{e^{\beta_{n_0}}}{\sum_{i=1}^{n_0} e^{\beta_i}}\right)^T
\end{aligned}
$$

See the next page for the matrices K_0 and D_0

BSVDD: Matrices for the Posterior

$$K_0 = \begin{pmatrix} K_{1,1} & K_{1,2} & \cdots & K_{1,n_0} \\ K_{2,1} & K_{2,2} & \cdots & K_{2,n_0} \\ \vdots & \vdots & \ddots & \vdots \\ K_{n_0,1} & K_{n_0,2} & \cdots & K_{n_0,n_0} \end{pmatrix},$$

$$D_0 = \begin{pmatrix} \sum_{j=1}^{n_0} K_{1,j} & 0 & \cdots & 0 \\ 0 & \sum_{j=1}^{n_0} K_{2,j} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{j=1}^{n_0} K_{n_0,j} \end{pmatrix}$$

where $K_{i,j} =< \phi(x_i^{(0)}), \phi(x_j^{(0)}) >, \ \forall i = 1, \cdots, n_0; \ \forall j = 1, \cdots, n_0$

BSVDD-M: Get the **a** with minor class from the KKT - 1

The Lagrangian form:

$$\min_{R,\mathbf{a},\xi_i^{(0)},\xi_l^{(1)}} L \leftrightarrow \max_{\alpha_i^{(0)},\alpha_l^{(1)},\beta_i^{(0)},\beta_l^{(1)}} L$$

where $L = R^2 + C_1 \sum_{i=1}^{n_0} \xi_i^{(0)} + C_2 \sum_{i=1}^{n_1} \xi_l^{(1)} - \sum_{i=1}^{n_0} \alpha_i^{(0)} (R^2 + \xi_i^{(0)} - ||\phi(x_i^{(0)}) - \mathbf{a}||^2)$

$$- \sum_{l=1}^{n_1} \alpha_l^{(1)} (||\phi(x_l^{(1)}) - \mathbf{a}||^2 - R^2 + \xi_l^{(1)}) - \sum_{i=1}^{n_0} \beta_i^{(0)} \xi_i^{(0)} - \sum_{i=1}^{n_1} \beta_l^{(1)} \xi_l^{(1)}$$

with $\alpha_i^{(0)} \geq 0,\ \alpha_l^{(1)} \geq 0,\ \beta_i^{(0)} \geq 0,\ \beta_l^{(1)} \geq 0$

where $C_1,\ C_2$ are some constants

BSVDD-M: Get the **a** with minor class from the KKT - 2

KKT condition for SVDD with the minor class

$$
\begin{aligned}
&\frac{\partial L}{\partial R} = 2R - 2R\sum_{i=1}^{n_0}\alpha_i^{(0)} + 2R\sum_{l=1}^{n_1}\alpha_l^{(1)} = 0 \\
&\Rightarrow \ \sum_{i=1}^{n_0}\alpha_i^{(0)} - \sum_{l=1}^{n_1}\alpha_l^{(1)} = 1 \\
&\frac{\partial L}{\partial \mathbf{a}} = -\sum_{i=1}^{n_0}\alpha_i^{(0)}\{2\phi(x_i^{(0)}) - 2\mathbf{a}\} - \sum_{l=1}^{n_1}\alpha_l^{(1)}\{-2\phi(x_l^{(1)}) + 2\mathbf{a}\} = 0 \\
&\Rightarrow \ \mathbf{a} = \sum_{i=1}^{n_0}\alpha_i^{(0)}\phi(x_i^{(0)}) - \sum_{l=1}^{n_1}\alpha_l^{(1)}\phi(x_l^{(1)}) : \textbf{Center} \\
&\frac{\partial L}{\partial \xi_i^{(0)}} = \sum_{i=1}^{n_0}(C_1 - \alpha_i^{(0)} - \beta_i^{(0)}) \ \Rightarrow \ C_1 - \alpha_i^{(0)} - \beta_i^{(0)} = 0, \ \forall i \\
&\frac{\partial L}{\partial \xi_l^{(1)}} = \sum_{l=1}^{n_1}(C_2 - \alpha_l^{(1)} - \beta_l^{(1)}) \ \Rightarrow \ C_2 - \alpha_l^{(1)} - \beta_l^{(1)} = 0, \ \forall l
\end{aligned}
$$

