Distance-Preserving Probabilistic Embeddings with Side Information: Variational Bayesian Multidimensional Scaling Gaussian Process

Harold Soh
University of Toronto
harold.soh@utoronto.ca

Abstract

Embeddings or vector representations of objects have been used with remarkable success in various machine learning and AI tasks—from dimensionality reduction and data visualization, to vision and natural language processing. In this work, we seek *probabilistic embeddings* that faithfully represent *observed relationships* between objects (e.g., physical distances, preferences). We derive a novel variational Bayesian variant of multidimensional scaling that (i) provides a posterior distribution over latent points without computationally-heavy Markov chain Monte Carlo (MCMC) sampling, and (ii) can leverage existing *side information* using sparse Gaussian processes (GPs) to learn a nonlinear mapping to the embedding. By partitioning entities, our method naturally handles incomplete side information from multiple domains, e.g., in product recommendation where ratings are available, but not all users and items have associated profiles. Furthermore, the derived approximate bounds can be used to discover the intrinsic dimensionality of the data and limit embedding complexity. We demonstrate the effectiveness of our methods empirically on three synthetic problems and on the real-world tasks of political unfolding analysis and multi-sensor localization.

1 Introduction

Recent achievements in multiple AI domains have been spearheaded by embeddings discovered by models trained on data. For example, word representations have been profitably applied to machine translation [Mikolov *et al.*, 2013] and question-answering [Iyyer *et al.*, 2014]. Embeddings of other entities (e.g., robots, products, and people) are also used to great effect in domains such as sensor localization [Shang *et al.*, 2003], recommender systems [Salakhutdinov and Mnih, 2008], and political analysis [Bakker and Poole, 2013].

In many applications, embeddings appear to require several properties in order to be useful. One of these properties is *distance preservation*—close objects in the original space should be similarly close in the embedding, and far objects similarly distant. Intuitively, this requirement arises when observed distances/similarities have intrinsic meaning (e.g., physical distances, preferences, and word counts) indicating the relationship between entities that we seek to capture.

In this work, we focus on learning *probabilistic embeddings* with this property. Although point-based embeddings are widely used, probabilistic representations offer distinct advantages. For example, uncertainty estimates in localization are essential for autonomous navigation, and Gaussian word embeddings can capture concept ambiguity and asymmetric relationships [Vilnis and McCallum, 2015].

To frame our contribution, we first discuss a general modeling framework—Bayesian Multidimensional Scaling—that encompasses methods that perform pairwise distance/similarity matching. As we will see, popular models such as Probabilistic Matrix Factorization [Salakhutdinov and Mnih, 2008] and word embedding [Hashimoto *et al.*, 2015] can be cast as specialized models within this framework.

Under this umbrella structure, we contribute a specific distance-preserving MDS model that (i) produces probabilistic embeddings under a log-normal likelihood, (ii) utilizes multiple, noisy distance measurements to derive more accurate representations, and (iii) can leverage *side information*. The last feature is particularly significant since side information is available in many domains, e.g., in political surveys, candidates are associated with news articles. However, this extra information is typically unavailable for *all* points (e.g., survey respondents are anonymized) and may have a non-trivial relationship to the embedding. Our model naturally handles such situations by using sparse GPs [Titsias, 2009] on point subsets to induce correlations and learn nonlinear mappings that can project new points to the embedding. These mappings enable "out-of-sample" predictions, e.g., for cold-start recommendations where ratings are not yet available.

To perform inference, we derive efficient approximate variational lower-bounds that allow us to (i) obtain posteriors missing from maximum a posteriori (MAP) solutions, and (ii) perform intrinsic dimensionality selection to limit embedding complexity. Empirical results on synthetic datasets and two real-world tasks demonstrate that our method produces useful embeddings and learns mappings at reasonable cost. In particular, our model was able to project political candidates absent from survey data to a coherent layout using side information (Wikipedia entries), and in a localization task, to pin-point new sensors quickly using beacon signals.

1.1 Multidimensional Scaling: A Brief Review

MDS encompasses a range of techniques that constrain embeddings to have pair-wise distances as close as possible to the data space. MDS has had a rich history, from its early beginnings in 1930s psychometrics at Chicago and Princeton [Shepard, 1980] to recent probabilistic incarnations [Bakker and Poole, 2013]. What is now known as classical metric MDS was developed by [Torgerson, 1952], and has since been extended to handle non-metric spaces and non-linear projections, yielding techniques such as Sammon mapping [Sammon, 1969], Isomap [Tenenbaum *et al.*, 2000] and SNE [Hinton and Roweis, 2002]. MDS has become a significant research area and we refer interested readers to [Borg and Groenen, 2005] for a more comprehensive treatment.

In brief, the essence of MDS lies in the minimization of loss functions (historically called stress or strain), defined on pairwise distances d_{ij}. The loss function and distances can be varied to induce different lower-dimensional representations. In classical metric MDS (CMDS), the distances are Euclidean $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ for data elements $\mathbf{x}_i \in X$ with the loss function being the residual sum of squares:

$$\mathcal{L}_{\text{CMDS}} = \sum_{i,j} (d_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\|)^2 \tag{1}$$

where $\mathbf{z}_i \in Z$ are the embedded points. Note that solving this minimization problem is equivalent to performing principal components analysis (PCA); the minimum configuration is the eigen-decomposition of the Gram matrix $\mathbf{X}\mathbf{X}^\top$.

