Incomplete data representation for SVM classification

Łukasz Struski  Marek Śmieja  Jacek Tabor

Jagiellonian University,  
Faculty of Mathematics and Computer Science,  
Lojasiewicza 6, 30-348 Kraków, Poland

e-mail: lukasz.struski@uj.edu.pl,  
marek.smieja@uj.edu.pl,  
jacek.tabor@uj.edu.pl.

December 6, 2016

Abstract

In this paper we propose two ways of incomplete data representation. The first one is a generalization of a flag representation, where a vector with missing attributes is filled with some values and joined with flag vectors indicating missing components. Our generalization uses pointed affine subspaces, which in addition to flag representation allows to perform various affine transformations of data, as whitening or dimensionality reduction. We show how to embed such affine subspaces into a vector space and how to define a proper scalar product. In the second approach, we represent missing data points by degenerated Gaussian densities, which additionally model the uncertainty connected with missing features. This representation allows to construct an analogue of RBF kernel on incomplete data space.

1 Introduction

Incomplete data analysis is an important part of data engineering and machine learning, since it appears in many practical problems. In medical diagnosis, a doctor may be unable to complete the patient examination due to the deterioration of health status or lack of patient’s compliance [2]; in object detection, the system has to recognize the shape from low resolution or corrupted images [1]; in chemistry, the complete analysis of compounds requires high financial costs [16]. In consequence, the understanding and the appropriate representation of such data is of great practical importance.

A missing data is typically viewed as a pair \( (x, J) \), where \( x \in \mathbb{R}^N \) is a vector with missing components \( J \subset \{1, \ldots, N\} \). In the most straightforward approach, one can fill the missing attributes with some statistic, e.g.
mean, taken from existing data, and add a flag indicating which components were missing. More precisely, we supply \( x \) with a binary vector \( 1_J \), in which 1 denotes absent feature while 0 means the present one, and perform the embedding \((x, J) \rightarrow (x, 1_J)\) of missing points into a vector space of extended complete data. Although this embedding allows for a direct application of typical classification methods, it is non-trivial how to perform a classification preprocessing, such as whitening or dimensionality reduction, on the flag vector.

In our first approach, we view the incomplete data as pointed affine subspaces, i.e. the subspace with a distinguished point called basepoint. In consequence, a missing data point \((x, J)\) is represented as

\[
x + \text{span}(e_j)_{j \in J},
\]

where \((e_j)_{j=1}^N\) denotes the canonical base of \(\mathbb{R}^N\) and \(x\) is a selected basepoint. Such a definition allows us to efficiently extend linear and affine operations from the standard points to missing ones, by taking the image of the subspace and the point. For example, a linear mapping \(F: w \rightarrow Aw + b\), can be extended to the case of pointed subspace \(x + V\) by

\[
F(x + V) = F(x) + AV.
\]

The above representation allows to embed missing data into a vector space by identifying a linear subspace \(V\) with an orthogonal projection \(p_V: \mathbb{R}^N \rightarrow V\) and apply typical classification methods on such embeddings.

In the second part, we assume that a (Gaussian) probability measure is given\(^1\) on a data space \(\mathbb{R}^N\). To use the information contained in missing features we can calculate conditional density on the subspace representing missing components. In other words, we model the uncertainty connected with missing features with a use of data distribution. We show how to construct an analogue of classical RBF kernel for incomplete data represented

\(^1\)Under MAR it can be reliably estimated by EM algorithm.
as probability measures and combine this representation together with SVM classifier.

\section{Related works}

The most common approach to learning from incomplete data is known as deterministic imputation \cite{12}. In this two-step procedure, the missing features are filled first, and only then a standard classifier is applied to the complete data \cite{11}. Although the imputation-based techniques are easy to use for practitioners, they lead to the loss of information which features were missing and do not take into account the reasons of missingness. To preserve the information of missing attributes, one can use an additional vector of binary flags, which was discussed in the introduction.