BSVDD-M: Posterior Calculation - 1

$$log\ P(\boldsymbol{\tau}, \boldsymbol{\alpha_1}|\Phi(\boldsymbol{x})) \propto log\ P(\Phi(\boldsymbol{x})|\boldsymbol{\alpha_0}(\boldsymbol{\alpha_1}, \boldsymbol{\tau}))P(\boldsymbol{\tau}|\boldsymbol{\alpha_1})P(\boldsymbol{\alpha_1})$$

where $\Phi(\boldsymbol{x}) = (\phi(x_1), \cdots, \phi(x_n)),\ \ n = n_0 + n_1$

$$= -\frac{1}{2}\sum_{i=1}^{n_0}\left|\left|\phi(x_i^{(0)}) - \left\{\sum_{j=1}^{n_0}\alpha_j^{(0)}\phi(x_j^{(0)}) - \sum_{l=1}^{n_1}\alpha_l^{(1)}\phi(x_l^{(1)})\right\}\right|\right|^2$$

$$- \frac{1}{2}(\boldsymbol{\tau} - \boldsymbol{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\tau} - \boldsymbol{m})$$

by taking $\alpha_l^{(1)} \sim Beta(1,1) \overset{d}{=} U(0,1),\ \ \forall l = 1, \ldots, n_1$

$$= -\frac{1}{2}\sum_{i=1}^{n_0}\Bigg[\cancel{K_{i,i}} - 2\phi(x_i^{(0)})^T\left\{\sum_{j=1}^{n_0}\alpha_j^{(0)}\phi(x_j^{(0)}) - \sum_{l=1}^{n_1}\alpha_l^{(1)}\phi(x_l^{(1)})\right\}$$

$$+ \sum_{i=1}^{n_0}\sum_{j=1}^{n_0}\alpha_i^{(0)}\alpha_j^{(0)}K_{i,j} - 2\sum_{i=1}^{n_0}\sum_{l=1}^{n_1}\alpha_i^{(0)}\alpha_l^{(1)}K_{i,l}$$

$$+ \sum_{l=1}^{n_1}\sum_{m=1}^{n_1}\alpha_l^{(1)}\alpha_m^{(1)}K_{l,m}\Bigg] - \frac{1}{2}(\boldsymbol{\tau} - \boldsymbol{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\tau} - \boldsymbol{m})$$

BSVDD-M: Posterior Calculation - 2

(cont'd)

$$\propto -\frac{1}{2}\{-2\boldsymbol{\alpha_0}^T D_0 \mathbf{1}_{\boldsymbol{n_0}} + n_0 \boldsymbol{\alpha_0}^T K_0 \boldsymbol{\alpha_0} + n_0 \boldsymbol{\alpha_1}^T K_1 \boldsymbol{\alpha_1} + 2\boldsymbol{\alpha_1}^T D_{01} \mathbf{1}_{\boldsymbol{n_1}} - 2n_0 \boldsymbol{\alpha_1}^T K_{01}^T \boldsymbol{\alpha_0}\} - \frac{1}{2}(\boldsymbol{\tau} - \boldsymbol{m})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\tau} - \boldsymbol{m})$$

$$\text{with } \boldsymbol{\alpha_0} = \left(\frac{e^{\tau_1}\left(1 + \sum_{l=1}^{n_1} \alpha_l^{(1)}\right)}{\sum_{i=1}^{n_0} e^{\tau_i}} \cdots \frac{e^{\tau_{n_0}}\left(1 + \sum_{l=1}^{n_1} \alpha_l^{(1)}\right)}{\sum_{i=1}^{n_0} e^{\tau_i}} \right)^T$$

See the next page for the matrices K_0, K_1, K_{01}, D_0, D_{01}.