In Sammon mapping, the distances remain Euclidean but the loss function is altered:

$$\mathcal{L}_{\text{Sammon}} = \frac{1}{\sum_{i<j} d_{ij}} \sum_{i<j} \frac{(d_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\|)^2}{d_{ij}}, \tag{2}$$

which causes small distances to be emphasized. In Isomap, the canonical stress function is used but the d_{ij}'s are geodesic distances, which are estimated using the shortest-path distances on a neighborhood graph where each point $\mathbf{x}_i$ is connected to its k nearest neighbors.

2 Probabilistic Bayesian MDS Framework

Classic MDS methods, while effective, do not provide probabilistic embeddings, but as will see, can be extended with the Bayesian framework. Recently, [Bakker and Poole, 2013] presented a Bayesian metric MDS model with MCMC inference. We take this approach one-step further and discuss how this perspective relates to a broader class of models.

From a probabilistic viewpoint, our primary problem is finding latent coordinates $\mathbf{z}_i \in Z$ given observed relationships between entities, $\mathcal{D} = \{o_k = (i_k, j_k, d_k)\}$. Each observation o_k comprises point labels, i_k and j_k, and the measurement d_k. Unlike the classical MDS setting, we do *not* assume access to the data space (i.e., node features from which the distances are computed). Furthermore, the dataset need not contain all pairs and there can be multiple (noisy) observations for any two points. Conventionally, we work with distances, but the measurement d_k can also reflect attraction/similarity.

Let $N = |\mathcal{D}|$ and to simplify notation, we drop the k subscript from the point labels. Adopting the Bayesian paradigm and assuming independent observational noise,

$$p(Z|\mathcal{D}) \propto p(\mathcal{D}|Z)p(Z) = \prod_{k=1}^{N} p(d_k|f_k = f_z(\mathbf{z}_i, \mathbf{z}_j))p(Z)$$

where we see the main elements comprise the likelihood $p(d_k|f_k)$, the pairwise function $f_z(\mathbf{z}_i, \mathbf{z}_j)$, and prior $p(Z)$. Applying this Bayesian MDS framework to a particular domain requires specification of these three ingredients, which control the qualitative and quantitative aspects of the embedding, and leads to contrasting models. For example,

- A Gaussian likelihood $p(d_k|f_k) = \prod_{k=1}^{N} \mathcal{N}(d_k|\hat{\mathbf{z}}_k, \sigma_n^2)$ with priors $p(Z) = \prod_i^N \mathcal{N}(0, \sigma_i^2)$, and a Euclidean distance link function, $f_k = f_z(\mathbf{z}_i, \mathbf{z}_j) = \|\mathbf{z}_i - \mathbf{z}_j\|$ defines a probabilistic version of classical metric MDS;
- If the measurements d_k are ratings and we change the Gaussian model above slightly by letting $f_z(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^\top \mathbf{z}_j$ (a similarity function), we recover the popular Probabilistic Matrix Factorization (PMF) model [Salakhutdinov and Mnih, 2008] for collaborative filtering;
- A negative binomial likelihood, $p(d_k|\mathbf{z}_k) = \prod_{k=1}^{N} \text{NegBin}(d_k|\theta, \theta f_k^{-1})$ with link function $f_z = \exp(-\|\mathbf{z}_i - \mathbf{z}_j\|^2/2)$, specifies co-occurance metric recovery [Hashimoto *et al.*, 2015], which is intimately linked to the successful word embedding method GLoVe [Pennington *et al.*, 2014].

Under the probabilistic MDS framework, we can interpret the above models to be performing distance/similarity matching under an assumed noise distribution. Moreover, we can compose new models in this family by selecting an appropriate likelihood, pairwise function and priors. If additional parameters are required, priors (e.g., GPs) can be placed over these variables. Applying the relevant Bayesian machinery allows us to perform inference to obtain the latent distributions, and to control model complexity, which can impact generalizability and predictive performance.

3 Log-Normal Distance-Preserving MDS

In this section, we specify a Bayesian MDS model that forms the foundation for our GP-model in Section 4. Similar to the above models, we use Gaussian priors $p(Z) = \prod_i^{|Z|} \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0 \mathbf{I})$. The Gaussian likelihood is popular due to its tractability, but is inappropriate for distance measurements. Thus, we specify the less conventional, but more realistic, log-normal, as advocated by [Bakker and Poole, 2013],

$$p(d_k|f_k) = \frac{1}{d_k \sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(\log d_k - f_k)^2}{2\sigma_n^2}\right) \tag{3}$$

and a log Euclidean distance function,

$$f_k = f_z(\mathbf{z}_i, \mathbf{z}_j) = \log \|\mathbf{z}_i - \mathbf{z}_j\| = \frac{1}{2} \log \|\mathbf{z}_i - \mathbf{z}_j\|^2. \tag{4}$$

3.1 Variational Inference for Log-Normal MDS

Variational Bayes is a "middle-ground" approach that allows us to obtain confidence estimates missing from the MAP solution, but at a much lower computational cost compared to MCMC. It transforms approximate inference into an optimization problem, and our goal will be to derive a lower bound (the objective function to maximize).