The second popular group of methods aims at building a probabilistic model of incomplete data which maximizes the likelihood by applying the EM algorithm \cite{7,13}. This allows to generate the most probable values from obtained probability distribution for missing attributes (random imputation) \cite{12} or to learn a decision function directly based on the distributional model. The second option was already investigated in the case of logistic regression \cite{18}, kernel methods \cite{15,17} or by using second order cone programming \cite{14}. One can also estimate the parameters of the probability model and the classifier jointly, which was considered in \cite{6,10}. This techniques work very well when the missing data is conditionally independent of the unobserved features given the observations, but there is no guarantee to get a reasonable estimation in more general missing not at random case.

There is also a group of methods, which does not make any assumptions about the missing data model and makes a prediction from incomplete data directly. In \cite{3} a modified SVM classifier is trained by scaling the margin according to observed features only. The alternative approaches to learning a linear classifier, which avoid features deletion or imputation, are presented in \cite{5,8}. Finally, in \cite{9} the embedding mapping of feature-value pairs is constructed together with a classification objective function.

In our contribution, we generalize the imputation-based techniques in such a way to preserve the information of missing features. To select a basepoint we propose to choose the most probable point form a subspace identifying a missing data point, however other imputation methods can be used as well. Constructed representation allows to apply various affine data transformations preserving classical scalar product before applying typical classification methods.
3 Pointed subspace approach

In this section, we introduce a pointed subspace subspace approach to incomplete data representation. First, we define a generalized missing data point, which allows to perform affine transformation of incomplete data. Then, we show how to embed generalized missing data into a vector space and discuss the selection of basepoint. Finally, we define a scalar product on the embedding space.

3.1 Incomplete data as pointed affine subspaces

Incomplete data $X$ can be understood as a sequence of pairs $(x_i, J_i)$, where $x_i \in \mathbb{R}^N$ and $J_i \subset \{1, \ldots, N\}$ indicates missing coordinates of $x_i$. Therefore, we can associate a missing data point $(x, J)$ with an affine subspace $x + \text{span}(e_j)_{j \in J}$, where $(e_j)_j$ is the canonical base of $\mathbb{R}^N$. Let us observe that $x + \text{span}(e_j)_{j \in J}$ is a set of all $N$-dimensional vectors which coincide with $x$ on the coordinates different from $J$.

We want to be able to transform incomplete data by affine mappings, e.g. whitening or dimensionality reduction. For this purpose, we generalize the above representation to arbitrary affine subspaces, or more precisely pointed affine subspaces, which do not have to be generated by canonical bases.

**Definition 3.1.** A generalized missing data point is defined as a pointed affine subspace $S_x = (S, x)$, where $S$ is an affine subspace of $\mathbb{R}^N$ and $x \in S$ is a basepoint.

A basepoint can be selected by filling missing attributes with a use of imputation methods, which will be discussed in the next subsection.

First, we show that the above definition is useful for defining linear mappings on incomplete data. Let $S_x = x + V$ be a generalized missing data point and let $f : \mathbb{R}^N \ni w \to Aw + b$ be an affine map. We can transform a generalized missing data point $x + V$ into another missing data point by the formula:

$$f(x + V) = \{Aw + b : w \in x + V\}.$$

The basepoint $x$ is mapped into $Ax + b$, while the linear part of $f(x + V)$ is given by

$$f(x + V) - f(x) = AV.$$

Consequently, we arrive at the definition:
Definition 3.2. For a generalized missing data point $S_x = x + V$ and an affine mapping $f : w \to Aw + b$ we put:

$$f(x + V) = (Ax + b) + AV,$$

where $Ax + b$ is a basepoint and $AV$ is a linear subspace.

One can easily compute and represent $AV$, if the orthonormal base $v_1, \ldots, v_n$ of $V$ is given, namely we simply orthonormalize the sequence $Av_1, \ldots, Av_n$.

3.2 Embedding of generalized missing data

The above representation is useful for understanding and performing affine transformations of incomplete data, such as whitening, dimensionality reduction or incorporating affine constraints to data. Nevertheless, typical machine learning methods require vectors or a kind of kernel (or similarity) matrix as the input. We show how to embed generalized missing data into a vector space.