BSVDD-M: Matrices for the Posterior - 1

$$K_0 = \begin{pmatrix} K_{1,1} & K_{1,2} & \cdots & K_{1,n_0} \\ K_{2,1} & K_{2,2} & \cdots & K_{2,n_0} \\ \vdots & \vdots & \ddots & \vdots \\ K_{n_0,1} & K_{n_0,2} & \cdots & K_{n_0,n_0} \end{pmatrix} \text{ consists of only } \phi(x_i^{(0)})$$

$$K_1 = \begin{pmatrix} K_{1,1} & K_{1,2} & \cdots & K_{1,n_1} \\ K_{2,1} & K_{2,2} & \cdots & K_{2,n_1} \\ \vdots & \vdots & \ddots & \vdots \\ K_{n_1,1} & K_{n_1,2} & \cdots & K_{n_1,n_1} \end{pmatrix} \text{ consists of only } \phi(x_l^{(1)})$$

$$K_{01} = \begin{pmatrix} K_{1,1} & K_{1,2} & \cdots & K_{1,n_1} \\ K_{2,1} & K_{2,2} & \cdots & K_{2,n_1} \\ \vdots & \vdots & \ddots & \vdots \\ K_{n_0,1} & K_{n_0,2} & \cdots & K_{n_0,n_1} \end{pmatrix} \text{ consists of } \phi(x_i^{(0)}) \text{ and } \phi(x_l^{(1)})$$

BSVDD-M: Matrices for the Posterior - 2

$$D_0 = \begin{pmatrix} \sum_{j=1}^{n_0} K_{1,j} & 0 & \cdots & 0 \\ 0 & \sum_{j=1}^{n_0} K_{2,j} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{j=1}^{n_0} K_{n_0,j} \end{pmatrix} \in R^{n_0 \times n_0}$$

$$D_{01} = \begin{pmatrix} \sum_{i=1}^{n_0} K_{i,l} & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^{n_0} K_{i,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{i=1}^{n_0} K_{i,n_1} \end{pmatrix} \in R^{n_1 \times n_1}$$

where $K_{i,j} =< \phi(x_i^{(0)}), \phi(x_j^{(1)}) >, \ \forall i = 1, \ldots, n_0; \ \forall j = 1, \ldots, n_1$

Data simulation explained

The toy dataset has structure below

$\boldsymbol{x_0} = (x_1^{(0)} \cdots x_{n_0}^{(0)})^T, \ \boldsymbol{x_1} = (x_1^{(1)} \cdots x_{n_1}^{(1)})^T, \ n_0 + n_1 = n.$

Set $\boldsymbol{x_0} \sim N(\boldsymbol{\mu_0}, \Sigma), \ \boldsymbol{x_1} \sim N(\boldsymbol{\mu_1}, \Sigma),$

where $\boldsymbol{\mu_0} = (7 \ 7 \ 7 \ 7 \ 0 \cdots 0)^T \in R^{50},$

$\boldsymbol{\mu_1} = (10 \ 10 \ 10 \ 10 \ 0 \cdots 0)^T \in R^{50},$

$\Sigma = diag(2, 2, 2, 2, 1, \cdots, 1) \in R^{50 \times 50}$

& Let $\eta = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots \beta_{50} x_{i,50}, \ \forall i = 1, \cdots, n$

Then only $\beta_0 = -31, \ \beta_1, \ \beta_2, \ \beta_3, \ \beta_4$ are meaningful.

Set $\beta_1 \sim N(0.96, v), \ \beta_2 \sim N(0.8, v), \ \beta_3 \sim N(1, v), \ \beta_4 \sim N(0.8, v)$

where $v =$ **5,10,20**

Set $P(x_i) := \dfrac{1}{1 + e^{-\eta}}, \ y_i \sim Ber(P(x_i))$

$\rightarrow \ y_i = 0$ *(major class)* or 1 *(minor class)*, $\forall i = 1, \cdots, n (= n_0 + n_1)$

References

Hearst et al. (1998).
Support vector machines.
IEEE Intelligent Systems and their applications, 13(4), 18-28..

Tax & Duin. (2004).
Support vector data description.
Machine learning, 54, 45-66.

Ghasemi, A., Rabiee, H. R., Manzuri, M. T., & Rohban, M. H. (2012).
A bayesian approach to the data description problem.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26, No. 1, pp. 907-913.

오정민. (2023).
이상치 탐지를 위한 베이지안 Support vector data description.
(국내석사학위논문).

배희진. (2024).
Minor class를 이용한 베이지안 Support vector data description.
(국내석사학위논문).

Robert, C. P., Casella, G., Robert, C. P., & Casella, G. (2004).
The metropolis—hastings algorithm.
Monte Carlo statistical methods, 267-320.

Turkoz, M., & Kim, S., (2022).
Multi-class Bayesian support vector data description with anomalies.
Annals of Operations Research, 1-26.