Obtaining the lower bound for the log-normal Bayesian MDS model is challenging due to its expression, which contains nonlinear transformations of random variables. Here, we provide a step-by-step derivation of an approximate lower bound and demonstrate practical techniques that should prove useful in the development of future variational Bayesian MDS models. To begin, we adopt a mean-field variational approximation:

$$q(Z) = \prod_{i}^{|Z|} q(\mathbf{z}_i) = \prod_{i}^{|Z|} \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{5}$$

With the above factorization, the variational lower-bound is:

$$\begin{aligned}\mathcal{L}_1(q) = &\int q(Z) \log \frac{p(\mathcal{D}|Z)p(Z)}{q(Z)} dZ \\ =&\mathbb{E}[\log p(\mathcal{D}|Z)] - \mathbb{D}_{\mathrm{KL}}[q(Z)\|p(Z)]\end{aligned} \tag{6}$$

We see that $\mathcal{L}_1$ consists of two components: the expectation of the log-likelihood under the variational distribution and the negative Kullback-Leibler (KL) divergence between the prior and variational distributions for Z (multivariate normals).

Computing $\mathbb{E}[\log p(\mathcal{D}|Z)]$: We first simplify the expression by introducing $\hat{\mathbf{z}}_k = \mathbf{z}_i - \mathbf{z}_j$. Since both $\mathbf{z}_i$ and $\mathbf{z}_j$ are normal, $\hat{\mathbf{z}}_k \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_k = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j, \hat{\boldsymbol{\Sigma}}_k = \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)$,

$$\begin{aligned}\mathbb{E}[\log p(\mathcal{D}|Z)] = &-\sum_{k=1}^{N} \log d_k - \frac{N}{2}\log 2\pi\sigma_n^2 - \\ &\frac{1}{2\sigma_n^2}\sum_{k=1}^{N} \mathbb{E}\left[\left(\log d_k - \frac{1}{2}\log\|\hat{\mathbf{z}}_k\|^2\right)^2\right]\end{aligned} \tag{7}$$

where the first two terms on the RHS are readily obtained, but the third requires additional effort— the unresolved expectation lacks a closed-form expression and typical solutions involve deriving an analytical approximation, or numerical estimation. In this work, we derive a second-order Taylor approximation,

$$\mathbb{E}\Big[g_k(\|\hat{\mathbf{z}}_k\|^2)\Big] \approx g_k(\mathbb{E}[\|\hat{\mathbf{z}}_k\|^2]) + \frac{g_k''(\mathbb{E}[\|\hat{\mathbf{z}}_k\|^2])\mathbb{V}[\|\hat{\mathbf{z}}_k\|^2]}{2} \tag{8}$$

where $g_k(x) = (\log d_k - \frac{1}{2}\log x)^2$. To obtain the necessary moments, the key "trick" is to recognize $\|\hat{\mathbf{z}}_k\|^2$ as a quadratic form of $\hat{\mathbf{z}}_k$ and introduce the random variable $y_k = \|\hat{\mathbf{z}}_k\|^2 = \hat{\mathbf{z}}_k^\top \hat{\mathbf{z}}_k$. Although the distribution for y_k is complex, it has a tractable moment generating function $M(t)$ [Mathai and Provost, 1992]:

$$M(t) = \exp\left(t\sum_{l=1}^{s} a_l^2\lambda_l(1-2t\lambda_l)^{-1}\right)\prod_{l=1}^{s}(1-2t\lambda_l)^{-\frac{1}{2}}$$

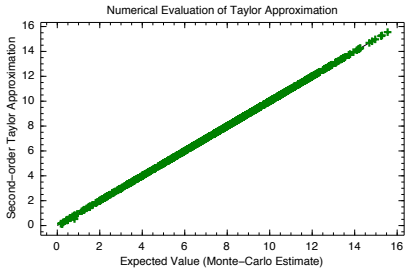


Figure 1: Comparison between approximate expected value and numerical expectations (5000 Monte-Carlo samples per estimate). The approximate expectations (green +'s) fall on the ideal outcome (solid diagonal black line $y = x$) with low mean squared error of 1.26×10^{-4}.

where the λ_l's are the eigenvalues of $\hat{\boldsymbol{\Sigma}}_k$ and $\mathbf{a} = \hat{\boldsymbol{\Sigma}}_k^{-\frac{1}{2}}\hat{\boldsymbol{\mu}}_k$. The corresponding derivatives of $M(t)$ at $t = 0$ yield simple equations for the mean and variance,

$$\mathbb{E}[y_k] = \hat{\boldsymbol{\mu}}_k^\top\hat{\boldsymbol{\mu}}_k + \mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_k) \tag{9}$$

$$\mathbb{V}[y_k] = 4\hat{\boldsymbol{\mu}}_k^\top\hat{\boldsymbol{\Sigma}}_k\hat{\boldsymbol{\mu}}_k + 2\,\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_k^2) \tag{10}$$

and completes our approximation. Numerical experiments (where $\hat{\mathbf{z}}_k$ was randomly sampled across a wide range of values) showed the approximation to be fast and to have low-error (Fig. 1). If higher precision is required, more costly numerical methods such as quadrature can be applied.

Combining the aforementioned elements and separating out the constants leads to the objective function $\mathcal{L}_2 \approx \mathcal{L}_1$,

$$\begin{aligned}\mathcal{L}_2 = &-\frac{N}{2}\log\sigma_n^2 - \frac{1}{2\sigma_n^2}\sum_{k=1}^{N}\left(\log d_k - \frac{\log\mathbb{E}[y_k]}{2}\right)^2 + \\ &\frac{\mathbb{V}[y_k]}{2\mathbb{E}[y_k]^2}\left(2\log d_k - \log\mathbb{E}[y_k] + 1\right) - \mathbb{D}_{\mathrm{KL}} + \text{const}\end{aligned} \tag{11}$$

Maximizing $\mathcal{L}_2$ (or equivalently, minimizing $-\mathcal{L}_2$) using an off-the-shelf optimizer gives us the variational posterior over the latent coordinates Z, and can also be used to estimate the intrinsic dimensionality of the embedding (demonstrated in the experiments). For large distance datasets or on-line settings, $\mathcal{L}_2$ can be optimized using stochastic gradient ascent.