A generalized missing data point $S_x = x + V$ consists of a basepoint $x \in S$ which is an element of vector space and a linear subspace $V$. To represent a subspace $V$, we propose to use a matrix of orthogonal projection $p_v$ onto $V$. To get an exact form of $p_v$, let us assume that $(v_j)_{j \in J}$ is an orthonormal base of $V$. Then, the projection of $y \in \mathbb{R}^N$ can be calculated by

$$p_v(y) = \sum_{j \in J} \langle y, v_j \rangle v_j = \sum_{j \in J} v_j v_j^T y = (\sum_{j \in J} v_j v_j^T) y,$$

which implies that

$$p_v = \sum_{j \in J} v_j v_j^T.$$

The selection of basepoint relies on filling missing attributes with some concrete values, which is commonly known as imputation. In our setting, by the imputation we denote a function $\Phi : X \to \mathbb{R}^N$ such that

$$\Phi(S_x) \in S,$$

for a generalized missing data $S_x$.

In the case of classical incomplete data, missing attributes are often filled with a mean or a median calculated from existing values for a given attribute. However, these imputations cannot be easily defined in a general case, because the linear part of generalized missing data point might be an arbitrary linear subspace (not necessarily a subspace generated by a subset of canonical base). Let us observe that another popular imputation method, which fills the missing coordinates with zeros can be defined for generalized incomplete data. This is performed by selecting a basepoint of an incomplete
data point $S_x = x + V$ as the orthogonal projection of missing data $x$ onto the subspace orthogonal to $V$, i.e.:

$$x_{V^\perp} = x - p_V(x) = x - \sum_{j \in J} \langle x_j, v_j \rangle v_j,$$

where $(v_j)_{j \in J}$ is an arbitrary orthonormal base of $V$. If $V$ is represented by canonical base then this is equivalent to filling missing attributes with zeros.

We propose another technique for setting missing values, which extends zero imputation method. Let us assume that $(m, \Sigma)$ are the mean and covariance matrix estimated for incomplete dataset $X$. In this method, a basepoint of $S_x = x + V$ is selected as the orthogonal projection of $m$ onto $x + V$ with respect to the Mahalanobis scalar product parametrized by $\Sigma$, i.e.

$$x^{(m, \Sigma)}_V = x + p^{\Sigma}_V(m - x),$$

where $p^{\Sigma}_V$ denotes a projection matrix onto $V$ with respect Mahalanobis scalar product given by $\Sigma$. To obtain the values for $m$ and $\Sigma$ in practice, one can use existing attributes of incomplete data for the calculation of a sample mean and a covariance matrix. Alternatively, if data satisfy missing at random assumption, then the EM algorithm can be applied to estimate the probability model describing data [13]. We call this technique by the most probable point imputation and examine its performance in the experimental section.

Summarizing, our embedding is defined as follows:

**Definition 3.3.** A generalized missing data point $S_x = x + V$ is embedded in a vector space by

$$x + V \rightarrow (x, p_V) \in \mathbb{R}^N \times \mathbb{R}^{N \times N}.$$

**Example 3.1.** To illustrate the effect of missing data imputation and transformation, let us consider the whitening operation:

$$\text{Whitening}(x) = \Sigma^{-1/2}(x - m),$$

where $\Sigma$ is the covariance, and $m$ the mean of $X$. For a generalized missing data the above operation is defined by:

$$\text{Whitening}(x + V) = \Sigma^{-1/2}(x - m) + \Sigma^{-1/2}V.$$

In other words, we map a basepoint in a classical way and transform a subspace $V$ into a linear subspace $\Sigma^{-1/2}V$. The illustration is given in Figure 2.

**Example 3.2.** In the case of high dimensional data, we sometimes reduce a dimension of input data space by applying the Principle Component Analysis, which is defined by:

$$\text{PCA}(x) = W^T(x - m),$$
where \( m \) is a mean of a dataset and \( k \) columns of \( W \) are the leading eigenvectors of covariance matrix \( \Sigma \). This operation can be extended to the case of generalized missing data by:

\[
\text{PCA}(x + V) = W^T(x - m) + W^TV.
\]

An example of the above operation is illustrated in the Figure 3.

### 3.3 Scalar product for SVM

To apply most of classification methods it is necessary to define a scalar product on a data space. As a natural choice, one could sum the scalar products between basepoints and embedding matrices, i.e.

\[
\langle x + V, y + W \rangle = \langle x, y \rangle + \langle p_V, p_W \rangle. \tag{1}
\]

However, for a data space of dimension \( N \), we have \( \|p_V\|^2 = N \), which implies that the weight of projection can dominate the first part of (1) concerning basepoints. Consequently, we decided to introduce an additional parameter to allow reducing the importance of projection part:

**Definition 3.4.** Let \( D \in [0,1] \) be fixed. As a scalar product between two generalized missing data points we put:

\[
\langle x + V, y + W \rangle_D = \langle x, y \rangle + D\langle p_V, p_W \rangle. \tag{2}
\]
Figure 3: The image 3(a) with two missing pixels and its projection onto two principal components 3(b). Image was represented by the feature vectors consisting of 8x8 blocks. Missing pixels are identified by the pointed subspaces with basepoints chosen by zero imputation strategy.

Let us observe that the above parametric scalar product can be implemented by taking the embedding \( x + V \rightarrow (x, \sqrt{D} p_V) \) and then using formula for a scalar product.

Remark 3.1. The above scalar product strictly depends on the section of basepoints. However, since the generalized missing data point is defined for a fixed basepoint (if we choose different basepoint, we get different generalized missing data point) then (2) defines a proper scalar product in the space of generalized missing data. If we did not use pointed subspaces, but ordinary affine subspaces to represent missing data, then (2) would not provide a well defined operation.

The following proposition shows how to calculate a scalar product between matrices defining two orthogonal projections onto linear subspaces.

Proposition 3.1. Let us consider subspaces

\[ V = \text{span}(v_j : j \in J), W = \text{span}(w_j : j \in K). \]

where \( v_j \) and \( w_k \) are orthonormal sequences. If \( p_V, p_W \) denote orthogonal projections onto \( V, W \), respectively, then

\[ \langle p_V, p_W \rangle = \sum_{j \in J, k \in K} \langle v_j, w_k \rangle^2. \]

Proof. By the definition of orthogonal projections and the scalar product between matrices, we have

\[ \langle p_V, p_W \rangle = \sum_{j \in J, k \in K} \text{tr}((v_j v_j^T)(w_k w_k^T)). \]  

(3)

Making use of \( \text{tr}(AB) = \text{tr}(BA) \), we get

\[ \text{tr}((v_j v_j^T)(w_k w_k^T)) = \text{tr}(v_j^T w_k w_k^T v_j) = \text{tr}(v_j^T w_k^T v_j) = \langle v_j, w_k \rangle^2. \]
Finally,\[
\langle p_V, p_W \rangle = \sum_{j \in J, k \in K} (v_j, w_k)^2.
\]

Concluding, the scalar product between embedding of two generalized missing data points given by Definition 3.4 can be calculated as:
\[
\langle x + V, y + W \rangle_D = \langle x, y \rangle + D \sum_{i,j} (p_V)_{ij} (p_W)_{ij} = \langle x, y \rangle + D \sum_{j \in J, k \in K} (v_j, w_k)^2,
\]
where \((v_j)_{j \in J}, (w_k)_{k \in K}\) are orthonormal bases of \(V, W\), respectively. The last expression can be more numerically efficient if the dimension of the subspaces (the number of missing attributes) is much smaller than the dimension of the whole space.