4 Using Side Information with Sparse GPs

At this point, our variational Bayesian MDS (VBMDS) model represents each point as an individual distribution, which may prove troublesome for large datasets. It also lacks a mapping function—given new data points, we would have to re-optimize $\mathcal{L}_2$ to obtain the latent coordinates. In this section, we extend our model using sparse GPs that can leverage on any existing side information to learn a non-linear mapping of points to latent coordinates. This enhancement can also lead to a more compact model, since some points are presented indirectly.

Consider that Z is split into two mutually exclusive sets, $Z = Z^p \cup Z^x$. Without loss of generality, assume that only

points in $\mathbf{z}_i^x \in Z^x$ possess side information $\mathbf{x}_i \in \mathcal{X}$[1]. Employing the variational inducing input scheme proposed by [Titsias, 2009] and [Hensman *et al.*, 2013], we introduce a set of r inducing variables $U = \{\mathbf{u}_1, \dots \mathbf{u}_r\}$ (one for each embedding dimension), and specify new priors,

$$p(z_{il}^x|\mathbf{u}_l) = \mathcal{N}(\mathbf{k}_i^\top \mathbf{K}_l^{-1}\mathbf{u}_l, k_{ii} - \mathbf{k}_i^\top \mathbf{K}_l^{-1}\mathbf{k}_i) \quad (12)$$

$$p(\mathbf{u}_l) = \mathcal{N}(\mathbf{0}, \mathbf{K}_l) \quad (13)$$

where $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}_i = [k(\mathbf{x}_i, \mathbf{x}_j)]_{j=1}^m$ and $\mathbf{K}_l^{-1} = [k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^m$. The inducing inputs introduce conditional independencies between the latent function variables. As before, we use a mean-field variational approximation,

$$\begin{aligned} q(Z^p, Z^x, U) &= q(Z^p)p(Z^x|U)q(U) \\ &= \prod_j^{|Z^p|} q(\mathbf{z}_j^p) \prod_i^{|Z^x|} \prod_l^r p(z_{il}^x|\mathbf{u}_l) \prod_l^r q(\mathbf{u}_l) \end{aligned} \quad (14)$$

where $q(\mathbf{u}_l) = \mathcal{N}(\mathbf{m}_l, \mathbf{S}_l)$ and $q(\mathbf{z}_j) = \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. The associated variational lower-bound is

$$\begin{aligned} &\mathcal{L}_3(q) = \\ &\iiint q(Z^p, Z^x, U) \log \frac{p(\mathcal{D}|Z^p, Z^x)p(U)}{q(Z^p)q(U)} dZ^p dZ^x dU \\ &= \mathbb{E}[\log p(\mathcal{D}|Z^p, Z^x)] - \mathbb{D}_{\mathrm{KL}}[q(U)\|p(U)] - \\ &\quad \mathbb{D}_{\mathrm{KL}}[q(Z^p)\|p(Z^p)] \end{aligned} \quad (15)$$

where $p(Z|U)$ inside the logarithm cancels out. The sparse GPs are assumed independent across the latent dimensions, which facilitates the derivation of the lower bound, and we have new variational parameters parameters $\mathbf{m}_l$ and $\mathbf{S}_l$ associated with the inducing variables (in addition to any GP kernel hyperparameters).

Again, the principal challenge is in computing the expectation of the likelihood, $\mathbb{E}[p(\mathcal{D}|Z^p, Z^x)]$. Fortunately, we can re-use the derivations in the previous section, except that we now compute the moments $\mathbb{E}[y_k]$ and $\mathbb{V}[y_k]$ with respect to $q(Z^p)p(Z^x|U)q(U)$ instead of $q(Z)$. Specifically, this approach results in four possible cases, each requiring expectations over different subsets of variables:

$$\mathbb{E}[y_k] = \begin{cases} \mathbb{E}_{z^p}[y_k] & \text{Case 1: } f_z(\mathbf{z}_i^p, \mathbf{z}_j^p) \\ \mathbb{E}_{z^x,u}[y_k] & \text{Case 2: } f_z(\mathbf{z}_i^x, z_j^x) \\ \mathbb{E}_{z^x,z^p,u}[y_k] & \text{Case 3: } f_z(\mathbf{z}_i^x, \mathbf{z}_j^p) \\ \mathbb{E}_{z^x,z^p,u}[y_k] & \text{Case 4: } f_z(\mathbf{z}_i^p, \mathbf{z}_j^x) \end{cases} \quad (16)$$

and similarly for the variances. **Case 1** arises when both points are in Z^p and thus, the moments are those presented in the previous section, i.e., eqns. (9) and (10).