Remark 3.2. One of typical representations of missing data \((x, J)\) relies on filling unknown attributes and supplying it with a binary flag vector \(1_J \in \mathbb{R}^N\), in which bit 1 denotes coordinate belonging to \(J\). This leads to the embedding of the missing data into a vector space given by
\[
(x, J) \rightarrow (x, 1_J) \in \mathbb{R}^N \times \mathbb{R}^N.
\]
Then, the scalar product of such embedding can be defined by
\[
\langle (x, 1_J), (y, 1_K) \rangle = \langle x, y \rangle + \langle 1_J, 1_K \rangle = \langle x, y \rangle + \text{card}(J \cap K). \tag{4}
\]

It is worth to noting that the formula (4) coincides with a scalar product defined for generalized missing data \((1)\) (for \(D = 1\)). Indeed, if \(V = \text{span}(e_j : j \in J)\) and \(W = \text{span}(e_k : k \in K)\), for \(J, K \subset \{1, \ldots, N\}\), then by Proposition 3.1 we have,
\[
\langle p_V, p_W \rangle = \sum_{j \in J, k \in K} (e_j, e_k)^2 = \sum_{l \in J \cap K} (e_l, e_l)^2 = \sum_{l \in J \cap K} 1 = \text{card}(J \cap K),
\]
which is exactly the RHS of (4).

Therefore, our approach generalizes and theoretically justifies the flag approach to missing data analysis. The importance of our construction lies in its generality, which in particular allows for performing typical affine transformations of data. In the case of flag representation, there is not obvious solution how to perform such mappings on flag vector.

4 Probabilistic approach

In previous section, we did not make any assumption about the data. Now, we will show how to represent incomplete data by probability measures and to construct the analogue of classical RBF kernel function, in the case we know a probability distribution on a data space.
4.1 Missing data as degenerated probability measures

Making use of the results from previous section, we assume that the set of
generalized missing data points is given. To be able to represent such data by
probability measures, let us further assume that we are given a probability
distribution on $\mathbb{R}^N$. For a simplicity and transparency, in this paper we
restrict our attention to the case of Gaussian densities $F = N(m, \Sigma)$. In
the missing at random (MAR) case, a density $F$ can be practically obtained
with a use of EM algorithm or Variational-Bayesian approach.

If we have a complete data point $x$ (with no missing coordinates), then it
always coincides with a Dirac measure $\delta_x$ (a measure that takes value $x$ with
probability 1), because there is no uncertainty connected with this example.
If we have an incomplete data point $S_x = x + V$ then the uncertainty is
connected with its missing part. To quantify it, we can condition a data
space density $F$ by an affine subspace $x + V$, which we denote by $F|_{x+V} =
N(m_V, \Sigma_V)$. This is a non-degenerated normal density in the space $\mathbb{R}^{\text{dim}(V)}$,
the subspace of $\mathbb{R}^N$. Since we work in $N$ dimensional space, we need to
extend this conditional density $F|_{x+V}$ back to the original space. In other
words, we define a degenerated density $N(m_V, \Sigma_V)$ from the conditional
density $N(m_V, \Sigma_V)$. This is a basic idea of probabilistic representation of
incomplete data.

To work with such data, we need to calculate the parameters $m_V, \Sigma_V$
from data space distribution. Before that, let us first introduce basic nota-
tions. Let us recall, that the standard scalar product in $L^2$ space is given
by
\[
\langle F, G \rangle = \int F(x)G(x)dx, \text{ for } F, G \in L^2, \text{ for } F, G \in L^2
\]
If we take Gaussian densities
\[
N(m, \Sigma)(x) = \frac{1}{(2\pi)^{D/2}\det^{1/2}\Sigma} \exp(-\frac{1}{2}\|x - m\|_\Sigma^2),
\]
where $\|w\|_\Sigma$ denotes the Mahalanobis norm given by $v^T\Sigma^{-1}v$, then the above
scalar product can be easily computed by [4 Eq.(6)]:
\[
\langle N(m_1, \Sigma_1), N(m_2, \Sigma_2) \rangle = N(m_1 - m_2, \Sigma_1 + \Sigma_2)(0), \quad (5)
\]
where $N(m_1, \Sigma_i)$ are non-degenerated Gaussians.

We also need the notion of convolution, which for densities $F, G \in L^2$ is
defined by
\[
(F * G)(y) = \int F(x - y)G(y)dx.
\]
If $F$ is a measure with mean $m_F$ and covariance $\Sigma_F$, then the convolution
$F * N(0, \gamma I)$, where $I$ is an identity matrix and $\gamma > 0$, is a measure with
mean $m_F$ and covariance $\Sigma_F + \rho I$. The convolution of normal densities is a normal density and,

$$N(m, \Sigma) * N(0, \gamma I) = N(m, \Sigma + \gamma I).$$

The above formula also holds for degenerated normal densities $N(m, \Sigma)$, that is we accept singular (not invertible) $\Sigma$.

We show how to calculate the conditional density and corresponding density in the original space from a data space distribution.

**Observation 4.1.** Assume that we have a density $N(m_V, \Sigma_V)$ in the sub-space $x + V$. Then the corresponding density in the original $\mathbb{R}^N$ space equals

$$N(x + vm_V, v\Sigma_V v^T),$$

where $v = [v_1, \ldots, v_n]$ is an orthonormal base of $V$.

**Proof.** If random vector $Z$ has mean $m_Z$ and covariance $\Sigma_Z$, then $\Phi(Z) = AZ + b$ has the mean $Am_Z + b$ and covariance $A\Sigma_Z A^T$. Therefore, we apply this fact to the map $\Phi : \mathbb{R}^n \ni \alpha = [\alpha_1, \ldots, \alpha_n]^T \rightarrow v\alpha + x \in \mathbb{R}^N$. □

Now we discuss the inverse problem:

**Observation 4.2.** Assume that we have a normal density $F = N(m, \Sigma)$ in $\mathbb{R}^N$, a subspace $x + V$, and an orthonormal base $v = [v_1, \ldots, v_n]$ of $V$. Then the conditional density $F|_{x+V}$ in the space $x + V$ in the base given by $v$ equals $N(m_V, \Sigma_V)$, where

$$\Sigma_V = (v^T\Sigma^{-1}v)^{-1} \quad \text{and} \quad m_V = \Sigma_V[v^T\Sigma^{-1}(m - x)].$$

**Proof.** Let us recall that the formula for normal density is given by

$$w \rightarrow Z \cdot \exp(-\frac{1}{2}(w - m)^T\Sigma^{-1}(w - m)),$$

where $Z$ is a normalization factor. Now, restricting the quadratic function $w \rightarrow (w - m)^T\Sigma^{-1}(w - m)$ to the space $x + V$ by putting $w = x + v\alpha$ we get

$$\alpha \rightarrow (x + v\alpha - m)^T\Sigma^{-1}(x + v\alpha - m)$$

$$= \alpha^T(v^T\Sigma^{-1}v)\alpha - 2[v^T\Sigma^{-1}(m - x)]^T\alpha + \text{const},$$

and by the canonical form of the quadratic function\footnote{Recall the formula $\alpha^T A \alpha + b^T \alpha + c$, for symmetric $A$, can be rewritten as $(\alpha - \alpha_0)^T A (\alpha - \alpha_0) + \text{const}$, for $\alpha_0 = -\frac{1}{2} A^{-1} b$.} we get

$$= (\alpha - m_V)^T \Sigma_V^{-1} (\alpha - m_V) + \text{const},$$

where

$$\Sigma_V = (v^T\Sigma^{-1}v)^{-1} \quad \text{and} \quad m_V = \Sigma_V[v^T\Sigma^{-1}(m - x)].$$

Thus the conditional density equals $N(m_V, \Sigma_V)$. □
Taking the above two observations together, we can calculate both densities from the original normal density defined on $\mathbb{R}^N$ and define the probabilistic representation of generalized missing data:

**Definition 4.1.** Let $F = N(m, \Sigma)$ be a density on $\mathbb{R}^N$ and let $S_x = x + V$ be a generalized missing data point, where $v = [v_1, \ldots, v_n]$ is an orthonormal base of $V$. We define the probabilistic representation of generalized missing data point $S_x = x + V$ as a (degenerated) Gaussian density $N(m^V, \Sigma^V)$, where

$$m^V = x + vm_V, \Sigma^V = v\Sigma_v v^T$$

and

$$\Sigma_V = (v^T\Sigma^{-1}v)^{-1} \text{ and } m_V = \Sigma_V[v^T\Sigma^{-1}(m - x)].$$

If $\text{dim}(V) = 0$ then $x$ is represented by a Dirac measure $\delta_x$.