Case 2: Here, both points have side information and are represented by the GPs. Starting from (9),

$$\mathbb{E}_{z^x,u}[y_k] = \mathbb{E}_u\left[\hat{\boldsymbol{\mu}}_k^\top \hat{\boldsymbol{\mu}}_k\right] + \mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_k)$$

with $\hat{\boldsymbol{\Sigma}}_k = \mathrm{Diag}([\hat{\sigma}_{kl}^2]_{l=1}^r))$ and $\hat{\sigma}_{kl}^2 = k_{ii} - \mathbf{k}_i^\top \mathbf{K}_l^{-1}\mathbf{k}_i + k_{jj} - \mathbf{k}_j^\top \mathbf{K}_l^{-1}\mathbf{k}_j$. It turns out that $\hat{\boldsymbol{\mu}}_k = [\hat{\mu}_{kl} = (\mu_{il} - \mu_{jl})]_{l=1}^r$, is normally distributed; recall that $\mu_{il} = \mathbf{k}_i^\top \mathbf{K}_l^{-1}\mathbf{u}_l$ and hence,

$$\hat{\mu}_{kl} = \mathbf{k}_i^\top \mathbf{K}_l^{-1}\mathbf{u}_l - \mathbf{k}_j^\top \mathbf{K}_l^{-1}\mathbf{u}_l = \mathbf{b}_{kl}^\top \mathbf{u}_l \quad (17)$$

where $\mathbf{b}_{kl}^\top = (\mathbf{k}_i - \mathbf{k}_j)^\top \mathbf{K}_l^{-1}$. Since each $\mathbf{u}_l$ is Gaussian, $\hat{\boldsymbol{\mu}}_k$ is also a Gaussian specified by,

$$\hat{\boldsymbol{\mu}}_k = \mathcal{N}\left(\hat{\boldsymbol{\beta}}_k = \left[\mathbf{b}_{kl}^\top \mathbf{m}_l\right]_{l=1}^r, \hat{\boldsymbol{\Lambda}}_k = \mathrm{Diag}\left(\left[\mathbf{b}_{kl}^\top \mathbf{S}_l \mathbf{b}_{kl}\right]_{l=1}^r\right)\right)$$

As in the previous section, the moments of the quadratic form $\hat{\boldsymbol{\mu}}_k^\top \hat{\boldsymbol{\mu}}_k$ are easily obtained via $M(t)$,

$$\mathbb{E}_u\left[\hat{\boldsymbol{\mu}}_k^\top \hat{\boldsymbol{\mu}}_k\right] = \hat{\boldsymbol{\beta}}_k^\top \hat{\boldsymbol{\beta}}_k + \mathrm{Tr}(\hat{\boldsymbol{\Lambda}}_k) \quad (18)$$

$$\mathbb{V}_u\left[\hat{\boldsymbol{\mu}}_k^\top \hat{\boldsymbol{\mu}}_k\right] = 4\hat{\boldsymbol{\beta}}_k^\top \hat{\boldsymbol{\Lambda}}_k \hat{\boldsymbol{\beta}}_k + 2\,\mathrm{Tr}(\hat{\boldsymbol{\Lambda}}_k^2) \quad (19)$$

The first moment is substituted into (17) to give,

$$\mathbb{E}_{z^x,u}[y_k] = \hat{\boldsymbol{\beta}}_k^\top \hat{\boldsymbol{\beta}}_k + \mathrm{Tr}(\hat{\boldsymbol{\Lambda}}_k) + \mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_k) \quad (20)$$

Turning our attention to the variance and using same reasoning as above (with some algebraic manipulation),

$$\begin{aligned} \mathbb{V}_{z^x,u}[y_k] &= \mathbb{E}_u\left[4\hat{\boldsymbol{\mu}}_k^\top \hat{\boldsymbol{\Sigma}}_k \hat{\boldsymbol{\mu}}_k\right] + 2\,\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_k^2) + \mathbb{V}_u\left[\hat{\boldsymbol{\mu}}_k^\top \hat{\boldsymbol{\mu}}_k\right] \\ &= 4\hat{\boldsymbol{\beta}}_k^\top (\hat{\boldsymbol{\Sigma}}_k \hat{\boldsymbol{\Lambda}}_k \hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\Lambda}}_k)\hat{\boldsymbol{\beta}}_k + 2\,\mathrm{Tr}((\hat{\boldsymbol{\Sigma}}_k \hat{\boldsymbol{\Lambda}}_k)^2 + \hat{\boldsymbol{\Sigma}}_k^2 + \hat{\boldsymbol{\Lambda}}_k^2) \end{aligned} \quad (21)$$

Cases 3 and 4: The third and fourth cases arise when the points come from each set Z^x and Z^p, and can be treated similarly. We build upon Case 2 by noting that

$$\hat{\mathbf{z}}_k \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_k = [\mathbf{b}_{kl}^\top \mathbf{u}_l]_{l=1}^r - \boldsymbol{\mu}_j, \hat{\boldsymbol{\Sigma}}_k = \mathrm{Diag}[\hat{\sigma}_{il}^2]_{l=1}^r + \boldsymbol{\Sigma}_j)$$

where $\hat{\sigma}_{il}^2 = k_{ii} - \mathbf{k}_i^\top \mathbf{K}_l^{-1}\mathbf{k}_i$. This leads to a slightly different Gaussian distribution for $\hat{\boldsymbol{\mu}}_k = \mathcal{N}(\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\Lambda}}_k)$ where

$$\hat{\boldsymbol{\beta}}_k = \left[\mathbf{b}_{kl}^\top \mathbf{m}_l\right]_{l=1}^r - \boldsymbol{\mu}_j \quad \text{and} \quad \hat{\boldsymbol{\Lambda}}_k = \mathrm{Diag}\left[\mathbf{b}_{kl}^\top \mathbf{S}_l \mathbf{b}_{kl}\right]_{l=1}^r \quad (22)$$

In the case that $\hat{\mathbf{z}}_k = \mathbf{z}_i^p - \mathbf{z}_j^x$, the signs in the mean are flipped, $\hat{\boldsymbol{\beta}}_k = \boldsymbol{\mu}_j - \left[\mathbf{b}_{kl}^\top \mathbf{m}_l\right]_{l=1}^r$. The above equations for $\hat{\boldsymbol{\beta}}_k$ and $\hat{\boldsymbol{\Lambda}}_k$ (along with $\hat{\boldsymbol{\Sigma}}_k$) can be plugged into (20) and (21) derived in Case 2 to give the required central moments of y_k.