### 4.2 Scalar product for probabilistic missing data

To define a scalar product on probabilistic representations of incomplete data, we will apply the reasoning used in classical RBF kernels. Let us recall that a type of RBF kernel can be constructed as follows. First, we map every point $x$ to Dirac measure $\delta_x$. Then we take the convolution $\delta_x \ast N(0, \gamma I) = N(x, \gamma I)$. Finally, we apply the standard scalar product $[5]$ in $L^2$ space.

To perform an analogue procedure in the case of probabilistic representations of incomplete data, let us calculate the formula for the convolution:

**Proposition 4.1.** Assume that we are given a density $F = N(m, \Sigma)$. Fix $\gamma > 0$ and affine subspace $x + V$. We assume that the base $v = [v_1, \ldots, v_k]$ is orthonormal in $V$ (important, as in the other case we would get different normalization constant for the singular Lebesgue measure on $x + V$). Put

$$\Sigma_V = (v^T\Sigma^{-1}v)^{-1} \text{ and } m_V = \Sigma_V[v^T\Sigma^{-1}(m - x)].$$

Next let

$$m^V = x + vm_V \text{ and } \Sigma^V = v\Sigma_v v^T.$$ 

Then

$$F|_{x+V} \ast N(0, \gamma I) = N(m^V, \Sigma^V + \gamma I).$$

**Proof.** The proof follows directly from Observation 4.1 and 4.2. \qed

We can now define the scalar product of generalized probabilistic missing data points:
Definition 4.2. Let \( \gamma > 0 \) be fixed. As a scalar product between two probabilistic representations of generalized missing data points \( S_x = x + V \) and \( S_y = y + V \) we put:

\[
\langle x + V, y + W \rangle_\gamma = \langle F|_{x+V} \ast N(0, \gamma I), F|_{y+W} \ast N(0, \gamma I) \rangle,
\]

where \( \gamma > 0 \) is fixed.

The following theorem gives exact formula for the above scalar product.

Theorem 4.1. Let \( F = N(m, \Sigma) \) be a density on \( \mathbb{R}^N \) and let \( \gamma > 0 \) be fixed. We assume that \( N(m^V, \Sigma^V) \) and \( N(m^W, \Sigma^W) \) are probabilistic representations of generalized missing data points \( x + V \) and \( y + W \) with orthonormal basis \( v, w \), respectively. The scalar product (6) equals

\[
\langle x + V, y + W \rangle_\gamma = \frac{1}{(2\pi)^{D/2}\det^{1/2}(\Sigma)} \exp\left(-\frac{1}{2}\|m^V - m^W\|_\Sigma^2\right),
\]

where \( \hat{\Sigma} = 2\gamma I + \Sigma^V + \Sigma^W \).

Let us observe that the above scalar product generalizes the classical RBF kernel to incomplete data. Indeed, complete data points \( x, y \) are represented by Dirac measures, i.e. \( m^V = x, m^W = y \) and \( \Sigma^V = \Sigma^W = 0 \). Then \( \hat{\Sigma} = 2\gamma I \) and we get a classical RBF kernel with different normalization.

4.3 Kernel function for SVM in \( L^2 \) space

One could probably use the scalar product (6) directly with SVM classifier. Nevertheless, since we work in \( L^2 \) space it might be more convenient to use the kernel function based on such probabilistic representation of missing data point that has a unit norm in \( L^2 \).