To complete $\mathcal{L}_4 \approx \mathcal{L}_3$, we replace the moments $\mathbb{E}[y_k]$ and $\mathbb{V}[y_k]$ in $\mathcal{L}_2$ (11), depending on the cases encountered (16). This extended model, the VBMDS-GP, completely generalizes the VBMDS since either set Z^x and Z^p can be empty. Distinct from $\mathcal{L}_2$, there are no separate distributions for elements of Z^x. Instead, the latent coordinates are represented indirectly using the sparse GPs, which enables prediction of latent coordinates using side information.

4.1 Relationship to other models

The VBMDS-GP builds upon a wide body of work and finds connections to a variety of models that also perform distance matching. For example, it is related to Neuroscale [Lowe and

[1]Extensions to more than two subsets (e.g., when points come from multiple domains) is straightforward, but we restrict our description to a two-subset model for expositional simplicity.

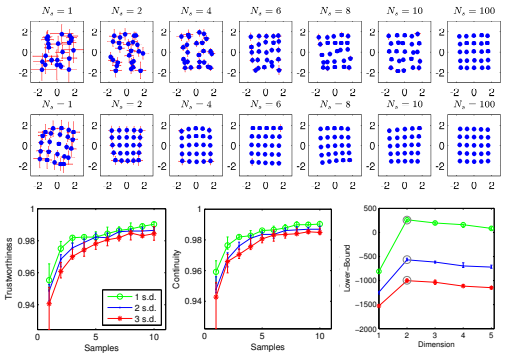
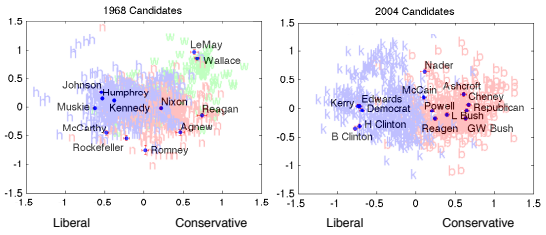


Figure 2: 2D-lattice embeddings obtained by individual distributions (top) and using the sparse GPs (middle). The point variances were initially large, reflecting the embedding impreciseness, but gradually decreased with more observations. (bottom) Both scores improved with the number of samples across the three noise levels, and the maximum $\mathcal{L}_4$ value matched the intrinsic dimensionality of the data $r = 2$.



Data-set	Comp. Time (s) VBMDS	MCMC	k-NN Error VBMDS	MCMC
1968	254.74	4371.42	0.231	0.261
2004	222.07	4888.60	0.085	0.075

Figure 3: Political unfolding VBMDS-GP configurations for the 1968 (left) and 2004 (right) elections are visually similar to MCMC; the political figures are lined up along their main political philosophies and surrounded by their respective voters. (Bottom table) 10-NN errors are comparable to MCMC, but with up to a 22-fold speedup.

Tipping, 1997], that optimizes a RBF network using MDS stress. But unlike Neuroscale, the VBMDS-GP is probabilistic in nature, giving uncertainties over latent coordinates.

As previously mentioned, our Bayesian MDS framework is related to PMF, which has a variational version [Lim and Teh, 2007]. Similar to the VBMDS-GP, Kernel Probabilistic Matrix Factorization (KPMF) [Zhou *et al.*, 2012], assumes a GP prior over the columns of the matrices (latent dimensions), which permits for the inclusion of side information. In addition to the difference in the likelihood and choice of f_z, KPMF uses MAP inference, whilst we provide a variational solution with a sparse GP model. Our model is also related to probabilistic *generative* dimensionality-reduction models such as GPLVM [Lawrence, 2005]. Modern variational GPLVM variants also use sparse GPs [Titsias and Lawrence, 2010] but attempt to find probabilistic embeddings that reproduce high-dimensional *observed features* (or matrices)—the GPLVM mapping is in the opposite direction $Z \rightarrow X$ compared to VBMDS-GP, and does not use side information, nor constrain distances unless back constraints are applied [Lawrence and Quinonero-Candela, 2006].

5 Experiments

In this section, we present empirical results, beginning with synthetic datasets to validate the approach, followed by two applications in political unfolding analysis and multi-sensor localization. VBMDS-GP source code is available at https://github.com/haroldsoh/vbmds.

5.1 Embeddings with Noisy Distances

In this first experiment, we validated our model by examining the embeddings generated given noisy distance observations. We used the 2D-lattice (5×5 grid), 3D-lattice ($5 \times 5 \times 5$ grid) and Oilflow (100 points, 12 dimensions) datasets. The 2D and 3D lattices are simple "toy" datasets with a well-defined shape—which simplified visual comparison—while Oilflow is a widely-used benchmark. The distances were corrupted with three log-normal noise levels corresponding to one, two and three times the standard deviation of the inter-point distance distribution. As quantitative measures, we use Trustworthiness and Continuity scores [Venna and Kaski, 2006].

Fig. 2 shows sample embeddings for the 2D-lattice under moderately high noise as more observations were provided. Note the high embedding uncertainty when $N = 1$, suggesting the embedding is imprecise and inaccurate. As N increased, embedding uncertainty decreased and the layout became more accurate, as reflected in the higher Trustworthiness and Continuity scores. Very similar plots were obtained for 3D-lattice and Oilflow (data not shown due to space constraints). If no artificial noise is introduced on the Oilflow dataset, the scores obtained by VBMDS-GP (99.92, 99.94) are better than variational-GPLVM (99.72, 99.86), indicating the effect of distance preservation as an objective. Lastly, the highest values for $\mathcal{L}_4$ correspond to the true intrinsic dimensionality of the datasets; two and three for the 2D and 3D lattices respectively, and three/four for oilflow, in agreement with previous studies [Titsias and Lawrence, 2010]. This result motivates its use in selecting embedding complexity.

5.2 Political Unfolding Analysis

Next, we applied VBMDS-GP to political unfolding where there are two sets of points—one set of respondents and one set of stimuli (political figures)—with distance information only available between *inter*-set points. We used thermometer datasets—subjective preferences on a scale from zero to a hundred—for the 1968 (20,076 observations) and 2004 (15,644 observations) US elections.

The configurations obtained (Fig. 3) are visually similar to those obtained via slice-sampling [Bakker and Poole, 2013]; the candidates spread along a main axis of their political philosophies, with liberals on one end and conservatives on the other. The uncertainties give us additional indications of the "spread" along these axes. The optimized $\mathcal{L}_4$ values indicated the datasets are best described by two dimensional embeddings. As a quantitative comparison, the nearest-neighbor errors for the respondents (vote choices) are

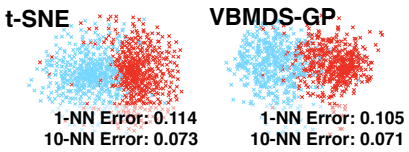


Figure 4: VBMDS-GP achieved good cluster separation, with slightly lower k-NN errors compared to the t-SNE embedding, which was more visually symmetrical.

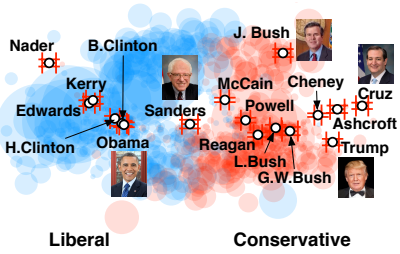


Figure 5: Projected individuals to the embedding with the VBMDS-GP model using side information (BoW features from Wikipedia entries). Respondents size indicate (scaled-up) variances. Portraits shown for political figures *not* in the 2004 survey and have no associated thermometer data. The projections are coherent with the overall mapping structure.

similar (error $\approx 1-3\%$), but MCMC (a C program provided by the authors) required more than 70 mins, while VBMDS-GP took only 3-4 minutes (10,000 iterations in MATLAB).

To investigate cluster separation, we transformed the distances into probabilities via $\exp(-d_{ij}/\sigma_p^2)$ with $\sigma_p^2 = 30$ for comparison to t-SNE [Van der Maaten and Hinton, 2008]. Under this transformation, both methods produced comparable embeddings (Fig. 4); t-SNE's embedding was visually more symmetrical, but VBMDS-GP achieved lower k-NN errors. Unlike t-SNE, VBMDS-GP also produced point uncertainties, which when used to compute the 1-NN error (via expected distance) gave a lower error of 0.098.

We constructed side information for each political figure consisting of PCA-reduced Bag-of-Words (BoW) features from Wikipedia entries, and used the VBMDS-GP to embed politicians that were *not included* in the 2004 survey. Although simple BoW features may only weakly indicate opinions on key issues, Fig. 5 shows the resultant embedding to be remarkably coherent: President Obama and Senator Sanders are close to their liberal base, while Jeb Bush, Senator Cruz, and Donald Trump are projected to the conservative region.

5.3 Multi-sensor Localization

The problem of multi-sensor localization arises in healthcare/environment monitoring where wireless ad-hoc sensor networks are deployed in a target areas without GPS (that can be too power-hungry, and perform poorly indoors). We applied VBMDS-GP to the Cricket dataset [Moore *et al.*, 2004], which comprises range readings from real-world Crickets—hardware platforms with ultrasonic transmitter and receivers. We used the ranging dataset as provided (5 clusters with missing and repeated distance readings) but filtered-out clearly erroneous readings, leaving a total of 35,312 distances. For side information, eight beacon nodes were associated with a signal propagation model $s_{bi} = \log \frac{A_b}{(1000+d_{bi})^{\alpha_b}}$ where A_b is the raw signal strength for beacon b, d_{bi} is the distance between the nodes, and α_b is the decay. These parameters were varied across the beacons and assumed unknown.

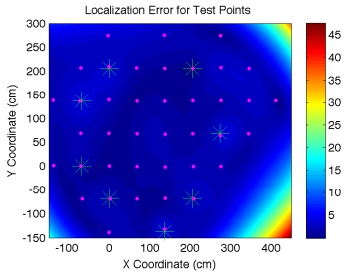


Figure 6: Localized positions of nodes. Beacons are shown as stars and the original sensor nodes shown as magenta dots.

The localization error achieved by VBMDS-GP was 3.1 cm (no sparsity) and 3.19 cm (GPs using 20 inducing inputs), compared to 6.73 cm using robust quadrilaterals [Moore *et al.*, 2004] and 6.63 cm for MDS-MAP [Shang *et al.*, 2003]. Combining all the clusters and training with VBMDS-GP leads to a lower error of 2.21 cm, which we posit was due to the elimination of the cluster stitching which tends to increase error [Whitehouse and Culler, 2006]. To further evaluate the learnt mapping, we localized $\approx$ 68k nodes in the target area only using the beacon features (Fig. 6). The average error was 3.47 cm; near the original sensor nodes, the error < 3 cm, with greater error and predictive uncertainty further away from the inducing inputs. The time required to localize all 68k nodes was 2.52s (3.7×10^{-5}s per prediction). As such, the VBMDS-GP can be used for new or mobile nodes; once a mapping is obtained, precise localization with uncertainty estimation can be performed at low computational cost.