We have

\[
\|N(m, \Sigma)\|_{L^2} = \frac{1}{(2\pi)^{D/4}\det^{1/4}(2\Sigma)}.
\]

Therefore, the normalized representation of \( S_x = x + V \) is given by

\[
(2\pi)^{D/4}\det^{1/4}(2\gamma I + 2\Sigma^V) N(m^V, \Sigma^V).
\]

Given two probabilistic representations \( N(m^V, \Sigma^V) \) and \( N(m^W, \Sigma^W) \) of generalized missing data points \( x + V \) and \( y + W \) with orthonormal basis \( v, w \), respectively, we define the following kernel function:

\[
K_{\gamma}(x+V, y+W) = \frac{\det^{1/4}(I + \frac{1}{2}\Sigma^V)\det^{1/4}(I + \frac{1}{2}\Sigma^W)}{\det^{1/2}(I + \frac{1}{2\gamma}(\Sigma^V + \Sigma^W))} \exp\left(-\frac{1}{2}\|m^V - m^W\|_\Sigma^2\right),
\]

where \( \hat{\Sigma} = 2\gamma I + \Sigma^V + \Sigma^W \), \( F = N(m, \Sigma) \) is a density on \( \mathbb{R}^N \) and \( \gamma > 0 \) is fixed.
References

[1] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 26–33. IEEE, 2005.

[2] L. E. Burke, J. M. Dunbar-Jacob, and M. N. Hill. Compliance with cardiovascular disease prevention strategies: a review of the research. Annals of Behavioral Medicine, 19(3):239–263, 1997.

[3] G. Chechik, G. Heitz, G. Elidan, P. Abbeel, and D. Koller. Max-margin classification of data with absent features. Journal of Machine Learning Research, 9:1–21, 2008.

[4] W. M. Czarnecki and J. Tabor. Multithreshold entropy linear classifier: Theory and applications. Expert Systems with Applications, 42(13):5591–5606, 2015.

[5] O. Dekel, O. Shamir, and L. Xiao. Learning to classify with missing and corrupted features. Machine Learning, 81(2):149–178, 2010.

[6] U. Dick, P. Haider, and T. Scheffer. Learning from incomplete data with infinite imputations. In Proceedings of the International Conference on Machine Learning, pages 232–239. ACM, 2008.

[7] Z. Ghahramani and M. I. Jordan. Supervised learning from incomplete data via an EM approach. In Advances in Neural Information Processing Systems, pages 120–127. Citeseer, 1994.

[8] A. Globerson and S. Roweis. Nightmare at test time: robust learning by feature deletion. In Proceedings of the International Conference on Machine Learning, pages 353–360. ACM, 2006.

[9] D. Grangier and I. Melvin. Feature set embedding for incomplete data. In Advances in Neural Information Processing Systems, pages 793–801, 2010.

[10] X. Liao, H. Li, and L. Carin. Quadratically gated mixture of experts for incomplete data classification. In Proceedings of the International Conference on Machine Learning, pages 553–560. ACM, 2007.

[11] R. J. A. Little and D. B. Rubin. Statistical analysis with missing data. John Wiley & Sons, 2014.

[12] P. E. McKnight, K. M. McKnight, S. Sidani, and A. J. Figueredo. Missing data: A gentle introduction. Guilford Press, 2007.
[13] J. L. Schafer. *Analysis of incomplete multivariate data*. CRC Press, 1997.

[14] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, 2006.

[15] A. J. Smola, S. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. Citeseer, 2005.

[16] F. L. Stahura and J. Bajorath. Virtual screening methods that complement HTS. *Combinatorial Chemistry & High Throughput Screening*, 7(4):259–269, 2004.

[17] D. Williams and L. Carin. Analytical kernel matrix completion with incomplete multi-view data. In *Proceedings of the ICML Workshop on Learning With Multiple Views*, 2005.

[18] D. Williams, X. Liao, Y. Xue, and L. Carin. Incomplete-data classification using logistic regression. In *Proceedings of the International Conference on Machine Learning*, pages 972–979. ACM, 2005.