6 Summary and Conclusion

This paper presented a computationally efficient variational Bayesian MDS model that leverages upon side information and produces distance-preserving probabilistic embeddings. In the political unfolding task, VBMDS-GP produced embeddings comparable to MCMC at a fraction of the computational cost, and projected novel political candidates coherently to the embedding using only side information. Positive results were also observed for the localization task where the VBMDS-GP errors were half that of competing methods.

Given appropriate distance data, VBMDS-GP is applicable to a variety of tasks with side information, e.g., collaborative filtering with product description and user profiles, social network visualization with node attributes, and document/image clustering with meta-data. In general, we expect VBMDS-GP to perform well when the distance observations (plus discrepancy due to dimensionality difference) are approximately log-normal. As future work, alternative likelihoods can be used within the Bayesian MDS framework to better suit other domains and to induce embeddings of different flavors.

Acknowledgments

Thank you to Scott Sanner for his helpful and constructive comments.

References

[Bakker and Poole, 2013] Ryan Bakker and Keith T Poole. Bayesian metric multidimensional scaling. *Political Analysis*, 21(1):125–140, 2013.

[Borg and Groenen, 2005] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[Hashimoto *et al.*, 2015] Tatsunori B. Hashimoto, David Alvarez-Melis, and Tommi S. Jaakkola. Word, graph and manifold embedding from Markov processes. In *NIPS 2015 Workshop on Nonparametric Methods for Large Scale Representation Learning*, 2015.

[Hensman *et al.*, 2013] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian Processes for Big Data. In *Uncertainty in Artificial Intelligence (UAI-13)*, 2013.

[Hinton and Roweis, 2002] Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *NIPS*, pages 833–840, 2002.

[Iyyer *et al.*, 2014] Mohit Iyyer, Jordan L Boyd-Graber, Leonardo Max Batista Claudino, Richard Socher, and Hal Daumé III. A neural network for factoid question answering over paragraphs. In *EMNLP*, pages 633–644, 2014.

[Lawrence and Quinonero-Candela, 2006] Neil D Lawrence and Joaquin Quinonero-Candela. Local distance preservation in the GP-LVM through back constraints. In *Proc. of the 23rd Intl. Conf on Machine learning*, pages 513–520, 2006.

[Lawrence, 2005] Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *The Journal of Machine Learning Research*, 6:1783–1816, 2005.

[Lim and Teh, 2007] Yew Jin Lim and Yee Whye Teh. Variational Bayesian Approach to Movie Rating Prediction. *Proceedings of KDD Cup and Workshop*, pages 15–21, 2007.

[Lowe and Tipping, 1997] David Lowe and Michael E Tipping. Neuroscale: novel topographic feature extraction using RBF networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 543–549, 1997.

[Mathai and Provost, 1992] Arakaparampil M Mathai and Serge B Provost. Quadratic forms in random variables: theory and applications. *Journal of the American Statistical Association*, 1992.

[Mikolov *et al.*, 2013] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[Moore *et al.*, 2004] David Moore, John Leonard, Daniela Rus, and Seth Teller. Robust distributed network localization with noisy range measurements. In *Proc. of the 2nd Intl. Conf. on Embedded Networked Sensor Systems*, pages 50–61. ACM, 2004.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[Salakhutdinov and Mnih, 2008] R Salakhutdinov and A Mnih. Probabilistic Matrix Factorization. *NIPS*, pages 1257–1264, 2008.

[Sammon, 1969] John W Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans. on Computers*, 18(5):401–409, 1969.

[Shang *et al.*, 2003] Yi Shang, Wheeler Ruml, Ying Zhang, and Markus PJ Fromherz. Localization from mere connectivity. In *Proc. of the 4th ACM Int. Symp. on Mobile ad hoc networking & computing*, pages 201–212. ACM, 2003.

[Shepard, 1980] Roger N Shepard. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398, 1980.

[Tenenbaum *et al.*, 2000] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[Titsias and Lawrence, 2010] Michalis K Titsias and Neil D Lawrence. Bayesian Gaussian Process Latent Variable Model. In *AISTATS*, pages 844–851, 2010.

[Titsias, 2009] Michalis Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *AISTATS*, pages 567–579, 2009.

[Torgerson, 1952] Warren S Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419, 1952.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *JMLR*, 9(2579-2605):85, 2008.

[Venna and Kaski, 2006] Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19(6):889–899, 2006.

[Vilnis and McCallum, 2015] Luke Vilnis and Andrew McCallum. Word Representations via Gaussian Embedding. In *ICLR*, 2015.

[Whitehouse and Culler, 2006] Kamin Whitehouse and David Culler. A robustness analysis of multi-hop ranging-based localization approximations. In *Proc. of the 5th Intl. Conf. on Information Processing in Sensor Networks*, pages 317–325. ACM, 2006.

[Zhou *et al.*, 2012] Tinghui Zhou, H Shan, A Banerjee, and Guillermo Sapiro. Kernelized Probabilistic Matrix Factorization: Exploiting Graphs and Side Information. In *SIAM International Conference on Data Mining (SDM)*, pages 914–925, 2012.